Energy-Efficient Group-Sparse Transceiver Design for Multiuser MIMO Relaying in C-RAN

Jiaxin Yang, Ayoub Saab, Alireza Morsali, *Student Member, IEEE*, Benoit Champagne, *Senior Member, IEEE*, and Ioannis Psaromiligkos, *Member, IEEE*

***Abstract*—This paper addresses the problem of centralized transceiver design for multiuser MIMO amplify-and-forward (AF) relaying within a cloud radio access network (C-RAN). The aim is to optimize AF matrices of remote radio heads (RRHs) acting as relays, in order to improve the reception quality at the destinations while reducing network power consumption and feedback overhead on the fronthaul links. A two-stage method is proposed to solve this problem efficiently. The first stage relies on interference leakage minimization subject to per-relay transmit power constraints along with signal preserving constraints. To reduce the total network power, RRH selection is achieved by incorporating in the objective function a regularization term that promotes group-sparsity among the RRHs. In the second stage, to reduce feedback overhead, a different penalty term is added that induces weight-level sparsity in the AF matrix of each active RRH. For both stages, low-complexity iterative algorithms based on the alternating direction method of multipliers (ADMM) are developed to solve the corresponding regularized problems with low complexity. Extensive simulations are performed to demonstrate the explicit benefits of the proposed design method, which results in notably lower power consumption, computational complexity and weight feedback overhead than conventional approaches.**

***Index Terms*—ADMM, amplify-and-forward, cloud-RAN, group-sparsity, multiuser MIMO, RRH, wireless relaying.**

I. Introduction

NETWORK densification along with massive multiple-input multiple-output (MIMO) processing are widely recognized as key enabling technologies to meet the exacting performance targets for fifth generation (5G) wireless networks [1]–[4]. To fully reap the benefits of these new technologies, however, advanced interference coordination and resource allocation schemes need to be employed. The low-cost implementation of these sophisticated schemes will be facilitated by an innovative centralized radio access network (RAN) architecture, called the cloud-RAN (C-RAN) [5].

Manuscript received August 29, 2019; revised February 1, 2020; accepted March 30, 2020. Date of publication April 8, 2020; date of current version August 19, 2020. This work was supported by an NSERC CRD Grant, with Industrial Support from InterDigital Canada. The associate editor coordinating the review of this article and approving it for publication was V. Mancuso. *(Corresponding author: Ioannis Psaromiligkos.)*

The authors are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0E9 Canada (e-mail: jiaxin.yang@mail.mcgill.ca; ayoub.saab@mail.mcgill.ca; alireza.morsali@mail.mcgill.ca; benoit.champagne@mcgill.ca; ioannis.psaromiligkos@mcgill.ca).

Digital Object Identifier 10.1109/TGCN.2020.2986587

In the C-RAN architecture, traditional base station functionalities are apportioned between a centralized pool of baseband units (BBUs) and spatially distributed remote radio heads (RRHs) [6]. The BBUs handle the demanding baseband signal processing functions while the RRHs provide wireless connectivity to the user equipments (UEs). Besides reducing deployment and maintenance costs, C-RAN can improve the network spectral efficiency by exploiting cloud computing to jointly process user data and perform interference coordination. To enable the exchange of vast amounts of data between the BBU pool and RRHs, comprising user data signals, channel state information (CSI), transceiver parameters and control signals, low-latency high-capacity fronthaul links (e.g., optical fibers) must be deployed. The use of powerful BBUs, multiple RRHs and high-speed transport links inevitably introduces additional power consumption across the network [7], which has motivated various research efforts devoted to designing energy-efficient, or "green", C-RAN.

A. Related Works

Multiple-antenna solutions for green C-RAN have been extensively studied in multi-cell downlink (DL) and uplink (UL) setups. A group sparse RRH beamforming framework was proposed in [8] with the aim to minimize the total power consumption in a C-RAN DL multicast scenario. The resulting design method reduces the number of active RRHs, consequently leading to lower network power consumption. The effect of imperfect CSI under the same network setup was addressed in [9] and a robust version of the group sparse beamforming method was proposed. To enhance group sparsity in DL multicast beamforming, a smoothed l_p minimization approach relying on the iterative reweighted l_2 minimization algorithm was developed in [10]. Group sparse beamforming still suffers from high computational complexity, especially in the solution of the RRH ordering criterion needed to determine the active RRHs and associated fronthaul links. In [11], by leveraging on Lagrange duality and random matrix theories, new approaches were proposed to reduce the complexity of this crucial step.

The works in [12]–[14] addressed the joint problem of user-RRH association and RRH DL beamforming design by optimizing some alternative network performance metrics under a set of finite-capacity constraints on the fronthaul or backhaul links. The joint DL and UL network power minimization problem under C-RAN was investigated in [15],

where the UL-DL duality was invoked to derive two near-optimal beamforming solutions, respectively based on group sparse optimization and relaxed integer programming. In [16], the joint problem of RRH association, sub-channel assignment, and power allocation for network sum-rate maximization in single-carrier frequency division multiple access (SC-FDMA)-based multi-tier C-RAN was studied, and a two-step iterative algorithm was conceived to solve the underlying non-linear mixed-integer problem.

A key requirement to fully leverage the performance benefits of C-RAN is the availability of a high capacity fronthaul connecting the RRHs and BBU pool. However, due to the dense deployment of RRHs in C-RAN, the cost for providing such high-speed links may be extremely high, while in certain locations, the installation of new fibers may not be possible. For these reasons, practical C-RAN may rely on fronthaul with limited bandwidth, including dedicated wireless links [5], [7], [17]. Therefore, given fixed fronthaul capacity, minimizing the performance loss due to quantization becomes a primary issue. In this regard, some previous works have focused on the design of robust quantization schemes for the efficient transmission of baseband user signals over fronthaul links in C-RAN, e.g., [18], [19].

The above works on green C-RAN focus mainly on coordinated beamforming design in DL or UL setups, without consideration of relaying strategies. In the context of existing wireless standards such as Evolved Universal Terrestrial Radio Access (E-UTRA) [20], Long-Term Evolution Advanced (LTE-Advanced) [21] and Worldwide Interoperability for Microwave Access (WiMAX) [22], cooperative relaying has proven to be extremely valuable in extending cell coverage, improving link quality, and increasing network capacity [23]. Consequently, it has attracted considerable attention in recent years and is still being studied under a variety of new scenarios, e.g., [24]–[28]. While it is anticipated that similar benefits can be reaped from the use of cooperative MIMO relaying in dense 5G networks[29], the use of RRHs as relays for further improving coverage and performance under C-RAN, along with the associated relay selection and transceiver design algorithms, remain largely undisclosed.

In [30], the joint design of the wireless fronthaul and access links for C-RAN was studied under the assumption that the multi-antenna RRHs use either the decode-and-forward (DF) or the decompress-and-forward (DCF) strategies to relay the user signals from a BBU to the users. In [31], the design of a multiuser amplify-and forward (AF) MIMO relaying subnetwork within C-RAN was investigated from an energy-efficient perspective, and a joint RRH selection and relay transceiver optimization algorithm was devised to minimize the network power subject to mean-square error-based quality-of-service (QoS) constraints. While the AF scheme leads to reduced processing delays and simplified implementation, the resultant block-coordinate descent type iterative algorithm still exhibits high computational complexity. Regarding the issue caused by limited fronthaul capacity, we note that different from the DL and UL scenarios, in a AF relaying scenario under C-RAN, the RRHs will collect their received signals and then process and forward them to the corresponding destination users. Hence, in this case, attention should be focused on the quantization of the optimized RRH transceiver parameters at the BBU pool, which are fed back to the RRHs over the fronthaul links at regular intervals.

B. Contributions and Paper Organization

In this paper, we present a full-fledged extension and study of the concepts introduced in [32], where we investigated the problem of RRH transceiver optimization in a multiuser amplify-and-forward (AF) relaying subnetwork within C-RAN. We consider the scenario where multiple source-destination pairs communicate with the aid of multiple cooperative MIMO RRHs connected to a BBU pool. There are three key aspects in designing practical green relay transceivers for the C-RAN architecture. The major concern in green communications is, of course, reducing the network power consumption. Since the calculated RHH weights are transmitted via fronthaul links, reducing the feedback overhead is another important aspect. Finally, since the BBU pool is simultaneously performing the transceiver design for multiple clusters of source and destination users under mobility condition, the required aggregate computational resources must be minimized by developing low complexity design procedures. Therefore, while focusing on improving the QoS at the destination users, our design approach aims at: 1) Reducing network power consumption by performing RRH selection; 2) Reducing fronthaul feedback overhead by resorting to RRH antenna weight selection and quantization; and 3) Developing efficient design algorithms with low computational complexity.

The novel transceiver design algorithms that we propose for MIMO AF relaying in C-RAN capitalize on sparse signal processing techniques to address these aspects. Specifically, to enforce a small number of active RRHs, *group-level sparsity* is first employed, whereby selected AF matrices (corresponding to inactive RRHs) are collectively zeroed. Subsequently, to reduce fronthaul overhead, a form of *weight-level sparsity* is utilized, whereby the number of non-zero weights associated with each active RRH is reduced, so that more bits can be allocated to the non-zero weight. Based on these two different levels of sparsity, a two-stage RRH transceiver optimization method with low complexity is proposed. Our main contributions are summarized as follows.

- In the first stage, we perform joint RRH selection and transceiver optimization by formulating the problem as a group sparsity-inducing interference leakage minimization subject to per-relay power constraints and linear signal preserving constraints at the destinations. The problem is converted into a form suitable for application of the alternating direction method of multipliers (ADMM) [33]. A simple closed-form solution can be derived for each one of the main ADMM steps, leading to a very low-complexity iterative algorithm.
- In the second stage, given the selected RRHs, we re-optimize their weights by solving a modified interference leakage minimization problem, where the objective function is now penalized with a sparsity-inducing term at the weight level. A low-complexity ADMM-based iterative

algorithm is developed here as well for efficiently solving the joint weight selection and design problem. Based on the statistical properties of the optimized weights, we also present a non-uniform scalar quantization scheme (based on Lloyd-Max algorithm) for the purpose of weight feedback on the fronthaul.

- We perform extensive simulation studies to evaluate the performance of the proposed design method. Results show that the latter can yield a satisfactory QoS level at all destinations while significantly reducing the total network power. When used in conjunction with non-uniform scalar quantization, the weight selection approach allows a significant reduction in the total number of bits needed for weight quantization. Finally, the processing time of the proposed algorithms is significantly reduced as compared to benchmark methods relying on external optimization solvers.

The rest of the paper is organized as follows. The C-RAN-based multiuser relaying system model is introduced in Section II. In Section III, the constrained and regularized interference leakage minimization problem is formulated for joint RRH selection and transceiver optimization, followed by the development of the low-complexity ADMM-based algorithm. In Section IV, the RRH weight selection method is developed by resorting to weight-level regularization; non-uniform scalar quantization is also investigated in this section. Simulation results are presented and discussed in Section V. Finally, we conclude the paper in Section VI.

II. System Model

We consider a multiuser relaying sub-network consisting of L RRHs serving as AF relays and K pairs of source and destination UEs, as depicted in Fig. 1. Each source UE is paired with a single destination UE, both modeled as single-antenna nodes due to their limited processing capabilities and low power budgets. By contrast, the l^{th} RRH, $l \in \mathcal{L} = \{1, 2, \ldots, L\}$, is equipped with $N_l \geq 1$ antennas. A narrowband flat-fading model is assumed for the radio channels between the UEs and RRHs. The CSI is assumed to be known and remain constant within a given transmission interval. There is no direct link between the source and destination UEs. In C-RAN architecture, all RRHs are connected to a central node, namely the BBU pool, whose role is to select and activate a subset of RRHs and design their AF transceiver matrix; the entries of the computed AF matrices are then quantized and fed back to the RRHs via the fronthaul links.

Communication is performed in a two-hop half-duplex manner. During the first hop, the k^{th} source UE, $k \in \mathcal{K} = \{1, 2, \ldots, K\}$, transmits its information symbol s_k, modeled as a zero-mean complex random variable with variance $\mathrm{E}\{|s_k|^2\} = 1$. Let $\mathbf{h}_{lk} \in \mathbb{C}^{N_l \times 1}$ denote the channel vector between the k^{th} source UE and the l^{th} RRH. The latter receives the superposition of the transmitted symbols from each source UE, corrupted by additive noise. This can be written as

$$\mathbf{r}_l = \sum_{k=1}^{K} \mathbf{h}_{lk} s_k + \mathbf{n}_l, \tag{1}$$

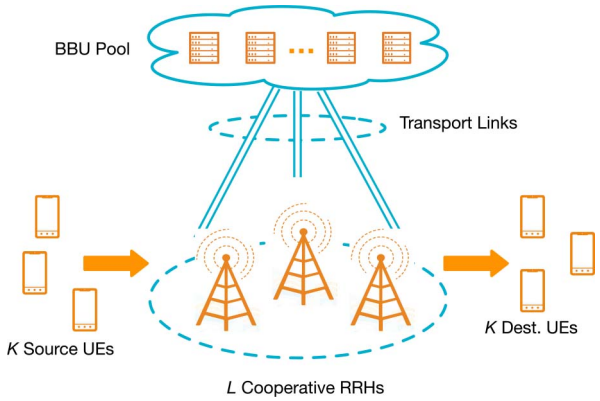


Fig. 1. Multiuser sub-network where communication between the source-destination pairs is assisted by cooperative MIMO relays; under C-RAN, the relay AF matrices are designed at the BBU pool and fed back in quantized form to the RRHs via fronthaul links.

where $\mathbf{n}_l \in \mathbb{C}^{N_l \times 1}$ is a spatially white noise vector, with zero mean and covariance matrix $\mathbf{\Sigma}_l = \sigma_l^2 \mathbf{I}_{N_l}$.

During the second hop, under the AF scheme, the l^{th} RRH first applies a linear transformation to $\mathbf{r}_l$, represented by the matrix $\mathbf{B}_l \in \mathbb{C}^{N_l \times N_l}$, and then transmits the resulting vector. We assume that the average transmission power of the RRH is constrained by an antenna power budget $P_l^{\max}$, that is,[1]

$$P_l^{\mathrm{tr}} = \mathrm{Tr}\left(\mathbf{B}_l \left(\sum_{k=1}^{K} \mathbf{h}_{lk}\mathbf{h}_{lk}^H + \sigma_l^2 \mathbf{I}_{N_l}\right) {\mathbf{B}_l}^H\right) \leq P_l^{\max}. \tag{2}$$

The k^{th} destination UE receives the sum of the transmitted signals from each RRH along with additive noise, which may be expressed as

$$d_k = \sum_{l=1}^{L} \mathbf{g}_{kl}^H \mathbf{B}_l \mathbf{r}_l + n_k = S_k + I_k + n_k, \tag{3}$$

where $\mathbf{g}_{kl}^H \in \mathbb{C}^{1 \times N_l}$ denotes the channel vector between the l^{th} RRH and the k^{th} destination UE, and n_k is an additive noise term with zero-mean and variance ς_k^2. In (3), the superimposed signal from the RRHs is expressed as the sum of two distinct components, i.e., the desired signal S_k and the interference leakage I_k:

$$S_k = \sum_{l=1}^{L} \mathbf{g}_{kl}^H \mathbf{B}_l \mathbf{h}_{lk} s_k \tag{4}$$

$$I_k = \sum_{\substack{j=1, \\ j \neq k}}^{K} \sum_{l=1}^{L} \mathbf{g}_{kl}^H \mathbf{B}_l \mathbf{h}_{lj} s_j + \sum_{l=1}^{L} \mathbf{g}_{kl}^H \mathbf{B}_l \mathbf{n}_l. \tag{5}$$

The objective of the transceiver optimization for multiuser MIMO relaying is to enhance the reception quality of the desired signal S_k at each destination UE subject to relaying power constraints, as specified by (2). Motivated by interference alignment techniques [34], to achieve this objective, we minimize the *total interference leakage* at all destination UEs, defined as $\sum_{k=1}^{K} \mathrm{E}\{|I_k|^2\}$, while enforcing

[1] User symbols and noise terms are assumed to be mutually independent.

a set of linear constraints meant to preserve the integrity of the desired signal S_k. Using (4), these constraints can be stated as

$$\sum_{l=1}^{L} \mathbf{g}_{kl}^H \mathbf{B}_l \mathbf{h}_{lk} = c_k, \quad \forall k \in \mathcal{K}, \tag{6}$$

where c_k are predefined positive constants. Hence, the interference leakage minimization problem can be written as

$$\min_{\{\mathbf{B}_l | l \in \mathcal{L}\}} \ \mathcal{I} \triangleq \sum_{k=1}^{K} \mathrm{E}\{|I_k|^2\} \quad \text{s.t.} \ \ (2) \text{ and } (6), (7)$$

where the total interference leakage can be expressed as

$$\mathcal{I} = \sum_{k=1}^{K} \left(\sum_{j \neq k} \left| \sum_{l=1}^{L} \mathbf{g}_{kl}{}^H \mathbf{B}_l \mathbf{h}_{lj} \right|^2 + \sum_{l=1}^{L} \sigma_l^2 \left\| \mathbf{B}_l^H \mathbf{g}_{kl} \right\|_2^2 \right). \tag{8}$$

In this work, our aim is to solve the above problem within the framework of green C-RAN. Consequently, minimization of the total power consumption is our primary objective, which is addressed in Section III by penalizing the objective function so as to promote group-sparsity among the RRHs. However, to further satisfy the requirements imposed by the C-RAN architecture, minimization of the fronthaul overhead is taken into consideration in Section IV by penalizing the objective function in a different way, so as to promote weight-level sparsity for the active RRHs. For each one of these sub-problems, an efficient algorithm with low-complexity is developed to obtain the desired solution.

III. RRH Selection Scheme

In this section, we first reformulate (7) in a convenient matrix form to simplify later derivations. We then incorporate a regularization term to promote the deactivation of a subset of RRHs and thus obtain an energy-efficient solution. Finally, we propose an ADMM-based algorithm with low complexity to solve the regularized problem.

A. Problem Reformulation

To obtain more compact expressions, we replace each AF matrix by its vectorized version $\mathbf{b}_l = \mathrm{vec}\{\mathbf{B}_l\}$ (obtained by stacking the columns of $\mathbf{B}_l$) and collect the resultant relaying vectors into a global relaying vector $\mathbf{b} \triangleq [\mathbf{b}_1^T, \ldots, \mathbf{b}_L^T]^T$. For ease of presentation, let us rewrite the total interference leakage $\mathcal{I}$ in (8) as $\mathcal{I} = \sum_{k=1}^{K} (\mathcal{I}_{1,k} + \mathcal{I}_{2,k})$. Applying the Kronecker product property $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$, $\mathcal{I}_{1,k}$ can be expressed as

$$\mathcal{I}_{1,k} = \sum_{j \neq k} \left| \sum_{l=1}^{L} \left(\mathbf{h}_{lj}^* \otimes \mathbf{g}_{kl} \right)^H \mathbf{b}_l \right|^2. \tag{9}$$

Defining $\boldsymbol{\delta}_k^{(j)} \triangleq [(\mathbf{h}_{1j}^* \otimes \mathbf{g}_{k1})^T, \ldots, (\mathbf{h}_{Lj}^* \otimes \mathbf{g}_{kL})^T]^T \in \mathbb{C}^{\sum N_l^2 \times 1}$, the above expression further simplifies to

$$\mathcal{I}_{1,k} = \sum_{j \neq k} \left| \mathbf{b}^H \boldsymbol{\delta}_k^{(j)} \right|^2 = \mathbf{b}^H \boldsymbol{\Delta}_k \mathbf{b}, \tag{10}$$

with $\boldsymbol{\Delta}_k \triangleq \sum_{j \neq k} \boldsymbol{\delta}_k^{(j)} \boldsymbol{\delta}_k^{(j)\,H}$. Noting that $\mathbf{I}_{N_l} = \sum_{i=1}^{N_l} \mathbf{e}_i \mathbf{e}_i^H$, where $\mathbf{e}_i$ are the standard basis vectors, $\mathcal{I}_{2,k}$ can be written as

$$\mathcal{I}_{2,k} = \sum_{l=1}^{L} \sum_{i=1}^{N_l} \sigma_l^2 \mathbf{g}_{kl}{}^H \mathbf{B}_l \mathbf{e}_i \mathbf{e}_i^H \mathbf{B}_l^H \mathbf{g}_{kl}. \tag{11}$$

Applying the same Kronecker product property as above, this latter expression simplifies to

$$\mathcal{I}_{2,k} = \sum_{l=1}^{L} \mathbf{b}_l^H \mathbf{G}_{kl} \mathbf{b}_l = \mathbf{b}^H \mathbf{G}_k \mathbf{b} \tag{12}$$

with $\mathbf{G}_{kl} \triangleq \sigma_l^2 \sum_{i=1}^{N_l} (\mathbf{e}_i \otimes \mathbf{g}_{kl})(\mathbf{e}_i \otimes \mathbf{g}_{kl})^H$ and $\mathbf{G}_k \triangleq \mathrm{blkdiag}\{\mathbf{G}_{k1}, \ldots, \mathbf{G}_{kL}\}$.

To complete the reformulation, we apply the property $\mathrm{Tr}(\mathbf{ABA}^H) = \mathrm{vec}(\mathbf{A})^H (\mathbf{B} \otimes \mathbf{I}) \mathrm{vec}(\mathbf{A})$ to the power constraint (2) and use the results in (9) and (12), which leads to

$$\min_{\mathbf{b}} \ \mathbf{b}^H \boldsymbol{\Theta} \mathbf{b} \tag{13a}$$

$$\text{s.t.} \ \ \mathbf{b}_l{}^H \boldsymbol{\Psi}_l \mathbf{b}_l \leq P_l^{\max}, \quad \forall l \in \mathcal{L} \tag{13b}$$

$$\boldsymbol{\Phi}^H \mathbf{b} = \mathbf{c}, \tag{13c}$$

where

$$\boldsymbol{\Theta} \triangleq \sum_{k=1}^{K} \boldsymbol{\Delta}_k + \mathbf{G}_k, \ \boldsymbol{\Psi}_l \triangleq \left(\sum_{k=1}^{K} \mathbf{h}_{lk} \mathbf{h}_{lk}^H + \boldsymbol{\Sigma}_l \right) \otimes \mathbf{I}_{N_l}, \tag{14}$$

$\mathbf{c} \triangleq [c_1, c_2, \ldots, c_K]^T$ and $\boldsymbol{\Phi} \triangleq [\boldsymbol{\phi}_{l,k}] \in \mathbb{C}^{\sum_{l=1}^{L} N_l^2 \times K}$ which is partitioned into $L \times K$ blocks, each given by $\boldsymbol{\phi}_{l,k} = \mathbf{h}_{lk}^* \otimes \mathbf{g}_{kl}$.

Before proceeding to the optimum transceiver design, we emphasize that at this stage of the formulation, nothing prevents any of the RRHs from participating in the relay-assisted transmission, potentially leading to a situation where all RRHs are activated. Hence, the solution to the optimization problem (13) is referred to as a *non-sparse solution*. In the following, we modify (13) by adding a regularization term, with the aim of reducing the number of active RRHs while still providing an acceptable QoS level.

B. Relay Selection via Group-Sparsity

We first note that according to (2) and (13b), the l^{th} RRH being inactive is equivalent to the condition $\|\mathbf{b}_l\|_2 = 0$. Consequently, to have a small subset of active RRHs requires that the solution vector $\mathbf{b}$ be group-sparse [8]. That is, $\mathcal{B} \triangleq [\|\mathbf{b}_1\|_2, \ldots, \|\mathbf{b}_L\|_2]^T$ only contains a reduced number of non-zero elements. This property can be captured by the l_0-"norm" constraint $\|\mathcal{B}\|_0 < L$.

Bearing in mind that our aim is to achieve energy efficiency across the network, a group-sparse solution is desired. Motivated by the widely used least absolute shrinkage and selection operator (LASSO) method in machine learning [35], an efficient way to promote group sparsity during optimization is to penalize the objective function (13a) with an l_1-norm term on $\mathcal{B}$, i.e., $\|\mathcal{B}\|_1 = \sum_{l=1}^{L} \|\mathbf{b}_l\|_2$. More generally, we penalize the objective function by the regularization term

$\sum_{l=1}^{L}\lambda_l\|\mathbf{b}_l\|_2$, where $\lambda_l > 0$ is an adjustable parameter representing the weight given to the l^{th} RRH.

To further simplify the presentation, we define $\mathbf{\Psi} \triangleq \mathrm{blkdiag}\{\mathbf{\Psi}_1,\ldots,\mathbf{\Psi}_L\}$ and $\mathbf{x} \triangleq \mathbf{\Psi}^{1/2}\mathbf{b} = [\mathbf{x}_1^T,\ldots,\mathbf{x}_L^T]^T$ with $\mathbf{x}_l = \mathbf{\Psi}_l^{1/2}\mathbf{b}_l$. Note that $\mathbf{\Psi}_l$ in (14) is non-singular, which implies that $\|\mathbf{x}_l\|_2 = 0$ if and only if $\|\mathbf{b}_l\|_2 = 0$. Hence, both vectors $\mathbf{x}$ and $\mathbf{b}$ share the same group-sparsity structure. Based on this observation, a regularized version of (13) can be formulated as

$$\min_{\mathbf{x}} \ \mathbf{x}^H\breve{\mathbf{\Theta}}\mathbf{x} + \sum_{l=1}^{L}\lambda_l\|\mathbf{x}_l\|_2 \tag{15a}$$

$$\text{s.t.} \ \ \mathbf{x}_l^H\mathbf{x}_l \le P_l^{\max}, \quad \forall l \in \mathcal{L} \tag{15b}$$

$$\breve{\mathbf{\Phi}}^H\mathbf{x} = \mathbf{c}, \tag{15c}$$

where $\breve{\mathbf{\Theta}} \triangleq \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Theta}\mathbf{\Psi}^{-\frac{1}{2}}$ and $\breve{\mathbf{\Phi}} \triangleq \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Phi}$.

It can be seen that (15) is convex, and therefore can be solved with global optimality using Newton-based interior point methods via standard optimization packages [36]. However, invoking general-purpose solvers becomes less computationally efficient as the problem dimension increases. In the current setup, the latter is defined by the number of RRHs and the size of the antenna arrays, i.e., $\sum_{l=1}^{L}N_l$, which could be significant, especially in future wireless networks employing millimeter-wave transmissions with massive antenna arrays. Besides, the BBUs are expected to implement the transceiver design procedure repeatedly in real-time as the channel conditions change due to user mobility, and this for a large scale C-RAN consisting of several sub-networks of the type described in Section II. Hence, in this work, we develop an alternative approach that leads to a low-complexity method to solve problem (15).

C. ADMM-Based Low-Complexity Algorithm

In what follows, we develop an algorithm for solving the regularized relaying optimization problem (15) based on ADMM [33]. In particular, we show that each one of the main steps in ADMM admits a closed-form solution, which significantly reduces the computational complexity of the resulting algorithm.

To rewrite (15) in a form amenable to ADMM, we introduce a synthesized copy of $\mathbf{x}$, namely $\mathbf{z}$, via the linear constraint $\mathbf{x} = \mathbf{z}$. Defining the constraint sets

$$\mathcal{C}_1 = \Big\{\mathbf{x}|\breve{\mathbf{\Phi}}^H\mathbf{x} = \mathbf{c}\Big\} \tag{16}$$

$$\mathcal{C}_2 = \Big\{\mathbf{z}|\mathbf{z}_l^H\mathbf{z}_l \le P_l^{\max}, \quad \forall l \in \mathcal{L}\Big\}, \tag{17}$$

(15) can be re-expressed as

$$\min_{\mathbf{x},\mathbf{z}} \ \mathbf{x}^H\breve{\mathbf{\Theta}}\mathbf{x} + \sum_{l=1}^{L}\lambda_l\|\mathbf{z}_l\|_2 \tag{18a}$$

$$\text{s.t.} \ \ \mathbf{x} \in \mathcal{C}_1, \ \mathbf{z} \in \mathcal{C}_2, \ \mathbf{x} = \mathbf{z}. \tag{18b}$$

The ADMM algorithm seeks to iteratively minimize the augmented Lagrangian given by

$$\mathcal{L}_\rho(\mathbf{x},\mathbf{z},\mathbf{y}) = \mathbf{x}^H\breve{\mathbf{\Theta}}\mathbf{x} + \sum_{l=1}^{L}\lambda_l\|\mathbf{z}_l\|_2 + \frac{\rho}{2}\|\mathbf{z}-\mathbf{x}\|_2^2 - 2\mathrm{Re}\Big(\big(\mathbf{z}-\mathbf{x}\big)^H\mathbf{y}\Big)\Big), \tag{19}$$

where $\rho > 0$ is an internal parameter which remains constant during the ADMM iterations, $\mathbf{y}$ denotes the Lagrange multiplier associated with the constraint $\mathbf{x} = \mathbf{z}$, and $\mathrm{Re}(\cdot)$ extracts the real part of its argument. Clearly, solving (18) becomes equivalent to the following problem,

$$\min_{\mathbf{x},\mathbf{z},\mathbf{y}} \ \mathcal{L}_\rho(\mathbf{x},\mathbf{z},\mathbf{y}) \ \ \text{s.t.} \ \ \mathbf{x} \in \mathcal{C}_1, \ \mathbf{z} \in \mathcal{C}_2, \ \mathbf{x} = \mathbf{z}. \tag{20a}$$

The basic idea behind ADMM is to solve the above problem with respect to $\mathbf{x}$ and $\mathbf{z}$ separately in an alternating manner, i.e., one variable at a time with the other being fixed. After each round of update of $\mathbf{x}$ and $\mathbf{z}$, a dual ascent update is performed on $\mathbf{y}$ to ensure that $\mathbf{x}$ and $\mathbf{z}$ become closer to each other. In effect, the optimization problem can now be decoupled into two separate steps, both of which, interestingly, admit a simple closed-form solution, as detailed below.

1) Updating $\mathbf{x}$*:* The subproblem for $\mathbf{x}$ can be expressed as $\min_{\mathbf{x}\in\mathcal{C}_1}\mathcal{L}_\rho(\mathbf{x},\mathbf{z},\mathbf{y})$, with $\mathbf{z}$ and $\mathbf{y}$ fixed. Using (19), this subproblem can further be expressed as the following linearly-constrained quadratic program after neglecting all terms that are independent of $\mathbf{x}$:

$$\mathbf{x}^{\mathrm{opt}} = \arg\min_{\mathbf{x}} \ \mathbf{x}^H\breve{\mathbf{\Theta}}\mathbf{x} + \frac{\rho}{2}\|\mathbf{z}-\mathbf{x}\|_2^2 - 2\mathrm{Re}\Big(\mathbf{y}^H(\mathbf{z}-\mathbf{x})\Big) \tag{21a}$$

$$\text{s.t.} \ \ \breve{\mathbf{\Phi}}^H\mathbf{x} = \mathbf{c}. \tag{21b}$$

It is observed that the objective function (21a) is strictly convex in $\mathbf{x}$ and the Slater's constraint qualification holds, i.e., (21) is strictly feasible. Hence, the Karush-Kuhn-Tucker (KKT) sufficient conditions hold for the optimal solution $\mathbf{x}^{\mathrm{opt}}$ together with some optimal dual variable $\boldsymbol{\nu}^{\mathrm{opt}}$, yielding:

$$\begin{cases} \mathbf{Q}\,\mathbf{x}^{\mathrm{opt}} = \frac{\rho}{2}\mathbf{z} - \mathbf{y} + \breve{\mathbf{\Phi}}\boldsymbol{\nu}^{\mathrm{opt}} \\ \breve{\mathbf{\Phi}}^H\,\mathbf{x}^{\mathrm{opt}} = \mathbf{c}, \end{cases} \tag{22}$$

where $\mathbf{Q} \triangleq \breve{\mathbf{\Theta}} + \frac{\rho}{2}\mathbf{I}_{\sum_{l=1}^{L}N_l^2}$.

Re-arranging the first equation in (22), we obtain

$$\mathbf{x}^{\mathrm{opt}} = \mathbf{Q}^{-1}\Big(\frac{\rho}{2}\mathbf{z} - \mathbf{y} + \breve{\mathbf{\Phi}}\boldsymbol{\nu}^{\mathrm{opt}}\Big). \tag{23}$$

To determine the value of the optimal dual variable $\boldsymbol{\nu}^{\mathrm{opt}}$, we substitute (23) back into the second equation of (22). After some matrix manipulations, $\boldsymbol{\nu}^{\mathrm{opt}}$ can be expressed as

$$\boldsymbol{\nu}^{\mathrm{opt}} = \breve{\mathbf{Q}}^{-1}\Big(\mathbf{c} - \breve{\mathbf{\Phi}}^H\mathbf{Q}^{-1}\Big(\frac{\rho}{2}\mathbf{z} - \mathbf{y}\Big)\Big), \tag{24}$$

where $\breve{\mathbf{Q}} \triangleq \breve{\mathbf{\Phi}}^H\mathbf{Q}^{-1}\breve{\mathbf{\Phi}}$. Then substituting (24) back into (23), the following closed-form solution is obtained

$$\mathbf{x}^{\mathrm{opt}} = \mathbf{Q}^{-1}\Big(\Big(\mathbf{I} - \breve{\mathbf{\Phi}}\breve{\mathbf{Q}}^{-1}\breve{\mathbf{\Phi}}^H\mathbf{Q}^{-1}\Big)\Big(\frac{\rho}{2}\mathbf{z} - \mathbf{y}\Big) + \breve{\mathbf{\Phi}}\breve{\mathbf{Q}}^{-1}\mathbf{c}\Big). \tag{25}$$

2) Updating $\mathbf{z}$*:* Similarly, the subproblem for $\mathbf{z}$ can be written as $\min_{\mathbf{z}\in\mathcal{C}_2}\mathcal{L}_p(\mathbf{x},\mathbf{z},\mathbf{y})$, where the values $\mathbf{x}$ and $\mathbf{y}$ are obtained from the previous iteration. Observing that the first term in (19) is independent of $\mathbf{z}$, we can write

$$\mathbf{z}^{\mathrm{opt}} = \arg\min_{\mathbf{z}\in\mathcal{C}_2}\sum_{l=1}^{L}\lambda_l\|\mathbf{z}_l\|_2 + \frac{\rho}{2}\|\mathbf{z}-\mathbf{x}\|_2^2 - 2\mathrm{Re}\Big(\mathbf{y}^H\mathbf{z}\Big). \tag{26}$$

Algorithm 1 ADMM for RRH Selection

1: **Initialization**: primal variable $\mathbf{z}^{(0)}$ (arbitrary non-zero vector); dual variable $\mathbf{y}^{(0)} = \mathbf{0}$; set ADMM iteration index $j = 0$
2: **repeat**
3: Update $\mathbf{x}^{(j+1)}$ using (25)
4: Update $\mathbf{z}_l^{(j+1)}$ using (31) for all $l \in \mathcal{L}$
5: Update the Lagrange multiplier $\mathbf{y}$ by

$$\mathbf{y}^{(j+1)} = \mathbf{y}^{(j)} + \frac{\rho}{2}\Big(\mathbf{x}^{(j+1)} - \mathbf{z}^{(j+1)}\Big)$$

6: $j \leftarrow j+1$
7: **until** $\|\mathbf{r}^{(j+1)}\|_2 \le \epsilon_r$ and $\|\mathbf{s}^{(j+1)}\|_2 \le \epsilon_s$

Decoupling (26) over each $\mathbf{z}_l$, we obtain L parallel subproblems each expressed by

$$\mathbf{z}_l^{\rm opt} = \arg \min_{\|\mathbf{z}_l\|_2^2 \le P_l^{\max}} \lambda_l \|\mathbf{z}_l\|_2 + \frac{\rho}{2}\|\mathbf{z}_l - \mathbf{x}_l\|_2^2 - 2{\rm Re}\Big(\mathbf{y}_l^H \mathbf{z}_l\Big), \tag{27}$$

where the original Lagrange multiplier is decomposed into L such multipliers, i.e., $\mathbf{y} = [\mathbf{y}_1^T, \ldots, \mathbf{y}_L^T]^T$. A closed-form solution to (28) can be obtained based on the following lemma.

Lemma 1: Consider the following convex minimization problem in $\mathbb{C}^n$, where $\mathbf{a}$ is constant:

$$\min_{\|\mathbf{x}\|_2^2 \le P} \lambda\|\mathbf{x}\|_2 + \frac{\rho}{2}\mathbf{x}^H\mathbf{x} - 2{\rm Re}\Big(\mathbf{x}^H\mathbf{a}\Big). \tag{28}$$

The global minimizer $\mathbf{x}^{\rm opt}$ admits the closed-form solution

$$\mathbf{x}^{\rm opt} = \frac{\mathbf{a}}{\|\mathbf{a}\|_2\big(\frac{\rho}{2} + \eta^{\rm opt}\big)}[\|\mathbf{a}\|_2 - \lambda]_+ \tag{29}$$

where $[c]_+ \triangleq \max\{0, c\}$ and the optimal dual variable $\eta^{\rm opt}$ associated with the quadratic constraint $\|\mathbf{x}\|_2^2 \le P$ is given by

$$\eta^{\rm opt} = \left[\frac{[\|\mathbf{a}\|_2 - \lambda]_+}{\sqrt{P}} - \frac{\rho}{2}\right]_+. \tag{30}$$

Proof: See the Appendix. ∎

Using Lemma 1, the solution to (27) is directly obtained as

$$\mathbf{z}_l^{\rm opt} = \frac{\mathbf{a}_l}{\|\mathbf{a}_l\|_2\big(\frac{\rho}{2} + \eta_l\big)}[\|\mathbf{a}_l\|_2 - \lambda_l]_+ \tag{31}$$

where $\eta_l = [\frac{\|\mathbf{a}_l\|_2 - \lambda_l}{\sqrt{P_l}} - \frac{\rho}{2}]_+$ and $\mathbf{a}_l = \frac{\rho}{2}\mathbf{x}_l + \mathbf{y}_l$.

In brief, both the updates of $\mathbf{x}$ and $\mathbf{z}$ are obtained in closed-form at each iteration with the aid of (25) and (31). In addition, the update step for $\mathbf{z}$ can be carried out in a parallel fashion over each $\mathbf{z}_l$. The ADMM-based algorithm is summarized in Algorithm 1, where the primal and dual residuals are defined as follows (see [33]),

$$\mathbf{r}^{(j+1)} = \mathbf{x}^{(j+1)} - \mathbf{z}^{(j+1)}, \tag{32}$$

$$\mathbf{s}^{(j+1)} = -\frac{\rho}{2}\Big(\mathbf{z}^{(j+1)} - \mathbf{z}^{(j)}\Big), \tag{33}$$

and $\epsilon_r, \epsilon_s > 0$ denote the tolerance parameters.

Referring to Step 4 of Algorithm 1, the thresholding operator in (31) accounts for the possible group sparse property associated with the outputs of the algorithm, i.e., some of the elements in $[\|\mathbf{z}_1\|_2, \ldots, \|\mathbf{z}_L\|_2]^T$ are zero. Therefore, the subset of active RRHs can be determined as:

$$\mathcal{A} = \{l \in \mathcal{L} : \|\mathbf{z}_l\|_2 > 0\}. \tag{34}$$

Only the RRHs in the above subset will be considered in the next step of RRH weight selection.

IV. RRH Weight Selection Scheme

In this section, we consider the problem of RRH weight selection with the aim of reducing the number of non-zero weights in the optimized RRH weight vectors. We then develop an ADMM-based algorithm to solve the corresponding regularized problem with low complexity. Finally, we examine the use of non-uniform quantization for efficient feedback of the optimized non-zero weights over the fronthaul links.

A. RRH Weight Selection via Weight-Level Sparsity

The objective of the RRH weight selection scheme is to reduce the number of non-zero weights associated with the active RRHs, which serves two complementary purposes. First, given a fixed total number of bits for RRH weight quantization, reducing the number of weights translates into more bits being assigned to each non-zero weight, so that the quantization error for these weights can be reduced. Second, when the number of quantization bits for each weight is fixed, reducing the number of weights leads to a corresponding reduction in the total number of bits that need to be fed back to the RRHs.

To achieve this objective, the following hard sparsity constraint can be imposed:

$$\|\mathbf{x}_l\|_0 \le I_l^{\max}, \quad l \in \mathcal{A}, \tag{35}$$

where $I_l^{\max}$ is a positive integer which constrains the number of non-zero elements in $\mathbf{x}_l$. Then, the weight vectors for the selected active RRHs can be optimized by solving

$$\min_{\breve{\mathbf{x}}} \ \breve{\mathbf{x}}^H \breve{\mathbf{\Theta}}_{\rm R} \breve{\mathbf{x}} \tag{36a}$$

$$\text{s.t.} \ \ \mathbf{x}_l^H\mathbf{x}_l \le P_l^{\max}, \quad \|\mathbf{x}_l\|_0 \le I_l^{\max}, \quad \forall l \in \mathcal{A} \tag{36b}$$

$$\breve{\mathbf{\Phi}}_{\rm R}^H \breve{\mathbf{x}} = \mathbf{c}, \tag{36c}$$

where $\breve{\mathbf{x}}$ now only consists of weights from active RRHs, i.e.,

$$\breve{\mathbf{x}} = \Big[\mathbf{x}_{l_1}, \mathbf{x}_{l_2}, \ldots, \mathbf{x}_{l_{|\mathcal{A}|}}\Big] \tag{37}$$

where $\{l_1, \ldots, l_{|\mathcal{A}|}\} = \mathcal{A}$, and $\breve{\mathbf{\Theta}}_{\rm R}$ and $\breve{\mathbf{\Phi}}_{\rm R}$ are the reduced versions of $\breve{\mathbf{\Theta}}$ and $\breve{\mathbf{\Phi}}$, respectively, with the elements corresponding to the inactive RRHs removed. Exploiting established results from the theory of sparse signal recovery [37], the effect of the hard sparsity constraint (35) can be equivalently captured by augmenting (36a) with an l_1-norm regularization term for each $\mathbf{x}_l$, yielding:

$$\min_{\breve{\mathbf{x}}} \ \breve{\mathbf{x}}^H \breve{\mathbf{\Theta}}_{\rm R} \breve{\mathbf{x}} + \sum_{l \in \mathcal{A}} \gamma_l \|\mathbf{x}_l\|_1 \tag{38a}$$

$$\text{s.t.} \ \ \mathbf{x}_l^H\mathbf{x}_l \le P_l^{\max}, \quad \forall l \in \mathcal{A} \tag{38b}$$

$$\breve{\mathbf{\Phi}}_{\rm R}^H \breve{\mathbf{x}} = \mathbf{c}, \tag{38c}$$

where γ_l is a positive tuning parameter which controls the sparsity level of each $\mathbf{x}_l$. At this point, it is worth noting that in (38), the l_1-norm term $\|\mathbf{x}_l\|_1$ is introduced to encourage the sparsity at the weight level while in (15), the l_2-norm term $\|\mathbf{x}_l\|_2$ is used to promote group sparsity, i.e., all weights in $\mathbf{x}_l$ are compelled to be zero simultaneously.

B. ADMM-Based Algorithm

Similar to Section III-C, we now proceed to derive an ADMM-based low-complexity algorithm to solve (38). We introduce a synthesized copy of $\breve{\mathbf{x}}$, i.e., $\boldsymbol{\eta} = \breve{\mathbf{x}}$, and express (38) in the following equivalent form:

$$\min_{\breve{\mathbf{x}},\boldsymbol{\eta}} \ \breve{\mathbf{x}}^H \breve{\boldsymbol{\Theta}}_{\mathrm{R}} \breve{\mathbf{x}} + \sum_{l\in\mathcal{A}} \gamma_l \|\boldsymbol{\eta}_l\|_1 \tag{39a}$$

$$\text{s.t.} \ \ \boldsymbol{\eta}_l^H \boldsymbol{\eta}_l \le P_l^{\max}, \quad \forall l \in \mathcal{A} \tag{39b}$$

$$\breve{\boldsymbol{\Phi}}_{\mathrm{R}}^H \breve{\mathbf{x}} = \mathbf{c}, \qquad \breve{\mathbf{x}} = \boldsymbol{\eta}. \tag{39c}$$

The augmented Lagrangian function associated to (39) can be expressed as

$$\mathcal{L}_\rho(\breve{\mathbf{x}}, \boldsymbol{\eta}, \boldsymbol{\nu}) = \breve{\mathbf{x}}^H \breve{\boldsymbol{\Theta}}_{\mathrm{R}} \breve{\mathbf{x}} + \frac{\rho}{2}\|\boldsymbol{\eta} - \breve{\mathbf{x}}\|_2^2 + \sum_{l\in\mathcal{A}} \gamma_l \|\boldsymbol{\eta}_l\|_1 - 2\mathrm{Re}\Big((\boldsymbol{\eta} - \breve{\mathbf{x}})^H \boldsymbol{\nu}\Big), \tag{40}$$

where $\rho > 0$ is a tuning parameter and $\boldsymbol{\nu}$ denotes the Lagrangian multiplier associated with the equality constraint $\breve{\mathbf{x}} = \boldsymbol{\eta}$. Our goal is to minimize $\mathcal{L}_\rho(\breve{\mathbf{x}}, \boldsymbol{\eta}, \boldsymbol{\nu})$ with respect to the primal variables $\breve{\mathbf{x}}$ and $\boldsymbol{\eta}$ and the dual variable $\boldsymbol{\nu}$ in a Gauss-Seidel way.

To begin, note that by removing all the terms independent of $\breve{\mathbf{x}}$ in (40), the resultant subproblem for the optimal $\breve{\mathbf{x}}$ becomes similar to (21). Therefore, after some manipulations, the following result can be obtained:

$$\breve{\mathbf{x}}^{\mathrm{opt}} = \mathbf{P}^{-1}\Big(\Big(\mathbf{I} - \breve{\boldsymbol{\Phi}}_{\mathrm{R}} \breve{\mathbf{P}}^{-1} \breve{\boldsymbol{\Phi}}_{\mathrm{R}}{}^H \mathbf{P}^{-1}\Big)\Big(\frac{\rho}{2}\boldsymbol{\eta} - \boldsymbol{\nu}\Big) + \breve{\boldsymbol{\Phi}}_{\mathrm{R}} \breve{\mathbf{P}}^{-1}\mathbf{c}\Big), \tag{41}$$

where $\mathbf{P} \triangleq \breve{\boldsymbol{\Theta}}_{\mathrm{R}} + \frac{\rho}{2}\mathbf{I}_{\sum_{l\in\mathcal{A}} N_l^2}$ and $\breve{\mathbf{P}} \triangleq \breve{\boldsymbol{\Phi}}_{\mathrm{R}}{}^H \mathbf{P}^{-1} \breve{\boldsymbol{\Phi}}_{\mathrm{R}}$.

Then, to update each $\boldsymbol{\eta}_l$ for $l \in \mathcal{A}$, we need to solve the following subproblem:

$$\boldsymbol{\eta}_l^{\mathrm{opt}} = \arg \min_{\|\boldsymbol{\eta}_l\|_2^2 \le P_l^{\max}} \gamma_l \|\boldsymbol{\eta}_l\|_1 + \frac{\rho}{2}\|\boldsymbol{\eta}_l - \mathbf{x}_l\|_2^2 - 2\mathrm{Re}\Big(\boldsymbol{\nu}_l^H \boldsymbol{\eta}_l\Big). \tag{42}$$

To this end, we make use of the following lemma, the proof of which is conceptually similar to the proof of Lemma 1 and is omitted due to lack of space.

Lemma 2: Consider the following convex minimization problem in $\mathbb{C}^n$, where $\mathbf{a} = [a_1, \ldots, a_n]^T \in \mathbb{C}^n$ is constant:

$$\min_{\|\mathbf{x}\|_2^2 \le P} \lambda\|\mathbf{x}\|_1 + \frac{\rho}{2}\mathbf{x}^H\mathbf{x} - 2\mathrm{Re}\Big(\mathbf{a}^H\mathbf{x}\Big). \tag{43}$$

The global minimizer $\mathbf{x}^{\mathrm{opt}}$ admits the closed-form solution

$$\mathbf{x}^{\mathrm{opt}} = \frac{\mathbf{a}}{\|\mathbf{a}\|_2\big(\frac{\rho}{2} + \mu^{\mathrm{opt}}\big)} \mathrm{soft}_\lambda(\mathbf{a}), \tag{44}$$

Algorithm 2 ADMM for RRH Weight Selection

1: **Initialization**: primal variable $\boldsymbol{\eta}^{(0)}$; dual variable $\boldsymbol{\nu}^{(0)} = \mathbf{0}$; set ADMM iteration index $j = 0$
2: **repeat**
3: Update $\breve{\mathbf{x}}^{(j+1)}$ using (41)
4: Update $\boldsymbol{\eta}_l^{(j+1)}$ using (46) for all $l \in \mathcal{A}$
5: Update the Lagrange multiplier $\boldsymbol{\nu}$ by
$$\boldsymbol{\nu}^{(j+1)} = \boldsymbol{\nu}^{(j)} + \frac{\rho}{2}\Big(\breve{\mathbf{x}}^{(j+1)} - \boldsymbol{\eta}^{(j+1)}\Big)$$
6: $j \leftarrow j+1$
7: **until** $\|\breve{\mathbf{r}}^{(j+1)}\|_2 \le \epsilon_r$ and $\|\breve{\mathbf{s}}^{(j+1)}\|_2 \le \epsilon_s$

where $\mathrm{soft}_\lambda(\mathbf{a})$ denotes the element-wise soft thresholding operator with threshold λ, i.e., $\mathrm{soft}_\lambda(\mathbf{a}) = [[|a_1| - \lambda]_+, \ldots, [|a_n| - \lambda]_+]^T$. The optimal dual variable μ^{opt} associated with the quadratic constraint $\|\mathbf{x}\|_2^2 \le P$ is given by

$$\mu^{\mathrm{opt}} = \left[\frac{\|\mathrm{soft}_\lambda(\mathbf{a})\|_2}{\sqrt{P}} - \frac{\rho}{2}\right]_+. \tag{45}$$

The solution of (42) follows directly from Lemma 2, i.e.,

$$\boldsymbol{\eta}_l^{\mathrm{opt}} = \frac{\mathbf{a}_l}{\|\mathbf{a}_l\|_2\big(\frac{\rho}{2} + \mu_l\big)} \mathrm{soft}_{\gamma_l}(\mathbf{a}_l), \tag{46}$$

where $\mu_l = [\frac{\|\mathrm{soft}_{\gamma_l}(\mathbf{a}_l)\|_2}{\sqrt{P_l^{\max}}} - \frac{\rho}{2}]_+$ and $\mathbf{a}_l = \frac{\rho}{2}\mathbf{x}_l + \boldsymbol{\nu}_l$.

The ADMM algorithm for RRH weight selection is summarized as Algorithm 2, where $\breve{\mathbf{r}}^{(j+1)} = \breve{\mathbf{x}}^{(j+1)} - \boldsymbol{\eta}^{(j+1)}$ and $\breve{\mathbf{s}}^{(j+1)} = -\frac{\rho}{2}(\boldsymbol{\eta}^{(j+1)} - \boldsymbol{\eta}^{(j)})$ denote the primal and dual residuals, respectively, and $\epsilon_r, \epsilon_s > 0$ denote the tolerance parameters.

C. Non-Uniform Scalar Quantization

We now investigate the potential performance benefits of the proposed weight selection algorithm when used in conjunction with a practical quantization scheme. Uniform scalar quantization has been considered in a number of prior works on RRH transceiver optimization within C-RAN due its flexibility and simplicity [7], [18]. However, it is well known that uniform quantization is optimal only for uniformly distributed input data. For non-uniformly distributed data, it can lead to an increase of the mean squared quantization error (MSQE) by several dB [38]. Therefore, we first examine the statistical properties of the optimized RRH weights, and on this basis, further consider a non-uniform scalar quantization scheme relying on the Lloyd-Max algorithm [39].

Let us denote by $\tilde{\mathbf{x}}$ the vector that collects all the selected non-zero weights from the output of Algorithm 2. For the purpose of statistical analysis, $\tilde{\mathbf{x}}$ is further normalized by

$$\bar{\mathbf{x}} = \frac{1}{\hat{\sigma}}(\tilde{\mathbf{x}} - \hat{\mu}\mathbf{1}), \tag{47}$$

where $\hat{\mu}$ and $\hat{\sigma}$ denote the sample mean and variance of the elements of $\tilde{\mathbf{x}}$, respectively, and $\mathbf{1} = [1, \ldots, 1]^T$ is an all-one vector of appropriate dimension. As $\bar{\mathbf{x}}$ is complex valued, it can be represented by the sum of its real and imaginary parts as $\bar{\mathbf{x}} = \bar{\mathbf{x}}_{\mathcal{I}} + \mathrm{j}\bar{\mathbf{x}}_{\mathcal{Q}}$, where $\mathrm{j} = \sqrt{-1}$. We analyze the statistical

properties of $\bar{\mathbf{x}}$ through simulations, as per the methodology described in Section V.

We first generate a set of M independent normalized weight vectors, where the real and imaginary parts of the i^{th} normalized vector are denoted by $\bar{\mathbf{x}}_{\mathcal{I}}^{(i)}$ and $\bar{\mathbf{x}}_{\mathcal{Q}}^{(i)}$, respectively, for $i \in \{1, \ldots, M\}$. We then construct two sets, i.e.,

$$\mathcal{X}_{\mathcal{I}} = \left\{\bar{\mathbf{x}}_{\mathcal{I}}^{(1)}, \ldots, \bar{\mathbf{x}}_{\mathcal{I}}^{(M)}\right\},\ \mathcal{X}_{\mathcal{Q}} = \left\{\bar{\mathbf{x}}_{\mathcal{Q}}^{(1)}, \ldots, \bar{\mathbf{x}}_{\mathcal{Q}}^{(M)}\right\}. \quad (48)$$

From the empirical distributions of $\mathcal{X}_{\mathcal{I}}$ and $\mathcal{X}_{\mathcal{Q}}$ it was observed that the real and imaginary parts of the non-zero RRH weights are approximately uncorrelated. Additionally, both the distributions of real and imaginary parts follow bell-shape curves similar to the Gaussian distribution. Hence, the uniform quantization is not an optimal scheme for the RRH weight feedback purpose.

We therefore consider the use of a non-uniform quantization scheme. To obtain an optimal non-uniform quantizer in terms of the partition regions and codewords, one can rely on the Lloyd-Max algorithm [39]. This algorithm, however, requires the distribution of the weight vector. We performed the Shapiro-Wilk test [40] of the null hypothesis that the samples from $\mathcal{X}_{\mathcal{I}}$ and $\mathcal{X}_{\mathcal{Q}}$ follow Gaussian distributions. The null hypothesis was rejected, so we do not assume that the samples are Gaussian distributed. Instead, we rely on the empirical distribution (i.e., the discrete distribution of $\mathcal{X}_{\mathcal{I}}$ and $\mathcal{X}_{\mathcal{Q}}$) for the determination of the optimal non-uniform scalar quantization levels.

For the real samples of the weight vector contained in $\mathcal{X}_{\mathcal{I}}$, given a predefined number of quantization levels, say $N_{\mathcal{I}}$, let us denote the initial codebook as $\mathbf{c}_{\mathcal{I}}^{(0)} = [c_{\mathcal{I},1}^{(0)}, \ldots, c_{\mathcal{I},N_{\mathcal{I}}}^{(0)}]^T$. The codebook is then improved by the Lloyd-Max algorithm, i.e., alternating between the following two steps, where the superscript m denotes the iteration index:

1) Find the optimal partitions using the nearest-neighbor condition with distance measure $d(x, y) \triangleq |x - y|$:

$$\mathcal{R}_i^{(m)} = \Big\{x \in \mathcal{X}_{\mathcal{I}} : d\Big(x, c_{\mathcal{I},i}^{(m)}\Big) \leq d\Big(x, c_{\mathcal{I},j}^{(m)}\Big), \forall j \neq i\Big\}. \quad (49)$$

2) Update the codewords based on the centroid condition:

$$c_{\mathcal{I},i}^{(m+1)} = \frac{1}{\left|\mathcal{R}_i^{(m)}\right|} \sum_{x \in \mathcal{R}_i^{(m)}} x, \quad i = 1, \ldots, N_{\mathcal{I}}. \quad (50)$$

The same procedure can be applied to obtain the codebook $\mathbf{c}_{\mathcal{Q}}$ for the imaginary samples in $\mathcal{X}_{\mathcal{Q}}$.

Finally, to reconstruct the weight vector, the RRHs require the knowledge of the sample mean and variance of the weight vector [see (47)] and a bit map (consisting of zeros and ones) indicating the selected non-zero RRH weights. Let us denote the number of bits required for quantizing the sample mean and variance by Q_{mean} and Q_{var}, respectively. Further, the number of bits required for transmitting the bit map is $Q_{\text{map}} = \sum_{l=1}^{L} N_l^2$. Therefore, given a total budget of Q bits for weight quantization and feedback, the number of bits allocated to each scalar weigh parameter (i.e., real or imaginary part of a normalized RRH weight) is given by[2]

$$Q_{\text{per}} = \frac{1}{2\|\bar{\mathbf{x}}\|_0}\big(Q - Q_{\text{mean}} - Q_{\text{var}} - Q_{\text{map}}\big), \quad (51)$$

where $\|\cdot\|_0$ gives the total number of non-zero RRH weights. It is then expected that more bits can be allocated to non-zero weights when weight selection is invoked compared to the non-sparse method.

Algorithm 3 Overall System Implementation at BBU Pool

1: Perform RRH selection using Algorithm 1
2: Perform RRH weight selection based on Algorithm 2
3: Re-optimize $\mathbf{x}_l$ for the selected weights for $l \in \mathcal{A}$
4: Quantize $\mathbf{x}_l$ and send back to the corresponding RRHs.

D. System Implementation

It is worth noting that the addition of the group and weight level regularization terms in the objective functions of (15) and (38) may lead to higher total interference leakage (also referred to as "bias-variance tradeoff" in the machine learning literature on LASSO method [37]). In practice, however, the final interference level can be reduced by solving for the optimal RRH weights once more for those selected weights of active RRHs from the previous stages. To conclude this section, we summarize as Algorithm 3 the resulting overall system implementation at the BBU pool, which comprises ADMM-based RRH and weight selection, re-optimization and weight quantization.

V. Results and Discussion

In this section, numerical simulations are performed to assess the performance of the proposed ADMM-based algorithms. After describing the methodology, the overall RRH transceiver optimization performance is studied in terms of average SIR, average number of active number of RHHs and average number of non-zero RHH weights. Moreover, in order to demonstrate that the proposed group-sparse relay transceiver design complies with the three main requirements of green C-RAN-based communications, i.e., reduced network power consumption, reduced fronthaul feedback overhead, and low computational complexity design procedure, corresponding simulation results are presented in separate subsections, respectively. Finally, comparisons with existing methods under a simplified beamforming scenario are presented.

A. Methodology

In our simulations, unless otherwise specified, we consider a RRH sub-network consisting of $K = 6$ user pairs and $L = 6$ RRHs. For simplicity, the number of antennas and power budget are identical for all RRHs, i.e., $N_l = 4$ and $P_l^{\max} = 2\,W$ for all $l \in \mathcal{L}$. We note that these values, corresponding to a small scale system, are of the same order as those used in related works [8], [11] on group sparse beamforming under

[2]If Q_{per} is fractional, we first assign $\lfloor Q_{\text{per}} \rfloor$ to each parameter, and then assign the remaining bits singly to a selected subset of the parameters.

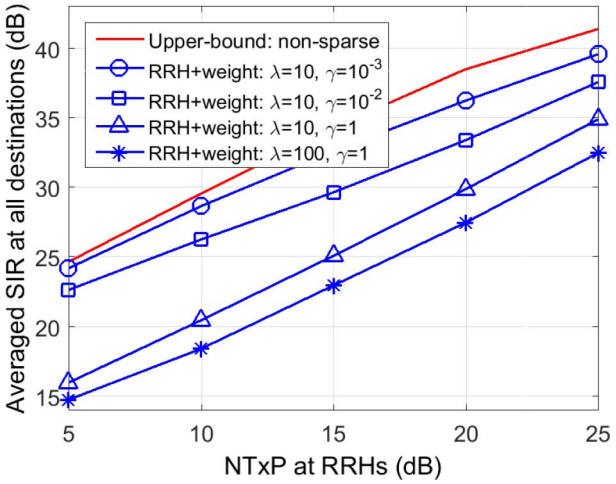


Fig. 2. Average SIR at destinations versus normalized RRH transmission power.

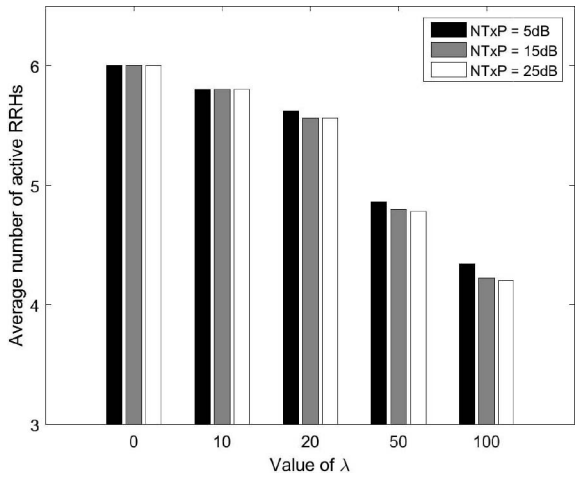
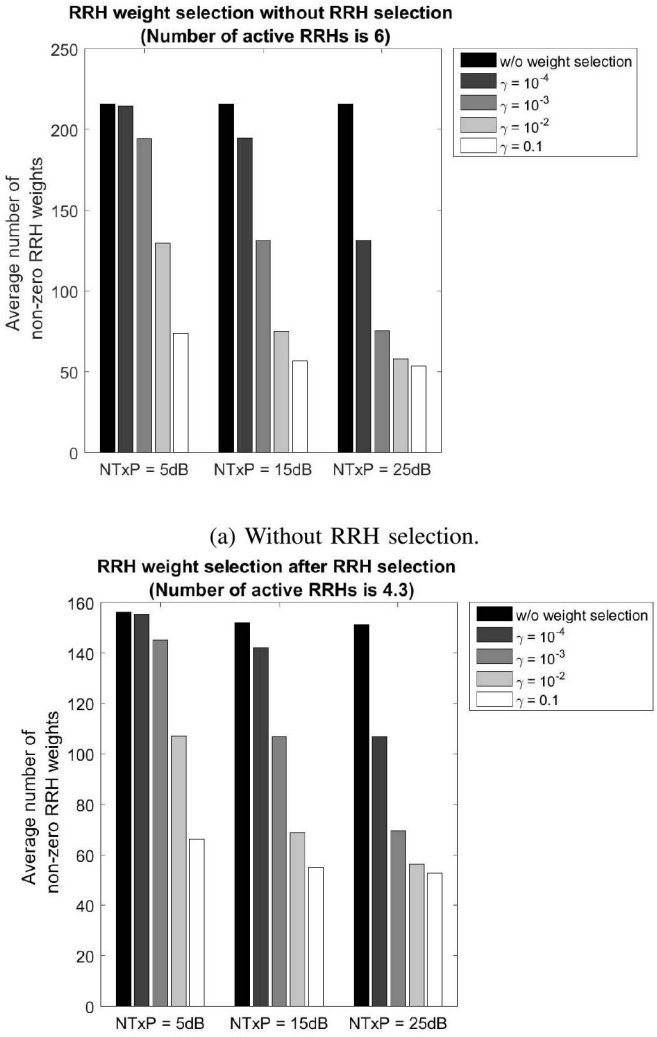


Fig. 3. Average number of active RRHs as a function of λ for different NTxP values.



(a) Without RRH selection.

(b) With RRH selection.

Fig. 4. Average number of non-zero RRH weights for different choices of γ and NTxP values.

C-RAN control. The channel coefficients and additive noise terms are generated as independent and identically distributed zero-mean complex circular Gaussian random variables. The variance of the channel coefficients is set to unity while the noise variance at the lth RRH, σ_l^2, is set according to the desired value of the normalized RRH transmission power (NTxP), defined as $\varrho_l = P_l^{\max}/N_l\sigma_l^2$. All the simulation results are averaged over 100 independent channel realizations. For the proposed ADMM-based algorithms, identical values of $\lambda_l \equiv \lambda$ and $\gamma_l \equiv \gamma$ are used for all RRHs to simplify the discussion (see (15a) and (38a)). Different algorithms are considered for the purpose of comparison, as explained in the corresponding subsections.

B. Performance of Overall System

The performance of the overall RRH transceiver optimization specified by Algorithm 3, which performs RRH selection followed by weight selection is evaluated in Figs. 2–5. At this stage, no quantization scheme is considered. The signal-to-interference ratio (SIR) averaged over all destinations is adopted as an indicator of quality for the received signals at the destinations, which is expressed as

$$\mathrm{SIR} = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{E}\{|S_k|^2\}}{\mathrm{E}\{|I_k|^2\}}, \tag{52}$$

where S_k and I_k are defined in (4) and (5), respectively. The achieved SIR of the proposed scheme with different combinations of (λ, γ), which affect the numbers of active RRHs and non-zero weights, respectively, is compared with an upper bound on performance, referred to as "non-sparse" approach in the sequel. The latter is obtained by solving for the optimal RRH weight vectors without employing any form of sparsity-inducing regularization, i.e., no RRH and weight selection, as obtained by setting $\lambda_l = 0\ \forall l$ in (15a). In Fig. 2, it is observed that, as the values of λ and γ increase, the gap between the proposed algorithm and the upper bound also increases owing to the reduced number of RRHs and weights.

To gain more insight into the observed performance, we further evaluate the average number of active RRHs involved in the transmission and the total number of non-zero RRH

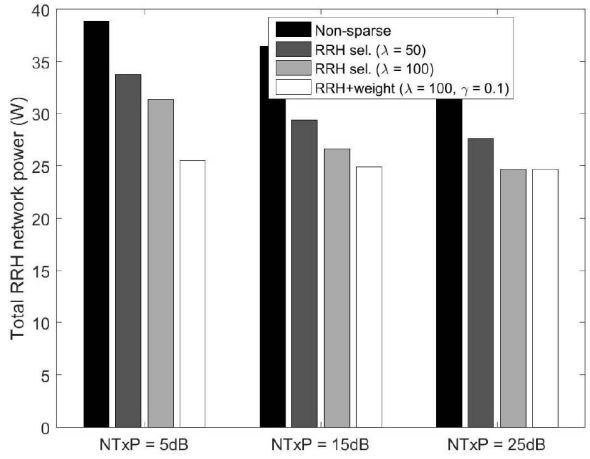


Fig. 5. Total RRH network power consumption for different NTxP values ($P_l^{\max} = 2W$, σ_l^2 is varied).

weights. In Fig. 3, the former is shown as a function of λ for different NTxP values. In each case, it is observed that the number of active RRHs monotonically decreases as λ increases. For instance, when λ reaches a value of 50, less than 5 of the 6 available RRHs participate in the transmission, on average, regardless of the RRH noise level; this number drops to around 4 when λ further increases to 100. The effect of varying the regularization parameter γ in controlling the number of non-zero RRH weights is investigated in Figs. 4(a) and 4(b), where the results are reported for different values of NTxP. In Fig. 4(a), it is assumed that no RRH selection is performed prior to weight selection with Algorithm 2, i.e., the number of active RRHs is 6 and the original number of non-zero weights is $6 \times 36 = 216$. It is seen that the proposed algorithm is capable of significantly reducing the number of non-zero weights, e.g., by over 65%, but the required value of γ depends on the NTxP level. In Fig. 4(b), RRH selection is first performed using Algorithm 1 with $\lambda = 100$, leading to an average of around 4.3 active RRHs, prior to the application of Algorithm 2. The latter can still significantly reduce the number of non-zero weights, e.g., by over 55% with an appropriate choice of γ. Overall, Figs. 4(a) and 4(b) show a similar trend, i.e., the number of non-zero weights decreases with an increase in γ, as expected.

C. Network Power Consumption

To better appreciate the benefits of the proposed algorithms in terms of energy efficiency, we examine the total power consumption of the RRH subnetwork within C-RAN. For each RRH, the following simplified power model is adopted from [10]:

$$P_l^{\rm rrh} = P_l^{\rm amp} + I(\|\mathbf{x}_l\|_2) P_l^{\rm sta}, \tag{53}$$

where $P_l^{\rm amp}$ denotes the power consumption of the RRH's power amplifier (PA), $P_l^{\rm sta}$ denotes the static power consumption of the fronthaul link and other RRH components, and $I(x)$ is the indicator function, i.e., $I(x) = 1$ for $x \neq 0$ and $I(x) = 0$ otherwise. The power consumption of the PA can be expressed as $P_l^{\rm amp} = P_l^{\rm tr}/\eta$, where $P_l^{\rm tr}$ denotes the transmission power of the RRH (see (2)) and η is the drain efficiency of the PA. A representative value of η for a high-efficiency broadband PA for next-generation wireless standards is 50% [41]. According to (53), when an RRH and its corresponding fronthaul link are switched off (sleep mode), we must have $\|\mathbf{x}_l\|_2 = 0$ and therefore, the static power consumption can be saved. A typical value of $P_l^{\rm sta} = 5.6W$ is used here [10]. Based on the above model, the total RRH network power consumption can be calculated as $P^{\rm net} = \sum_{l=1}^{L} P_l^{\rm rrh}$.

In Fig. 5, the values of $P^{\rm net}$ achieved by the proposed algorithms are compared for different RRH noise levels, where "RRH sel." refers to Algorithm 1 and "RRH+weight" refers to Algorithm 3 (but without quantization). It is observed that in all cases, the non-sparse algorithm yields the highest network power consumption while the proposed RRH selection Algorithm 1 effectively reduces the network power by around 18%–25% depending on the noise level. When weight selection is used in conjunction with relay selection, i.e., Algorithm 3, the network power can be reduced further for medium to high RRH noise levels.

D. Fronthaul Feedback Overhead

In this subsection, we verify the performance benefits of the proposed RRH weight selection algorithm when the non-uniform quantization scheme described in Section IV-C is employed. Referring to (51), the numbers of bits for quantizing the mean and variance are set to $Q_{\rm mean} = 20$ and $Q_{\rm var} = 10$, while the number of bits for transmitting the bit map is $Q_{\rm map} = 216$. Fig. 6 shows the SIR performance of different transceiver design using Algorithm 3 and possible variants. The upper-bound represents the SIR value obtained from the non-sparse solution without any quantization (i.e., analog RRH weights), "w/o weight selection" refers to Algorithm 1 with quantization, while "weight selection" refers to the complete Algorithm 3 with Step 4. It is observed that the proposed weight selection algorithm yields a significantly higher SIR than that of the solution obtained without weight selection. For instance, when the total number of bits is $Q = 3000$, use of quantization with the weight selection algorithm ($\gamma = 10^{-3}$) leads to a SIR loss of around 2 dB compared to the upper bound, while direct quantization of the full set of RRH weights yields a SIR loss of more than 15 dB. Note also that the choice of $\gamma = 10^{-4}$ leads to a lower SIR than $\gamma = 10^{-3}$, due to the resulting larger number of non-zero RRHs weights, and hence less available bits per sample for quantization.

E. Computational Complexity

The convergence of the proposed ADMM-based algorithms is studied in Figs. 7(a) and 7(b). The values of the objective function, primal residual, dual residual and sparsity patterns are shown in the four subfigures within each figure. Fig. 7(a) illustrates the convergence behavior of Algorithm 1, which jointly performs RRH selection and weight vector optimization, for one specific channel realization. It can be observed from the top-left figure that the value of the objective function, i.e., (15a), decreases sharply within a few iterations, owing to the exact nature of the closed-form solutions in the

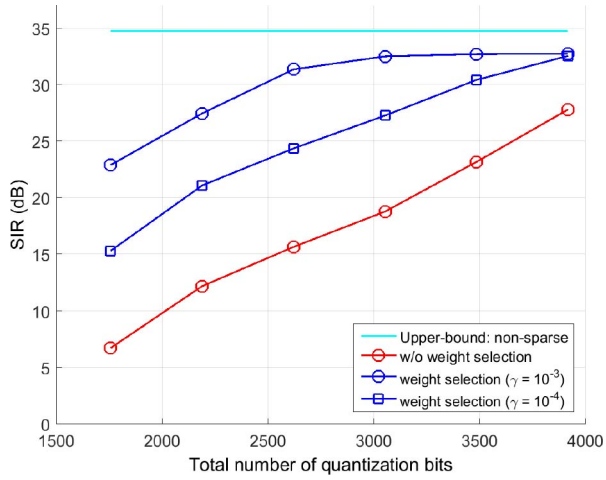
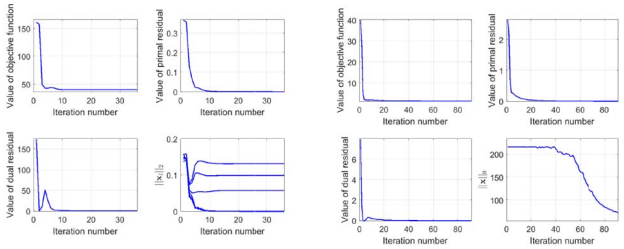


Fig. 6. SIR performance of the proposed weight compression algorithm with non-uniform quantization.



(a) Algorithm 1 with $\varrho_l = 15$ dB, $\lambda = 100$. (b) Algorithm 2 with $\varrho_l = 15$ dB, $\gamma = 0.5$.

Fig. 7. Convergence behavior of Algorithm 1 and 2 for RRH selection with $\epsilon_r = \epsilon_s = 10^{-3}$.

ADMM steps. The algorithm converges (i.e., stops) within less than 40 iterations when both values of the primal and dual residuals become smaller than the predefined tolerance, set to 10^{-3} in this case. In practice, when a larger tolerance is used, the algorithm may stop even faster. It is interesting to observe that the weight vectors of three RRHs converge to zero, which means that for this specific realization, the algorithm yields a subset of only $L - 3 = 3$ active RRHs.

Fig. 7(b) illustrates the convergence of Algorithm 2, which jointly performs weight selection and optimization. Here, it is assumed that prior to the weight selection, no RRH selection is performed and therefore, the initial number of non-zero weights is $L \times N_l^2 = 216$. It is seen that by setting $\gamma = 0.5$, the number of non-zero RRH weights decreases from 216 to around 75, which in turn leads to a reduction of the fronthaul overhead during the weight feedback stage. Finally, we compare the processing time of the proposed ADMM-based Algorithm 1 for group-sparse RRH transceiver design, with two alternative solution approaches relying on external optimization solvers, namely, `SeDuMi` and `MOSEK`. In effect, these alternative approaches implement Algorithm 1 by directly solving problem (15) with the assistance of the corresponding solver. The results,[3] listed in Table I, show that

[3]Based on the use of a desktop computer equipped with 8th Generation Intel i7-8700 6-core processor (12M Cache, 4.6 GHz) and 32GB RAM.

TABLE I
PROCESSING TIME OF DIFFERENT SOLUTIONS (IN SECONDS)

	Regularization parameter λ				
	1	10	20	50	100
SeDuMi	0.1658	0.2992	0.2293	0.2367	0.2391
MOSEK	0.0349	0.0459	0.0473	0.0506	0.0509
Proposed	0.0056	0.0077	0.0098	0.0118	0.0130

the complexity of the proposed ADMM-based algorithms is only a small fraction of that of the solver-based approaches, owing to the use of simple, exact closed-form solutions in the main ADMM steps. Similar gains in processing time have been observed in the case of Algorithm 2.

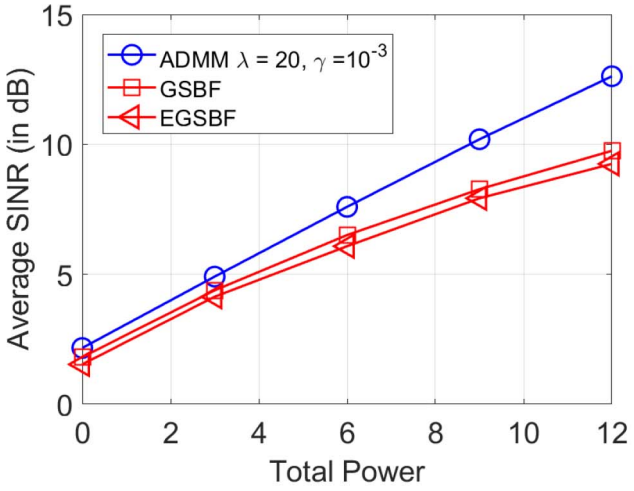
F. Comparison With Other Methods

Direct comparison of our method with previous studies is not possible since the functionality of RRHs as AF MIMO relays within a C-RAN architecture has not yet been addressed in the literature. However, the problem of energy efficient multi-user DL beamforming under C-RAN has been recently studied in [8] and [11], where a group sparse beamforming (GSBF) algorithm based on weighted mixed ℓ_1/ℓ_2-norm, and an enhanced GSBF (EGSBF) algorithm based on iterative reweighted ℓ_2-norm are presented, respectively. Therefore, we consider the multi-user DL scenario to further examine and compare the performance of our proposed ADMM-based algorithms to existing solutions. Our system model applies to this scenario by assuming that the destination users are served by data readily available at the RRHs, which amounts to the special choice $\mathbf{h}_{lk} = \mathbf{e}_k$ and $\mathbf{n}_l = 0$ in (1), where $K \leq N_l$ is implied. Under this setting, the selected RRHs and AF matrices obtained from the application of our algorithms can be used for group sparse multi-user DL beamforming as in [8] and [11]. Here, we compare the performance of our proposed Algorithm 3 (without quantization) to that of the above GSBF and EGSBF algorithms. For the simulations, we consider a sub-network consisting of $L = 5$ RRHs, each equipped with $N_l = 10$ antennas, and serving $K = 10$ single antenna users. To achieve a similar level of sparsity with the different methods, the QoS constraint parameter is set to 2 for the GSBF and EGSBF algorithms, while the parameters γ and λ in Algorithm 3 are set to 10^{-3} and 20, respectively. The performance is evaluated in terms of the average signal-to-interference plus noise ratio (SINR) and the average bit error rate (BER) at the destination users, The former is defined as

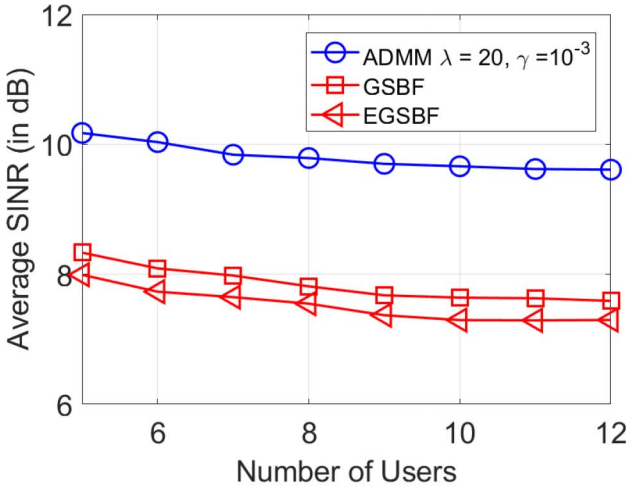
$$\text{SINR} = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{E}\{|S_k|^2\}}{\mathrm{E}\{|I_k|^2\} + \varsigma_k^2} \tag{54}$$

while the latter is measured by transmitting 4-ary quadrature amplitude modulation (4-QAM) symbols from the RRHs to the destinations.

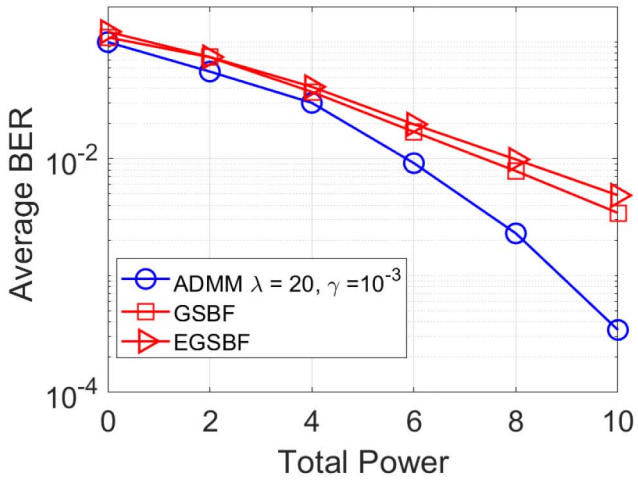
In Fig. 8(a), we plot the average SINR of the three algorithms under comparison as a function of the total transmit power at the RRHs, defined as $P^{\text{tr}} = \sum_{l=1}^{L} P_l^{\text{tr}}$. It can be observed that the proposed ADMM-based Algorithm 3



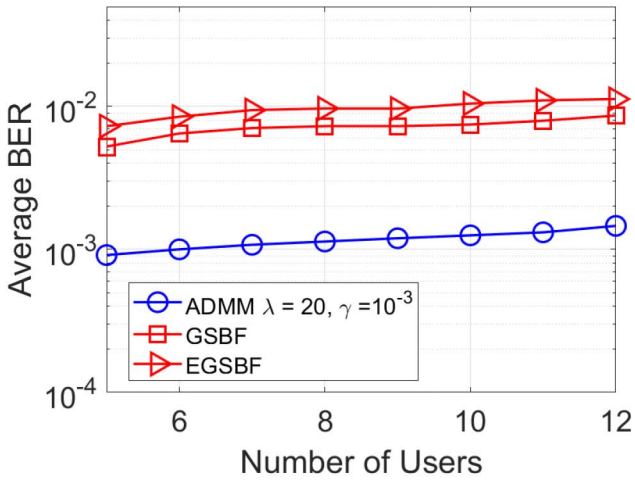
(a) SINR performance versus total transmit power.



(a) SINR performance versus number of users.



(b) Average BER performance versus total transmit power.



(b) Average BER performance versus number of users.

Fig. 8. Comparison of the proposed ADMM Algorithm 3, GSBF in [8] and EGSBF in [11].

Fig. 9. Comparison of the proposed ADMM Algorithm 3, GSBF in [8] and EGSBF in [11] for $P^{\mathrm{tr}} = 8$ *dB*, and $L = 5$.

achieves the highest SINR for all the considered values of P^{tr}. To further evaluate the performance of the proposed method, the average BER of the beamforming solutions designed with the different algorithms is plotted as a function of P^{tr} in Fig. 8(b). Again, the results indicate that our proposed Algorithm 3 can lead to significant performance improvement over GSBF and EGSBF in the design of group sparse beamforming solutions for green C-RAN.

In Fig. 9(a) and 9(b), we show the average SINR and BER versus the number of users for a total transmit power of 8 dB at the RHHs, respectively. We consider $L = 5$ RHHs, with a number of antennas at each RHH equal to the number of users. We can observe from Fig. 9(a) that our design achieves a SINR gain of more than 1.5 dB over GSBF and EGSBF, regardless of the number of users. A performance improvement is also observed in 9(b), where our design achieves a BER performance of 10^{-3} for $K = 6$ users, and therefore significantly outperforms GSBF and EGSBF with BER performance of around 10^{-2}.

VI. Conclusion

We addressed the problem of centralized transceiver design for multiuser MIMO AF relaying within C-RAN, where the aim is to optimize the AF matrices of selected RRHs acting as relays, in order to improve the signal quality at the destination users, while reducing the network power consumption and the fronthaul feedback overhead. A novel two-stage method was conceived to solve the problem with low complexity. The first stage seeks to minimize the interference leakage subject to per-relay transmit power constraints along with signal preserving constraints at the destinations. To reduce the total network power, the objective function was penalized with a regularization term that promotes group-sparsity among the RRHs. A low-complexity iterative algorithm based on the alternating direction method of multipliers (ADMM) was then developed to solve the regularized problem. In the second stage, the interference minimization problem was extended by penalizing the objective function with a weight-level term inducing sparsity in the AF matrix. A low-complexity ADMM-based

algorithm was also conceived to solve this problem, which leads to a reduction in the number of non-zero weights for each active RRH. The performance of the proposed ADMM-based two-stage method was studied by simulations, showing that it can lead to substantial reductions in network power consumption and in computational design complexity over benchmark approaches.

APPENDIX
PROOF OF LEMMA 1

Consider the following minimization problem in $\mathbf{x}$:

$$\min_{\|\mathbf{x}\|_2^2 \le P} \lambda\|\mathbf{x}\|_2 + \frac{\rho}{2}\mathbf{x}^H\mathbf{x} - 2\mathrm{Re}\Big(\mathbf{x}^H\mathbf{a}\Big), \tag{55}$$

which is convex and strictly feasible. Therefore, strong duality holds, i.e., the optimal primal and dual pair $(\mathbf{x}^*, \mu^*)$ must satisfy the following KKT conditions:

$$\begin{cases} \mathbf{0} \in \lambda\partial\|\mathbf{x}^*\|_2 + \frac{\rho}{2}\mathbf{x}^* - \mathbf{a} + \mu^*\mathbf{x}^* \\ \|\mathbf{x}^*\|_2^2 \le P,\ \mu^* \ge 0 \\ \mu^*\big(\|\mathbf{x}^*\|_2^2 - P\big) = 0 \end{cases} \tag{56}$$

where $\partial\|\mathbf{x}^*\|_2$ is the subdifferential of $\|\mathbf{x}\|_2$ at $\mathbf{x}^*$, which is given by

$$\partial\|\mathbf{x}\|_2 = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|_2} & \text{if } \mathbf{x} \ne \mathbf{0}, \\ \{\mathbf{g} \mid \|\mathbf{g}\|_2 \le 1\} & \text{if } \mathbf{x} = \mathbf{0}. \end{cases} \tag{57}$$

To find the optimal solution we distinguish between two cases, namely, when $\mathbf{x}^* \ne \mathbf{0}$ and $\mathbf{x}^* = \mathbf{0}$.

If $\mathbf{x}^* \ne \mathbf{0}$, the first relationship in (56) yields

$$\left(\frac{\lambda}{\|\mathbf{x}^\star\|_2} + \frac{\rho}{2} + \mu^\star\right)\mathbf{x}^\star = \mathbf{a}. \tag{58}$$

In order for (58) to hold, $\mathbf{x}^*$ must take the form of $\mathbf{x}^* = z\mathbf{a}$ for some $z > 0$. Substituting $\mathbf{x}^* = z\mathbf{a}$ back into (58), we find

$$z = \frac{\|\mathbf{a}\|_2 - \lambda}{\|\mathbf{a}\|_2\big(\frac{\rho}{2} + \mu^\star\big)}. \tag{59}$$

Since $z > 0$, we must have $\|\mathbf{a}\|_2 \ge \lambda$. Therefore, we have

$$\mathbf{x}^* = \frac{\mathbf{a}}{\|\mathbf{a}\|_2\big(\frac{\rho}{2} + \mu^*\big)}(\|\mathbf{a}\|_2 - \lambda), \quad \text{if } \|\mathbf{a}\|_2 \ge \lambda. \tag{60}$$

Now, if $\mathbf{x}^* = \mathbf{0}$, the first relationship in (56) becomes $\mathbf{0} \in \lambda\|\mathbf{g}\|_2 - \mathbf{a}$, which means

$$\mathbf{x}^* = \mathbf{0}, \quad \text{if } \|\mathbf{a}\|_2 < \lambda. \tag{61}$$

Combining (60) and (61), $\mathbf{x}^*$ can be written as

$$\mathbf{x}^* = \frac{\mathbf{a}}{\|\mathbf{a}\|_2\big(\frac{\rho}{2} + \mu^*\big)}[\|\mathbf{a}\|_2 - \lambda]_+. \tag{62}$$

To solve for the optimal dual variable μ^*, we substitute (62) back into the second inequality in (56) and obtain

$$\mu^* \ge \frac{\|\mathbf{a}\|_2 - \lambda}{\sqrt{P}} - \frac{\rho}{2}. \tag{63}$$

Since $\mu^* \ge 0$, we must have

$$\mu^* = \begin{cases} \frac{\|\mathbf{a}\|_2 - \lambda}{\sqrt{P}} - \frac{\rho}{2} & \text{if } \frac{\|\mathbf{a}\|_2 - \lambda}{\sqrt{P}} - \frac{\rho}{2} > 0 \\ 0 & \text{if } \frac{\|\mathbf{a}\|_2 - \lambda}{\sqrt{P}} - \frac{\rho}{2} \le 0. \end{cases} \tag{64}$$

In conclusion, the optimal variables may be succinctly expressed by

$$\mathbf{x}^* = \frac{\mathbf{a}}{\|\mathbf{a}\|_2\big(\frac{\rho}{2} + \mu^*\big)}[\|\mathbf{a}\|_2 - \lambda]_+,\ \mu^* = \left[\frac{\|\mathbf{a}\|_2 - \lambda}{\sqrt{P}} - \frac{\rho}{2}\right]_+. \tag{65}$$

REFERENCES

[1] "3rd Generation partnership project; technical report; Small cell enhancements for E-UTRA and E-UTRAN—Physical layer aspects (release 12)," 3GPP, Sophia Antipolis, France, Rep. TR 36.872, 2013.

[2] J. Liu, M. Sheng, L. Liu, and J. Li, "Network densification in 5G: From the short-range commumnications perspective," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 96–102, Dec. 2017.

[3] A. Morsali, S. S. Hosseini, B. Champagne, and X. Chang, "Design criteria for omnidirectional STBC in massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1435–1439, Oct. 2019.

[4] A. Morsali, A. Haghighat, and B. Champagne, "Realizing fully digital precoders in hybrid A/D architecture with minimum number of RF chains," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2310–2313, Oct. 2017.

[5] A. Checko *et al.*, "Cloud RAN for mobile network—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[6] C. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: A 5G perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.

[7] (Dec. 2013). *C-RAN White Paper: The Road Towards Green RAN*. [Online]. Available: http://labs.chinamobile.com/cran

[8] 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.

[9] Y. Shi, J. Zhang, and K. B. Letaief, "Robust group sparse beamforming for multicast green cloud-RAN with imperfect CSI," *IEEE Trans. Signal Process.*, vol. 63, no. 17, pp. 4647–4659, Sep. 2015.

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

[11] Y. Shi, J. Zhang, W. Chen, and K. B. Letaief, "Enhanced group sparse beamforming for green cloud-RAN: A random matrix approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2511–2524, Apr. 2018.

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

[13] L. Liang and Z. Rui, "Downlink SINR balancing in C-RAN under limited fronthaul capacity," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2016, pp. 3506–3510.

[14] V. N. Ha and L. B. Le, "Joint coordinated beamforming and admission control for fronthaul constrained cloud-RANs," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 4054–4059.

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

[16] S. Ali, A. Ahmad, R. Iqbal, S. Saleem, and T. Umer, "Joint RRH-association, sub-channel assignment and power allocation in multi-tier 5G C-RANs," *IEEE Access*, vol. 6, pp. 34393–34402, 2018.

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

[18] L. Liu and R. Zhang, "Optimized uplink transmission in multi-antenna C-RAN with spatial compression and forward," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5083–5095, Oct. 2015.

[19] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[20] "Technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); relay architectures for E-UTRA (LTE-advanced) (Release 9)," 3GPP, Sophia Antipolis, France, Rep. TR 36.806, 2009.

[21] S. W. Peters, A. Y. Panah, K. T. Truong, and R. W. Heath, "Relay architectures for 3GPP LTE-Advanced," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, no. 1, Jul. 2009, Art. no. 618787.

[22] C. Nie, P. Liu, T. Korakis, E. Erkip, and S. S. Panwar, "Cooperative relaying in next-generation mobile WiMAX networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 3, pp. 1399–1405, Mar. 2013.
[23] Y. Hu, M. C. Gursoy, and A. Schmeink, "Relaying-enabled ultra-reliable low-latency communications in 5G," *IEEE Netw.*, vol. 32, no. 2, pp. 62–68, Mar. 2018.
[24] Y. Zhou and Y. Chen, "Novel energy-harvesting AF relaying schemes with channel estimation errors," *IEEE Syst. J.*, vol. 14, no. 1, pp. 333–342, Mar. 2020.
[25] Y. Shao and T. A. Gulliver, "Precoding design for two-way MIMO full-duplex amplify-and-forward relay communication systems," *IEEE Access*, vol. 7, pp. 76458–76469, 2019.
[26] D. Chen, Y. Cheng, X. Wang, W. Yang, J. Hu, and Y. Cai, "Energy-efficient secure multiuser scheduling in energy harvesting untrusted relay networks," *J. Commun. Netw.*, vol. 21, no. 4, pp. 365–375, Aug. 2019.
[27] J. Ma, C. Huang, S. Cui, and Q. Li, "Energy efficiency of amplify-and-forward full-duplex relay channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1365–1368, Oct. 2019.
[28] B. Rong, Z. Zhang, X. Zhao, and X. Yu, "Robust superimposed training designs for MIMO relaying systems under general power constraints," *IEEE Access*, vol. 7, pp. 80404–80420, 2019.
[29] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 118–127, Jun. 2014.
[30] S. Park, K. Lee, C. Song, and I. Lee, "Joint design of fronthaul and access links for C-RAN with wireless fronthauling," *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1657–1661, Nov. 2016.
[31] J. Yang, B. Champagne, Y. Zou, and L. Hanzo, "Centralized energy-efficient multiuser multiantenna relaying in next-generation radio access networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7913–7924, Sep. 2017.
[32] A. Saab, J. Yang, B. Champagne, and I. Psaromiligkos, "Efficient group-sparse transceiver design for multiuser MIMO relaying in C-RAN," in *Proc. IEEE 88th Veh. Tech. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–6.
[33] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
[34] P. Mohapatra, K. E. Nissar, and C. R. Murthy, "Interference alignment algorithms for the K user constant MIMO interference channel," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5499–5508, Nov. 2011.
[35] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc. B (Stat. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.
[36] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[37] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
[38] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. Hoboken, NJ, USA: Wiley, 2003.
[39] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
[40] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, nos. 3–4, pp. 591–611, 1965.
[41] "Sumitomo electric develops high efficiency broadband power amplifier compliant with next generation mobile communication standards," Sumitomo Elect. Ind., Ltd., Osaka, Japan, Feb. 2013. [Online]. Available: http://global-sei.com/news/press/13/prs012_s.html

Jiaxin Yang received the B.Eng. degree in information engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009, the M.E.Sc. degree in electrical and computer engineering from the University of Western Ontario, London, ON, Canada, in 2011, and the Ph.D. degree in electrical and computer engineering from McGill University, Montreal, QC, Canada. His research interests include optimization theory, statistical signal processing, detection and estimation, and the applications thereof in wireless communications, such as MIMO systems, cooperative communications, and physical-layer security. He is a recipient of several awards and scholarships, including the Best Paper Award of 27th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, the McGill Engineering Doctoral Award, the FRQNT International Internship Scholarship, and the NSERC CREATE Ph.D. Scholarship.

Ayoub Saab received the M.Eng. degree (with thesis) from McGill University in 2018. His research interests include optimization theory, statistical signal processing, and their applications in wireless communications.

Alireza Morsali (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Shahid Bahonar University, Kerman, Iran, in 2009 and 2011, respectively. He is current pursuing the Ph.D. degree with McGill University, Montreal, QC, Canada. His research interests include coding theory, signal processing, wireless communications, artificial intelligence, and blockchain. He is a recipient of several awards and scholarships, including, the McGill Engineering Doctoral Award, the STARaCom Collaborative Grant and the BLUE Fellowship.

Benoit Champagne (Senior Member, IEEE) received the B.Ing. degree in engineering physics from the École Polytechnique of Montreal in 1983, the M.Sc. degree in physics from the University of Montreal in 1985, and the Ph.D. degree in electrical engineering from the University of Toronto in 1990. From 1990 to 1999, he was an Assistant and then an Associate Professor with INRS Telecommunications, Montreal. In 1999, he joined McGill University, where he is currently a Full Professor with the ECE Department. His research interests include statistical signal processing and wireless communications, where he has coauthored more than 300 publications. He has been an Associate Editor of IEEE SIGNAL PROCESSING LETTERS and the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Ioannis Psaromiligkos (Member, IEEE) received the Diploma degree in computer engineering and science from the University of Patras, Patras, Greece, in 1995, and the M.S. and Ph.D. degrees in electrical engineering from the State University of New York, Buffalo, NY, USA, in 1997 and 2001, respectively. Since 2001, he has been with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada, where he is currently an Associate Professor. His research interests are in the areas of statistical detection and estimation, adaptive signal processing, machine learning, and wireless communications. He has served as an Associate Editor for IEEE COMMUNICATIONS LETTERS and IEEE SIGNAL PROCESSING LETTERS. He is currently an Associate Editor of the *EURASIP Journal on Advances in Signal Processing*.