First-Order Algorithm for Content-Centric Sparse Multicast Beamforming in Large-Scale C-RAN

Yang Li, *Student Member, IEEE*, Minghua Xia, *Member, IEEE*, and Yik-Chung Wu, *Senior Member, IEEE*

***Abstract*— In multimedia-rich communication scenarios, popular contents are requested by many users. This calls for the communication system design perspective transferring from user-centric to content-centric. To realize the content-centric paradigm, one of the dominant approaches is the multi-group multicast transmission. However, different content groups may cause interference with each other, and the quality of service is difficult to be guaranteed without coordination. Fortunately, a cloud radio access network (C-RAN) perfectly fills this gap as all the computations in the network are off-loaded to the computation center, making the central coordination possible. But a major challenge that C-RAN faces is that the resultant problem size could be extremely large, invalidating many existing second-order algorithms. In this paper, content-centric sparse multicast beamforming in a large-scale C-RAN is studied. In addition to the large-scale nature, this problem is further complicated by the discontinuity and non-convexity of the cost function and constraints. Despite the challenges, a first-order algorithm is proposed. Not only is the proposed algorithm guaranteed to converge to a critical point, but its complexity order is only linear with respect to the problem size. This is in sharp contrast to the cubic order of an existing solution, making the proposed algorithm indispensable for large-scale C-RAN with hundreds or thousands of users.**

***Index Terms*— Firsr-order algorithm, large-scale cloud radio access network (C-RAN), content-centric, sparse multicast beamforming, caching.**

I. Introduction

WHILE modern wireless data traffic is dominated by videos and other multimedia data, a prominent feature is that many users may be requesting the same content. To fully exploit the broadcast nature of the wireless medium, content-centric multicast transmission [1]–[3], in which users requesting the same content are served in the same multicast group, is a viable solution. On the other hand, to effectively manage the interference among different multicast groups, cloud radio access network (C-RAN) [4], where all the base stations (BSs) are connected to a computation center via high speed backhaul links and multi-BS cooperative beamforming design [5]–[11] is performed on the computation center, is a promising architecture for next-generation wireless systems.

However, putting multicast transmission and C-RAN together requires the joint design of BS clustering and multicast beamforming (also known as sparse multicast beamforming [12]). By BS clustering [13]–[17], each user is assigned to a subset of BSs instead of all the BSs to alleviate the backhaul burden of the network [18]–[21]. By multicast beamforming [22]–[27], a single beamformer is designed to serve each group of users requesting the same content. While BS clustering involves combinatorial optimization and is challenging by itself, adding multicast beamforming design further complicates the problem as it may involve nonconvex quality of service (QoS) constraints. Worse still, the optimization problem to be solved is usually in very large scale, making the task even more formidable.

For small/medium-scale C-RAN with no more than dozens of users, a classical flow for solving this problem [12], [24], [26] is to approximate and transform the original problem to a sequence of second-order cone programming (SOCP) subproblems, and then solve each subproblem by the interior-point method. Nevertheless, since the complexity order of the interior-point method is $\mathcal{O}(N^3)$ where N is the problem size [28], such a paradigm is not scalable as N is larger than 10^4. For instance, for a large-scale C-RAN with 50 multicast groups and 100 four-antenna BSs, the dimension of beamforming variables is $N = 2 \times 10^4$, and hence the complexity of the interior-point method to such a problem is about 8×10^{12}.

To solve a large-scale optimization problem, a basic principle is to decompose the problem by recognizing its structures and then build first-order algorithms for parallel implementations [29], [30]. Drawing on this principle, algorithms based on alternating direction method of multipliers (ADMM) [31] have been developed for different applications in large-scale networks [32]–[34]. While ADMM has been frequently applied and its convergence guarantee only depends on mild conditions for convex optimization problems [35], [36], it is not directly applicable to the problem of joint BS clustering and multicast beamforming design in C-RAN, since the combinatorial backhaul cost is discontinuous and nonconvex, and the QoS constraints are nonconvex [12]. To tackle the non-convexity, a recent trend is to first relax or approximate the problem

Manuscript received December 4, 2017; revised April 17, 2018; accepted June 26, 2018. Date of publication July 11, 2018; date of current version September 10, 2018. This work was supported by the National Natural Science Foundation of China under Grant 61671488. The associate editor coordinating the review of this paper and approving it for publication was M. Payaró.

Y. Li and Y.-C. Wu are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: liyang@eee.hku.hk; ycwu@eee.hku.hk).

M. Xia is with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: xiamingh@mail.sysu.edu.cn).

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

Digital Object Identifier 10.1109/TWC.2018.2852300

into a convex form, and then apply ADMM by exploiting the structures in the transformed problem [37]–[39].

Instead, in this paper, we directly tackle the joint design of BS clustering and multicast beamforming in large-scale C-RAN without relaxing the problem into a convex form. In particular, we observe that each component of the discontinuous nonconvex term is only determined by an individual beamforming vector, thus the discontinuous nonconvex term is block fully separable. Moreover, each component of the discontinuous nonconvex term is discontinuous only at a single point. Based on these two structures, we first decouple the discontinuous nonconvex cost function and the nonconvex constraints by introducing auxiliary variables. Then based on alternating minimization (AM), the nonconvex constraints and the discontinuous nonconvex cost function are handled separately. For the subproblem involving the non-convex constraints, the strong duality is exploited to obtain the closed-form optimal solution. On the other hand, for the subproblem involving the discontinuous nonconvex cost function, by exploiting the block separability and single-point discontinuity property, a majorization-minimization (MM) [40] based algorithm is proposed. Due to the judicious decomposition based on the problem structures, the overall algorithm only involves first-order differentiation, making its complexity order linear with respect to the problem size. Furthermore, it is proved that the overall algorithm is guaranteed to converge to a critical point,[1] which is important for the considered discontinuous nonconvex problem, since without such theoretical guarantee, AM based algorithms may find some point of no interest. Finally, simulation results are presented to show that in medium-scale networks, the proposed algorithm achieves the same performance as an existing second-order based approach, but with much shorter computation time. As the network size increases, the proposed algorithm becomes indispensable, as second-order algorithms (with cubic complexity order) are too complicated to run on a computer, while the proposed algorithm still returns a solution within reasonable time.

The rest of the paper is organized as follows. In Section II, we introduce the system model and the problem formulation. In Section III, a first-order algorithm is proposed to solve the large-scale discontinuous nonconvex problem. In Section IV, the proposed algorithm is proved to converge to a critical point. Simulation results are presented in Section V. Finally, Section VI concludes the paper.

II. System Model and Problem Formulation

Consider a C-RAN, where K single-antenna users are served by N L-antenna BSs, each with a local cache of limited size. Through backhaul links, the N BSs are connected to a computation center, which also has access to a database containing all the contents potentially required by the users. To effectively utilize the broadcast nature of the wireless medium, we consider multi-group multicast services in which users desiring the same content are served in the same multicast group [1]–[3].

In particular, let $\mathcal{M} = \{1, 2, \ldots, M\}$ be the set of multicast groups, with each group requesting the same content from the content set $\mathcal{F} = \{1, 2, \ldots, F\}$. Moreover, it is assumed that each user requests one content at a time and hence belongs to at most one group at any time [12], [26], i.e., there exists a many-to-one mapping from user $k \in \{1, 2, \ldots, K\}$ to group $m_k \in \mathcal{M}$. Denoting the aggregate network-wide multicast beamforming vector from all the BSs to the m-th group as $\mathbf{w}_m = \left[\mathbf{w}_{m,1}^H, \mathbf{w}_{m,2}^H, \ldots, \mathbf{w}_{m,N}^H\right]^H \in \mathbb{C}^{NL\times 1}$, where $\mathbf{w}_{m,n} \in \mathbb{C}^{L\times 1}$ is the beamforming vector from the n-th BS to the m-th group, the received signal at user k can be expressed as

$$y_k = \mathbf{h}_k^H \mathbf{w}_{m_k} x_{m_k} + \sum_{m\neq m_k} \mathbf{h}_k^H \mathbf{w}_m x_m + n_k, \quad \forall k = 1, 2, \ldots, K, \tag{1}$$

where $\mathbf{h}_k \in \mathbb{C}^{NL\times 1}$ is the network-wide channel vector from all the BSs to user k, $x_{m_k} \in \mathbb{C}$ is the data symbol sent to the group m_k with $\mathbb{E}\left[|x_{m_k}|^2\right] = 1$, and $n_k \sim \mathcal{CN}\left(0, \sigma_k^2\right)$ is the additive white Gaussian noise. In (1), the first term is the desired signal and the second term is the inter-group interference due to the multi-group multicast transmission. Consequently, the received signal-to-interference-plus-noise ratio (SINR) of user k is

$$\mathrm{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{w}_{m_k}|^2}{\sum_{m\neq m_k} |\mathbf{h}_k^H \mathbf{w}_m|^2 + \sigma_k^2}, \quad \forall k = 1, 2, \ldots, K. \tag{2}$$

In C-RAN, a central issue is to alleviate the backhaul burden by clustering the BSs to serve different multicast groups so that each group is only served by a subset of BSs. In this way, the computation center only distributes the required data of each group to its serving BSs instead of all the BSs. Such a BS clustering can be achieved by imposing a sparse structure to the aggregate network-wide beamformer $\mathbf{w}_m$. Furthermore, in modern content-centric communications, highly popular contents are cached in the BSs to further reduce the traffic in the backhaul links. If the content requested by the m-th group $f_m \in \mathcal{F}$ is cached in BS n, we define a caching status $c_{f_m,n} = 1$ (otherwise $c_{f_m,n} = 0$). Consequently, the amount of traffic on a particular backhaul link can be categorized into the following three cases. If BS n does not serve group m, the corresponding backhaul traffic is 0. If BS n serves group m and has stored the required content f_m on its local cache, the corresponding backhaul traffic is also 0. Only when BS n serves group m but has not stored the required f_m, it is necessary to fetch f_m from the database to BS n at a required transmission rate R_m. Combining the three cases yields the total backhaul cost $\sum_{m=1}^{M}\sum_{n=1}^{N} \left\| \|\mathbf{w}_{m,n}\|_2^2 \right\|_0 \left(1 - c_{f_m,n}\right) R_m$, where $\left\| \|\mathbf{w}_{m,n}\|_2^2 \right\|_0$ takes either 1 or 0 depending on whether BS n serves group m ($\mathbf{w}_{m,n} \neq \mathbf{0}$) or not.

In multicast transmission, to decode the message successfully for the users in the group m_k, we need $\mathrm{SINR}_k \geq \gamma_{m_k}$, where γ_{m_k} is the target SINR such that

[1]For general nonconvex nonsmooth functions, the points with limiting subdifferential (see Appendix C) containing zero are called critical points [41]. If the function is differentiable, critical point is the same as stationary point whose gradient equals to zero, since the limiting subdifferential of a differentiable function is just its gradient.

$R_{m_k} = B\log_2(1+\gamma_{m_k})$ and B is the available bandwidth. With the SINR requirement satisfied, we should minimize both the backhaul cost and the transmit power, giving the optimization problem as

$$\min_{\mathcal{W}} \underbrace{\sum_{m=1}^{M}\sum_{n=1}^{N}\left\|\|\mathbf{w}_{m,n}\|_2^2\right\|_0 (1-c_{f_m,n}) R_m}_{\text{backhaul cost}} + \eta \underbrace{\sum_{m=1}^{M}\sum_{n=1}^{N}\|\mathbf{w}_{m,n}\|_2^2}_{\text{transmit power cost}}, \tag{3a}$$

$$\text{s.t. } \frac{|\mathbf{h}_k^H\mathbf{w}_{m_k}|^2}{\sum_{m\neq m_k}|\mathbf{h}_k^H\mathbf{w}_m|^2+\sigma_k^2} \geq \gamma_{m_k}, \quad \forall k=1,2,\ldots,K, \tag{3b}$$

where $\mathcal{W} = \{\mathbf{w}_{m,n}|m\in\mathcal{M}, n=1,2,\ldots,N\}$ is the set of all the beamformers and η is a positive weighting parameter reflecting the trade-off between the backhaul cost and the transmit power cost. Since the caching status $c_{f_m,n}$ depends on a specific caching placement that happens in a much larger timescale than scheduling and transmission [42]–[45], we assume that the caching status $c_{f_m,n}$ is given and fixed during the design of BS clustering and multicast beamforming [12], [26]. The channel state information and user requests are also assumed to be available at the computation center.

The challenges in solving problem (3) lie in the discontinuity/non-convexity of the cost function and the non-convexity of the constraints. Worse still, the challenges in the cost function and the K constraints are coupled as they all depend on the same variable set $\mathcal{W}$. A partial solution [12] is to approximate the discontinuous cost function by a smooth function and then apply convex-concave procedure (CCP) [46] to solve a sequence of non-trivial SOCP subproblems with the interior-point method. However, such an approach does not decouple the cost function and the constraints, making its complexity order cubic with respect to the problem size. For a large-scale C-RAN with huge variable dimension, such a method would incur prohibitive computational burden to the computation center.

To overcome the challenges mentioned above, we strive to decouple the cost function and the K constraints by introducing auxiliary variables $\mathcal{V} = \{v_{k,m}|v_{k,m}=\mathbf{h}_k^H\mathbf{w}_m, m\in\mathcal{M}, k=1,2,\ldots,K\}$ such that problem (3) becomes

$$\min_{\mathcal{W},\mathcal{V}} \sum_{m=1}^{M}\sum_{n=1}^{N}\left\|\|\mathbf{w}_{m,n}\|_2^2\right\|_0 \alpha_{m,n} + \eta\sum_{m=1}^{M}\sum_{n=1}^{N}\|\mathbf{w}_{m,n}\|_2^2 + \rho\sum_{m=1}^{M}\sum_{k=1}^{K}\left|v_{k,m}-\mathbf{h}_k^H\mathbf{w}_m\right|^2, \tag{4a}$$

$$\text{s.t. } \gamma_{m_k}\left(\sum_{m\neq m_k}|v_{k,m}|^2+\sigma_k^2\right) - |v_{k,m_k}|^2 \leq 0, \quad \forall k=1,2,\ldots,K, \tag{4b}$$

where $\alpha_{m,n} \triangleq (1-c_{f_m,n}) R_m$ and ρ is a penalty parameter to control the degree of matching between the original variables and the auxiliary variables. When ρ tends to infinity, the solution of problem (4) converges to that of problem (3) [47]. For practical implementation, ρ is usually either set as a large value or increased with continuation.

With the constraints decoupled in (4), it is tempting to introduce a set of dual variables to make problem (4) resemble the ADMM formulation. However, even if we take the trouble to introduce the dual variables, ADMM cannot guarantee its convergence due to the discontinuity/non-convexity of the cost function and the non-convexity of the constraints. In contrast, we develop a novel first-order algorithm in Section III, with its guarantee to converge to a critical point proved in Section IV.

Remark 1: While the general formulation of the problem provides flexibility of all BSs serving a single group, due to path loss, the channels from far-away BSs would be much weaker than those of nearby BSs. Correspondingly, the solution of the optimization problem would automatically downplay the importance of the beamforming vectors corresponding to weak channels, meaning that we can treat the channels of those far-away BSs as zero without sacrificing the performance of the system. This results in a network with channel spatial sparsity, and hence compressed sensing based methods can be used to achieve good CSI estimation performance with limited training resources [48]–[50]. Similarly, efficient synchronization can be achieved by exploiting the sparse connection of the large-scale wireless network [51].

III. Proposed First-Order Algorithm

Since the cost function and the constraints in (4) are decoupled when either $\mathcal{W}$ or $\mathcal{V}$ is fixed, AM is a suitable framework in solving (4). However, each subproblem under the AM framework is still challenging. In particular, when $\mathcal{W}$ is fixed, the subproblem over $\mathcal{V}$ is a nonconvex quadratically constrained quadratic programming (QCQP) problem. Prevalent techniques for solving nonconvex QCQP problems include semi-definite relaxation [22] and CCP [46], which however lead to squaring the number of variables or iteratively solving a sequence of non-trivial convex optimization problems. On the other hand, when $\mathcal{V}$ is fixed, the subproblem over $\mathcal{W}$ is combinatorial due to the discontinuous nonconvex term, which requires exhaustively enumerating all $2^N M$ possibilities and solving a convex optimization problem with dimension NL in each possible configuration. Below, we will propose a closed-form solution for updating $\mathcal{V}$ when $\mathcal{W}$ is fixed, and a first-order algorithm for updating $\mathcal{W}$ when $\mathcal{V}$ is fixed, rendering the overall algorithm first-order under the AM framework.

A. Updating $\mathcal{V}$: Tackling the Nonconvex Constraints

When $\mathcal{W}$ is fixed, problem (4) becomes a nonconvex QCQP problem, which can be decomposed into K subproblems, with the k-th subproblem written as

$$\min_{\{v_{k,m}\}_{m=1}^{M}} \sum_{m=1}^{M}\left|v_{k,m}-\mathbf{h}_k^H\mathbf{w}_m\right|^2, \tag{5a}$$

TABLE I
ALGORITHM 1 - COMPUTING CLOSED-FORM OPTIMAL SOLUTION OF (5)

if $\gamma_{m_k}\left(\sum_{m\neq m_k}|\mathbf{h}_k^H\mathbf{w}_m|^2+\sigma_k^2\right)-|\mathbf{h}_k^H\mathbf{w}_{m_k}|^2\leq 0$,

$v_{k,m}=\mathbf{h}_k^H\mathbf{w}_m,\ \ \forall m\in\mathcal{M}$;

else if $\mathbf{h}_k^H\mathbf{w}_{m_k}=0$,

$v_{k,m}=\frac{\mathbf{h}_k^H\mathbf{w}_m}{1+\gamma_{m_k}},\ \ \forall m\neq m_k,\ m\in\mathcal{M}$,

$v_{k,m_k}=\sqrt{\gamma_{m_k}\left(\sum_{m\neq m_k}|v_{k,m}|^2+\sigma_k^2\right)}$;

else

$v_{k,m}=\frac{\mathbf{h}_k^H\mathbf{w}_m}{1+\mu_k\gamma_{m_k}},\ \ \forall m\neq m_k,\ m\in\mathcal{M}$,

$v_{k,m_k}=\frac{\mathbf{h}_k^H\mathbf{w}_{m_k}}{1-\mu_k}$,

where $\mu_k\in(0,1)$ satisfying

$\frac{\gamma_{m_k}\sum_{m\neq m_k}|\mathbf{h}_k^H\mathbf{w}_m|^2}{(1+\mu_k\gamma_{m_k})^2}-\frac{|\mathbf{h}_k^H\mathbf{w}_{m_k}|^2}{(1-\mu_k)^2}+\gamma_{m_k}\sigma_k^2=0$

is found by the bisection method.

$$\text{s.t. } \gamma_{m_k}\left(\sum_{m\neq m_k}|v_{k,m}|^2+\sigma_k^2\right)-|v_{k,m_k}|^2\leq 0. \tag{5b}$$

Since the resulting subproblem (5) has only one constraint, it is a QCQP-1 (QCQP with only one constraint) problem. Consequently, the strong duality of (5) holds despite the non-convexity of the constraint and hence it can be solved optimally [52], [53]. Further coupled with the fact that the dual variable of a QCQP-1 is a scalar, the optimal solution of (5) can be efficiently found by analyzing its Karush-Kuhn-Tucker (KKT) conditions. As shown in Appendix A, by working with the dual variable in the KKT conditions, the closed-form optimal solution of (5) can be obtained, and the resultant algorithm is shown in Table I.

B. Updating $\mathcal{W}$: Tackling the Discontinuous Nonconvex Cost Function

When $\mathcal{V}$ is fixed, and due to the fact that all terms in the cost function (4a) contain a summation on m, problem (4) reduces to M subproblems, with each subproblem given by

$$\min_{\mathbf{w}_m}\ \sum_{n=1}^{N}\left\|\|\mathbf{w}_{m,n}\|_2^2\right\|_0\alpha_{m,n}+\eta\sum_{n=1}^{N}\|\mathbf{w}_{m,n}\|_2^2 +\rho\sum_{k=1}^{K}\left|v_{k,m}-\mathbf{h}_k^H\mathbf{w}_m\right|^2. \tag{6}$$

Inherited from (4), the challenges in solving (6) is the discontinuity and the non-convexity of the first term. In the sequel, we strive to solve such a discontinuous nonconvex problem (6) by exploiting two particular structures. Firstly, notice that each component $\left\|\|\mathbf{w}_{m,n}\|_2^2\right\|_0\alpha_{m,n}$ is only determined by an individual beamforming vector $\mathbf{w}_{m,n}$, thus the discontinuous nonconvex term in (6) is block fully separable. Secondly, each component $\left\|\|\mathbf{w}_{m,n}\|_2^2\right\|_0\alpha_{m,n}$ is discontinuous only at a single point $\mathbf{w}_{m,n}=\mathbf{0}$. These two structures imply that if we can further decompose the problem (6) into N smaller-scale subproblems, with each subproblem only depending on a single $\mathbf{w}_{m,n}$, exhaustive combinatorial enumeration can be avoided when dealing with the first term of (6).

To achieve further decomposition to problem (6), we construct a separable upper bound of the cost function in (6), where each component of the upper bound depends on a single $\mathbf{w}_{m,n}$. To be specific, given any $\tilde{\mathbf{w}}_m\in\mathbb{C}^{NL\times 1}$, the second order Taylor expansion of the last term in (6) around $\tilde{\mathbf{w}}_m$ is

$$\begin{aligned}\sum_{k=1}^{K}&\left|v_{k,m}-\mathbf{h}_k^H\mathbf{w}_m\right|^2\\ &=\sum_{k=1}^{K}\left|v_{k,m}-\mathbf{h}_k^H\tilde{\mathbf{w}}_m\right|^2\\ &\quad+2\Re\left\{\left(\sum_{k=1}^{K}\mathbf{h}_k\left(\mathbf{h}_k^H\tilde{\mathbf{w}}_m-v_{k,m}\right)\right)^H(\mathbf{w}_m-\tilde{\mathbf{w}}_m)\right\}\\ &\quad+(\mathbf{w}_m-\tilde{\mathbf{w}}_m)^H\left(\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^H\right)(\mathbf{w}_m-\tilde{\mathbf{w}}_m),\end{aligned} \tag{7}$$

where higher order terms are zero since (7) is a quadratic function over $\mathbf{w}_m$. Let $H_{\max}$ denote the maximum eigenvalue of the matrix $\mathbf{H}\triangleq\sum_{k=1}^{K}\mathbf{h}_k\mathbf{h}_k^H$ in (7) and define $H=H_{\max}+\delta$ for any $\delta>0$. Then we have $\mathbf{H}\preceq H_{\max}\mathbf{I}_{NL}\preceq H\mathbf{I}_{NL}$, which means that $H_{\max}\mathbf{I}_{NL}-\mathbf{H}$ and $H\mathbf{I}_{NL}-H_{\max}\mathbf{I}_{NL}$ are positive semi-definite. This implies

$$\begin{aligned}(\mathbf{w}_m-\tilde{\mathbf{w}}_m)^H\mathbf{H}(\mathbf{w}_m-\tilde{\mathbf{w}}_m)&\leq H_{\max}\|\mathbf{w}_m-\tilde{\mathbf{w}}_m\|_2^2\\ &\leq H\|\mathbf{w}_m-\tilde{\mathbf{w}}_m\|_2^2,\end{aligned} \tag{8}$$

with the equalities holding at $\mathbf{w}_m=\tilde{\mathbf{w}}_m$. Substituting (8) into (7), we can obtain the following inequalities:

$$\begin{aligned}\sum_{k=1}^{K}&\left|v_{k,m}-\mathbf{h}_k^H\mathbf{w}_m\right|^2\\ &\leqslant\sum_{k=1}^{K}\left|v_{k,m}-\mathbf{h}_k^H\tilde{\mathbf{w}}_m\right|^2\\ &\quad+2\Re\left\{\left(\sum_{k=1}^{K}\mathbf{h}_k\left(\mathbf{h}_k^H\tilde{\mathbf{w}}_m-v_{k,m}\right)\right)^H(\mathbf{w}_m-\tilde{\mathbf{w}}_m)\right\}\\ &\quad+H\|\mathbf{w}_m-\tilde{\mathbf{w}}_m\|_2^2 \qquad (9)\\ &=H\|\mathbf{w}_m-\mathbf{u}_m\|_2^2+C, \qquad (10)\end{aligned}$$

where $C\triangleq\sum_{k=1}^{K}\left|v_{k,m}-\mathbf{h}_k^H\tilde{\mathbf{w}}_m\right|^2-\frac{1}{H}\left\|\sum_{k=1}^{K}\mathbf{h}_k\left(\mathbf{h}_k^H\tilde{\mathbf{w}}_m-v_{k,m}\right)\right\|_2^2$ is a constant and $\mathbf{u}_m\triangleq\tilde{\mathbf{w}}_m-\frac{1}{H}\sum_{k=1}^{K}\mathbf{h}_k\left(\mathbf{h}_k^H\tilde{\mathbf{w}}_m-v_{k,m}\right)$ is also independent of $\mathbf{w}_m$. Based on (10), an upper bound of the cost function in (6) can be constructed as

$$\sum_{n=1}^{N}\left\{\left\|\|\mathbf{w}_{m,n}\|_2^2\right\|_0\alpha_{m,n}+\eta\|\mathbf{w}_{m,n}\|_2^2 +\rho H\|\mathbf{w}_{m,n}-\mathbf{u}_{m,n}\|_2^2+\frac{\rho C}{N}\right\}, \tag{11}$$

which can be split into N parts, with each component only depending on a single $\mathbf{w}_{m,n}$ as shown inside the brace in (11) and $\mathbf{u}_{m,n}\in\mathbb{C}^{L\times 1}$ being the n-th segment of $\mathbf{u}_m$, i.e., $\mathbf{u}_m=\left[\mathbf{u}_{m,1}^H,\mathbf{u}_{m,2}^H,\ldots,\mathbf{u}_{m,N}^H\right]^H$.

With the constructed separable upper bound (11), we can minimize (6) in the MM framework [40], in which a sequence of upper bounds expanded around the solution in the last iteration is successively minimized. In this way, a sequence of solutions decreasing the cost function in (6) can be obtained. Specifically, in the t-th ($t \geq 1$) iteration, we need to solve N smaller-scale subproblems in parallel, with each written as

$$\mathbf{w}_{m,n}^{(t+1)} = \arg\min_{\mathbf{w}_{m,n}} \left\| \|\mathbf{w}_{m,n}\|_2^2 \right\|_0 \alpha_{m,n} + \eta \|\mathbf{w}_{m,n}\|_2^2 + \rho H \left\| \mathbf{w}_{m,n} - \mathbf{u}_{m,n}^{(t)} \right\|_2^2, \quad (12)$$

where $\mathbf{u}_{m,n}^{(t)}$ is the corresponding $\mathbf{u}_{m,n}$ with the expansion point being the solution in the last iteration $\tilde{\mathbf{w}}_m = \mathbf{w}_m^{(t)}$.

Note that each subproblem (12) involves only one discontinuous term $\left\| \|\mathbf{w}_{m,n}\|_2^2 \right\|_0 \alpha_{m,n}$, which only takes the value either 0 or $\alpha_{m,n}$ depending on whether the variable $\mathbf{w}_{m,n} = \mathbf{0}$ or not. Therefore, we consider two cases:

1) If $\mathbf{w}_{m,n} = \mathbf{0}$, the corresponding cost function value of (12) is $\rho H \left\| \mathbf{u}_{m,n}^{(t)} \right\|_2^2$.
2) If $\mathbf{w}_{m,n} \neq \mathbf{0}$, the first term of the cost function in (12) is a constant $\alpha_{m,n}$ and hence subproblem (12) can be rewritten as

$$\min_{\mathbf{w}_{m,n}} \eta \|\mathbf{w}_{m,n}\|_2^2 + \rho H \left\| \mathbf{w}_{m,n} - \mathbf{u}_{m,n}^{(t)} \right\|_2^2, \quad (13)$$

which is a convex quadratic programming problem whose optimal solution can be found by differentiating the cost function in (13) and setting it to zero: $\eta \mathbf{w}_{m,n} + \rho H \left(\mathbf{w}_{m,n} - \mathbf{u}_{m,n}^{(t)} \right) = \mathbf{0}$, which gives the optimal closed-form solution as $\mathbf{w}_{m,n} = \frac{\rho H}{\eta + \rho H} \mathbf{u}_{m,n}^{(t)}$. By substituting it into (12), we can obtain the corresponding cost function value as $\alpha_{m,n} + \frac{\eta \rho H}{\eta + \rho H} \left\| \mathbf{u}_{m,n}^{(t)} \right\|_2^2$.

By comparing the resultant cost function values in the above two cases, the closed-form optimal solution to subproblem (12) can be obtained as

$$\mathbf{w}_{m,n}^{(t+1)} = \begin{cases} \mathbf{0}, & \text{if } \left\| \mathbf{u}_{m,n}^{(t)} \right\|_2 \leq \frac{\sqrt{\alpha_{m,n}(\eta + \rho H)}}{\rho H}, \\ \frac{\rho H}{\eta + \rho H} \mathbf{u}_{m,n}^{(t)}, & \text{if } \left\| \mathbf{u}_{m,n}^{(t)} \right\|_2 > \frac{\sqrt{\alpha_{m,n}(\eta + \rho H)}}{\rho H}. \end{cases} \quad (14)$$

To sum up, the above MM procedure for solving problem (6) is given in Table II. Since in each iteration, only the first-order differentiation is involved, Algorithm 2 belongs to the class of first-order algorithms. Notice that when the cost function is continuous, the convergence property of MM based algorithms has been analyzed in [54]. However, the cost function in (6) is discontinuous and hence it requires further analysis to establish its convergence property, which is shown in the following theorem.

Theorem 1: Using Algorithm 2, the cost function of problem (6) is monotonic decreasing as the iteration number t increases. In particular, as t tends to infinity, it can be guaranteed to converge to at least a local minimum of problem (6).

Proof: See Appendix B. ■

TABLE II
ALGORITHM 2 - FINDING LOCAL OPTIMAL SOLUTION OF (6)

Initialization of $\mathbf{w}_m$.
for $t = 1, 2, \cdots, T$
1) $\mathbf{u}_m^{(t)} = \mathbf{w}_m^{(t)} - \frac{1}{H} \sum_{k=1}^{K} \mathbf{h}_k \left(\mathbf{h}_k^H \mathbf{w}_m^{(t)} - v_{k,m} \right)$.
2) $\forall n \in \mathcal{N}$
  if $\left\| \mathbf{u}_{m,n}^{(t)} \right\|_2 \leq \frac{\sqrt{\alpha_{m,n}(\eta + \rho H)}}{\rho H}$, $\mathbf{w}_{m,n}^{(t+1)} = \mathbf{0}$;
  else, $\mathbf{w}_{m,n}^{(t+1)} = \frac{\rho H}{\eta + \rho H} \mathbf{u}_{m,n}^{(t)}$.
end

TABLE III
ALGORITHM 3 - THE OVERALL AM BASED FIRST-ORDER ALGORITHM

Initialization of $\mathcal{V}$ and $\mathcal{W}$.
repeat
1) Update $\mathcal{V}$: Compute the closed-form solution of $\{v_{k,m}\}_{m=1}^{M}$ using Algorithm 1 $\forall k = 1, 2, \ldots, K$.
2) Update $\mathcal{W}$: Update $\mathbf{w}_m$ using Algorithm 2 $\forall m \in \mathcal{M}$.
until convergence

C. Overall Algorithm and Parallel Implementation

With the updates of $\mathcal{V}$ and $\mathcal{W}$ given by Algorithms 1 and 2 respectively, the overall algorithm for solving problem (4) is to alternatively execute the two algorithms under the AM framework (see Algorithm 3 in Table III). The computation of the overall Algorithm 3 is dominated by the calculation of $\sum_{k=1}^{K} \mathbf{h}_k \left(\mathbf{h}_k^H \mathbf{w}_m^{(t)} - v_{k,m} \right), \forall m \in \mathcal{M}$, thus its complexity order is $\mathcal{O}(KMNL)$. Since both Algorithms 1 and 2 only involve first-order differentiation, the overall Algorithm 3 is a first-order algorithm. Furthermore, since each update of $\mathcal{V}$ or $\mathcal{W}$ leads to a monotonic decrease in the cost function (4a) which is also bounded below by 0, Algorithm 3 can guarantee the convergence of the cost function values of problem (4). Notice that due to Theorem 1, when updating $\mathcal{W}$, any number of iterations T in Algorithm 2 induces a monotonic decrease in the cost function in (6) as well as in (4a), therefore T can be chosen as any positive integer in the implementation.

A closer look at Algorithm 3 reveals that updating $\mathcal{V}$ consists of K parallel subproblems, with each in the form of (5). On the other hand, updating $\mathcal{W}$ consists of M parallel subproblems, and each subproblem can be further decomposed into N smaller-scale subproblems in the form of (12). This revelation suggests that the overall Algorithm 3 can be implemented in a parallel manner as in Fig. 1. In particular, the k-th node on the left computes $\{v_{k,m}\}_{m=1}^{M}$ (corresponding to Algorithm 1) after receiving $\{\mathbf{h}_k^H \mathbf{w}_m\}_{m=1}^{M}$ from the M nodes in the middle. The result of $v_{k,m}$ will be sent to the m-th node in the middle. On the other hand, after receiving $\{v_{k,m}\}_{k=1}^{K}$, the m-th node in the middle computes $\mathbf{u}_m^{(1)}$ (corresponding to step 1 of Algorithm 2). Then $\mathbf{u}_m^{(1)}$ is split into N segments and its n-th segment $\mathbf{u}_{m,n}^{(1)}$ is sent to the (m,n)-th node on the right. Once $\mathbf{u}_{m,n}^{(1)}$ is received at the (m,n)-th node on the right, $\mathbf{w}_{m,n}^{(2)}$ is computed (corresponding to step 2 of Algorithm 2). Then $\mathbf{w}_{m,n}^{(2)}$ will be returned to the nodes in the middle to construct $\mathbf{w}_m^{(2)}$. Algorithm 2 will be executed

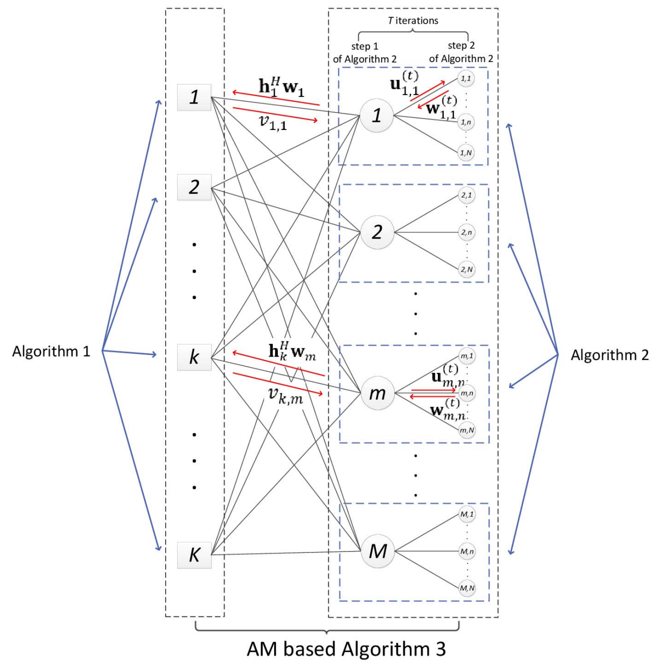


Fig. 1. Parallel architecture of Algorithm 3.

for T iterations before returning to Algorithm 1. The whole procedure is repeated until the AM converges. Notice that the computations in the K nodes on the left can be carried out in parallel, while the computations in the M nodes in the middle can also be carried out in parallel. Furthermore, within Algorithm 2, the step 2 can also be executed in parallel for the N subproblems on the right. Therefore, the proposed algorithm has the potential of leveraging the modern multi-thread and multi-core computing architecture for speeding up the computation.

IV. Convergence Guarantee to Critical Point

As shown in Section III-C, the cost function values of problem (4) can be guaranteed to converge using Algorithm 3. If the cost function is continuously differentiable and under certain additional assumptions, we can further conclude that the resulting sequence of solutions has limit points and every limit point is a critical point [55]–[57]. Recently, the convergence analysis is extended to a more general setting, where the cost function is a sum of a differentiable part and a nonsmooth convex part [58]. However, all the above analyses cannot be applied to problem (4), since the nonsmooth term in the cost function (4a) is neither convex nor continuous. Consequently, it is necessary to answer the question of whether the sequence of solutions would converge, and if so where it converges? In this section, we prove that, despite the discontinuity and non-convexity of problem (4), the sequence of solutions can still be guaranteed to converge to a critical point.

To begin with, we define $(\mathcal{V}_i, \mathcal{W}_i)$ as the solution generated by Algorithm 3 at the i-th iteration, with the corresponding cost function value (4a) denoted as $g(\mathcal{V}_i, \mathcal{W}_i)$. First, we show that the sequence of solutions $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ generated by Algorithm 3 has convergent subsequences $\left\{\mathcal{V}_{i_j}, \mathcal{W}_{i_j}\right\}_{j\in\mathbb{N}}$ with the index sequence $\{i_j\}_{j\in\mathbb{N}} \subseteq \mathbb{N}$ satisfying $i_j < i_{j+1}, \forall j \in \mathbb{N}$.

Theorem 2: (Existence of Convergent Subsequences) The sequence of solutions $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ generated by Algorithm 3 has convergent subsequences, i.e., there must exist a subsequence $\left\{\mathcal{V}_{i_j}, \mathcal{W}_{i_j}\right\}_{j\in\mathbb{N}}$ converging to a limit point $(\mathcal{V}^*, \mathcal{W}^*)$ as j tends to infinity.

Proof: Since the cost function in (4a) monotonically decreases as iteration number increases, we have

$$g\left(\mathcal{V}_{i-1}, \mathcal{W}_{i-1}\right) \geq g\left(\mathcal{V}_i, \mathcal{W}_{i-1}\right) \geq g\left(\mathcal{V}_i, \mathcal{W}_i\right), \forall i \in \{1, 2, \ldots\}. \tag{15}$$

It follows that

$$g\left(\mathcal{V}_i, \mathcal{W}_i\right) \leq g\left(\mathcal{V}_0, \mathcal{W}_0\right), \quad \forall i \in \{1, 2, \ldots\}, \tag{16}$$

where $g\left(\mathcal{V}_0, \mathcal{W}_0\right)$ is any finite initial value of the cost function. Now we prove the boundedness of $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ by contradiction. Suppose that there exists an unbounded $(\mathcal{V}_i, \mathcal{W}_i)$. Due to the quadratic terms with respect to $\mathcal{V}$ and $\mathcal{W}$ in (4a), this hypothesis will lead to an infinite $g\left(\mathcal{V}_i, \mathcal{W}_i\right)$, which is contradictory to (16). Thus, the sequence $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ is bounded.

With the boundedness of $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$, we can directly establish the existence of convergent subsequences [59]. Consequently, there must exist a subsequence $\left\{\mathcal{V}_{i_j}, \mathcal{W}_{i_j}\right\}_{j\in\mathbb{N}}$ converging to a limit point $(\mathcal{V}^*, \mathcal{W}^*)$ as j tends to infinity. ∎

Up to now, we have shown that $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ has convergent subsequences. To investigate the property of the corresponding limit point $(\mathcal{V}^*, \mathcal{W}^*)$, we need two more properties given below. To facilitate the discussion, we define $\mathcal{W}_i^{(t+1)}$ as the $\mathcal{W}$ after the t-th iteration of Algorithm 2 within the i-th iteration of Algorithm 3. Furthermore, since Algorithm 2 is executed T iterations within each iteration of Algorithm 3, we have $\mathcal{W}_i^{(T+1)} = \mathcal{W}_{i+1}^{(1)} = \mathcal{W}_i$.

Lemma 1: (Sufficient Decrease) Using Algorithm 3, the cost function $g\left(\mathcal{V}, \mathcal{W}\right)$ has a sufficient decrease property:

$$g\left(\mathcal{V}_{i-1}, \mathcal{W}_{i-1}\right) - g\left(\mathcal{V}_i, \mathcal{W}_i\right) \geq \sum_{t=1}^{T} \rho\delta \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2^2, \quad \forall i \in \{1, 2, \ldots\}, \tag{17}$$

where ρ is defined in (4a) and $\delta = H - H_{\max}$ is defined before (8).

Proof: Define

$$\begin{aligned} &g_\beta\left(\mathcal{V}, \mathcal{W}; \tilde{\mathcal{W}}\right) \\ &= \sum_{m=1}^{M}\sum_{n=1}^{N} \left\| \|\mathbf{w}_{m,n}\|_2^2 \right\|_0 \alpha_{m,n} + \eta \sum_{m=1}^{M}\sum_{n=1}^{N} \|\mathbf{w}_{m,n}\|_2^2 \\ &\quad + \rho \sum_{m=1}^{M}\sum_{k=1}^{K} \left| v_{k,m} - \mathbf{h}_k^H \tilde{\mathbf{w}}_m \right|^2 \\ &\quad + 2\rho \sum_{m=1}^{M} \Re\left\{ \left(\sum_{k=1}^{K} \mathbf{h}_k \left(\mathbf{h}_k^H \tilde{\mathbf{w}}_m - v_{k,m} \right) \right)^H \left(\mathbf{w}_m - \tilde{\mathbf{w}}_m \right) \right\} \\ &\quad + \rho\beta \sum_{m=1}^{M} \|\mathbf{w}_m - \tilde{\mathbf{w}}_m\|_2^2, \end{aligned} \tag{18}$$

with β being a parameter. Applying (8) to $g_\beta\left(\mathcal{V}_i, \mathcal{W}; \mathcal{W}_i^{(t)}\right)$, we have

$$g(\mathcal{V}_i, \mathcal{W}) \le g_{H_{\max}}\left(\mathcal{V}_i, \mathcal{W}; \mathcal{W}_i^{(t)}\right) \le g_H\left(\mathcal{V}_i, \mathcal{W}; \mathcal{W}_i^{(t)}\right), \quad \forall \mathcal{W}, \tag{19}$$

with the equalities holding at $\mathcal{W} = \mathcal{W}_i^{(t)}$. It follows that

$$\begin{aligned} & g\left(\mathcal{V}_i, \mathcal{W}_i^{(t)}\right) - g\left(\mathcal{V}_i, \mathcal{W}_i^{(t+1)}\right) \\ & \ge g\left(\mathcal{V}_i, \mathcal{W}_i^{(t)}\right) - g_{H_{\max}}\left(\mathcal{V}_i, \mathcal{W}_i^{(t+1)}; \mathcal{W}_i^{(t)}\right) \\ & = g_H\left(\mathcal{V}_i, \mathcal{W}_i^{(t)}; \mathcal{W}_i^{(t)}\right) - g_{H_{\max}}\left(\mathcal{V}_i, \mathcal{W}_i^{(t+1)}; \mathcal{W}_i^{(t)}\right) \\ & \ge g_H\left(\mathcal{V}_i, \mathcal{W}_i^{(t+1)}; \mathcal{W}_i^{(t)}\right) - g_{H_{\max}}\left(\mathcal{V}_i, \mathcal{W}_i^{(t+1)}; \mathcal{W}_i^{(t)}\right) \\ & = \rho\delta \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2^2, \end{aligned} \tag{20}$$

where the second last step is due to the fact that with $\mathcal{V}$ fixed at $\mathcal{V}_i$, the component of $\mathcal{W}_i^{(t+1)}$ is the minimizer of subproblem (12) or equivalently $\mathcal{W}_i^{(t+1)}$ is the minimizer of $g_H\left(\mathcal{V}_i, \mathcal{W}; \mathcal{W}_i^{(t)}\right)$, and the last equality is due to the definition in (18). Consequently,

$$\begin{aligned} g(\mathcal{V}_i, \mathcal{W}_{i-1}) - g(\mathcal{V}_i, \mathcal{W}_i) &= g\left(\mathcal{V}_i, \mathcal{W}_i^{(1)}\right) - g\left(\mathcal{V}_i, \mathcal{W}_i^{(T+1)}\right) \\ &\ge \sum_{t=1}^{T} \rho\delta \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2^2. \end{aligned} \tag{21}$$

Applying (15) to the left hand side of (21) yields (17). ■

Lemma 1 reveals how much the cost function $g(\mathcal{V}, \mathcal{W})$ decreases as iteration number increases. Based on this, we can further establish the square summability of $\{\mathcal{V}_i, \mathcal{W}_i\}_{i \in \mathbb{N}}$ with the following lemma.

Lemma 2: (Square Summability) Using Algorithm 3, the sequence $\left\{\mathcal{W}_i^{(t)}\right\}_{i \in \mathbb{N}}$ satisfies

$$\sum_{i=1}^{\infty} \sum_{t=1}^{T} \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2^2 < \infty, \tag{22}$$

which implies

$$\begin{aligned} & \lim_{i \to \infty} \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2 = \lim_{i \to \infty} \|\mathcal{W}_i - \mathcal{W}_{i+1}\|_2 \\ & = \lim_{i \to \infty} \|\mathcal{V}_i - \mathcal{V}_{i+1}\|_2 = 0, \quad \forall t = 1, 2, \ldots, T. \end{aligned} \tag{23}$$

Proof: According to the sufficient decrease property of Lemma 1, we have

$$g(\mathcal{V}_0, \mathcal{W}_0) - g(\mathcal{V}_I, \mathcal{W}_I) \ge \sum_{i=1}^{I} \sum_{t=1}^{T} \rho\delta \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2^2, \tag{24}$$

where I is any positive integer. Letting $I \to \infty$, we have

$$\begin{aligned} & \sum_{i=1}^{\infty} \sum_{t=1}^{T} \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2^2 \\ & \le \lim_{I \to \infty} \frac{1}{\rho\delta} \left(g(\mathcal{V}_0, \mathcal{W}_0) - g(\mathcal{V}_I, \mathcal{W}_I) \right) \\ & \le \frac{1}{\rho\delta} g(\mathcal{V}_0, \mathcal{W}_0) < \infty, \end{aligned} \tag{25}$$

where the second inequality follows from the fact that $g(\mathcal{V}, \mathcal{W})$ is bounded below by 0.

Furthermore, due to the square summability (25), we have $\lim_{i \to \infty} \sum_{t=1}^{T} \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2^2 = 0$ and hence $\lim_{i \to \infty} \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2 = 0, \forall t = 1, 2, \ldots, T$ [59]. On the other hand, notice that

$$\begin{aligned} \|\mathcal{W}_i - \mathcal{W}_{i+1}\|_2 &= \left\| \mathcal{W}_{i+1}^{(1)} - \mathcal{W}_{i+1}^{(T+1)} \right\|_2 \\ &= \left\| \sum_{t=1}^{T} \left(\mathcal{W}_{i+1}^{(t)} - \mathcal{W}_{i+1}^{(t+1)} \right) \right\|_2 \\ &\le \sum_{t=1}^{T} \left\| \mathcal{W}_{i+1}^{(t)} - \mathcal{W}_{i+1}^{(t+1)} \right\|_2, \end{aligned} \tag{26}$$

thus we have $\lim_{i \to \infty} \|\mathcal{W}_i - \mathcal{W}_{i+1}\|_2 = \lim_{i \to \infty} \left\| \mathcal{W}_i^{(t)} - \mathcal{W}_i^{(t+1)} \right\|_2 = 0$, i.e., $\mathcal{W}_i \to \mathcal{W}_{i+1}$ as $i \to \infty$. Notice that the closed-form solution of $\mathcal{V}$ is continuous over $\mathcal{W}$, thus we have $\mathcal{V}_{i+1} \to \mathcal{V}_{i+2}$ as $\mathcal{W}_i \to \mathcal{W}_{i+1}$ and hence (23) holds. ■

Lemma 2 reveals that the difference between $\mathcal{W}_i$ and $\mathcal{W}_{i+1}$ (also $\mathcal{V}_i$ and $\mathcal{V}_{i+1}$) goes to zero as the iteration number goes to infinity. Based on this lemma, we can investigate the property of the limit points of $\{\mathcal{V}_i, \mathcal{W}_i\}_{i \in \mathbb{N}}$ in the following theorem.

Theorem 3: (Subsequence Converges to Critical Point) Using Algorithm 3, any convergent subsequence $\{\mathcal{V}_{i_j}, \mathcal{W}_{i_j}\}_{j \in \mathbb{N}}$ is guaranteed to converge to a critical point of (4), i.e., the limit point $(\mathcal{V}^*, \mathcal{W}^*)$ satisfies the first-order optimality condition of (4).

Proof: To begin with, we show that the limit point $(\mathcal{V}^*, \mathcal{W}^*)$ satisfies the first-order optimality condition of (4) over $\mathcal{W}$. As shown after (20), $\mathcal{W}_i^{(t+1)}$ is the minimizer of $g_H\left(\mathcal{V}_i, \mathcal{W}; \mathcal{W}_i^{(t)}\right)$, thus we have

$$g_H\left(\mathcal{V}_{i_j}, \mathcal{W}; \mathcal{W}_{i_j}^{(T)}\right) \ge g_H\left(\mathcal{V}_{i_j}, \mathcal{W}_{i_j}^{(T+1)}; \mathcal{W}_{i_j}^{(T)}\right), \quad \forall \mathcal{W}. \tag{27}$$

Notice that as $j \to \infty$, $\mathcal{W}_{i_j}^{(T+1)} = \mathcal{W}_{i_j} \to \mathcal{W}^*$, and from (23) of Lemma 2 we also have $\mathcal{W}_{i_j}^{(T)} \to \mathcal{W}_{i_j}^{(T+1)}$, therefore $\mathcal{W}_{i_j}^{(T)} \to \mathcal{W}^*$ and hence (27) leads to

$$\begin{aligned} g_H(\mathcal{V}^*, \mathcal{W}; \mathcal{W}^*) &= \liminf_{j \to \infty} g_H\left(\mathcal{V}_{i_j}, \mathcal{W}; \mathcal{W}_{i_j}^{(T)}\right) \\ &\ge \liminf_{j \to \infty} g_H\left(\mathcal{V}_{i_j}, \mathcal{W}_{i_j}^{(T+1)}; \mathcal{W}_{i_j}^{(T)}\right) \\ &\ge g_H(\mathcal{V}^*, \mathcal{W}^*; \mathcal{W}^*), \quad \forall \mathcal{W}, \end{aligned} \tag{28}$$

where the first and third steps are due to the fact that $g_H\left(\mathcal{V}, \mathcal{W}; \tilde{\mathcal{W}}\right)$ is continuous over $\mathcal{V}$ and $\tilde{\mathcal{W}}$, and lower semi-continuous over $\mathcal{W}$. From (28), it follows that $\mathcal{W}^*$ is the minimizer of $g_H(\mathcal{V}^*, \mathcal{W}; \mathcal{W}^*)$. Thus, the limiting subdifferential[2] of $g_H(\mathcal{V}^*, \mathcal{W}; \mathcal{W}^*)$ satisfies

$$\mathbf{0} \in \partial_{\mathcal{W}} g_H(\mathcal{V}^*, \mathcal{W}; \mathcal{W}^*)|_{\mathcal{W} = \mathcal{W}^*}. \tag{29}$$

[2]The limiting subdifferential of a nonsmooth nonconvex function is defined in Appendix C.

On the other hand, by substituting (7) into (4a), it can be proved that

$$g(\mathcal{V},\mathcal{W}) = g_H\left(\mathcal{V},\mathcal{W};\tilde{\mathcal{W}}\right) + \rho\sum_{m=1}^{M}(\mathbf{w}_m - \tilde{\mathbf{w}}_m)^H(\mathbf{H} - H\mathbf{I}_{NL})(\mathbf{w}_m - \tilde{\mathbf{w}}_m). \quad (30)$$

Notice that the gradient of the last term in (30) over $\mathcal{W}$ is zero at $\mathcal{W} = \tilde{\mathcal{W}}$. Therefore, applying the limiting subdifferentiation over $\mathcal{W}$ on both sides of (30) and putting $\mathcal{W} = \tilde{\mathcal{W}} = \mathcal{W}^*$, we have

$$\partial_{\mathcal{W}} g(\mathcal{V},\mathcal{W})|_{\mathcal{W}=\mathcal{W}^*} = \partial_{\mathcal{W}} g_H(\mathcal{V},\mathcal{W};\mathcal{W}^*)|_{\mathcal{W}=\mathcal{W}^*}. \quad (31)$$

Putting $\mathcal{V} = \mathcal{V}^*$ in (31) and comparing with (29) yield the first-order optimality condition of (4) over $\mathcal{W}$:

$$\mathbf{0} \in \partial_{\mathcal{W}} g(\mathcal{V}^*,\mathcal{W})|_{\mathcal{W}=\mathcal{W}^*}. \quad (32)$$

Next we show that $(\mathcal{V}^*,\mathcal{W}^*)$ satisfies the first-order optimality condition of (4) over $\mathcal{V}$. To deal with the constraints in (4b), we define an indicator function as

$$\phi(\mathcal{V}) = \begin{cases} 0, & \text{if } \mathcal{V} \text{ satisfies (4b)}, \\ \infty, & \text{else}, \end{cases} \quad (33)$$

which is lower semi-continuous over $\mathcal{V}$ since (4b) represents a closed constraint set. With (33), the constrained problem (4) can be equivalently written as an unconstrained form, with the cost function expressed as

$$g(\mathcal{V},\mathcal{W}) + \phi(\mathcal{V}) = g^{\text{NS}}(\mathcal{W}) + g^{\text{S}}(\mathcal{V},\mathcal{W}) + \phi(\mathcal{V}), \quad (34)$$

where $g^{\text{NS}}(\mathcal{W}) = \sum_{m=1}^{M}\sum_{n=1}^{N}\left\|\|\mathbf{w}_{m,n}\|_2^2\right\|_0 \alpha_{m,n}$ is the nonsmooth part of $g(\mathcal{V},\mathcal{W})$ in (4a) and $g^{\text{S}}(\mathcal{V},\mathcal{W}) = g(\mathcal{V},\mathcal{W}) - g^{\text{NS}}(\mathcal{W})$ is the smooth part. With $\mathcal{W}$ fixed at $\mathcal{W}_{i_j-1}$, the component of $\mathcal{V}_{i_j}$ is the minimizer of subproblem (5) and equivalently $\mathcal{V}_{i_j}$ is the minimizer of problem (4), thus we have

$$g^{\text{NS}}\left(\mathcal{W}_{i_j-1}\right) + g^{\text{S}}\left(\mathcal{V},\mathcal{W}_{i_j-1}\right) + \phi(\mathcal{V}) \geq g^{\text{NS}}\left(\mathcal{W}_{i_j-1}\right) + g^{\text{S}}\left(\mathcal{V}_{i_j},\mathcal{W}_{i_j-1}\right) + \phi\left(\mathcal{V}_{i_j}\right), \quad \forall \mathcal{V}. \quad (35)$$

Notice that as $j \to \infty$, $\mathcal{W}_{i_j} \to \mathcal{W}^*$, and from (23) of Lemma 2 we also have $\mathcal{W}_{i_j-1} \to \mathcal{W}_{i_j}$, therefore $\mathcal{W}_{i_j-1} \to \mathcal{W}^*$ and hence (35) leads to

$$\begin{aligned} g^{\text{S}}(\mathcal{V},\mathcal{W}^*) + \phi(\mathcal{V}) &= \liminf_{j\to\infty} g^{\text{S}}\left(\mathcal{V},\mathcal{W}_{i_j-1}\right) + \phi(\mathcal{V}) \\ &\geq \liminf_{j\to\infty}\left(g^{\text{S}}\left(\mathcal{V}_{i_j},\mathcal{W}_{i_j-1}\right) + \phi\left(\mathcal{V}_{i_j}\right)\right) \\ &\geq g^{\text{S}}(\mathcal{V}^*,\mathcal{W}^*) + \phi(\mathcal{V}^*), \quad \forall \mathcal{V}, \end{aligned} \quad (36)$$

where the first and third steps are due to the fact that $g^{\text{S}}(\mathcal{V},\mathcal{W})$ is continuous over both $\mathcal{V}$ and $\mathcal{W}$, and $\phi(\mathcal{V})$ is lower semi-continuous over $\mathcal{V}$. From (36), it follows that $\mathcal{V}^*$ is the minimizer of $g^{\text{S}}(\mathcal{V},\mathcal{W}^*) + \phi(\mathcal{V})$. Therefore, we can obtain $\mathbf{0} \in \nabla_{\mathcal{V}} g^{\text{S}}(\mathcal{V},\mathcal{W}^*)|_{\mathcal{V}=\mathcal{V}^*} + \partial_{\mathcal{V}}\phi(\mathcal{V})|_{\mathcal{V}=\mathcal{V}^*}$. Together with the fact that $g^{\text{NS}}(\mathcal{W})$ does not depend on $\mathcal{V}$ and hence $\nabla_{\mathcal{V}} g^{\text{NS}}(\mathcal{W}^*)|_{\mathcal{V}=\mathcal{V}^*}$ is zero, we finally obtain the first-order optimality condition of (4) over $\mathcal{V}$:

$$\mathbf{0} \in \nabla_{\mathcal{V}} g^{\text{NS}}(\mathcal{W}^*)|_{\mathcal{V}=\mathcal{V}^*} + \nabla_{\mathcal{V}} g^{\text{S}}(\mathcal{V},\mathcal{W}^*)|_{\mathcal{V}=\mathcal{V}^*} + \partial_{\mathcal{V}}\phi(\mathcal{V})|_{\mathcal{V}=\mathcal{V}^*}. \quad (37)$$

Combining (32) and (37), the proof is completed. ■

Theorem 3 establishes the convergence property of the subsequence $\left\{\mathcal{V}_{i_j},\mathcal{W}_{i_j}\right\}_{j\in\mathbb{N}}$. However, it is still not sufficient to guarantee the convergence of the whole sequence $\{\mathcal{V}_i,\mathcal{W}_i\}_{i\in\mathbb{N}}$. For example, if there exist two different subsequences that converge to different critical points of problem (4), the whole sequence $\{\mathcal{V}_i,\mathcal{W}_i\}_{i\in\mathbb{N}}$ would still diverge. In the following, by further examining the critical points of (4), we can establish the convergence property of the whole sequence $\{\mathcal{V}_i,\mathcal{W}_i\}_{i\in\mathbb{N}}$.

Theorem 4: (Whole Sequence Converges to Critical Point) Using Algorithm 3, the whole sequence $\{\mathcal{V}_i,\mathcal{W}_i\}_{i\in\mathbb{N}}$ can be guaranteed to converge to a critical point $(\mathcal{V}^*,\mathcal{W}^*)$. Moreover, the sequence of the cost function values $\{g(\mathcal{V}_i,\mathcal{W}_i)\}_{i\in\mathbb{N}}$ converges to $g(\mathcal{V}^*,\mathcal{W}^*)$.

Proof: We begin by showing that the number of limit points of $\{\mathcal{V}_i,\mathcal{W}_i\}_{i\in\mathbb{N}}$ is finite. From the optimality condition (32), any limit point satisfies $\mathbf{0} \in \partial_{\mathcal{W}} g^{\text{NS}}(\mathcal{W}) + \nabla_{\mathcal{W}} g^{\text{S}}(\mathcal{V},\mathcal{W})$, or equivalently

$$\mathbf{0} \in \partial_{\mathbf{w}_{m,n}} g^{\text{NS}}(\mathcal{W}) + \nabla_{\mathbf{w}_{m,n}} g^{\text{S}}(\mathcal{V},\mathcal{W}), \quad \forall m \in \mathcal{M}, \quad \forall n = 1,2,\ldots,N. \quad (38)$$

By noticing $g^{\text{NS}}(\mathcal{W}) = \sum_{m=1}^{M}\sum_{n=1}^{N}\left\|\|\mathbf{w}_{m,n}\|_2^2\right\|_0 \alpha_{m,n}$, it is proved in Appendix C that the limiting subdifferential of $g^{\text{NS}}(\mathcal{W})$ over $\mathbf{w}_{m,n}$ is

$$\partial_{\mathbf{w}_{m,n}} g^{\text{NS}}(\mathcal{W}) = \begin{cases} \mathbb{C}^{L\times 1}, & \text{if } \mathbf{w}_{m,n} = \mathbf{0} \text{ and } \alpha_{m,n} > 0, \\ \{\mathbf{0}\}, & \text{if } \mathbf{w}_{m,n} \neq \mathbf{0} \text{ or } \alpha_{m,n} = 0, \end{cases} \quad \forall m \in \mathcal{M}, \quad \forall n = 1,2,\ldots,N. \quad (39)$$

For the first case of (39), since the limiting subdifferential $\partial_{\mathbf{w}_{m,n}} g^{\text{NS}}(\mathcal{W})|_{\mathbf{w}_{m,n}=\mathbf{0}}$ is the set of all vectors in $\mathbb{C}^{L\times 1}$, $\mathbf{w}_{m,n} = \mathbf{0}$ always makes (38) hold regardless of gradient $\nabla_{\mathbf{w}_{m,n}} g^{\text{S}}(\mathcal{V},\mathcal{W})|_{\mathbf{w}_{m,n}=\mathbf{0}}$. For the second case of (39), since $\partial_{\mathbf{w}_{m,n}} g^{\text{NS}}(\mathcal{W}) = \{\mathbf{0}\}$, (38) reduces to $\nabla_{\mathbf{w}_{m,n}} g^{\text{S}}(\mathcal{V},\mathcal{W}) = \mathbf{0}$. Combining the two cases, it follows that the component of any limit point satisfies either $\mathbf{w}_{m,n} = \mathbf{0}$ or $\nabla_{\mathbf{w}_{m,n}} g^{\text{S}}(\mathcal{V},\mathcal{W}) = \mathbf{0}$. For the MN components in $\mathcal{W}$, there are no more than 2^{MN} configurations depending on each $\mathbf{w}_{m,n} = \mathbf{0}$ or not. Putting a particular configuration into (38), the optimality condition reduces to

$$\nabla_{\mathbf{w}_{m,n}} g^{\text{S}}(\mathcal{V},\mathcal{W}) = \mathbf{0} \quad \text{for those } (m,n) \text{ with } \mathbf{w}_{m,n} \neq \mathbf{0} \text{ or } \alpha_{m,n} = 0. \quad (40)$$

Since $g^{\text{S}}(\mathcal{V},\mathcal{W})$ is in a positive definite quadratic form over $\mathcal{W}$ as shown in (4a), there is at most one $\mathcal{W}$ satisfying (40). Therefore, considering all the configurations, the number of limit points of $\{\mathcal{W}_i\}_{i\in\mathbb{N}}$ is finite. On the other hand, since Algorithm 1 generates a unique $\mathcal{V}_i$ given $\mathcal{W}_{i-1}$, for any limit point of $\{\mathcal{W}_i\}_{i\in\mathbb{N}}$ there is a unique corresponding limit point of $\{\mathcal{V}_i\}_{i\in\mathbb{N}}$. Thus, we can conclude that the number of limit points of $\{\mathcal{V}_i,\mathcal{W}_i\}_{i\in\mathbb{N}}$ is finite.

Then we show the convergence of the whole sequence $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ by contradiction. Suppose that the whole sequence $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ does not converge. From Theorem 3, we know that there exists a subsequence $\left\{\mathcal{V}_{i_j}, \mathcal{W}_{i_j}\right\}_{j\in\mathbb{N}}$ converging to a critical point $(\mathcal{V}^*, \mathcal{W}^*)$. Together with the boundedness of $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ (proved below (16)), there must exist another subsequence $\left\{\mathcal{V}_{i_j+\Delta_j}, \mathcal{W}_{i_j+\Delta_j}\right\}_{j\in\mathbb{N}}$ converging to another critical point other than $(\mathcal{V}^*, \mathcal{W}^*)$, where Δ_j is a finite positive integer satisfying $i_j + \Delta_j < i_{j+1} + \Delta_{j+1}, \forall j \in \mathbb{N}$. Since there are a finite number of limit points, we can define the minimum distance between any two limit points as $3\epsilon > 0$. Then there exists a $J_1 > 0$ such that

$$\sqrt{\left\|\mathcal{V}_{i_j+\Delta_j} - \mathcal{V}^*\right\|_2^2 + \left\|\mathcal{W}_{i_j+\Delta_j} - \mathcal{W}^*\right\|_2^2} > 2\epsilon, \quad \forall j \geqslant J_1. \tag{41}$$

From (41), if we choose a particular $\Delta_j, \forall j \geqslant J_1$ as

$$\Delta_j = \min\left\{ \begin{array}{l} l : \sqrt{\left\|\mathcal{V}_{i_j+l} - \mathcal{V}^*\right\|_2^2 + \left\|\mathcal{W}_{i_j+l} - \mathcal{W}^*\right\|_2^2} > 2\epsilon, \\ l = 1, 2, \ldots \end{array} \right\}, \tag{42}$$

it is proved in Appendix D that the subsequence $\left\{\mathcal{V}_{i_j+\Delta_j-1}, \mathcal{W}_{i_j+\Delta_j-1}\right\}_{j\in\mathbb{N}}$ must converge to $(\mathcal{V}^*, \mathcal{W}^*)$ and hence there exists a $J_2 > 0$ such that

$$\sqrt{\left\|\mathcal{V}_{i_j+\Delta_j-1} - \mathcal{V}^*\right\|_2^2 + \left\|\mathcal{W}_{i_j+\Delta_j-1} - \mathcal{W}^*\right\|_2^2} \leq \epsilon, \quad \forall j \geqslant J_2. \tag{43}$$

Subtracting (43) from (41), and applying the reverse triangle inequality, we have

$$\sqrt{\left\|\mathcal{V}_{i_j+\Delta_j} - \mathcal{V}_{i_j+\Delta_j-1}\right\|_2^2 + \left\|\mathcal{W}_{i_j+\Delta_j} - \mathcal{W}_{i_j+\Delta_j-1}\right\|_2^2} > \epsilon, \quad \forall j \geqslant \max\{J_1, J_2\}, \tag{44}$$

which is contradictory to Lemma 2. Thus, the whole sequence $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ converges to a critical point $(\mathcal{V}^*, \mathcal{W}^*)$.

Finally, we show that the limit value of $g(\mathcal{V}_i, \mathcal{W}_i)$ is $g(\mathcal{V}^*, \mathcal{W}^*)$. Notice that if $g(\mathcal{V}, \mathcal{W})$ is continuous, it immediately holds that $g(\mathcal{V}_i, \mathcal{W}_i) \to g(\mathcal{V}^*, \mathcal{W}^*)$ as $(\mathcal{V}_i, \mathcal{W}_i) \to (\mathcal{V}^*, \mathcal{W}^*)$. However, $g(\mathcal{V}, \mathcal{W})$ in (4a) is discontinuous, thus the limit value of the cost function needs to be further investigated. Since $g^{\text{NS}}(\mathcal{W})$ is lower semi-continuous over $\mathcal{W}$, we have

$$\liminf_{i\to\infty} g^{\text{NS}}(\mathcal{W}_i) \geq g^{\text{NS}}(\mathcal{W}^*). \tag{45}$$

On the other hand, as shown after (20), $\mathcal{W}_i^{(t+1)}$ is the minimizer of $g_H\left(\mathcal{V}_i, \mathcal{W}; \mathcal{W}_i^{(t)}\right)$, thus we have $g_H\left(\mathcal{V}_i, \mathcal{W}_i^{(T+1)}; \mathcal{W}_i^{(T)}\right) \leq g_H\left(\mathcal{V}_i, \mathcal{W}^*; \mathcal{W}_i^{(T)}\right)$. Letting $i \to \infty$ yields

$$\limsup_{i\to\infty} g_H\left(\mathcal{V}_i, \mathcal{W}_i^{(T+1)}; \mathcal{W}_i^{(T)}\right) \leq \limsup_{i\to\infty} g_H\left(\mathcal{V}_i, \mathcal{W}^*; \mathcal{W}_i^{(T)}\right). \tag{46}$$

Notice that $(\mathcal{V}_i, \mathcal{W}_i) \to (\mathcal{V}^*, \mathcal{W}^*)$ as $i \to \infty$, thus the continuous parts of both sides of (46) are equal. After subtracting the continuous parts from both sides of (46), it reduces to

$$\limsup_{i\to\infty} g^{\text{NS}}(\mathcal{W}_i) \leq g^{\text{NS}}(\mathcal{W}^*). \tag{47}$$

Combining (45) and (47) yields $\lim_{i\to\infty} g^{\text{NS}}(\mathcal{W}_i) = g^{\text{NS}}(\mathcal{W}^*)$, and hence

$$\begin{aligned} \lim_{i\to\infty} g(\mathcal{V}_i, \mathcal{W}_i) &= \lim_{i\to\infty} g^{\text{NS}}(\mathcal{W}_i) + \lim_{i\to\infty} g^{\text{S}}(\mathcal{V}_i, \mathcal{W}_i) \\ &= g^{\text{NS}}(\mathcal{W}^*) + g^{\text{S}}(\mathcal{V}^*, \mathcal{W}^*) \\ &= g(\mathcal{V}^*, \mathcal{W}^*). \end{aligned} \tag{48}$$

■

V. Simulation Results

In this section, we evaluate the performance of the proposed first-order algorithm through simulations. For comparison, the simulation results of G-CCP [12] and G-CCP-1 which performs G-CCP for one iteration only, are provided as baselines. All simulations are performed on MATLAB R2014a running on a Windows x64 machine with 2.4 GHz CPU and 8 GB RAM.

The simulation set-up is as follows. The large-scale C-RAN consists $N = 20$ BSs [36], [60] each with $L = 8$ antennas, and $K = 40$ to 200 users. The BSs and the users are independently and uniformly distributed in the region $[-1, 1] \times [-1, 1]$ in km [36]. The path loss at distance d km follows $PL = 148.1 + 37.6\log_{10}(d)$ in dB [12], [14], [26]. The log-normal shadowing is 8 dB and the small-scale channel is subject to Rayleigh fading. The transmit antenna power gain at each BS is 10 dBi. The bandwidth is 10 MHz and the noise power spectral density is -172 dBm/Hz. Each user requests a content independently from a database of $F = 100$ contents [12], [26], where the most popular content is requested with probability 0.5 and the rest 99 contents follows a Zipf distribution with parameter 1 [12]. Each BS stores 10 most popular contents. The target SINR for each multicast group is set as 10 dB. All the simulation results are obtained by averaging over 100 simulation trials, with independent BSs' and users' locations, channel and noise realizations, and content requests in each trial. Unless otherwise specified, this set-up will be used for the following simulations.

The parameters regarding to the proposed algorithm are set as follows. Since the convergence of the overall algorithm is valid for any inner iteration number T when updating $\mathcal{W}$ (Algorithm 2), without loss of generality, we set $T = 1$. The iteration of the overall algorithm terminates when the relative change of the cost function in (4a) is less than 10^{-4}. In principle, ρ should be chosen as a very large number to make problems (3) and (4) to be equivalent. To retain a relatively large ρ regardless of the variation of η, we set $\rho = \eta\sqrt{K}$. In the case that $\rho = \eta\sqrt{K}$ is not large enough to obtain a feasible solution of (3), inspired by the continuation method [47], we can increase ρ by a factor of 10.

Notice that when the SINR requirement is too stringent or the channels of users in different multicast groups are highly correlated, problem (3) can be infeasible [61]. In this paper, we focus on solving the problem when it is feasible [12]. If the problem is infeasible, admission control

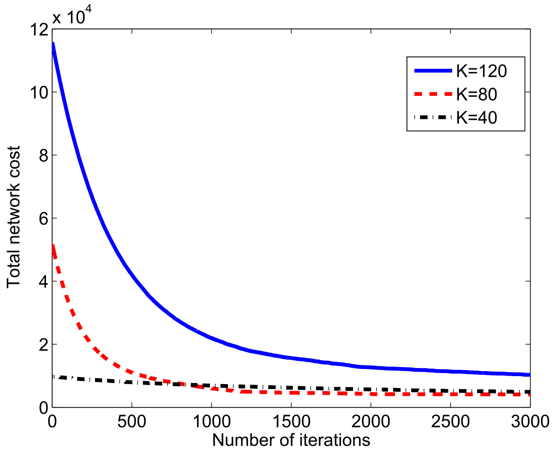


Fig. 2. Convergence behavior of the proposed algorithm with $\eta = 1$.

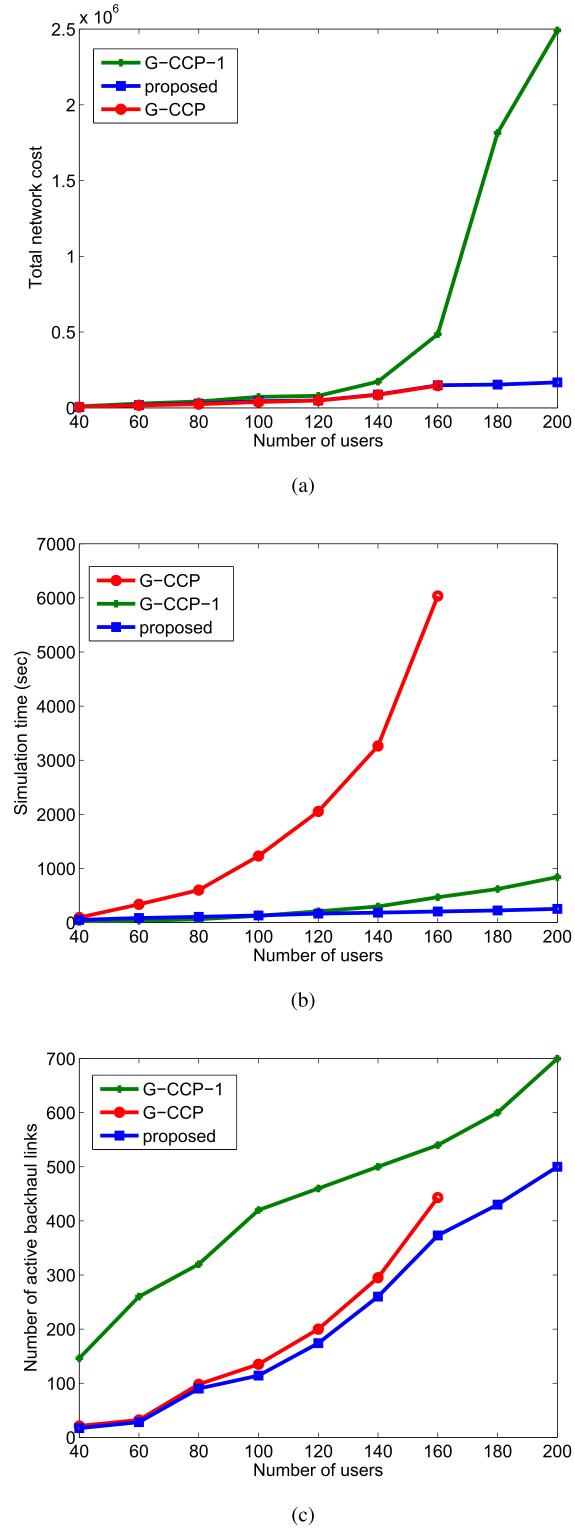
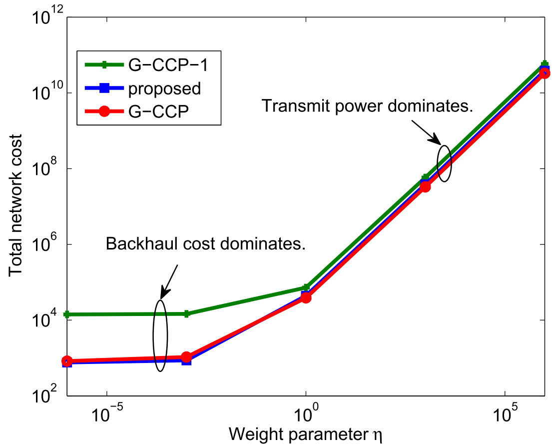


Fig. 3. (a) Total network cost (b) simulation running time (c) active number of backhaul links versus user number with $\eta = 1$.

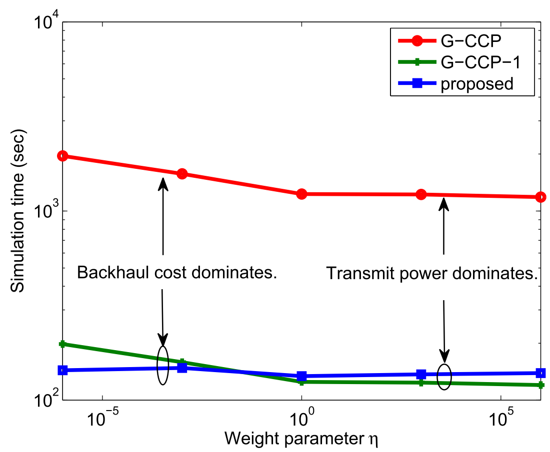
can be considered [24], [62], [63], but this is not the focus of this paper. For fair comparison, all the approaches in the simulations are initialized with the same feasible point of (3) obtained by [53].

First, we show the convergence behavior of the proposed algorithm under a single simulation trial in Fig. 2. For illustration purpose, the weighting parameter η is fixed as 1 and the user number is set as 40, 80, and 120, respectively. It can be seen that the total network cost is monotonic decreasing and convergent as the iteration number increases under different numbers of users. Notice that in Fig. 2, the convergence behavior is observed under a single simulation trial. Due to the randomness of BSs' and users' locations, channel and noise realizations, and content requests, the cost with $K = 80$ may saturate to a lower level than that of $K = 40$. However, if the result is averaged over 100 simulation trials, the cost with $K = 80$ would be higher than that of $K = 40$ (e.g., see Fig. 3(a) below).

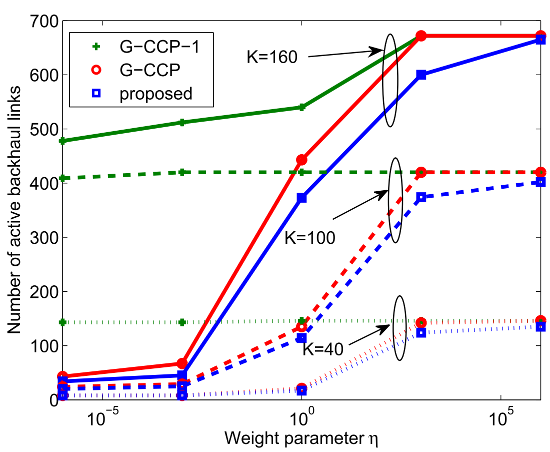
Next, the performance in terms of the total network cost and the simulation time of different approaches versus user number are shown in Fig. 3(a) and Fig. 3(b), respectively. From Fig. 3(a) we can see that the proposed algorithm and G-CCP achieve nearly the same performance in terms of the total network cost, and they are much better than G-CCP-1. However, it can be seen from Fig. 3(b) that the simulation time of the proposed algorithm is about 30 times shorter than that of G-CCP and 2 times shorter than that of G-CCP-1 when the user number $K = 160$. Furthermore, it is obvious that the larger the number of users, the larger the gap between the simulation time of the proposed algorithm and that of G-CCP/G-CCP-1. In fact, due to the extremely long running time, the simulation results of G-CCP for $K > 160$ are not obtained in reasonable time, thus they are omitted in the figures. This demonstrates that the proposed algorithm is more suitable to large-scale C-RAN with hundreds or thousands of users. Notice that the simulation time is measured based on MATLAB implementation. This is for fair comparison in computational complexity with existing algorithms. In commercial/industrial implementation, more efficient programming languages such as C/C++ or Python would be adopted and the codes would also be optimized to achieve even more efficient execution. Furthermore, thanks to the parallel architecture of the proposed algorithm as shown in Fig. 1, its running time can be further reduced by leveraging the modern multi-core computing architecture.



(a)



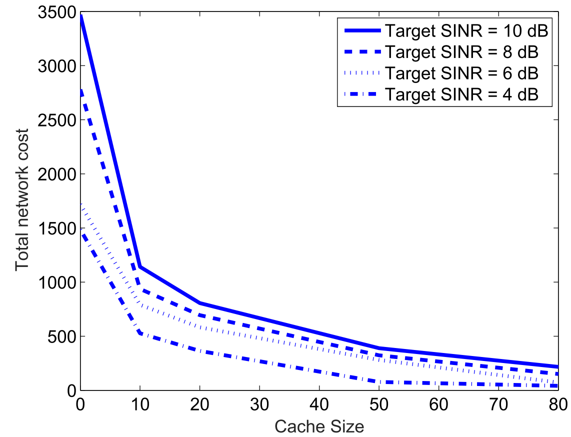
(b)



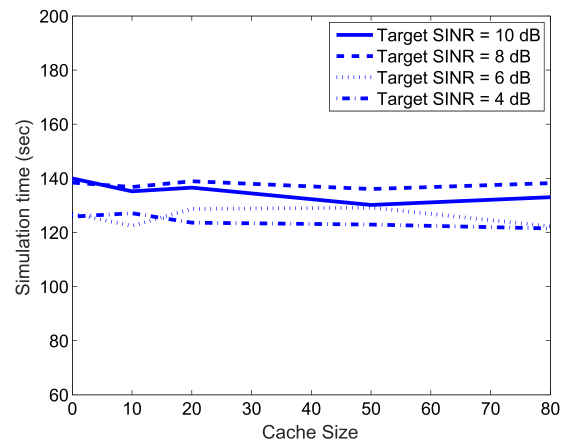
(c)

Fig. 4. (a) Total network cost and (b) simulation running time (c) active number of backhaul links versus weighting parameter.

In addition to the superiority of the simulation time, the solution obtained from the proposed algorithm is also more sparse compared to that obtained from G-CCP. This is evidenced from Fig. 3(c). Moreover, as the user number increases, the sparsity gap between the two approaches increases, which means that the proposed algorithm can better preserve the sparsity of the solutions especially when the number of users is large.



(a)



(b)

Fig. 5. (a) Total network cost and (b) simulation running time versus cache size under different target SINRs, with $\eta = 10^{-3}$ and $K = 100$.

Fig. 4 shows the corresponding simulation results versus the parameter η. From Fig. 4(a), it can be seen that no matter how we set the weighting of the transmit power term with respect to that of the backhaul cost term, the resulting total network cost of the proposed algorithm is close to that of G-CCP, and lower than that of G-CCP-1. On the other hand, as shown in Fig. 4(b) and Fig. 4(c), compared to G-CCP, the proposed algorithm is much faster and the resulting solution is more sparse.

Finally, to demonstrate the impacts of different cache sizes and target SINRs, the total network cost and simulation time with $\eta = 10^{-3}$ and $K = 100$ are shown in Fig. 5. As shown in Fig. 5(a), as the cache size increases, the resulting total network cost becomes lower. This is because caching a larger number of contents at BSs would help to reduce the backhaul cost on content delivery from the computation center. Moreover, the total network cost becomes higher as the target SINR increases. This increase is due to two reasons: 1) higher transmit power is needed to satisfy the higher SINR requirement, 2) higher target SINR means higher target rate which in turn requires more backhaul resource. On the other hand, as shown in Fig. 5(b), the simulation times do not change much since the complexity order depends on the problem size, which remains unchanged under different cache sizes and target SINRs.

VI. Conclusions

A first-order algorithm for content-centric sparse multicast beamforming in large-scale C-RAN was proposed in this paper. The proposed algorithm has linear complexity order with respect to the problem size, and was proved to converge to a critical point. Simulation results were presented to demonstrate that the proposed algorithm performs identically to an existing second-order based approach in terms of the total network cost, but requires much shorter running time, making the proposed algorithm indispensable in truly large-scale networks.

Appendix A

Subproblem (5) is strictly feasible, since the strict inequality in (5b) holds when $v_{k,m} = 0, \forall m \neq m_k$, and $v_{k,m_k} = \sqrt{\gamma_{m_k}}\sigma_k + \varepsilon_k, \forall \varepsilon_k > 0$. Together with the fact that (5) is a QCQP-1 problem, its strong duality holds [52], [53]. Then we can find its optimal solution by analyzing the following KKT conditions:

$$\begin{cases} v_{k,m} - \mathbf{h}_k^H \mathbf{w}_m + \mu_k \gamma_{m_k} v_{k,m} = 0, \quad \forall m \neq m_k,\ m \in \mathcal{M}, \\ v_{k,m_k} - \mathbf{h}_k^H \mathbf{w}_{m_k} - \mu_k v_{k,m_k} = 0, \end{cases} \tag{A.1}$$

$$\gamma_{m_k}\left(\sum_{m\neq m_k} |v_{k,m}|^2 + \sigma_k^2\right) - |v_{k,m_k}|^2 \leq 0, \tag{A.2}$$

$$\mu_k \geq 0, \tag{A.3}$$

$$\mu_k\left(\gamma_{m_k}\left(\sum_{m\neq m_k} |v_{k,m}|^2 + \sigma_k^2\right) - |v_{k,m_k}|^2\right) = 0, \tag{A.4}$$

where μ_k is the dual variable. By working with the dual variable in (A.3), we divide the discussion into $\mu_k = 0$ or $\mu_k > 0$.

If $\mu_k = 0$, (A.3) and (A.4) always hold. Putting $\mu_k = 0$ into (A.1), we have $v_{k,m} = \mathbf{h}_k^H \mathbf{w}_m, \forall m \in \mathcal{M}$. Furthermore, putting $v_{k,m} = \mathbf{h}_k^H \mathbf{w}_m$ into (A.2), we obtain a simplified condition $\gamma_{m_k}\left(\sum_{m\neq m_k} |\mathbf{h}_k^H \mathbf{w}_m|^2 + \sigma_k^2\right) - |\mathbf{h}_k^H \mathbf{w}_{m_k}|^2 \leq 0$ for checking whether $v_{k,m} = \mathbf{h}_k^H \mathbf{w}_m$ is the optimal solution of (5). If this condition is satisfied, it means that $v_{k,m} = \mathbf{h}_k^H \mathbf{w}_m, \forall m \in \mathcal{M}$ is the optimal solution. Otherwise, it cannot be the optimal solution, and we need to consider the other case that $\mu_k > 0$.

If $\mu_k > 0$, it can be proved that there exists one and only one μ_k with $\mathbf{A}_0 + \mu_k \mathbf{A}_1 \succeq 0$ satisfied, where $\mathbf{A}_0$ and $\mathbf{A}_1$ are the matrices of the quadratic terms in the cost function and the constraint of (5), respectively [52]. From (5), $\mathbf{A}_0 = \mathbf{I}_M$ and $\mathbf{A}_1$ is a diagonal matrix whose m_k-th diagonal element is -1, thus we can derive that $1 - \mu_k \geq 0$ and hence (A.2)-(A.4) reduce to the following two conditions:

$$\gamma_{m_k}\left(\sum_{m\neq m_k} |v_{k,m}|^2 + \sigma_k^2\right) - |v_{k,m_k}|^2 = 0, \tag{A.5}$$

$$0 < \mu_k \leq 1. \tag{A.6}$$

Below, we further divide the discussion into two subcases:

1) If $\mathbf{h}_k^H \mathbf{w}_{m_k} = 0$, the second line of (A.1) leads to $\mu_k = 1$ and hence (A.6) is satisfied. Substituting $\mu_k = 1$ into (A.1) and (A.5), we can obtain

$$\begin{cases} v_{k,m} = \dfrac{\mathbf{h}_k^H \mathbf{w}_m}{1+\gamma_{m_k}}, \quad \forall m \neq m_k,\ m \in \mathcal{M}, \\ v_{k,m_k} = \sqrt{\gamma_{m_k}\left(\sum_{m\neq m_k} |v_{k,m}|^2 + \sigma_k^2\right)} e^{j\theta_k}, \end{cases} \tag{A.7}$$

where $\theta_k \in [0, 2\pi]$ is arbitrary. Without loss of generality, we set $\theta_k = 0$.

2) On the other hand, if $\mathbf{h}_k^H \mathbf{w}_{m_k} \neq 0$, the second line of (A.1) leads to $\mu_k \neq 1$, and hence from (A.1) we have

$$\begin{cases} v_{k,m} = \dfrac{\mathbf{h}_k^H \mathbf{w}_m}{1+\mu_k\gamma_{m_k}}, \quad \forall m \neq m_k,\ m \in \mathcal{M}, \\ v_{k,m_k} = \dfrac{\mathbf{h}_k^H \mathbf{w}_{m_k}}{1-\mu_k}. \end{cases} \tag{A.8}$$

Substituting (A.8) into (A.5), we obtain a condition that μ_k needs to satisfy:

$$f(\mu_k) \triangleq \frac{\gamma_{m_k}\sum_{m\neq m_k} |\mathbf{h}_k^H \mathbf{w}_m|^2}{(1+\mu_k\gamma_{m_k})^2} - \frac{|\mathbf{h}_k^H \mathbf{w}_{m_k}|^2}{(1-\mu_k)^2} + \gamma_{m_k}\sigma_k^2 = 0. \tag{A.9}$$

For finding μ_k, notice that

$$f'(\mu_k) = \frac{-2\gamma_{m_k}^2\sum_{m\neq m_k} |\mathbf{h}_k^H \mathbf{w}_m|^2}{(1+\mu_k\gamma_{m_k})^3} - \frac{2|\mathbf{h}_k^H \mathbf{w}_{m_k}|^2}{(1-\mu_k)^3}. \tag{A.10}$$

Combining $\mu_k \neq 1$ with (A.6) yields $0 < \mu_k < 1$. Putting this into (A.10), we have $f'(\mu_k) < 0$ and hence $f(\mu_k)$ is monotonic decreasing in the region $0 < \mu_k < 1$. Thus, μ_k can be found by the bisection method [52].

Appendix B

We begin by showing that the cost function of problem (6) is monotonic decreasing as iteration number t increases. For notational simplicity, we denote $F_m(\mathbf{w}_m)$ and $U_m(\mathbf{w}_m; \tilde{\mathbf{w}}_m)$ as the functions in (6) and (11), respectively, i.e.,

$$F_m(\mathbf{w}_m) \triangleq \underbrace{\sum_{n=1}^{N} \left\| \|\mathbf{w}_{m,n}\|_2^2 \right\|_0 \alpha_{m,n}}_{G_m(\mathbf{w}_m)} + \underbrace{\eta \sum_{n=1}^{N} \|\mathbf{w}_{m,n}\|_2^2 + \rho \sum_{k=1}^{K} \left| v_{k,m} - \mathbf{h}_k^H \mathbf{w}_m \right|^2}_{\tilde{F}_m(\mathbf{w}_m)}, \tag{B.1}$$

$$U_m(\mathbf{w}_m; \tilde{\mathbf{w}}_m) \triangleq \underbrace{\sum_{n=1}^{N} \left\| \|\mathbf{w}_{m,n}\|_2^2 \right\|_0 \alpha_{m,n}}_{G_m(\mathbf{w}_m)}$$

$$\underbrace{+\eta\sum_{n=1}^{N}\|\mathbf{w}_{m,n}\|_2^2+\rho H\sum_{n=1}^{N}\|\mathbf{w}_{m,n}-\mathbf{u}_{m,n}\|_2^2+\rho C}_{\tilde{U}_m(\mathbf{w}_m;\tilde{\mathbf{w}}_m)}, \tag{B.2}$$

where $G_m(\mathbf{w}_m)$ is the discontinuous part, and $\tilde{F}_m(\mathbf{w}_m)$ and $\tilde{U}_m(\mathbf{w}_m;\tilde{\mathbf{w}}_m)$ are the continuous parts of $F_m(\mathbf{w}_m)$ and $U_m(\mathbf{w}_m;\tilde{\mathbf{w}}_m)$, respectively. Applying (9) to (B.1) and putting the definition of C (defined below (10)) into (B.2), we can establish that $F_m(\mathbf{w}_m)\le U_m\left(\mathbf{w}_m;\mathbf{w}_m^{(t)}\right),\forall\mathbf{w}_m$, with the equality holding at $\mathbf{w}_m=\mathbf{w}_m^{(t)}$. It follows that

$$\begin{aligned}F_m\left(\mathbf{w}_m^{(t+1)}\right)&\le U_m\left(\mathbf{w}_m^{(t+1)};\mathbf{w}_m^{(t)}\right)\\&\le U_m\left(\mathbf{w}_m^{(t)};\mathbf{w}_m^{(t)}\right)=F_m\left(\mathbf{w}_m^{(t)}\right),\end{aligned} \tag{B.3}$$

where the second step is due to the fact that $\mathbf{w}_m^{(t+1)}$ is the minimizer of $U_m\left(\mathbf{w}_m;\mathbf{w}_m^{(t)}\right)$. From (B.3), we can conclude that $F_m\left(\mathbf{w}_m^{(t)}\right)$ is monotonic decreasing as t increases.

Next we show that $\left\{F_m\left(\mathbf{w}_m^{(t)}\right)\right\}_{t\in\mathbb{N}}$ converges to $F_m(\mathbf{w}_m^*)$ as $t\to\infty$, where $\mathbf{w}_m^*$ is any limit point of $\left\{\mathbf{w}_m^{(t)}\right\}_{t\in\mathbb{N}}$. Notice that, by using a similar argument under (16), it can be proved that the sequence $\left\{\mathbf{w}_m^{(t)}\right\}_{t\in\mathbb{N}}$ is bounded, thus there exists a subsequence $\left\{\mathbf{w}_m^{(t_j)}\right\}_{j\in\mathbb{N}}$ converging to $\mathbf{w}_m^*$ [59], where $t_j\in\mathbb{N}$ satisfies $t_{j+1}>t_j,\forall j\in\mathbb{N}$. From (B.1) and (B.2), it can be noticed that $G_m(\mathbf{w}_m)$ is lower semi-continuous over $\mathbf{w}_m$, therefore

$$\liminf_{j\to\infty}G_m\left(\mathbf{w}_m^{(t_j)}\right)\ge G_m(\mathbf{w}_m^*). \tag{B.4}$$

On the other hand, noticing that $t_j\in\mathbb{N}$ satisfies $t_{j+1}>t_j,\forall j\in\mathbb{N}$, we have $t_{j+1}\ge t_j+1$. Combining with (B.3) yields

$$\begin{aligned}U_m\left(\mathbf{w}_m^{(t_{j+1})};\mathbf{w}_m^{(t_{j+1})}\right)&=F_m\left(\mathbf{w}_m^{(t_{j+1})}\right)\\&\le F_m\left(\mathbf{w}_m^{(t_j+1)}\right)\\&\le U_m\left(\mathbf{w}_m^{(t_j+1)};\mathbf{w}_m^{(t_j)}\right)\\&\le U_m\left(\mathbf{w}_m;\mathbf{w}_m^{(t_j)}\right),\quad\forall\mathbf{w}_m,\end{aligned} \tag{B.5}$$

where the second step is due to the monotonic decreasing of $F_m\left(\mathbf{w}_m^{(t)}\right)$, and the last step follows from the fact that $\mathbf{w}_m^{(t_j+1)}$ is the minimizer of $U_m\left(\mathbf{w}_m;\mathbf{w}_m^{(t_j)}\right)$. Putting $\mathbf{w}_m=\mathbf{w}_m^*$ into (B.5) and focusing on the left-most and the right-most terms yield

$$\limsup_{j\to\infty}U_m\left(\mathbf{w}_m^{(t_{j+1})};\mathbf{w}_m^{(t_{j+1})}\right)\le\limsup_{j\to\infty}U_m\left(\mathbf{w}_m^*;\mathbf{w}_m^{(t_j)}\right). \tag{B.6}$$

Notice that $\mathbf{w}_m^{(t_j)}\to\mathbf{w}_m^*$ as $j\to\infty$, thus the continuous parts of both sides of (B.6) are equal to $\tilde{U}_m(\mathbf{w}_m^*;\mathbf{w}_m^*)$. After subtracting $\tilde{U}_m(\mathbf{w}_m^*;\mathbf{w}_m^*)$ from both sides of (B.6), it reduces to

$$\limsup_{j\to\infty}G_m\left(\mathbf{w}_m^{(t_{j+1})}\right)\le G_m(\mathbf{w}_m^*). \tag{B.7}$$

Combining (B.4) and (B.7) yields

$$\lim_{j\to\infty}G_m\left(\mathbf{w}_m^{(t_j)}\right)=G_m(\mathbf{w}_m^*). \tag{B.8}$$

Consequently, (B.1) yields $\lim_{j\to\infty}F_m\left(\mathbf{w}_m^{(t_j)}\right)=\lim_{j\to\infty}G_m\left(\mathbf{w}_m^{(t_j)}\right)+\lim_{j\to\infty}\tilde{F}_m\left(\mathbf{w}_m^{(t_j)}\right)=G_m(\mathbf{w}_m^*)+\tilde{F}_m(\mathbf{w}_m^*)=F_m(\mathbf{w}_m^*)$, which means $F_m(\mathbf{w}_m^*)$ is the limit point of $\left\{F_m\left(\mathbf{w}_m^{(t_j)}\right)\right\}_{j\in\mathbb{N}}$. Together with the fact that $F_m\left(\mathbf{w}_m^{(t)}\right)$ is monotonic decreasing and bounded below by 0, we can conclude that $F_m(\mathbf{w}_m^*)$ is the unique limit point of $\left\{F_m\left(\mathbf{w}_m^{(t)}\right)\right\}_{t\in\mathbb{N}}$, i.e., $\lim_{t\to\infty}F_m\left(\mathbf{w}_m^{(t)}\right)=F_m(\mathbf{w}_m^*)$.

Then we prove that $F_m(\mathbf{w}_m^*)$ is at least a local minimum of $F_m(\mathbf{w}_m)$. For this purpose, we first establish that $\mathbf{w}_m^*$ is the minimizer of $U_m(\mathbf{w}_m;\mathbf{w}_m^*)$. From (B.2), it can be noticed that the continuous part $\tilde{U}_m\left(\mathbf{w}_m^{(t_{j+1})};\mathbf{w}_m^{(t_{j+1})}\right)\to\tilde{U}_m(\mathbf{w}_m^*;\mathbf{w}_m^*)$ as $\mathbf{w}_m^{(t_j)}\to\mathbf{w}_m^*$. Adding this result to (B.8), we obtain

$$\lim_{j\to\infty}U_m\left(\mathbf{w}_m^{(t_{j+1})};\mathbf{w}_m^{(t_{j+1})}\right)=U_m(\mathbf{w}_m^*;\mathbf{w}_m^*). \tag{B.9}$$

On the other hand, from (B.5) we have

$$\begin{aligned}\lim_{j\to\infty}U_m\left(\mathbf{w}_m^{(t_{j+1})};\mathbf{w}_m^{(t_{j+1})}\right)&\le\lim_{j\to\infty}U_m\left(\mathbf{w}_m;\mathbf{w}_m^{(t_j)}\right)\\&=U_m(\mathbf{w}_m;\mathbf{w}_m^*),\quad\forall\mathbf{w}_m,\end{aligned} \tag{B.10}$$

where the last equality follows from (B.9). Comparing (B.9) with (B.10), we obtain $U_m(\mathbf{w}_m^*;\mathbf{w}_m^*)\le U_m(\mathbf{w}_m;\mathbf{w}_m^*),\forall\mathbf{w}_m$, which means that $\mathbf{w}_m^*$ is the minimizer of $U_m(\mathbf{w}_m;\mathbf{w}_m^*)$.

With $\mathbf{w}_m^*$ being the minimizer of $U_m(\mathbf{w}_m;\mathbf{w}_m^*)$, we need to further establish the local optimality of $\mathbf{w}_m^*$ in its neighbourhood. Specifically, for any $\boldsymbol{\varepsilon}_m\to\mathbf{0}$ and $\boldsymbol{\varepsilon}_m\neq\mathbf{0}$, using (B.2) we have

$$\begin{aligned}&U_m(\mathbf{w}_m^*+\boldsymbol{\varepsilon}_m;\mathbf{w}_m^*)-U_m(\mathbf{w}_m^*;\mathbf{w}_m^*)\\&=\sum_{n=1}^{N}\alpha_{m,n}\left(\left\|\|\mathbf{w}_{m,n}^*+\boldsymbol{\varepsilon}_{m,n}\|_2^2\right\|_0-\left\|\|\mathbf{w}_{m,n}^*\|_2^2\right\|_0\right)\\&\quad+\eta\sum_{n=1}^{N}\left(\|\mathbf{w}_{m,n}^*+\boldsymbol{\varepsilon}_{m,n}\|_2^2-\|\mathbf{w}_{m,n}^*\|_2^2\right)\\&\quad+\rho H\sum_{n=1}^{N}\left(\|\boldsymbol{\varepsilon}_{m,n}\|_2^2+2\Re\left\{\left(\mathbf{w}_{m,n}^*-\mathbf{u}_{m,n}^*\right)^H\boldsymbol{\varepsilon}_{m,n}\right\}\right)\\&\triangleq\sum_{n=1}^{N}\varphi_{m,n},\end{aligned} \tag{B.11}$$

where $\boldsymbol{\varepsilon}_{m,n}$ is the n-th segment of $\boldsymbol{\varepsilon}_m$, i.e., $\boldsymbol{\varepsilon}_m=\left[\boldsymbol{\varepsilon}_{m,1}^H,\boldsymbol{\varepsilon}_{m,2}^H,\ldots,\boldsymbol{\varepsilon}_{m,N}^H\right]^H$. When $\boldsymbol{\varepsilon}_{m,n}=\mathbf{0}$, from (B.11) we have $\varphi_{m,n}=0$. On the other hand, when $\boldsymbol{\varepsilon}_{m,n}\neq\mathbf{0}$, since $\mathbf{w}_m^*$ is the minimizer of $U_m(\mathbf{w}_m;\mathbf{w}_m^*)$, or equivalently $\mathbf{w}_{m,n}^*$ is the minimizer of (12), by using (14) we can divide the discussion into two cases:

1) If $\left\|\mathbf{u}_{m,n}^*\right\|_2\le\frac{\sqrt{\alpha_{m,n}(\eta+\rho H)}}{\rho H}$, we have $\mathbf{w}_{m,n}^*=\mathbf{0}$, and consequently $\left\|\|\mathbf{w}_{m,n}^*+\boldsymbol{\varepsilon}_{m,n}\|_2^2\right\|_0-\left\|\|\mathbf{w}_{m,n}^*\|_2^2\right\|_0=1$.

Thus, from (B.11) we have

$$\begin{aligned}\varphi_{m,n} &= \alpha_{m,n} + (\eta+\rho H)\left\|\boldsymbol{\varepsilon}_{m,n}\right\|_2^2 \\ &\quad -2\rho H \Re\left\{\left(\mathbf{u}_{m,n}^*\right)^H \boldsymbol{\varepsilon}_{m,n}\right\} \\ &\geq \alpha_{m,n} + (\eta+\rho H)\left\|\boldsymbol{\varepsilon}_{m,n}\right\|_2^2 \\ &\quad -2\sqrt{\alpha_{m,n}(\eta+\rho H)}\left\|\boldsymbol{\varepsilon}_{m,n}\right\|_2 \\ &\geq (\eta+\rho H)\left\|\boldsymbol{\varepsilon}_{m,n}\right\|_2^2, \end{aligned} \tag{B.12}$$

where the second step is due to $\left\|\mathbf{u}_{m,n}^*\right\|_2 \leq \frac{\sqrt{\alpha_{m,n}(\eta+\rho H)}}{\rho H}$, and the last inequality holds as long as $\alpha_{m,n} - 2\sqrt{\alpha_{m,n}(\eta+\rho H)}\left\|\boldsymbol{\varepsilon}_{m,n}\right\|_2 \geq 0$, which obviously holds when $\left\|\boldsymbol{\varepsilon}_{m,n}\right\|_2$ is small enough.

2) If $\left\|\mathbf{u}_{m,n}^*\right\|_2 > \frac{\sqrt{\alpha_{m,n}(\eta+\rho H)}}{\rho H}$, we have $\mathbf{w}_{m,n}^* = \frac{\rho H}{\eta+\rho H}\mathbf{u}_{m,n}^* \neq \mathbf{0}$. Putting this into (B.11), and noticing that $\left\|\left\|\mathbf{w}_{m,n}^* + \boldsymbol{\varepsilon}_{m,n}\right\|_2^2\right\|_0 = \left\|\left\|\mathbf{w}_{m,n}^*\right\|_2^2\right\|_0 = 1$ as long as $\left\|\boldsymbol{\varepsilon}_{m,n}\right\|_2$ is small enough, we have $\varphi_{m,n} = (\eta+\rho H)\left\|\boldsymbol{\varepsilon}_{m,n}\right\|_2^2$.

Therefore, no matter $\boldsymbol{\varepsilon}_{m,n} \neq \mathbf{0}$ or $\boldsymbol{\varepsilon}_{m,n} = \mathbf{0}$, we have

$$U_m\left(\mathbf{w}_m^* + \boldsymbol{\varepsilon}_m; \mathbf{w}_m^*\right) - U_m\left(\mathbf{w}_m^*; \mathbf{w}_m^*\right) \geq (\eta+\rho H)\left\|\boldsymbol{\varepsilon}_m\right\|_2^2. \tag{B.13}$$

On the other hand, by substituting (7) into (B.1) and comparing with (B.2), it can be proved that

$$\begin{aligned} &U_m\left(\mathbf{w}_m^* + \boldsymbol{\varepsilon}_m; \mathbf{w}_m^*\right) \\ &\quad = F_m\left(\mathbf{w}_m^* + \boldsymbol{\varepsilon}_m\right) + \rho\boldsymbol{\varepsilon}_m^H\left(H\mathbf{I}_{NL} - \mathbf{H}\right)\boldsymbol{\varepsilon}_m. \end{aligned} \tag{B.14}$$

Finally, putting (B.14) into the first term of (B.13) and noticing $U_m\left(\mathbf{w}_m^*; \mathbf{w}_m^*\right) = F_m\left(\mathbf{w}_m^*\right)$ yield

$$F_m\left(\mathbf{w}_m^* + \boldsymbol{\varepsilon}_m\right) - F_m\left(\mathbf{w}_m^*\right) \geq \eta\left\|\boldsymbol{\varepsilon}_m\right\|_2^2 + \rho\boldsymbol{\varepsilon}_m^H\mathbf{H}\boldsymbol{\varepsilon}_m \geq 0, \tag{B.15}$$

which means that $F_m\left(\mathbf{w}_m^*\right)$ is at least a local minimum of $F_m\left(\mathbf{w}_m\right)$.

Appendix C

For $x \in \mathrm{dom} f$, the Fréchet subdifferential of $f(x)$ is defined as [41]

$$\hat{\partial} f(x) = \left\{u : \liminf_{y\to x, y\neq x} \frac{f(y) - f(x) - \langle u, y-x\rangle}{\|y-x\|_2} \geq 0\right\},$$

and the limiting subdifferential is $\partial f(x) = \left\{u : \exists x^k \to x,\ f(x^k) \to f(x) \text{ and } u^k \in \hat{\partial} f(x^k) \to u\right\}$.

Rewriting $g^{\mathrm{NS}}(\mathcal{W}) = \sum_{m=1}^M \sum_{n=1}^N \left\|\left\|\mathbf{w}_{m,n}\right\|_2^2\right\|_0 \alpha_{m,n}$ as

$$g^{\mathrm{NS}}(\mathcal{W}) = \begin{cases} \sum_{m'\neq m}\sum_{n'\neq n}\left\|\left\|\mathbf{w}_{m',n'}\right\|_2^2\right\|_0 \alpha_{m',n'}, & \text{if } \mathbf{w}_{m,n} = \mathbf{0}, \\ \sum_{m'\neq m}\sum_{n'\neq n}\left\|\left\|\mathbf{w}_{m',n'}\right\|_2^2\right\|_0 \alpha_{m',n'} + \alpha_{m,n}, & \text{if } \mathbf{w}_{m,n} \neq \mathbf{0}, \end{cases}$$

we can obtain its Fréchet subdifferential as

$$\begin{aligned} &\hat{\partial}_{\mathbf{w}_{m,n}} g^{\mathrm{NS}}(\mathcal{W}) \\ &= \left\{\mathbf{p} : \liminf_{\boldsymbol{\varepsilon}_{m,n}\to\mathbf{0}, \boldsymbol{\varepsilon}_{m,n}\neq\mathbf{0}} \frac{\varpi_{m,n} - \Re\left\{\mathbf{p}^H\boldsymbol{\varepsilon}_{m,n}\right\}}{\left\|\boldsymbol{\varepsilon}_{m,n}\right\|_2} \geq 0\right\}, \end{aligned} \tag{C.1}$$

where

$$\varpi_{m,n} \triangleq \begin{cases} \alpha_{m,n}, & \text{if } \mathbf{w}_{m,n} = \mathbf{0} \text{ and } \alpha_{m,n} > 0, \\ 0, & \text{if } \mathbf{w}_{m,n} \neq \mathbf{0} \text{ or } \alpha_{m,n} = 0. \end{cases} \tag{C.2}$$

Notice that for the first case of (C.2), the inequality in (C.1) holds $\forall \mathbf{p} \in \mathbb{C}^{L\times 1}$. Furthermore, for the second case of (C.2), only $\mathbf{p} = \mathbf{0}$ makes the inequality in (C.1) hold. Thus we have

$$\hat{\partial}_{\mathbf{w}_{m,n}} g^{\mathrm{NS}}(\mathcal{W}) = \begin{cases} \mathbb{C}^{L\times 1}, & \text{if } \mathbf{w}_{m,n} = \mathbf{0} \text{ and } \alpha_{m,n} > 0, \\ \{\mathbf{0}\}, & \text{if } \mathbf{w}_{m,n} \neq \mathbf{0} \text{ or } \alpha_{m,n} = 0. \end{cases} \tag{C.3}$$

Taking the limit, we have $\partial_{\mathbf{w}_{m,n}} g^{\mathrm{NS}}(\mathcal{W}) = \hat{\partial}_{\mathbf{w}_{m,n}} g^{\mathrm{NS}}(\mathcal{W})$ as shown in (C.3).

Appendix D

With Δ_j chosen as (42), suppose that $\left\{\mathcal{V}_{i_j+\Delta_j-1}, \mathcal{W}_{i_j+\Delta_j-1}\right\}_{j\in\mathbb{N}}$ does not converge to $(\mathcal{V}^*, \mathcal{W}^*)$. As the sequence $\{\mathcal{V}_i, \mathcal{W}_i\}_{i\in\mathbb{N}}$ is bounded, there exists a subsequence of $\left\{\mathcal{V}_{i_j+\Delta_j-1}, \mathcal{W}_{i_j+\Delta_j-1}\right\}_{j\in\mathbb{N}}$ converging to a limit point other than $(\mathcal{V}^*, \mathcal{W}^*)$ [59]. Thus, there must exist a $j > J_1$ with $\Delta_j - 1 > 0$ such that $\sqrt{\left\|\mathcal{V}_{i_j+\Delta_j-1} - \mathcal{V}^*\right\|_2^2 + \left\|\mathcal{W}_{i_j+\Delta_j-1} - \mathcal{W}^*\right\|_2^2} > 2\epsilon$. Therefore, we find an l which is smaller than Δ_j but makes the inequality in (42) satisfied. This contradicts with the definition of Δ_j, and by contradiction, $\left\{\mathcal{V}_{i_j+\Delta_j-1}, \mathcal{W}_{i_j+\Delta_j-1}\right\}_{j\in\mathbb{N}}$ converges to $(\mathcal{V}^*, \mathcal{W}^*)$.

References

[1] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: Overview and rel-11 enhancements," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 68–74, Nov. 2012.

[2] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[3] H. Liu, Z. Chen, X. Tian, X. Wang, and M. Tao, "On content-centric wireless delivery networks," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 118–125, Dec. 2014.

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

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

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

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

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

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

[10] Y. Li, Y. Tian, and C. Yang, "Beamforming design with proactive interference cancelation in MISO interference channels," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 67, pp. 1–11, Aug. 2015.

[11] S.-H. Park, O. Simeone, and S. Shamai (Shitz), "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

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

[13] 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.

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

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

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

[17] C. Pan, H. Zhu, N. J. 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.

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

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

[20] 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.

[21] H. Ren, N. Liu, C. Pan, and L. Hanzo, "Joint fronthaul link selection and transmit precoding for energy efficiency maximization of multiuser MIMO-aided distributed antenna systems," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5180–5196, Dec. 2017.

[22] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[23] N. Bornhorst, M. Pesavento, and A. B. Gershman, "Distributed beamforming for multi-group multicasting relay networks," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 221–232, Jan. 2012.

[24] Y. Shi, J. Cheng, J. Zhang, B. Bai, W. Chen, and K. B. Letaief, "Smoothed L_p-minimization for green cloud-RAN with user admission control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1022–1036, Apr. 2016.

[25] O. Tervo, L.-N. Tran, H. Pennanen, S. Chatzinotas, M. Juntti, and B. Ottersten, "Energy-efficient coordinated multi-cell multi-group multicast beamforming with antenna selection," in *Proc. IEEE Int. Conf. Commun. Workshops*, May 2017, pp. 1209–1214.

[26] X. Peng, Y. Shi, J. Zhang, and K. B. Letaief, "Layered group sparse beamforming for cache-enabled green wireless networks," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5589–5603, Dec. 2017.

[27] B. Hu, C. Hua, J. Zhang, C. Chen, and X. Guan, "Joint fronthaul multicast beamforming and user-centric clustering in downlink C-RANs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5395–5409, Aug. 2017.

[28] Y. Ye, "Interior-point polynomial algorithms in convex programming (Y. Nesterov and A. Nemirovskii)," *SIAM Rev.*, vol. 36, no. 6, pp. 682–683, 1994.

[29] Z. Peng, T. Wu, Y. Xu, M. Yan, and W. Yin, "Coordinate friendly structures algorithms and applications," *Ann. Math. Sci. Appl.*, vol. 1, pp. 57–119, Aug. 2016.

[30] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.

[31] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[32] M. Leinonen, M. Codreanu, and M. Juntti, "Distributed joint resource and routing optimization in wireless sensor networks via alternating direction method of multipliers," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5454–5467, Nov. 2013.

[33] H. Tabrizi, B. Peleato, G. Farhadi, J. M. Cioffi, and G. Aldabbagh, "Spatial reuse in dense wireless areas: A cross-layer optimization approach via ADMM," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 7083–7095, Dec. 2015.

[34] B. Luo, L. Cheng, and Y.-C. Wu, "Fully distributed clock synchronization in wireless sensor networks under exponential delays," *Signal Process.*, vol. 125, pp. 261–273, Aug. 2016.

[35] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. New York, NY, USA: Athena Scientific, 1997.

[36] Y. Shi, J. Zhang, B. O'Donoghue, and K. B. Letaief, "Large-scale convex optimization for dense wireless cooperative networks," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4729–4743, Jun. 2015.

[37] C. Shen *et al.* "Distributed robust multicell coordinated beamforming with imperfect CSI: An ADMM approach," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2988–3003, Jun. 2012.

[38] W.-C. Liao, M. Hong, Y.-F. Liu, and Z.-Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3939–3952, Aug. 2014.

[39] E. Chen and M. Tao, "ADMM-based fast algorithm for multi-group multicast beamforming in large-scale wireless systems," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2685–2698, Jun. 2017.

[40] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, 2004.

[41] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, nos. 1–2, pp. 459–494, 2014.

[42] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[43] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[44] B. Chen, C. Yang, and A. F. Molisch, "Cache-enabled device-to-device communications: Offloading gain and energy cost," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4519–4536, Jul. 2017.

[45] B. Zhou, Y. Cui, and M. Tao, "Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2956–2970, Jul. 2017.

[46] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, 2003.

[47] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*. New York, NY, USA: Springer, 2008.

[48] X. Liu, Y. Shi, J. Zhang, and K. B. Letaief, "Massive CSI acquisition for dense cloud-RANs with spatial-temporal dynamics," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2557–2570, Apr. 2018.

[49] T. Van Nguyen, T. Q. S. Quek, and H. Shin, "Joint channel identification and estimation in wireless network: Sparsity and optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3141–3153, May 2018.

[50] Q. He, T. Q. S. Quek, Z. Chen, Q. Zhang, and S. Li, "Compressive channel estimation and multi-user detection in C-RAN with low-complexity methods," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3931–3944, Jun. 2018.

[51] J. Du and Y.-C. Wu, "Network-wide distributed carrier frequency offsets estimation and compensation via belief propagation," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5868–5877, Dec. 2013.

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

[53] K. Huang and N. D. Sidiropoulos, "Consensus-ADMM for general quadratically constrained quadratic programming," *IEEE Trans. Signal Process.*, vol. 64, no. 20, pp. 5297–5310, Oct. 2016.

[54] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

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

[56] Z.-Q. Luo and P. Tseng, "Error bounds and convergence analysis of feasible descent methods: A general approach," *Ann. Oper. Res.*, vol. 46, no. 1, pp. 157–178, 1993.

[57] L. Grippo and M. Sciandrone, "Globally convergent block-coordinate techniques for unconstrained optimization," *Optim. Methods Softw.*, vol. 10, no. 4, pp. 587–637, 1999.
[58] A. Beck, "On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes," *SIAM J. Optim.*, vol. 25, no. 1, pp. 185–209, Jan. 2015.
[59] R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*, 3rd ed. New York, NY, USA: Wiley, 2000.
[60] "C-RAN: The road towards green RAN," China Mobile Res. Inst., Beijing, China, White Paper ver 2.5, Oct. 2011.
[61] Z. Xiang, M. Tao, and X. Wang, "Coordinated multicast beamforming in multicell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 12–31, Jan. 2013.
[62] E. Matskani, N. D. Sidiropoulos, Z. Q. Luo, and L. Tassiulas, "Efficient batch and adaptive approximation algorithms for joint multicast beamforming and admission control," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4882–4894, Dec. 2009.
[63] X. Zhai, L. Zheng, and C. W. Tan, "Energy-infeasibility tradeoff in cognitive radio networks: Price-driven spectrum access algorithms," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 3, pp. 528–538, Mar. 2014.

Yang Li (S'14) received the B.E. and M.E. degrees in electronics engineering from Beihang University, Beijing, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree in electrical and electronic engineering with The University of Hong Kong. His research interests are in the general areas of wireless communications and signal processing.

Minghua Xia (M'12) received the Ph.D. degree in telecommunications and information systems from Sun Yat-sen University, Guangzhou, China, in 2007. Since 2015, he has been a Professor with Sun Yat-sen University.

From 2007 to 2009, he was with the Electronics and Telecommunications Research Institute of South Korea, Beijing Research and Development Center, Beijing, China, where he was a Member and then a Senior Member of Engineering Staff and participated in projects on the physical layer design of 3GPP LTE mobile communications. From 2010 to 2014, he was in sequence with The University of Hong Kong, Hong Kong; the King Abdullah University of Science and Technology, Jeddah, Saudi Arabia; and the Institut National de la Recherche Scientifique, the University of Quebec, Montréal, QC, Canada, as a Post-Doctoral Fellow. He holds two patents granted in China. His research interests are in the general area of 5G wireless communications, and in particular the design and performance analysis of multi-antenna systems, cooperative relaying systems, and cognitive relaying networks, and recently focus on the design and analysis of wireless power transfer and/or energy harvesting systems, as well as massive MIMO and small cells. He received the Professional Award at the IEEE TENCON, held in Macau, in 2015. He was also recognized as an Exemplary Reviewer by the IEEE TRANSACTIONS ON COMMUNICATIONS in 2014, the IEEE COMMUNICATIONS LETTERS in 2014, and the IEEE WIRELESS COMMUNICATIONS LETTERS in 2014 and 2015.

Yik-Chung Wu (S'99–M'05–SM'14) received the B.Eng. degree in electrical and electronics engineering in 1998, the M.Phil. degree from The University of Hong Kong in 2001, and the Ph.D. degree from Texas A&M University, College Station, in 2005. He received the Croucher Foundation Scholarship in 2002 to study for the Ph.D. degree. From 2005 to 2006, he was with the Thomson Corporate Research, Princeton, NJ, USA, as a Member of Technical Staff. Since 2006, he has been with HKU, where he is currently an Associate Professor. He was a Visiting Scholar with Princeton University, in 2015 and 2017. His research interests are in the general areas of signal processing, machine learning, and communication systems. He served as an Editor for the IEEE COMMUNICATIONS LETTERS and the IEEE TRANSACTIONS ON COMMUNICATIONS, and is currently an Editor for the *Journal of Communications and Networks*.