Energy Efficiency Maximization for C-RANs: Discrete Monotonic Optimization, Penalty, and ℓ_0-Approximation Methods

Kien-Giang Nguyen, *Student Member, IEEE*, Quang-Doanh Vu, *Member, IEEE*, Markku Juntti, *Senior Member, IEEE*, and Le-Nam Tran, *Senior Member, IEEE*

***Abstract*—We study downlink of multiantenna cloud radio access networks (C-RANs) with finite-capacity fronthaul links. The aim is to propose joint designs of beamforming and remote radio head (RRH)-user association, subject to constraints on users' quality-of-service, limited capacity of fronthaul links and transmit power, to maximize the system energy efficiency. To cope with the limited-capacity fronthaul we consider the problem of RRH-user association to select a subset of users that can be served by each RRH. Moreover, different to the conventional power consumption models, we take into account the dependence of baseband signal processing power on the data rate, as well as the dynamics of the efficiency of power amplifiers. The considered problem leads to a mixed binary integer program (MBIP) which is difficult to solve. Our first contribution is to derive a globally optimal solution for the considered problem by customizing a discrete branch-reduce-and-bound (DBRB) approach. Since the global optimization method requires a high computational effort, we further propose two suboptimal solutions able to achieve the near optimal performance but with much reduced complexity. To this end, we transform the design problem into continuous (but inherently nonconvex) programs by two approaches: penalty and ℓ_0-approximation methods. These resulting continuous nonconvex problems are then solved by the successive convex approximation framework. Numerical results are provided to evaluate the effectiveness of the proposed approaches.**

***Index Terms*—Energy efficiency, cloud radio access network, limited fronthaul capacity, rate-dependent signal processing power, nonlinear power amplifier, beamforming, mixed binary integer program, discrete branch-reduce-and-bound, successive convex approximation.**

I. Introduction

Coordinated multipoint joint transmission (CoMP-JT) [1] has been proposed in the current LTE standards to deal with the inter-cell interference, which is one of the key factors limiting the capacity of modern wireless communications systems. The central ideal of CoMP-JT is to allow for joint processing of the data symbols by multiple transmitters, thereby exploiting the cooperative gains efficiently. Thus, CoMP-JT is expected to improve the system performance significantly, especially for the cell-edge users. However, CoMP-JT requires a low-latency and high-capacity backhaul network, and a strict synchronization mechanism among transmitters [2]. These requirements are hard to implement in practice.

Cloud radio access networks (C-RANs) are emerging as a revolutionary solution that can deliver the same benefits as CoMP-JT [3], [4] but with less stringent synchronization requirements. In C-RANs, the baseband (BB) signal processing units are no longer installed at base stations (BSs) but relocated at a central cloud computing platform, which is referred to as BB unit (BBU) pool. Thus, BSs on C-RANs are solely responsible for wireless interface of the network, and now called remote radio heads (RRHs). By these particular features, C-RANs can potentially facilitate tight synchronization issue of BB signals required for CoMP-JT technique, and also leverage powerful computing capabilities for full cooperation [5]. However, BB signals from the BBU pool still need to be transported to the RRHs through the fronthaul links of limited capacity. In addition, the fronthaul links should support the strict latency and jitter requirements in order to perform the synchronization across the connected RRHs. Those are the main challenges of the C-RAN design in practice [5]–[7].

Due to the growing concern over the power consumption in existing mobile networks, recent research in wireless communications has shifted its focus on energy efficiency (EE) approaches [8]. In the past, wireless communications systems were mainly developed to maximize the spectral efficiency, i.e., with the aim to transmit at high data rates at any cost. This leads to a huge amount of power consumption on current wireless networks, since high data rate transmission essentially requires high transmit power. The notion of EE on the other hand, measured in bits/Joule, considers the data rate and total power consumption simultaneously.

C-RANs are a promising solution to address the problem of EE in future wireless networks, i.e., 5G and beyond. The potential gains of C-RANs on delivering the EE performance will be explored in this paper. Since the RRHs are controlled by the common BBU pool, they can be switched off to reduce the power consumption, thereby increasing the EE. In doing so, we also need to assign users to a proper set of serving RRHs. This should be done taking into account the limited

This work has been financially supported by Academy of Finland under the projects "Wireless Connectivity for Internet of Everything–Energy Efficient Transceiver and System Design (WiConIE)" (grant 297803), "Flexible Uplink-Downlink Resource Management for Energy and Spectral Efficiency Enhancing in Future Wireless Networks (FURMESFuN)" (grant 31089), and "6Genesis Flagship" (grant 318927). This publication has also emanated from research supported in part by a Grant from Science Foundation Ireland under Grant number 17/CDA/4786. A part of this paper was presented at 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Alberta, Canada, April 15-20, 2018.

K.-G. Nguyen, Q.-D. Vu, and M. Juntti are with the Centre for Wireless Communications, University of Oulu, P.O.Box 4500, FI-90014, Oulu, Finland. Email: {giang.nguyen, doanh.vu, markku.juntti}@oulu.fi.

L.-N. Tran is with School of Electrical and Electronic Engineering, University College Dublin, Ireland. Email: nam.tran@ucd.ie).

capacity of fronthaul links.

A. Related Works

The problem of EE maximization (EEmax) has been studied in prior publications [9]–[16] for different contexts. In the noise-limited scenarios, *parametric fractional programming* (PFP), i.e., Dinkelbach's algorithms were used to globally solve the EE power control problems with a linear (even super-linear) convergence [9]. In multiuser interference channels, Dinkelbach's algorithms cannot be applied to the EEmax problems here since Dinkelbach's assumptions are not met [9]. Thus, [14] and [15] resorted to using monotonic optimization in order to achieve optimal EE solution for multiple-input single-output (MISO) or single-input single-output (SISO) systems. As such global optimization methods require prohibitively high complexity, efficient suboptimal solutions were also of particular interest. Among them, the heuristic approaches developed based on PFP and the *successive convex approximation* (SCA) have been widely used in many wireless applications [10]–[16]. It is observed that the former approach often leads to a multi-stage iterative procedure [11], [12], and the convergence may not be guaranteed [9, Section 4.1]. On the other hand, the latter usually results in one-layer iterative procedures provably converging to stationary points with a small number of iterations [13], [14]. In fact, extensive numerical experiments conducted for some EEmax problems in multiuser MISO systems showed that the SCA-based methods outperform the heuristic PFP-based methods in terms of computational complexity [13], [14].

The aforementioned and other related studies assume that signal processing power is independent of the data rate. However, this is a simplification as different data rates require different modulation and coding schemes. In fact, signal processing power increases proportionally with the data rates [17]–[19], [20]. Moreover, the efficiency of the power amplifiers (PAs) is also assumed to be a constant in previous studies on EEmax [10]–[16], [21]. As shown in many works, PA's efficiency is often dynamic and it is degraded when operating in the back-off region of the maximum power [22]–[24]. Thus, it is practically relevant to investigate the impact of the dynamic of PA's efficiency and rate dependent power on EEmax designs.

C-RAN designs concerning limited fronthaul have been considered in some recent works [25]–[29]. A simple, but effective and widely used, method is to reduce the amount of BB signals exchanged through the fronthaul links. This is done by selecting a set of users that can be served by a RRH, giving rise to the RRH-user association problem that is often jointly designed with the transmit beamforming to optimize a network performance measure such as sum rate, power consumption or EE [25]–[29]. In the multicarrier transmission, jointly optimizing RRH selection and spectrum allocation would improve the network performances [30], [31]. The RRH-user association and RRH selection problems are usually modeled by a set of binary preference variables, leading to a *mixed binary integer program* (MBIP). As a result, optimal solutions to C-RANs with RRH-user associations and RRH selection are difficult to derive. On the other hand, in the EE perspective, there exists an approach of minimizing total power consumption for improving EE for C-RANs, e.g., [21], [26] for SISO or [32] for MISO. However, since the achieved data rate is not jointly considered in the objective, this approach may be far from the optimal [14], [16].

B. Contributions

We investigate the EEmax problem in C-RANs with capacity-limited fronthaul. Specifically, we propose a joint design of transmit beamforming and RRH-user association to maximize the network EE, while satisfying per-RRH fronthaul capacity, transmit power budget and users' quality-of-service (QoS). Towards a more realistic power consumption model, we account for the rate-dependent signal processing power and the dynamics of PA's efficiency. The considered problem is modeled as an MBIP. Our contributions include the following:

- We propose a globally optimal solution to the considered MBIP problem by customizing the discrete branch-reduce-and-bound (DBRB) framework introduced in [33]. To this end, we present transformations to reformulate the design problem into a form that is amendable to the application of the DBRB algorithm. Special modifications are made to improve the convergence performance of the proposed method.
- As global optimization methods are always of great concerns for an MBIP, we also propose two suboptimal solutions to the joint design problem that can achieve near-optimal solutions but with remarkably reduced complexity. In the first method, we use a set of continuous constraints to represent the binary variables, and then apply the penalty method to solve the resulting problem. In the second one, we approximate the binary variables by a piecewise linear function which is inspired by [34]. In both suboptimal methods, the obtained continuous problems are nonconvex, which are solved by the framework of the SCA.
- We provide extensive numerical results to justify the proposed solutions. The achievement of near-optimal performance by the proposed suboptimal methods is demonstrated by benchmarking against the optimal one. We compare the proposed solutions to other known methods in the literature. The impacts of rate-dependent power and dynamic PA's efficiency are also numerically investigated.

The rest of the paper is organized as follows. System model, design constraints, power consumption model and problem formulation are described in Section II. Section III presents the preliminaries of the DBRB framework in solving an MBIP, followed by the customization to solve the considered problem. Two suboptimal solutions are presented in Section IV. Numerical results are provided in Section V and Section VI concludes the paper.

Notation: We follow the standard notations in this paper. Lowercase letters, bold lowercase letters and bold uppercase letters represent the scalars, column (row) vectors and matrices, respectively. $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{C}$ represent the integer, real and complex domains, respectively. $(.)^T$ and $(.)^H$ represent

the transpose and Hermitian transpose operator, respectively. $\Re(.)$ and $|.|$ represent the real part and absolute value of a complex number, respectively. $\|.\|_2$ represents the ℓ_2 norm. The expectation of random variable is denoted as $\mathbb{E}[\cdot]$. $\{\mathbf{a}_b\}_b$ and $\{a_b\}_b$ refer to a set of vectors and scalars with different index b, respectively. $[\mathbf{a}]_i$ is the ith element of vector $\mathbf{a}$. $\mathbf{e}_i$ denotes the ith unit vector, i.e., the vector such that $e_i = 1$, $e_j = 0\ \forall j \neq i$. Finally, $\lceil a \rceil_{\mathcal{S}}$ and $\lfloor a \rfloor_{\mathcal{S}}$ are the upper and lower nearest neighbor elements of a in set $\mathcal{S}$.

II. System Model and Problem Formulation

A. System Model

We consider a multiuser MISO wireless system consisting of a set of B RRHs, denoted by $\mathcal{B} \triangleq \{1, \ldots, B\}$, each equipped with I antennas[1], and a set of K single-antenna users, denoted by $\mathcal{K} \triangleq \{1, \ldots, K\}$. The RRHs are connected to a common BBU pool through finite-capacity fronthaul links. The BBU pool is assumed to achieve perfect channel state information (CSI) associated with all the users in the network.[2] In this paper, CoMP-JT is considered, i.e., any user can simultaneously receive data from multiple RRHs [1]. Let d_k denote the data symbol intended for user k which has unit-energy, i.e., $\mathbb{E}[|d_k|^2] = 1$, and $\mathbf{w}_{b,k} \in \mathbb{C}^{I\times 1}$ denote the beamforming vector from RRH b to user k. Assuming a flat fading channel model, the received signal at user k can be written as

$$y_k = \underbrace{\left(\sum_{b\in\mathcal{B}} \mathbf{h}_{b,k}\mathbf{w}_{b,k}\right) d_k}_{\text{desired signal}} + \underbrace{\sum_{j\in\mathcal{K}\setminus k} \left(\sum_{b\in\mathcal{B}} \mathbf{h}_{b,k}\mathbf{w}_{b,j}\right) d_j}_{\text{interference}} + n_k \tag{1}$$

where $\mathbf{h}_{b,k} \in \mathbb{C}^{1\times I}$ is the channel between RRH b and user k, and $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ is the additive white Gaussian noise at user k. For notational convenience, let $\mathbf{h}_k \triangleq [\mathbf{h}_{1,k}, \mathbf{h}_{2,k}, \ldots, \mathbf{h}_{B,k}] \in \mathbb{C}^{1\times IB}$ and $\mathbf{w}_k \triangleq [\mathbf{w}_{1,k}^T, \mathbf{w}_{2,k}^T, \ldots, \mathbf{w}_{B,k}^T]^T \in \mathbb{C}^{IB\times 1}$ be the aggregate vectors of all channels and beamformers from all RRHs to user k, respectively. We also denote by $\mathbf{w}$ the beamforming vector stacking all $\mathbf{w}_k$. Assuming single-user decoding, i.e. interference among users is treated as Gaussian noise, the SINR at user k can be written as

$$\begin{aligned}\gamma_k(\mathbf{w}) \triangleq& \frac{|\sum_{b\in\mathcal{B}} \mathbf{h}_{b,k}\mathbf{w}_{b,k}|^2}{\sum_{j\in\mathcal{K}\setminus k} |\sum_{b\in\mathcal{B}} \mathbf{h}_{b,k}\mathbf{w}_{b,j}|^2 + \sigma_k^2} \\ =& \frac{|\mathbf{h}_k\mathbf{w}_k|^2}{\sum_{j\in\mathcal{K}\setminus k} |\mathbf{h}_k\mathbf{w}_j|^2 + \sigma_k^2}.\end{aligned} \tag{2}$$

Let r_k be the achievable data rate transmitted to user k. By the Shannon's coding theory, we have

$$r_k \leq \log(1 + \gamma_k(\mathbf{w})).$$

[1]Herein, the same number of equipped antennas for all RRHs is assumed purely for notational simplicity.

[2]From a practical perspective, overhead and accuracy of channel estimation should be considered, since they have major impacts on the scale of coordination and performance of C-RANs (see detail discussion in [6], [7], [27]). Also, there are some channel estimation techniques proposed for C-RANs which are summarized in [7, Section V].

B. Fronthaul Constraints

In practice the fronthaul link from the BBU pool to RRH b has a finite capacity, denoted by $\bar{C}_b$. To be feasible, the total data rate of the wireless physical layer of RRH b should not be larger than $\bar{C}_b$. For the problem formulation purposes, let us define $x_{b,k} \in \{0, 1\}$ to be the preference variable representing the connection between RRH b and user k, i.e., $x_{b,k} = 1$ indicates that user k receives data from RRH b and $x_{b,k} = 0$ otherwise. Then it is clear that the total data rate which can be reliably transmitted by the wireless interface of RRH b is $\sum_{k\in\mathcal{K}} x_{b,k} r_k$, and thus the following constraint

$$\sum_{k\in\mathcal{K}} x_{b,k} r_k \leq \bar{C}_b$$

should hold for RRH b.

C. Power Consumption Model

We consider the power consumption model based on those in [17], [24], [35], [36] which includes the power consumed by the electronic circuits in the network and the PAs on RRHs. Specifically, the circuit power consumption is divided into two parts as detailed below.

1) Rate-independent Circuit Power Consumption: The rate-independent power consumption is modeled as [35], [36]

$$\begin{aligned}P_{\text{I}} \triangleq\ & KP_{\text{ms}} + \sum_{b\in\mathcal{B}} s_b \underbrace{(P_{\text{RRH}}^{\text{active}} + P_{\text{NU}}^{\text{active}})}_{\text{active mode}} \\ & + \sum_{b\in\mathcal{B}} (1 - s_b) \underbrace{(P_{\text{RRH}}^{\text{sleep}} + P_{\text{NU}}^{\text{sleep}})}_{\text{sleep mode}} + P_{\text{OLT}}.\end{aligned} \tag{3}$$

In (3), P_{ms} is the circuit power consumed by a user device, $P_{\text{RRH}}^{\text{active}}$ and $P_{\text{RRH}}^{\text{sleep}}$ are the power consumption at a RRH corresponding to the active and sleep modes, respectively. In particular, $P_{\text{RRH}}^{\text{active}}$ consists of power for feeding signal processing circuits of transceiver chains, and operating RRHs (e.g. main supply, site-cooling) and hardware elements for RF parts (e.g. converters, filters, mixers, etc) [35]. It is assumed that all RRHs connect to the BBU pool through a passive optical network which consists of an optical line terminal (OLT) and a set of network units (NUs) [36]. The OLT is always active and consumes a fixed power, i.e. P_{OLT} in (3). On the other hand, NUs are switchable between the active and sleep modes for power saving purposes, each consuming a power $P_{\text{NU}}^{\text{active}}$ and $P_{\text{NU}}^{\text{sleep}}$, respectively. In order to represent the operating mode of RRH b and the associated NU, we introduce binary preference variables $\{s_b\}_b$ such that $s_b = 1$ when RRH and NU b is active and $s_b = 0$ otherwise. The relationship between s_b and $x_{b,k}$ (introduced in the previous subsection) can be represented as

$$s_b = \max_{k\in\mathcal{K}} \{x_{b,k}\} \Leftrightarrow \begin{cases} s_b \geq x_{b,k}, \forall k \in \mathcal{K} \\ s_b \leq \sum_{k\in\mathcal{K}} x_{b,k} \end{cases}, \forall b \in \mathcal{B} \tag{4}$$

i.e., $s_b = 1$ when RRH b serves at least one user and $s_b = 0$ otherwise.

2) Rate-dependent BB Signal Processing Power: The power consumed by the signal processing operations at the BBU pool such as channel encoding, decoding and fronthauling expenditure depends on the data rate [17]–[20]. For RRH b, this power consumption is measured by a continuous function of the fronthaul rate $\tilde{r}_b$ denoted as $\psi_b(\tilde{r}_b)$ where $\tilde{r}_b \triangleq \sum_{k\in\mathcal{K}} x_{b,k} r_k$. According to [17], [20], $\psi_b(\tilde{r}_b)$ is linearly scaled w.r.t. $\tilde{r}_b$, i.e.,

$$\psi_b(\tilde{r}_b) = p_{\text{SP}} \tilde{r}_b \tag{5}$$

where p_{SP} is a constant coefficient in W/(Gnats/s).

3) Dynamic Power Amplifier: Many existing approaches in relation to energy-efficient design assume a constant efficiency of PAs in their problem formulation [10]–[15]. However, in practice, the efficiency of PAs depends on their operating conditions, and thus is dynamic [22]–[24]. We can model the PA's efficiency of RF chain i at RRH b as [24]

$$\epsilon_{b,i}(\{\mathbf{w}_{b,k}\}_k) \triangleq \frac{1}{\tilde{\epsilon}} \sqrt{\sum_{k\in\mathcal{K}} |[\mathbf{w}_{b,k}]_i|^2} \tag{6}$$

where $\tilde{\epsilon} \triangleq \sqrt{P_{\text{a}}}/\epsilon_{\max}$, and P_{a} and $\epsilon_{\max} \in [0,1]$ are the maximum power of the PA and the maximum PA's efficiency, respectively. Let $\phi_b(\{\mathbf{w}_{b,k}\}_k)$ be a function of beamforming vectors which measures the amount of power consumed by the PAs for radiating the transmitted signals outwards the antennas at RRH b. From (6), $\phi_b(\{\mathbf{w}_{b,k}\}_k)$ is expressed as

$$\phi_b(\{\mathbf{w}_{b,k}\}_k) = \sum_{i=1}^{I} \frac{\sum_{k\in\mathcal{K}} |[\mathbf{w}_{b,k}]_i|^2}{\epsilon_{b,i}(\{\mathbf{w}_{b,k}\}_k)} = \tilde{\epsilon} \sum_{i=1}^{I} ||\tilde{\mathbf{w}}_{b,i}||_2 \tag{7}$$

where $\tilde{\mathbf{w}}_{b,i} \triangleq [[\mathbf{w}_{b,1}]_i; [\mathbf{w}_{b,2}]_i; ...; [\mathbf{w}_{b,K}]_i] \in \mathbb{C}^{K\times 1}$.

4) Total Power Consumption: For notational convenience, let us define $\mathbf{x} \triangleq \{x_{b,k}\}_{b\in\mathcal{B}k\in\mathcal{K}}$, $\mathbf{s} \triangleq \{s_b\}_{b\in\mathcal{B}}$, and $\mathbf{r} \triangleq \{r_k\}_{k\in\mathcal{K}}$. Based on the above discussions, the total consumed power in the considered system is denoted by $f_{\text{P}}(\mathbf{w},\mathbf{x},\mathbf{r},\mathbf{s})$ and can be expressed as

$$\begin{aligned} f_{\text{P}}(\mathbf{w},\mathbf{x},\mathbf{r},\mathbf{s}) &\triangleq P_{\text{I}} + \sum_{b\in\mathcal{B}} \left(\psi_b(\tilde{r}_b) + \phi_b(\{\mathbf{w}_{b,k}\}_k)\right) \\ &= \sum_{b\in\mathcal{B}} \Big(\tilde{\epsilon} \sum_{i=1}^{I} ||\tilde{\mathbf{w}}_{b,i}||_2 + \Delta P s_b + p_{\text{SP}} \sum_{k\in\mathcal{K}} x_{b,k} r_k \Big) \\ &\quad + \underbrace{B P^{\text{sleep}} + K P_{\text{ms}} + P_{\text{OLT}}}_{P_{\text{const}}} \end{aligned} \tag{8}$$

in which $P^{\text{active}} \triangleq P^{\text{active}}_{\text{RRH}} + P^{\text{active}}_{\text{NU}}$, $P^{\text{sleep}} \triangleq P^{\text{sleep}}_{\text{RRH}} + P^{\text{sleep}}_{\text{NU}}$ and $\Delta P \triangleq P^{\text{active}} - P^{\text{sleep}}$ which are constants.

D. Problem Formulation

We consider the problem of joint beamforming and RRH-user association design where the overall network EE is maximized. Mathematically, the problem of interest reads

$$\underset{\mathbf{w},\mathbf{x},\mathbf{s},\mathbf{r}}{\text{maximize}} \quad \frac{\sum_{k\in\mathcal{K}} r_k}{f_{\text{P}}(\mathbf{w},\mathbf{x},\mathbf{r},\mathbf{s})} \tag{9a}$$

$$\text{subject to} \quad r_k \le \log(1+\gamma_k(\mathbf{w})),\ \forall k \in \mathcal{K} \tag{9b}$$

$$r_k \ge r_0,\ \forall k \in \mathcal{K} \tag{9c}$$

$$\sum_{k\in\mathcal{K}} x_{b,k} r_k \le \bar{C}_b,\ \forall b \in \mathcal{B} \tag{9d}$$

$$\sum_{k\in\mathcal{K}} \|\mathbf{w}_{b,k}\|_2^2 \le \bar{P}_b,\ \forall b \in \mathcal{B} \tag{9e}$$

$$||\tilde{\mathbf{w}}_{b,i}||_2^2 \le P_{\text{a}},\ \forall b \in \mathcal{B}, i = 1,...,I \tag{9f}$$

$$\|\mathbf{w}_{b,k}\|_2^2 \le x_{b,k} \bar{P}_b,\ \forall k \in \mathcal{K}, b \in \mathcal{B} \tag{9g}$$

$$\sum_{b\in\mathcal{B}} x_{b,k} \ge 1, \forall k \in \mathcal{K} \tag{9h}$$

$$s_b \ge x_{b,k}, \forall k \in \mathcal{K};\ s_b \le \sum_{k\in\mathcal{K}} x_{b,k},\ \forall b \in \mathcal{B} \tag{9i}$$

$$\mathbf{x} \in \{0,1\}^{BK},\ \mathbf{s} \in \{0,1\}^{B}. \tag{9j}$$

We impose (9c) to guarantee that the data rate of user k is not smaller than r_0 to meet the required QoS. The constraints (9e) and (9f) represent the total transmit power and per antenna power constraints at each individual RRH, respectively. The constraints in (9g) guarantee that if RRH b does not serve user k, i.e. $x_{b,k} = 0$ then it holds that $\|\mathbf{w}_{b,k}\|_2^2 = 0$. The constraints in (9h) imply that each user is served by at least one RRH (due to the required QoS).

We remark that Dinkelbach's algorithm cannot be applied to find optimal solutions of (9), since (9a) is intractable [9, Section 3].[3] In fact, problem (9) is a nonconvex MBIP generally known to be NP-hard. In the following sections we first derive an optimal algorithm to solve (9) by customizing the DBRB framework, and then propose low-complexity suboptimal approaches that can achieve the near-optimal performance.

III. Optimal Jointly Energy-efficient Beamforming and RRH-user Association Design

General monotonic optimization (GMO) is a widely-used global continuous optimization technique [37] for solving numerous wireless communications nonconvex problems [14], [15], [38], [39]. For MBIP problems, the GMO principle is inapplicable, since it outputs only approximate solutions of discrete variables at convergence [37]. In recent work of [40], Luong *et al.* combined GMO with mixed integer programming (MIP) to solve their considered problem which is also an MBIP. Particularly, the GMO works on the continuous domain of their problem, and at each iteration of GMO, a mixed integer program is solved. In this paper, we propose below a new globally optimal approach to solve (9) based on the so-called *discrete monotonic optimization* (DMO) [33].

A. Preliminaries: Discrete Branch-reduce-and-bound

To proceed we provide some background of DMO and briefly review the DBRB procedure. In this paper we follow the definitions of *box*, *increasing function*, and *normal cone* in [33]. The standard form of a DMO problem is given by [33]

$$\max_{\mathbf{y}}\ f(\mathbf{y}) \text{ subject to } \{\mathbf{y} \in \mathcal{S} \subseteq D \triangleq [\mathbf{a};\mathbf{b}]\} \tag{10}$$

where $f(\mathbf{y})$ is an increasing function w.r.t. variable $\mathbf{y}$; $\mathbf{y} \triangleq [\mathbf{y}_{\text{d}}^T, \mathbf{y}_{\text{c}}^T]^T \in \mathbb{R}^{N_{\text{d}}+N_{\text{c}}}$, $\mathbf{y}_{\text{d}} \in \mathbb{Z}^{N_{\text{d}}}$ and $\mathbf{y}_{\text{c}} \in \mathbb{R}^{N_{\text{c}}}$ are the discrete

[3] Applying Dinkelbach's method to (9) results in the parametric subproblem (solved in each iteration) which is still nonconvex.

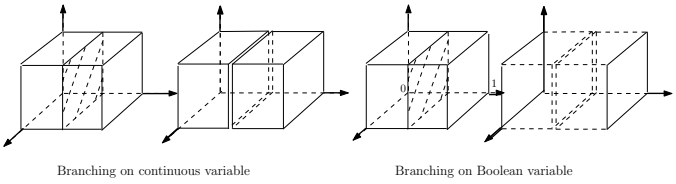


Fig. 1. Illustration for branching operator.

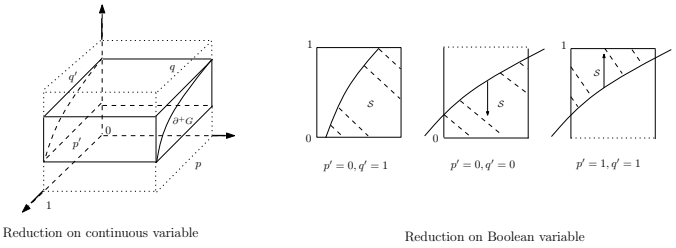


Fig. 2. Illustration for reduction operator.

and continuous variables respectively; $\mathcal{S}$ is normal feasible set of $\mathbf{y}$; and D is the box containing $\mathcal{S}$ with lower and upper vertices $\mathbf{a}$ and $\mathbf{b}$, respectively.

1) DBRB Procedure: Similar to the standard branch-reduce-and-bound (BRB) algorithm [37], DBRB is an iterative procedure performing three basic operations at each iteration: *branching*, *reduction,* and *bounding*. Starting from original box $[\mathbf{a};\mathbf{b}]$, we iteratively divide it into smaller and smaller ones, remove boxes that do not contain an optimal solution, search over remaining boxes for an improved solution until an error tolerance is met. Since the feasible set of the discrete optimization problem is smaller than that of its continuous relaxation, DBRB is modified from the standard BRB procedure in order to efficiently remove those regions not belonging to the discrete constraints, thereby achieving exact solutions [33]. In particular, during the branching and reduction steps, elements corresponding to discrete constraints are adjusted to stay in the discrete set. Details of these three operations are presented next.

Branching: At iteration n we select a box in the set of candidate boxes, denoted by $\mathcal{R}_n$, and split it into two new boxes, which are of equal size. To be bound improving we pick a box $V_{\rm c} \triangleq [\mathbf{p};\mathbf{q}] \in \mathcal{R}_n$, which has the largest upper bound, i.e., $V_{\rm c} = \arg\max_{V\in\mathcal{R}_n} f_{\rm U}(V)$ ($f_{\rm U}(V)$ denotes the upper bound of V), and bisect along the longest edge, i.e., $l = \arg\max_{1\le j\le N_{\rm d}+N_{\rm c}}(q_j - p_j)$ to create two smaller boxes $V_{\rm c}^1 = [\mathbf{p};\mathbf{q}']$ and $V_{\rm c}^2 = [\mathbf{p}';\mathbf{q}]$, in which $\mathbf{q}'$ and $\mathbf{p}'$ are given by

$$q'_j = \begin{cases} q_j & \forall j \neq l \\ \lfloor q_j - (q_j - p_j)/2 \rfloor_{\mathbb{Z}} & \text{if } j = l \le N_{\rm d}, \\ q_j - (q_j - p_j)/2 & \text{if } j = l > N_{\rm d}, \end{cases} \tag{11}$$

and

$$p'_j = \begin{cases} p_j & \forall j \neq l \\ \lceil p_j + (q_j - p_j)/2 \rceil_{\mathbb{Z}} & \text{if } j = l \le N_{\rm d}, \\ p_j + (q_j - p_j)/2 & \text{if } j = l > N_{\rm d}, \end{cases} \tag{12}$$

respectively.

Remark 1. (*Branching over Binary variables*) If $p_j, q_j \in \{0,1\}$ and $q_j - p_j = 1$ for $j \le N_{\rm d}$, then $\lfloor q_j - (q_j - p_j)/2 \rfloor_{\{0,1\}} = 0$ and $\lceil p_j + (q_j - p_j)/2 \rceil_{\{0,1\}} = 1$ (e.g. see Fig. 1).

Reduction: For any box, it possibly contains segments either infeasible to (10) or resulting in an objective smaller than the *current best objective* (CBO), i.e. the known feasible point that offers the best objective value at current iteration. Reduction is to remove those portions of no interest to reduce the search space in the next iterations. Given a box $V = [\mathbf{p};\mathbf{q}]$, we wish to shrink the size of V without loss of optimality by creating a smaller box $\mathrm{r}(V) \triangleq [\mathbf{p}';\mathbf{q}'] \subset V$ such that an optimal solution (if exists in V) must be contained in $\mathrm{r}(V)$. To do so we eliminate the portions $[\mathbf{p};\mathbf{p}')$ and $(\mathbf{q}';\mathbf{q}]$ that result in an objective value smaller than the CBO and/or are infeasible to (10). Mathematically, we can replace $\mathbf{p}$ by $\mathbf{p}' \ge \mathbf{p}$ where $\mathbf{p}' = \mathbf{q} - \sum_{j=1}^{N_{\rm d}+N_{\rm c}} \alpha_j (q_j - p_j)\mathbf{e}_j$ and

$$\alpha_j = \sup\{\alpha \mid 0 \le \alpha \le 1,\ \mathbf{q} - \alpha(q_j - p_j)\mathbf{e}_j \in D\backslash\mathcal{S}, \\ f(\mathbf{q} - \alpha(q_j - p_j)\mathbf{e}_j) \ge \text{CBO}\} \tag{13}$$

for each $j = 1, \ldots, N_{\rm d} + N_{\rm c}$. Similarly, vertex $\mathbf{q}$ is replaced by $\mathbf{q}' \le \mathbf{q}$ where $\mathbf{q}' = \mathbf{p}' + \sum_{j=1}^{N_{\rm d}+N_{\rm c}} \beta_j (q_j - p'_j)\mathbf{e}_j$ and

$$\beta_j = \sup\{\beta \mid 0 \le \beta \le 1,\ \mathbf{p}' + \beta(q_j - p'_j)\mathbf{e}_j \in \mathcal{S}\}. \tag{14}$$

The values of α_j and β_j in (13) and (14) can be found easily by the bisection method. Note that for $j \le N_{\rm d}$, the output of the reduction procedure is then adjusted into the discrete set, i.e., $p'_j = \lceil p'_j \rceil_{\mathbb{Z}}$ and $q'_j = \lfloor q'_j \rfloor_{\mathbb{Z}}$.

Remark 2. (*Reduction over Binary variables*) If $p_j, q_j \in \{0,1\}$ and $q_j - p_j = 1$ for $j \le N_{\rm d}$, we can quickly set that $p'_j = \begin{cases} 1 & \text{if } \mathbf{q} - \mathbf{e}_j \in D\backslash\mathcal{S} \\ 0 & \text{otherwise,} \end{cases}$. If $p'_j = 0$, we then replace $q_j - p'_j = 1$ into (14) and obtain $q'_j = \begin{cases} 1 & \text{if } \mathbf{p}' + \mathbf{e}_j \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$ (e.g. see Fig. 2).

The reduction procedure above does not drop off any feasible solution of (10) as shown in [33].

Bounding: Bounding is another basic operation for the DBRB to ensure the convergence. The main purpose of this step is to improve the upper and lower bounds of $f(\mathbf{y})$. Due to its monotonicity, the upper and lower bounds of a box $V = [\mathbf{p};\mathbf{q}]$ can be easily found as $f(\mathbf{p})$ and $f(\mathbf{q})$, respectively. These bounds are then used to update the CBO as mentioned above and to remove the boxes whose upper bound is smaller than the CBO [33].

We are now ready to customize the DBRB procedure to solve problem (9). Algorithm 1 outlines our proposed optimal method and its details are presented in the sequel.

B. Customization of DBRB for Solving (9)

We remark that (9) is not a DMO problem in a standard form, since the objective in (9a) is not an increasing function w.r.t. the involved variables. To apply the DBRB algorithm we first reformulate (9) as

$$\underset{\eta,\mathbf{w},\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t}}{\text{maximize}} \quad \eta \tag{15a}$$

Algorithm 1 The proposed DBRB algorithm

Initialization: Compute $\mathbf{a}$, $\mathbf{b}$ and apply box reduction to box $[\mathbf{a};\mathbf{b}]$. Let $n:=1$, $\mathcal{R}_1=\mathrm{r}([\mathbf{a};\mathbf{b}])$ and $\eta_1^{\text{best}}=0$
repeat $\{n:=n+1.\}$
  Branching: select a box $V_{\mathrm{c}}=[\mathbf{p};\mathbf{q}]\subset\mathcal{R}_{n-1}$ and branch V_{c} into two smaller ones V_{c}^1 and V_{c}^2, then remove V_{c} from $\mathcal{R}_{n-1}$.
  Reduction: apply box reduction to each box V_{c}^m ($m=\{1,2\}$) and obtain reduced box $\mathrm{r}(V_{\mathrm{c}}^m)$.
  Bounding: for each box $\mathrm{r}(V_{\mathrm{c}}^m)$ not violating (18)
  if solving (17) is feasible **then**
    Achieve $\mathbf{w}^*,\mathbf{u}^*$, calculate $\mathbf{t}^*$ and extract $\mathbf{x}^*$.
    Update $\underline{\mathbf{t}}:=\mathbf{t}^*$ and calculate $\eta_{\mathrm{U}}(\mathrm{r}(V_{\mathrm{c}}^m))$ by (20).
    Check $\mathbf{x}^*$ with (22), if true, obtain $\eta_{\mathrm{L}}(\mathrm{r}(V_{\mathrm{c}}^m))$ as (21) and update CBO $\eta_n^{\text{best}}:=\max\{\eta_{\mathrm{L}}(\mathrm{r}(V_{\mathrm{c}}^m)),\eta_{n-1}^{\text{best}}\}$, otherwise $\eta_{\mathrm{L}}(\mathrm{r}(V_{\mathrm{c}}^m))=\frac{\sum_{k\in\mathcal{K}}\underline{r}_k}{\hat{f}_{\mathrm{P}}(\overline{\mathbf{s}},\overline{\mathbf{x}},\overline{\mathbf{r}},\overline{\mathbf{t}})}$.
    Update $\mathcal{R}_n:=\mathcal{R}_{n-1}\cup\{\mathrm{r}(V_{\mathrm{c}}^m)|\eta_{\mathrm{U}}(\mathrm{r}(V_{\mathrm{c}}^m))\geq\eta_n^{\text{best}}\}$.
  end if
until Convergence
Output: With $(\eta_n^{\text{best}},\mathbf{x}^*,\mathbf{s}^*,\mathbf{r}^*,\mathbf{t}^*)$, recover $\mathbf{w}^*$ by (16) to achieve the globally optimal solution of (9), i.e. $(\mathbf{w}^*,\mathbf{x}^*,\mathbf{s}^*,\mathbf{r}^*)$.

$$\begin{aligned}\text{subject to}\quad & \eta\hat{f}_{\mathrm{P}}(\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t})-\textstyle\sum_{k\in\mathcal{K}}r_k\leq 0 && \text{(15b)}\\ & \textstyle\sum_{i=1}^{I}||\tilde{\mathbf{w}}_{b,i}||_2\leq t_b,\ \forall b\in\mathcal{B} && \text{(15c)}\\ & \text{(9b)}-\text{(9j)} && \text{(15d)}\end{aligned}$$

where η and $\mathbf{t}\triangleq\{t_b\}_b$ are newly introduced variables and $f_{\mathrm{P}}(\mathbf{w},\mathbf{x},\mathbf{s},\mathbf{r})$ is redefined as $\hat{f}_{\mathrm{P}}(\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t})\triangleq\sum_{b\in\mathcal{B}}\left(\tilde{\epsilon}t_b+\Delta P s_b+p_{\mathrm{SP}}\sum_{k=1}^{K}x_{b,k}r_k\right)+P_{\text{const}}$. The equivalence between (9) and (15) in terms of optimal solution set can be easily proved, since (15) is indeed the epigraph of (9). Towards solving (15) we have the following lemma.

Lemma 1. *Let $(\eta^*,\mathbf{w}^*,\mathbf{x}^*,\mathbf{s}^*,\mathbf{r}^*,\mathbf{t}^*)$ denote an optimal solution to* (15)*. Given the value of $(\mathbf{x}^*,\mathbf{s}^*,\mathbf{r}^*,\mathbf{t}^*)$, then the optimal beamforming vector, denoted by $\mathbf{w}^*$, can be computed as*

$$\mathbf{w}^*=\mathrm{find}\{\mathbf{w}|(9\mathrm{b}),(9\mathrm{e})-(9\mathrm{g}),(15\mathrm{c})\} \tag{16}$$

in which we replace $(\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t})$ by $(\mathbf{x}^*,\mathbf{s}^*,\mathbf{r}^*,\mathbf{t}^*)$.

Proof: See Appendix A. ∎

The lemma implies that we can obtain $\mathbf{w}^*$ if $(\mathbf{x}^*,\mathbf{s}^*,\mathbf{r}^*,\mathbf{t}^*)$ are known. We remark that η is easily determined when $(\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t})$ is fixed as $\eta=\frac{\sum_{k=1}^{K}r_k}{\hat{f}_{\mathrm{P}}(\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t})}$. Also, the feasibility of $\mathbf{r}$ depends on $\mathbf{t}$, $\mathbf{x}$ and $\mathbf{s}$ as can be seen in (9d) and (15b). Furthermore constraints (9d), (9h), (9i) and (15b) are monotone w.r.t. $\mathbf{x}$, $\mathbf{s}$, $\mathbf{r}$ and $\mathbf{t}$. Thus we can develop a DBRB algorithm to solve (15) by branching over $(\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t})$, which is the central idea of the proposed algorithm as described next.

Let $\mathcal{S}$ be the feasible set of problem (15), i.e.,

$$\begin{aligned}\mathcal{S}\triangleq\{[\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t}]|&(9\mathrm{b}),(9\mathrm{c}),(9\mathrm{f})-(9\mathrm{j}),(15\mathrm{b}),\\ &\sum_{k\in\mathcal{K}}x_{b,k}r_k\leq s_b\bar{C}_b,\sum_{k\in\mathcal{K}}\|\mathbf{w}_{b,k}\|_2^2\leq s_b\bar{P}_b,\\ &\sum_{i=1}^{I}||\tilde{\mathbf{w}}_{b,i}||_2\leq s_b t_b,\forall b\in\mathcal{B}\}.\end{aligned}$$

Remark that we have equivalently rewritten (9d), (9e) and (15c) by introducing s_b to the right hand side of these constraints so as to improve the proposed algorithm's efficiency. Specifically, if $s_b=0$ we can skip examining the constraints involving s_b. Because $\mathcal{S}$ is upper bounded by the power and fronthaul constraints, it satisfies the normal and finite properties required by a DBRB algorithm. Let $D=[\mathbf{a};\mathbf{b}]\in\mathbb{R}_+^{BK+2B+K}$ be the box such that $\mathcal{S}\subseteq D$, where the upper and lower vertices of D are defined as $\mathbf{a}\triangleq[\underline{\mathbf{x}},\underline{\mathbf{s}},\underline{\mathbf{r}},\underline{\mathbf{t}}]$ and $\mathbf{b}\triangleq[\overline{\mathbf{x}},\overline{\mathbf{s}},\overline{\mathbf{r}},\overline{\mathbf{t}}]$, respectively. Vertices in $\mathbf{a}$ and $\mathbf{b}$ are calculated as follows. It is obvious that $\underline{s}_b=0,\overline{s}_b=1,\underline{x}_{b,k}=0,\overline{x}_{b,k}=1$. We can immediately see that $r_k\geq\underline{r}_k=r_0$ due to (9b) and

$$\begin{aligned}r_k\leq\overline{r}_k&=\min\{\bar{C}_b,\log(1+|\mathbf{h}_k\mathbf{w}_k|^2/\sigma_k^2)\}\\ &\leq\min\{\bar{C}_b,\log(1+B\bar{P}_b\|\mathbf{h}_k\|_2^2/\sigma_k^2)\}\end{aligned}$$

as $|\mathbf{h}_k\mathbf{w}_k|^2\leq\|\mathbf{h}_k\|_2^2\|\mathbf{w}_k\|_2^2$ by the Cauchy-Schwarz inequality, and $\|\mathbf{w}_k\|_2^2\leq B\bar{P}_b$. We also have $t_b\geq\underline{t}_b=0$ and $t_b\leq\overline{t}_b=I\sqrt{P_{\mathrm{a}}}$.

As mentioned above, we can solve (15) by branching over $(\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t})$. Recall that branching and reduction for binary variables $\mathbf{x}$ and $\mathbf{s}$ follow Remarks 1 and 2. In bounding step, because the objective η is determined via $(\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t})$, the upper and lower bounds of η over a specific box $V=[\underline{\mathbf{x}},\underline{\mathbf{s}},\underline{\mathbf{r}},\underline{\mathbf{t}};\overline{\mathbf{x}},\overline{\mathbf{s}},\overline{\mathbf{r}},\overline{\mathbf{t}}]\subset\mathcal{R}_n$ can be simply calculated as $\eta_{\mathrm{L}}(V)\triangleq\frac{\sum_{k\in\mathcal{K}}\underline{r}_k}{\hat{f}_{\mathrm{P}}(\overline{\mathbf{x}},\overline{\mathbf{s}},\overline{\mathbf{r}},\overline{\mathbf{t}})}$ and $\eta_{\mathrm{U}}(V)\triangleq\frac{\sum_{k\in\mathcal{K}}\overline{r}_k}{\hat{f}_{\mathrm{P}}(\underline{\mathbf{x}},\underline{\mathbf{s}},\underline{\mathbf{r}},\underline{\mathbf{t}})}$. Note that we need to verify whether box V potentially contains a feasible beamforming solution to (15) before bounding. For the considered problem, we provide a better way of computing the lower and upper bounds, and checking the feasibility of candidate box V during the bounding process. In what follows, we present modifications (compared to the generic framework) made in Algorithm 1 to improve its efficiency.

Improved Branching: Normally each entry of $(\mathbf{x},\mathbf{s},\mathbf{r},\mathbf{t})$ is branched at each iteration, and thus the total number of iterations may increase quickly with the problem size. For (15), it turns out that we can skip branching on $\mathbf{t}$ while still guaranteeing the convergence. In particular, let us consider the following SOCP

$$\begin{aligned}\underset{\mathbf{w},\mathbf{u}}{\text{minimize}}\quad & \sum_{b\in\mathcal{B}}\sum_{i=1}^{I}u_{b,i} && \text{(17a)}\\ \text{subject to}\ & \mathbf{h}_k\mathbf{w}_k\geq\sqrt{(e^{\underline{r}_k}-1)(\textstyle\sum_{j\neq k}^{K}|\mathbf{h}_k\mathbf{w}_j|^2+\sigma^2)} && \text{(17b)}\\ & ||\tilde{\mathbf{w}}_{b,i}||_2\leq u_{b,i},\ \underline{s}_b\underline{t}_b\leq\sum_{i=1}^{I}u_{b,i}\leq\overline{s}_b\overline{t}_b,b\in\mathcal{B} && \text{(17c)}\\ & ||\tilde{\mathbf{w}}_{b,i}||_2^2\leq\overline{s}_bP_{\mathrm{a}},\ \|\mathbf{w}_{b,k}\|_2^2\leq\overline{x}_{bk}\bar{P}_b,b\in\mathcal{B} && \text{(17d)}\end{aligned}$$

$$\sum_{k\in\mathcal{K}} \|\mathbf{w}_{bk}\|_2^2 \le \overline{s}_b \bar{P}_b \ \forall b \in \mathcal{B} \tag{17e}$$

which can be viewed as minimizing the power consumption subject to minimum users' rate requirement $\underline{\mathbf{r}}$. Let us denote by $\mathbf{u}^*$ the optimal solution if (17) is feasible and $\mathbf{t}^* \triangleq \{t_b^*\}_b$ with $t_b^* = \sum_{i=1}^{I} u_{b,i}^*$. Obviously $\mathbf{t}^*$ is the minimum power required to achieve $\underline{\mathbf{r}}$, and it holds $\underline{\mathbf{t}} \le \mathbf{t}^*$. Also, t_b^* is unique solution because the objective in (17) is the epigraph of the function $\sum_{b\in\mathcal{B}}\sum_{i=1}^{I} ||\tilde{\mathbf{w}}_{b,i}||_2$ [41, Chapter 3]. At this point, we can replace $\underline{\mathbf{t}}$ by $\mathbf{t}^*$ to obtain a tighter lower bound on $\mathbf{t}$. Thus, it is sufficient to only branch $(\mathbf{x}, \mathbf{s}, \mathbf{r})$ as the lower bound on $\mathbf{t}$ is always improved with $\underline{\mathbf{r}}$. The property significantly accelerates the convergence of the proposed algorithm.

Improved Branching Order: Essentially, in each iteration of a DBRB algorithm we can randomly select a variable to perform branching. Exploiting the specifics of the considered problem, we can potentially reduce the computational complexity if we opt to branch $\mathbf{s}$ first due to its dependency on other factors. Intuitively, the number of active RRHs provides the degree-of-freedoms that can make the desired data rate $\underline{\mathbf{r}}$ achievable. Moreover, we can immediately obtain $x_{b,k} = 0, \forall k \in \mathcal{K}$ whenever $s_b = 0$, implying that the effective dimension in V is reduced by K times. Therefore by first keeping branching on $\mathbf{s}$ until $\underline{\mathbf{s}} = \overline{\mathbf{s}}$, we can quickly remove combinations of $\{s_b\}_b$ infeasible to (15). This is done by solving (17) with given $\overline{\mathbf{s}}$ and target rate r_0 for all users. Moreover, since the length of $\mathbf{s}$ is much smaller than that of $\mathbf{x}$ in most of wireless communications applications, branching on $\mathbf{s}$ may take a relatively small number of iterations.

Improved Memory Requirement: A DBRB algorithm basically stores a sequence of boxes until an optimal solution is found, which requires some memory capacity. To reduce this memory requirement we can eliminate boxes that contain no feasible solution. Recall that the feasible set of (15) is determined by the users' rate requirement, power and fronthaul constraints. It is easily seen that the rate and power feasibility of box V is equivalent to solving problem (17). For fronthaul constraints, we have the following feasibility condition, i.e., if the inequality below does not hold

$$\sum_{k\in\mathcal{K}} \underline{r}_k \le \sum_{b\in\mathcal{B}} \overline{s}_b \bar{C}_b \tag{18}$$

then V contains no feasible solution. In fact, (18) is due to $\sum_{b\in\mathcal{B}} \overline{s}_b \bar{C}_b \ge \sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}} x_{b,k} r_k \ge \sum_{k\in\mathcal{K}} r_k \sum_{b\in\mathcal{B}} x_{b,k} \ge \sum_{k\in\mathcal{K}} \underline{r}_k$ where the last inequality follows (9h). In Algorithm 1, we check (18) prior to (17) for saving computational efforts. We remark that the computational complexity of checking the feasibility of V is dominated by solving (17), and is independent of the dimension of binary variables.

Improved Bounds: Using monotonicity to compute bounds as mentioned above is inefficient for our considered problem. We now present a way to obtain tighter bounds which can improve the convergence rate of Algorithm 1 in practice. First recall that $\hat{f}_{\text{P}}(\underline{\mathbf{x}}, \underline{\mathbf{s}}, \underline{\mathbf{r}}, \underline{\mathbf{t}}) = \sum_{b\in\mathcal{B}}(\tilde{\epsilon}\underline{t}_b + \Delta P \underline{s}_b + p_{\text{SP}} \sum_{k=1}^{K} \underline{x}_{b,k}\underline{r}_k) + P_{\text{const}}$ and observe that the terms involving binary variables are zero if $\underline{s}_b = 0$ and $\underline{x}_{b,k} = 0$ for some b, k, whereas ΔP and p_{SP}, i.e., the power for operating RRHs and signal processing circuits are much larger than the power consumption on the PAs. Let us consider the following bound

$$\begin{aligned}\underline{\hat{f}}_{\text{P}}(\mathbf{x}, \mathbf{s}, \mathbf{r}, \mathbf{t}^*) \triangleq \sum_{b\in\mathcal{B}} \tilde{\epsilon} t_b^* + \Delta P \max\{1, \sum_{b\in\mathcal{B}} \underline{s}_b\} \\ + p_{\text{SP}} \max\{\sum_{k\in\mathcal{K}} \underline{r}_k, \sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}} \underline{x}_{b,k}\underline{r}_k\} + P_{\text{const}}\end{aligned} \tag{19}$$

in which the first term is a result of solving (17) (if feasible); the second term is due to the fact that at least one RRH is active for transmission; the third term is achieved by $\sum_{b\in\mathcal{B}}(\sum_{k\in\mathcal{K}} x_{b,k} r_k) \ge \sum_{k\in\mathcal{K}} r_k (\sum_{b\in\mathcal{B}} x_{b,k}) \ge \sum_{k\in\mathcal{K}} r_k$. Obviously, $\hat{f}_{\text{P}}(\underline{\mathbf{x}}, \underline{\mathbf{s}}, \underline{\mathbf{r}}, \underline{\mathbf{t}}) \le \underline{\hat{f}}_{\text{P}}(\mathbf{x}, \mathbf{s}, \mathbf{r}, \mathbf{t}^*)$ and replacing $\hat{f}_{\text{P}}(\underline{\mathbf{x}}, \underline{\mathbf{s}}, \underline{\mathbf{r}}, \underline{\mathbf{t}})$ by $\underline{\hat{f}}_{\text{P}}(\mathbf{x}, \mathbf{s}, \mathbf{r}, \mathbf{t}^*)$ does not remove any feasible solution. A tighter upper bound on η over V can be recalculated as

$$\eta_{\text{U}}(V) = \frac{\sum_{k\in\mathcal{K}} \overline{r}_k}{\underline{\hat{f}}_{\text{P}}(\mathbf{x}, \mathbf{s}, \mathbf{r}, \mathbf{t}^*)}. \tag{20}$$

Similarly, suppose $(\hat{\mathbf{x}}, \hat{\mathbf{s}}, \underline{\mathbf{r}}, \hat{\mathbf{t}})_V$ to be some feasible point within V. We can easily check that $\hat{f}_{\text{P}}(\hat{\mathbf{x}}, \hat{\mathbf{s}}, \underline{\mathbf{r}}, \hat{\mathbf{t}})_V \le \hat{f}_{\text{P}}(\overline{\mathbf{x}}, \overline{\mathbf{s}}, \overline{\mathbf{r}}, \overline{\mathbf{t}})$ due to the monotonicity property of $\hat{f}_{\text{P}}(\mathbf{x}, \mathbf{s}, \mathbf{r}, \mathbf{t})$. Then an improved lower bound on η over V can be obtained as

$$\eta_{\text{L}}(V) = \frac{\sum_{k\in\mathcal{K}} \underline{r}_k}{\hat{f}_{\text{P}}(\hat{\mathbf{x}}, \hat{\mathbf{s}}, \underline{\mathbf{r}}, \hat{\mathbf{t}})_V}. \tag{21}$$

Remark that if $\eta_{\text{L}}(V) \ge \eta_n^{\text{best}}$ where η_n^{best} denotes the CBO at iteration n, we can update $\eta_{\text{L}}(V)$ as the new CBO and then remove boxes whose upper bounds are smaller than η_n^{best} (see Step 10 in Algorithm 1). Thus, obtaining a feasible point is vital for improving the algorithm's efficiency. For this purpose we present in the following a heuristic way.

Heuristic Method for Finding a Feasible Solution : We propose a simple trick which may quickly find a feasible solution in V. It is worth noting that a feasible point $(\hat{\mathbf{x}}, \hat{\mathbf{s}}, \underline{\mathbf{r}}, \hat{\mathbf{t}})_V$ of problem (15) must satisfy two conditions: $\underline{\mathbf{r}}$ is achievable by $(\hat{\mathbf{x}}, \hat{\mathbf{s}}, \hat{\mathbf{t}})_V$; and

$$\hat{\mathbf{x}} \in \{\mathbf{x} \mid \sum_{b\in\mathcal{B}} x_{b,k} \ge 1, k \in \mathcal{K}, \sum_{k\in\mathcal{K}} x_{b,k}\underline{r}_k \le \bar{C}_b, b \in \mathcal{B}\}. \tag{22}$$

As can be easily seen, the feasible solution returned by solving (17) always satisfies the former condition. Thus, our idea is to extract $\hat{\mathbf{x}}$ from the optimal point of (17) and verify (22). Specifically, we can compute $\hat{\mathbf{x}}$ by setting $\hat{x}_{b,k} = 0$ if $\|\mathbf{w}_{b,k}^*\|_2 = 0$ and vice versa $\hat{x}_{b,k} = 1$ if $\|\mathbf{w}_{b,k}^*\|_2 > 0$ where $\mathbf{w}^*$ is an optimal solution obtained by solving (17).

Convergence Analysis of Algorithm 1

Algorithm 1 is guaranteed to yield a globally optimal solution of (9) which can be justified following the same arguments in the convergence analysis of generic DBRB [33]. Specifically, we first recall that the branching and reduction operations follow the same manner as in [33], [37]. These guarantee that the upper and lower bounds of η in each box are always improved after every iteration (branching rule), and that no feasible point in a box being lost (reduction operations) [33]. On the other hand, during the bounding step, it is easy to

check that the feasibility conditions (i.e., (17) and (18)) and the calculation of tighter upper and lower bounds (i.e., (20) and (21)) do not eliminate any feasible point, and the upper bound (20) (resp. lower bound (21)) is non-increasing (resp. non-decreasing). We note that the feasible set is upper bounded by the power and fronthaul constraints, and lower bounded by the users' QoS constraints. Therefore, following the proof of [33, Theorem 17], Algorithm 1 generates a sequence of boxes such that the gap between the upper bound and lower bound is guaranteed to converge to a single point, which is a globally optimal solution of (15). Recall that (9) and (15) is optimally equivalent, thus Algorithm 1 achieves globally optimal solution of (9).

IV. Suboptimal Designs

In general a global optimization algorithm often takes enormous complexity to output a solution. In this section, we propose two sub-optimal approaches that are more practically appealing.

A. Penalty Method

In the first method, a binary variable is equivalently represented by a set of continuous functions and then a penalty method is applied. Note that we can rewrite (9) as

$$\underset{\substack{\eta,t,\mathbf{w},\mathbf{s},\mathbf{x},\\ \mathbf{r},\mathbf{g},\mathbf{q},\boldsymbol{\vartheta}}}{\text{maximize}} \quad \eta \tag{23a}$$

$$\text{subject to } \eta t \le \sum_{k\in\mathcal{K}} r_k \tag{23b}$$

$$t \ge \tilde{f}_{\text{P}}(\mathbf{w},\mathbf{x},\mathbf{s},\boldsymbol{\vartheta}) \tag{23c}$$

$$\log(1+g_k) \ge r_k \ \forall k\in\mathcal{K} \tag{23d}$$

$$q_k \ge \|[\sigma_k,\ \{\mathbf{h}_k\mathbf{w}_j\}_{j\in\mathcal{K}\setminus k}]\|_2^2, \forall k\in\mathcal{K} \tag{23e}$$

$$q_k g_k \le |\mathbf{h}_k\mathbf{w}_k|^2 \ \forall k\in\mathcal{K} \tag{23f}$$

$$\sum_{k\in\mathcal{K}} x_{b,k} r_k \le \vartheta_b,\ \vartheta_b\in[0,\bar{C}_b],\ \forall b\in\mathcal{B} \tag{23g}$$

$$(9c), (9e)-(9j) \tag{23h}$$

where η, t, $\mathbf{g} \triangleq \{g_k\}_k$, $\mathbf{q} \triangleq \{q_k\}_k$, $\boldsymbol{\vartheta} \triangleq \{\vartheta_b\}_b$ are newly introduced slack variables, and $\tilde{f}_{\text{P}}(\mathbf{w},\mathbf{x},\mathbf{s},\boldsymbol{\vartheta}) \triangleq \sum_{b\in\mathcal{B}}(\sum_{i=1}^{I}\tilde{\epsilon}||\tilde{\mathbf{w}}_{b,i}||_2 + \Delta P s_b + p_{\text{SP}}\vartheta_b) + P_{\text{const}}$. We can further reformulate (23) as

$$\underset{\substack{\eta,t,\mathbf{w},\mathbf{s},\mathbf{x},\\ \mathbf{r},\mathbf{g},\mathbf{q},\boldsymbol{\vartheta}}}{\text{maximize}} \quad \eta \tag{24a}$$

$$\text{subject to } (\eta+t)^2 \le \|[\eta,\ t]\|_2^2 + 2\sum_{k\in\mathcal{K}} z_k \tag{24b}$$

$$(q_k+g_k)^2 \le \|q_k, g_k, \sqrt{2}\mathbf{h}_k\mathbf{w}_k]\|_2^2,\ k\in\mathcal{K} \tag{24c}$$

$$\sum_{k\in\mathcal{K}}(x_{b,k}+r_k)^2 \le \|[\{x_{b,k}\}_k,\{r_k\}_k]\|_2^2 + 2\vartheta_b \tag{24d}$$

$$(9c), (9e)-(9j), (23c)-(23e). \tag{24e}$$

Clearly (24) maintains the feasible set of (9). To invoke continuous optimization, we now represent binary variables $\mathbf{x}$ and $\mathbf{s}$ by a continuous constraint. To this end, we can use the well-known relaxation of binary variables which is given as [33, Section 1]

$$x_{b,k}\in\{0,1\}, \forall b,k \Leftrightarrow \sum_{b\in\mathcal{B},k\in\mathcal{K}} x_{b,k}^2 - x_{b,k} \ge 0,\ x_{b,k}\in[0,1]. \tag{25}$$

The above representation is justified by the fact that $x_{b,k}^2 - x_{b,k} < 0$ for $x_{b,k}\in(0,1)$. We note that s_b is automatically binary when $x_{b,k}$ is so, which is due to (9i). Thus we can simply relax $s_b\in[0,1]$ and equivalently rewrite (24) as

$$\max_{\boldsymbol{\Omega}\in\mathcal{S}_{\text{c}}\cap\mathcal{S}_{\text{nc}}} \ \eta \quad \text{subject to} \quad \{(25),\ s_b\in[0,1]\} \tag{26}$$

where $\boldsymbol{\Omega} \triangleq \{\eta,t,\mathbf{w},\mathbf{s},\mathbf{x},\mathbf{r},\mathbf{g},\mathbf{q},\boldsymbol{\vartheta}\}$ and

$$\mathcal{S}_{\text{c}} \triangleq \{\boldsymbol{\Omega}|(9c),(9e)-(9i),(23c)-(23e)\}$$
$$\mathcal{S}_{\text{nc}} \triangleq \{\boldsymbol{\Omega}|(24b)-(24d)\}$$

which are the set of convex and nonconvex constraints of (26), respectively. From this point onwards, $x_{b,k}$'s and s_b's are understood to be continuous over $[0,1]$. Now (26) is a continuous nonconvex problem, for which one can basically apply the SCA method to solve. However, finding an initial point of the iterative process is usually difficult. To overcome the issue, we apply a penalty method which results in the following regularized problem

$$\begin{aligned}\max_{\boldsymbol{\Omega}\in\mathcal{S}_{\text{c}}\cap\mathcal{S}_{\text{nc}}} \psi(\boldsymbol{\Omega},\alpha,\xi) \triangleq \eta &+ \alpha\sum_{b\in\mathcal{B},k\in\mathcal{K}}(x_{b,k}^2 - x_{b,k}) \\ &+ \xi\sum_{b\in\mathcal{B}}\min\{0,\bar{C}_b-\vartheta_b\}\end{aligned} \tag{27}$$

where $\alpha,\xi>0$ are the penalty parameters. Intuitively, the second term in $\psi(\boldsymbol{\Omega},\alpha,\xi)$ represents the cost when $x_{b,k}$'s are not binary, while the last term represents the cost when the fronthaul constraints are violated. Our expectation is that solving (27) will eventually produce binary solutions. In this regard we replace (9g) by

$$\|\mathbf{w}_{b,k}\|_2^2 \le x_{b,k}^q \bar{P}_b. \tag{28}$$

We can check that (28) is equivalent to (9g) for $x_{b,k}\in\{0,1\}$. To appreciate the above maneuver, let $\mathcal{F}_q$ denote the feasible set of (27) when (9g) is replaced by (28) and $\tilde{\eta}_q$ is the resulting optimal objective. For $x_{b,k}\in[0,1]$ it is clear that $x_{b,k}^q \ge x_{b,k}^{q+1}$ for any $q>0$, meaning

$$\mathcal{F}_{q+1}\subseteq\mathcal{F}_q\subseteq\cdots\subseteq\mathcal{F}_1 \triangleq \mathcal{S}_{\text{c}}\cap\mathcal{S}_{\text{nc}} \tag{29}$$

and thus

$$\tilde{\eta}^* \le \tilde{\eta}_{q+1} \le \tilde{\eta}_q \le \cdots \le \tilde{\eta}_1 \tag{30}$$

where $\tilde{\eta}^*$ is the optimal value of (27) for $x_{b,k}\in\{0,1\}$. The above inequality simply implies that a tighter continuous relaxation can be obtained with higher values of q. However we also note that (28) for $q>1$ is noncovex and thus it has not been used in the development of the proposed global optimization algorithm.

Now we can apply the SCA to solve (27). In the light of the SCA principle [42], the nonconvex constraints in $\mathcal{S}_{\text{nc}}$ and (28) can be approximated as

$$(\eta+t)^2 \le 2[\eta^n,t^n][\eta,t]^T - \|[\eta^n,t^n]\|_2^2 + 2\sum_{k\in\mathcal{K}} r_k \tag{31}$$

Algorithm 2 Proposed method for solving (23)

1: **Initialization:** Set $n := 0$, choose initial values for $\mathbf{\Omega}^0$ and set α^0 small
2: **repeat** $\{n := n + 1\}$
3: Solve (35) and achieve $\mathbf{\Omega}^*$
4: Update $\mathbf{\Omega}^n := \mathbf{\Omega}^*$
5: Update $\alpha^n := \min\{\alpha_{\max}; \alpha^{n-1} + \varepsilon\}$ for small ε
6: **until** Convergence

$$(q_k + g_k)^2 \leq 2\Re([q_k^n, g_k^n, \sqrt{2}\mathbf{h}_k\mathbf{w}_k^n][q_k, g_k, \sqrt{2}\mathbf{h}_k\mathbf{w}_k]^H) - \|[q_k^n, g_k^n, \sqrt{2}\mathbf{h}_k\mathbf{w}_k^n]\|_2^2, \ \forall k \tag{32}$$

$$\sum_{k\in\mathcal{K}}(x_{b,k} + r_k)^2 \leq 2[\{x_{b,k}^n\}_k, \{r_k^n\}_k][\{x_{b,k}\}_k, \{r_k\}_k]^T - \|[\{x_{b,k}\}_k, \{r_k\}_k]\|_2^2 + 2\vartheta_b, \ \forall b \tag{33}$$

$$\|\mathbf{w}_{b,k}\|_2^2 \leq (q(x_{b,k}^n)^{q-1}x_{b,k} + (1-q)(x_{b,k}^n)^q)\bar{P}_b, \ \forall b, k. \tag{34}$$

Herein, the superscript n denotes the iteration. Moreover, we also convexify $\psi(\mathbf{\Omega}, \alpha, \xi)$ using the first order as $\psi(\mathbf{\Omega}, \alpha, \xi; \mathbf{\Omega}^n) \triangleq \eta + \alpha\sum_{b\in\mathcal{B},k\in\mathcal{K}}(2x_{b,k}x_{b,k}^n - (x_{b,k}^n)^2 - x_{b,k}) + \xi\sum_{b\in\mathcal{B}}\min\{0, \bar{C}_b - \vartheta_b\}$. In summary, at iteration $n+1$ of the proposed method, we solve the following approximate convex program of (27)

$$\max_{\mathbf{\Omega}\in\mathcal{S}_c\backslash(9g)} \ \psi(\mathbf{\Omega}, \alpha, \xi; \mathbf{\Omega}^n) \quad \text{subject to} \quad \{(31) - (34)\}. \tag{35}$$

The convergence of Algorithm 2 can be proved following the arguments in [43, Section 2]. We also refer the interested reader to [34], [42], [44] for other convergence results.

An important point in Algorithm 2 is that the value of penalty parameter α is increased at each iteration, i.e., step 5. We note that a high value of α will encourage $x_{b,k}$ to take on binary values. The idea is to start Algorithm 2 with a small value of α to focus on maximizing the original objective, and then increase α in subsequent iterations to force $x_{b,k}$ to be binary.

B. ℓ_0-Approximation Method

In the second suboptimal method, we view the problem of RRH selection and RRH-user association as finding a sparse solution of beamformer vector $\mathbf{w}$. In particular, no binary variables are introduced to formulate the considered problem. Instead, RRH selection and RRH-user association are concluded from the values of beamformers. To clarify this point, let us consider the inequality $\|\mathbf{w}_{b,k}\|_2 \leq v_{b,k}$. Then it is clear that RRH b is switched off if $\sum_{k\in\mathcal{K}} v_{b,k} = 0$, and switched on if $\sum_{k\in\mathcal{K}} v_{b,k} > 0$. In other words, whether RRH b is active or not is the step function of $\sum_{k\in\mathcal{K}} v_{b,k}$. The central idea of the second proposed method is to approximate the step function by a continuous function to which continuous optimization can be applied. In fact there are many functions proposed in the literature for this purpose in different contexts (see [34] for further discussions on approximations). For the considered problem, we find the following approximation function is very efficient

$$\varphi_\beta(y) \triangleq \min\{1, \beta y\} = \begin{cases} 1 & \text{if } y \geq \frac{1}{\beta} \\ \beta y & \text{if otherwise} \end{cases} \tag{36}$$

where β is the approximation parameter. In fact the above approximation function is a special case of (nonconcave) piecewise linear function presented in [34], [45], which is modified to be concave for the purpose of applying the SCA later on. We can easily see that $\varphi_\beta(y)$ well approximates the step function when β is sufficiently large. Based on the above discussion, we formulate the joint design problem as

$$\underset{\mathbf{w},\mathbf{r},\mathbf{v}}{\text{maximize}} \quad \frac{\sum_{k\in\mathcal{K}} r_k}{\check{f}_{\text{P}}(\mathbf{w}, \mathbf{r}, \mathbf{v})} \tag{37a}$$

$$\text{subject to } \|\mathbf{w}_{b,k}\|_2 \leq v_{b,k}, \ \sum_{k\in\mathcal{K}} v_{b,k}^2 \leq \bar{P}_b, \ \forall b \in \mathcal{B} \tag{37b}$$

$$\sum_{k\in\mathcal{K}} \varphi_\beta(v_{b,k})r_k \leq \bar{C}_b, \ \forall b \in \mathcal{B} \tag{37c}$$

$$(9b), (9c), (9f) \tag{37d}$$

where $\mathbf{v} \triangleq \{v_{b,k}\}$ and $\check{f}_{\text{P}}(\mathbf{w}, \mathbf{r}, \mathbf{v}) \triangleq \sum_{b\in\mathcal{B}}\big(\sum_{i=1}^I \tilde{\epsilon}||\tilde{\mathbf{w}}_{b,i}||_2 + \Delta P\varphi_\beta(\sum_{k\in\mathcal{K}} v_{b,k}) + p_{\text{SP}}\sum_{k\in\mathcal{K}}\varphi_\beta(v_{b,k})r_k\big) + P_{\text{const}}$. We note that (37) is still nonconvex but it has fewer optimization variables than (9). Next we rewrite (37) as

$$\underset{\substack{\eta,t,\mathbf{w},\mathbf{r},\mathbf{v}\\ \mathbf{g},\mathbf{q},\boldsymbol{\vartheta},\boldsymbol{\mu},\boldsymbol{\nu}}}{\text{maximize}} \quad \eta \tag{38a}$$

$$\text{subject to } t \geq \sum_{b\in\mathcal{B}}(\sum_{i=1}^I \tilde{\epsilon}||\tilde{\mathbf{w}}_{b,i}||_2 + \Delta P\nu_b + p_{\text{SP}}\vartheta_b) + P_{\text{const}} \tag{38b}$$

$$\sum_{k\in\mathcal{K}} \mu_{b,k}r_k \leq \vartheta_b, \ \vartheta_b \in [0, \bar{C}_b], \ \forall b \in \mathcal{B} \tag{38c}$$

$$\mu_{b,k} \geq \varphi_\beta(v_{b,k}), \ \forall b \in \mathcal{B}, k \in \mathcal{K} \tag{38d}$$

$$\nu_b \geq \varphi_\beta(\textstyle\sum_{k\in\mathcal{K}} v_{b,k}), \ \forall b \in \mathcal{B} \tag{38e}$$

$$(9c), (9f), (23d), (23e), (24b), (24c), (37b) \tag{38f}$$

where $\boldsymbol{\mu} \triangleq \{\mu_{b,k}\}_{b,k}$ and $\boldsymbol{\nu} \triangleq \{\nu_b\}_b$, and the introduction of $\eta, t, \mathbf{g}, \mathbf{q}, \boldsymbol{\vartheta}$ follows exactly the same arguments as those in the previous subsection. For the ease of description, we define

$$\tilde{\mathcal{S}}_{\text{c}} \triangleq \{\tilde{\mathbf{\Omega}}|(9c), (9f), (23d), (23e), (37b), (38b)\}$$
$$\tilde{\mathcal{S}}_{\text{nc}} \triangleq \{\tilde{\mathbf{\Omega}}|(24b), (24c), (38c) - (38e)\}$$

where $\tilde{\mathbf{\Omega}} \triangleq \{\eta, t, \mathbf{w}, \mathbf{r}, \mathbf{v}, \mathbf{g}, \mathbf{q}, \boldsymbol{\vartheta}, \boldsymbol{\mu}, \boldsymbol{\nu}\}$. Note that $\tilde{\mathcal{S}}_{\text{c}}$ and $\tilde{\mathcal{S}}_{\text{nc}}$ are the convex and nonconvex parts of (38), respectively. Now the application of SCA to solve (38) is straightforward. The nonconvex constraints (24b), (24c) and (38c) in $\tilde{\mathcal{S}}_{\text{nc}}$ can be convexified using the same way as done in the previous subsection, given in (31)–(33). Convex approximation of $\varphi_\beta(y)$ deserves a remark. Note that $\varphi_\beta(y)$ is concave and continuous but not smooth at $y = \frac{1}{\beta}$. However we can use the subdifferential of $\varphi_\beta(y)$ to derive a convex upper bound. It is easy to check that a subgradient of $\varphi_\beta(y)$ is given by

$$\partial\varphi_\beta(y) = \begin{cases} 0 & \text{if } y \geq \frac{1}{\beta} \\ \beta & \text{if otherwise} \end{cases} \tag{39}$$

Algorithm 3 Proposed method for solving (37)

1: **Initialization:** Set $n := 0$, choose initial values for $\tilde{\mathbf{\Omega}}^0$ and set β^0 small
2: **repeat** $\{n := n + 1\}$
3: Solve (41) and achieve $\tilde{\mathbf{\Omega}}^*$
4: Update $\tilde{\mathbf{\Omega}}^n := \tilde{\mathbf{\Omega}}^*$
5: Update $\beta^n := \min\{\beta_{\max}; \beta^{n-1} + \varepsilon\}$ for small ε
6: **until** Convergence and output $\tilde{\mathbf{\Omega}}^*$

and thus we can approximate (38d) and (38e) as

$$\mu_{b,k} \geq \bar{\varphi}_\beta(v_{b,k}; v_{b,k}^n) \tag{40a}$$
$$\nu_b \geq \bar{\varphi}_\beta(\textstyle\sum_{k\in\mathcal{K}} v_{b,k}; \sum_{k\in\mathcal{K}} v_{b,k}^n) \tag{40b}$$

where $\bar{\varphi}_\beta(y; y^n) \triangleq \begin{cases} 1 & \text{if } y^n \geq \frac{1}{\beta} \\ \beta y & \text{otherwise} \end{cases}$. Finally, we arrive at the approximate convex program of problem (38), i.e.,

$$\max_{\tilde{\mathbf{\Omega}}\in\tilde{\mathcal{S}}_c} \quad \eta \quad \text{subject to } \{(31)-(33), (40a), (40b)\}. \tag{41}$$

We describe the second proposed suboptimal method in Algorithm 3. Similar to Algorithm 2, the approximation parameter β is also updated after each iteration. The idea is the same as β is viewed to provide the tightness of the binary approximation function (36). In Algorithm 3 we start with a small value of β and then increase β after each iteration. Numerical results provided in the next section demonstrate the impact of updating β. To avoid the problem of the initial guess, we can add the penalty of violating the fronthaul constraints to the objective of (41) similarly as with Algorithm 2. Convergence of Algorithm 3 is guaranteed, which is discussed in Appendix B. It is worth mentioning that the achieved limit point is not ensured to hold the first-order optimality of (38) since the approximation of the step function is not smooth.

C. Second-order-cone Representation

This subsection presents a more efficient way to treat (23d). First we remark that (23d) is indeed a convex constraint and thus convex approximation is not required. However, since (23d) involves an exponential cone, (35) and (41) are generic nonlinear programs, while other constraints are SOC presentable. This prevents us from exploiting powerful conic convex solvers such as MOSEK or GUROBI. To take full advantage of these solvers we will also approximate (23d) using the SCA framework. More explicitly, we will approximate $\log(1 + g_k)$ by a lower bound that makes the resulting constraint SOC representable. To this end we recall the following inequality

$$\log(1 + g_k) \geq g_k(1 + g_k)^{-1}. \tag{42}$$

Substituting g_k in both sides of (42) by $\frac{g_k - g_k^n}{g_k^n + 1}$ results in

$$\log(1 + g_k) \geq \log(1 + g_k^n) + (g_k - g_k^n)(1 + g_k)^{-1}. \tag{43}$$

Now we can approximate (23d) as

$$\log(1 + g_k^n) + (g_k - g_k^n)(1 + g_k)^{-1} \geq r_k. \tag{44}$$

Table I
SIMULATION PARAMETERS

PARAMETERS	VALUE
Inter-RRH distance	200 m
Active power for RRH and NU P^{active} [35], [36]	10.65 W
Sleep power for RRH and NU P^{sleep} [35], [36]	5.05 W
Circuit power for user P_{ms}	0.1 W
Max. power efficiency $\epsilon_{\max}$ [24]	0.55
Number of Tx antennas N	2
Min. rate requirement r_0	1 nat/s/Hz
Bandwidth	10 MHz
Noise power	-143 dBW

The above constraint can be reformulated as an SOC constraint as

$$\|2\sqrt{1 + g_k^n},\ \log(1 + g_k^n) - r_k - g_k\|_2 \leq \log(1 + g_k^n) - r_k + g_k + 2,\ \forall k \in \mathcal{K}. \tag{45}$$

Using (45), the convex program obtained at each iteration of Algorithms 2 and 3 is an SOCP which is much easier to solve.

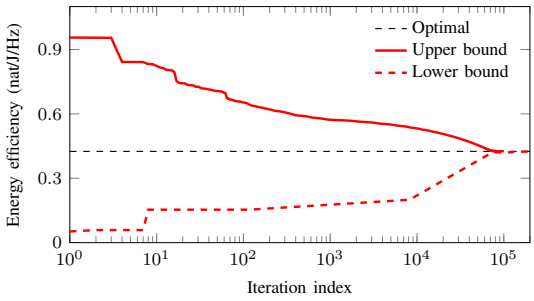
D. Complexity Analysis of Algorithms 2 and 3

We now discuss the worst-case per-iteration complexity of the two proposed suboptimal algorithms. For Algorithm 2, the SOCP consists of $2BKI + BK + 2B + 3K + 2$ real-valued variables and $B(K+I)+2B+3K+2$ conic constraints. Thus, the per-iteration complexity for solving the SOCP problems corresponding to Algorithms 2 by path-following interior-point method are $\mathcal{O}(\sqrt{B(K+I)}B^3K^3I^3)$ [46]. Similarly, the per-iteration complexity for solving the SOCP in Algorithm 3, which contains $2BKI + 2BK + 2B + 3K + 2$ real-valued variables and $B(K + I) + 2B + 3K + 1$ conic constraints, is $\mathcal{O}(\sqrt{B(K+I)}B^3K^3I^3)$ [46].

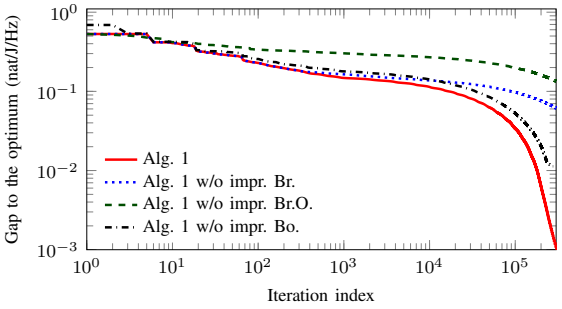
V. NUMERICAL RESULTS

In this section we provide numerical demonstration to evaluate the effectiveness of the proposed methods. The simulation parameters shown in Table I are used, unless mentioned otherwise. The channel $\mathbf{h}_{b,k}$ between RRH b and user k is assumed to be flat fading which is generated following Gaussian distribution, i.e., $\mathbf{h}_{b,k} \sim \mathcal{CN}(0, \rho_{b,k}\mathbf{I}_I)$, where $\rho_{b,k}$ represents the large-scale fading and is calculated as $\rho_{b,k}[\text{dB}] = 30\log_{10}(d_{b,k}) + 38 + \mathcal{N}(0, 8)$ ($d_{b,k}$ is the distance in meters). P_{a} is set to be same for all antenna chains, and we take $\bar{P} = \bar{P}_b = IP_{\text{a}}, \forall b$. For the maximum fronthaul capacity, we set $\bar{C}_b = \bar{C}, \forall b$.

We generate initial point $\mathbf{\Omega}^0$ for starting Algorithm 2 by solving the power minimization problem (17) with selection vectors being fixed as $x_{b,k}^0 = 1$ and $s_b^0 = 1\ \forall b, k$ to obtain $\mathbf{w}^0$; then the values for the remaining variables are determined based on (23b)–(23g). The initial point $\tilde{\mathbf{\Omega}}^0$ for starting Algorithm 3 is generated similarly with $\mu_{b,k}^0 = 1$ and $\nu_b^0 = 1\ \forall b, k$. For the penalty parameters, we take $\xi = 1$ and initialize $\alpha^0 = 10^{-5}$ and $\beta^0 = 0.1$. Algorithms 2 and 3 are terminated when the increase in the objective between two consecutive iterations is less than 10^{-6}.

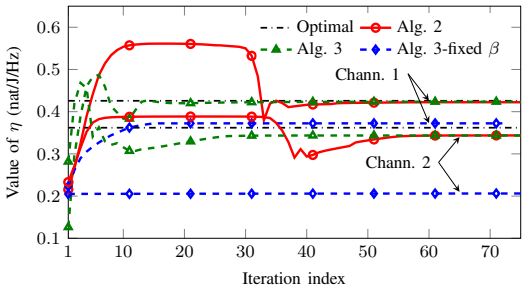


(a) Convergence of the upper and lower bounds.

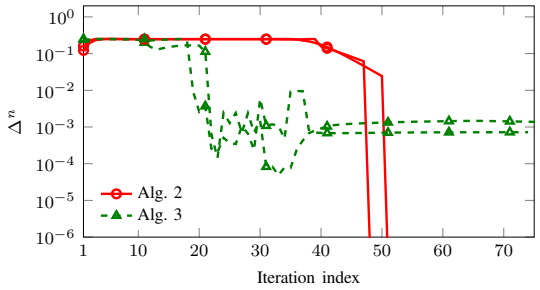


(b) Convergence speed with and without the proposed mofications.

Fig. 3. Convergence behavior of Algorithm 1 for one channel realization with $B = 3$, $K = 4$, $\bar{P} = 30$ dBm, $\bar{C} = 10$ nats/s/Hz and $p_{\text{SP}} = 10$ W/(Gnats/Hz).



(a) Convergence behavior of the proposed suboptimal algorithms.



(b) Gap to the binary of relaxed variables obtained by the proposed suboptimal algorithms.

Fig. 4. Convergence behavior of the proposed suboptimal algorithms for two random channel realizations with $B = 3$, $K = 4$, $\bar{P} = 30$ dBm, $\bar{C} = 10$ nats/s/Hz and $p_{\text{SP}} = 10$ W/(Gnats/Hz).

A. Convergence Results

The first set of experiments examines the convergence behavior of the proposed methods. We consider the network setting where $B = 3$, $K = 4$, $\bar{P} = 30$ dBm, $\bar{C} = 10$ nats/s/Hz and $p_{\text{SP}} = 10$ W/(Gnats/Hz).

The convergence performance of Algorithm 1 for a random channel realization is demonstrated in Fig. 3. Particularly, Fig. 3(a) depicts the upper and lower bounds returned by the algorithm. We can see that the bounds monotonically converge to the optimal value. Fig. 3(b) shows the convergence speed of Algorithm 1 by the gap between the upper bound and the optimal value over iterations. In this figure, we also provide the performance of other schemes to confirm the effectiveness of the proposed modifications made to the DBRB. Specifically, the schemes labelled '*w/o. impr. Br.*', '*w/o. impr. Br.O.*', and '*w/o. impr. Bo.*' represent for Algorithm 1 without applying improved branching, improved branching order and improved bounding, respectively. The results clearly demonstrate that applying the proposed modifications significantly improves the convergence performance.

In Fig. 4, we show the convergence behavior of Algorithms 2 and 3 for two random channel realizations. In order to illustrate the advantages of updating parameter β in Algorithm 3, we also provide the convergence results of Algorithm 3 without updating β dubbed as '*Alg. 3-fixed* β'. For this scheme, we fix $\beta = 1000$. Fig. 4(a) shows the variation of η over iterations. It is observed that Algorithms 2 and 3 converge to the points close to the optimal values within a few tens of iterations. This behavior proves that the proposed algorithms are fast convergent and effective methods. Another observation is that, with a fixed β, Algorithm 3 converges very fast but results in poor performance. Whereas, by updating β, the algorithm needs a bit more iterations to achieve near-optimal performance. In Fig. 4(b), we study how close the obtained values of the relaxed variables are to 0 or 1. Let us define

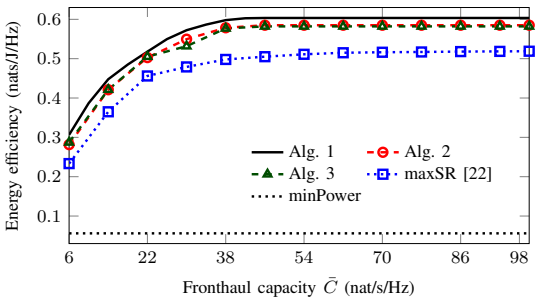
$$\Delta^n \triangleq \begin{cases} \max_{b,k}\{x^n_{b,k} - (x^n_{b,k})^2\} & \text{for Algorithm 2} \\ \max_{b,k}\{\mu^n_{b,k} - (\mu^n_{b,k})^2\} & \text{for Algorithm 3} \end{cases}$$

and note that a smaller Δ^n indicates a closer gap between $\{x^n_{b,k}\}_{b,k}$ (or $\{\mu^n_{b,k}\}_{b,k}$) and binary values. As can be seen, $\Delta^n \approx 0$ at convergence for Algorithm 2 which implies that the penalty method can achieve binary solutions. On the other hand, although the ℓ_0-approximation method (i.e. Algorithm 3) cannot derive exact binary solutions (for all relaxed variables), it still returns $\{\mu^n_{b,k}\}_{b,k}$ very close to 0 or 1 at convergence (the maximum gap is about 10^{-3}).

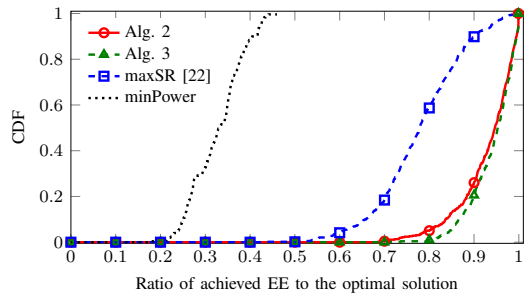
For the solving time, the corresponding average per-iteration runtime of solver MOSEK [47] with Algorithms 1, 2 and 3 are 0.006 s, 0.008 s, and 0.007 s, respectively. As can be seen, the per-iteration runtime is relatively small due to the fact that only an SOCP is solved in each iteration for all algorithms.

B. EE Comparison between Optimal and Suboptimal Algorithms

In Fig. 5, we illustrate the effectiveness of the proposed suboptimal algorithms by comparing their average EE performances with that of Algorithm 1 and the existing schemes, those are sum rate maximization (maxSR) [25] and power consumption minimization (minPower) [21], [26], [32]. Fig.



(a) Average EE versus $\bar{C}$.



(b) CDF of the ratio of achieved EE (of the considered schemes) to the optimal solution.

Fig. 5. Average performances of the considered schemes with $B = 3$, $K = 4$, $\bar{P} = 30$ dBm and $p_{\text{SP}} = 10$ W/(Gnats/Hz).

5(a) plots the average EE of the considered schemes as a function of the fronthaul capacity $\bar{C}$. It is seen that the proposed methods remarkably outperform the existing schemes. The important observation is that the gaps between the curves of the optimal and suboptimal algorithms are really small in all cases of $\bar{C}$ demonstrating the validity of the proposed suboptimal schemes in terms of the average EE. We can also observe that the performance of Algorithms 2 and 3 almost agree with each other. Fig. 5(b) shows the cumulative distribution function (CDF) of the ratio of achieved EE (of Algorithms 2 and 3, maxSR, and minPower) to the optimal solution. As can be observed, the probability that Algorithms 2 and 3 achieve more than 90% of the optimal values is up to 75%. In the worst case, these schemes also achieve about 70% of the optimal performance. We can also see that most of the solutions obtained by maxSR and minPower are far from the optimal values.

C. Performances of the Proposed Suboptimal Algorithms in Large Network Settings

In the following set of experiments, we consider a larger network setting and evaluate the impacts of the fronthaul capacity, the signal processing power and the dynamics of the PA's efficiency on the EE performance. In particular, we evaluate the suboptimal methods in a 7-cell wrap-around topology with $B = 7$ RRHs in which a total of $K = 14$ users are randomly placed across the network's coverage.

1) Impact of Fronthaul Capacity: Fig. 6 shows the achieved EE of Algorithms 2, 3, and maxSR versus the fronthaul capacity $\bar{C}$. The corresponding average number of served users per RRH and average number of serving RRHs per user are

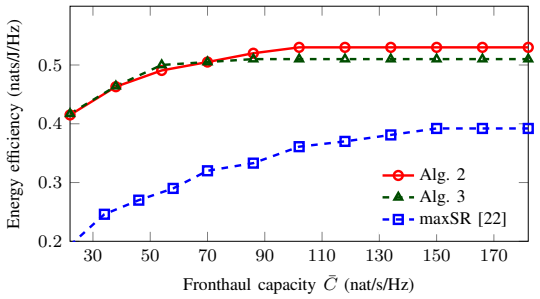


Fig. 6. Average EE performance of the considered schemes versus $\bar{C}$ with $B = 7$, $K = 14$, $\bar{P} = 30$ dBm and $p_{\text{SP}} = 10$ W/(Gnats/Hz).

provided in Table II. We can see from Fig. 6 that EE increases as $\bar{C}$ increases for all considered schemes. However, after a certain large value of $\bar{C}$, further increasing $\bar{C}$ does not change the performance. This observation is consistent with that in Fig. 5(a). The result can be explained as follows. For the EE schemes, to increase $\bar{C}$ is to expand feasible set of (9). When $\bar{C}$ is small, it is the primary constraint on the network performances. Thus the expanded feasible set results in performance improvement. When $\bar{C}$ is large enough, other constraints (e.g. transmit power constraints) become the primary restriction on the network performance. In this case, increasing $\bar{C}$ has no impact on the objective value. For a physical interpretation, increasing the fronthaul capacity allows a RRH to serve more users, i.e. the number of RRHs cooperating to transmit data to a user increases (as can be seen from Table II). This increases the cooperation gain, and thus improves the system performance. When the fronthaul capacity is large enough such that either the additional cooperation gain provides no gain in the achieved performance or the full connection (each user is served by all RRHs) is arrived, increasing the fronthaul capacity does not change the performance. Therefore, in the large fronthaul capacity regime, we can observe from the table that maxSR arrives at full connection, since this topology provides the maximum capacity for wireless transmission. On the other hand, for the EE schemes, the average number of serving RRHs per user is smaller than B even when $\bar{C}$ is sufficiently large. This is because adding more serving RRHs for a user degrades the EE performance, if the benefit from the cooperation gain cannot compensate for the additional signal processing power.

2) Impact of Rate-dependent Signal Processing Power: Fig. 7 depicts the EE performance of the considered schemes versus different values of p_{SP}. We recall that, for a fixed data rate, a larger p_{SP} leads to larger power consumed in signal processing. As expected, the EE decreases when p_{SP} increases for all considered schemes. For SRmax, the sum rate performance is independent of p_{SP}. Thus, its EE performance is a decreasing function of p_{SP} due to the increase in the total consumed power with respect to p_{SP}. The results clearly show that parameter p_{SP} has a significant impact on the EE performance, indicating that the model of rate-dependent signal processing power should be considered for proper EE designs and evaluation.

Table II
AVERAGE NUMBER OF SERVED USERS PER RRH AND AVERAGE NUMBER OF SERVING RRHS PER USER CORRESPONDING TO THE SIMULATION RESULTS SHOWN IN FIG. 6.

$\bar{C}$ (nats/s/Hz)		22	30	38	46	54	66	70	78	86	118	150
Algorithm 2	Num. of served users per RRH	8.4	9.8	10.3	11.1	11.3	11.4	11.5	11.5	11.6	11.8	12.0
	Num. of serving RRHs per user	4.2	4.9	5.1	5.5	5.6	5.7	5.7	5.7	5.8	5.9	6.0
Algorithm 3	Num. of served users per RRH	8.0	8.6	8.9	9.4	9.6	9.7	9.8	9.8	9.8	9.8	9.8
	Num. of serving RRHs per user	4.0	4.3	4.5	4.7	4.8	4.9	4.9	4.9	4.9	4.9	4.9
maxSR [25]	Num. of served users per RRH	11.3	12.0	12.3	12.6	13.0	13.6	13.8	13.8	13.8	14	14
	Num. of serving RRHs per user	5.6	6.0	6.1	6.3	6.5	6.8	6.9	6.9	6.9	7	7

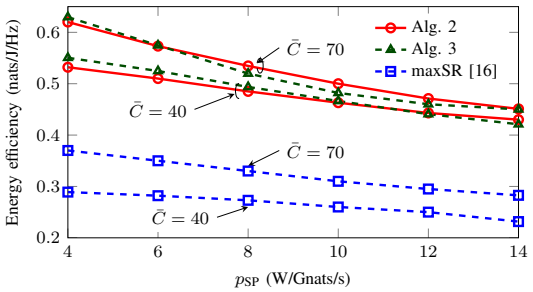
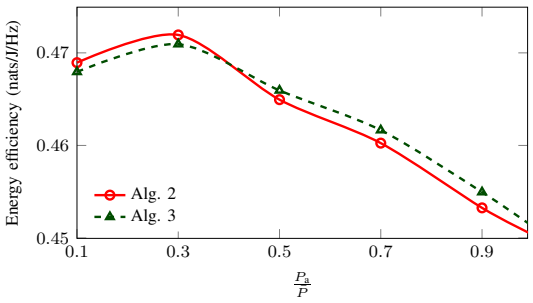
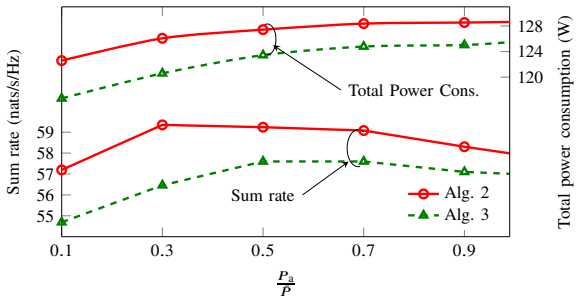


Fig. 7. Average EE performance of the considered schemes versus p_{SP} with $B = 7$, $K = 14$ and $\bar{P} = 30$ dBm.



(a) Average EE



(b) Average sum rate and total power consumption

Fig. 8. Achieved performances versus the ratio $\frac{P_a}{\bar{P}}$ with $B = 7$, $K = 14$, $\bar{C} = 40$ nats/s/Hz, $p_{SP} = 10$ W/(Gnats/Hz) and $\bar{P} = 30$ dBm.

3) Impact of the Dynamics of PA's Efficiency: In the final experiment, we fix $\bar{P} = 30$ dBm and let P_a vary to investigate the impact of the dynamics of PA's efficiency on the EE performance. We recall that with some given ϵ_{max} and input power, the PA's efficiency is a decreasing function with respect to P_a (see (6)). Fig. 8(a) plots the EE performances of Algorithms 2 and 3 versus P_a. The corresponding sum rate and consumed power are shown in Fig. 8(b). As can be seen from Fig. 8(a), when P_a increases, the EE performance first increases and then decreases. This observation can be explained as follows. In the small regime of P_a, the transmit power is small and an increase in the transmit power results in a significant increase in the data rate, due to the logarithmic behavior of the data rate w.r.t the transmit power. For this situation, the PA's efficiency is still sufficiently high. Therefore, as P_a increases, the additional transmit power increases the sum rate more significantly than the power consumption on PAs, and thus, it achieves better EE (as can be seen from Fig. 8(b)). However, after a certain value of P_a, the effective PA's efficiency becomes small and its negative impact outweighs the benefit of increasing transmit power. In this case, the reduced PA's efficiency due to the increase of P_a degrades EE performance.

VI. CONCLUSION

This paper has studied the joint designs of beamforming, RRH-user association and RRH selection in C-RANs to maximize the system EE subject to per-RRH fronthaul capacity, transmit power budget and per-user QoS. Specially, we have adopted relatively realistic power consumption model compared to the previous works where the impacts of rate-dependent signal processing power and the dynamics of PA's efficiency are considered. To investigate the optimal performance of the formulated problem, we have developed the new globally optimal method by customizing the DBRB algorithm. We have also proposed novel modifications on the generic framework of the DBRB method to improve the optimal algorithm's efficiency. Towards practically appealing methods, we have proposed two suboptimal approaches which can achieve very close to optimal performance with much reduced complexity. Numerical evaluations have been provided to demonstrate the effectiveness of the proposed schemes. Specifically, the proposed modifications made on the DBRB framework remarkably reduce the complexity of the globally optimal method. On the other hand, the two proposed suboptimal approaches can achieve a near-optimal solutions with a reasonable complexity and outperform the other known methods. The impacts of the limited fronthaul capacity, rate-dependent signal processing power and the dynamic of PA's efficiency on the EE performance have also been demonstrated.

APPENDIX

A. Proof of Lemma 1

First we show that (9b) is active, which is proved by the contradiction. Let $(\eta^*, \mathbf{w}^*, \mathbf{r}^*, \mathbf{t}^*)$ be an optimal solution of

(15) and suppose that (9b) is not active at the optimum, i.e., $r_k^* < \log(1+\gamma_k(\mathbf{w}^*))$ for some k. Then we can scale down the transmit power for user k, i.e., $\|\mathbf{w}_k\|_2^2$, to achieve a new beamformer $\|\hat{\mathbf{w}}_k\|_2^2$ such that $\|\hat{\mathbf{w}}_k\|_2^2 = \tau\|\mathbf{w}_k\|_2^2 < \|\mathbf{w}_k\|_2^2$ for $\tau \in (0,1)$ while keeping the others unchanged. By this way, we can achieve $r_k^* < \log(1+\gamma_k(\hat{\mathbf{w}}))$ for all k, since interference power at all users has reduced. However, the new set of beamformers also generates a new power consumption vector on PAs $\sum_{b\in\mathcal{B}} \hat{t}_b < \sum_{b\in\mathcal{B}} t_b^*$, which immediately implies the increase of EE objective, i.e., $\eta > \eta^*$. This contradicts to the fact that $(\eta^*, \mathbf{w}^*, \mathbf{r}^*, \mathbf{t}^*)$ is the optimal solution and thus completes the proof. Now for fixed $(\mathbf{x}^*, \mathbf{s}^*, \mathbf{r}^*, \mathbf{t}^*)$, problem (9) reduces to a beamforming design subject to the desired data rate $\mathbf{r}^*$ and the power constraint $\sum_{i=1}^{I} ||\tilde{\mathbf{w}}_{b,i}^*||_2 = t_b^*$ such that at the output of (16), (9b) must be binding.

B. Convergence Analysis of Algorithm 3

We justify the convergence of Algorithm 3 by showing the following facts: (i) when $\beta^n < \beta_{\max}$, the update of β (see Step 5) tightens the approximations (40a) and (40b) after every iteration; and (ii) let $\bar{n}$ be the iteration such that $\beta^{\bar{n}-1} < \beta_{\max}$ and $\beta^{\bar{n}} = \beta_{\max}$, then we have the sequence $\{\eta^n\}_{n>\bar{n}}$ being non-decreasing, which is guaranteed to converge.

To prove (i), let us consider the non-smooth constraint (40a), i.e., $\mu_{b,k} \geq \bar{\varphi}_\beta(v_{b,k}; v_{b,k}^n)$ with arbitrary β. Since it holds that $\bar{\varphi}_{\bar{\beta}}(.) \geq \bar{\varphi}_\beta(.)$ for any $\bar{\beta} \geq \beta$, we can replace $\bar{\varphi}_\beta(v_{b,k}; v_{b,k}^n)$ by $\bar{\varphi}_{\bar{\beta}}(v_{b,k}; v_{b,k}^n)$ in (40a) to obtain a tighter approximation, i.e., $\mu_{b,k} \geq \bar{\varphi}_{\bar{\beta}}(v_{b,k}; v_{b,k}^n)$. Similarly, we can use the same argument for (40b).

Next, we prove (ii). We recall that the feasible set of (41) is bounded by power, fronthaul and users' QoS constraints. Thus it is sufficient to prove that solution of (41) returned at iteration n (i.e., $\tilde{\mathbf{\Omega}}^n$) is feasible to the problem at iteration $n+1$ for $n > \bar{n}$, as such we yield $\eta^{n+1} \geq \eta^n$ [13], [14]. To this end we note that $\tilde{\mathbf{\Omega}}^n$ satisfies (smooth) constraints (31)–(33) at iteration $n+1$. This fact follows from the properties of convex approximation which is also discussed in [42, Properties (i) and (ii)]. On the other hand, for non-smooth constraint (40a), we have $\bar{\varphi}_{\beta_{\max}}(v_{b,k}^n; v_{b,k}^n) = \min\{1, \beta_{\max} v_{b,k}^n\} \leq \mu_{b,k}^n$. This is because $(v_{b,k}^n, \mu_{b,k}^n)$ is the solution of (41) at iteration n. The result for (40b) can be obtained following the same manner. At this point, we accomplish the argument (ii).

REFERENCES

[1] P. Marsch and G. P. Fettweis, *Coordinated Multi-Point in Mobile Communications: from theory to practice*. Cambridge University Press, 2011.

[2] 3GPP, "Coordinated multi-point operation for LTE physical layer aspects (Release 11)," 3rd Generation Partnership Project, TR 36.819. [Online]. Available: http://www.3gpp.org/technologies

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

[4] J. Wu, "Green wireless communications: from concept to reality [industry perspectives]," *IEEE Wireless Commun.*, vol. 19, no. 4, pp. 4–5, Aug. 2012.

[5] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks – A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

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

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

[8] D. Feng, C. Jiang, G. Lim, J. Cimini, L. J., G. Feng, and G. Li, "A survey of energy-efficient wireless communication," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 167–178, Feb. 2013.

[9] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Foundations and Trends in Communications and Information Theory*, vol. 11, no. 3-4, pp. 185–396, 2015.

[10] D. Nguyen, L.-N. Tran, P. Pirinen, and M. Latva-aho, "Precoding for full duplex multiuser MIMO systems: Spectral and energy efficiency maximization," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4038–3050, Aug. 2013.

[11] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in multi-cell OFDMA systems with limited backhaul capacity," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3618–3631, Oct. 2012.

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

[13] K.-G. Nguyen, L.-N. Tran, O. Tervo, Q.-D. Vu, and M. Juntti, "Achieving energy efficiency fairness in multicell multiuser MISO downlink," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1426 – 1429, Aug. 2015.

[14] 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 – 5588, Oct. 2015.

[15] A. Zappone, E. Björnson, L. Sanguinetti, and E. Jorswieck, "Globally optimal energy-efficient power control and receiver design in wireless networks," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2844–2859, Jun. 2017.

[16] S. He, Y. Huang, S. Jin, F. Yu, and L. Yang, "Max-min energy efficient beamforming for multicell multiuser joint transmission systems," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1956–1959, Oct. 2013.

[17] C. Isheden and G. P. Fettweis, "Energy-efficient multi-carrier link adaptation with sum rate-dependent circuit power," in *2010 IEEE GLOBECOM 2010*, Dec. 2010, pp. 1–6.

[18] C. Xiong, G. Y. Li, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient resource allocation in OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 60, no. 12, pp. 3767–3778, Dec. 2012.

[19] T. Wang and L. Vandendorpe, "On the optimum energy efficiency for flat-fading channels with rate-dependent circuit power," *IEEE Trans. Commun.*, vol. 61, no. 12, pp. 4910–4921, Dec. 2013.

[20] T. X. Tran, A. Younis, and D. Pompili, "Understanding the computational requirements of virtualized baseband units using a programmable cloud radio access network testbed," in *2017 IEEE International Conference on Autonomic Computing (ICAC)*, Columbus, Ohio,USA, July 2017, pp. 221–226.

[21] A. Hajisami, T. X. Tran, and D. Pompili, "Elastic-net: Boosting energy efficiency and resource utilization in 5G C-RANs," in *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, Orlando, FL, USA, Oct 2017, pp. 466–470.

[22] S. Mikami, T. Takeuchi, H. Kawaguchi, C. Ohta, and M. Yoshimoto, "An efficiency degradation model of power amplifier and the impact against transmission power control for wireless sensor networks," in *2007 IEEE Radio and Wireless Symposium,*, 2007, pp. 447–450.

[23] E. Björnemo, "Energy constrained wireless sensor networks: Communication principles and sensing aspects," Ph.D. dissertation, Institutionen för teknikvetenskaper, 2009.

[24] D. Persson, T. Eriksson, and E. G. Larsson, "Amplifier-aware multiple-input multiple-output power allocation," *IEEE Commun. Lett.*, vol. 17, no. 6, pp. 1112–1115, Jun. 2013.

[25] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, no. 6, pp. 1326–1339, Oct. 2014.

[26] ——, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, April 2016.

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

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

[29] K. G. Nguyen, Q. D. Vu, M. Juntti, and L. N. Tran, "Globally optimal beamforming design for downlink CoMP transmission with

limited backhaul capacity," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.
[30] A. Hajisami and D. Pompili, "Dynamic joint processing: Achieving high spectral efficiency in uplink 5G cellular networks," *Computer Networks*, vol. 126, pp. 44 – 56, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128617302700
[31] ——, "Joint virtual edge-clustering and spectrum allocation scheme for uplink interference mitigation in C-RAN," *Ad Hoc Networks*, vol. 72, pp. 91 – 104, 2018.
[32] 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.
[33] H. Tuy, M. Minoux, and N. Hoai-Phuong, "Discrete monotonic optimization with application to a discrete location problem," *SIAM Journal on Optimization*, vol. 17, no. 1, pp. 78–97, 2006.
[34] H. A. Le Thi, T. P. Dinh, H. M. Le, and X. T. Vo, "DC approximation approaches for sparse optimization," *European Journal of Operational Research*, vol. 244, no. 1, pp. 26–46, 2015.
[35] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.
[36] A. R. Dhaini, P. H. Ho, G. Shen, and B. Shihada, "Energy efficiency in TDMA-Based next-generation passive optical access networks," *IEEE/ACM Trans. Netw.*, vol. 22, no. 3, pp. 850–863, Jun. 2014.
[37] H. Tuy, F. Al-Khayyal, and P. T. Thach, "Monotonic optimization: Branch and cut methods," in *Essays and Surveys in Global Optimization*. Springer, 2005, pp. 39–78.
[38] E. Björnson, G. Zheng, M. Bengtsson, and B. Ottersten, "Robust monotonic optimization framework for multicell MISO systems," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2508–2523, May 2012.
[39] E. A. Jorswieck and E. G. Larsson, "Monotonic optimization framework for the two-user MISO interference channel," *IEEE Trans. Commun.*, vol. 58, no. 7, pp. 2159–2168, Jul. 2010.
[40] P. Luong, F. Gagnon, C. Despins, and L. N. Tran, "Optimal joint remote radio head selection and beamforming design for limited fronthaul C-RAN," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5605–5620, Nov. 2017.
[41] J. Dattorro, *Convex optimization & Euclidean distance geometry*. Lulu. com, 2010.
[42] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Operations Research*, vol. 26, no. 4, pp. 681–683, Jul.-Aug. 1978.
[43] H. A. Le Thi, T. P. Dinh *et al.*, "DC programming and DCA for general DC programs," in *Advanced Computational Methods for Knowledge Engineering*. Springer, 2014, pp. 15–35.
[44] H. A. Le Thi, T. Pham Dinh, and H. V. Ngai, "Exact penalty and error bounds in DC programming," *Journal of Global Optimization*, vol. 52, no. 3, pp. 509–535, 2012.
[45] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *Journal of Machine Learning Research*, vol. 11, no. Mar, pp. 1081–1107, 2010.
[46] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization*. Philadelphia: MPS-SIAM Series on Optimization, SIAM, 2001.
[47] I. MOSEK ApS, 2014, [Online]. Available: www.mosek.com.