Optimal Joint Remote Radio Head Selection and Beamforming Design for Limited Fronthaul C-RAN

Phuong Luong, *Student Member, IEEE*, François Gagnon, *Senior Member, IEEE*, Charles Despins, *Senior Member, IEEE*, and Le-Nam Tran, *Senior Member, IEEE*

***Abstract*—This paper considers the downlink transmission of cloud-radio access networks (C-RANs) with limited fronthaul capacity. We formulate a joint design of remote radio head (RRH) selection, RRH-user association, and transmit beamforming for simultaneously optimizing the achievable sum rate and total power consumption, using the multi-objective optimization concept. Due to the non-convexity of per-fronthaul capacity constraints and introduced binary selection variables, the formulated problem lends itself to a mixed-integer non-convex program, which is generally NP-hard. Motivated by powerful computing capability of C-RAN and for benchmarking purposes, we propose a branch and reduce and bound based algorithm to attain a globally optimal solution. For more practically appealing approaches, we then propose three iterative low-complexity algorithms. In the first method, we iteratively approximate the continuous non-convex constraints by convex conic ones using successive convex approximation (SCA) framework. More explicitly, the problem obtained at each iteration is a mixed-integer second order cone program (MI-SOCP) for which dedicated solvers are available. In the second method, we first relax the binary variables to be continuous to arrive at a sequence of SOCPs and then perform a post-processing procedure on the relaxed variables to search for a high-performance solution. In the third method, we solve the considered problem in view of sparsity-inducing regularization. Numerical results show that our proposed algorithms converge rapidly and achieve near-optimal performance as well as outperform the known algorithms.**

***Index Terms*—Base station selection, beamforming, cloud radio access networks, limited fronthaul, mixed integer second order cone programming, optimization.**

I. Introduction

Recently, cloud radio access networks (C-RANs) have been considered as a key technology to significantly enhance network performance in order to cope with the explosive demand expected in the foreseen 5G networks [2]. By merging the capability of cloud computing and radio frequency (RF) transmissions, C-RAN architectures are anticipated to use low-cost low-power base stations for radio services while embracing coordinated and centralized computational tasks at the cloud center to achieve higher network performance. Generally, C-RAN systems contain several low-power RRHs that are all connected to a baseband unit (BBU) pool through fronthaul links [3], as illustrated in Fig. 1. In C-RANs, RRHs equipped with RF modules only account for compression and transmission/reception of radio signals to/from user equipment (UEs). The fronthaul links connecting RRHs and the BBU pool play a role as a data signal transportation media towards/backwards the BBU pool from/to RRHs. On the BBU side, the joint centralized processing task powered by multiple advanced computer processing units (CPUs) is executed to handle all the relevant baseband signals. With this architectural advantage, C-RANs are able to cater both effective interference management and cooperative gains, thereby increasing system capacity. However, the performance of a C-RAN is heavily restricted by the limited fronthaul capacity between RRHs and the BBU pool [4], [5]. This creates a fundamental bottleneck on the network operation, which requires appropriate management on the selection and transmit power design at RRHs to attain the optimal performance.

Achieving the maximal achievable sum rate with minimal amount of available resources is a vital problem in wireless networks in general and in C-RANs in particular. The number of RRHs, together with the associated fronthaul links, in C-RANs can be very high, which results in huge power consumption. In this regard, RRH selection and RRH-user association problem is of particular interest. This should be done in accordance with limited fronthaul capacity constraints, which present a new challenge in the design of C-RANs. Consequently, the existing design techniques for conventional wireless communication networks are no longer applicable and thus new design methods for C-RANs are required.

There have been several pioneer works that study the joint design of RRH-user (UE) association and beamforming in C-RAN with limited fronthaul capacity. For example, the works in [6], [7] proposed various compression techniques to minimize the transmitted data delivered over the fronthaul network. In [8], Fan *et. al.* developed a low-complexity and efficient algorithm to form clusters of RRHs so that the number of centralized computational processing tasks at the BBU pool was greatly reduced. In [9], [10], the authors employed a sparsity inducing-norm to develop a joint beamforming and base station (BS) selection design to minimize the power consumption in C-RANs so that the related fronthaul capacity required to transport data was implicitly minimized. Inspired by these works, the authors in [11] further addressed the coupling factor of uplink (UL) and downlink (DL) transmissions in C-RANs to resolve the problem of [9] by exploiting the UL-DL duality and MI-SOCP framework. In [12], the authors employed a generalized Bender decomposition (GBD) method to develop a decentralized algorithm that jointly optimizes the

Phuong Luong, Charles Despins, and François Gagnon are with École de Technologie Supérieure (ÉTS), Montréal, QC, Canada. Le-Nam Tran is with the School of Electrical and Electronic Enginering, University College Dublin, Ireland. He was with the Department of Electronic Engineering, Maynooth University, Ireland (email: {thi-thu-phuong.luong.1@ens.etsmtl.ca}, {francois.gagnon,charles.despins}@etsmtl.ca, {nam.tran@ucd.ie}).

Preliminary results of this work were presented in IEEE VTC-Fall 2016, Montréal, Canada, Sep. 2016 [1].

beamforming and BS clustering under the limited message exchange assumption in cognitive radio networks. The work of [13] considered the limited backhaul constraint and formulated a power minimization problem as a combinatorial non-convex problem, where different resource allocation algorithms based on GBD combined with semidefinite programming and difference of convex programming were derived. In [14], an increment-based greedy allocation algorithm was proposed to solve the problem of resource allocation and user association through a user-centric resource sharing scheme for a C-RAN with fronthaul capacity constraint. In addition, the authors in [15], [16] explicitly incorporated the per-fronthaul capacity constraints in their optimization problems and applied different methods based on group sparsity inducing norms to attain their designs. In [40], the authors assigned the fixed rates in the previous iteration to overcome the non-convexity of fronthaul capacity constraints and applied a generalized WMMSE method to solve the problem of energy efficiency maximization in queue-aware H-CRAN. The works of [33], [36] developed the coalitional formation game based algorithm to form an RRH cluster, while a contract game based interference coordination was proposed in [39]. Using the approximation Bellman equation, the authors in [38] derived a close-form approximation function for the problem of power-delay trade-off for MU-MIMO systems. To develop an optimal algorithm for resource allocation in C-RANs, branch and bound method was used in [21], [24] and the dual decomposition method was exploited in [34], [37]. An exhaustive search was adopted in [16], [35] to find the optimal RRH cluster. However, the authors in [11], [12], [21], [22], [24], [34], [37], [39] did not explicitly consider the fronthaul capacity constraints, while the user rates were set to be constant to overcome the non-convex fronthaul capacity constraints in [13]–[16], [35], [40].

From a network optimization perspective, total power consumption minimization and achievable sum rate maximization are the two most common performance metrics when designing wireless communications. However, these two design criteria have been often considered separately as their goals are conflicting. Note that, by weighing the two objectives, we can find the whole rate and power region of the system [17], [18]. This is in close relation to energy-efficient transmission strategies [19]. The research on transmit power-throughput trade-off was considered in [20], where the convex hull of the entire achievable power-rate region of MIMO heterogeneous networks was obtained. By an MI-SOCP approach, a mechanism to find the optimal trade-off between the overall BS power consumption and power consumption overhead associated with CoMP transmission was proposed in [21]. The work in [1] was the first to employ the MI-SOCP approach to study the power minimization in limited C-RANs. The works of [22] adopted the reweighted ℓ_1-norm to study the trade-off between total power consumption and fronthaul capacity for data sharing and compression strategies in C-RANs.

It is worth mentioning that the studies in relation to the overall power consumption minimization in C-RANs implicitly imply the minimization of the fronthaul capacity usage. This also helps the cloud center to use the least computational resource to satisfy QoS requirements. Investigations on how C-RANs can benefit from cloud computing capabilities have been reported recently. For example, the works of [24], [23] proposed a joint design of virtual machine computation capacity, RRH selection and beamforming to minimize the total power consumption in C-RANs. In this paper, we focus on the communications part of C-RANs rather than the cloud computing capabilities.

In this paper, we consider the downlink transmission of a C-RAN with limited fronthaul capacity. In the considered system model, digital data is transmitted from the BBU pool to RRHs using fronthaul links of finite capacity, and beamforming technique is used to send data to UEs. Under this context, we study a joint design of RRH selection, RRH-UE association and beamformer that simultaneously maximizes the achievable sum rate and minimizes the total power consumption. The main motivation for jointly designing beamforming with RRH selection, and RRH-UE association is due to the design goal. It is true that for spectral efficiency maximization, we do not need to consider the RRH-UE association and RRH selection since maximum degree of freedom is achieved if all RRHs are allowed to serve all the UEs in the system. We note that the objective function in our problem strikes the balance between spectral efficiency maximization and total power minimization. Thus, RRH-UE association and RRH selection are particularly relevant. Intuitively, the optimization of RRH-UE association and RRH selection is important because there exists a situation where some RRHs of severe fading conditions can be switched off and each UE can be served by a small subset of active RRHs to save power.

To deal with two conflicting targets, we formulate the problem of interest as a multi-objective (or vector) optimization problem, directly solving which is cumbersome. To overcome this difficulty, we propose to employ the scalarization approach for the formulated problem by linearly combining each weighted element of the vector objective function to result in a standard scalar optimization problem. As shown later on, two challenges arise in the considered problem: (i) the non-convexity of per-fronthaul capacity constraints, and (ii) the combinatorial nature of the selection procedure. To deal with the latter one, we naturally introduce binary selection variables to represent the selection status of RRHs and associated users. The formulated problem is basically a combinatorial one, which is generally NP-hard. Moreover, another problem is that even if these binary selection variables are relaxed to be continuous, the resulted problem is still non-convex because of the non-convexity of the objective function and per-fronthaul capacity constraints. This attribute makes the considered problem much more difficult to solve, and the methods presented in previous studies such as those in [11], [12], [21], [22], [24], [34], [37], [39] are no longer applicable. Moreover, different from [13]–[16], [35], [40] where the authors simply assign a predetermined achievable rate to overcome the non-convex fronthaul constraint, we directly tackle it by proposing novel transformations to arrive at an equivalent but more tractable form. Based on that, we develop a new framework using SCA method [42] to solve the considered problem efficiently. The main contributions of this paper are summarized as follows.

- We formulate the joint design of RRH selection, RRH-

UE association and beamforming for achievable sum rate-total power maximization problem by employing the concept of multi-objective optimization [17]. The problem is formulated as a mixed integer nonlinear program. We then present a novel transformation to rewrite the design problem in a form that facilitates a customized branch-and-reduce-and-bound (BRB) algorithm to find a globally optimal solution based on monotonic optimization.

- To overcome the high complexity inherently in a global optimization method, we propose novel transformations and convex approximation techniques to derive two sub-optimal low-complexity algorithms aiming at attaining a high-quality feasible solution. More specifically, in the first method, we iteratively approximate the continuous non-convex constraints by convex ones using SCA framework. By using a quadratic bound of the logarithm function, we are able to arrive at a sequence of MI-SOCPs, for which dedicated solvers are available and efficient. The second method is a simplified version of the first one where we further relax the binary variables in each iteration to be continuous. That is to say, each iteration of the second method merely requires solving an SOCP. After convergence, we then perform a post-processing procedure on the relaxed selection variables to search for a high-performance solution.
- From a different viewpoint, we reformulate the considered problem under the concept of sparsity-inducing regularization. The connection status of a particular pair of RRH and UE is represented by the norm of the associated beamforming vector, which is encouraged to be zero if doing so improves the objective. By exploiting a ℓ_2-norm based logarithm approximation, the new optimization problem basically shares the same non-convex structure as the previous one. Applying similar steps in the proposed methods mentioned above, we arrive at an SOCP, but of smaller size, in each iteration. Then, RRH selection and RRH-UE association can be decided by ignoring the zero elements in the obtained sparse solution, after the convergence of the iterative algorithm.
- Extensive numerical results are presented to show the efficiency of our proposed algorithms, compared to known solutions in the literature, especially for the cases of sum achievable rate maximization and power minimization. In particular, compared to the WMMSE approach in [15], our proposed SCA-based methods converge much faster, while still achieving a better performance.

The rest of the paper is organized as follows. Section II introduces the system model and formulates our joint RRH selection, RRH-UE association and transmit beamformers into an achievable sum rate-total power consumption optimization problem. Section III provides the proposed globally optimal design. In Section IV, we present the proposed low-complexity algorithms, followed by numerical results and insight discussions under various simulation setups in Section V. Finally, the conclusion of the paper is given in Section VI.

Notation: We use bold uppercase and lowercase letters to denote matrices and vectors, respectively. $\mathbb{C}$ and $\mathbb{R}$ represent the space of complex and real numbers. $\mathbf{x}^T$ and $\mathbf{x}^H$ stand for the transpose and Hermitian operation of vector $\mathbf{x}$. $|x|$ represents the modulus of $x \in \mathbb{C}$, while $\|\mathbf{x}\|_2$ is the ℓ_2-norm of the vector $\mathbf{x}$. For a scalar x, we denote by $\chi(x)$ the indicator function if x is zero or not. That is, $\chi(x) = 1$ if $x \neq 0$ and $\chi(x) = 0$ if $x = 0$. The notation $\mathbb{E}\{\cdot\}$ denotes the expectation operator; x^* represents the complex conjugate of $x \in \mathbb{C}$; $\mathrm{Re}(.)$ and $\mathrm{Im}(.)$ stand for the real and imaginary part of the argument, respectively; $\mathcal{O}$ represents the big O notation.

II. System Model and Problem Formulation

A. Transmission Model

We consider the DL of C-RAN consisting of I RRHs and K single antenna UEs. For notational convenience, we denote $\mathcal{I} = \{1, \ldots, I\}$ and $\mathcal{K} = \{1, \ldots, K\}$ as the set of RRHs and UEs, respectively. We assume that the i^{th} RRH is equipped with M_i antennas, $\forall i \in \mathcal{I}$. As shown in Fig. 1, we assume that all the RRHs are connected to BBU pool via the fronthaul links, i.e., high-speed optical ones, where the i^{th} link has a predetermined maximum capacity C_i. Each UE is served by a specific group of RRHs but one RRH can serve more than one users simultaneously. Let us denote by s_k the signal with unit power, i.e., $\mathbb{E}\{s_k s_k^*\} = 1$, intended for the k^{th} UE and by $\mathbf{w}_{i,k} \in \mathbb{C}^{M_i \times 1}$ the transmit beamforming vector from the i^{th} RRH to the k^{th} UE. The vector of channel coefficients encompassing small-scale fading and pathloss from the i^{th} RRH to the k^{th} UE is represented by $\mathbf{h}_{i,k} \in \mathbb{C}^{M_i \times 1}$. In this work, we assume perfect channel state information (CSI) between the RRHs and the UEs.[1] For notational convenience we denote the set of beamforming vectors intended for the k^{th} UE as $\mathbf{w}_k \triangleq [\mathbf{w}_{1,k}^T, \mathbf{w}_{2,k}^T, \ldots, \mathbf{w}_{I,k}^T]^T \in \mathbb{C}^{M \times 1}$, and the vector including the channels from all RRHs to the k^{th} UE as $\mathbf{h}_k \triangleq [\mathbf{h}_{1,k}^T, \mathbf{h}_{2,k}^T, \ldots, \mathbf{h}_{I,k}^T]^T \in \mathbb{C}^{M \times 1}$, where $M = \sum_{i \in \mathcal{I}} M_i$. Using these notations, the received signal at the k^{th} UE is given by

$$y_k = \mathbf{h}_k^H \mathbf{w}_k s_k + \sum_{j \in \mathcal{K} \setminus k} \mathbf{h}_k^H \mathbf{w}_j s_j + z_k \tag{1}$$

where $z_k \sim \mathcal{CN}(0, \sigma_0^2)$ is the additive white Gaussian noise (AWGN) and σ_0^2 is the noise power. Note that in (1), we have assumed that the k^{th} UE is connected to all the RRHs, but the i^{th} RRH serves the k^{th} UE only if $\left\|\mathbf{w}_{i,k}\right\|_2 > 0$. By treating interference as noise, the achievable rate in b/s/Hz at the k^{th} UE is given by

$$R_k(\mathbf{w}) = \log_2(1 + \Gamma_k(\mathbf{w})) \tag{2}$$

where

$$\Gamma_k(\mathbf{w}) = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{j \in \mathcal{K} \setminus k} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_0^2} \tag{3}$$

and $\mathbf{w} \triangleq [\mathbf{w}_1^T, \mathbf{w}_2^T, \ldots, \mathbf{w}_k^T]^T \in \mathbb{C}^{(KM) \times 1}$ is vector stacking the beamformers for all users.

[1] In practice, CSI between the RRHs and the UEs can be estimated by exploiting the channel reciprocity property between the UL and DL transmissions in the time division duplexing (TDD) system or by feedback channels (from users) in the frequency division duplexing (FDD) system. Then, the CSI can be transferred to the BBU pool from all RRHs via the corresponding fronthaul links to design the resource allocation.

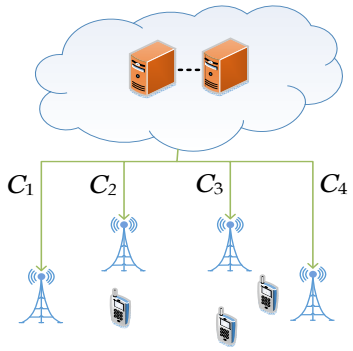


Fig. 1: Limited fronthaul C-RAN.

B. Fronthaul Capacity Constraint

After the BBU pool performs a relevant radio resource management algorithm to determine the beamforming vectors, data for the k^{th} UE is routed from the BBU pool to the i^{th} RRH via the i^{th} fronthaul link only if $\left\|\mathbf{w}_{i,k}\right\|_2 > 0$. For the transmission to be feasible, the capacity of the i^{th} fronthaul link should be ξ_i times greater than or equal to the total achievable rate at the i^{th} RRH where $\xi_i \geq 1$, $\forall i \in \mathcal{I}$ [25]. Herein, we assume that the channel conditions are slow varying. Thus, the transportation of CSI via the fronthaul link occurs less frequently than that of data. As a result, conveying CSI consumes much less fronthaul capacity than conveying the users' data, and thus can be neglected for the sake of simplicity. For the purpose of problem formulation, let us introduce binary variables $a_{i,k} \in \{0,1\}, \forall i \in \mathcal{I}$ and $k \in \mathcal{K}$ to represent the association status between the i^{th} RRH and the k^{th} UE, i.e., $a_{i,k} = 1$ implies that the k^{th} UE is served by the i^{th} RRH and $a_{i,k} = 0$, otherwise. Then, the per-fronthaul capacity constraints can be written as

$$\sum_{k\in\mathcal{K}} a_{i,k} R_k(\mathbf{w}) \leq \frac{C_i}{\xi_i}, \forall i \in \mathcal{I}. \tag{4}$$

C. Power consumption

In this subsection, we present a power consumption model that accounts for the power consumption at RRHs as well as for transmitting digital data from the BBU pool to the corresponding RRHs. According to [9], the power consumption at a RRH consists of two types, namely, data dependent power and data independent power. The former is the power dispatched at the power amplifiers in an RRH which is a function of transmitted signals, while the latter is mostly due to electronic components. The data independent power can be sub-categorized into two types, one, denoted by P_i^{ra}, representing the fixed amount of power when the i^{th} RRH is in active mode, and one, denoted by P_i^{ri}, accounting for the power required to keep the i^{th} RRH in sleep mode. More specifically, P_i^{ra} and P_i^{ri} are the power that is consumed by the circuit and to maintain the operation of the fronthaul optical link in the active and sleep mode of the i^{th} RRH, respectively. The power consumption for forwarding information data and beamformers related to the transmission from the i^{th} RRH to the k^{th} UE via fronthaul transmission is denoted by $P_{i,k}^{\text{FH}}$. From the introduction of $a_{i,k}$, it is obvious that when $a_{i,k} = 0$, then $P_{i,k}^{\text{FH}} = 0$. To represent the operation mode of the i^{th} RRH, we introduce a binary variable $b_i = \{0,1\}, \forall i \in \mathcal{I}$. In particular, $b_i = 0$ states that the i^{th} RRH is in sleep mode and $b_i = 1$ means otherwise. In summary, the sum power consumption at all RRHs and corresponding fronthaul links can be written as

$$P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) = \frac{1}{\eta_i}\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}}\left\|\mathbf{w}_{i,k}\right\|_2^2 + \underbrace{\sum_{i\in\mathcal{I}} b_i P_i^{\text{ra}} + \sum_{i\in\mathcal{I}}(1-b_i)P_i^{\text{ri}} + \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} a_{i,k}P_{i,k}^{\text{FH}}}_{P^{\text{circ}}(\mathbf{a},\mathbf{b})} \tag{5}$$

where $\eta_i \in [0,1]$ is the power amplifier efficiency, $\mathbf{b} = [b_1,\ldots,b_I]^T$ and $\mathbf{a} = \left[\mathbf{a}_1^T,\ldots,\mathbf{a}_K^T\right]^T$ where $\mathbf{a}_k = [a_{1,k}\ldots,a_{I,k}]^T$. For simplicity, we denote $P^{\text{circ}}(\mathbf{a},\mathbf{b}) = \sum_{i\in\mathcal{I}} b_i P_i^{\text{ra}} + \sum_{i\in\mathcal{I}}(1-b_i)P_i^{\text{ri}} + \sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} a_{i,k}P_{i,k}^{\text{FH}}$.

D. Problem Formulation

We are ready to formulate the problem of simultaneously maximizing the achievable sum rate and minimizing the total power consumption of the considered C-RAN model. By optimizing the two performance measures in a single framework, we can achieve maximal sum rate with minimal total power consumption, and also easily trade-off between the two conflicting objectives. In general, this problem is categorized as a multi-objective optimization one, where the objective is a vector-valued function. A common method to solve it is to apply the scalarization method by taking a linear combination of individual components [17], [26]. Motivated by this method, we study a joint design of beamforming, RRH selection and RRH-UE association given by

$$\max_{\mathbf{b},\mathbf{a},\mathbf{w},\boldsymbol{\nu}} \ \alpha\frac{R^{\text{tot}}(\mathbf{w})}{R_0} - (1-\alpha)\frac{P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b})}{P_0} \tag{6a}$$

$$\text{s.t.} \quad \Gamma_k(\mathbf{w}) \geq \Gamma_k^{\min}, \forall k \in \mathcal{K} \tag{6b}$$

$$\sum_{k\in\mathcal{K}}\left\|\mathbf{w}_{i,k}\right\|_2^2 \leq b_i P^{\max}, \forall i \in \mathcal{I} \tag{6c}$$

$$\left\|\mathbf{w}_{i,k}\right\|_2^2 \leq a_{i,k}\nu_{i,k}, \forall i \in \mathcal{I}, \forall k \in \mathcal{K} \tag{6d}$$

$$\nu_{i,k} \leq a_{i,k}P^{\max}, \forall i \in \mathcal{I}, \forall k \in \mathcal{K} \tag{6e}$$

$$a_{i,k} \leq b_i, \forall i \in \mathcal{I}, \forall k \in \mathcal{K} \tag{6f}$$

$$\sum_{i\in\mathcal{I}} a_{i,k} \geq 1, \forall k \in \mathcal{K} \tag{6g}$$

$$\sum_{k\in\mathcal{K}} a_{i,k}R_k(\mathbf{w}) \leq \frac{C_i}{\xi_i}, \forall i \in \mathcal{I} \tag{6h}$$

$$b_i \in \{0,1\}, a_{i,k} \in \{0,1\}, \forall i \in \mathcal{I}, \forall k \in \mathcal{K} \tag{6i}$$

where $R^{\text{tot}}(\mathbf{w}) \triangleq \sum_{k\in\mathcal{K}} R_k(\mathbf{w})$. In (6), we have introduced the weight $\alpha \in [0,1]$ to strike the balance between sum rate maximization and total power minimization. It is worth mentioning that if $\alpha = 1$ (or $\alpha = 0$), we arrive at the sum rate maximization problem (or total power minimization). In addition, due to the different physical meaning of rate and power in the objective, we divide $R^{\text{tot}}(\mathbf{w})$ and $P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b})$

by a reference value R_0 b/s/Hz and P_0 W, respectively. The values of R_0 and P_0 are provided in Section V. Before proceeding further, we note that there also exist other scalarization techniques such as weighted Tchebycheff [27], weighted exponential and other methods introduced in [26] to solve a multi-objective problem. However, the weighted Tchebycheff method is inefficient to the considered problem in this paper because optimizing individual objectives is already intractable. The weighted exponential and other methods in [26] essentially lead to a formulation similar to (6), and thus the proposed solutions in the subsequent sections are still applicable. In this paper, we adopt the linear scalarization method for its popularity and simplicity.

The introduction of the set of auxiliary variables $\boldsymbol{\nu} = \{\nu_{i,k}\}, \forall i \in \mathcal{I}, k \in \mathcal{K}$ and the constraints in (6) deserve further explanation. Intuitively, $\nu_{i,k}$ represents the *soft* power transmitted from the i^{th} RRH to the k^{th} UE . Constraint (6b) is to ensure the QoS requirement for the k^{th} user, where $\Gamma_k^{\min}$ is the predetermined SINR requirement for the k^{th} user. Moreover, constraint (6c) implies that the total transmit power at each RRH is limited by a given budget power $P^{\max}$. The constraints (6c) and (6f) are to make sure that when the i^{th} RRH is in sleep mode, i.e., $b_i = 0$, no power will be transmitted from it. This can be easily seen as $b_i = 0$, then $a_{i,k} = 0$ for all $k \in \mathcal{K}$ and $\sum_{k\in\mathcal{K}} \left\|\mathbf{w}_{i,k}\right\|_2^2 = 0$. Similarly, in (6d) we also guarantee that the transmit power $\left\|\mathbf{w}_{i,k}\right\|_2^2$ from the i^{th} RRH to the k^{th} user is zero if $a_{i,k} = 0$. The constraint in (6e) means that the soft power from the i^{th} RRH to the k^{th} user should not exceed $P^{\max}$. We also impose the constraint (6g) to ensure that each user is served by at least one RRH. Finally, the per-fronthaul capacity constraint is explicitly presented in (6h).

We remark that problem (6) includes, as a special case, RRH clustering [12], [15], [21]. Specifically, a dynamic cluster of RRHs can be formed by posing an extra constraint on the variable $\{a_{i,k}\}_{\forall i,k}$, i.e., $\sum_{i\in\mathcal{I}} a_{i,k} \leq \kappa$ where $\kappa \leq I$ to require that each user can only connect to at most κ RRHs instead of all RRHs. In this way, dynamic RRH cluster formation can be optimized through binary variables $a_{i,k} \in \{0,1\}$, $\forall i \in \mathcal{I}$ and $\forall k \in \mathcal{K}$ in each scheduling slot. Exploring the potential gains offered by dynamic RRH clustering deserves a thorough study, and thus is left as future work.

Towards solving (6) optimally, we note that the constraint (6d) is called a rotated second order cone [17], [21]. It is trivial to see that (6d) can be rewritten as $\left(\frac{a_{i,k}+\nu_{i,k}}{2}\right)^2 - \left(\frac{a_{i,k}-\nu_{i,k}}{2}\right)^2 \geq \left\|\mathbf{w}_{i,k}\right\|_2^2$. Thus (6d) is equivalent to the following SOC constraint

$$\frac{a_{i,k}+\nu_{i,k}}{2} \geq \left\| \left[\frac{a_{i,k}-\nu_{i,k}}{2}, \mathbf{w}_{i,k}^T\right]^T \right\|_2, \forall i \in \mathcal{I}, \forall k \in \mathcal{K} \quad (7)$$

III. Global Optimization Method for (6)

In this section we present a solution to solve (6) optimally. First, we provide some comments on the complexity of (6). Problem (6) is a mixed integer non-linear program (MINLP) due to binary variables $\mathbf{a}$ and $\mathbf{b}$, which is generally NP-hard. Moreover, even if $\mathbf{a}$ and $\mathbf{b}$ are relaxed to be continuous, the obtained problem is still non-convex because of the non-convexity of (6a) and (6h). In mathematical programming, (6) is categorized as a mixed integer (MI) non-convex problem for which such a method in [21] is not applicable to find a globally optimal solution. To the best of our knowledge, there is no off-the-shelf solver for (6). In what follows, we present an equivalent formulation of (6), based on which a BRB algorithm using monotonic optimization is customized to solve it optimally.

A. Equivalent Formulation

Consider the following problem

$$\max_{\mathbf{b},\mathbf{a},\mathbf{w},\boldsymbol{\nu},\mathbf{u}} \; f(\mathbf{u}) \triangleq \tilde{\alpha} \sum_{k\in\mathcal{K}} u_k - \bar{\alpha} u_0^{-1} \quad (8a)$$

$$\text{s.t. } R_k(\mathbf{w}) \geq u_k, \forall k \in \mathcal{K} \quad (8b)$$

$$u_k \geq \log(1+\Gamma_k^{\min}) \quad (8c)$$

$$P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \leq u_0^{-1} \quad (8d)$$

$$\sum_{k\in\mathcal{K}} a_{i,k} u_k \leq \frac{C_i}{\xi_i} \quad (8e)$$

$$u_k \geq 0, k = 0,1,2,\ldots,K \quad (8f)$$

$$(7), (6c), (6e) - (6g), (6i). \quad (8g)$$

where $\tilde{\alpha} \triangleq \alpha/R_0$ and $\bar{\alpha} \triangleq (1-\alpha)/P_0$. The key to the development of our proposed optimal solution is due to the following lemma.

Lemma 1. *The formulations in* (6) *and* (8) *are equivalent, i.e., they have the same optimal objective and solution set.*

Proof: The equivalence is due to the observation that at optimality of (8), the inequalities $R_k(\mathbf{w}) \geq u_k$ and $P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \leq u_0^{-1}$ must hold with equality. The details of the proof are presented in Appendix A. □

B. Proposed BRB Solution

While the formulation in (8) does not reduce the non-convexity of the considered problem, it facilitates the development of an optimal design based on monotonic optimization. More specifically, it is easy to see that the objective in (8) monotonically increases with respect to each entry of $\mathbf{u}$. Thus we can apply a BRB method to solve (8) optimally as done in [28], [41]. We refer the interested reader to [28], [41] for a detailed description of a monotonic optimization-based BRB. Herein we present the customized steps required for solving the considered problem. For this purpose, we reuse the definitions and concepts in [28], [41] relevant to the development of the proposed BRB. Specifically, we define the compact normal set $\mathcal{Q} = \{\mathbf{u} \in \mathbb{R}_+^{K+1} | (8b) - (8g)\}$ and $\mathcal{U} = [\underline{\mathbf{u}}, \bar{\mathbf{u}}]$ as the box that contains all $\mathbf{u}$ feasible to (8). The values of $\underline{\mathbf{u}}$ and $\bar{\mathbf{u}}$ are can be computed as follows. From (8c), it holds that $u_k \geq \log(1+\Gamma_k^{\min}) = \underline{u}_k$, $\forall k = \{1,\ldots,K\}$. Moreover, we have

$$u_0 \geq \frac{1}{\frac{1}{\eta_i} I \times P^{\max} + I \times P^{\text{ra}} + I \times K \times P_{i,k}^{\text{FH}}} = \underline{u}_0. \quad (9)$$

Similarly, an upper bound of $\mathbf{u}$ can be given by

$$u_0 \leq \frac{1}{I \times P^{\text{ri}}} = \overline{u}_0 \tag{10}$$

$$u_k \overset{(a)}{\leq} \log(1 + \frac{\left|\mathbf{h}_k^H \mathbf{w}_k\right|^2}{\sigma_0^2}) \overset{(b)}{\leq} \log\left(1 + \frac{\|\mathbf{h}_k\|_2^2 \|\mathbf{w}_k\|_2^2}{\sigma_0^2}\right) \overset{(c)}{\leq} \log\left(1 + \frac{I \times P_{\max} \|\mathbf{h}_k\|_2^2}{\sigma_0^2}\right) = \overline{u}_k, \forall k \in \mathcal{K}. \tag{11}$$

where (a) is due to omitting the inter-user interference, (b) is the result of applying Cauchy-Schwarz inequality, and (c) is obvious from the power constraint for each $\mathbf{w}_{i,k}$. The main problem in a BRB algorithm using monotonic optimization framework is to check if a given $\mathbf{u}$ belongs to $\mathcal{Q}$ or not. Mathematically we need to solve the following feasibility problem for a given $\mathbf{u}$

$$\text{find} \quad \mathbf{w}, \mathbf{a}, \mathbf{b}, \boldsymbol{\nu} \tag{12a}$$

$$\text{s.t.} \quad (8b), (8c), (8d), (8e), (8f), (8g). \tag{12b}$$

Similar to (26) we can equivalently rewrite (8b) as

$$c' \operatorname{Re}(\mathbf{h}_k^H \mathbf{w}_k) \geq \left\|\mathbf{h}_k^H \mathbf{w}_1, \ldots, \mathbf{h}_k^H \mathbf{w}_K, \sigma_0\right\|_2 \tag{13}$$

where $c' = \sqrt{\frac{1}{2^{u_k}-1} + 1}$. Furthermore (8d) is equivalent to

$$\frac{u_0^{-1} - P^{\text{circ}}(\mathbf{a}, \mathbf{b}) + 1}{2} \geq \left\| \left[\frac{\mathbf{w}_{1,1}^T}{\sqrt{\eta_1}}, \ldots, \frac{\mathbf{w}_{I,K}^T}{\sqrt{\eta_I}}, \frac{u_0^{-1} - P^{\text{circ}}(\mathbf{a}, \mathbf{b}) - 1}{2} \right]^T \right\|_2 \tag{14}$$

From the above transformations, it is clear that when $\mathbf{u}$ is fixed, (12) is an MI-SOCP feasibility problem, which can be solved optimally by modern MI-SOCP solvers such as MOSEK and Gurobi. Despite exponential worst-case complexity, these mixed integer solvers can solve (12) reasonably fast in practice, especially when leveraging the distributed and parallel optimization capability in a cloud computing platform.

Based on the above analysis, problem (8) can now be expressed as $\max\{f(\mathbf{u}) | \mathbf{u} \in \mathcal{Q} \subset \mathcal{U}\}$. First, we check whether $\underline{\mathbf{u}}$ is feasible or not. If so, we apply the proposed BRB algorithm, which is outlined in Algorithm 1, to find a globally optimal solution to (8). In principle, the proposed BRB algorithm recursively branches a box $\mathcal{B}$ into two smaller boxes, checks the feasibility of each new box, update the current upper and lower bounds by the box reduction and bound computation process, and removes the boxes that do not contain an optimal solution. The details of these operations can be found in [28], [41], and thus not presented here for space limitation. The upper and lower bound of a box $\mathcal{B} = [\underline{\mathbf{u}}, \bar{\mathbf{u}}]$ are computed by $UB(\mathcal{B}) = f(\bar{\mathbf{u}})$ and $LB(\mathcal{B}) = f(\underline{\mathbf{u}})$ due to the monotonic increase of $f(\mathbf{u})$, respectively. After updating the current best lower bound ζ_n, the pruning is performed to delete the boxes whose upper bounds are smaller than ζ_n. According to [41], the proposed BRB algorithm is bound–improving and guaranteed to terminate after a finite number of iterations for a given desired accuracy level ϵ.

The proposed algorithm to find an optimal solution to (8) is summarized in Algorithm 1 and its convergence is presented in the following.

Algorithm 1 Proposed BRB algorithm.

1: Apply box reduction to $\mathcal{U}$ to obtain redu $(\mathcal{U})$
2: $n = 1$; $\mathcal{B}_1 = \text{redu}(\mathcal{U})$; $\mathcal{D}_1 = \{\mathcal{B}_1\}$; $\zeta_1 = LB(\mathcal{B}_1)$;
3: **repeat**
4: Select the box with the largest upper bound to branch: $\mathcal{B}_n = \arg\max_{\mathcal{B}_i \subset \mathcal{D}_n} UB(\mathcal{B}_i)$;
5: Branch the box $\mathcal{B}_n$ into two small boxes $\mathcal{B}_n^{(1)}$ and $\mathcal{B}_n^{(2)}$; // *Box Branching* //
6: **for** $j = 1 : 2$ **do**
7: Compute lower bound set of $\mathcal{B}_n^{(j)}$, denoted as $\underline{\mathbf{X}}_n^{(j)} = \left\{\underline{\mathbf{u}}_n^{(j)}\right\}$;
8: **if** $\underline{\mathbf{X}}_n^{(j)}$ is feasible **then**
9: Apply box reduction to $\mathcal{B}_n^{(j)}$ to obtain redu $\left(\mathcal{B}_n^{(j)}\right)$; // *Box Reduction* //
10: **else** $\underline{\mathbf{X}}_n^{(j)} = \emptyset$;
11: **end if**
12: Compute lower bound $LB\left(\text{redu}\left(\mathcal{B}_n^{(j)}\right)\right)$, upper bound $UB\left(\text{redu}\left(\mathcal{B}_n^{(j)}\right)\right)$ from the reduced box; // *Bound Computation* //
13: **end for**
14: Update the lower bound: $\zeta_{n+1} = \max\left(LB\left(\text{redu}\left(\mathcal{B}_n^{(1)}\right)\right), LB\left(\text{redu}\left(\mathcal{B}_n^{(2)}\right)\right), \zeta_n\right)$;
15: Update the set of boxes: $\mathcal{D}_{n+1} = \left\{\mathcal{D}_n, \mathcal{B}_n^{(1)}, \mathcal{B}_n^{(2)}\right\}$;
16: Delete the box that do not contain optimal solution: $\mathcal{D}_{n+1} = \mathcal{D}_n \setminus \{\mathcal{B}_i | \zeta_{n+1} > UB(\text{redu}(\mathcal{B}_i)), \forall i = 1, \ldots, \text{cardinal}(\mathcal{D}_n)\}$; // *Pruning* //
17: $n = n + 1$;
18: **until** $|\max_{\mathcal{B}_i \subset \mathcal{D}_n} UB(\text{redu}(\mathcal{B}_i)) - \zeta_n| \leq \epsilon$;

C. Convergence analysis

Algorithm 1 is guaranteed to compute an optimal solution to (8) and its convergence can be proved using the same arguments as those in [41], which can be explained as follows. First, the branching rule improves the lower and upper bounds of the objective (8a) after every iteration. Specifically, by the updating rule in Step 14, the lower bound is non-decreasing after each iteration. Due to the box reduction and bound computation rule, the upper bound is non-increasing. After a finite number of iterations, Algorithm 1 will create a set of boxes that contain an optimal solution, and the gap between the upper bound and lower bound is less than or equal to ϵ, where ϵ is a predetermined desired accuracy level.

IV. Low-complexity Algorithms

In general, computing a globally optimal solution to (6) is very difficult and even if possible, it is of little practical use in wireless communications since the channel conditions can change quickly. Thus, the proposed optimal solution is

mostly useful for benchmarking purposes. For more practically appealing methods, we derive in this section three iterative low-complexity approaches to find a high-quality feasible solution to (6). In the first approach, we employ SCA method to convexify the non-convex continuous part of problem (6). The problem at each iteration of the proposed algorithm is an MI-SOCP. However, the number of MI-SOCPs that needs to be solved is significantly reduced, compared to the optimal BRB method, since the SCA-based convexification converges rapidly. In the second approach, we further lower the computational complexity of the first one by allowing the binary variables to be continuous. This results in a series of SOCP being solved until convergence. For a continuous relaxation method, it is generally known that the obtained solution may not produce a high-performance (or even a feasible) solution. To this end, we carry out a post-processing procedure over the obtained solution to search for a high-quality solution. In the final method, the problem is reformulated from the viewpoint of sparsity-inducing regularization and solved iteratively by applying a ℓ_2-norm based logarithm approximation in combination with the SCA-based approximation.

A. New Equivalent Transformation

We first remark that, although (6) and (8) are equivalent as shown in Lemma 1, their feasible sets are different. It means a feasible solution of (8) might be infeasible to (6). This can be verified by observing that in (8b), we can find a feasible solution $\bar{\mathbf{w}}, \bar{u}_k, \bar{a}_{i,k}, \forall i,k$ such that $R_k(\bar{\mathbf{w}}) > \bar{u}_k$ and $\sum_{k\in\mathcal{K}} \bar{a}_{i,k} R_k(\bar{\mathbf{w}}) > \frac{C_i}{\xi_i}$, violating constraint (6h). In this section we are about to apply SCA optimization to find low-complexity algorithms that provide suboptimality of (6). Thus a new transformation with an equivalent feasible set is necessary. To this end, we consider the following formulation

$$\max_{\substack{\mathbf{w},\mathbf{a},\mathbf{b},\\ \boldsymbol{\nu},\boldsymbol{\mu},\boldsymbol{\gamma}}} \ \tilde{\alpha}\sum_{k\in\mathcal{K}}\mu_k - \bar{\alpha}P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \tag{15a}$$

$$\text{s.t.} \ \ \log(1+\gamma_k) \geq \mu_k \tag{15b}$$

$$\Gamma_k(\mathbf{w}) \geq \gamma_k \tag{15c}$$

$$\gamma_k \geq \Gamma_k^{\min} \tag{15d}$$

$$(7), (6c), (6e)-(6i) \tag{15e}$$

It is easy to see that a solution feasible to (15) is also feasible to (6). Moreover, all the constraints (15b)–(15c) are active at optimality. Thus (15) is an equivalent formulation of (6) that serves the purpose mentioned above. From the previous discussions, it is clear that the continuous nonconvexity of (15) is due to (6h) and (15c). First, we rewrite (6h) as

$$\sum_{k=1}^{K} a_{i,k}\log(1+t_k) \leq \frac{C_i}{\xi_i}, \tag{16a}$$

$$\frac{|\mathbf{h}_k^H\mathbf{w}_k|^2}{\sum_{j\neq k}^{K}|\mathbf{h}_k^H\mathbf{w}_j|^2+\sigma_0^2} \leq t_k, \tag{16b}$$

where $\mathbf{t} = \{t_k \geq 0\}_{\forall k}$ is the set of newly introduced variables. Moreover, with the introduction of additional variables $\mathbf{z} = \{z_k \geq 0\}_{\forall k}$, we can rewrite (16a) as

$$\sum_{k\in\mathcal{K}} a_{i,k}^2/z_k \leq \frac{C_i}{\xi_i}, \tag{17}$$

$$1 + t_k \leq e^{1/z_k}. \tag{18}$$

A subtle point should be made here. In fact, to arrive at (17), we have used the fact that $a_{i,k} = a_{i,k}^2$ for $a_{i,k} \in \{0,1\}$. This maneuver has two purposes. Firstly, (17) is SOC representable. Secondly, if $a_{i,k}$ is allowed to be continuous on $[0,1]$, then it holds that $a_{i,k} \geq a_{i,k}^2$. Thus, if $a_{i,k}$ satisfies (16a), then it also does for (17). As a result, continuous relaxation based on $a_{i,k}^2$ will yield a tighter bound, compared to using $a_{i,k}$. This important property will be exploited to derive a high-performance solution based on the continuous relaxation. To summary, we can equivalently rewrite (6) as

$$\max_{\substack{\mathbf{w},\mathbf{a},\mathbf{b},\\ \boldsymbol{\nu},\boldsymbol{\mu},\boldsymbol{\gamma}\\ \mathbf{t},\mathbf{z}}} \ \tilde{\alpha}\sum_{k\in\mathcal{K}}\mu_k - \bar{\alpha}P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \tag{19a}$$

$$\text{s.t.} \ \ \log(1+\gamma_k) \geq \mu_k \tag{19b}$$

$$\sum_{j\in\mathcal{K}\setminus k}|\mathbf{h}_k^H\mathbf{w}_j|^2 + \sigma_0^2 \leq \frac{|\mathbf{h}_k^H\mathbf{w}_k|^2}{\gamma_k} \tag{19c}$$

$$1 + t_k \leq e^{1/z_k}, \tag{19d}$$

$$\frac{|\mathbf{h}_k^H\mathbf{w}_k|^2}{t_k} \leq \sum_{j\neq k}^{K}|\mathbf{h}_k^H\mathbf{w}_j|^2 + \sigma_0^2, \tag{19e}$$

$$(6e)-(6g), (6i), (6c), (7), (15d), (17). \tag{19f}$$

We remark that problem (19) is still non-convex but its non-convexity is easier to handle in light of SCA as demonstrated in the following.

B. SCA-MISOCP Algorithm

In the first iterative method we preserve the Boolean variables, and only approximate the continuous nonconvex parts of (19). In particular, we do so by applying the framework of SCA optimization. Explicitly, at iteration n of the proposed algorithm, the right side of (19c) is simply replaced by its first order Taylor approximation around the points $\mathbf{w}_k^{(n)}$ and $\gamma_k^{(n)}$

$$\frac{\left|\mathbf{h}_k^H\mathbf{w}_k\right|^2}{\gamma_k} \overset{(a)}{\geq} H\big(\mathbf{w}_k,\gamma_k;\mathbf{w}_k^{(n)},\gamma_k^{(n)}\big) = \frac{2\,\mathrm{Re}\big(\mathbf{w}_k^{(n)H}\mathbf{H}_k\mathbf{w}_k\big)}{\gamma_k^{(n)}} - \frac{\left|\mathbf{h}_k^H\mathbf{w}_k^{(n)}\right|^2}{\gamma_k^{(n)2}}\gamma_k \tag{20}$$

where $\mathbf{H}_k \triangleq \mathbf{h}_k\mathbf{h}_k^H$, and we have denoted $\mathbf{w}_k^{(n)H} = \big(\mathbf{w}_k^{(n)}\big)^H$ and $\gamma_k^{(n)2} = \big(\gamma_k^{(n)}\big)^2$ to lighten the notation. In the same way we convexify the right sides of (19d) and (19e) by the first order Taylor approximation as

$$e^{\frac{1}{z_k}} \overset{(b)}{\geq} F\big(z_k;z_k^{(n)}\big) = e^{1/z_k^{(n)}} - \frac{e^{1/z_k^{(n)}}}{z_k^{(n)2}}\big(z_k - z_k^{(n)}\big) \tag{21}$$

$$\sum_{j\neq k}^{K}\left|\mathbf{h}_k^H\mathbf{w}_j\right|^2 + \sigma_0^2 \overset{(c)}{\geq} G\big(\mathbf{w};\mathbf{w}^{(n)}\big) = \sum_{j\neq k}^{K} 2\,\mathrm{Re}\big(\mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j\big) - \sum_{j\neq k}^{K}\mathbf{w}_j^{(n)H}\mathbf{H}_k\mathbf{w}_j^{(n)} + \sigma_0^2 \tag{22}$$

By applying these approximations into the non-convex constraints (19c),(19d) and (19e), we can formulate the MI-convex approximation of problem (19) at iteration $n+1$ as below

$$\max_{\substack{\mathbf{w},\mathbf{a},\mathbf{b},\\ \boldsymbol{\nu},\boldsymbol{\mu},\boldsymbol{\gamma}\\ \mathbf{t},\mathbf{z}}} \ \tilde{\alpha}\sum_{k\in\mathcal{K}}\mu_k - \bar{\alpha}P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b}) \tag{23a}$$

$$\text{s.t.}\ \log\left(1+\gamma_k\right) \geq \mu_k \tag{23b}$$

$$\sum_{j\in\mathcal{K}\setminus k}|\mathbf{h}_k^H\mathbf{w}_j|^2+\sigma_0^2 \leq H\big(\mathbf{w}_k,\gamma_k;\mathbf{w}_k^{(n)},\gamma_k^{(n)}\big) \tag{23c}$$

$$1+t_k \leq F\big(z_k;z_k^{(n)}\big) \tag{23d}$$

$$\left|\mathbf{h}_k^H\mathbf{w}_k\right|^2/t_k \leq G\big(\mathbf{w};\mathbf{w}^{(n)}\big) \tag{23e}$$

$$(6e)-(6g),(6i),(6c),(7),(15d),(17). \tag{23f}$$

where $\mathbf{w}^{(n)},\mathbf{z}^{(n)},\boldsymbol{\gamma}^{(n)}$ are the parameters to be updated at the $(n+1)^{\text{th}}$ iteration.

Remark 2. Note that all the continuous constraints in (23), except (23b), are convex quadratic representable. Thus (23) is recognized as a generic convex mixed-integer program for which dedicated solvers are quite limited. We note that (23b) is a convex constraint, and thus convex approximation is not required and in fact convexity should be preserved in general. In an effort to do so, while still able to avail of more powerful solvers, the authors in [28] approximate (23b) by a system of SOC constraints. In this way, (23) reduces to an MI-SOCP for which modern solvers such as MOSEK or GUROBI have proved to be very efficient. However, the number of SOC constraints (and the number of slack variables) required to approximate the exponential cone in (23b) increases quickly with the accuracy.

In this paper we propose a novel approach to transform (23) into an MI-SOCP. The key is due to the following inequality. For any $\gamma_k \geq 0$ it holds that

$$\log(1+\gamma_k) \geq U\big(\gamma_k;\gamma_k^{(n)}\big) = \log\big(1+\gamma_k^{(n)}\big) + \frac{1}{1+\gamma_k^{(n)}}\big(\gamma_k-\gamma_k^{(n)}\big) - \frac{1}{2}\big(\gamma_k-\gamma_k^{(n)}\big)^2. \tag{24}$$

In fact $U\big(\gamma_k;\gamma_k^{(n)}\big)$ is a quadratic lower bound of $\log(1+\gamma_k)$ around $\gamma_k^{(n)}$, which is derived from the Lipschitz continuity of the derivative of $\log(1+\gamma_k)$. The proof is given in Appendix B. To obtain an MI-SOCP formulation of (23), we replace (23b) by a SOC representation as

$$U(\gamma_k;\gamma_k^{(n)})) \geq \mu_k \tag{25}$$

The first proposed algorithm, referred to as the SCA-MISOCP based algorithm, is outlined in Algorithm 2.

Convergence Analysis: We now prove that Algorithm 2 is guaranteed to converge. This can be established by showing that the sequence of objectives returned by Algorithm 2 is monotonically convergent. Towards this end, let $\theta^{(n)}$ and $\boldsymbol{\Theta}^{(n)}$ denote the optimal objective value and the achieved optimal solution at the n^{th} iteration of Algorithm 2, respectively. Due to the first order approximation in (20), (21) and (22), it holds that equalities occur at (a) when $(\mathbf{w}_k^{(n)},\gamma_k^{(n)})=(\mathbf{w}_k,\gamma_k)$, at (b) when $z_k^{(n)}=z_k$, and (c) when $\mathbf{w}^{(n)}=\mathbf{w}$, respectively.

Algorithm 2 SCA-MISOCP based algorithm.

1: Set $n := 0$ and initialize starting points of $\mathbf{w}^{(n)},z^{(n)},\boldsymbol{\gamma}^{(n)}$;
2: **repeat**
3: Solve the approximated problem (23) with the SOC approximation (25) at $\mathbf{w}^{(n)},\mathbf{z}^{(n)},\boldsymbol{\gamma}^{(n)}$ to achieve the optimal solution $\mathbf{a}^\star,\mathbf{b}^\star,\boldsymbol{\gamma}^\star,\mathbf{t}^\star,\boldsymbol{\mu}^\star,\boldsymbol{\nu}^\star,\mathbf{w}^\star,\mathbf{z}^\star$;
4: Update $\mathbf{w}^{(n+1)}=\mathbf{w}^\star,\mathbf{z}^{(n+1)}=\mathbf{z}^\star,\boldsymbol{\gamma}^{(n+1)}=\boldsymbol{\gamma}^\star$;
5: Set $n:=n+1$;
6: **until** Convergence;

Then, the updating rule in Algorithm 2 (cf. Step 4 in Algorithm 2) ensures that $\boldsymbol{\Theta}^{(n)}$ is also feasible to problem (23) at the $(n+1)^{\text{th}}$ iteration. This subsequently leads to $\theta^{(n+1)}\geq\theta^{(n)}$, meaning that Algorithm 2 generates a non-decreasing sequence of objective function values. Due to the power budget constraint (6c), the sequence of objectives $\{\theta^{\{n\}}\}$ is upper bounded and thus, is convergent.

Generation of Initial Point : To start the iterative process in Algorithm 2, it is essential to find a feasible point in Step 1 of Algorithm 2. For this purpose, we can simply set $t_k=\Gamma_{\min}$, $\gamma_k=\Gamma_{\min}$, $\mu_k=\log(1+\Gamma_{\min})$ for all $k\in\mathcal{K}$, and then solve the following feasibility problem (Pini) $=$ find$\{\mathbf{a},\mathbf{z}|z_k\leq 1/\log(1+\Gamma_{\min}),(17),(6g),a_{i,k}\in\{0,1\}\}$. We remark that the problem (Pini) is a feasibility MI-SOCP program which can be solved optimally by off-the-shelf solvers such as MOSEK or GUROBI. Next, from the obtained value of $\mathbf{a},\mathbf{t},\mathbf{z},\boldsymbol{\gamma},\boldsymbol{\mu}$, we consider the following mixed integer program (Pmin) $=\min_{\mathbf{w},\mathbf{b},\boldsymbol{\nu}}\{P^{\text{tot}}(\mathbf{w},\mathbf{a},\mathbf{b})|(6b)-(6f),b_i\in\{0,1\}\}$. Note that the constraints (6c)-(6f) are SOC representable as discussed earlier. In fact, (6b) can also be reformulated by a SOC constraint as shown in [28], [29], which can be briefly explained as follows. It is easy to check that if $\mathbf{w}_k,\forall k\in\mathcal{K}$ is feasible to (6), then a phase rotation on $\mathbf{w}_k$ (i.e., replace $\mathbf{w}_k$ by $\mathbf{w}_k e^{j\phi_k}$ for some $\phi_k\in[0,2\pi]$) creates another feasible solution of the same objective value. Therefore, without loss of optimality, $\mathbf{w}_k$ can be chosen such that $\mathbf{h}_k^H\mathbf{w}_k$ is real and non-negative $\forall k\in\mathcal{K}$. As a result, (6b) is equivalent to the following two constraints

$$c\text{Re}\left(\mathbf{h}_k^H\mathbf{w}_k\right) \geq \left\|\left[\mathbf{h}_k^H\mathbf{w}_1,\ldots,\mathbf{h}_k^H\mathbf{w}_K,\sigma_0^2\right]^T\right\|_2 \tag{26a}$$

$$\text{Im}\left(\mathbf{h}_k^H\mathbf{w}_k\right)=0 \tag{26b}$$

where $c=\sqrt{(\Gamma_{\min}+1)/\Gamma_{\min}}$. Now, it is clear that (Pmin) is a MI-SOCP problem, and thus can be solved optimally. The obtained values of $\mathbf{w},\mathbf{z},\boldsymbol{\gamma}$ by solving (Pini) and (Pmin) are then used to start Algorithm 2. Alternative option is to initialize Algorithm 2 from a feasible solution that can be found by the suboptimal algorithms presented in the subsequent subsections.

C. Continuous relaxation and inflation based algorithm

To develop a more practically appealing algorithm, we further consider the continuous relaxation of (23), i.e., $0\leq b_i\leq 1, 0\leq a_{i,k}\leq 1$ for $\forall i\in\mathcal{I},\forall k\in\mathcal{K}$. As a result, the continuous relaxation of (23), denoted as $(\mathcal{P}^r)$, becomes an

SOCP which can be solved in polynomial time. The second proposed iterative method combines two stages: (i) continuous relaxation and (ii) post-processing. In the first stage, we follow an iterative algorithm similar to Algorithm 2, but simply solve $(\mathcal{P}^r)$ in Step 3. The post-processing process is then used to map the obtained b_i's and $a_{i,k}$'s to the binary values, which is required due to the continuous relaxation. Towards this end, we apply the inflation procedure in [21] to refine the achieved solution. In particular, we rely on the solution to the continuous relaxation at convergence as an incentive measure to make a decision on the binary value of $\mathbf{a}$ and $\mathbf{b}$. Let us denote $\tilde{\mathbf{a}}, \tilde{\mathbf{b}}$ and $\tilde{\mathbf{w}}$ as the solution achieved after the first stage. Intuitively, the connection between the i^{th} RRH and the k^{th} UE is more likely if the channel of the link is in better condition and the power consumed to transmit fronthaul data $P_{i,k}^{\text{FH}}$ is smaller than the others. Consequently, solving the continuous relaxation would possibly yield higher $\widetilde{b}_i$ for the i^{th} RRH and higher $\widetilde{a}_{i,k}$ for the connection between the i^{th} RRH and the k^{th} UE. Based on the above intuitive observations, we propose an iterative procedure to determine the set of active RRHs and RRH-UE association based on $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$. The process starts by assuming that all the RRHs are off and there is no association between RRH and UE. In each iteration, $(\mathcal{P}^r)$ is solved with a set of remaining inactive RRHs and RRH-UE association that is not connected. The RRH-UE association with the largest $\tilde{a}_{i,k}$ will be made connected and the resulting RRH will be set active, following the relationship in (6f). The overall algorithm is presented in Algorithm 3.

Algorithm 3 Relax continuous and inflation based algorithm

1: Set $m := 0$, $\pi^{(m)}$ is significantly small, and initialize the set $\mathcal{R}_{\text{off}}^{(m)} = \{(i,k) \times i \in (\mathcal{I}, \mathcal{K}) \times \mathcal{I}\}$.
2: **repeat**
3: Set $m := m + 1$;
4: Solve $(\mathcal{P}^r)$ with $a_{i',k'} = 1$ and $b_{i'} = 1, \forall \{(i',k') \times i'\} \notin \mathcal{R}_{\text{off}}^{(m-1)}$;
5: Update $\mathcal{R}_{\text{off}}^{(m)} = \mathcal{R}_{\text{off}}^{(m-1)} \setminus \left\{(i',k') \times i' = \arg\max_{i,k \in \mathcal{R}_{\text{off}}^{(m-1)}} \widetilde{a}_{i,k}\right\}$;
6: Solve (23) with (25) given $a_{i',k'} = 1, b_{i'} = 1, \forall \{(i',k') \times i'\} \notin \mathcal{R}_{\text{off}}^{(m)}$ and $a_{i,k} = 0, b_i = 0, \forall \{(i,k) \times i\} \in \mathcal{R}_{\text{off}}^{(m)}$, denoted as $\left(\mathcal{P}^{int}\right)$. If $\left(\mathcal{P}^{int}\right)$ is feasible, set $\pi^{(m)}$ as the value of objective function achieved at the convergence. If not, set $\pi^{(m)} = \pi^{(0)}$.
7: **until** $(\mathcal{P}^r)$ starts to be infeasible or $\left(\mathcal{P}^{int}\right)$ is feasible and $\pi^{(m)} < \pi^{(m-1)}$;
8: Solve (23) with (25) given $a_{i',k'} = 1, b_{i'} = 1, \forall \{(i',k') \times i'\} \notin \mathcal{R}_{\text{off}}^{(m-1)}$ and $a_{i,k} = 0, b_i = 0, \forall \{(i,k) \times i\} \in \mathcal{R}_{\text{off}}^{(m-1)}$ to obtain $\mathbf{w}^\star, \boldsymbol{\nu}^\star, \mathbf{t}^\star, \mathbf{z}^\star, \mu^\star, \gamma^\star$;

Convergence Analysis: Algorithm 3 is provably convergent due to two facts. First, the SCA-based algorithm to solve $(\mathcal{P}^r)$ is guaranteed to converge and this can be proved in the same way as done for Algorithm 2. Second, the post-processing procedure is executed $(I-1)\,K$ times in the worst case. In the last step when all the binary variables have been fixed, the SCA-based algorithm is applied to solve (23) until convergence. Note that in this case, we deal with a continuous optimization problem and a stronger convergence result can be achieved. Specifically, every limit point of the SCA-based algorithm is a stationary solution to the continuous optimization problem. However, we remark that a stationary point is not necessarily a locally optimal solution. Exploring further properties of the obtained stationary solution is beyond the scope of the paper.

Generation of Initial Point: To apply the SCA-based algorithm to solve $(\mathcal{P}^r)$ in the first iteration of Algorithm 3 (i.e, when $0 \le a_{i,k} \le 1$ and $0 \le b_i \le 1$), we may need a feasible point. However, as mentioned earlier, the challenge is that $(\mathcal{P}^r)$ (cf. (23)) is nonconvex with the remaining other continuous variables, making it difficult to find a feasible point. There is in fact a penalty method to allow the SCA-based procedure to start from an infeasible point, which is described in [30]. The idea is to introduce slack variables into each constraint as the violations and penalizing the sum of these violations in the objective. In this way, first iterations of Algorithm 3 may be infeasible to (23), but violations are forced to be zero as the iterative process progresses. We refer the interested reader to [30, Algorithm 3.1] for a complete description of this initialization method.

D. Sparsity-inducing Norm Approach

In the final low-complexity approach, we reformulate the sum rate-power maximization from a viewpoint of group sparsity. Note that the i^{th} RRH will not be selected if the vector $\tilde{\mathbf{w}}_i = \left[\mathbf{w}_{i,1}^H, \ldots, \mathbf{w}_{i,K}^H\right]$ which includes all beamformers related to the i^{th} RRH is a zero vector. Let us rewrite the total power consumption as

$$P_{\text{sparse}}^{\text{tot}}(\mathbf{w}) = \frac{1}{\eta_i} \sum_{\forall i \in \mathcal{I}} \left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \sum_{\forall i \in \mathcal{I}} \chi\left(\left\|\tilde{\mathbf{w}}_i\right\|_2^2\right) \left(P_i^{\text{ra}} - P_i^{\text{ri}}\right) + \sum_{\forall i \in \mathcal{I}} P^{\text{ri}} + \sum_{\forall k \in \mathcal{K}} \sum_{\forall i \in \mathcal{I}} \chi\left(\left\|\mathbf{w}_{i,k}\right\|_2^2\right) P_{i,k}^{\text{FH}}. \quad (27)$$

The sum rate–power optimization can now be written as

$$\max_{\mathbf{w}} \quad \tilde{\alpha} \sum_{k=1}^{K} R_k(\mathbf{w}) - \bar{\alpha} P_{\text{sparse}}^{\text{tot}}(\mathbf{w}) \quad (28a)$$

$$\text{s.t. } \Gamma_k(\mathbf{w}) \ge \Gamma_k^{\min}, \forall k \in \mathcal{K} \quad (28b)$$

$$\sum_{\forall k \in \mathcal{K}} \chi\left(\left\|\mathbf{w}_{i,k}\right\|_2^2\right) R_k(\mathbf{w}) \le \frac{C_i}{\xi_i}, \forall i \in \mathcal{I}, \quad (28c)$$

$$\left\|\tilde{\mathbf{w}}_i\right\|_2^2 \le P^{\max}, \forall i \in \mathcal{I}. \quad (28d)$$

In fact we can impose sparsity on the soft power vector $\boldsymbol{\nu}$ to derive the sparsity-inducing norm method. However our idea is to impose sparsity directly on the beamforming vector $\mathbf{w}$ to arrive at (28). Thus, all slack variables are not introduced to reduce the complexity of the resulting formulation. However, problem (28) is still non-convex due to the presence of the indication functions, which are intractable. To deal with this problem, we will replace $\chi(x)$ by $\log(\tau + x)$ for a small $\tau > 0$, following the result in [31]. In this way we can approximate

$$\chi\left(\left\|\tilde{\mathbf{w}}_i\right\|_2^2\right) \approxeq \log\left(\left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \tau_1\right) \quad (29)$$

$$\chi\left(\left\|\mathbf{w}_{i,k}\right\|_2^2\right) \approxeq \log\left(\left\|\mathbf{w}_{i,k}\right\|_2^2+\tau_2\right) \tag{30}$$

where $\tau_1, \tau_2 > 0$ are small positive parameters. Consequently, we can obtain a continuous approximation of (28) as

$$\max_{\mathbf{w}} \quad \tilde{\alpha}\sum_{k=1}^{K} R_k(\mathbf{w}) - \bar{\alpha}\tilde{P}^{\text{tot}}_{\text{sparse}}(\mathbf{w},\mathbf{p},\mathbf{q}) \tag{31a}$$

$$\text{s.t.} \quad \sum_{\forall k\in\mathcal{K}} q_{i,k}^2 R_k(\mathbf{w}) \le \frac{C_i}{\xi_i}, \forall i\in\mathcal{I}, \tag{31b}$$

$$\log\left(\left\|\tilde{\mathbf{w}}_i\right\|_2^2+\tau_1\right) \le p_i, \forall i\in\mathcal{I}, \tag{31c}$$

$$\log\left(\left\|\mathbf{w}_{i,k}\right\|_2^2+\tau_2\right) \le q_{i,k}^2 \forall i\in\mathcal{I}, \forall k\in\mathcal{K} \tag{31d}$$

$$(28\text{b}), (28\text{d}) \tag{31e}$$

where we have introduced $\mathbf{q} = \{q_{i,k} \ge 0, \forall k\in\mathcal{K}, \forall i\in\mathcal{I}\}$ and $\mathbf{p} = \{p_i \ge 0, \forall i\in\mathcal{I}\}$ and defined

$$\tilde{P}^{\text{tot}}_{\text{sparse}}(\mathbf{w},\mathbf{p},\mathbf{q}) = \frac{1}{\eta_i}\sum_{\forall i\in\mathcal{I}}\left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \sum_{\forall i\in\mathcal{I}} p_i\left(P_i^{\text{ra}} - P_i^{\text{ri}}\right) + \sum_{\forall i\in\mathcal{I}} P_i^{\text{ri}} + \sum_{\forall k\in\mathcal{K}}\sum_{\forall i\in\mathcal{I}} q_{i,k}^2 P_{i,k}^{\text{FH}}. \tag{32}$$

Note that $\tilde{P}^{\text{tot}}_{\text{sparse}}(\mathbf{w},\mathbf{p},\mathbf{q})$ is convex with the involving variables, and that the purpose of using the second order on $q_{i,k}$ is to reuse the approximations presented previously, as we will show shortly. The constraint in (31c) can be equivalently rewritten as

$$\left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \tau_1 \le e^{p_i}, \forall i\in\mathcal{I} \tag{33}$$

and thus can be approximated by

$$\left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \tau_1 \le e^{p_i^{(n)}} + e^{p_i^{(n)}}(p_i - p_i^{(n)}) \triangleq \tilde{F}(p_i; p_i^{(n)}). \tag{34}$$

In the same way (31d) can be approximated as

$$\left\|\mathbf{w}_{i,k}\right\|_2^2+\tau_2 \le e^{q_{i,k}^{(n)2}} + 2q_{i,k}^{(n)} e^{q_{i,k}^{(n)2}}(q_{i,k} - q_{i,k}^{(n)}) \triangleq \bar{F}(q_{i,k}; q_{i,k}^{(n)}). \tag{35}$$

Here we write $e^{q_{i,k}^{(n)2}}$ instead of $e^{\left(q_{i,k}^{(n)}\right)^2}$ to lighten the notation. Unlike to approach that fixes the rate function $R_k(\mathbf{w})$ in each iteration in [15], [16], here, we deal with the nonconvexity in (31b) by equivalently rewriting it as

$$\begin{cases} \displaystyle\sum_{k\in\mathcal{K}} \frac{q_{i,k}^2}{z_k} \le \frac{C_i}{\xi_i}, \\ (16\text{b}), (18). \end{cases} \tag{36}$$

where $\mathbf{t}$ and $\mathbf{z}$ are introduced as done similarly in (16b)–(18). Now the approximations used to deal with (16b)–(18) can be applied, which results in the following convex approximated problem (28) at the $(n+1)^{\text{th}}$ iteration of the sparsity-based iterative algorithm

$$\max_{\substack{\mathbf{w},\boldsymbol{\mu},\boldsymbol{\gamma},\mathbf{t}\\ \mathbf{z},\mathbf{p},\mathbf{q}}} \ \tilde{\alpha}\sum_{k\in\mathcal{K}}\mu_k - \bar{\alpha}\tilde{P}^{\text{tot}}_{\text{sparse}}(\mathbf{w};\mathbf{p};\mathbf{q}) \tag{37a}$$

$$\text{s.t.} \ \left\|\tilde{\mathbf{w}}_i\right\|_2^2 + \tau_1 \le \tilde{F}\left(p_i; p_i^{(n)}\right) \tag{37b}$$

$$U\left(\gamma_k; \gamma_k^{(n)}\right) \ge \mu_k \tag{37c}$$

$$\gamma_k \ge \Gamma_k^{\min} \tag{37d}$$

$$\sum_{j\in\mathcal{K}\setminus k} |\mathbf{h}_k^H\mathbf{w}_j|^2 + \sigma_0^2 \le H\left(\mathbf{w}_k, \gamma_k; \mathbf{w}_k^{(n)}, \gamma_k^{(n)}\right) \tag{37e}$$

$$1 + t_k \le F\left(z_k; z_k^{(n)}\right) \tag{37f}$$

$$\left\|\mathbf{w}_{i,k}\right\|_2^2 + \tau_2 \le \bar{F}\left(q_{i,k}; q_{i,k}^{(n)}\right) \tag{37g}$$

$$\left|\mathbf{h}_k^H\mathbf{w}_k\right|^2 / t_k \le G\left(\mathbf{w}; \mathbf{w}^{(n)}\right) \tag{37h}$$

$$\left\|\tilde{\mathbf{w}}_i\right\|_2^2 \le P^{\max}, \forall i\in\mathcal{I}, \tag{37i}$$

$$\sum_{k\in\mathcal{K}} \frac{q_{i,k}^2}{z_k} \le \frac{C_i}{\xi_i}, \forall i\in\mathcal{I} \tag{37j}$$

where $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \boldsymbol{\gamma}^{(n)}, \mathbf{p}^{(n)}, \mathbf{q}^{(n)}$ are the parameters that are updated at the $(n+1)^{\text{th}}$ iteration. The proposed iterative approach to solve problem (28) is given in Algorithm 4. Note that the convergence of Algorithm 4 can be established following the same arguments as those in Algorithms 2 and 3 above. Also, the generation of an initial point for Algorithm 4 can be carried out the same way as for Algorithm 3.

Algorithm 4 SCA-Sparsity based algorithm.

1: Set $n := 0$ and initialize starting points of $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \boldsymbol{\gamma}^{(n)}, \mathbf{p}^{(n)}, \mathbf{q}^{(n)}$;
2: **repeat**
3: Solve the approximated problem (37) at $\mathbf{w}^{(n)}, \mathbf{z}^{(n)}, \boldsymbol{\gamma}^{(n)}, \mathbf{p}^{(n)}, \mathbf{q}^{(n)}$ to achieve the optimal solution $\mathbf{w}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\gamma}^\star, \mathbf{t}^\star, \mathbf{z}^\star, \mathbf{p}^\star, \mathbf{q}^\star$;
4: Update $\mathbf{w}^{(n+1)} = \mathbf{w}^\star, \mathbf{z}^{(n+1)} = \mathbf{z}^\star, \boldsymbol{\gamma}^{(n+1)} = \boldsymbol{\gamma}^\star, \mathbf{p}^{(n+1)} = \mathbf{p}^\star, \mathbf{q}^{(n+1)} = \mathbf{q}^\star$;
5: Set $n := n+1$;
6: **until** Convergence;

E. Complexity Analysis

We now discuss the complexity of each proposed algorithm in this section. For the optimal design based on a BRB method, i.e., Algorithm 1, the complexity is extremely high since the number of the boxes needs to be considered increases exponentially with the problem dimension. Moreover, in each iteration, an MI-SOCP feasibility problem is solved. For Algorithm 2, the overall complexity mainly depends on that of solving the MI-SOCP problem in (23) which is indeed a combinatorial optimization problem. In particular, there are IK binary variables $a_{i.k}$'s and I binary variables b_i's, resulting in 2^{IK+I} combinations for all the binary variables. Given fixed $\mathbf{a}$ and $\mathbf{b}$, all the constraints in problem (23) approximately consist of a total number of $KM+2IK+4K+1$ variables and a number of $3IK+2I+4K+1$ SOC constraints of dimension $KM+1$. Thus, the worst-case complexity of Algorithm 2 in each iteration can be written as $\mathcal{O}\left(2^{IK+K}(K^4M^3I)\right)$. Compared to Algorithm 1, Algorithm 2 has less complexity due to the continuous approximation converging rapidly.

Next, we analyze the complexity of Algorithms 3 and 4. First we remark that in the worst case, Algorithm 3 must iteratively solve and update the resulting parameters for the SOCP problem $(\mathcal{P}^r)$ and $\left(\mathcal{P}^{int}\right)$ for $(I-1)K$ times. In each step, the complexity of solving $(\mathcal{P}^r)$ and $\left(\mathcal{P}^{int}\right)$ is

TABLE I: Simulation parameters

Description	Notation	Value
Number of RRHs	I	6
Number of users	K	4
Number of antennas per RRH	M_i	2
Power amplifier efficiency	η_i	0.35
Maximum transmit power	$P^{\max}$	10 dBW
Active power for RRH and fronthaul	P_i^{ra}	12.5 dBW
Sleep power for RRH and fronthaul	P_i^{ri}	2.5 dBW
Reference rate	R_0	1 b/s/Hz
Reference power	P_0	0 dBW
Noise power	σ_0^2	-143 dBW
Fronthaul power	P_i^{FH}	0 dBW
Maximum fronthaul capacity	$C_i = C, \forall i$	500 b/s/Hz
Fronthaul capacity factor	$\xi_i, \forall i$	10
Reweighted parameter	τ_1, τ_2	10^{-3}

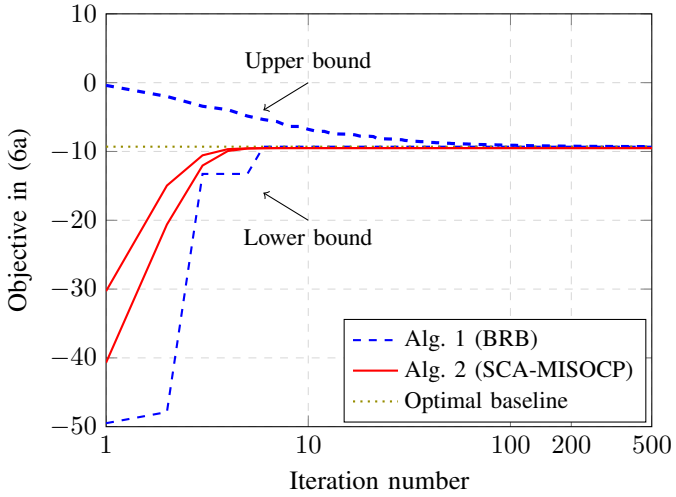
approximately $\mathcal{O}\left(K^4M^3I\right)$, resulting the overall complexity of $\mathcal{O}\left(2(I-1)K\left(K^4M^3I\right)\right)$ for Algorithm 3. In Section V, the numerical results show that Algorithm 3 yields a performance very close to that of Algorithm 2 but with much lower computation time. Finally, for Algorithm 4, the worst-case complexity is given by $\mathcal{O}\left(K^4M^3I\right)$.

V. Numerical Results

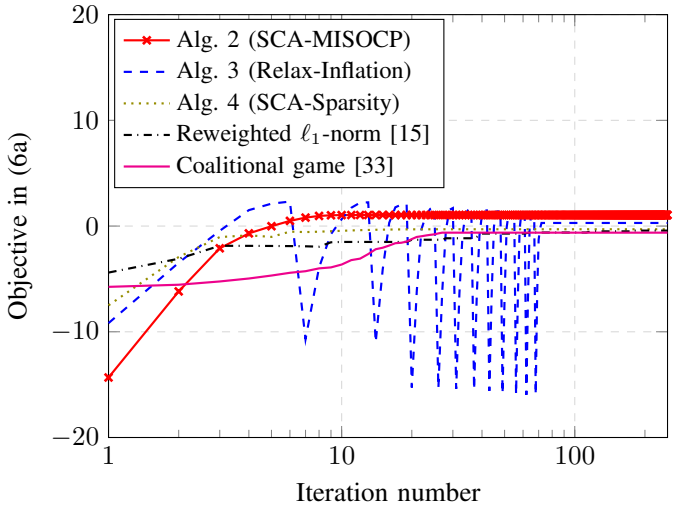
In this section, we numerically evaluate the performance of the proposed algorithms. For most numerical experiments, we use the simulation parameters listed in Table I. In particular, the parameters in the RRH and fronthaul power consumption model are taken from [21]. For the spatial model, we assume a network consisting of I RRHs that are uniformly located around the considered coverage and K UEs are randomly scattered across the considered network coverage. Moreover, we assume Rayleigh fading channel and the pathloss component is calculated as $(d_{ik}/d_0)^{-3}$where d_{ik} is the distance between the i^{th} RRH and the k^{th} user and $d_0 = 100$ m is the reference distance. To simplify the notations, we also assume the fronthaul link capacity $C_i = C, \forall i \in \mathcal{I}$, can be achieved up to 10 Gbps over 20 MHz bandwidth, which is equivalent to 500 b/s/Hz. In our simulations, Algorithms 2, 3, and 4 are terminated when the increase in the objective between two consecutive iterations is less than 10^{-5}.

In Fig. 2(a), we show the convergence of the objective function in (6) of Algorithms 1 and 2 for a set of random channel realizations. In this numerical experiment, we set $\alpha = 0.7$ and consider a small network setting with $I = 4$, $K = 3$. For Algorithm 2, we show the convergence behavior of the objective function with two different initial points $\mathbf{w}^{(0)}, \mathbf{z}^{(0)}, \boldsymbol{\gamma}^{(0)}$. As expected, Algorithm 1 requires much more iterations to update the upper and lower bounds, and thus converges after many iterations. On the other hand, Algorithm 2 converges much faster, just after a few iterations, and achieves the same objective value as return by the optimal algorithm despite the choice of initial points. This clearly demonstrates the effectiveness of Algorithm 2 which is used for benchmarking in the next experiments.

For onwards, we will consider the network setting as mentioned in Table I. In Figs. 2(b) and 3(a), we compare the convergence performance of our proposed low-complexity



(a)



(b)

Fig. 2: (a)-(b): Convergence behavior between different algorithms for a set of random channel realizations .

algorithms with the iterative WMMSE-reweighted ℓ_1-norm algorithm introduced in [15] and coalitional game based algorithm in [33] for $\alpha = 0.9$, in terms of both number of iterations and the overall runtime. Although only the sum rate maximization problem was studied in [15] and [33], we can easily modify their algorithm to solve the sum rate-power maximization problem considered in this paper. As can be clearly seen, our proposed solutions need a much smaller number of iterations to converge (possibly to different objectives), compared the reweighted ℓ_1-norm algorithm and coalitional game based algorithm. We note that in Fig. 2(b), the convergence of each SOCP during the inflation process is plotted, which explains the uphill and downhill effect in the figure. As can be seen, Algorithm 2 just takes a few iterations to stabilize. However its overall runtime is very high since the problem in each iteration is an MI-SOCP. On the other hand, Algorithms 3 and 4 require more iterations to converge but the per-iteration problem is an SOCP, which can be solved with

much computational effort. Thus their eventual computation time is much lower than that of Algorithm 2. Due to the fast converging property, Algorithms 3 and 4 outperform the reweighted ℓ_1-norm method, which is shown in Fig. 3(a). In Fig. 2(b), we can also see that the reweighted ℓ_1-norm and coalitional game based optimization methods converge to a smaller value, compared to our proposed solutions. This will be elaborated in the following experiments.

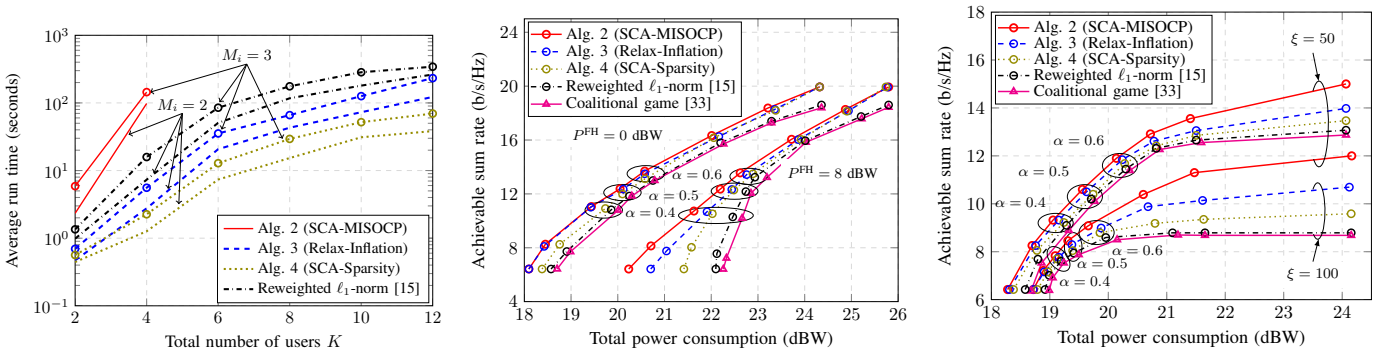
In Figs. 3(b) and 3(c), we study the trade-off between achievable sum rate (ASR) and total power consumption (TPC) by varying the parameter α over the interval $[0, 1]$ for the algorithms of comparison. The end points on bottom left of Figs. 3(b) and 3(c) show the smallest possible value of TPC without any consideration of the ASR (i.e., $\alpha = 0$). On the contrary, the end points on top right represent the largest possible ASR that can be obtained without any consideration of the TPC (i.e., $\alpha = 1$). As expected and shown in Figs. 3(b) and 3(c), when the TPC increases, so does the ASR. Moreover, it can be clearly seen that our proposed algorithms outperform the reweighted ℓ_1-norm and coalitional game based algorithm. Algorithm 2 is shown to attain the best performance among all the algorithms. Noticeably, the differences in the TPC between our proposed algorithms and the reweighted ℓ_1-norm algorithm as well as coalitional game based algorithm are significant. The reason is that, the method in [15] does not take into the fronthaul power while it becomes significant for large P^{FH} and the method in [33] does not consider the RRH selection in their coalition formation algorithm.

In Fig. 3(c), we investigate the trade-off between ASR and TPC for different values of fronthaul capacity factor $\xi_i = 50, 100, \forall i \in \mathcal{I}$. We first observe that there exists a "strong" trade-off between ASR and TPC in the high power regime when the fronthaul capacity is small. That is to say, a large amount of TPC is consumed just for a negligible improvement in the ASR. For high transmit power, the ASR over the wireless medium may be high but the small fronthaul capacity will act as a bottleneck. The practical guidance here is to avoid transmitting at full power for small fronthaul capacity to improve the network energy efficiency. Fig. 3(c) also demonstrates the increase of the ASR with the fronthaul capacity. To achieve the same ASR, more TPC is required for the networks with smaller fronthaul capacities. Intuitively, for smaller fronthaul capacities, the number of cooperative RRHs is reduced (cf. Table II for further insights). In fact, when each fronthaul capacity limit is small, common data shared by cooperative RRHs are limited to be transported via the fronthaul link to the RRHs, which in turn allows lesser number of cooperative RRHs. This results in more concurrent transmissions from the non-cooperative RRHs, which increases interference at each UE and subsequently leads to an increase in TPC to achieve the desired ASR. Furthermore, the results in Fig. 3(c) again show that the proposed algorithms achieve an improvement in the sum rate by up to 3.2 bits/s/Hz for the same TPC in the case $\xi_i = 100$, compared to the reweighted ℓ_1-norm and coalitional game based algorithms. In addition, the performance of coalitional game method is worse than other methods since RRHs are formed disjoint coalitions, thereby increasing the cell-edge interference and decreasing the ASR.

To gain more insights into the considered problem, we list the average number of active RRHs and number of RRH-UE associations versus the fronthaul power and fronthaul capacity in Table II. In this table, when $P^{\text{FH}} = 0, 8$ dBW, we choose $C = 500$ b/s/Hz and when $C = 60, 100$ b/s/Hz, we choose $P^{\text{FH}} = 0$ dBW. We can see that when the fronthaul power consumption increases, fewer RRH-UE associations are active and more RRHs are turned on. We observe that our proposed algorithms switch on only 50% of RRHs and 29.17% of user-RRHs associations to further reduce the total transmit power, while the referred algorithm switches on 66.67% of RRHs as well as RRH-UE associations.

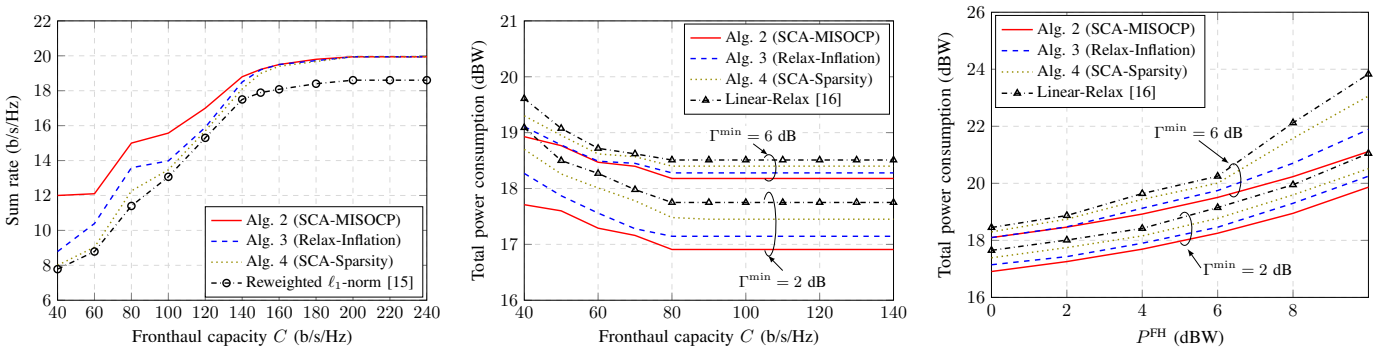
In Fig. 4(a), we compare the ASR (i.e., $\alpha = 1$) as a function of the fronthaul capacity C. As can be seen, when C increases, the ASR obtained by all the algorithms in comparison increase accordingly. This is because more data will be transported via the fronthaul links for large C, which then enables more cooperation among RRHs and improves the overall ASR of the system. However, the ASR become saturated at the high regime of fronthaul capacity, since the multi-user interference always exists even as more cooperation can be attained among all RRHs. For interference limited situations, there is an upper bound on the achievable rate for all users so that increasing more fronthaul capacity basically provides no benefit to the system performance. In Fig. 4(b) and 4(c), we show the total transmit power versus the fronthaul capacity and fronthaul power with $\Gamma_k^{\min} = 2$, 6 dB and $P^{\text{FH}} = 0$ dBW in case $\alpha = 0$. As expected, the TPC decreases and increases with the increase in C and P^{FH}, respectively. This can be explained by the fact that when C becomes higher, the number of UEs which are served by each RRH is larger, resulting in less power consumption in each RRH. Importantly, in Fig. 4(a), 4(b) and 4(c), our proposed algorithms achieve a better ASR and less TPC than that of algorithms in [15] and [16], respectively.

In Figs. 5(a) and 5(b), we compare the performance of the optimal solution attained by Algorithm 1 and the suboptimal solution by Algorithm 2 versus the required minimum SINR $\Gamma_k^{\min} = \Gamma^{\min}$, $\forall k \in \mathcal{K}$ with parameter $\alpha = 0.2$, 0.5, 0.7 and 0.9. Here, we consider a small network setting with $I = 4$, $K = 3$. As can be seen, when $\Gamma^{\min}$ increases, the ASR and TPC increase for all values of α. Particularly, the ASR and the TPC rapidly increase in the regime of low $\Gamma^{\min}$ and slightly increase in the high regime of $\Gamma^{\min}$. The increase of the ASR when $\Gamma^{\min}$ grows in Fig. 5(a) can be explained as follows. At low value of α, problem (6) has more priority to minimize TPC under the minimum rate constraint. Each user's rate achieved by solving the optimization (6) in this case is almost equal to the minimum rate, so that when $\Gamma^{\min}$ grows, the ASR increases proportionally. However, at high value of α, the problem of sum rate maximization is dominant. As a result, increasing $\Gamma^{\min}$ has less impact on the ASR performance. Similar explanation can be applied for the increase of the TPC at different α when $\Gamma^{\min}$ increases in Fig. 5(b). Moreover, in Fig. 5(c), the ASR is shown with three different values of the noise power $\sigma_0^2 = -143$, -140, and -130 dBW for $\alpha = 0.8$. It is obvious that when the noise power increases, the ASR decreases since the SINR of all users is eventually



(a) (b) (c)

Fig. 3: (a): The average run time comparison between different algorithms with number of antenna per RRH $M_i = 2, 3$, (b)-(c): Tradeoff between ASR and TPC by varying $\alpha \in [0, 1]$, some values of $\alpha = 0.4, 0.5, 0.6$ are marked with (b): $P^{\text{FH}} = 0$ and 8 dBW, (c): $\xi = 50$ and 100.



(a) $\alpha = 1$ (b) $\alpha = 0$ (c) $\alpha = 0$

Fig. 4: (a): Performance of ASR maximization problem, (b)-(c): TPC minimization problem.

reduced. Regarding the optimality of the proposed suboptimal solutions, it is shown numerically in Fig. 5(a), 5(b) and 5(c) that the suboptimal solution achieved by Algorithm 2 is very close to the optimal solution obtained by Algorithm 1. This again demonstrates the effectiveness of Algorithm 2.

In Fig. 6(a), we compare the objective of different algorithms for different values of α. We note that the variation of the objective in (6a) depends not only on α but also on the values of R_0 and P_0. For the chosen R_0 and P_0 stated in Table I, we observe that when α increases, the objective first decreases and then increases. From the results shown in Figs. 3(b) and 3(c), it is clear that when α increases, the TPC increases, and this make the objective decrease. However, after a certain point, the term $\alpha R^{\text{tot}}(\mathbf{w})/R_0$ will become dominant $(1-\alpha)P^{\text{tot}}(\mathbf{w}, \mathbf{a}, \mathbf{b})/P_0$ since the weight associated with power consumption is small, resulting in the objective increasing. Fig. 6(a) again demonstrates that our proposed algorithms outperform the reweighted ℓ_1-norm algorithm.

In the final numerical experiment, we consider a relatively large network setting with the number of RRHs $I = 60$ for the number of UEs $K = 50$ and $K = 60$. In Fig. 6(b), the trade-off between the ASR and TPC of low-complexity algorithms is plotted by varying $\alpha \in [0, 1]$. As shown in Fig. 6(b), the ASR and TPC increase when the weight associated with sum achievable rate (i.e., α) increases. The reasons is that in this case, the objective is in favor of sum rate maximization rather than power consumption minimization. This leads to more power consumption needed to obtain the higher ASR. Moreover, when the number of UEs increases, so do the ASR and TPC. It can be clearly explained that for higher number of UEs, more RRHs should be active to provide sufficient degree of freedom, leading to an increase in the TPC and also ASR. Again, Fig. 6(b) shows that Algorithms 3 and 4 attain a better performance compared to reweighted ℓ_1-norm algorithm. This demonstrates the effectiveness of our proposed framework.

TABLE II: Average number of active RRH-UE associations (Avr.RRH-UE) and active RRHs (Avr.RRHs).

P^{FH}/C		Alg. 2	Alg. 3	Alg. 4	[15]
$P^{\text{FH}} = 0$ dBW	Avr.RRH-UE	0.4167	0.4167	0.5	0.5
	Avr.RRHs	0.5	0.5	0.5	0.5
$P^{\text{FH}} = 8$ dBW	Avr.RRH-UE	0.2917	0.4167	0.5	0.6667
	Avr.RRHs	0.5	0.5	0.6667	0.6667
$C = 60$ b/s/Hz	Avr.RRH-UE	0.375	0.4583	0.5	0.5
	Avr.RRHs	0.5	0.6667	0.6667	0.6667
$C = 100$ b/s/Hz	Avr.RRH-UE	0.4583	0.4583	0.5	0.5
	Avr.RRHs	0.5	0.5	0.5	0.5

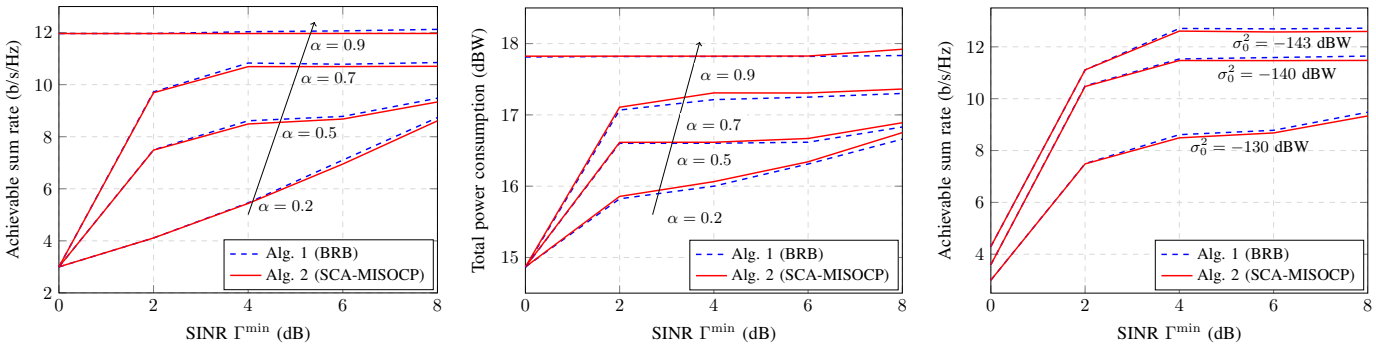


Fig. 5: (a)-(b): ASR and TPC versus required SINR $\Gamma^{\min}$ with parameter $\alpha = 0.2$, 0.5, 0.7 and 0.9;(c): ASR versus required SINR $\Gamma^{\min}$ with some different values of parameter $\sigma_0^2 = -143$, -140, -130 dBW.

VI. Conclusion

In this paper, joint beamforming, RRHs selection and RRH-UE association design has been proposed to maximize achievable sum rate and minimize total power consumption in the DL of C-RAN with limited capacity fronthaul links. In order to solve this multi-objective optimization problem, we have employed the scalarization method to form a scalar weighted sum objective function. Then, by novel transformations, we have transformed the combinatorial optimization problem into a more tractable form based on which a BRB algorithm has been customized to find an optimal solution. Further, by presenting novel methods to approximate the formulated non-convex problem, we have developed three more practically appealing and computationally efficient algorithms with much low complexity compared to the benchmarking scheme. By extensive numerical results, these algorithms have been shown to achieve a good convergence rate under various simulation settings and obtain a performance close to that of the optimal algorithm. Moreover, the proposed algorithms are also superior to other known algorithms.

Appendix A
Proof of Lemma 1

We prove that the constraints in (8b) and (8d) of problem (8) are active at optimality by contradiction. Let $(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{w}^\star, \boldsymbol{\nu}^\star, \mathbf{u}^\star)$ denote an optimal solution of (8). By contradiction, suppose that (8d) is inactive, i.e., $P^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star) < 1/u_0^\star$. Then there exists u_0' such that $u_0' > u_0^\star$ and $P^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star) < 1/u_0'$. That is, u_0' is feasible to (8) but yields a strictly larger objective, which contradicts with the fact that $(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{w}^\star, \boldsymbol{\nu}^\star, \mathbf{u}^\star)$ is an optimal solution. Similarly, assume that $R_k(\mathbf{w}^\star) > u_k^\star$ for some k. We then create a new set of beamformers as $\mathbf{w}' = [\mathbf{w}_1'^T, \mathbf{w}_2'^T, \ldots, \mathbf{w}_k'^T]^T$ where

$$\mathbf{w}_i' = \begin{cases} \mathbf{w}_i^\star & i \neq k \\ \zeta \mathbf{w}_k^\star & i = k \end{cases} \tag{38}$$

for some $0 < \zeta < 1$. Intuitively, the beamforming vector of user k is scaled down by a factor of ζ and the beamforming vectors of other users remain the same. From (3), it is easy to see that there exists $\zeta \in (0,1)$ such that $R_k(\mathbf{w}') > u_k^\star$ for all k. Note that $\|\mathbf{w}'\|_2 < \|\mathbf{w}\|_2$ and thus $P^{\text{tot}}(\mathbf{w}', \mathbf{a}^\star, \mathbf{b}^\star) < P^{\text{tot}}(\mathbf{w}^\star, \mathbf{a}^\star, \mathbf{b}^\star) \leq 1/u_0^\star$. Consequently, we find u_0' such that $u_0' > u_0^\star$ and $P^{\text{tot}}(\mathbf{w}', \mathbf{a}^\star, \mathbf{b}^\star) \leq 1/u_0'$, meaning that a strictly larger objective can be obtained. Again, this contradicts with the fact that $(\mathbf{b}^\star, \mathbf{a}^\star, \mathbf{w}^\star, \boldsymbol{\nu}^\star, \mathbf{u}^\star)$ is an optimal solution, and thus completes the proof.

Appendix B
Proof of (24)

We first show that the gradient of the function $g(x) = -\log(1+x)$ for $x \geq 0$ is Lipschitz continuous with parameter $L = 1$. This can be easily proved since

$$\begin{aligned} \|\nabla g(x_1) - \nabla g(x_2)\|_2 &= \left| -\frac{1}{1+x_1} + \frac{1}{1+x_2} \right| \\ &= \left| \frac{x_1 - x_2}{(1+x_1)(1+x_2)} \right| \overset{(a)}{\leq} |x_1 - x_2| \end{aligned} \tag{39}$$

where (a) is due to $(1+x_1)(1+x_2) > 1$ for $x_1, x_2 > 0$. Due to the Lipschitz continuity of $\nabla g(x)$, it holds that [32]

$$g(\gamma_k) \leq g(\gamma_k) + \nabla g(\gamma_k^{(n)}) \left(\gamma_k - \gamma_k^{(n)}\right) + \frac{1}{2\lambda} \left(\gamma_k - \gamma_k^{(n)}\right)^2 \tag{40}$$

for $\lambda \in (0, 1]$, and thus completes the proof by noting that (40) is actually (24) when $\lambda = 1$.

References

[1] P. Luong *et al.*, "Joint beamforming and remote radio head selection in limited fronthaul C-RAN," in *Proc. IEEE Veh. Technol. Conf. (VTC'16 Fall)*, Montreal, Canada, Sep. 2016, pp. 1–5.

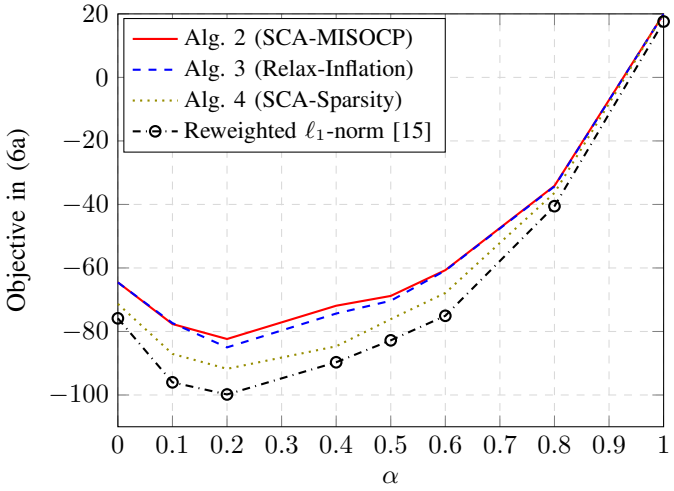
[2] P. Rost *et al.*, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[3] J. Wu et al., "Cloud radio access network (C-RAN): A primer," *IEEE Network*, vol. 29, no. 1, pp. 35–41, Jan. 2015.

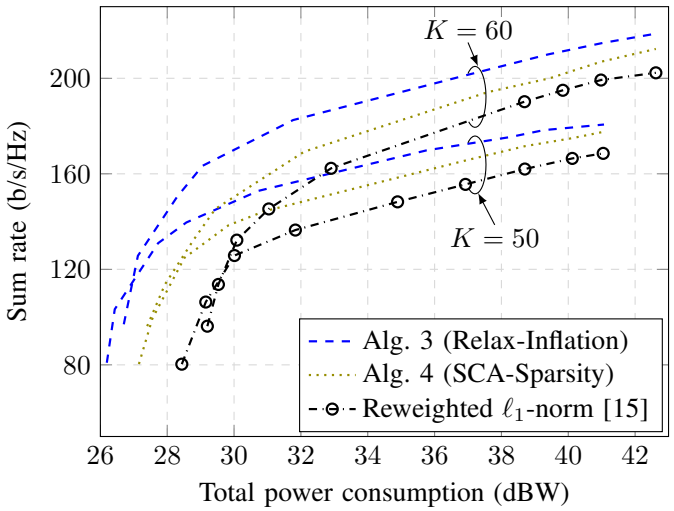
[4] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun. Mag.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[5] C. J. Bernardos et al., "Challenges of designing jointly the backhaul and radio access network in a cloud-based mobile network," in *Proc. IEEE Future Netw. and Mobile Summit*, Lisboa, 2013, pp. 1–10.

[6] S.-H. Park et al., "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.



(a)



(b)

Fig. 6: (a): Objective in (6a) versus parameter α, (b): Trade-off curves with $K = 50$ and $K = 60$.

[7] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access networkk," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, Jun. 2014.

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

[9] 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.

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

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

[12] R. Ramamonjison et al., "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.

[13] 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. 2015.

[14] B. Niu, Y. Zhou, H. Shah-Mansouri, and V. W. S. Wong, "A dynamic resource sharing mechanism for cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8325–8338, Dec. 2016.

[15] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 31, no. 2, pp. 1326–1339, Oct. 2014.

[16] V. N. Ha, L. B. Le, and N.-D. Dao, "Coordinated multipoint (CoMP) transmission design for cloud-RANs with limited fronthaul capacity constraints," *IEEE Trans. Veh. Technol.*, 2015.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[18] Y. Wu *et al.*, "Green transmission technologies for balancing the energy efficiency and spectrum efficiency trade-off," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 112–120, Nov. 2014.

[19] Q.-D. Vu, L.-N. Tran, R. Farrell, and E.-K. Hong, "Energy-efficient zero-forcing precoding design for small-cell networks," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 790–804, Feb. 2016.

[20] K. B. S. Manosha, M. Codreanu, N. Rajatheva, and M. Latva-aho, "Power–throughput tradeoff in MIMO heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4309–4322, Aug. 2014.

[21] Y. Cheng et al., "Joint network optimization and downlink beamforming for CoMP transmission using mixed integer conic programming," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3972–3987, Aug. 2013.

[22] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. PP, no. 99, pp. 1–14, Mar. 2016.

[23] D. Pompili, A. Hajisami, and T. X. Tran, "Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN," *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 26–32, Jan. 2016.

[24] 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.

[25] M. Peng *et al.*, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, Thridquater 2016.

[26] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and Multidisciplinary Optim.*, vol. 26, no. 6, pp. 369–395, Apr. 2004.

[27] D. W. K. Ng, E. S. Lo, and R. Schober, "Multiobjective resource allocation for secure communication in cognitive radio networks with wireless information and power transfer," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3166–3184, May 2016.

[28] O. Tervo, L.-N. Tran, and M. Juntti, "Optimal energy-efficient transmit beamforming for multi-user MISO downlink," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5574–5587, Oct. 2015.

[29] A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.

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

[31] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[32] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in optimization*, vol. 1, no. 3, Nov. 2014.

[33] Y. Sun, T. Dang and J. Zhou, "User scheduling and cluster formation in Fog Computing based Radio access networks," in *Proceedings of ICUWB*, Nanjing, China, Dec. 2016, pp. 1-4.

[34] M. Peng et al., "Energy-Efficient resource assignment and power allocation in Heterogeneous Cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275-5287, Nov. 2015.

[35] K. Guo, M. Sheng, J. Tang, T. Q.S. Quek, and Z. Qiu, "Exploiting Hybrid Clustering and Computation Provisioning for Green C-RAN," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 4063-4076, Dec. 2016.

[36] Z. Zhao et al., "Joint Design of Iterative Training-Based Channel Estimation and Cluster Formation in Cloud-Radio Access Networks," *IEEE Access*, vol. 4, pp. 9643–9658, Oct. 2016.

[37] J. Li et al., "Energy-Efficient Joint Congestion Control and Resource Optimization in Heterogeneous Cloud Radio Access Networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9873-9887, Dec. 2016.

[38] V. K. N. Lau, F. Zhang, and Y. CuiLow, "Complexity Delay-Constrained Beamforming for Multi-User MIMO Systems With Imperfect CSIT," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4090-4099, Aug. 2013.

[39] M. Peng et al., "Contract-Based Interference Coordination in Heterogeneous Cloud Radio Access Networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1140-1153, June 2015.

[40] M. Peng, Y. Yu, H. Xiang, and H. V. Poor, "Energy-Efficient Resource Allocation Optimization for Multimedia Heterogeneous Cloud Radio Access Networks," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 879-892, May 2016.

[41] H. Tuy and F. A. khayya-and P. Thach, "Monotonic optimization: Branch and cut methods," *Essays and Surveys in Global Optimization*, pp. 39–78, C. Audet, P. Hansen, and G. Savard, Eds. , Srpinger US 2005.

[42] A. Beck *et al.*, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Opti.*, vol. 47, no. 1, pp. 29–51, May 2010.

Phuong Luong (S'15) received the B.Eng. degree in telecommunications and electrical engineering from the Hanoi University of Science and Technology, Vietnam, in 2009, and the M.E. degree in the Department of Computer Science from Kyung Hee University, Yongin, South Korea, in 2012. She is currently working towards the Ph.D. degree with the École de Technologie Supérieure (ÉTS), Montréal, QC, Canada. Her current research interests include radio resource management in cloud radio access networks (C-RAN), cloud computing, network virtualization, green communications.

François Gagnon holds a Bachelor of Engineering degree and a Doctorate in Electrical Engineering from the École Polytechnique de Montréal, and has been a professor in the Department of Electrical Engineering at the École de technologie Supérieure (ÉTS) since 1991. He served as director of this department from 1999 to 2001. He has held industrial research chairs since 2001. In addition to holding the Richard J. Marceau Industrial Research Chair for Wireless Internet in developing countries, François Gagnon also holds the NSERC-Ultra Electronics Chair in Wireless Emergency and Tactical Communication, the most prestigious industrial chair program in Canada. He also founded the Communications and Microelectronic Integration Laboratory (LACIME) and was its first director. He has been very involved in the creation of the new generation of high-capacity line-of-sight military radios offered by the Canadian Marconi Corporation, which is now Ultra Electronics Tactical Communication Systems. Ultra-Electronics TCS and ÉTS have obtained the NSERC Synergy prize for this collaboration. Professor Gagnon serves on the boards of funding agencies and companies, he specializes in wireless communications, modulation, coding, microelectronics, signal processing, equalization, software defined radio, mobile communication and fading channels. He is actively involved in the SmartLand project of UTPL, Ecuador, the STARACOM strategic research network and the Réseau Québec Maritime.

Charles Despins received the bachelor's degree in electrical engineering from McGill University, Montreal, QC, Canada, in 1984, and the master's and Ph.D. degrees from Carleton University, Ottawa, ON, Canada, in 1987 and 1991, respectively. He was with CAE Electronics as a member of the Technical Staff from 1984 to 1985, the Department of Electrical and Computer Engineering, École Polytechnique de Montréal, Canada, as a Lecturer and a Research Engineer from 1991 to 1992, and a Faculty Member with the Institut National de la Recherche Scientifique, Université du Québec, Montréal, from 1992 to 1996. From 1996 to 1998, he was with Microcell Telecommunications Inc., a Canadian GSM operator, and was responsible for industry standard and operator working groups, as well as for technology trials and technical support for joint venture deployments in China and India. From 1998 to 2003, he was Vice President and Chief Technology Officer of Bell Nordiq Group Inc., a wireless and wireline network operator in northern and rural areas of Canada. From 2003 to 2016, he was the President and CEO of Prompt Inc., a university-industry research consortium in the field of information and communications technologies. He is currently Dean of Research and a Faculty Member at the École de Technologie Supérieure (Université du Québec), with research interests in wireless communications. He is also a Guest Lecturer in the M.B.A. program with McGill University, Montreal. He received the IEEE Vehicular Technology Society Best Paper of the Year prize in 1993, and the Outstanding Engineer Award in 2006 from the IEEE Canada. He is a member of the Order of Engineers of Québec and is also a Fellow (2005) of the Engineering Institute of Canada.

Le-Nam Tran (M'10–SM'17) received the B.S. degree in electrical engineering from Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2003 and the M.S. and Ph.D. degrees in radio engineering from Kyung Hee University, Seoul, Korea, in 2006 and 2009, respectively. He is currently a Lecturer/Assistant Professor at the School of Electrical and Electronic Engineering, University College Dublin, Ireland. Prior to this, he was a Lecturer at the Department of Electronic Engineering, Maynooth University, Co. Kildare, Ireland. From 2010 to 2014, he had held postdoc positions at the Signal Processing Laboratory, ACCESS Linnaeus Centre, KTH Royal Institute of Technology, Stockholm, Sweden (2010-2011), and at Centre for Wireless Communications and the Department of Communications Engineering, University of Oulu, Finland (2011-2014). His research interests are mainly on applications of optimization techniques on wireless communications design. Some recent particular topics include energy-efficient communications, cloud radio access networks, massive MIMO, and full-duplex transmission. He has authored or co-authored in some 70 papers published in international journals and conference proceedings.

Dr. Tran is an Associate Editor of EURASIP Journal on Wireless Communications and Networking. He was Symposium Co-Chair of Cognitive Computing and Networking Symposium of International Conference on Computing, Networking and Communication (ICNC 2016).