2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC): Track 2: MAC and Cross---Layer Design

Energy Efficiency Maximization in Massive MIMO-aided, Fronthaul-constrained C-RAN

Jobin Francis and Gerhard Fettweis, *Fellow, IEEE*

***Abstract*—Cloud radio access network (C-RAN) and massive multiple-input-multiple-output (MIMO) are two key enabling technologies for 5G as they improve radio performance while lowering the cost of operation. In a C-RAN system with massive MIMO-based remote radio units (RRUs), fronthaul is often the bottleneck in practice due to its finite capacity. To reduce the capacity requirements on fronthaul, precoding is done at the RRUs. In this paper, we maximize the energy efficiency (EE) of such a system by optimizing the transmit powers while explicitly incorporating the capacity constraints on fronthaul. We develop a successive convex approximation (SCA) algorithm, which is guaranteed to converge to a local optimum. Towards this, we propose novel bounds on the user rate function, which facilitates a convex approximation of the EE maximization problem. The convex problem is solved in each SCA iteration through Dinkelbach's algorithm and dual decomposition. Numerical results show that the proposed algorithm significantly improves EE compared to the case with no power control and outperforms the weighted minimum mean square error algorithm.**

I. Introduction

The 5G cellular systems are envisioned to improve spectral efficiency, energy efficiency (EE), and coverage while lowering latency [1]. Two promising technologies to achieve these challenging goals while reducing the capital and operational expenditures are cloud radio access network (C-RAN) and massive multiple-input-multiple-output (MIMO).

In C-RAN, the radio frequency processing is done at remote radio units (RRUs) while the baseband processing for all the cells is done centrally at the baseband unit (BBU) pool. The BBU pool is connected to the RRUs via fronthaul. Advantages of C-RAN include higher SE, EE, and cell-edge performance through centralized baseband processing. Further, C-RAN facilitates antenna site simplifications and pooling gains by sharing the co-located computing resources of multiple cells.

In massive MIMO, the base stations are equipped with a large number of antennas. Compared to regular MIMO, it significantly improves the throughput, EE, and coverage through higher spatial multiplexing and array gains. Furthermore, it is quite robust to hardware impairments enabling the use of low cost, low precision components. To avail these benefits in a C-RAN system, massive MIMO-based RRUs are used.

Despite the many advantages of C-RAN, realizing them in practice is challenging due to non-ideal fronthaul. Often, fronthaul is capacity or latency constrained. This can limit the potential gains of C-RAN given the stringent requirements on fronthaul. For example, the required fronthaul capacity to transport a 20 MHz LTE signal for 2×2 MIMO, which gives a peak downlink data rate of 150 Mbps, is 2.46 Gbps [2]. Also, it increases linearly with the number of antennas. Hence, it is infeasible to provision resources in fronthaul to meet the capacity requirements of massive MIMO-based RRUs.

Functional splitting [2] has been used to reduce the fronthaul requirements, in which some baseband functions are moved to the RRU. A functional split suggested for massive MIMO involves moving transmit precoding to the RRUs [3] and thereby, the traffic in fronthaul depends only on users' data rates and not on the number of antennas. Although coordinated beamforming is no longer possible, network performance can still be improved by cooperative power control and coordinated scheduling between multiple cells.

In this paper, we optimize the power allocated to the users to maximize EE while explicitly incorporating the constraints on fronthaul capacity. This makes the problem challenging and different from most of the existing works on C-RAN design. The power control can be done at a slower time-scale leveraging the fact that the user rates can be accurately approximated in terms of their large-scale fading coefficients due to the channel hardening effect in massive MIMO [4]. Further, power control implicitly does coordinated scheduling as a user is not scheduled if the power allocated to it is zero.

A. Related Literature

We briefly discuss the prior works on EE maximization that explicitly incorporates fronthaul capacity constraints. EE is maximized in [5] by optimizing transmit precoding, fronthaul compression covariance matrices, and user association. In [6], EE is maximized in a dynamic traffic scenario while ensuring the stability of the queues. Transmit beamforming, RRU selection, and user association are optimized in [7] to maximize EE. It proposes a successive convex approximation (SCA) algorithm in which a sequence of computationally intensive mixed integer second order cone programs are solved. In [8], transmit powers and subcarrier allocation of an orthogonal frequency division multiple access system is optimized to maximize EE. It employs Dinkelbach's (DB) algorithm and duality to solve the problem. However, we note that the models and the optimization problems considered in [5]–[8] are quite different from ours. Specifically, fronthaul compression is considered in [5] and dynamic traffic and queuing dynamics are considered [6]. Further, the algorithm in [8] lacks any convergence guarantees as the duality gap in it is not zero. We also note that the above works do not consider massive MIMO, in which signal-to-interference-plus-noise ratio (SINR) has a different functional form due to pilot contamination.

The pertinent resource allocation works in massive MIMO systems are as follows: The transmit powers, number of users

978-1-5386-8110-7/19/$31.00 ©2019 IEEE

and antenna elements are optimized to maximize EE in [9]. The trade-off between SE and EE is explored in [10] by optimizing user association and transmit powers. The transmit powers and coefficients of reflecting surface elements are optimized to maximize EE in [11]. However, the works in [9]–[11] do not consider any fronthaul constraints. The sum rate of a C-RAN system with massive MIMO is maximized in [12] by optimizing transmit powers, user association, link selection between RRUs and BBUs, and number of antennas serving a user. However, it considers a simpler fronthaul constraint that the number of users served over a fronthaul link is below a pre-defined value instead of a capacity constraint. Also, it does not consider EE maximization.

B. Contributions

We present an optimization framework to maximize EE by coordinating the transmit powers of users in a C-RAN system, where RRUs are equipped with massive MIMO and fronthaul has limited capacity. The framework is based on SCA, the main idea behind which is to tackle a difficult non-convex problem by solving a sequence of easier convex problems. To apply SCA to our problem, we develop a novel upper bound and a lower bound for the non-convex user rate function. These bounds are needed as user rates appear in the EE expression and in the fronthaul capacity constraint. The convex problem in each SCA iteration is then solved using the DB algorithm and duality. The convergence of the SCA algorithm to a point satisfying the Karush-Kuhn-Tucker (KKT) conditions of the EE maximization problem is guaranteed.

We evaluate the performance of the proposed algorithm for maximal ratio transmission (MRT) and zero-forcing (ZF) precoding schemes and for two different user association rules, namely, distance-based and signal power-based association. We benchmark against a baseline scheme with no power coordination and an extension of the weighted minimum mean square error (WMMSE) algorithm in [13], which, however, has no convergence guarantees. The proposed algorithm significantly improves EE compared to the baseline scheme and outperforms the extension of the WMMSE algorithm.

C. Organization and Notations

The paper is organized as follows. Section II describes the system model and Section III develops algorithms for EE maximization. Numerical results are presented in Section IV. Our conclusions follow in Section V.

We use boldface lowercase and uppercase letters to denote vectors and matrices, respectively. A vector of zeros, a vector of ones, and an $n \times n$ identity matrix are denoted by $\mathbf{0}$, $\mathbf{1}$, and $\mathbf{I}_n$, respectively. Define $[a]^+ = \max\{0, a\}$. Let $\mathbf{x}^H$, $||\mathbf{x}||$ and $\mathbf{x} \geq \mathbf{0}$ denote, respectively, the hermitian, the Euclidean norm, and that each element of $\mathbf{x}$ is non-negative. The expectation operator is denoted by $\mathbb{E}[\cdot]$ and indicator of event A by I_A.

II. System Model

The C-RAN system consists of a BBU, L RRUs and K users as shown in Figure 1. We consider a fronthaul network with a switch that demultiplexes the data from BBU to the RRUs. The users are equipped with a single antenna while RRUs are equipped with $N \gg 1$ antennas. The users associate statically and uniquely to the RRUs. Let $\mathcal{K}_l$ and l_k respectively denote the set of users associated to RRU l, for $l = 1, \ldots, L$, and the RRU to which user k is associated to. We shall consider different association rules in Section IV.

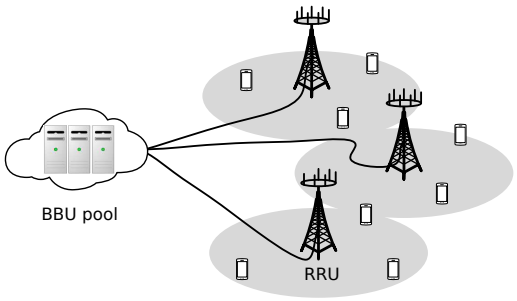

Fig. 1. Illustration of a C-RAN system with fronthaul connecting BBU pool to massive MIMO-based RRUs, which serves multiple users in the downlink.

The system operates in time division duplex mode and channel reciprocity is assumed. We consider a coherence interval of T_c symbols and $T_p < T_c$ symbols are used for uplink training. Then, $\tau = \kappa(1 - T_p/T_c)$ is the fraction of symbols available for downlink data transmission, where $0 \leq \kappa \leq 1$ determines the partitioning of data symbols between uplink and downlink. Let $\mathbf{h}_{jk} \in \mathcal{C}^N$ denote the complex baseband channel gain vector between RRU j and user k. It is modeled as a zero mean complex Gaussian random vector with covariance matrix $\beta_{jk}\mathbf{I}_N$, where β_{jk} is the large-scale fading coefficient that models pathloss and shadowing[1].

A. Uplink Channel Estimation

In order to estimate the uplink channel, users transmit orthogonal pilot sequences of length T_p to the RRUs. The pilot sequences are shared by the users since the number of pilot sequences is typically less than the number of users. Let b_k denote the pilot sequence used by user k and $\mathcal{B}_k$ denote the set of users that use the pilot sequence b_k. As shown in [4], the minimum mean square estimate $\hat{\mathbf{h}}_{jk}$ of $\mathbf{h}_{jk}$ is given by $\hat{\mathbf{h}}_{jk} = \beta_{jk}\tilde{\mathbf{h}}_{jb_k}$, where $\tilde{\mathbf{h}}_{jb_k}$ is as follows:

$$\tilde{\mathbf{h}}_{jb_k} = \frac{\sum_{i \in \mathcal{B}_k} \mathbf{h}_{ji} + \sqrt{\sigma^2/T_p P_{\text{tr}}}\mathbf{n}_{jb_k}}{\sum_{i \in \mathcal{B}_k} \beta_{ji} + \sigma^2/T_p P_{\text{tr}}}, \tag{1}$$

where $\mathbf{n}_{jb_k}$ is an independent and identical complex Gaussian random vector of unit variance, σ^2 is the thermal noise variance, and P_{tr} is the uplink transmit power. Further, $\hat{\mathbf{h}}_{jk}$ is a complex Gaussian random vector with covariance matrix $\theta_{jk}\mathbf{I}_N$, where $\theta_{jk} = \beta_{jk}^2 / \left(\sum_{i \in \mathcal{B}_k} \beta_{ji} + \sigma^2/(T_p P_{\text{tr}})\right)$.

B. Downlink Data Transmission

The estimated uplink channel gains are used to generate the downlink precoding vectors. Let $\mathbf{f}_k \in \mathbb{C}^N$ and p_k denote the

[1]Although we have ignored spatial correlation, extension to the case with spatial correlation is possible through deterministic equivalents [14].

precoder and power applied to the data s_k intended for user k by RRU j, respectively. Then, the transmitted signal $\mathbf{z}_j$ of RRU j is $\mathbf{z}_j = \sum_{k\in\mathcal{K}_j} s_k p_k \mathbf{f}_k$. The power of $\mathbf{z}_j$ should be below the power budget of RRU j. Without loss of generality, we assume the power budget P_{tx} to be the same for all the RRUs. The users' data are uncorrelated and are normalized such that $\mathbb{E}\left[|s_k|^2\right] = 1$, for all $k \in \mathcal{K}_j$. Then, the power constraint of RRU j, for $j = 1, \ldots, L$, is as follows:

$$\mathbb{E}\left[\mathbf{z}_j^H \mathbf{z}_j\right] = \sum_{k\in\mathcal{K}_j} p_k \mathbb{E}\left[\mathbf{f}_k^H \mathbf{f}_k\right] \leq P_{\text{tx}}. \tag{2}$$

The received signal y_k at user k is $y_k = \sum_{l=1}^{L} \mathbf{h}_{lk}^H \mathbf{z}_l + n_k$, where n_k is circularly symmetric complex Gaussian noise with variance σ^2. An achievable rate R_k is $R_k = \tau \log_2(1+\gamma_k)$, where SINR γ_k is given by [4]

$$\gamma_k = \frac{p_k \left|\mathbb{E}\left[\mathbf{h}_{l_k k}^H \mathbf{f}_k\right]\right|^2}{\sigma^2 + \sum_{i=1}^{K} p_i \mathbb{E}\left[\left|\mathbf{h}_{l_i k}^H \mathbf{f}_i\right|^2\right] - p_k \left|\mathbb{E}\left[\mathbf{h}_{l_k k}^H \mathbf{f}_k\right]\right|^2}. \tag{3}$$

This SINR expression is widely used in the massive MIMO literature [4], [9], [14] and is only a function of large-scale fading coefficients.

We consider MRT and ZF precoding schemes. For user k, the precoding vector with MRT is $\mathbf{f}_k = \hat{\mathbf{h}}_k / \sqrt{\mathbb{E}[||\hat{\mathbf{h}}_k||^2]}$ and with ZF is $\mathbf{f}_k = \mathbf{G}_l \mathbf{g}_{b_k} / \sqrt{\mathbb{E}\left[||\mathbf{G}_l \mathbf{g}_{b_k}||^2\right]}$. Note that the precoding vectors are normalized so that the average power is unity. Here, $\mathbf{g}_{b_k}$ is the b_k^{th} column of the matrix $\left(\mathbf{G}_l^H \mathbf{G}_l\right)^{-1}$ with $\mathbf{G}_l = [\tilde{\mathbf{h}}_{l1}, \ldots, \tilde{\mathbf{h}}_{lT_p}]$. Under perfect channel estimation, this ZF precoding scheme cancels out interference even to users not associated with RRU l.

The SINR expression in (3) can be evaluated in closed-form for MRT and ZF precoders. They have a similar form [4]:

$$\gamma_k = \frac{v p_k \theta_{l_k k}}{\sum_{i=1}^{K} p_i \left(w_{l_i k} + v\theta_{l_i k} s_{ik}\right) - v p_k \theta_{l_k k} + \sigma^2}, \tag{4}$$

where $s_{ik} = 1$ if $i \in \mathcal{B}_k$ and 0 otherwise. Here, v and $w_{l_i k}$ depend on the precoding scheme. For MRT, $v = N$ and $w_{l_i k} = \beta_{l_i k}$, and for ZF, $v = N - T_p$ and $w_{l_i k} = \beta_{l_i k} - \theta_{l_i k}$.

The SINR expression in (4) is quite intuitive. The numerator term represents the received signal power while $\sum_{i=1, i\neq k}^{K} v\theta_{l_i k} s_{ik} p_i$ and $\sum_{i=1}^{K} w_{l_i k} p_i$ in the denominator represent the interference due to pilot contamination and the multi-user interference, respectively. Note that the interference due to pilot contamination increases with N similar to the numerator. Thus, pilot contamination becomes the limiting factor asymptotically in N. Comparing the SINR expressions of MRT and ZF, we see that the array gain v is higher for MRT while the multi-user interference gain $w_{l_i k}$ is lower for ZF. This is because ZF attempts to nullify the multi-user interference, while MRT tries to maximize the signal-to-noise-ratio. Since the SINR expression in (4) depends only on the large-scale fading coefficients β_{jk} and not on the instantaneous channel realizations $\mathbf{h}_{jk}$, power control needs to be carried out only at a slower time scale. This reduces the system complexity and the computational requirements at the BBU.

Fronthaul Constraints: These constraints ensure that the data rate provided by each RRU in the downlink is less than the capacity C_{FH} of their fronthaul link. We assume C_{FH} to be the same for all the fronthaul links. Let η denote the ratio of fronthaul bandwidth to downlink bandwidth. Then, we have

$$\tau \sum_{k\in\mathcal{K}_l} \log_2(1+\gamma_k) \leq \eta C_{\text{FH}}, \text{ for } l = 1, \ldots, L. \tag{5}$$

C. Power Consumption Model

The power consumed in a C-RAN system has static and dynamic power components. The static component P_S involves the power consumed by users during uplink training and data transmission, the circuit power in RRUs, and the power expended in fronthaul. The power consumed by a user during uplink training and data transmission is $P_{\text{UE}} = (1-\tau)P_{\text{tr}}/\omega_{\text{UE}}$, where ω_{UE} is the power amplifier efficiency of the user equipment. As in [9], the circuit power in massive MIMO is given by $\varrho + N\varsigma$. Here, ς models the power components that scales with the number of antennas, which involves the power consumed by converters, mixers, and filters in each antenna branch and ϱ models the power components that do not scale with the number of antennas. The power expended in fronthaul P_{FH} is assumed to be fixed as has been assumed in [6], [8]. This is justified as fronthaul is never idle and P_{FH} does not depend on the precoding vectors as has been assumed in [5], [15] since precoding is done at RRU. Thus, the static component is $P_S = KP_{\text{UE}} + L(\varrho + N\varsigma) + P_{\text{FH}}$.

The dynamic component depends on the signal transmitted in the downlink and is given by $(\tau/\omega_{\text{RRU}})\sum_{i=1}^{K} p_i$, where ω_{RRU} is the power amplifier efficiency of RRUs. Therefore, the totalevaluated. We then transform $\tilde{\boldsymbol{p}}$ in $B_k(\tilde{\boldsymbol{p}})$ back to $\boldsymbol{p}$. The difference between the bounds for $A_k(\boldsymbol{p})$ and $B_k(\boldsymbol{p})$ yields $H_k\left(\boldsymbol{p}; \boldsymbol{p}^{(r)}\right)$. Since this bound follows from concavity and convexity of $A_k(\boldsymbol{p})$ and $B_k(\tilde{\boldsymbol{p}})$, it satisfies the properties in [16, Prop. 4.2].

For $G_k\left(\boldsymbol{p}; \boldsymbol{p}^{(r)}\right)$, we obtain a lower bound on $A_k(\boldsymbol{p})$ and an upper bound on $B_k(\boldsymbol{p})$ similar to above. We skip the details. power consumed is $P_{\text{C-RAN}} = P_S + (\tau/\omega_{\text{RRU}})\sum_{i=1}^{K} p_i$.

III. EE Maximization

In this section, we focus on maximizing the EE of the C-RAN system. We first formulate the optimization problem and then develop an SCA algorithm.

Problem Formulation: Using the above power consumption model, EE is defined as $\text{EE} = \sum_{k=1}^{K} R_k / P_{\text{C-RAN}}$. Let $\boldsymbol{p} = [p_1, \ldots, p_K]$. Then, the EE maximization problem subject to transmit power and fronthaul capacity constraints is as follows:

$$\mathbf{P}1: \max_{\boldsymbol{p}\geq\mathbf{0}} \frac{\tau \sum_{k=1}^{K} \log(1+\gamma_k)}{P_S + (\tau/\omega_{\text{RRU}})\sum_{k=1}^{K} p_k}, \tag{6}$$

$$\text{s.t.} \sum_{k\in\mathcal{K}_l} \log(1+\gamma_k) \leq \tilde{\eta} C_{\text{FH}}, \tag{7}$$

$$\sum_{k\in\mathcal{K}_l} p_k \leq P_{\text{tx}}, \text{ for } l = 1, \ldots, L, \tag{8}$$

where $\tilde{\eta}=\eta\log(2)/\tau$. This problem has non-convex objective function and non-convex constraints in (7). Hence, global solutions are hard to find. We develop an iterative algorithm that is guaranteed to converge to a local optimum.

Proposed SCA Algorithm: In order to apply the SCA algorithm, we need a locally tight lower bound and an upper bound, which are convex functions of $\boldsymbol{p}$, for the non-convex function $\log\left(1+\gamma_k\right)$ as it appears in (6) and in (7). For this, we have the following key result.

Result *1:* For any two non-negative power vectors $\boldsymbol{p}$ and $\boldsymbol{p}^{(r)}$, we have $\log\left(1+\gamma_k\right)\geq G_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$ and $\log\left(1+\gamma_k\right)\leq H_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$, where $G_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$ and $H_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$ are

$$\begin{aligned}G_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)&=\log\left(1+\gamma_k^{(r)}\right)\\&-\sum_{i=1}^{K}\left[\frac{\left(p_i-p_i^{(r)}\right)\left(w_{l_ik}+v\theta_{l_ik}s_{ik}I_{\{i\neq k\}}\right)}{\sum_{n=1}^{K}p_n^{(r)}\left(w_{l_nk}+v\theta_{l_nk}s_{nk}\right)-vp_k^{(r)}\theta_{l_kk}+\sigma^2}\right.\\&\left.-\frac{\left(\log(p_i)-\log(p_i^{(r)})\right)\left(w_{l_ik}+v\theta_{l_ik}s_{ik}\right)}{\sum_{n=1}^{K}p_n^{(r)}\left(w_{l_nk}+v\theta_{l_nk}s_{nk}\right)+\sigma^2}\right],\end{aligned}\tag{9}$$

$$\begin{aligned}H_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)&=\log\left(1+\gamma_k^{(r)}\right)\\&+\sum_{i=1}^{K}\left[\frac{\left(p_i-p_i^{(r)}\right)\left(w_{l_ik}+v\theta_{l_ik}s_{ik}\right)}{\sum_{n=1}^{K}p_n^{(r)}\left(w_{l_nk}+v\theta_{l_nk}s_{nk}\right)+\sigma^2}\right.\\&\left.-\frac{\left(\log(p_i)-\log(p_i^{(r)})\right)\left(w_{l_ik}+v\theta_{l_ik}s_{ik}I_{\{i\neq k\}}\right)}{\sum_{n=1}^{K}p_n^{(r)}\left(w_{l_nk}+v\theta_{l_nk}s_{nk}\right)-vp_k^{(r)}\theta_{l_kk}+\sigma^2}\right],\end{aligned}\tag{10}$$

where $\gamma_k^{(r)}$ is SINR of user k at $\boldsymbol{p}^{(r)}$. Further, $G_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$ and $H_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$ satisfy the properties in [16, Prop. 4.2].

Proof: The proof is relegated to Appendix VI-A. ∎

We note that these bounds differ from those in [17] and they yield simpler convex problems as we show below. Using the above bounds, we get the following convex approximation for $\mathbf{P}1$ at $\boldsymbol{p}^{(r)}$ in the r^{th} SCA iteration:

$$\mathbf{P}2:\ \min_{\boldsymbol{p}\geq\mathbf{0}}\ -\frac{\sum_{i=1}^{K}G_i\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)}{P_S+(\tau/\omega_{\text{RRU}})\sum_{i=1}^{K}p_i},\tag{11}$$

$$\text{s.t. (8)},\ \sum_{k\in\mathcal{K}_l}H_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)\leq\tilde{\eta}C_{\text{FH}},\ l=1,\ldots,L.\tag{12}$$

Since $\mathbf{P}2$ involves the minimization of the ratio of a convex and an affine function over a convex set, its optimal solution can be found by the DB algorithm. In each iteration of it, the difference between the numerator and scaled denominator of the objective function in (11) is solved subject to constraints (8) and (12). The scaling factor is the value of the objective function in (11) evaluated at the solution from the previous DB iteration. Thus, the minimization problem solved in the m^{th} DB iteration given the solution $\boldsymbol{p}^*\left(q_{m-1},\boldsymbol{p}^{(r)}\right)$ of the $(m-1)^{\text{th}}$ iteration is as follows:

$$\mathbf{P}3:\ \min_{\{\boldsymbol{p}\geq\mathbf{0}\}}\ -\sum_{i=1}^{K}G_i\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)-\frac{q_m\tau}{\omega_{\text{RRU}}}\sum_{i=1}^{K}p_i,\ \text{s.t. (8), (12),}$$

where

$$q_m=-\frac{\sum_{i=1}^{K}G_i\left(\boldsymbol{p}^*\left(q_{m-1},\boldsymbol{p}^{(r)}\right);\boldsymbol{p}^{(r)}\right)}{P_S+(\tau/\omega_{\text{RRU}})\sum_{i=1}^{K}p_i^*\left(q_{m-1},\boldsymbol{p}^{(r)}\right)}.\tag{13}$$

The problem $\mathbf{P}3$ can be solved by the Lagrange multipliers method. Let $\boldsymbol{\lambda}=[\lambda_1,\ldots,\lambda_L]\geq\mathbf{0}$ denote the Lagrange multipliers associated with the constraints in (12). Then, the Lagrange function $\mathcal{L}\left(\boldsymbol{p},\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)$ is given by

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{p},\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)&=-\sum_{i=1}^{K}G_i\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)-\frac{q_m\tau}{\omega_{\text{RRU}}}\sum_{i=1}^{K}p_i\\&+\sum_{l=1}^{L}\lambda_l\left(\sum_{k\in\mathcal{K}_l}H_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)-\tilde{\eta}C_{\text{FH}}\right).\end{aligned}\tag{14}$$

Let $\mathcal{G}\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)$ denote the dual function. It is then given by $\mathcal{G}\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)=\min_{\mathcal{P}}\mathcal{L}\left(\boldsymbol{p},\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)$, where $\mathcal{P}=\left\{\boldsymbol{p}\geq\mathbf{0},\sum_{k\in\mathcal{K}_l}p_k\leq P_{\text{tx}},\forall l\right\}$. Let $p_i^*\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)$, for $i=1,\ldots,K$, denote the optimal power that minimizes $\mathcal{L}\left(\boldsymbol{p},\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)$. From the KKT conditions, which are necessary and sufficient in this case, it can be shown to be

$$p_i^*\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)=\left[\frac{p_i^{(r)}U_i\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)}{\mu_{l_i}-\frac{q_m\tau}{\omega_{\text{RRU}}}+V_i\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)}\right]^+,\tag{15}$$

where μ_{l_i} is chosen to ensure that the power constraint is satisfied. Since $p_i^*\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)$ is non-increasing in μ_{l_i}, bisection method can be used to efficiently compute μ_{l_i}. Here, $U_i\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)$ and $V_i\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)$ are given by

$$\begin{aligned}U_i\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)&=\sum_{k=1}^{K}\frac{\alpha_k\left(w_{l_ik}+v\theta_{l_ik}s_{ik}\right)}{\sum_{n=1}^{K}p_n^{(r)}\left(w_{l_nk}+v\theta_{l_nk}s_{nk}\right)+\sigma^2}\\&+\frac{\lambda_{l_k}\left(w_{l_ik}+v\theta_{l_ik}s_{ik}I_{\{i\neq k\}}\right)}{\sum_{n=1}^{K}p_n^{(r)}\left(w_{l_nk}+v\theta_{l_nk}s_{nk}\right)-vp_k^{(r)}\theta_{l_kk}+\sigma^2},\end{aligned}\tag{16}$$

$$\begin{aligned}V_i\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)&=\sum_{k=1}^{K}\frac{\lambda_{l_k}\left(w_{l_ik}+v\theta_{l_ik}s_{ik}\right)}{\sum_{n=1}^{K}p_n^{(r)}\left(w_{l_nk}+v\theta_{l_nk}s_{nk}\right)+\sigma^2}\\&+\frac{\alpha_k\left(w_{l_ik}+v\theta_{l_ik}s_{ik}I_{\{i\neq k\}}\right)}{\sum_{n=1}^{K}p_n^{(r)}\left(w_{l_nk}+v\theta_{l_nk}s_{nk}\right)-vp_k^{(r)}\theta_{l_kk}+\sigma^2}.\end{aligned}\tag{17}$$

The dual problem is now given by $\max_{\boldsymbol{\lambda}\geq\mathbf{0}}\mathcal{G}\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)$, where $\mathcal{G}\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)=\mathcal{L}\left(\boldsymbol{p}^*\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right),\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)$. Since $\mathcal{G}\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)$ is a concave function of $\boldsymbol{\lambda}$, the dual problem can be solved via gradient-ascent methods [18]. While $\mathcal{G}\left(\boldsymbol{\lambda};q_m,\boldsymbol{p}^{(r)}\right)$ may not be differentiable, we can compute a subgradient as $\mathbf{d}\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)=\left[d_1\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right),\ldots,d_L\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)\right]$, where $d_l\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)$, for $l=1,\ldots,L$, is given by [19]

$$d_l\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right)=\sum_{k\in\mathcal{K}_l}H_k\left(\boldsymbol{p}^*\left(\boldsymbol{\lambda};\boldsymbol{p}^{(r)}\right);\boldsymbol{p}^{(r)}\right)-\tilde{\eta}C_{\text{FH}}.\tag{18}$$

The dual variables are now updated as follows:

$$\boldsymbol{\lambda}_{t+1}=\left[\boldsymbol{\lambda}_t+\delta_t\mathbf{d}\left(\boldsymbol{\lambda}_t;\boldsymbol{p}^{(r)}\right)\right]^+,\tag{19}$$

where δ_t is the step-size for iteration t. The dual variables $\boldsymbol{\lambda}_t$ converges to the optimal variables $\boldsymbol{\lambda}^*$ as $t\to\infty$. Since the

Algorithm 1: Maximizing EE with Fronthaul Constraints

Set $r = 0$. Initialize with a feasible power vector $\boldsymbol{p}^{(r)}$
repeat [SCA algorithm]
  Set $m = 0$. Initialize $q_m = 0$
  repeat [DB algorithm]
    repeat [Dual subgradient algorithm]
      Compute $\boldsymbol{p}^*\left(\boldsymbol{\lambda}; q_m, \boldsymbol{p}^{(r)}\right)$ using (15)
      Update the dual variables according to (19)
    until: relative change in dual objective is below ϵ
    Update q_m using (13) and $m = m + 1$
  until: relative change in (11) is less than ϵ
  Update $\boldsymbol{p}^{(r)} = \boldsymbol{p}^*\left(\boldsymbol{\lambda}^*; q^*, \boldsymbol{p}^{(r)}\right)$ and $r = r + 1$
until: relative change in EE is less than ϵ

duality gap for $\mathbf{P}3$ is zero and the dual solution is unique, the optimal solution to $\mathbf{P}3$ is given by $\boldsymbol{p}^*\left(\boldsymbol{\lambda}^*; q_m, \boldsymbol{p}^{(r)}\right)$. It is then used to compute q_{m+1} using (13). The DB algorithm is guaranteed to converge monotonically. Let q^* denote the limit of q_m. Then, the solution to $\mathbf{P}2$ is given by $\boldsymbol{p}^*\left(\boldsymbol{\lambda}^*; q^*, \boldsymbol{p}^{(r)}\right)$.

In the $(r+1)^{\text{th}}$ iteration of the SCA algorithm, the problem $\mathbf{P}2$ is solved with $\boldsymbol{p}^{(r)}$ replaced by $\boldsymbol{p}^{(r+1)} = \boldsymbol{p}^*\left(\boldsymbol{\lambda}^*; q^*, \boldsymbol{p}^{(r)}\right)$. This is repeated until EE converges. The overall SCA algorithm for EE maximization is summarized in Algorithm 1.

Convergence and Complexity: Since the bounds in Result 1 satisfy the properties in [16, Prop. 4.2], EE monotonically increases and the power updates converge to a KKT point. We note that we apply the DB algorithm differently from that in [8], where the DB algorithm is applied directly to the non-convex EE maximization problem. Therefore, the convergence of the DB algorithm is not guaranteed in [8] as it does not ensure that the non-convex problems in each DB iteration is solved globally. This is unlike our approach in which convex problems are solved in each DB iteration.

The computational complexity of each iteration of dual subgradient algorithm is $\mathcal{O}(K^3)$ and the number of iterations needed is dependent on the step size δ_t and the initialization point $\boldsymbol{\lambda}_0$. We have observed that the optimal dual variables for the previous iteration of the SCA algorithm is a good initialization point for the dual subgradient algorithm. Since the DB algorithm has super-linear convergence [8], it needs very few iterations to converge.

IV. Numerical Results

We carry out Monte Carlo simulations to evaluate the performance of the proposed SCA algorithms. We consider a 7-cell hexagonal cellular layout with wrap-around. The cell radius is 500 m. We drop $K = 70$ users randomly in the network area. The pathloss in dB at distance d in km is $128.1 + 3.76\log_{10}(d)$. The standard deviation of lognormal shadowing is 8 dB. The other simulation parameters are $N = 200$, $T_{\text{c}} = 200$, $T_p = 10$, $P_{\text{tr}} = 23$ dBm, $P_{\text{dl}} = 46$ dBm, $\varsigma = 0.2$ W, $\varrho = 1.8$ W, $\omega_{\text{RRU}} = \omega_{\text{UE}} = 0.3$, $\kappa = \eta = 1$, and $\epsilon = 0.01$. The signal bandwidth is 10 MHz. The pilot sequences are reused in each cell and they are allocated randomly to the users.

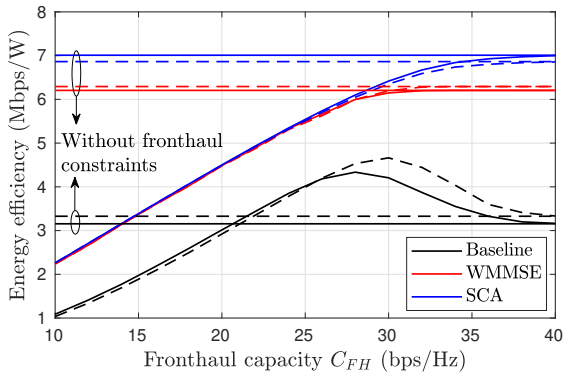

Fig. 2. MRT precoding: EE as a function of C_{FH} for different schemes. The solid and dashed lines denote signal power- and distance-based association rules, respectively.

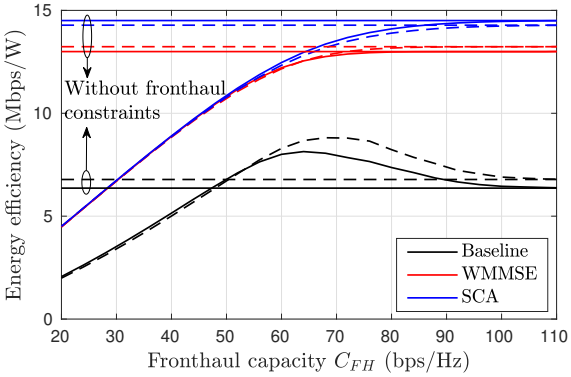

Fig. 3. ZF precoding: EE as a function of C_{FH} for different schemes. The solid and dashed lines denote signal power- and distance-based association rules, respectively.

We consider distance-based and signal power-based [4] association rules. In the former, a user associates to the RRU that is closest to it. In the latter, a user associates to the RRU from which it receives the highest signal power when transmitting at full power. Thus, the association here is based on both pathloss and shadowing. We also note that it is not straightforward to implement the commonly considered *max-SINR* association rule, in which a user associates to the RRU that provides the highest SINR. This is because SINR depends on pilot allocation, which, in turn, depends on user association.

We benchmark the proposed SCA algorithms against a baseline scheme and an extension of the algorithm in [13]. In the baseline scheme, there is no coordination between the RRUs to maximize the network performance. The RRUs transmit at equal power and the power value is chosen to ensure that the fronthaul constraints are satisfied. We employ the DB algorithm as used in [8] to extend the WMMSE algorithm in [13] for sum rate maximization to maximize EE. However, this algorithm lacks any convergence guarantee and we set a limit of 100 on the maximum number of iterations.

Figures 2 and 3 show the EE results with per-link capacity constraints for the MRT and ZF precoding schemes, respectively. We see that the proposed SCA algorithm achieves a

significantly higher EE than the baseline scheme for both the association rules. For example, the SCA algorithm improves EE by 49%, 53%, and 122% for $C_{\text{FH}} = 20$, 30, and 40 bps/Hz, respectively, with signal power-based association and MRT precoding. The corresponding gains for ZF precoding at $C_{\text{FH}} = 40$, 60, and 80 bps/Hz are 73%, 56%, and 91%, respectively. While the WMMSE algorithm achieves a similar EE as the SCA algorithm at low fronthaul capacities, the SCA algorithm achieves a higher EE at high fronthaul capacities. For example, the SCA algorithm improves the EE by 12% at $C_{\text{FH}} = 100$ over the WMMSE algorithm for ZF precoding.

For SCA and WMMSE algorithms, EE increases as C_{FH} increases before saturating, in which case EE is determined only by the power constraints. For the baseline scheme, however, EE first increases and then decreases. This is because the RRUs transmit at higher powers as C_{FH} increases. However, beyond a certain point, this increase in transmit power does not translate to an increase in sum rate. This causes the EE to decrease. We also see that the difference in EE with the two association rules is more pronounced for the baseline scheme than the SCA and WMMSE algorithms. Thus, with power control, the initial user association is not too critical. Notice that ZF precoding achieves a higher EE than MRT precoding since the former manages the interference better than the latter.

V. Conclusions

We developed an SCA-based optimization framework to maximize EE of a C-RAN system with massive MIMO-based RRUs and capacity constrained fronthaul. The framework employed a novel upper bound and a lower bound on the user rate function to obtain a convex approximation for the non-convex EE maximization problem. The convex problem is then solved through DB algorithm and dual decomposition in each SCA iteration. The SCA algorithm is guaranteed to converge to a KKT point. We observed that the proposed algorithm significantly improved EE compared to the case with no power control. Further, it achieved an EE comparable to or better than the WMMSE algorithm for all values of fronthaul capacity.

VI. Appendix

A. Derivation of $G_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$ and $H_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$

We first derive the upper bound $H_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$. We express user rate $\log\left(1+\gamma_k\right)$ as $\log\left(1+\gamma_k\right) = A_k(\boldsymbol{p}) - B_k(\boldsymbol{p})$, where $A_k(\boldsymbol{p}) = \log\left(\sum_{n=1}^{K} p_n\left(w_{l_nk} + v\theta_{l_nk}s_{nk}\right) + \sigma^2\right)$ and $B_k(\boldsymbol{p}) = \log\left(\sum_{n=1}^{K} p_n\left(w_{l_nk} + v\theta_{l_nk}s_{nk}\right) - vp_k\theta_{l_kk} + \sigma^2\right)$.

We now provide an upper bound for $A_k(\boldsymbol{p})$ and a lower bound for $B_k(\boldsymbol{p})$. Since $A_k(\boldsymbol{p})$ is concave in $\boldsymbol{p}$, for any $\boldsymbol{p}^{(r)}$, we have $A_k(\boldsymbol{p}) \le A_k\left(\boldsymbol{p}^{(r)}\right) + \sum_{i=1}^{K}(p_i - p_i^{(r)})\frac{d}{dp_i}A_k(\boldsymbol{p})|_{\boldsymbol{p}^{(r)}}$.

For $B_k(\boldsymbol{p})$, we use the variable transformation $\tilde{p}_i = \log(p_i)$, for $i = 1,\ldots,K$. Since $B_k(\tilde{\boldsymbol{p}})$ is the composition of affine and log-sum-exp functions, it is convex in $\tilde{\boldsymbol{p}}$ [18]. Thus, we have $B_k(\tilde{\boldsymbol{p}}) \ge B_k\left(\tilde{\boldsymbol{p}}^{(r)}\right) + \sum_{i=1}^{K}(\tilde{p}_i - \tilde{p}_i^{(r)})\frac{d}{d\tilde{p}_i}B_k(\tilde{\boldsymbol{p}})|_{\tilde{\boldsymbol{p}}^{(r)}}$.

The derivatives in the above bounds can be easily evaluated. We then transform $\tilde{\boldsymbol{p}}$ in $B_k(\tilde{\boldsymbol{p}})$ back to $\boldsymbol{p}$. The difference between the bounds for $A_k(\boldsymbol{p})$ and $B_k(\boldsymbol{p})$ yields $H_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$. Since this bound follows from concavity and convexity of $A_k(\boldsymbol{p})$ and $B_k(\tilde{\boldsymbol{p}})$, it satisfies the properties in [16, Prop. 4.2].

For $G_k\left(\boldsymbol{p};\boldsymbol{p}^{(r)}\right)$, we obtain a lower bound on $A_k(\boldsymbol{p})$ and an upper bound on $B_k(\boldsymbol{p})$ similar to above. We skip the details.

References

[1] ITU-R Rec. M.2083-0, "IMT vision - framework and overall objectives of the future development of IMT for 2020 and beyond," Tech. Rep., Sep. 2015.

[2] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[3] J. Bartelt, N. Vucic, D. Camps-Mur, E. Garcia-Villegas, I. Demirkol, A. Fehske, M. Grieger, A. Tzanakaki, J. Gutiérrez, E. Grass, G. Lyberopoulos, and G. Fettweis, "5G transport network requirements for the next generation fronthaul interface," *European J. Wireless Commun. Netw.*, vol. 2017, no. 1, p. 89, May 2017.

[4] T. V. Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and user association optimization for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6384–6399, Sep. 2016.

[5] K. G. Nguyen, Q. D. Vu, M. Juntti, and L. N. Tran, "Energy efficient preceding C-RAN downlink with compression at fronthaul," in *Proc. ICC*, May 2017, pp. 1–6.

[6] M. Peng, Y. Yu, H. Xiang, and H. V. Poor, "Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 879–892, May 2016.

[7] P. Luong, C. Despins, F. Gagnon, and L. N. Tran, "Designing green C-RAN with limited fronthaul via mixed-integer second order cone programming," in *Proc. ICC*, May 2017, pp. 1–6.

[8] 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.

[9] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3075, Jun. 2015.

[10] Y. Hao, Q. Ni, H. Li, and S. Hou, "Energy and spectral efficiency tradeoff with user association and power coordination in massive MIMO enabled HetNets," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2091–2094, Oct. 2016.

[11] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Large intelligent surfaces for energy efficiency in wireless communication," 2018. [Online]. Available: http://arxiv.org/abs/1810.06934

[12] S. Parsaeefard, R. Dawadi, M. Derakhshani, T. Le-Ngoc, and M. Baghani, "Dynamic resource allocation for virtualized wireless networks in massive-MIMO-aided and fronthaul-limited C-RAN," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9512–9520, Oct. 2017.

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

[14] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[15] 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.

[16] A. Zappone and E. Jorswieck, *Energy Efficiency in Wireless Networks via Fractional Programming Theory*. Now Foundations and Trends, 2015.

[17] T. Wang and L. Vandendorpe, "Successive convex approximation based methods for dynamic spectrum management," in *Proc. ICC*, Jun. 2012, pp. 4061–4065.

[18] S. Boyd and L. Vandenberge, *Convex Optimization*. Cambridge University Press, 2003.

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