Caching at Base Stations with Multi-Cluster Multicast Wireless Backhaul via Accelerated First-Order Algorithm

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

***Abstract*—Cloud radio access network (C-RAN) has been recognized as a promising architecture for next-generation wireless systems to support the rapidly increasing demand for higher data rate. However, the performance of C-RAN is limited by the backhaul capacities, especially for the wireless deployment. While C-RAN with fixed BS caching has been demonstrated to reduce backhaul consumption, it is more challenging to further optimize the cache allocation at BSs with multi-cluster multicast backhaul, where the inter-cluster interference induces additional non-convexity to the cache optimization problem. Despite the challenges, we propose an accelerated first-order algorithm, which achieves much higher content downloading sum-rate than a second-order algorithm running for the same amount of time. Simulation results demonstrate that, by simultaneously delivering the required contents to different multicast clusters, the proposed algorithm achieves significantly higher downloading sum-rate than those of time-division single-cluster transmission schemes. Moreover, it is found that the proposed algorithm allocates larger cache sizes to the farther BSs within the nearer clusters, which provides insight to the superiority of the proposed cache allocation.**

***Index Terms*—Caching, cloud radio access network (C-RAN), first-order algorithm, large-scale nonsmooth nonconvex optimization, multi-cluster multicast beamforming (MCMB), wireless backhaul.**

I. Introduction

To meet the dramatically increasing demand for higher data rate, cloud radio access network (C-RAN), where the base stations (BSs) are connected to a computation center via high-speed backhaul links for multi-BS cooperation, is a promising architecture for next-generation wireless systems [1]–[3]. However, the performance of C-RAN is mainly limited by the backhaul capacities from the computation center to the BSs, especially for the small-cell deployment, where high-speed optical fiber connections may not be available [4], and wireless backhaul is the only option.

On the other hand, with modern wireless data traffic being more and more dominated by videos and other multimedia data, content-centric communications exploiting multicast transmission and BS caching draw a lot of attention lately [5]–[7]. As multiple BSs in the same cluster share the same users' data for BS cooperation, by multicasting users' messages from the computation center to these BSs simultaneously, the broadcast nature of the wireless backhaul channels can be efficiently exploited. Furthermore, by proactively caching a fraction of popular contents at each BS, the amount of data to be delivered through the wireless backhaul is reduced, thus improving the system efficiency in terms of content downloading rate [8].

While C-RAN with BS caching has been investigated in [9]–[12], they all assume fixed cache allocation among BSs and focus on how BS caching helps to improve the system performance. In particular, [9]–[11] investigate how BS caching facilitates the reduction of backhaul burden and power consumption. In [12], data-sharing and compression are combined to examine how BS caching help in improving the spectral efficiency. On the other hand, although cache optimization has been studied in [13]–[17], they only focus on the layer between the BS and the users, without considering the limitation of the backhaul efficiency. To the best of our knowledge, only the pioneering work [8] investigates cache optimization at BSs aiming at improving the backhaul efficiency between the computation center and the BSs.

Unfortunately, since [8] only considered a simplified C-RAN setup with a single cluster of BSs, the resulting caching scheme could not directly generalize to the more practical scenario with multiple BS clusters. For the multi-cluster scenario, the computation center is required to transmit different multicast data to different BS clusters simultaneously, resulting in inter-cluster interference, which in turn induces additional non-convexity to the cache optimization problem.

Despite the challenges mentioned above, this paper optimizes the cache allocation at BSs for a C-RAN with multi-cluster multicast backhaul, aiming to maximize the content downloading sum-rate of the wireless backhaul under a total cache budget constraint. Since cache placement impacts a much larger timescale than that of channel variations [4], [18], [19], the cache allocation should be optimized based on a

Manuscript received March 7, 2019; revised August 20, 2019 and October 29, 2019; accepted January 18, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61671488, in part by the Major Science and Technology Special Project of Guangdong Province under Grant 2018B010114001, and in part by the Fundamental Research Funds for the Central Universities under Grant 191gjc04. The associate editor coordinating the review of this paper and approving it for publication was A. Abrardo.

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, ycwu}@eee.hku.hk).

M. Xia is with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, 510006, China, and with Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (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

large number of potential channel realizations. Furthermore, to maximize the content downloading sum-rate, various channel realizations requires tailored optimal beamformers, which are coupled in the optimization of cache sizes. Consequently, with a large number of beamfomers being nuisance variables, the cache allocation is a large-scale nonsmooth nonconvex problem. To solve this problem, we first tackle the non-smoothness and non-convexity by introducing auxiliary variables and constructing a sequence of quadratic convex functions in the successive convex approximation (SCA) framework. But instead of directly solving each convexified problem with the interior-point method, we further construct a strongly convex upper bound of the cost function, so that an accelerated first-order algorithm can be developed for solving each SCA subproblem in its dual domain.

Simulation results show that the proposed accelerated first-order algorithm achieves much higher content downloading sum-rate than a second-order algorithm running for the same amount of time. Moreover, by simultaneously delivering the required contents to different multicast clusters, the proposed algorithm achieves significantly higher downloading sum-rate than those of time-division single-cluster transmission schemes. Finally, it is found that the proposed algorithm proactively allocates larger cache sizes to the farther BSs within the nearer clusters, which provides insight to the superiority of the proposed cache allocation.

The remainder of this paper is organized as follows. System model and problem formulation are introduced in Section II. In Section III, an accelerated first-order algorithm is proposed for the cache allocation. The multi-cluster multicast beamforming (MCMB) design for content delivery is presented in Section IV. Simulation results and discussions are provided in Section V. Section VI concludes the paper.

Throughout this paper, scalars, vectors, and matrices are denoted by lower-case letters (e.g., a), lowercase bold letters (e.g., $\mathbf{a}$), and upper bold letters (e.g., $\mathbf{A}$), respectively. The complex domain is denoted by $\mathbb{C}$. We denote the transpose and conjugate transpose of a vector/matrix by $(\cdot)^T$ and $(\cdot)^H$, respectively. The real part, trace, and Frobenius-norm of a matrix are denoted by $\Re(\cdot)$, $\mathrm{Tr}(\cdot)$, and $\|\cdot\|_F$, respectively. The expectation of a random variable is denoted by $\mathbb{E}[\cdot]$, and the complex Gaussian distribution is represented as $\mathcal{CN}(\cdot,\cdot)$.

II. System Model and Problem Formulation

Consider a downlink C-RAN with G clusters of BSs connected to a computation center through wireless backhaul. To effectively utilize the wireless medium, the computation center adopts multicast beamforming to deliver users' intended messages to each cluster, and the BSs in each cluster serve their users through cooperative transmission with data sharing [20], [21]. Furthermore, to alleviate the backhaul burden, each BS is equipped with a local cache to pre-store a subset of popular files. An example of such a cache enabled downlink C-RAN is illustrated in Fig. 1, where the BSs are clustered into G disjoint clusters [22], [23]. In this paper, we assume that the BSs have been clustered, and focus on how to optimally allocate the cache sizes among the BSs to improve the backhaul efficiency.

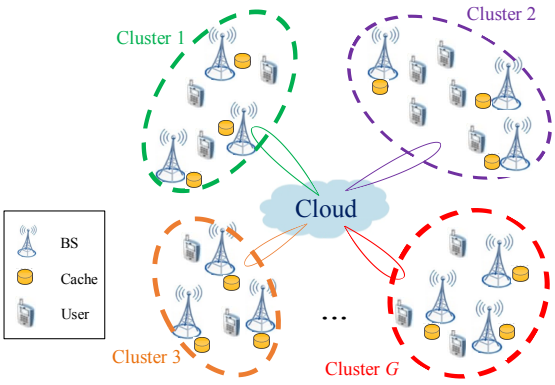


Fig. 1. A downlink C-RAN consists of G clusters of BSs, where each BS is equipped with a local cache.

Let M and N ($M > N$) denote the numbers of antennas at the computation center and each BS, respectively. Then the maximum number of independent data streams of each cluster is $d = \min\{M, N\} = N$, and the multicast beamforming matrix from the computation centre to the g-th cluster of BSs is denoted as $\mathbf{V}_g \in \mathbb{C}^{M\times d}$. Denoting the total number of BSs as K, we can express the received signal at the k-th BS as

$$\mathbf{y}_k = \underbrace{\mathbf{H}_k\mathbf{V}_{g_k}\mathbf{x}_{g_k}}_{\text{desired multicast signal}} + \underbrace{\sum_{g'\neq g_k}\mathbf{H}_k\mathbf{V}_{g'}\mathbf{x}_{g'}}_{\text{inter-cluster interference}} + \mathbf{n}_k, \tag{1}$$

where $\mathbf{H}_k \in \mathbb{C}^{N\times M}$ is the channel matrix from the computation centre to BS k, $g_k \in \{1, 2, \ldots, G\}$ is the index of the group to which BS k belongs, $\mathbf{x}_{g_k} \in \mathbb{C}^{d\times 1}$ is the data vector sent to cluster g_k, and $\mathbf{n}_k \sim \mathcal{CN}\left(\mathbf{0}, \sigma_k^2\mathbf{I}_N\right)$ is the additive white Gaussian noise. Based on (1), the mutual information between the transmit signal $\mathbf{V}_{g_k}\mathbf{x}_{g_k}$ and the received signal $\mathbf{y}_k$ can be written as

$$I\left(\mathbf{V}_{g_k}\mathbf{x}_{g_k};\mathbf{y}_k\right) = \log\det\left(\mathbf{I}_N + \mathbf{H}_k\mathbf{V}_{g_k}\mathbf{V}_{g_k}^H\mathbf{H}_k^H\mathbf{J}_k\right), \tag{2}$$

where $\mathbf{J}_k \triangleq \left(\sum_{g'\neq g_k}\mathbf{H}_k\mathbf{V}_{g'}\mathbf{V}_{g'}^H\mathbf{H}_k^H + \sigma_k^2\mathbf{I}_N\right)^{-1}$.

A central issue in C-RAN is to alleviate the backhaul burden during the peak traffic time [21]. To address this issue, caching highly popular files at BSs during off-peak hours provides a viable solution [10] [24]. However, the network operator has a fixed budget to deploy only a limited amount of total cache size. Due to the limited cache size, each BS pre-stores fractions of popular contents during off-peak hours, and requests the rest from the computation center via wireless backhaul [8] [25]. Specifically, BS k caches the first C_k bits of the file requested by cluster g_k. Denote F_{g_k} as the total size of the file requested by cluster g_k, then BS k requires to receive the rest $F_{g_k} - C_k$ bits of the file from the computation center when the file is delivered to mobile users [8] [25]. With the knowledge of cached content at the BSs, an efficient joint cache-channel coding strategy [25] results in the content downloading rate of cluster g as $R_g = \min_{k\in\mathcal{K}_g}\left\{\frac{I(\mathbf{V}_g\mathbf{x}_g;\mathbf{y}_k)}{1-C_k/F_g}\right\}$ [8], where $\mathcal{K}_g$ denotes the set of BSs in cluster g. By substituting the mutual information $I\left(\mathbf{V}_g\mathbf{x}_g;\mathbf{y}_k\right)$ into R_g, the downloading sum-rate

of all the G clusters of BSs can be written as

$$R_{\text{sum}} = \sum_{g=1}^{G} \min_{k\in\mathcal{K}_g} \left\{ \frac{F_g}{F_g - C_k} \cdot \log\det\left(\mathbf{I}_N + \mathbf{H}_k\mathbf{V}_g\mathbf{V}_g^H\mathbf{H}_k^H\mathbf{J}_k\right) \right\}. \quad (3)$$

To improve the backhaul efficiency, the cache sizes $\{C_k\}$ should be allocated to maximize the content downloading sum-rate in (3). However, since cache placement happens in a much larger timescale than scheduling and transmission [14], [26], [27], cache sizes optimization should be based on long-term channel statistics. Furthermore, to maximize the content downloading sum-rate, the cache sizes should also be optimized together with the optimal beamformers. This gives the following cache size allocation problem:

$$\max_{\{C_k\}} \quad \mathbb{E}_{\{\mathbf{H}_k\}}\left[\max_{\{\mathbf{V}_g\}} \sum_{g=1}^{G} \min_{k\in\mathcal{K}_g} \left\{ \frac{F_g}{F_g - C_k} \cdot \log\det\left(\mathbf{I}_N + \mathbf{H}_k\mathbf{V}_g\mathbf{V}_g^H\mathbf{H}_k^H\mathbf{J}_k\right) \right\} \right], \quad (4a)$$

$$\text{s.t.} \sum_{g=1}^{G} \text{Tr}\left(\mathbf{V}_g\mathbf{V}_g^H\right) \le P_{\text{tot}}, \quad (4b)$$

$$\sum_{k=1}^{K} C_k \le C_{\text{tot}}, \quad (4c)$$

$$0 \le C_k \le F_{g_k}, \quad \forall k = 1, 2, \ldots, K, \quad (4d)$$

where P_{tot} is the total budget of the transmit power at the computation center, and C_{tot} is the total budget of the cache size for the whole network. A common approach to tackle the expectation in (4a) is the sample approximation [28], which reformulates (4) as

$$\max_{\{C_k\},\{\mathbf{V}_{g,t}\}} \quad \sum_{t=1}^{T}\sum_{g=1}^{G} \min_{k\in\mathcal{K}_g} \left\{ \frac{F_g}{F_g - C_k} \log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t}\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right) \right\}, \quad (5a)$$

$$\text{s.t.} \sum_{g=1}^{G} \text{Tr}\left(\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\right) \le P_{\text{tot}}, \quad \forall t = 1, 2, \ldots, T, \quad (5b)$$

$$\sum_{k=1}^{K} C_k \le C_{\text{tot}}, \quad (5c)$$

$$0 \le C_k \le F_{g_k}, \quad \forall k = 1, 2, \ldots, K, \quad (5d)$$

where T is the sample size, $\{\mathbf{H}_{k,t}\}_{k=1}^{K}$ are the t-th channel samples, which can be drawn from any given channel distribution or historical channel realizations, $\{\mathbf{V}_{g,t}\}_{g=1}^{G}$ are the corresponding beamformers, and $\mathbf{J}_{k,t} \triangleq \left(\sum_{g'\neq g_k} \mathbf{H}_{k,t}\mathbf{V}_{g',t}\mathbf{V}_{g',t}^H\mathbf{H}_{k,t}^H + \sigma_k^2\mathbf{I}_N\right)^{-1}$.

However, problem (5) is challenging to solve due to three reasons. Firstly, the objective function (5a) is nonsmooth since the content downloading rate of each BS cluster is the minimum over $|\mathcal{K}_g|$ terms. Secondly, the objective function (5a) is also nonconcave due to the nonconcave coupling between $\frac{F_g}{F_g - C_k}$ and $\log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t}\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right)$, and the involvement of $\{\mathbf{V}_{g,t}\}$ in the inter-cluster interference inside the expression of $\mathbf{J}_{k,t}$. Thirdly, since the sample size T is generally large for good approximation, problem (5) is imposed by a large number of variables and constraints, which induces a heavy computational burden.

Remark 1: For multicast transmission, the file requests in the same multicast group should arrive within a very short time period. Although this assumption is a little strong for mobile users, it makes more sense for the considered scenario in this paper, where the multicast receivers are BSs that require data sharing for coordinated multipoint joint processing rather than mobile users. Therefore, it is reasonable to assume that the BSs in the same cluster request the content simultaneously.

Remark 2: If the computation center transmits data to each BS directly, it either transmits the data in a time-division fashion, or transmits multiple beams at the same time. For using the time division transmission, it avoids interference among beams, but it would take a long time to transmit, as one BS is served after another. On the other hand, if multiple beams are transmitted at the same time, the transmission time is shortened, but the interference among beams would cause severe decoding error.

Remark 3: Strictly speaking, C_k is a discrete variable, which makes the optimization problem (5) combinatorial and highly complex. To make it more tractable, as in the most relevant work [8], we set C_k as a continuous variable. Consequently, it is much easier to reveal the insight of caching as shown in Section VI. For practical implementation of the scheme, the solution of C_k can be rounded off to the nearest integer after the continuous optimization problem is solved.

III. Accelerated First-Order Algorithm for Cache Allocation

In this section, we strive to solve the large-scale nonsmooth nonconvex cache size allocation problem (5). Specifically, we first tackle the non-smoothness and non-convexity of problem (5) by introducing auxiliary variables and constructing a sequence of quadratic convex functions in the SCA framework. Then, instead of directly solving each convexified problem with the interior-point method, we further construct a strongly convex upper bound of the cost function, so that an accelerated first-order algorithm is developed for solving each SCA subproblem in its dual domain.

A. Tackling Non-Smoothness and Non-Convexity

We first tackle the non-smoothness of the objective function (5a). Since (5a) is the minimum over $|\mathcal{K}_g|$ terms, by introducing a set of auxiliary variables $\{\eta_{g,t}\}$ such that $\eta_{g,t} \le \frac{1}{F_g - C_k}\log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t}\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right)$, $\forall k \in \mathcal{K}_g$, problem (5) can be equivalently transformed into the

following smooth problem:

$$\min_{\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}} -\sum_{t=1}^{T}\sum_{g=1}^{G} F_g \eta_{g,t}, \tag{6a}$$

$$\text{s.t.} \sum_{g=1}^{G} \mathrm{Tr}\left(\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\right) \le P_{\text{tot}}, \quad \forall t = 1,2,\ldots,T, \tag{6b}$$

$$\sum_{k=1}^{K} C_k \le C_{\text{tot}}, \tag{6c}$$

$$0 \le C_k \le F_{g_k}, \quad \forall k = 1,2,\ldots,K, \tag{6d}$$

$$(F_{g_k} - C_k)\eta_{g_k,t} - \log\det\Big(\mathbf{I}_N + \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\mathbf{V}_{g_k,t}^H \cdot \mathbf{H}_{k,t}^H \mathbf{J}_{k,t}\Big) \le 0, \forall k = 1,2,\ldots,K, \forall t = 1,2,\ldots,T. \tag{6e}$$

However, due to the nonconvex coupling between C_k and $\eta_{g_k,t}$, and the involvement of $\{\mathbf{V}_{g,t}\}$ in the inter-cluster interference inside the expression of $\mathbf{J}_{k,t}$, the constraint (6e) is nonconvex, making problem (6) still challenging to solve. For the special case of $G = 1$, $\mathbf{J}_k$ would reduce to $1/\sigma_k^2\mathbf{I}_N$, which is independent of the variable $\mathbf{V}_1$. Therefore, by introducing an auxiliary variable $\mathbf{W}_1 = \mathbf{V}_1\mathbf{V}_1^H$ and dropping the rank constraint of $\mathbf{W}_1$ [8], (6e) would be convex over $\mathbf{W}_1$. However, for the general case of $G > 1$, since $\mathbf{J}_k$ involves variables $\{\mathbf{V}_g\}$, the above convexity over $\{\mathbf{W}_g\}$ does not hold. Thus, for the general setting of $G > 1$, the non-convexity of (6e) is difficult to tackle.

A prevalent technique to tackle nonconvex constraints is the successive convex approximation (SCA) [29], in which nonconvex constraints are approximated by a sequence of convex constraints. When the nonconvex constraints are in difference of convex (DC) forms, a common approach for convex approximation is the convex-concave procedure (CCP) [30]. Nevertheless, CCP is not applicable to (6e), since it is not in a DC form. To address this issue, we construct a sequence of convex constraints to approximate (6e) by quadratically convexifying the left-hand-side of (6e). Specifically, given any fixed $C_k^{(i)}$, $\eta_{g_k,t}^{(i)}$, and $\left\{\mathbf{V}_{g,t}^{(i)}\right\}_{g=1}^{G}$, we define a convex quadratic function

$$f_{k,t}^{(i)}\left(C_k, \eta_{g_k,t}, \{\mathbf{V}_{g,t}\}_{g=1}^{G}\right) \triangleq \sum_{g=1}^{G} \mathrm{Tr}\left(\mathbf{V}_{g,t}^H \mathbf{A}_{k,t}^{(i)} \mathbf{V}_{g,t}\right) + 2\Re\left\{\mathrm{Tr}\left(\mathbf{B}_{k,t}^{(i)}\mathbf{V}_{g_k,t}\right)\right\} + \frac{\eta_{g_k,t}^2 + C_k^2}{2} + F_{g_k}\eta_{g_k,t} - \left(\eta_{g_k,t}^{(i)} + C_k^{(i)}\right)\left(\eta_{g_k,t} + C_k\right) + b_{k,t}^{(i)}, \tag{7}$$

where $\mathbf{A}_{k,t}^{(i)}$, $\mathbf{B}_{k,t}^{(i)}$, and $b_{k,t}^{(i)}$ are given by

$$\mathbf{A}_{k,t}^{(i)} \triangleq \mathbf{H}_{k,t}^H \mathbf{U}_{k,t}^{(i)} \left(\mathbf{I}_d - \left(\mathbf{U}_{k,t}^{(i)}\right)^H \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}^{(i)}\right)^{-1} \cdot \left(\mathbf{U}_{k,t}^{(i)}\right)^H \mathbf{H}_{k,t}, \tag{8}$$

$$\mathbf{B}_{k,t}^{(i)} \triangleq -\left(\mathbf{I}_d - \left(\mathbf{U}_{k,t}^{(i)}\right)^H \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}^{(i)}\right)^{-1} \cdot \left(\mathbf{U}_{k,t}^{(i)}\right)^H \mathbf{H}_{k,t}, \tag{9}$$

$$b_{k,t}^{(i)} \triangleq \mathrm{Tr}\Bigg(\left(\mathbf{I}_d - \left(\mathbf{U}_{k,t}^{(i)}\right)^H \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}^{(i)}\right)^{-1} \cdot \left(\mathbf{I}_d + \sigma_k^2\left(\mathbf{U}_{k,t}^{(i)}\right)^H \mathbf{U}_{k,t}^{(i)}\right)\Bigg) + \log\det\left(\mathbf{I}_d - \left(\mathbf{U}_{k,t}^{(i)}\right)^H \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}^{(i)}\right) + \frac{\left(\eta_{g_k,t}^{(i)} + C_k^{(i)}\right)^2}{2} - d, \tag{10}$$

with

$$\mathbf{U}_{k,t}^{(i)} \triangleq \left(\sum_{g=1}^{G} \mathbf{H}_{k,t}\mathbf{V}_{g,t}^{(i)}\left(\mathbf{V}_{g,t}^{(i)}\right)^H \mathbf{H}_{k,t}^H + \sigma_k^2\mathbf{I}_N\right)^{-1} \cdot \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}^{(i)}. \tag{11}$$

Then, we can establish two properties of $f_{k,t}^{(i)}\left(C_k, \eta_{g_k,t}, \{\mathbf{V}_{g,t}\}_{g=1}^{G}\right)$ with the following proposition.

Proposition 1. *The defined function* $f_{k,t}^{(i)}\left(C_k, \eta_{g_k,t}, \{\mathbf{V}_{g,t}\}_{g=1}^{G}\right)$ *in* (4) *satisfies:*

(1.1) $f_{k,t}^{(i)}\left(C_k, \eta_{g_k,t}, \{\mathbf{V}_{g,t}\}_{g=1}^{G}\right) \ge (F_{g_k} - C_k)\eta_{g_k,t} - \log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\mathbf{V}_{g_k,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right)$, *where the equality holds at* $C_k = C_k^{(i)}$, $\eta_{g_k} = \eta_{g_k}^{(i)}$, *and* $\mathbf{V}_{g,t} = \mathbf{V}_{g,t}^{(i)}$, $\forall g = 1,2,\ldots,G$.

(1.2) $\frac{\partial}{\partial\tau} f_{k,t}^{(i)}\left(C_k^{(i)}, \eta_{g_k,t}^{(i)}, \left\{\mathbf{V}_{g,t}^{(i)}\right\}_{g=1}^{G}\right) = \frac{\partial}{\partial\tau}\Bigg((F_{g_k} - C_k^{(i)})\eta_{g_k,t}^{(i)} - \log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}^{(i)}\left(\mathbf{V}_{g_k,t}^{(i)}\right)^H \mathbf{H}_{k,t}^H\mathbf{J}_{k,t}^{(i)}\right)\Bigg)$, *where* τ *represents* C_k, η_{g_k}, *or any element of* $\mathbf{V}_{g,t}$, $\forall g = 1,2,\ldots,G$, *and* $\mathbf{J}_{k,t}^{(i)} = \mathbf{J}_{k,t}\big|_{\left\{\mathbf{V}_{g,t}\right\}_{g=1}^{G} = \left\{\mathbf{V}_{g,t}^{(i)}\right\}_{g=1}^{G}}$.

Proof: See Appendix A. ∎

In particular, property **(1.1)** means that the left-hand-sides of the original nonconvex constraints are upper bounded by the left-hand-sides of the constructed convex constraints; while property **(1.2)** means that the gradients of the left-hand-sides of both the constructed convex constraints and the original nonconvex constraints are equal at the expansion points. Based on the two properties in Proposition 1, the left-hand-side of the nonconvex constraint (6e) is upper bounded by $f_{k,t}^{(i)}\left(C_k, \eta_{g_k,t}, \{\mathbf{V}_{g,t}\}_{g=1}^{G}\right)$, and hence (6e) can be successive-

ly approximated by

$$f_{k,t}^{(i)}\left(C_k,\eta_{g_k,t},\{\mathbf{V}_{g,t}\}_{g=1}^G\right)\le 0,\quad \forall k=1,2,\ldots,K,\quad \forall t=1,2,\ldots,T, \tag{12}$$

which is convex since $f_{k,t}^{(i)}\left(C_k,\eta_{g_k,t},\{\mathbf{V}_{g,t}\}_{g=1}^G\right)$ is convex quadratic. With the sequence of convex constraints constructed in (12), problem (6) can be iteratively solved in the SCA framework, with the i-th SCA subproblem written as

$$\begin{aligned}&\left[\left\{C_k^{(i+1)}\right\},\left\{\mathbf{V}_{g,t}^{(i+1)},\eta_{g,t}^{(i+1)}\right\}\right]\\ =\ &\arg\min_{\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}} -\sum_{t=1}^T\sum_{g=1}^G F_g\eta_{g,t}, \end{aligned}\tag{13a}$$

$$\text{s.t.}\ \sum_{g=1}^G \mathrm{Tr}\left(\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\right)\le P_{\text{tot}},\quad \forall t=1,2,\ldots,T, \tag{13b}$$

$$\sum_{k=1}^K C_k\le C_{\text{tot}}, \tag{13c}$$

$$0\le C_k\le F_{g_k},\quad \forall k=1,2,\ldots,K, \tag{13d}$$

$$f_{k,t}^{(i)}\left(C_k,\eta_{g_k,t},\{\mathbf{V}_{g,t}\}_{g=1}^G\right)\le 0,\quad \forall k=1,2,\ldots,K,\quad \forall t=1,2,\ldots,T. \tag{13e}$$

B. First-Order Algorithm in Dual Domain

While problem (13) can be optimally solved with the interior point method, due to the large number of variables and constraints (induced by the large sample size T), such a method would incur a heavy computational cost. To avoid such a heavy computational burden, we strive to develop a first-order algorithm, which alternatively performs a gradient step and a projection step. However, since problem (13) is imposed by coupling constraints (13b), (13c), and (13e), the projection onto (13b)-(13e) would be highly complicated.

To address this issue, we develop another form of the i-th SCA subproblem of (6) by majorizing the cost function (6a) with a strongly convex upper bound. Specifically, given any fixed $\left\{C_k^{(i)}\right\}$ and $\left\{\mathbf{V}_{g,t}^{(i)},\eta_{g,t}^{(i)}\right\}$, the cost function (6a) can be strongly convexified by adding three positive quadratic terms:

$$\begin{aligned}&\Upsilon^{(i)}\left(\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}\right)\\ =\ &-\sum_{t=1}^T\sum_{g=1}^G F_g\eta_{g,t}+\frac{\rho_1}{2}\sum_{t=1}^T\sum_{g=1}^G\left(\eta_{g,t}-\eta_{g,t}^{(i)}\right)^2\\ &+\rho_2\sum_{t=1}^T\sum_{g=1}^G\left\|\mathbf{V}_{g,t}-\mathbf{V}_{g,t}^{(i)}\right\|_F^2\\ &+\frac{\rho_3}{2}\sum_{k=1}^K\left(C_k-C_k^{(i)}\right)^2,\end{aligned}\tag{14}$$

where ρ_1, ρ_2, and ρ_3 are fixed positive parameters. Consequently, (14) serves as a tight upper bound of (6a), with their function values equal at $\{C_k\}=\left\{C_k^{(i)}\right\}$ and $\{\mathbf{V}_{g,t},\eta_{g,t}\}=\left\{\mathbf{V}_{g,t}^{(i)},\eta_{g,t}^{(i)}\right\}$. Following the same procedure for convexifying the constraints of (6) in the last section, another valid i-th SCA subproblem of (6) can be written as

$$\begin{aligned}&\left[\left\{C_k^{(i+1)}\right\},\left\{\mathbf{V}_{g,t}^{(i+1)},\eta_{g,t}^{(i+1)}\right\}\right]\\ =\ &\arg\min_{\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}}\Upsilon^{(i)}\left(\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}\right),\\ &\text{s.t.}\ \ (13\text{b}),(13\text{c}),(13\text{d}),(13\text{e}).\end{aligned}\tag{15}$$

Based on the strong convexity of $\Upsilon^{(i)}\left(\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}\right)$, we can derive the dual problem of (15) in closed-form with the following proposition.

Proposition 2. *The dual problem of* (15) *is*

$$\begin{aligned}\max_{\{\delta_t\},\{\lambda_{k,t}\},\mu}\quad &\Upsilon^{(i)}\left(\{C_k^\diamond\},\{\mathbf{V}_{g,t}^\diamond,\eta_{g,t}^\diamond\}\right)\\ &+\sum_{t=1}^T\delta_t\left(\sum_{g=1}^G\mathrm{Tr}\left(\mathbf{V}_{g,t}^\diamond\left(\mathbf{V}_{g,t}^\diamond\right)^H\right)-P_{tot}\right)\\ &+\sum_{t=1}^T\sum_{k=1}^K\lambda_{k,t}f_{k,t}^{(i)}\left(C_k^\diamond,\eta_{g_k,t}^\diamond,\{\mathbf{V}_{g,t}^\diamond\}_{g=1}^G\right)\\ &+\mu\left(\sum_{k=1}^K C_k^\diamond-C_{tot}\right),\end{aligned}\tag{16a}$$

$$\begin{aligned}&\textit{s.t.}\ \mu\ge 0,\quad \delta_t\ge 0,\quad \lambda_{k,t}\ge 0,\\ &\forall k=1,2,\ldots,K,\quad \forall t=1,2,\ldots,T,\end{aligned}\tag{16b}$$

where $\left\{\mathbf{V}_{g,t}^\diamond,\eta_{g,t}^\diamond\right\}$ *and* $\{C_k^\diamond\}$ *are uniquely given in the following closed forms:*

$$\eta_{g,t}^\diamond=\eta_{g,t}^{(i)}+\frac{\left(1-\sum_{k\in\mathcal{K}_g}\lambda_{k,t}\right)F_g+\sum_{k\in\mathcal{K}_g}\lambda_{k,t}C_k^{(i)}}{\rho_1+\sum_{k\in\mathcal{K}_g}\lambda_{k,t}},\quad \forall g=1,2,\ldots,G,\quad \forall t=1,2,\ldots,T, \tag{17}$$

$$\begin{aligned}\mathbf{V}_{g,t}^\diamond\ =\ &\left((\rho_2+\delta_t)\mathbf{I}_M+\sum_{k=1}^K\lambda_{k,t}\mathbf{A}_{k,t}^{(i)}\right)^{-1}\\ &\cdot\left(\rho_2\mathbf{V}_{g,t}^{(i)}-\sum_{k\in\mathcal{K}_g}\lambda_{k,t}\left(\mathbf{B}_{k,t}^{(i)}\right)^H\right),\\ &\forall g=1,2,\ldots,G,\quad \forall t=1,2,\ldots,T,\end{aligned}\tag{18}$$

$$C_k^\diamond=\min\left\{\max\left\{\left(C_k^{(i)}+\frac{\sum_{t=1}^T\lambda_{k,t}\eta_{g_k,t}^{(i)}-\mu}{\rho_3+\sum_{t=1}^T\lambda_{k,t}}\right),0\right\},F_{g_k}\right\},\forall k=1,2,\ldots,K. \tag{19}$$

Proof: See Appendix B. ∎

Denoting the the dual objective function in (16a) as $D\left(\{\delta_t\},\{\lambda_{k,t}\},\mu\right)$, and noticing that the values of $\left\{\mathbf{V}_{g,t}^\diamond,\eta_{g,t}^\diamond\right\}$ and $\{C_k^\diamond\}$ are uniquely determined by (17)-(19), we can obtain the partial derivatives as

$$\begin{cases}\frac{\partial D}{\partial\delta_t}=\sum_{g=1}^G\mathrm{Tr}\left(\mathbf{V}_{g,t}^\diamond\left(\mathbf{V}_{g,t}^\diamond\right)^H\right)-P_{\text{tot}},\\ \frac{\partial D}{\partial\lambda_{k,t}}=f_{k,t}^{(i)}\left(C_k^\diamond,\eta_{g_k,t}^\diamond,\{\mathbf{V}_{g,t}^\diamond\}_{g=1}^G\right),\\ \frac{\partial D}{\partial\mu}=\sum_{k=1}^K C_k^\diamond-C_{\text{tot}},\end{cases}\tag{20}$$

Algorithm 1 First-Order Algorithm for Solving (15)

1: Compute $\mathbf{A}_{k,t}^{(i)}$, $\mathbf{B}_{k,t}^{(i)}$, and $b_{k,t}^{(i)}$ with (8)-(10), $\forall k = 1, 2, \ldots, K$, $\forall t = 1, 2, \ldots, T$.
2: Initialize $\delta_t = 1$, $\forall t = 1, 2, \ldots, T$; $\lambda_{k,t} = 1$, $\forall k = 1, 2, \ldots, K$, $\forall t = 1, 2, \ldots, T$; $\mu = 1$.
3: **repeat** $(s = 1, 2, \ldots)$
4: Update $\left\{\mathbf{V}_{g,t}^{\diamond}, \eta_{g,t}^{\diamond}\right\}$ and $\left\{C_k^{\diamond}\right\}$ with (17)-(19).
5: Update $\{\delta_t\}$, $\{\lambda_{k,t}\}$, and μ with (21).
6: **until** convergence
7: Output $\left\{\mathbf{V}_{g,t}^{(i+1)}, \eta_{g,t}^{(i+1)}\right\} = \left\{\mathbf{V}_{g,t}^{\diamond}, \eta_{g,t}^{\diamond}\right\}$ and $\left\{C_k^{(i+1)}\right\} = \left\{C_k^{\diamond}\right\}$.

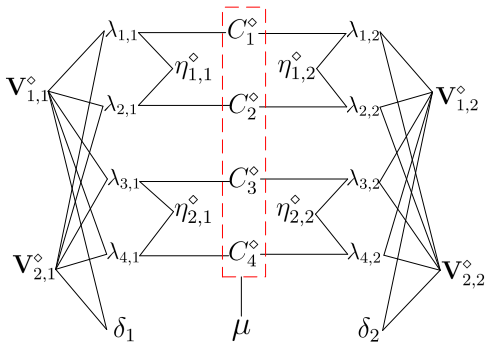


Fig. 2. Parallel structure of Algorithm 1 when $T = 2$, $G = 2$, $K = 4$, $\mathcal{K}_1 = \{1, 2\}$, and $\mathcal{K}_2 = \{3, 4\}$.

thus the projected gradient step increasing $D\left(\{\delta_t\}, \{\lambda_{k,t}\}, \mu\right)$ can be expressed as

$$\begin{cases} \left(\delta_t + \beta_s \left(\sum_{g=1}^{G} \mathrm{Tr}\left(\mathbf{V}_{g,t}^{\diamond}\left(\mathbf{V}_{g,t}^{\diamond}\right)^H\right) - P_{\mathrm{tot}}\right)\right)^+, \\ \left(\lambda_{k,t} + \beta_s f_{k,t}^{(i)}\left(C_k^{\diamond}, \eta_{g_k,t}^{\diamond}, \left\{\mathbf{V}_{g,t}^{\diamond}\right\}_{g=1}^{G}\right)\right)^+, \\ \left(\mu + \beta_s \left(\sum_{k=1}^{K} C_k^{\diamond} - C_{\mathrm{tot}}\right)\right)^+, \end{cases} \tag{21}$$

where β_s is the step size at the s-th iteration, and $(\cdot)^+ \triangleq \max(\cdot, 0)$ is the non-negativity projection over (16b).

By iteratively updating $\{\delta_t\}$, $\{\lambda_{k,t}\}$, and μ with (21), we can obtain the optimal solution to the dual problem (16). Correspondingly, the optimal $\{\mathbf{V}_{g,t}, \eta_{g,t}\}$ and $\{C_k\}$ to the primal problem (15) is given by substituting the optimal $\{\delta_t\}$, $\{\lambda_{k,t}\}$, and μ into (17)-(19). We summarize this procedure for solving problem (15) in Algorithm 1, which is guaranteed to converge to the global optimum of (16) at a rate of $\mathcal{O}(1/s)$, if the step size β_s is smaller than the inverse of the Lipschitz constant of ∇D [31]. Moreover, notice that the primal problem (15) is convex, thus the convergent optimum of (16) is also the global optimum of (15), provided that (15) is strictly feasible [32].

Notice that each iteration of Algorithm 1 can be executed in parallel. In particular, both the $2GT + K$ primal variables in line 4 and the $(K + 1)T + 1$ dual variables in line 5 can be updated in parallel. An example of the parallel structure of Algorithm 1 is shown in Fig. 2, where $T = 2$, $G = 2$, $K = 4$, $\mathcal{K}_1 = \{1, 2\}$, and $\mathcal{K}_2 = \{3, 4\}$. The 12 primal variables $\mathbf{V}_{1,1}^{\diamond}$, $\mathbf{V}_{2,1}^{\diamond}$, $\mathbf{V}_{1,2}^{\diamond}$, $\mathbf{V}_{2,2}^{\diamond}$, $\eta_{1,1}^{\diamond}$, $\eta_{2,1}^{\diamond}$, $\eta_{1,2}^{\diamond}$, $\eta_{2,2}^{\diamond}$, $C_1^{\diamond}$, $C_2^{\diamond}$, $C_3^{\diamond}$, and $C_4^{\diamond}$ can be simultaneously updated. Furthermore, the update of each primal variable only depends on a few dual variables (e.g., the update of $\eta_{1,1}^{\diamond}$ only depends on $\lambda_{1,1}$ and $\lambda_{2,1}$), thus the message passing overhead is small. Similarly, the dual variables can also be simultaneously updated and each depends on only a few primal variables. Due to this parallel structure, Algorithm 1 has the potential of leveraging the modern multi-core multi-thread processor architecture for speeding up the computation.

Algorithm 2 Accelerated First-Order Algorithm for Solving (15)

1: Compute $\mathbf{A}_{k,t}^{(i)}$, $\mathbf{B}_{k,t}^{(i)}$, and $b_{k,t}^{(i)}$ with (8)-(10), $\forall k = 1, 2, \ldots, K$, $\forall t = 1, 2, \ldots, T$.
2: Initialize $\delta_t = 1$, $\forall t = 1, 2, \ldots, T$; $\lambda_{k,t} = 1$, $\forall k = 1, 2, \ldots, K$, $\forall t = 1, 2, \ldots, T$; $\mu = 1$; $\theta^{(0)} = 1$.
3: **repeat** $(s = 1, 2, \ldots)$
4: Update $\theta^{(s)}$ with (24).
5: Update $\left\{\mathbf{V}_{g,t}^{\diamond}, \eta_{g,t}^{\diamond}\right\}$ and $\left\{C_k^{\diamond}\right\}$ with (17)-(19).
6: Update $\left\{\tilde{\delta}_t^{(s)}\right\}$, $\left\{\tilde{\lambda}_{k,t}^{(s)}\right\}$, and $\tilde{\mu}^{(s)}$ with (22).
7: Update $\{\delta_t\}$, $\{\lambda_{k,t}\}$, and μ with (23).
8: **until** convergence
9: Output $\left\{\mathbf{V}_{g,t}^{(i+1)}, \eta_{g,t}^{(i+1)}\right\} = \left\{\mathbf{V}_{g,t}^{\diamond}, \eta_{g,t}^{\diamond}\right\}$ and $\left\{C_k^{(i+1)}\right\} = \left\{C_k^{\diamond}\right\}$.

C. Acceleration with Momentum Technique

Although Algorithm 1 only involves the gradient information and can be executed in parallel, as a first-order algorithm, it may require a large number of iterations to converge. To improve the convergence speed, we further apply the momentum technique [33] to accelerate Algorithm 1. In particular, the projected gradient step in (21) is modified by updating $\tilde{\delta}_t^{(s)}$, $\tilde{\lambda}_{k,t}^{(s)}$, and $\tilde{\mu}^{(s)}$:

$$\begin{cases} \left(\delta_t + \beta_s \left(\sum_{g=1}^{G} \mathrm{Tr}\left(\mathbf{V}_{g,t}^{\diamond}\left(\mathbf{V}_{g,t}^{\diamond}\right)^H\right) - P_{\mathrm{tot}}\right)\right)^+, \\ \left(\lambda_{k,t} + \beta_s f_{k,t}^{(i)}\left(C_k^{\diamond}, \eta_{g_k,t}^{\diamond}, \left\{\mathbf{V}_{g,t}^{\diamond}\right\}_{g=1}^{G}\right)\right)^+, \\ \left(\mu + \beta_s \left(\sum_{k=1}^{K} C_k^{\diamond} - C_{\mathrm{tot}}\right)\right)^+, \end{cases} \tag{22}$$

$$\begin{cases} \delta_t \leftarrow \tilde{\delta}_t^{(s)} + \frac{\theta^{(s-1)}-1}{\theta^{(s)}}\left(\tilde{\delta}_t^{(s)} - \tilde{\delta}_t^{(s-1)}\right), \\ \lambda_{k,t} \leftarrow \tilde{\lambda}_{k,t}^{(s)} + \frac{\theta^{(s-1)}-1}{\theta^{(s)}}\left(\tilde{\lambda}_{k,t}^{(s)} - \tilde{\lambda}_{k,t}^{(s-1)}\right), \\ \mu \leftarrow \tilde{\mu}^{(s)} + \frac{\theta^{(s-1)}-1}{\theta^{(s)}}\left(\tilde{\mu}^{(s)} - \tilde{\mu}^{(s-1)}\right), \end{cases} \tag{23}$$

where $\theta^{(s)}$ is the weighting parameter to dynamically control the momentums $\tilde{\delta}_t^{(s)} - \tilde{\delta}_t^{(s-1)}$, $\tilde{\lambda}_{k,t}^{(s)} - \tilde{\lambda}_{k,t}^{(s-1)}$, and $\tilde{\mu}^{(s)} - \tilde{\mu}^{(s-1)}$. To achieve fast convergence, $\theta^{(s)}$ is updated by [33]

$$\theta^{(s)} = \frac{1 + \sqrt{1 + 4\left(\theta^{(s-1)}\right)^2}}{2}. \tag{24}$$

By using (22)-(24), the accelerated first-order algorithm for solving problem (15) is summarized in Algorithm 2. The key insight of the acceleration lies in the momentums $\tilde{\delta}_t^{(s)} - \tilde{\delta}_t^{(s-1)}$, $\tilde{\lambda}_{k,t}^{(s)} - \tilde{\lambda}_{k,t}^{(s-1)}$, and $\tilde{\mu}^{(s)} - \tilde{\mu}^{(s-1)}$ in (23) (without these momentums, Algorithm 2 would reduce to Algorithm 1). These momentums utilize previous updates to generate an overshoot, so that the update using (22) and (23) in Algorithm 2 is more aggressive than the conventional gradient step (21) in Algorithm 1. On the other hand, to ensure these overshoots to be well behaved, the momentums are controlled by a sequence of weighting parameters $\left\{\theta^{(s)}\right\}$. With $\left\{\theta^{(s)}\right\}$ updated using

Algorithm 3 Overall Algorithm for Solving (5)

1: Initialization:
$\mathbf{V}_{g,t}^{(0)} = \sqrt{\frac{P_{\text{tot}}}{GMd}}\mathbf{1}_{M\times d}, \quad \forall g = 1,2,\ldots,G, \quad \forall t = 1,2,\ldots,T;$
$C_k^{(0)} = C_{\text{tot}}/K, \quad \forall k = 1,2,\ldots,K;$
$\eta_{g,t}^{(0)} = \min_{k\in\mathcal{K}_g}\left\{\frac{1}{F_g - C_k^{(0)}}\log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t}\mathbf{V}_{g,t}^{(0)}\cdot\left(\mathbf{V}_{g,t}^{(0)}\right)^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}^{(0)}\right)\right\}, \quad \forall g = 1,2,\ldots,G, \quad \forall t = 1,2,\ldots,T.$
2: **repeat** $(i = 0,1,\ldots)$
3: Solve the i-th SCA subproblem (15) with Algorithm 2.
4: **until** convergence

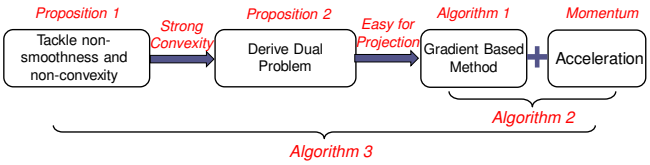


Fig. 3. Proposed algorithm framework for the cache allocation problem (5).

(24), Algorithm 2 is guaranteed to converge to the the global optimum of (16) at a rate of $\mathcal{O}\left(1/s^2\right)$ [33].

D. Overall Algorithm for Solving (5)

With the i-th SCA subproblem (15) solved by Algorithm 2, the overall algorithm for solving the cache size allocation problem (5) is summarized in Algorithm 3. Since the constructed convex constraint (12) satisfies properties **(1.1)** and **(1.2)** of Proposition 1, and the constructed upper bound $\Upsilon^{(i)}\left(\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}\right)$ tightly approximates the cost function (6a), Algorithm 3 is guaranteed to converge to a stationary point of problem (6) [34], which is equivalent to problem (5).

Notice that Algorithm 3 is based on the SCA framework, thus it requires to be initialized from a feasible point. For simplicity, as shown in line 1, we provide a feasible initial point for Algorithm 3 by equally allocating the total transmit power P_{tot} and the total cache sizes C_{tot} to each $\mathbf{V}_{g,t}^{(0)}$ and $C_k^{(0)}$ respectively, and $\eta_{g,t}^{(0)}$ is obtained from (6e) by substituting $\mathbf{V}_{g,t} = \mathbf{V}_{g,t}^{(0)}$ and $C_k = C_k^{(0)}$.

In summary, the extension from the single-cluster scenario [8] to the more general multi-cluster scenario brings two new challenges. First, the inter-cluster interference induces non-convexity in the optimization problem. Although SCA provides a general framework to tackle the non-convexity, it is still challenging to convexify the nonconvex constraint (6e), which is not in the common difference of convex form. To tackle this challenge, we establish Proposition 1, which tightly approximates (6e) by quadratically convexifying its left-hand-side. Second, even after tackling the non-convexity, it is still challenging to solve the resulting SCA subproblem (13) due to its large numbers of coupling constraints, which make the first-order projected gradient descent method not applicable. To get around the coupling constraints, we further establish Proposition 2, which exploits the strong convexity of (14), so that the dual problem (16) can be efficiently solved by the proposed accelerated gradient based method. The overall framework for solving the cache allocation problem (5) is depicted in Fig. 3.

Algorithm 4 Proposed Algorithm for Solving (25)

1: Initialize $\mathbf{V}_g^{(0)} = \sqrt{\frac{P_{\text{tot}}}{GMd}}\mathbf{1}_{M\times d}, \quad \forall g = 1,2,\ldots,G.$
2: **repeat** $(i = 0,1,\ldots)$
3: Compute $\hat{\mathbf{A}}_k^{(i)}$, $\hat{\mathbf{B}}_k^{(i)}$, and $\hat{b}_k^{(i)}$ with (C.2)-(C.4), $\forall k = 1,2,\ldots,K.$
4: Solve the i-th SCA subproblem (27) with CVX.
5: **until** convergence

IV. MCMB Design for Content Delivery

To evaluate the performance of the proposed cache size allocation in Section III, we further design the MCMB for content delivery with fixed $\{C_k\}$. Since only the multicast beamformers $\{\mathbf{V}_g\}$ are optimized for maximizing the instantaneous downloading sum-rate R_{sum} in (3), the MCMB design problem becomes

$$\max_{\{\mathbf{V}_g\}}\sum_{g=1}^{G}\min_{k\in\mathcal{K}_g}\left\{\frac{F_g}{F_g - C_k}\log\det\left(\mathbf{I}_N + \mathbf{H}_k\mathbf{V}_g\mathbf{V}_g^H\mathbf{H}_k^H\mathbf{J}_k\right)\right\}, \tag{25a}$$

$$\text{s.t.} \sum_{g=1}^{G}\text{Tr}\left(\mathbf{V}_g\mathbf{V}_g^H\right) \le P_{\text{tot}}, \tag{25b}$$

which can be equivalently transformed into the following smooth problem:

$$\min_{\{\mathbf{V}_g,\eta_g\}} -\sum_{g=1}^{G}F_g\eta_g, \tag{26a}$$

$$\text{s.t.} \sum_{g=1}^{G}\text{Tr}\left(\mathbf{V}_g\mathbf{V}_g^H\right) \le P_{\text{tot}}, \tag{26b}$$

$$(F_{g_k} - C_k)\eta_{g_k} - \log\det\left(\mathbf{I}_N + \mathbf{H}_k\mathbf{V}_{g_k}\mathbf{V}_{g_k}^H\mathbf{H}_k^H\mathbf{J}_k\right) \le 0, \forall k = 1,2,\ldots,K. \tag{26c}$$

Notice that problem (26) is a simplified version of (6) when $T = 1$ and $\{C_k\}$ are fixed. Following similar derivations to that of the cache allocation problem in the last section (see Appendix C), the i-th SCA subproblem of (26) can be written as

$$\left\{\mathbf{V}_g^{(i+1)},\eta_g^{(i+1)}\right\} = \arg\min_{\{\mathbf{V}_g,\eta_g\}} -\sum_{g=1}^{G}F_g\eta_g, \tag{27a}$$

$$\text{s.t.} \sum_{g=1}^{G}\text{Tr}\left(\mathbf{V}_g\mathbf{V}_g^H\right) \le P_{\text{tot}}. \tag{27b}$$

$$(F_{g_k} - C_k)\eta_{g_k} + h_k^{(i)}\left(\{\mathbf{V}_g\}\right) \le 0, \quad \forall k = 1,2,\ldots,K, \tag{27c}$$

where $h_k^{(i)}\left(\{\mathbf{V}_g\}\right)$ is given in (C.1) of Appendix C. Since subproblem (27) is convex, it can be optimally solved by the standard optimization toolbox based on the interior-point method (e.g., CVX [35]). Notice that different from (13), the SCA subproblem (27) involves only a single channel realization, thus the interior-point method would not induce a heavy computational burden.

The overall framework for solving the MCMB design problem (25) is depicted in Fig. 4, and the algorithm is summarized in Algorithm 4. Without loss of generality, the initialization of

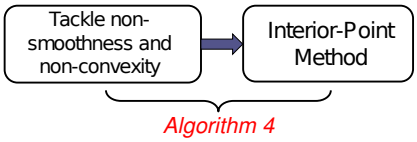


Fig. 4. Proposed algorithm framework for the content delivery problem (25).

Algorithm 4 is obtained by equally allocating the total transmit power P_{tot} to each $\mathbf{V}_g^{(0)}$. Since the constructed convex constraint in (C.6) satisfies properties **C.a** and **C.b** in Appendix C, Algorithm 4 is guaranteed to converge to a stationary point of problem (26) [34].

V. Caching Multiple Files with Different Popularities

While the files were assumed to have the same popularity in Section III, it can be extended to the more general case with multiple files having different popularities. In particular, denote $f_g \in \mathcal{F}_g = \{c_{g,1}, c_{g,2}, \ldots, c_{g,|\mathcal{F}_g|}\}$ as a potential file to be requested by a BS group g with the file's request probability p_{f_g} (with $\sum_{f_g \in \mathcal{F}_g} p_{f_g} = 1$). Let $\mathbf{f} \triangleq [f_1, f_2, \ldots, f_G]^H$ and $\mathcal{F} \triangleq \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \times \mathcal{F}_G$, then the request probability of $\mathbf{f}$ is $p_{\mathbf{f}} = \prod_{g=1}^{G} p_{f_g}$ and we have $\sum_{\mathbf{f}\in\mathcal{F}} p_{\mathbf{f}} = 1$. Each BS k caches $C_{k,f_{g_k}}/F_{f_{g_k}}$ of the file f_{g_k}. Consequently, under the total cache size constraint $\sum_{k=1}^{K}\sum_{f_{g_k}\in\mathcal{F}_{g_k}} C_{k,f_{g_k}} \le C_{\text{tot}}$, the cache size allocation problem with different popularities can be formulated as

$$\max_{\{C_{k,f_g}\},\{\mathbf{V}_{g,t,\mathbf{f}}\}} \sum_{t=1}^{T}\sum_{g=1}^{G}\sum_{\mathbf{f}\in\mathcal{F}} p_{\mathbf{f}} \min_{k\in\mathcal{K}_g} \left\{ \frac{F_{f_g}}{F_{f_g} - C_{k,f_g}} \cdot \log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t,\mathbf{f}}\mathbf{V}_{g,t,\mathbf{f}}\mathbf{V}_{g,t,\mathbf{f}}^H\mathbf{H}_{k,t,\mathbf{f}}^H\mathbf{J}_{k,t,\mathbf{f}}\right)\right\}, \quad (28a)$$

$$\text{s.t.} \sum_{g=1}^{G} \text{Tr}\left(\mathbf{V}_{g,t,\mathbf{f}}\mathbf{V}_{g,t,\mathbf{f}}^H\right) \le P_{\text{tot}}, \quad \forall \mathbf{f}\in\mathcal{F}, \quad \forall t = 1,2,\ldots,T, \quad (28b)$$

$$\sum_{k=1}^{K}\sum_{f_{g_k}\in\mathcal{F}_{g_k}} C_{k,f_{g_k}} \le C_{\text{tot}}, \quad (28c)$$

$$0 \le C_{k,f_{g_k}} \le F_{f_{g_k}}, \quad \forall f_{g_k}\in\mathcal{F}_{g_k}, \quad \forall k = 1,2,\ldots,K, \quad (28d)$$

where $\{\mathbf{H}_{k,t,\mathbf{f}}\}_{k=1}^{K}$ and $\{\mathbf{V}_{g,t,\mathbf{f}}\}_{g=1}^{G}$ are the corresponding channel matrices and beamformers, respectively. Notice that when $p_{\mathbf{f}}$ with different $\mathbf{f}$ are equal, (28) will reduce to (5). With the only difference between (28) and (5) lying in the weighting parameter $p_{\mathbf{f}}$ in (28a), (28) can be solved using the same algorithm framework as Algorithm 3 proposed in Section III. The corresponding algorithm is summarized as Algorithm 5 in Appendix D. Finally, notice that an unpopular file would lead to a small $C_{k,f_{g_k}}$. But no matter $C_{k,f_{g_k}}$ is big or small, for content delivery, as shown in Section IV, whenever a file (even an unpopular file) is requested by the BSs, the computation center must provide multicast transmission immediately. Therefore, there is no delay during the transmission.

VI. Simulation Results and Discussions

In this section, we evaluate the performance of the proposed cache size allocation in terms of downloading sum-rate through simulations. All simulations are performed on MATLAB R2017b running on a Windows x64 machine with 3.3 GHz CPU and 8 GB RAM.

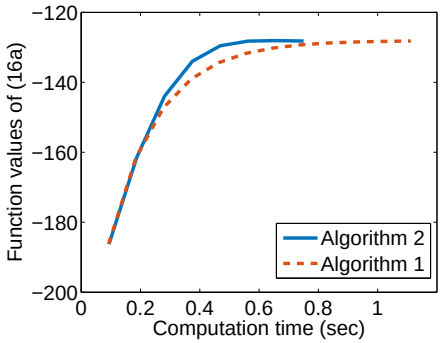
We consider a downlink C-RAN with $G = 4$ clusters of BSs, where each cluster consists of 3 BSs. In particular, the distances between the computation center and the BSs in the 4 clusters are $\{160, 260, 360\}$, $\{200, 280, 360\}$, $\{160, 280, 400\}$, and $\{240, 320, 400\}$ in meters, respectively. The computation center is equipped with $M = 20$ antennas and each BS is equipped with $N = 2$ antennas. The path loss at distance D kilometers follows $128.1 + 37.6\log_{10}(D)$ in dB, and the small-scale channel is subject to Rayleigh fading. We use $T = 100$ sets of channel realizations for solving the cache allocation problem (5). The maximum transmit power at the computation center is $P_{\text{tot}} = 40$ W and the antenna gain is 17 dBi [8]. The backhaul channel bandwidth is 20 MHz and the background noise power spectral density is -150 dBm/Hz. The content size is $F_g = 100, \forall g = 1,2,3,4$, and the budget of the total cache size is $C_{\text{tot}} = 120$.

The step size β_s in Algorithm 1 and Algorithm 2 is fixed as 1, and the parameters in (14) are fixed as $\rho_1 = 10^5$, $\rho_2 = 10^4$, and $\rho_3 = 1$. The iterations of Algorithm 1, Algorithm 2, and Algorithm 4 terminate when the relative changes of the corresponding objective functions between two consecutive iterations are less than 10^{-3}. The SCA iteration of Algorithm 3 stops when the relative decrease of the cost function (6a) in the last 100 iterations is less than 10^{-2}.

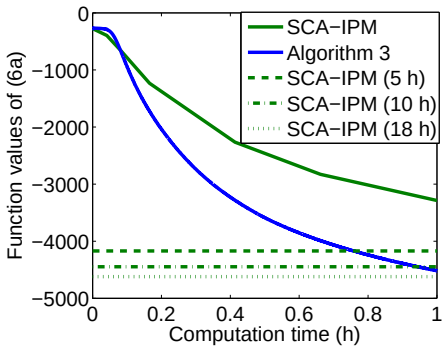
A. Convergence Behaviors of Proposed Algorithms

First, we show the convergence behavior of Algorithm 3 for solving the cache size allocation problem (5). Since the inner iteration of Algorithm 3 is based on Algorithm 1 or Algorithm 2, we illustrate the convergence behaviors of Algorithm 1 and Algorithm 2 in Fig. 5(a). It can be seen that they both converge to the same objective function value, but Algorithm 2 achieves much faster convergence than Algorithm 1, with computation time roughly half of that of Algorithm 1. This demonstrates the effectiveness of the acceleration technique exploited in Algorithm 2. Therefore, we only adopt Algorithm 2 as the inner iteration of Algorithm 3 for the rest of simulations.

On the other hand, the convergence behavior of the outer iteration of Algorithm 3 is shown in Fig. 5(b). To show the complexity advantage of Algorithm 3, we compare it with a second-order algorithm (termed as SCA-IPM), which also applies the SCA framework but solves each SCA subproblem (13) with the interior-point method. It can be seen that with the proposed algorithm and SCA-IPM running for the same amount of time, the proposed algorithm achieves a much lower function value of (6a). Furthermore, in order to achieve the same accuracy, SCM-IPM requires 5 to 18 times more computation time. This demonstrates that Algorithm 3 is more suitable for the cache size allocation problem (5) with a large number of variables and constraints. Notice that the computation time is measured based on MATLAB



(a)



(b)

Fig. 5. (a) Convergence behavior of the inner iteration of Algorithm 3. (b) Convergence behavior of the outer iteration of Algorithm 3.

implementation. In industrial implementation, more efficient programming languages (e.g., C++) would be adopted to achieve even more efficient execution.

B. Performance of Proposed Algorithms on Downloading Sum-Rate

In this subsection, we demonstrate the effectiveness of the proposed cache size allocation in terms of downloading sum-rate (with MCMB design using Algorithm 4). The sum-rate achieved by the proposed cache allocation is compared to that of the uniform cache size allocation. Furthermore, since the recent work [8] is designed for the single-cluster multicast scenario without inter-cluster interference, we compare with a benchmark by extending [8] to the multi-cluster multicast scenario in a time-division multiplexing manner. Specifically, for a G-cluster multicast scenario, to avoid the inter-cluster interference, each cluster utilizes $1/G$ of the total time resources. Consequently, without inter-cluster interference, the cache allocation problem can be directly solved by the proposed algorithm in [8]. To show the importance of modeling the inter-cluster interference, we also add a baseline, which directly applies [8] without modeling the interference for optimization.

Figure 6 illustrates the cumulative distribution functions of downloading sum-rates obtained by different schemes under 400 channel realizations. It can be seen that, by simultaneously delivering the required contents to different multicast clusters, the multi-cluster transmission schemes achieve significantly higher downloading sum-rate than the time-division single-cluster transmission schemes. This demonstrates the necessity

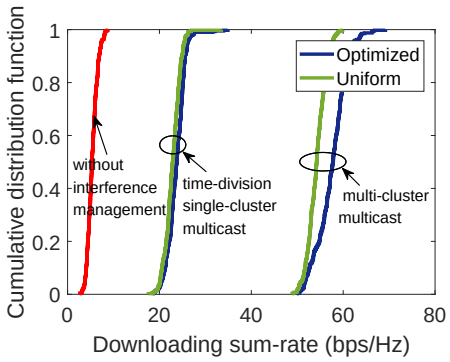


Fig. 6. Cumulative distribution functions of downloading sum-rates obtained by different schemes under 400 channel realizations.

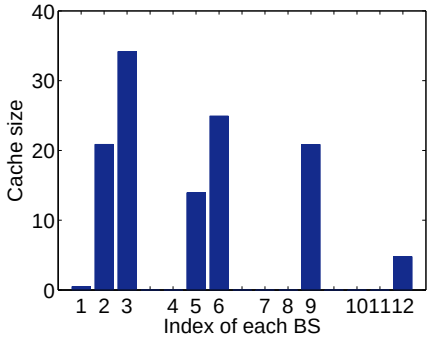


Fig. 7. Cache size allocation results among different BSs. The distances between the computation center and the BSs in the 4 clusters are $\{160, 260, 360\}$, $\{200, 280, 360\}$, $\{160, 280, 400\}$, and $\{240, 320, 400\}$ in meters.

of the multi-cluster multicast transmission for improving the backhaul efficiency. Moreover, with the optimized cache size, the downloading sum-rate achieved by the proposed algorithm is much higher than that of the uniform caching scheme. On the other hand, without effective interference management, the direct application of [8] results in the lowest downloading sum-rate. This demonstrates the necessity of modeling the interference for the multi-cluster multicast scenario. The comparison among various schemes is summarized in Table I. It can be seen that the average downloading sum-rate achieved by the proposed approach is more than 2 times of that of the extension of [8], and more than 10 times of that of the direct application of [8], respectively.

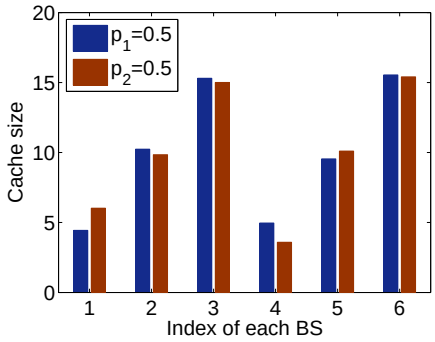
To reveal the insight of the superiority of the proposed cache size allocation, we show the optimized cache sizes among different BSs in Fig. 7. It can be seen that, within a cluster, the cache size allocated by the proposed algorithm increases with the distance between the BS and the computation center. For instance, in Cluster 1, BS 3 is allocated with the largest cache sizes. On the other hand, among different clusters, the proposed algorithm allocates larger cache sizes to the nearer clusters. For example, Cluster 1 and Cluster 4 are allocated with the the largest and the smallest cache sizes, respectively.

Finally, to investigate the impact of file popularity, the allocated cache sizes among different files are shown in Fig. 8, where there are 2 clusters of BSs, and each cluster consists of 3

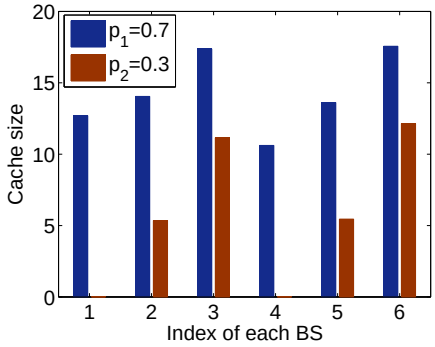
TABLE I
AVERAGE DOWNLOADING SUM-RATE COMPARISON

Schemes	The way to handle multi-cluster interference	Average downloading sum-rate
Proposed approach	Multi-cluster beamforming	59 bps/Hz
Direct application of [8]	No mitigation	5 bps/Hz
Extending [8] to multi-cluster scenario	Time-division multiplexing	26 bps/Hz

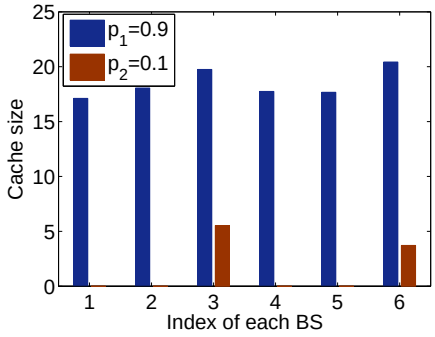
BSs situated in 160, 260, and 360 meters from the computation center. In Fig. 8(a) to Fig. 8(c), each BS requests 2 files with popularities $(p_1 = 0.5, p_2 = 0.5)$, $(p_1 = 0.7, p_2 = 0.3)$, and $(p_1 = 0.9, p_2 = 0.1)$, respectively. It can be seen that the weakest BS 3 and BS 6 are always allocated the largest cache sizes in all the three cases. Furthermore, as the difference between the popularities of the two files increases, the file with higher popularity is allocated much larger cache size. For instance, as shown in Fig. 8(c), when $p_1 = 0.9$ and $p_2 = 0.1$, the file with higher popularity is allocated almost all the cache.



(a) $p_1 = 0.5, p_2 = 0.5$



(b) $p_1 = 0.7, p_2 = 0.3$



(c) $p_1 = 0.9, p_2 = 0.1$

Fig. 8. Cache size allocation results among different files with different popularities.

VII. CONCLUSIONS

Caching at BSs was studied for a C-RAN with multi-cluster multicast backhaul, with the aim of maximizing the content downloading sum-rate of the wireless backhaul under a total cache budget constraint. To solve this large-scale nonsmooth nonconvex problem, we proposed an accelerated first-order algorithm, which achieves much higher content downloading sum-rate than a second-order algorithm running for the same amount of time. Moreover, with multi-cluster multicast transmission, the proposed algorithm achieves significantly higher downloading sum-rate than those of time-division single-cluster transmission schemes. In addition, simulation results revealed that the proposed algorithm allocates larger cache sizes to farther BSs within nearer clusters, which provides insight to the superiority of the proposed cache allocation.

APPENDIX A
PROOF OF PROPOSITION 1

From (4), we can rewrite $f_{k,t}^{(i)}\left(C_k, \eta_{g_k,t}, \{\mathbf{V}_{g,t}\}_{g=1}^G\right) = \phi_{k,t}^{(i)}\left(\{\mathbf{V}_{g,t}\}_{g=1}^G\right) + \psi_{k,t}^{(i)}\left(C_k, \eta_{g_k,t}\right)$, where

$$\phi_{k,t}^{(i)}\left(\{\mathbf{V}_{g,t}\}_{g=1}^G\right) = \sum_{g=1}^{G} \mathrm{Tr}\left(\mathbf{V}_{g,t}^H \mathbf{A}_{k,t}^{(i)} \mathbf{V}_{g,t}\right) + 2\Re\left\{\mathrm{Tr}\left(\mathbf{B}_{k,t}^{(i)} \mathbf{V}_{g_k,t}\right)\right\} + b_{k,t}^{(i)} - \frac{\left(\eta_{g_k,t}^{(i)} + C_k^{(i)}\right)^2}{2}, \quad \text{(A.1)}$$

$$\psi_{k,t}^{(i)}\left(C_k, \eta_{g_k,t}\right) = \frac{\eta_{g_k,t}^2 + C_k^2}{2} - \left(\eta_{g_k,t}^{(i)} + C_k^{(i)}\right) \cdot \left(\eta_{g_k,t} + C_k\right) + F_{g_k}\eta_{g_k,t} + \frac{\left(\eta_{g_k,t}^{(i)} + C_k^{(i)}\right)^2}{2}. \quad \text{(A.2)}$$

From (A.2), we have

$$\psi_{k,t}^{(i)}\left(C_k, \eta_{g_k,t}\right) - \left(F_{g_k} - C_k\right)\eta_{g_k,t} = \frac{1}{2}\left(\eta_{g_k,t} + C_k - \eta_{g_k,t}^{(i)} - C_k^{(i)}\right)^2 \geq 0, \quad \text{(A.3)}$$

where the equality holds at $C_k = C_k^{(i)}$ and $\eta_{g_k} = \eta_{g_k}^{(i)}$.

Next, we prove

$$\phi_{k,t}^{(i)}\left(\{\mathbf{V}_{g,t}\}_{g=1}^G\right) \geq -\log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\mathbf{V}_{g_k,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right). \quad \text{(A.4)}$$

More specifically, applying $\det\left(\mathbf{I} + \mathbf{XY}\right) = \det\left(\mathbf{I} + \mathbf{YX}\right)$, we have

$$\log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\mathbf{V}_{g_k,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right) = \log\det\left(\mathbf{I}_d + \mathbf{V}_{g_k,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\right). \quad \text{(A.5)}$$

Applying the Woodbury matrix identity to the right-hand-side of (A.5), we have

$$\begin{aligned}&\log\det\left(\mathbf{I}_N+\mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\mathbf{V}_{g_k,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right)\\ =\;&\log\det\left(\left(\underbrace{\mathbf{I}_d-\mathbf{U}_{k,t}^H\mathbf{H}_{k,t}\mathbf{V}_{g_k,t}}_{\mathbf{Q}_{k,t}}\right)^{-1}\right),\end{aligned}\tag{A.6}$$

where $\mathbf{U}_{k,t}$ is defined as

$$\mathbf{U}_{k,t}\triangleq\left(\sum_{g=1}^{G}\mathbf{H}_{k,t}\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\mathbf{H}_{k,t}^H+\sigma_k^2\mathbf{I}_N\right)^{-1}\mathbf{H}_{k,t}\mathbf{V}_{g_k,t}.\tag{A.7}$$

Notice that the right-hand-side of (A.6) is convex over $\mathbf{Q}_{k,t}$, thus by using first-order Taylor expansion at $\mathbf{Q}_{k,t}^{(i)}\triangleq\mathbf{I}_d-\left(\mathbf{U}_{k,t}^{(i)}\right)^H\mathbf{H}_{k,t}\mathbf{V}_{g_k,t}^{(i)}$, we have

$$\begin{aligned}&\log\det\left(\mathbf{I}_N+\mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\mathbf{V}_{g_k,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right)\\ \geq\;&\log\det\left(\left(\mathbf{Q}_{k,t}^{(i)}\right)^{-1}\right)-\mathrm{Tr}\left(\left(\mathbf{Q}_{k,t}^{(i)}\right)^{-1}\mathbf{Q}_{k,t}\right)+d.\end{aligned}\tag{A.8}$$

Furthermore, we majorize $\mathbf{Q}_{k,t}$ by

$$\mathbf{E}_{k,t}=\mathbf{Q}_{k,t}+\left(\mathbf{U}_{k,t}^{(i)}-\mathbf{U}_{k,t}\right)^H\mathbf{X}_{k,t}\left(\mathbf{U}_{k,t}^{(i)}-\mathbf{U}_{k,t}\right),\tag{A.9}$$

where $\mathbf{X}_{k,t}\triangleq\sum_{g=1}^{G}\mathbf{H}_{k,t}\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\mathbf{H}_{k,t}^H+\sigma_k^2\mathbf{I}_N$. Substituting (A.9) into (A.8) yields

$$\begin{aligned}&-\log\det\left(\mathbf{I}_N+\mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\mathbf{V}_{g_k,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right)\\ \leq\;&\mathrm{Tr}\left(\left(\mathbf{Q}_{k,t}^{(i)}\right)^{-1}\mathbf{E}_{k,t}\right)-\log\det\left(\left(\mathbf{Q}_{k,t}^{(i)}\right)^{-1}\right)-d.\end{aligned}\tag{A.10}$$

Then we show the right-hand-side of (A.10) is equal to $\phi_{k,t}^{(i)}\left(\{\mathbf{V}_{g,t}\}_{g=1}^G\right)$. Substituting the expressions of $\mathbf{Q}_{k,t}$ in (A.6) and $\mathbf{U}_{k,t}$ in (A.7) into $\mathbf{E}_{k,t}$, we have

$$\mathbf{E}_{k,t}=\left(\mathbf{U}_{k,t}^{(i)}\right)^H\mathbf{X}_{k,t}\mathbf{U}_{k,t}^{(i)}-2\Re\left\{\left(\mathbf{U}_{k,t}^{(i)}\right)^H\mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\right\}+\mathbf{I}_d.\tag{A.11}$$

By applying the expressions of $\mathbf{E}_{k,t}$, $\mathbf{A}_{k,t}^{(i)}$, $\mathbf{B}_{k,t}^{(i)}$, and $b_{k,t}^{(i)}$ in (A.11), (8), (9), and (10), the right-hand-side of (A.10) is equal to $\phi_{k,t}^{(i)}\left(\{\mathbf{V}_{g,t}\}_{g=1}^G\right)$, thus (A.4) is proved.

Now, we show the equality in (A.4) holds at $\{\mathbf{V}_{g,t}\}_{g=1}^G=\left\{\mathbf{V}_{g,t}^{(i)}\right\}_{g=1}^G$. When $\{\mathbf{V}_{g,t}\}_{g=1}^G=\left\{\mathbf{V}_{g,t}^{(i)}\right\}_{g=1}^G$, we have $\mathbf{U}_{k,t}=\mathbf{U}_{k,t}^{(i)}$ and $\mathbf{Q}_{k,t}=\mathbf{Q}_{k,t}^{(i)}$, thus the equality in (A.8) holds. Moreover, from (A.9), seeing that $\mathbf{Q}_{k,t}=\mathbf{E}_{k,t}$ at $\mathbf{U}_{k,t}=\mathbf{U}_{k,t}^{(i)}$, the equality in (A.10) also holds. Since we have shown that the right-hand-side of (A.10) is equal to $\phi_{k,t}^{(i)}\left(\{\mathbf{V}_{g,t}\}_{g=1}^G\right)$, the equality in (A.4) also holds at $\{\mathbf{V}_{g,t}\}_{g=1}^G=\left\{\mathbf{V}_{g,t}^{(i)}\right\}_{g=1}^G$. Therefore, adding up the left-hand-sides of (A.3) and (A.4), we complete the proof for **(1.1)** of Proposition 1.

Finally, we prove **(1.2)** of Proposition 1. From (A.2) we have

$$\frac{\partial\psi_{k,t}^{(i)}\left(C_k^{(i)},\eta_{g_k,t}^{(i)}\right)}{\partial C_k}=-\eta_{g_k,t}^{(i)},\tag{A.12}$$

$$\frac{\partial\psi_{k,t}^{(i)}\left(C_k^{(i)},\eta_{g_k,t}^{(i)}\right)}{\partial\eta_{g_k,t}}=F_{g_k}-C_k^{(i)}.\tag{A.13}$$

On the other hand, since (A.8) is the first-order expansion of $\log\det\left(\mathbf{I}_N+\mathbf{H}_{k,t}\mathbf{V}_{g_k,t}\mathbf{V}_{g_k,t}^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}\right)$ at $\mathbf{Q}_{k,t}=\mathbf{Q}_{k,t}^{(i)}$, we have

$$\begin{aligned}&\frac{\partial}{\partial v}\left(\log\det\left(\mathbf{I}_N+\mathbf{H}_{k,t}\mathbf{V}_{g_{k,t}}^{(i)}\left(\mathbf{V}_{g_{k,t}}^{(i)}\right)^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}^{(i)}\right)\right)\\ =\;&-\mathrm{Tr}\left(\left(\mathbf{Q}_{k,t}^{(i)}\right)^{-1}\frac{\partial\mathbf{Q}_{k,t}^{(i)}}{\partial v}\right),\end{aligned}\tag{A.14}$$

where v represents any element of $\mathbf{V}_{g,t}$, $\forall g=1,2,\ldots,G$. From (A.9), we have

$$\begin{aligned}\frac{\partial\mathbf{E}_{k,t}^{(i)}}{\partial v}=\;&\frac{\partial\mathbf{Q}_{k,t}^{(i)}}{\partial v}+\frac{\partial}{\partial v}\Bigg(\left(\mathbf{U}_{k,t}^{(i)}-\mathbf{U}_{k,t}\right)^H\\ &\cdot\mathbf{X}_{k,t}\Bigg)\Bigg|_{\left\{\mathbf{V}_{g,t}\right\}_{g=1}^G=\left\{\mathbf{V}_{g,t}^{(i)}\right\}_{g=1}^G}\left(\mathbf{U}_{k,t}^{(i)}-\mathbf{U}_{k,t}^{(i)}\right)\\ &+\left(\mathbf{U}_{k,t}^{(i)}-\mathbf{U}_{k,t}^{(i)}\right)^H\mathbf{X}_{k,t}\\ &\frac{\partial}{\partial v}\left(\mathbf{U}_{k,t}^{(i)}-\mathbf{U}_{k,t}\right)\Bigg|_{\left\{\mathbf{V}_{g,t}\right\}_{g=1}^G=\left\{\mathbf{V}_{g,t}^{(i)}\right\}_{g=1}^G}\\ =\;&\frac{\partial\mathbf{Q}_{k,t}^{(i)}}{\partial v}.\end{aligned}\tag{A.15}$$

Substituting (A.15) into (A.14) yields

$$\begin{aligned}&\frac{\partial}{\partial v}\left(\log\det\left(\mathbf{I}_N+\mathbf{H}_{k,t}\mathbf{V}_{g_{k,t}}^{(i)}\left(\mathbf{V}_{g_{k,t}}^{(i)}\right)^H\mathbf{H}_{k,t}^H\mathbf{J}_{k,t}^{(i)}\right)\right)\\ =\;&-\mathrm{Tr}\left(\left(\mathbf{Q}_{k,t}^{(i)}\right)^{-1}\frac{\partial\mathbf{E}_{k,t}^{(i)}}{\partial v}\right)=-\frac{\partial}{\partial v}\phi_{k,t}^{(i)}\left(\{\mathbf{V}_{g,t}\}_{g=1}^G\right),\end{aligned}\tag{A.16}$$

where the second equality follows from the fact that the right-hand-side of (A.10) is equal to $\phi_{k,t}^{(i)}\left(\{\mathbf{V}_{g,t}\}_{g=1}^G\right)$. Combining (A.12), (A.13), and (A.16), we complete the proof for **(1.2)** of Proposition 1.

Appendix B
Proof of Proposition 2

The partial Lagrangian function of (15) is

$$\begin{aligned}&L\left(\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\},\{\delta_t\},\{\lambda_{k,t}\},\mu\right)\\ =\;&\Upsilon^{(i)}\left(\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}\right)\\ &+\sum_{t=1}^{T}\delta_t\left(\sum_{g=1}^{G}\mathrm{Tr}\left(\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\right)-P_{\mathrm{tot}}\right)\\ &+\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_{k,t}f_{k,t}^{(i)}\left(C_k,\eta_{g_k,t},\{\mathbf{V}_{g,t}\}_{g=1}^G\right)\end{aligned}$$

$$+\mu\left(\sum_{k=1}^{K}C_k-C_{\text{tot}}\right), \tag{B.1}$$

where $\{\delta_t\}$, μ, and $\{\lambda_{k,t}\}$ are the non-negative dual variables corresponding to the coupling constraints (13b), (13c), and (13e). Consequently, the dual function of (15) is defined as

$$\begin{aligned}&D\left(\{\delta_t\},\{\lambda_{k,t}\},\mu\right)\triangleq\\&\min_{\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}} L\left(\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\},\{\delta_t\},\{\lambda_{k,t}\},\mu\right),\\&\text{s.t. (13d).}\end{aligned} \tag{B.2}$$

By substituting the expressions of $f_{k,t}^{(i)}\left(C_k,\eta_{g_k,t},\{\mathbf{V}_{g,t}\}_{g=1}^{G}\right)$ in (4) and $\Upsilon^{(i)}\left(\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}\right)$ in (14) into (B.2), problem (B.2) becomes

$$\begin{aligned}&\min_{\{C_k\},\{\mathbf{V}_{g,t},\eta_{g,t}\}} -\sum_{t=1}^{T}\sum_{g=1}^{G}F_g\eta_{g,t}+\frac{\rho_3}{2}\sum_{k=1}^{K}\left(C_k-C_k^{(i)}\right)^2\\&+\frac{\rho_1}{2}\sum_{t=1}^{T}\sum_{g=1}^{G}\left(\eta_{g,t}-\eta_{g,t}^{(i)}\right)^2\\&+\rho_2\sum_{t=1}^{T}\sum_{g=1}^{G}\left\|\mathbf{V}_{g,t}-\mathbf{V}_{g,t}^{(i)}\right\|_F^2+\mu\left(\sum_{k=1}^{K}C_k-C_{\text{tot}}\right)\\&+\sum_{t=1}^{T}\delta_t\left(\sum_{g=1}^{G}\text{Tr}\left(\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\right)-P_{\text{tot}}\right)\\&+\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_{k,t}\Bigg(\sum_{g=1}^{G}\text{Tr}\left(\mathbf{V}_{g,t}^H\mathbf{A}_{k,t}^{(i)}\mathbf{V}_{g,t}\right)\\&+2\Re\left\{\text{Tr}\left(\mathbf{B}_{k,t}^{(i)}\mathbf{V}_{g_k,t}\right)\right\}+\frac{\eta_{g_k,t}^2+C_k^2}{2}+F_{g_k}\eta_{g_k,t}\\&-\left(\eta_{g_k,t}^{(i)}+C_k^{(i)}\right)\left(\eta_{g_k,t}+C_k\right)+b_{k,t}^{(i)}\Bigg),\end{aligned} \tag{B.3a}$$

$$\text{s.t. } 0\le C_k\le F_{g_k},\quad \forall k=1,2,\ldots,K, \tag{B.3b}$$

Due to the variable separability of (B.3a), problem (B.3) can be decomposed into $2GT+K$ subproblems in parallel. Specifically, there are GT subproblems over $\{\eta_{g,t}\}$, with each written as

$$\begin{aligned}\min_{\eta_{g,t}}\quad&-F_g\eta_{g,t}+\frac{\rho_1}{2}\left(\eta_{g,t}-\eta_{g,t}^{(i)}\right)^2\\&+\sum_{k\in\mathcal{K}_g}\lambda_{k,t}\left(\frac{\eta_{g,t}^2}{2}+\left(F_g-\eta_{g,t}^{(i)}-C_k^{(i)}\right)\eta_{g,t}\right).\end{aligned} \tag{B.4}$$

Since the cost function (B.4) is strongly convex over $\eta_{g,t}$, by setting its gradient to zero, the minimizer $\eta_{g,t}^{\diamond}$ is uniquely given by (17). Moreover, there are GT subproblems over $\{\mathbf{V}_{g,t}\}$, with each written as

$$\begin{aligned}\min_{\mathbf{V}_{g,t}}\quad&\rho_2\left\|\mathbf{V}_{g,t}-\mathbf{V}_{g,t}^{(i)}\right\|_F^2+\sum_{k=1}^{K}\lambda_{k,t}\text{Tr}\left(\mathbf{V}_{g,t}^H\mathbf{A}_{k,t}^{(i)}\mathbf{V}_{g,t}\right)\\&+2\sum_{k\in\mathcal{K}_g}\lambda_{k,t}\Re\left\{\text{Tr}\left(\mathbf{B}_{k,t}^{(i)}\mathbf{V}_{g,t}\right)\right\}+\delta_t\text{Tr}\left(\mathbf{V}_{g,t}\mathbf{V}_{g,t}^H\right).\end{aligned} \tag{B.5}$$

Since the cost function (B.5) is strongly convex over $\mathbf{V}_{g,t}$, by setting its gradient to zero, the minimizer $\mathbf{V}_{g,t}^{\diamond}$ is uniquely given by (18). In addition, there are K subproblems over $\{C_k\}$, with each written as

$$\begin{aligned}\min_{0\le C_k\le F_{g_k}}\quad&\mu C_k+\frac{\rho_3}{2}\left(C_k-C_k^{(i)}\right)^2\\&+\sum_{t=1}^{T}\lambda_{k,t}\left(\frac{C_k^2}{2}-\left(\eta_{g_k,t}^{(i)}+C_k^{(i)}\right)C_k\right).\end{aligned} \tag{B.6}$$

Since the cost function (B.6) is strongly convex quadratic over the scalar variable C_k, by setting its gradient to zero and then projecting the solution to $0\le C_k\le F_{g_k}$, the minimizer $C_k^{\diamond}$ is uniquely given by (19). By substituting the optimal solution $\left\{\mathbf{V}_{g,t}^{\diamond},\eta_{g,t}^{\diamond}\right\}$ and $\{C_k^{\diamond}\}$ into (B.2), the dual function $D\left(\{\delta_t\},\{\lambda_{k,t}\},\mu\right)$ is expressed in a closed-form as shown in (16a).

Appendix C
SCA Subproblem of (25)

To tackle the non-convexity of the constraint (26c), we apply the SCA framework by quadratically convexifying $-\log\det\left(\mathbf{I}_N+\mathbf{H}_k\mathbf{V}_{g_k}\mathbf{V}_{g_k}^H\mathbf{H}_k^H\mathbf{J}_k\right)$. Specifically, given any fixed $\left\{\mathbf{V}_g^{(i)}\right\}$, we define a convex quadratic function:

$$\begin{aligned}h_k^{(i)}\left(\{\mathbf{V}_g\}\right)\quad\triangleq\quad&\sum_{g=1}^{G}\text{Tr}\left(\mathbf{V}_g^H\hat{\mathbf{A}}_k^{(i)}\mathbf{V}_g\right)\\&+2\Re\left\{\text{Tr}\left(\hat{\mathbf{B}}_k^{(i)}\mathbf{V}_{g_k}\right)\right\}+\hat{b}_k^{(i)},\end{aligned} \tag{C.1}$$

where $\hat{\mathbf{A}}_k^{(i)}$, $\hat{\mathbf{B}}_k^{(i)}$, and $\hat{b}_k^{(i)}$ are given by

$$\begin{aligned}\hat{\mathbf{A}}_k^{(i)}\quad\triangleq\quad&\mathbf{H}_k^H\hat{\mathbf{U}}_k^{(i)}\left(\mathbf{I}_d-\left(\hat{\mathbf{U}}_k^{(i)}\right)^H\mathbf{H}_k\mathbf{V}_{g_k}^{(i)}\right)^{-1}\\&\cdot\left(\hat{\mathbf{U}}_k^{(i)}\right)^H\mathbf{H}_k,\end{aligned} \tag{C.2}$$

$$\begin{aligned}\hat{\mathbf{B}}_k^{(i)}\quad\triangleq\quad&-\left(\mathbf{I}_d-\left(\hat{\mathbf{U}}_k^{(i)}\right)^H\mathbf{H}_k\mathbf{V}_{g_k}^{(i)}\right)^{-1}\\&\cdot\left(\hat{\mathbf{U}}_k^{(i)}\right)^H\mathbf{H}_k,\end{aligned} \tag{C.3}$$

$$\begin{aligned}\hat{b}_k^{(i)}\quad\triangleq\quad&\text{Tr}\Bigg(\left(\mathbf{I}_d-\left(\hat{\mathbf{U}}_k^{(i)}\right)^H\mathbf{H}_k\mathbf{V}_{g_k}^{(i)}\right)^{-1}\\&\left(\mathbf{I}_d+\sigma_k^2\left(\hat{\mathbf{U}}_k^{(i)}\right)^H\hat{\mathbf{U}}_k^{(i)}\right)\Bigg)\\&+\log\det\left(\mathbf{I}_d-\left(\hat{\mathbf{U}}_k^{(i)}\right)^H\mathbf{H}_k\mathbf{V}_{g_k}^{(i)}\right)-d,\end{aligned} \tag{C.4}$$

with

$$\hat{\mathbf{U}}_k^{(i)}\triangleq\left(\sum_{g=1}^{G}\mathbf{H}_k\mathbf{V}_g^{(i)}\left(\mathbf{V}_g^{(i)}\right)^H\mathbf{H}_k^H+\sigma_k^2\mathbf{I}_N\right)^{-1}\mathbf{H}_k\mathbf{V}_{g_k}^{(i)}. \tag{C.5}$$

Notice that $h_k^{(i)}\left(\{\mathbf{V}_g\}\right)$ is in a similar form to $\phi_{k,t}^{(i)}\left(\{\mathbf{V}_{g,t}\}_{g=1}^{G}\right)$ in (A.1), thus by using similar arguments as in Appendix A, we can establish two properties of $h_k^{(i)}\left(\{\mathbf{V}_g\}\right)$:

C.a $h_k^{(i)}\left(\{\mathbf{V}_g\}\right)\quad\ge\quad-\log\det\left(\mathbf{I}_N+\mathbf{H}_k\mathbf{V}_{g_k}\mathbf{V}_{g_k}^H\mathbf{H}_k^H\mathbf{J}_k\right)$, where the equality holds at $\mathbf{V}_g=\mathbf{V}_g^{(i)}$, $\forall g=1,2,\ldots,G$.

C.b $\frac{\partial}{\partial v}h_k^{(i)}\left(\left\{\mathbf{V}_g^{(i)}\right\}\right) = -\frac{\partial}{\partial v}\Big(\log\det\Big(\mathbf{I}_N + \mathbf{H}_k\mathbf{V}_{g_k}^{(i)}\left(\mathbf{V}_{g_k}^{(i)}\right)^H\mathbf{H}_k^H\mathbf{J}_k^{(i)}\Big)\Big)$, where v represents any element of $\mathbf{V}_g$, $\forall g = 1,2,\ldots,G$, and $\mathbf{J}_k^{(i)} = \mathbf{J}_k\big|_{\{\mathbf{V}_g\}=\{\mathbf{V}_g^{(i)}\}}$.

Consequently, the nonconvex constraint (26c) can be tightly approximated by

$$(F_{g_k} - C_k)\eta_{g_k} + h_k^{(i)}\left(\{\mathbf{V}_g\}\right) \le 0, \quad \forall k = 1,2,\ldots,K. \quad \text{(C.6)}$$

Since $h_k^{(i)}\left(\{\mathbf{V}_g\}\right)$ in (C.1) is convex quadratic over $\{\mathbf{V}_g\}$, and $(F_{g_k} - C_k)\eta_{g_k}$ is linear over η_{g_k}, the constructed constraint in (C.6) is jointly convex over $\{\mathbf{V}_g\}$ and η_{g_k}. With the sequence of convex constraints constructed in (C.6), problem (25) can be iteratively solved in the SCA framework, with the i-th SCA subproblem shown in (27).

Appendix D
Proposed Algorithm for Solving (28)

Algorithm 5 Proposed Algorithm for Solving (28)

1: Initialize $\left\{\mathbf{V}_{g,t,\mathbf{f}}^{(0)}, \eta_{g,t,\mathbf{f}}^{(0)}\right\}$ and $\left\{C_{k,f_{g_k}}^{(0)}\right\}$:

$$C_{k,f_{g_k}}^{(0)} = C_{\text{tot}}/\left(K\left|\mathcal{F}_{g_k}\right|\right), \quad \mathbf{V}_{g,t,\mathbf{f}}^{(0)} = \sqrt{\frac{P_{\text{tot}}}{GMd}}\mathbf{1}_{M\times d},$$
$$\eta_{g,t,\mathbf{f}}^{(0)} = \min_{k\in\mathcal{K}_g}\Bigg\{\frac{1}{F_{f_g} - C_{k,f_g}^{(0)}} \cdot \log\det\left(\mathbf{I}_N + \mathbf{H}_{k,t,\mathbf{f}}\mathbf{V}_{g,t,\mathbf{f}}^{(0)}\left(\mathbf{V}_{g,t,\mathbf{f}}^{(0)}\right)^H\mathbf{H}_{k,t,\mathbf{f}}^H\mathbf{J}_{k,t,\mathbf{f}}^{(0)}\right)\Bigg\}.$$

2: **repeat** $(i = 0,1,\ldots)$

3: Compute $\left\{\mathbf{U}_{k,t,\mathbf{f}}^{(i)}, \mathbf{A}_{k,t,\mathbf{f}}^{(i)}, \mathbf{B}_{k,t,\mathbf{f}}^{(i)}, b_{k,t,\mathbf{f}}^{(i)}\right\}$:

$$\mathbf{U}_{k,t,\mathbf{f}}^{(i)} = \left(\sum_{g=1}^{G}\mathbf{H}_{k,t,\mathbf{f}}\mathbf{V}_{g,t,\mathbf{f}}^{(i)}\left(\mathbf{V}_{g,t,\mathbf{f}}^{(i)}\right)^H\mathbf{H}_{k,t,\mathbf{f}}^H + \sigma_k^2\mathbf{I}_N\right)^{-1}\mathbf{H}_{k,t,\mathbf{f}}\mathbf{V}_{g_k,t,\mathbf{f}}^{(i)},$$
$$\mathbf{A}_{k,t,\mathbf{f}}^{(i)} = \mathbf{H}_{k,t,\mathbf{f}}^H\mathbf{U}_{k,t,\mathbf{f}}^{(i)}\left(\mathbf{I}_d - \left(\mathbf{U}_{k,t,\mathbf{f}}^{(i)}\right)^H\mathbf{H}_{k,t,\mathbf{f}}\mathbf{V}_{g_k,t,\mathbf{f}}^{(i)}\right)^{-1}\cdot\left(\mathbf{U}_{k,t,\mathbf{f}}^{(i)}\right)^H\mathbf{H}_{k,t,\mathbf{f}},$$
$$\mathbf{B}_{k,t,\mathbf{f}}^{(i)} = -\left(\mathbf{I}_d - \left(\mathbf{U}_{k,t,\mathbf{f}}^{(i)}\right)^H\mathbf{H}_{k,t,\mathbf{f}}\mathbf{V}_{g_k,t,\mathbf{f}}^{(i)}\right)^{-1}\cdot\left(\mathbf{U}_{k,t,\mathbf{f}}^{(i)}\right)^H\mathbf{H}_{k,t,\mathbf{f}},$$
$$b_{k,t,\mathbf{f}}^{(i)} = \mathrm{Tr}\Bigg(\left(\mathbf{I}_d - \left(\mathbf{U}_{k,t,\mathbf{f}}^{(i)}\right)^H\mathbf{H}_{k,t,\mathbf{f}}\mathbf{V}_{g_k,t,\mathbf{f}}^{(i)}\right)^{-1}\cdot\left(\mathbf{I}_d + \sigma_k^2\left(\mathbf{U}_{k,t,\mathbf{f}}^{(i)}\right)^H\mathbf{U}_{k,t,\mathbf{f}}^{(i)}\right)\Bigg) + \log\det\left(\mathbf{I}_d - \left(\mathbf{U}_{k,t,\mathbf{f}}^{(i)}\right)^H\mathbf{H}_{k,t,\mathbf{f}}\mathbf{V}_{g_k,t,\mathbf{f}}^{(i)}\right) + \frac{\left(\eta_{g_k,t,\mathbf{f}}^{(i)} + C_{k,f_{g_k}}^{(i)}\right)^2}{2} - d.$$

4: Initialize $\theta^{(0)} = 1, \mu = 1, \delta_{t,\mathbf{f}} = 1, \lambda_{k,t,\mathbf{f}} = 1, \forall\mathbf{f}\in\mathcal{F}, \forall k = 1,2,\ldots,K, \forall t = 1,2,\ldots,T$.

5: **repeat** $(s = 1,2,\ldots)$

6: $\theta^{(s)} = \frac{1+\sqrt{1+4\left(\theta^{(s-1)}\right)^2}}{2}$.

7: Update $\left\{\mathbf{V}_{g,t,\mathbf{f}}^{\diamond}, \eta_{g,t,\mathbf{f}}^{\diamond}\right\}$ and $\left\{C_{k,f_{g_k}}^{\diamond}\right\}$:

$$\mathbf{V}_{g,t,\mathbf{f}}^{\diamond} = \left(\left(\rho_2 + \delta_{t,\mathbf{f}}\right)\mathbf{I}_M + \sum_{k=1}^{K}\lambda_{k,t,\mathbf{f}}\mathbf{A}_{k,t,\mathbf{f}}^{(i)}\right)^{-1}\cdot\left(\rho_2\mathbf{V}_{g,t,\mathbf{f}}^{(i)} - \sum_{k\in\mathcal{K}_g}\lambda_{k,t,\mathbf{f}}\left(\mathbf{B}_{k,t,\mathbf{f}}^{(i)}\right)^H\right),$$
$$\eta_{g,t,\mathbf{f}}^{\diamond} = \frac{\left(p_{f_g} - \sum_{k\in\mathcal{K}_g}\lambda_{k,t,\mathbf{f}}\right)F_{f_g} + \sum_{k\in\mathcal{K}_g}\lambda_{k,t,\mathbf{f}}C_{k,f_{g_k}}^{(i)}}{\rho_1 + \sum_{k\in\mathcal{K}_g}\lambda_{k,t,\mathbf{f}}} + \eta_{g,t,\mathbf{f}}^{(i)},$$
$$C_{k,f_{g_k}}^{\diamond} = \min\left\{\max\left\{\left(C_{k,f_{g_k}}^{(i)} + \frac{\sum_{t=1}^{T}\sum_{f_{g_k}}\lambda_{k,t,\mathbf{f}}\eta_{g_k,t,\mathbf{f}}^{(i)} - \mu}{\rho_3 + \sum_{t=1}^{T}\sum_{f_{g_k}}\lambda_{k,t,\mathbf{f}}}\right), 0\right\}, F_{f_{g_k}}\right\}.$$

8: Update $\left\{\tilde{\delta}_{t,\mathbf{f}}^{(s)}\right\}$, $\left\{\tilde{\lambda}_{k,t,\mathbf{f}}^{(s)}\right\}$, and $\tilde{\mu}^{(s)}$:

$$\tilde{\delta}_{t,\mathbf{f}}^{(s)} = \left(\delta_{t,\mathbf{f}} + \beta_s\left(\sum_{g=1}^{G}\mathrm{Tr}\left(\mathbf{V}_{g,t,\mathbf{f}}^{\diamond}\left(\mathbf{V}_{g,t,\mathbf{f}}^{\diamond}\right)^H\right) - P_{\text{tot}}\right)\right)^+,$$
$$\tilde{\lambda}_{k,t,\mathbf{f}}^{(s)} = \left(\lambda_{k,t,\mathbf{f}} + \beta_s f_{k,t,\mathbf{f}}^{(i)}\left(C_{k,f_{g_k}}^{\diamond}, \eta_{g_k,t,\mathbf{f}}^{\diamond}, \left\{\mathbf{V}_{g,t,\mathbf{f}}^{\diamond}\right\}_{g=1}^{G}\right)\right)^+,$$
$$\tilde{\mu}^{(s)} = \left(\mu + \beta_s\left(\sum_{k=1}^{K}\sum_{f_{g_k}\in\mathcal{F}_{g_k}}C_{k,f_{g_k}}^{\diamond} - C_{\text{tot}}\right)\right)^+.$$

9: Update $\left\{\delta_{t,\mathbf{f}}\right\}$, $\left\{\lambda_{k,t,\mathbf{f}}\right\}$, and μ:

$$\delta_{t,\mathbf{f}} = \tilde{\delta}_{t,\mathbf{f}}^{(s)} + \frac{\theta^{(s-1)} - 1}{\theta^{(s)}}\left(\tilde{\delta}_{t,\mathbf{f}}^{(s)} - \tilde{\delta}_{t,\mathbf{f}}^{(s-1)}\right),$$
$$\lambda_{k,t,\mathbf{f}} = \tilde{\lambda}_{k,t,\mathbf{f}}^{(s)} + \frac{\theta^{(s-1)} - 1}{\theta^{(s)}}\left(\tilde{\lambda}_{k,t,\mathbf{f}}^{(s)} - \tilde{\lambda}_{k,t,\mathbf{f}}^{(s-1)}\right),$$
$$\mu = \tilde{\mu}^{(s)} + \frac{\theta^{(s-1)} - 1}{\theta^{(s)}}\left(\tilde{\mu}^{(s)} - \tilde{\mu}^{(s-1)}\right).$$

10: **until** convergence

11: $\left\{\mathbf{V}_{g,t,\mathbf{f}}^{(i+1)}, \eta_{g,t,\mathbf{f}}^{(i+1)}\right\} = \left\{\mathbf{V}_{g,t,\mathbf{f}}^{\diamond}, \eta_{g,t,\mathbf{f}}^{\diamond}\right\}$ and $\left\{C_{k,f_{g_k}}^{(i+1)}\right\} = \left\{C_{k,f_{g_k}}^{\diamond}\right\}$.

12: **until** convergence

References

[1] P. Rost, C. Bernardos, A. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[2] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *J. Commun. Netw.*, vol. 18, no. 2, pp. 135–149, Apr. 2016.

[3] T. Q. S. Quek, M. Peng, O. Simeone, and W. Yu, *Cloud Radio Access Networks: Principles, Technologies, and Applications*. Cambridge University Press, 2017.

[4] 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.

[5] 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.

[6] 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.

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

[8] B. Dai, Y.-F. Liu, and W. Yu, "Optimized base-station cache allocation for cloud radio access network with multicast backhaul," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1737–1750, Aug. 2018.

[9] Y. Ugur, Z. H. Awan, and A. Sezgin, "Cloud radio access networks with coded caching," in *Proc. 20th Int. ITG Workshop on Smart Antennas*, 2016.

[10] 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.

[11] Y. Li, M. Xia, and Y.-C. Wu, "First-order algorithm for content-centric sparse multicast beamforming in large-scale C-RAN," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5959–5974, Sep. 2018.

[12] S.-H. Park, O. Simeone, and S. S. 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.

[13] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2760–2776, May 2013.

[14] 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.

[15] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Net.*, vol. 2015, no. 41, pp. 1–11, Feb. 2015.

[16] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, Jul. 2016.

[17] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug 2017.

[18] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[19] P. Patil, B. Dai, and W. Yu, "Performance comparison of data-sharing and compression strategies for cloud radio access networks," in *Proc. EUSIPCO*, 2015.

[20] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, "Downlink multicell processing with limited-backhaul capacity," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, pp. 1–10, Feb. 2009.

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

[22] "Coordinated multi-point operation for LTE physical layer aspects," Tech. Rep. 3GPP TR 36.819 V11.0.0, 3rd Generation Partnership Project (3GPP), Sep. 2011. [Online]. Available: http://www.qtc.jp/3GPP/Specs/36819-b10.pdf.

[23] H. Huang, M. Trivellato, A. Hottinen, M. Shafi, P. J. Smith, and R. Valenzuela, "Increasing downlink cellular throughput with limited network MIMO coordination," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 2983–2989, Jun. 2009.

[24] 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.

[25] S. S. Bidokhti, M. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 6996–7016, Nov. 2018.

[26] 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.

[27] D. Liu and C. Yang, "Caching at base stations with heterogeneous user demands and spatial locality," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1554–1569, Feb. 2019.

[28] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*. 2nd ed. Springer, 2011.

[29] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to non-convex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, Jul. 2010.

[30] T. Lipp and S. Boyd, "Variations and extension of the convex-concave procedure," *Optim. Eng.*, vol. 17, no. 2, pp. 263–287, Jun. 2016.

[31] A. Beck and M. Teboulle, "Gradient-based algorithms with application to signal recovery problems," in *Convex Optimization in Signal Processing and Communications*. Y. C. Eldar and D. Palomar, Eds., Cambridge Univ. Press, 2010.

[32] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univercity Press, 2004.

[33] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[34] M. Razaviyayn, "Successive convex approximation: Analysis and applications," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Minnesota, Minneapolis, MN, USA, 2014.

[35] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," Sep. 2013. [Online]. Available: http://cvxr.com/cvx.

Yang Li (S'14) received the B.E degree and the M.E degree in electronics engineering from Beihang University (BUAA), Beijing, China, in 2012 and 2015, respectively. He received the Ph.D. degree in electrical and electronic engineering at the University of Hong Kong (HKU), Hong Kong, China, in 2019. His research interests are in general areas of wireless communications and signal processing.

Minghua Xia (M'12) received his 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 (ETRI) of South Korea, Beijing R&D Center, Beijing, China, where he worked as a member and then as a senior member of engineering staff. From 2010 to 2014, he was in sequence with The University of Hong Kong, Hong Kong, China; King Abdullah University of Science and Technology, Jeddah, Saudi Arabia; and the Institut National de la Recherche Scientifique (INRS), University of Quebec, Montreal, Canada, as a Postdoctoral Fellow. His research interests are in the general areas of wireless communications and signal processing.

Dr. Xia received the Professional Award at the IEEE TENCON, held in Macau, in 2015. He was recognized as Exemplary Reviewer by IEEE TRANSACTIONS ON COMMUNICATIONS in 2014, IEEE COMMUNICATIONS LETTERS in 2014, and IEEE WIRELESS COMMUNICATIONS LETTERS in 2014 and 2015. Dr. Xia serverd as TPC Symposium Chair of IEEE ICC'2019 and now serves as Associate Editor for the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING and the IET SMART CITIES.

Yik-Chung Wu received the B.Eng. (EEE) degree in 1998 and the M.Phil. degree in 2001 from the University of Hong Kong (HKU). He received the Croucher Foundation scholarship in 2002 to study Ph.D. degree at Texas A&M University, College Station, and graduated in 2005. From August 2005 to August 2006, he was with the Thomson Corporate Research, Princeton, NJ, as a Member of Technical Staff. Since September 2006, he has been with HKU, currently as an Associate Professor. He was a visiting scholar at Princeton University, in summers of 2015 and 2017. His research interests are in general areas of signal processing, machine learning and communication systems. Dr. Wu served as an Editor for IEEE COMMUNICATIONS LETTERS and IEEE TRANSACTIONS ON COMMUNICATIONS, and is currently an editor for JOURNAL OF COMMUNICATIONS AND NETWORKS.