IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 12, 2021, accepted August 1, 2021, date of publication August 11, 2021, date of current version August 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104259

Energy-Efficient User Clustering and Downlink Beamforming for MIMO-SCMA in C-RAN

SARA NOROUZI[1], YUNLONG CAI[2], (Senior Member, IEEE), AND BENOIT CHAMPAGNE[1], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0E9, Canada
[2]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China

Corresponding author: Sara Norouzi (sara.norouzi@mcgill.ca)

This work was supported by the Collaborative Research and Development Program (CRD) Grant through the Natural Sciences and Engineering Research Council (NSERC) Canada, with industrial sponsor InterDigital.

ABSTRACT Multiple-input multiple-output (MIMO) sparse code multiple access (SCMA) is of great interest for future wireless networks to achieve higher spectral efficiency and support massive connectivity. In this paper, we investigate the key problems of user clustering and downlink beamforming for MIMO-SCMA in a cloud radio access network (C-RAN). Using channel state information available at the central processor, an efficient user clustering algorithm based on the constrained K-means method is proposed. Subsequently, two iterative algorithms for beamforming design are developed by minimizing the total transmission power under quality-of-service (QoS) and fronthaul capacity constraints. In the first approach, we approximate the continuous non-convex constraints by convex conic ones using first-order Taylor expansion and iteratively solve a sequence of mixed-integer second order cone programs (MI-SOCPs) to achieve high quality solution, but with higher complexity. In the second approach, a two-stage low-complexity solution is developed in which beamforming matrices obtained from each stage are combined to form a single beamformer for each user. In the first stage, *cluster* beamformers are designed by taking advantage of block diagonalization, while in the second stage, *user-specific* beamformers are determined by minimizing transmission power. The performance of the proposed user clustering and downlink beamforming approaches for MIMO-SCMA in C-RAN is validated through simulations over mmWave channels. Compared to benchmark approaches, the results show significant improvements in terms of transmit power and spectral efficiency.

INDEX TERMS MIMO-SCMA, mmWave, C-RAN, downlink beamforming, user clustering, constrained K-means.

I. INTRODUCTION

In mobile wireless networks, multiple access technologies are of crucial importance to meet performance requirements in terms of data throughput, network capacity, device connectivity and energy consumption. Recently, the application of non-orthogonal multiple access (NOMA) techniques to fifth generation (5G) and beyond 5G (B5G) wireless networks has received considerable attention. In effect, NOMA allows multiple users to access overlapping time and frequency resource elements in the same spatial layer [1]. Hence, this technology has the potential to provide higher spectral efficiency and meet the massive connectivity demand needed for machine-to-machine (M2M) communications and internet of things (IoT) in future wireless networks [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Jiayi Zhang.

NOMA techniques can be classified into three main categories, namely: code domain, power domain and multiple domain [3]. In code domain NOMA, different codes are applied to modulate the data streams of the users over multiple resource elements in a sparse manner. Hence, the processed data of different users can be multiplexed over the same resource elements, wherein the induced sparsity allows the control of interference. Power domain NOMA relies on the use of superposition coding strategies, wherein user signals are simultaneously broadcast with different power levels at the transmitter, while successive interference cancellation (SIC) techniques are employed to separate them at the receiver. In multiple domain NOMA, such as pattern division multiple access (PDMA) and lattice partition multiple access (LPMA), multiple user signals are superimposed in multiple domains, including power, code, and spatial domains.

Sparse code multiple access (SCMA) is a code domain NOMA scheme inspired from the well-known code division multiple access (CDMA) technique. While CDMA extends each information symbol (taken, e.g., from a quadrature amplitude modulation (QAM) constellation) into a finite sequence of complex symbols by using orthogonal or near orthogonal spreading codes, SCMA directly maps each group of bits into a sequence of complex symbols by merging together the symbol mapper and the CDMA spreader. The overall process can be interpreted as a coding procedure from the binary domain to a multidimensional complex domain, which in turn raises new problems in terms of codebook design [4].

As an emerging network architecture for 5G and B5G, cloud radio access network (C-RAN) offers several benefits, e.g., improved energy efficiency, ability to handle interference on a larger scale and increased network capacity [5]. The C-RAN architecture consists of three main components, namely: the central processor, the remote radio heads (RRH) and the fronthaul links. The central processor, which is located in one or more data centers within the cloud, is responsible for all the baseband processing. The RRHs connect wireless devices to the network, alike base stations in current cellular networks. The fronthaul link provides connectivity (e.g., via dedicated optical fiber or microwave links) between the central processor and the RRHs. The C-RAN architecture concentrates the baseband processing in the central processor and coordinates the operation of the RRHs. This separation of the central processor and RRHs functionalities reduces the power consumption and complexity of the RRHs, since the latter only need to perform basic transceiver operations.

A. RELATED WORKS

There have been extensive studies devoted to the design of multidimensional constellations for downlink and uplink SCMA systems. In [6], the performance of a systematic sub-optimal design for the mother constellations (from which the individual user codebooks are derived) is investigated and a unified metric is proposed to obtain the optimum codebooks using a specific mother constellation. The authors in [7] evaluate the average bit error rate (BER) performance of SCMA systems in which codebooks are based on star-QAM signaling constellations. Multidimensional constellations with a low number of projections are designed in [8] based on the extrinsic information transfer (EXIT) chart using a multistage optimization. Subsequently, an appropriate labeling method based on the EXIT chart is optimized for the resulting constellation. In [9], the design of SCMA codebooks based on star-QAM constellations is addressed and an analytical approach to obtain the theoretical BER performance over Rayleigh fading channels is proposed. The design of an efficient suboptimal SCMA codebook is proposed in [10] for a large scale scenario with growing number of resources and users.

The use of multiple antennas along with multiple-input multiple-output (MIMO) techniques can lead to significant performance improvements in terms of user capacity, spectral efficiency, and peak data rates, by taking advantage of spatial diversity, multiplexing or beamforming gains. In [11], a joint sparse graph is constructed for a MIMO-SCMA system model, and the corresponding virtual SCMA codebooks are designed for the detector, wherein the message passing algorithm (MPA) is employed to reconstruct the transmitted data bits. In [12], a joint decoding algorithm is proposed for MIMO-SCMA systems based on space frequency block codes (SFBC), which exhibits lower computational complexity than MPA and yet achieves a similar block error rate (BLER). A novel downlink MIMO mixed-SCMA scheme is proposed in [13], such that the transmitted codewords for each user over different antennas come from different codebooks. The authors in [14] propose near-optimal low-complexity iterative receivers based on factor graph for a downlink MIMO-SCMA system over frequency selective fading channels.

Recently, the C-RAN architecture has aroused great interest for the implementation of MIMO-NOMA transmission schemes. In [15], a novel framework for C-RAN is proposed in which two users are scheduled over the same resources according to power domain NOMA, while the performance of cell-edge users is further enhanced by means of coordinated beamforming. Stochastic geometry is used to analyze the outage probability of NOMA under C-RAN in [16], where power domain multipexing along with SIC are employed to increase downlink system capacity. The application of beamforming along with power domain NOMA is investigated for cache-enabled C-RAN in [17]. The design of robust radio resource allocation and beamforming approaches for MIMO-SCMA systems under C-RAN is studied in [18], where the aim is to maximize the total sum rate of users subject to a minimum required rate for each slice.

B. MOTIVATIONS AND CONTRIBUTIONS

MIMO-SCMA combines MIMO techniques, which increase capacity by transmitting different signals over multiple antennas, and SCMA which improves spectral efficiency and device connectivity by transmitting multiple user signals over the same radio resources. As seen in works related to power domain NOMA [19], [20], the joint application of spatial user clustering along with beamforming techniques in MIMO-SCMA systems has the potential to improve spectral efficiency and reduce the total transmit power. Additionally, when considered within a C-RAN architecture, this approach makes it possible to increase the number of supported users in the network by using a common codebook for users in different clusters, while the effect of inter-cluster interference can be eliminated by centralized beamformer design and coordinated RRH operation. In spite of its importance, the joint problem of user clustering and beamforming has not received considerable attention in the literature on MIMO-SCMA, let alone C-RAN.

Motivated by the above considerations, we propose energy-efficient user clustering and downlink beamforming approaches for MIMO-SCMA in C-RAN. Our main contributions in addressing the above challenges are summarized as follows:

1) We approach the user clustering problem by modifying the widely-used K-means method from the field of machine learning, in order to limit the number of users in each cluster. Specifically, the proposed *constrained* K-means algorithm uses the Euclidian metric to characterize the similarities between the user channel vectors and the cluster centers, and seeks to group users with channel vectors exhibiting large correlation. The elbow method is utilized to find the optimum number of clusters for the network.
2) We formulate the beamforming design and RRH selection as a non-convex mixed-integer nonlinear programing (MINLP) optimization problem, aiming to minimize the total transmit power while satisfying the signal-to-interference-plus-noise ratio (SINR) and fronthaul capacity constraints. We then propose transformations to reformulate the problem as a difference of convex functions (DC) program and derive two algorithms for solving the problem. In the first algorithm, we iteratively approximate the continuous non-convex constraints by convex ones using first-order Taylor expansion and solve a sequence of mixed-integer second-order cone programing (MI-SOCP) using dedicated solvers. This algorithm entails high computational complexity, yet it can achieve high quality solution.
3) The second algorithm is based on a two-stage low-complexity beamforming approach wherein the beamforming matrices obtained from each stage are multiplied to form the final beamformer. In the first stage, specifically, a block diagonalization (BD) technique is adopted to design the cluster beamformers (one for each cluster), which remove the inter-cluster interference and thus enhance the quality-of-service (QoS) for intra-cluster users. In the second stage, the user-specific beamformers are designed along with RRH selection by employing a smoothed ℓ_0-norm approximation. The resulting optimization problem is solved via the convex-concave procedure (CCCP) with guaranteed convergence [21].
4) We evaluate the performance of the proposed algorithms for user clustering and downlink beamforming using in-depth simulations of MIMO-SCMA in C-RAN with mmWave channel models and different parameter configurations. The results illustrate the convergence behavior of the new algorithms and the effect of various parameters on the system performance, while providing useful insights into the advantages of the proposed approaches over competing ones from the literature.

C. ORGANIZATION

The rest of the paper is organized as follows: Section II introduces the MIMO-SCMA system model under C-RAN and describes the problem under consideration. The proposed constrained K-means algorithm for user clustering is introduced in Section III. The two-stage energy-efficient beamforming approach for eliminating inter-cluster interference and minimizing total transmit power is developed In Section IV. The results of our simulation experiments are presented in Section V, followed by the conclusion in Section VI.

Notations: Scalars, vectors and matrices are respectively denoted by lower case, boldface lower case and boldface upper case letters. For a matrix $\mathbf{A}$, $[\mathbf{A}]_{i,j}$ denotes its (i,j)th entry, while $\mathbf{A}^T$ and $\mathbf{A}^H$ denote its transpose and conjugate transpose, respectively. The operators $\|.\|_2$ and $\|.\|_0$ denote the Euclidean and zero norms of a vector, respectively. For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes its cardinality. $\mathbb{C}^{m\times n}$ ($\mathbb{R}^{m\times n}$) denotes the space of $m \times n$ complex (real) matrices. $\mathbb{B}^{m\times n}$ denotes binary matrices of size $m\times n$ where the set $\mathbb{B} = \{0, 1\}$. We use $\mathcal{CN}(\mu, \sigma^2)$ to denote a complex circular Gaussian random variable with mean μ and variance σ^2.

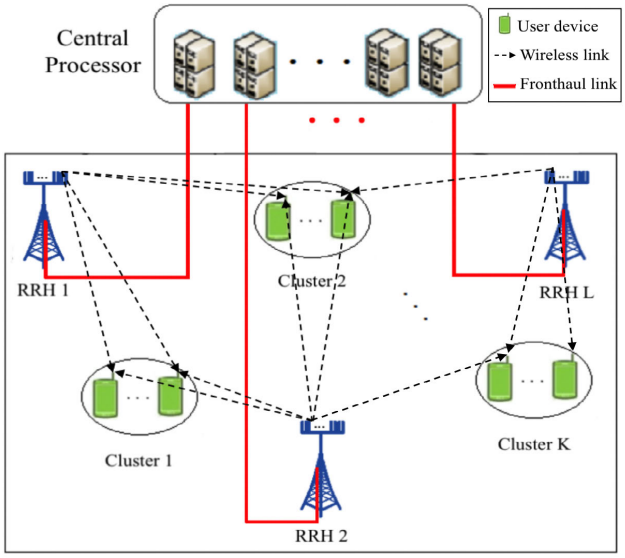


FIGURE 1. The MIMO-SCMA system model under C-RAN.

II. SYSTEM MODEL AND PROBLEM DESCRIPTION

We consider downlink transmission in a MIMO-SCMA system under C-RAN, as illustrated in Figure 1. The system consists of L RRHs, each equipped with M antennas, and J single-antenna users. The RRHs indexed by $l \in \mathcal{L} \triangleq \{1, \ldots, L\}$, are connected to the central processor via limited-capacity fronthaul links. Due to the fronthaul constraint, each user is cooperatively served by a specific subset of RRHs through joint beamforming. Moreover, the users are partitioned into K non-overlapping clusters, indexed by $k \in \mathcal{K} \triangleq \{1, \ldots, K\}$ with the kth cluster comprising J_k users

such that $J = \sum_{k=1}^{K} J_k$. Below, we provide further details on the SCMA encoder, mmWave channel, received signal model, and problem description. For convenience, we list the key notations of this paper in Table 1.

A. SCMA ENCODER

In SCMA, contiguous groups of data bits from each user are directly mapped to sparse N-dimensional codewords selected from a predefined codebook and then transmitted over N radio resources, e.g., orthogonal frequency division multiple access (OFDMA) subcarriers. The SCMA encoder for the ith user can be defined as $f_i : \mathbb{B}^U \to \mathcal{X}_i$ which is a one-to-one mapping from the set of U-bit tuples to a codebook $\mathcal{X}_i \subset \mathbb{C}^N$ of N-dimensional codewords, with cardinality $|\mathcal{X}_i| = 2^U$. Specifically, for $\mathbf{b} = [b_1, \ldots, b_U] \in \mathbb{B}^U$, the corresponding codeword is obtained as,

$$\mathbf{x} = f_i(\mathbf{b}) = [x(1), \ldots, x(N)] \tag{1}$$

where $\mathbf{x}$ is a sparse vector with $C < N$ non-zero elements.

Each user is assigned C subcarriers such that no two users occupy the same set of subcarriers. Hence, only q users can be supported by SCMA, as given by [22],

$$q = \binom{N}{C} = \frac{N!}{C!(N-C)!}. \tag{2}$$

In this work, we group users into K clusters of size $J_k \leq q$ and remove inter-cluster interference so that the users in different clusters can use common codebooks.

TABLE 1. Summary of key notations.

Notation	Description
L	Number of RRHs
M	Number of transmit antennas per RRH
K	Number of non-overlapping clusters
J	Total number of users
J_k	Number of users in the kth cluster
$\mathcal{L}, \mathcal{J}, \mathcal{K}$	Index set of RRHs/users/clusters
N	Number of subcarriers
C	Number of non-zero elements for each codeword
q	Maximum number of supported users via SCMA
$\mathbf{F}$	Factor graph matrix
P	Number of NLOS paths in mmWave channel
$\alpha^{(p)}, a_{jk}^{(lp)}(n), \theta_{jk}^{(lp)}$	Path loss exponent/complex gain/normalized direction for the pth path in mmWave channel
$\mathbf{h}_{jk}(n), \mathbf{w}_{jk}(n)$	Network-wide channel/beamforming vector for the jth user in the kth cluster over the nth subcarrier
σ_{jk}^2	Noise power
$\mathbf{d}_j$	Normalized channel vector for clustering
$\mathbf{c}_k$	Center of the kth cluster
W_K	Sum of the normalized within-cluster SSE distance
$\gamma_{jk}(n)$	SINR of the jth user in the kth cluster over the nth subcarrier
$\gamma_{\min}$	Minimum required SINR
$R_{jk}(n)$	Transmission rate of the jth user in the kth cluster over the nth subcarrier
$C_{\max}$	Maximum capacity constraint for each RRH
$P_{\max}$	Maximum available total transmit power
$\mathbf{B}_k(n)$	First stage (cluster) beamforming matrix
$\mathbf{v}_{jk}(n)$	Second stage (user-specific) beamforming vector

Referring to (1), we can associate to each codeword $\mathbf{x}$ a vector $\mathbf{y}$ containing its C non-zero elements in the same order, i.e., $\mathbf{y}$ is obtained from $\mathbf{x}$ by removing its zero elements. For convenience, we represent this operation by the function $\phi : \mathbb{C}^N \to \mathbb{C}^C$, so that $\mathbf{y} = \phi(\mathbf{x}) = [y(1), \ldots, y(C)]$. Through this operation, the original codebook $\mathcal{X}_i \subset \mathbb{C}^N$ is transformed into a constellation of C-dimensional codewords, i.e., $\mathcal{Y}_i \subset \mathbb{C}^C$, where $\mathcal{Y}_i = \{\phi(\mathbf{x}) : \mathbf{x} \in \mathcal{X}_i\}$. We also let $g_i = \phi \circ f_i : \mathbb{B}^U \to \mathcal{Y}_i$ denote the composite mapping of f_i and ϕ, so that for any $\mathbf{b} \in \mathbb{B}^U$, and $\mathbf{x} = f_i(\mathbf{b})$, we have,

$$\mathbf{y} = \phi(\mathbf{x}) = g_i(\mathbf{b}). \tag{3}$$

From this perspective, the SCMA encoder can be redefined as $f_i(\mathbf{b}) = \mathbf{S}_i g_i(\mathbf{b})$, where matrix $\mathbf{S}_i \in \mathbb{B}^{N \times C}$ maps a C-dimensional constellation point to an N-dimensional codeword. Note that $\mathbf{S}_i$ contains $N - C$ all-zero rows and hence, all the codewords in codebook $\mathcal{X}_i$ contain 0 in the same $N - C$ positions. Moreover, an identity matrix of order C is obtained by removing the all-zero rows from $\mathbf{S}_i$.

The set of resources occupied by user i is determined by the positions (or indices) of the non-zero elements of the binary indicator vector $\mathbf{f}_i = \text{diag}(\mathbf{S}_i \mathbf{S}_i^T) \in \mathbb{B}^{N \times 1}$. In effect, the complete SCMA encoder structure for q users and N subcarriers can be represented by a factor graph, with associated matrix $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_q] \in \mathbb{B}^{N \times q}$. In this interpretation, subcarrier node n and user node i are connected if and only if the corresponding element of matrix $\mathbf{F}$ is equal to 1, i.e., $[\mathbf{F}]_{n,i} = 1$.

B. CHANNEL MODEL

Due to the propagation characteristics at such high frequencies, the application of MIMO-SCMA communication in the mmWave band is more challenging than in a conventional low-frequency scenario. The mmWave-based channel vector $\mathbf{h}_{jk}^{(l)}(n) \in \mathbb{C}^{1 \times M}$ from the lth RRH to the jth user in the kth cluster over the nth subcarrier can be expressed as the discrete sum of a line-of-sight (LOS) and P non line-of-sight (NLOS) components [23], [24], i.e.,

$$\mathbf{h}_{jk}^{(l)}(n) = \sum_{p=0}^{P} \frac{\sqrt{M} a_{jk}^{(lp)}(n) \mathbf{a}(\theta_{jk}^{(lp)})}{\sqrt{P+1}(1 + (d_{jk}^{(l)})^{\alpha^{(p)}})} \tag{4}$$

where: p is the path index, with $p = 0$ corresponding to LOS and $p \geq 1$ to NLOS paths; $d_{jk}^{(l)}$ is the distance between the RRH and the user; $\alpha^{(p)}$ is the path loss exponent; $a_{jk}^{(lp)}(n)$ denotes the complex gain for the pth path which follows a complex circular Gaussian distribution, i.e., $a_{jk}^{(lp)}(n) \sim \mathcal{CN}(0, 1)$; and $\mathbf{a}(\theta_{jk}^{(lp)}) \in \mathbb{C}^{1 \times M}$ is the antenna array steering vector. In the case of a uniform linear antenna array, the steering vector is given by,

$$\mathbf{a}(\theta_{jk}^{(lp)}) = \frac{1}{\sqrt{M}} [1, e^{-j\pi\theta_{jk}^{(lp)}}, \ldots, e^{-j\pi(M-1)\theta_{jk}^{(lp)}}] \tag{5}$$

where $\theta_{jk}^{(lp)}$ is the normalized direction of the pth path. The latter can be expressed as,

$$\theta_{jk}^{(lp)} = \frac{2d}{\lambda} \sin(\phi_{jk}^{(lp)}) \tag{6}$$

where $\phi_{jk}^{(lp)} \in [0, 2\pi]$ is the angle of departure (AoD) of the pth path, d is the inter-antenna element spacing, and λ is the wavelength at the operating frequency.

In MIMO systems operating at mmWave frequencies, a single-path model is often adopted for the channel vectors by retaining only one dominant path in (4) [25]. In most cases, the latter will be the LOS path, whose gain can be as much as 20dB stronger than that of NLOS paths [26]. However, when there is no LOS path due to blockage, the dominant NLOS path can be considered instead. Hence, the mmWave channel model can be simplified to,

$$\mathbf{h}_{jk}^{(l)}(n) = \frac{\sqrt{M} a_{jk}^{(l)}(n)}{(1 + (d_{jk}^{(l)})^{\alpha})} \mathbf{a}(\theta_{jk}^{(l)}) \tag{7}$$

where, for simplicity of notation, the superscript p for the path index has been removed.

C. SIGNAL MODEL

Let $x_{jk}(n) \in \mathbb{C}$ denote the codeword element intended for the jth user in the kth cluster over the nth subcarrier. Due to the sparsity of the SCMA encoder, $x_{jk}(n)$ can be either 0, or a non-zero element with normalized power, i.e., $E\{|x_{jk}(n)|^2\} = 1$. Codeword element $x_{jk}(n)$ is transmitted from the M antennas of the lth RRH by employing the beamforming vector $\mathbf{w}_{jk}^{(l)}(n) \in \mathbb{C}^{M \times 1}$. Hence, the transmit signal of the lth RRH over the nth subcarrier can be expressed as,

$$\mathbf{z}^{(l)}(n) = \sum_{k=1}^{K} \sum_{j \in \mathcal{U}_{n,k}} \mathbf{w}_{jk}^{(l)}(n) x_{jk}(n) \tag{8}$$

where $\mathcal{U}_{n,k}$ denotes the set of users in the kth cluster occupying the nth subcarrier. Owing to the limited-capacity fronthaul link, only a selected group of RRHs serve a specific user cooperatively. The process of RRH selection for transmission can be performed through beamforming. That is, $\|\mathbf{w}_{jk}^{(l)}(n)\|_2 = 0$ implies that the lth RRH does not participate in the transmission for that user over its assigned subcarrier. Hence, the corresponding network-wide beamforming vector, $\mathbf{w}_{jk}(n) = [\mathbf{w}_{jk}^{(1)}(n)^T, \ldots, \mathbf{w}_{jk}^{(L)}(n)^T]^T \in \mathbb{C}^{LM \times 1}$ may be sparse.

Let $\mathbf{h}_{jk}(n) = [\mathbf{h}_{jk}^{(1)}(n), \ldots, \mathbf{h}_{jk}^{(L)}(n)] \in \mathbb{C}^{1 \times LM}$ denote the network-wide channel vector for the jth user in the kth cluster and $\mathbf{z}(n) = [\mathbf{z}^{(1)}(n)^T, \ldots, \mathbf{z}^{(L)}(n)^T]^T \in \mathbb{C}^{LM \times 1}$ denote the network-wide transmit signal over the nth subcarrier. The received signal at the jth user in the kth cluster over the nth subcarrier is given by,

$$r_{jk}(n) = \mathbf{h}_{jk}(n)\mathbf{z}(n) + n_{jk} \tag{9}$$

where $n_{jk} \sim \mathcal{CN}(0, \sigma_{jk}^2)$ is an additive noise term. We can express the received signal of this user as a sum of the desired signal, the interference from the other users in that cluster (intra-cluster interference), the inter-cluster interference and the noise, i.e.,

$$\begin{aligned} r_{jk}(n) = {} & \mathbf{h}_{jk}(n)\mathbf{w}_{jk}(n)x_{jk}(n) \\ & + \underbrace{\sum_{j' \neq j, j' \in \mathcal{U}_{n,k}} \mathbf{h}_{jk}(n)\mathbf{w}_{j'k}(n)x_{j'k}(n)}_{\text{Intra-cluster Interference}} \\ & + \underbrace{\sum_{k' \neq k} \sum_{j \in \mathcal{U}_{n,k'}} \mathbf{h}_{jk}(n)\mathbf{w}_{jk'}(n)x_{jk'}(n)}_{\text{Inter-cluster Interference}} + n_{jk}. \end{aligned} \tag{10}$$

The SINR of the jth user in the kth cluster over the nth subcarrier with non-zero codeword element is given by,

$$\gamma_{jk}(n) = \frac{|\mathbf{h}_{jk}(n)\mathbf{w}_{jk}(n)|^2}{I_{jk}^{(1)}(n) + I_{jk}^{(2)}(n) + \sigma_{jk}^2} \tag{11}$$

where the first term in the denominator represents the intra-cluster interference and the second term represents the inter-cluster interference, i.e.,

$$I_{jk}^{(1)}(n) = \sum_{j' \neq j, j' \in \mathcal{U}_{n,k}} |\mathbf{h}_{jk}(n)\mathbf{w}_{j'k}(n)|^2 \tag{12}$$

$$I_{jk}^{(2)}(n) = \sum_{k' \neq k} \sum_{j' \in \mathcal{U}_{n,k'}} |\mathbf{h}_{jk}(n)\mathbf{w}_{j'k'}(n)|^2. \tag{13}$$

The total transmit power for the whole network over N subcarriers is given by,

$$P_T = \sum_{n=1}^{N} E\{\mathbf{z}(n)^H \mathbf{z}(n)\} = \sum_{n=1}^{N} \sum_{l=1}^{L} E\{\mathbf{z}^{(L)}(n)^H \mathbf{z}^{(L)}(n)\}. \tag{14}$$

Upon substitution of (8) into (14) and assuming that the transmitted codewords $x_{jk}(n)$ from different sources are uncorrelated and have zero mean and unit variance, we can write the total transmit power as,

$$P_T = \sum_{n} \sum_{l} \sum_{k} \sum_{j} \|\mathbf{w}_{jk}^{(l)}\|^2 = \sum_{n} \sum_{k} \sum_{j} \|\mathbf{w}_{jk}\|^2, \tag{15}$$

where the last equality follows from the definition of the network-wide beamforming vector.

D. PROBLEM DESCRIPTION

In this work, our objective is to group users into non-overlapping clusters and design beamformers such that the total transmit power is minimized while constraining the inter-cluster interference, the user SINRs and the fronthaul capacity. Indeed, removing inter-cluster interference not only enhances the SINR at the user terminal, but also allows the transmitter to use a common SCMA codebook to serve users in different clusters, which in turn boosts network capacity. To further satisfy the requirements imposed by the limited-capacity fronthaul links of C-RAN, dynamic RRH selection is taken into consideration in our formulation.

In order to address the above challenges and obtain the desire solution, we conceive efficient algorithms for user clustering and beamforming design with low complexity.

Specifically, we propose an efficient user clustering algorithm based on the constrained K-means method in Section III. Then, the beamformer design is addressed in Section IV by means of a two-stage energy-efficient approach wherein the inter-cluster interference is removed using a BD technique in the first stage and the total transmit power is optimized under SINR and fronthaul capacity constraints in the second stage.

III. USER CLUSTERING

In this section, we first introduce the proposed constrained K-means algorithm for user clustering. We then apply the elbow method to determine the number of clusters. Finally, we evaluate the computational complexity of the proposed algorithm.

A. CONSTRAINED K-MEANS CLUSTERING

K-means is a celebrated method for grouping inharmonious multi-dimensional data points into K clusters such that a similarity criterion within clusters is maximized [27], [28]. In effect, K-means attempts to group J data points (or vectors) $\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_J\}$ into K clusters by finding cluster centers $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$ such that similarities between the points in the same group are high while similarities between the points in different groups are low. Two key factors in the K-means method are the number of clusters K, which is predetermined, and the similarity metric [29].

In the current MIMO-SCMA application, high correlation between the channel vectors of the users in a cluster can provide a better beamforming performance. Indeed, if users in a cluster have highly correlated channels, more degrees of freedom will be left for the inter-cluster interference cancellation (as explained in Section IV). In this work, we utilize the Euclidian distance as a similarity metric to measure the correlation between a user's channel vector and the cluster centers. Moreover, to account for variations of channel gains due to fading and other propagation effects, the channel vectors are normalized, averaged over subcarriers, and treated as the data points in the application of the K-means method, i.e.,

$$\mathbf{d}_j = \frac{1}{N}\sum_{n=1}^{N}\frac{\mathbf{h}_j(n)}{\|\mathbf{h}_j(n)\|_2} \tag{16}$$

where $\mathbf{h}_j(n) \in \mathbb{C}^{1\times LM}$ for $j \in \mathcal{J} \triangleq \{1, \ldots, J\}$ are the known network-wide channel vectors of all users prior to clustering.

The K-means method can be presented as an optimization problem for finding the K best centers such that the sum of squared Euclidean (SSE) distance between the data points and their nearest cluster centers is minimized. Specifically, this optimization problem can be expressed as follows,

$$\min_{\mathbf{C}} \sum_{j} \min_{k\in\mathcal{K}} \|\mathbf{d}_j - \mathbf{c}_k\|_2^2 \tag{17}$$

where $\mathbf{C} \triangleq \{\mathbf{c}_k | k \in \mathcal{K}\}$.

Proposition 1: Given $\mathbf{d}_j$ and $\mathbf{c}_k \in \mathbb{C}^{1\times LM}$, we have,

$$\min_{k\in\mathcal{K}} \|\mathbf{d}_j - \mathbf{c}_k\|_2^2 = \min_{\{\iota_{j,k}|k\in\mathcal{K}\}} \sum_{k=1}^{K} \iota_{j,k}\|\mathbf{d}_j - \mathbf{c}_k\|_2^2 \tag{18a}$$

$$\text{s.t.} \sum_{k=1}^{K} \iota_{j,k} = 1 \tag{18b}$$

$$\iota_{j,k} \geq 0, \quad \forall k \in \mathcal{K}. \tag{18c}$$

Proof: The result follows directly from the linear programming duality theory [30]. □

By introducing selection variables $\boldsymbol{\iota} \triangleq \{\iota_{j,k} | j \in \mathcal{J}, k \in \mathcal{K}\}$ and using Proposition 1, we can reformulate problem (15) as the following problem,

$$\min_{\boldsymbol{\iota},\mathbf{C}} \sum_{j}\sum_{k} \iota_{j,k}\|\mathbf{d}_j - \mathbf{c}_k\|_2^2 \tag{19a}$$

$$\text{s.t.} \sum_{k} \iota_{j,k} = 1, \quad \forall j \in \mathcal{J} \tag{19b}$$

$$\iota_{j,k} \geq 0, \quad \forall j \in \mathcal{J}, k \in \mathcal{K} \tag{19c}$$

where $\iota_{j,k} = 1$ if the jth data point is closest to the kth cluster center, i.e., belongs to the kth cluster, and $\iota_{j,k} = 0$ otherwise.

While the K-means method does not involve *a priori* constraint on the number of users in each cluster [31], the SCMA encoder in the current application can support at most q users over N subcarriers. To avoid solutions with more than q data points in a cluster, we propose adding explicit constraints to problem (19) so that each cluster contains at most q data points, i.e.,

$$\min_{\boldsymbol{\iota},\mathbf{C}} \sum_{j}\sum_{k} \iota_{j,k}\|\mathbf{d}_j - \mathbf{c}_k\|_2^2 \tag{20a}$$

$$\text{s.t. (17b), (17c)} \tag{20b}$$

$$\sum_{i} \iota_{j,k} \leq q, \quad \forall k \in \mathcal{K}. \tag{20c}$$

The constrained K-means algorithm solves problem (20) iteratively by uncoupling cluster center and selection variables. Specifically, in each iteration, this algorithm alternates between solving a linear program for variable $\boldsymbol{\iota}$ with fixed $\mathbf{c}$ and solving a problem for $\mathbf{c}$ with fixed $\boldsymbol{\iota}$. The overall constrained K-means algorithm for solving problem (20) is summarized in Algorithm 1, where the superscript t denotes the iteration index.

Proposition 2: There exists an optimal solution for the cluster assignment subproblem in Algorithm 1 such that $\iota_{j,k} \in \{0, 1\}$.

Proof: See Appendix A. □

According to Proposition 2 and Appendix A, we can use the network simplex algorithm which is faster than mixed integer solvers for tackling the cluster assignment subproblem.

Proposition 3: The constrained K-means algorithm terminates in a finite number of iterations at a cluster assignment that is locally optimal. That is, the limit point of the iterates

Algorithm 1 The Proposed Constrained K-Means Algorithm for User Clustering

Initialization: Initialize cluster centers $\mathbf{c}^{(0)} = \{\mathbf{c}_1^{(0)}, \mathbf{c}_2^{(0)}, \ldots, \mathbf{c}_K^{(0)}\}$ by selecting K data points from the dataset randomly. Set $t = 0$.

Repeat:

1) **Cluster assignment:** Solve the following linear program with fixed $\mathbf{c}^{(t)}$.

$$\boldsymbol{\iota}^{(t)} = \arg\min_{\boldsymbol{\iota}} \sum_j \sum_k \iota_{j,k} \|\mathbf{d}_j - \mathbf{c}_k^{(t)}\|_2^2$$
$$\text{s.t. (19b), (19c), (20c).}$$

2) **Cluster update:** Update the cluster centers as,

$$\mathbf{c}_k^{(t+1)} = \frac{\sum_j \iota_{j,k}^{(t)} \mathbf{d}_j}{\sum_j \iota_{j,k}^{(t)}}, \quad \forall k \in \mathcal{K}.$$

3) Set $t \leftarrow t + 1$.

Until: $\mathbf{c}_k^{(t)} = \mathbf{c}_k^{(t-1)}, \forall k \in \mathcal{K}$.

generated by the constrained K-means algorithm is a stationary point that satisfies the Karush-Kuhn-Tucker (KKT) conditions for problem (20).

Proof: At each iteration, the cluster assignment step cannot increase the objective function of (20). The cluster update step will either strictly decrease the value of the objective function of (20) or the algorithm will terminate since,

$$\mathbf{c}^{(t+1)} = \arg\min_{\mathbf{c}} \sum_j \sum_k \iota_{j,k}^{(t)} \|\mathbf{d}_j - \mathbf{c}_k\|_2^2 \tag{21}$$

is a strictly convex optimization problem with a unique global solution (as shown in the cluster update step in Algorithm 1). Thus, the objective of (20) is strictly non-increasing and bounded below by zero. Moreover, there are a finite number of ways to assign J points to K clusters such that each cluster has at most q points and Algorithm 1 does not permit repeated assignments. Consequently, the algorithm must terminate at some cluster assignment that is locally optimal. □

B. NUMBER OF CLUSTERS

The choice of the number of clusters K plays a key role in the performance of K-means clustering [32]. An appropriate number of clusters can accurately reflect specific distribution characteristics of users in the network. While the number of clusters cannot exceed the number of users, it should also satisfy the constraint on the maximum number of users in each cluster. However, finding the optimal K is a major challenge in clustering analysis, and there is no definitive solution. To address this problem, a number of approaches have been proposed such as the elbow [33], silhouette [34], and gap statistic [35] methods. Among these, the elbow method is possibly the most well-known and utilized as it entails the lowest computational complexity while providing very good performance.

Herein, we employ the elbow method to determine the number of clusters. The elbow method is a heuristic method which involves running the clustering algorithm on the dataset and evaluating a clustering criterion for different values of K. The plot of the clustering criterion versus the number of clusters resembles an arm in which the elbow point (the point of discontinuity in the slope of the curve) determines the appropriate number of clusters for the dataset. The sum of the normalized within-cluster SSE distance is a common clustering criterion for applying the elbow method along with K-means.

In a given cluster $\mathcal{C}_k$, the within-cluster SSE distance between the data points is given by,

$$D_k = \frac{1}{2} \sum_{\mathbf{d}_i \in \mathcal{C}_k} \sum_{\mathbf{d}_{i'} \in \mathcal{C}_k} \|\mathbf{d}_i - \mathbf{d}_{i'}\|_2^2. \tag{22}$$

Hence, the sum of the normalized within-cluster SSE distances can be expressed as,

$$W_K = \sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} D_k \tag{23}$$

where $|\mathcal{C}_k|$ shows the cardinality of the cluster $\mathcal{C}_k$. It should be noted that although the sum of the normalized within-cluster SSE distance can give a proper measure of the compactness of the clustering, we may encounter cases with more than one elbow point or no elbow point. In such cases, other reliable methods mentioned before can be used to find the best K.

C. COMPLEXITY ANALYSIS

In this subsection, we analyze the computational complexity of the proposed constrained K-means algorithm by considering the number of required operations (e.g. complex addition and multiplication) in each step and in each iteration of the algorithm. Specifically, we divide the operations for each iteration into three steps:

- Calculation of Euclidean distances: The complexity of calculating the Euclidean distance between the data points and the cluster centers is $O(JKLM)$.
- Cluster assignment: The complexity of solving cluster assignment subproblem via network simplex algorithm is $O(J^3K^2(\log(J))^2)$ (See Appendix A).
- Cluster update: The complexity of updating the cluster centers is $O(JKLM)$.

Assuming that the algorithm converges after T_K iterations. The overall complexity of Algorithm 1 can be expressed as,

$$C_C \triangleq O(T_K J^3 K^2 (\log(J))^2 + T_K JKLM). \tag{24}$$

IV. DOWNLINK BEAMFORMING

In this section, we first formulate the beamforming design as a non-convex mixed-integer nonlinear programing (MINLP) optimization problem, aiming to minimize the total transmit

power while satisfying the QoS and fronthaul capacity constraints. We then propose transformations and convex approximation techniques to derive two iterative algorithms for solving the problem. In the first algorithm, we approximate the continuous non-convex constraints by convex ones using first-order Taylor expansion. Hence, we are able to arrive at a sequence of mixed-integer second-order cone programming (MI-SOCP), for which dedicated solvers are available. Although the MI-SOCP algorithm entails high computational complexity, it is shown that it can achieve high quality solution [36]. Hence, in this paper, we use MI-SOCP algorithm as a benchmark. A simplified suboptimal approach is also proposed which designs the beamformers in two stages to achieve lower complexity. In the first stage, the cluster beamformers are determined by taking advantage of BD to remove intercluster interference. In the second stage, we obtain the user-specific beamformers with the aid of CCCP method to minimize the total transmit power. Finally, the convergence and the computational complexity of the proposed algorithms are discussed.

A. BEAMFORMING PROBLEM

Our objective is to optimize the total transmit power through joint design of the dynamic RRH selection scheme and the beamforming vectors subject to the QoS and fronthaul capacity constraints on each individual RRH. Let the binary variable $s_{j,k}^{(l)}(n) = 1$ indicate that the lth RRH participates in transmission for the jth user in the kth cluster over the nth subcarrier and $s_{j,k}^{(l)}(n) = 0$ otherwise. Hence, our optimization problem can be mathematically formulated as,

$$\min_{\mathbf{w}_{jk}(n), s_{j,k}^{(l)}(n)} \sum_n \sum_k \sum_j \|\mathbf{w}_{jk}(n)\|^2 \tag{25a}$$

$$\text{s.t. } \mathbf{h}_{j'k'}(n)\mathbf{w}_{jk}(n) = \mathbf{0}, \quad \forall\, k' \neq k \tag{25b}$$

$$\gamma_{jk}(n) \geq \gamma_{\min}, \quad \forall n \in \mathcal{N}_{jk} \tag{25c}$$

$$\sum_k \sum_j s_{j,k}^{(l)}(n) R_{jk}(n) \leq C_{\max}, \quad \forall l, n \tag{25d}$$

$$\|\mathbf{w}_{jk}^{l}(n)\|_2 \leq s_{j,k}^{(l)}(n) P_{\max}, \quad \forall l, n \tag{25e}$$

$$\sum_l s_{j,k}^{(l)}(n) \geq 1, \quad \forall n \in \mathcal{N}_{jk} \tag{25f}$$

$$s_{j,k}^{(l)}(n) \in \{0, 1\} \tag{25g}$$

where $R_{jk}(n) = \log_2(1 + \gamma_{jk}(n))$ denotes the transmission rate, $\mathcal{N}_{jk}$ shows the set of subcarriers occupied by the jth user in the kth cluster, $\gamma_{\min}$, $C_{\max}$, and $P_{\max}$ are the minimum required SINR for the user over the subcarrier, the maximum capacity constraint for each RRH over the subcarrier, the maximum available total transmit power, respectively. Constraints (25b) and (25c) guarantee QoS by removing the inter-cluster interference and satisfying SINR requirements, respectively. The constraint (25d) shows that the sum-rate of the users served by the lth RRH over the nth subcarrier should be smaller than the maximum fronthaul capacity $C_{\max}$. Constraint (25e) utilizes the so-called Big M method which indicates that the beamformer $\|\mathbf{w}_{jk}^{l}(n)\|_2 = 0$ if the lth RRH does not participate in transmission for the jth user in the kth cluster over the nth subcarrier, i.e., $s_{j,k}^{(l)}(n) = 0$, but leaves the beamformer "open" otherwise. Therefore, $P_{\max}$ can be any large number. Constraint (25f) guarantees that each user is served by at least one RRH. Although constraint (25f) appears to be redundant, it is added to reduce the size of the feasible set of the associated problem which in turn improves the convergence time of the solver. We refer the interested reader to [37] for additional details.

Problem (25) is a non-convex MINLP problem, which can be considered as an NP-hard problem in general and is one of the most challenging class of mathematical optimization problems [38]. Obtaining its optimal solution is challenging due to the non-convexity of the SINR constraints, the combinatorial nature of the RRH selection variable $s_{j,k}^{(l)}(n)$, and the coupling between the variables $s_{j,k}^{(l)}(n)$ and $R_{jk}(n)$ in the fronthaul constraint. Even when the RRH selection scheme $s_{j,k}^{(l)}(n)$ is given, problem (23) is still non-convex and computationally difficult. In the following subsections, we develop two beamforming approaches to find a suboptimal solution.

B. MI-SOCP BEAMFORMING APPROACH

In this section, we first reformulate the problem (25) into a more tractable form. We then solve the resulting optimization problem via a CCCP-based algorithm with guaranteed convergence to a local stationary solution of the transformed problem.

Without loss of optimality, SINR constraint (25c) can be rewritten as the following second-order cone (SOC) constraint,

$$\sqrt{I_{jk}^{(1)}(n) + I_{jk}^{(2)}(n) + \sigma_{jk}^2} \leq \frac{\mathbf{h}_{jk}(n)\mathbf{w}_{jk}(n)}{\sqrt{\gamma_{\min}}} \tag{26}$$

where $I_{jk}^{(1)}(n)$ and $I_{jk}^{(2)}(n)$ are the intra- and inter-cluster interference as expressed in (12) and (13) respectively. We have restricted $\mathbf{h}_{jk}(n)\mathbf{w}_{jk}(n)$ to be positive real, which incurs no loss of optimality since we can always phase-rotate the vector $\mathbf{w}_{jk}(n)$ such that $\mathbf{h}_{jk}(n)\mathbf{w}_{jk}(n)$ is positive real without affecting the cost function or the constraints.

Let us introduce the auxiliary variables $u_{j,k}(n)$ and $v_{j,k}(n)$ as the upper bounds on the SINR and transmission rate for the jth user in the kth cluster over the nth subcarrier. Hence, constraint (25d) can be rewritten as follows,

$$\gamma_{jk}(n) \leq u_{jk}(n) \tag{27}$$

$$\log_2(1 + u_{jk}(n)) \leq v_{jk}(n) \tag{28}$$

$$\sum_k \sum_j s_{j,k}^{(l)}(n) v_{jk}(n) \leq C_{\max}. \tag{29}$$

Since the expression of $\gamma_{jk}(n)$ is in fractional form, the constraint in (27) is difficult to handle. Therefore, we introduce the auxiliary variables $l_{j,k}(n)$ as the lower bound of the denominator, and then equivalently transform (27) as the following two constraints,

$$|\mathbf{h}_{jk}(n)\mathbf{w}_{jk}(n)|^2 \leq l_{jk}(n) u_{jk}(n), \tag{30}$$

$$l_{jk}(n) \le I_{jk}^{(1)}(n) + I_{jk}^{(2)}(n) + \sigma_{jk}^2. \tag{31}$$

From the above discussion, we can finally reformulate the problem (25) into an equivalent problem as given below,

$$\min \sum_n \sum_k \sum_j \|\mathbf{w}_{jk}(n)\|^2 \tag{32a}$$

$$\text{s.t. (25b),(25e)-(25g),(26)} \tag{32b}$$

$$\sqrt{4|\mathbf{h}_{jk}(n)\mathbf{w}_{jk}(n)|^2 + (l_{jk}(n) - u_{jk}(n))^2} \le l_{jk}(n) + u_{jk}(n) \tag{32c}$$

$$l_{jk}(n) \le I_{jk}^{(1)}(n) + I_{jk}^{(2)}(n) + \sigma_{jk}^2 \tag{32d}$$

$$1 + u_{jk}(n) \le 2^{v_{jk}(n)} \tag{32e}$$

$$\sum_k \sum_j (s_{j,k}^{(l)}(n) + v_{jk}(n))^2 - 4C_{\max} \le \sum_k \sum_j (s_{j,k}^{(l)}(n) - v_{jk}(n))^2. \tag{32f}$$

where the identity $4xy = (x+y)^2 - (x-y)^2$ is used to obtain (32f). We note that even by continuous relaxation of binary variables $s_{j,k}^{(l)}(n)$, optimization problem (32) is still non-convex due to constraints (32d)-(32f). However, the latter can be expressed as differences of two convex functions. Thus, the obtained optimization problem can be efficiently solved using the iterative CCCP.

Basically, CCCP iteratively solves a sequence of convex subproblems, each of which is constructed by linearizing the concave part of the DC constraints using their first-order Taylor expansions [21]. Specifically, the first-order Taylor expansion of the right side of constraint (32d) around the current point $\hat{\mathbf{w}}_{jk}(n)$ is expressed as,

$$\begin{aligned} F(\mathbf{w}_{j'k}(n); \hat{\mathbf{w}}_{j'k}(n)) = & \sum_{j' \ne j, j' \in \mathcal{U}_{n,k}} [-|\mathbf{h}_{jk}(n)\hat{\mathbf{w}}_{j'k}(n)|^2 \\ & + 2\Re\{\hat{\mathbf{w}}_{j'k}^H(n)\mathbf{h}_{jk}^H(n)\mathbf{h}_{jk}(n)\mathbf{w}_{j'k}(n)\}] \\ & + \sum_{k' \ne k} \sum_{j' \in \mathcal{U}_{n,k'}} [-|\mathbf{h}_{jk}(n)\hat{\mathbf{w}}_{j'k'}(n)|^2 \\ & + 2\Re\{\hat{\mathbf{w}}_{j'k'}^H(n)\mathbf{h}_{jk}^H(n)\mathbf{h}_{jk}(n)\mathbf{w}_{j'k'}(n)\}] \end{aligned} \tag{33}$$

where $\Re\{.\}$ denotes the real part of its argument. In the same way, we convexify the right side of constraints (32d) and (32e) by using the first-order Taylor expansions around the current points $\hat{v}_{jk}(n)$, $\hat{s}_{j,k}^{(l)}(n)$, and $\hat{v}_{jk}(n)$ as,

$$\Gamma(v_{jk}(n); \hat{v}_{jk}(n)) = 2^{\hat{v}_{jk}(n)} + (\ln 2) 2^{\hat{v}_{jk}(n)} (v_{jk}(n) - \hat{v}_{jk}(n)), \tag{34}$$

$$\begin{aligned} &\Omega(s_{j,k}^{(l)}(n), v_{jk}(n); \hat{s}_{j,k}^{(l)}(n), \hat{v}_{jk}(n)) \\ &\quad = -(\hat{s}_{j,k}^{(l)}(n) - \hat{v}_{jk}(n))^2 + 2(\hat{s}_{j,k}^{(l)}(n) - \hat{v}_{jk}(n)) \\ &\qquad \times (s_{j,k}^{(l)}(n) - v_{jk}(n)). \end{aligned} \tag{35}$$

By applying the above approximations to the non-convex constraints (32d)-(32f), we can formulate the convex approximation of problem (32) as shown below,

$$\min \sum_n \sum_k \sum_j \|\mathbf{w}_{jk}(n)\|^2 \tag{36a}$$

$$\text{s.t. (23b),(23e)-(23g),(24),(30c)} \tag{36b}$$

$$l_{jk}(n) \le F(\mathbf{w}_{j'k}(n); \hat{\mathbf{w}}_{j'k}(n)) + \sigma_{jk}^2 \tag{36c}$$

$$1 + u_{jk}(n) \le \Gamma(v_{jk}(n); \hat{v}_{jk}(n)) \tag{36d}$$

$$\sum_k \sum_j (s_{j,k}^{(l)}(n) + v_{jk}(n))^2 - 4C_{\max} \le \Omega(s_{j,k}^{(l)}(n), v_{jk}(n); \hat{s}_{j,k}^{(l)}(n), \hat{v}_{jk}(n)). \tag{36e}$$

Hence, based on CCCP, we solve subproblem (36) at each iteration. Problem (36) is a MI-SOCP which can be solved via modern solvers such as MOSEK [39] or GUROBI [40]. The proposed iterative algorithm is summarized in Algorithm 2. The algorithm terminates if the variation of the total transit power, i.e., ΔP_T, is less than a preset threshold ϵ.

Algorithm 2 MI-SOCP Beamforming Algorithm

Initialize the algorithm with feasible points $\hat{\mathbf{w}}_{jk}(n)$, $\hat{s}_{j,k}^{(l)}(n)$, and $\hat{v}_{jk}(n)$. Set iteration index $t = 0$ and termination threshold $\epsilon > 0$.

Repeat

1) Update $\hat{\mathbf{w}}_{jk}(n)$, $\hat{s}_{j,k}^{(l)}(n)$, and $\hat{v}_{jk}(n)$ by solving problem (36).

2) Set $t = t + 1$.

Until: Termination criterion is met: $\Delta P_T < \epsilon$.

Initialization: Choosing a feasible point for initialization of Algorithm 2 is essential. For this purpose, we simply set $\hat{v}_{jk}(n) = \log_2(1 + \gamma_{\min})$ and then solve the following feasibility problem $P = \text{find}\{s_{j,k}^{(l)}(n) | (25\text{f}),(25\text{g}), \sum_k \sum_j s_{j,k}^{(l)}(n)\hat{v}_{jk}(n) \le C_{\max}\}$ which is a mixed-integer linear program which can be solved optimally by off-the-shelf solvers such as MOSEK or GUROBI. Subsequently, we solve the following quadratic program with fixed $\hat{s}_{j,k}^{(l)}(n)$ via any general-purpose solver using interior-point method,

$$\begin{aligned} \hat{\mathbf{w}}_{jk}(n) = \arg \min_{\mathbf{w}_{jk}(n)} & \sum_n \sum_k \sum_j \|\mathbf{w}_{jk}(n)\|^2 \\ \text{s.t. } & (25\text{b}), (26), \\ & \|\mathbf{w}_{jk}^l(n)\|_2 \le \hat{s}_{j,k}^{(l)}(n) P_{\max}. \end{aligned} \tag{37}$$

C. TWO-STAGE BEAMFORMING APPROACH

In order to reduce the computational complexity, we propose a two-stage energy-efficient beamforming approach such that,

$$\mathbf{w}_{jk}(n) = \mathbf{B}_k(n)\mathbf{v}_{jk}(n) \tag{38}$$

where $\mathbf{B}_k(n) \in \mathbb{C}^{LM \times a}$ is the kth cluster beamformer obtained in the first stage which should eliminate the inter-cluster interference and $\mathbf{v}_{jk}(n) \in \mathbb{C}^{a \times 1}$ is the user-specific beamformer for the jth user in the kth cluster optimized in the second stage.

Using channel state information (CSI) available at the central processor, BD beamforming can be adopted in a MIMO-SCMA system to remove the inter-cluster interference and enhance the QoS for intra-cluster users [41]. Hence, the users in different clusters can share codebooks. Although BD algorithm does not work well in the presence of imperfect CSI, we considered a second stage for beamforming in which the QoS can be guaranteed. Specifically, the BD beamforming projects the transmitted signal onto the null-space of the interfering channels and hence eliminates the inter-cluster interference.

To find the corresponding null-space, let us define,

$$\mathbf{H}_k(n) = [\mathbf{h}_{1k}(n)^T \dots \mathbf{h}_{Jk}(n)^T] \in \mathbb{C}^{LM\times J_k} \tag{39}$$

$$\mathbf{H}_{-k}(n) = [\mathbf{H}_1(n) \dots \mathbf{H}_{k-1}(n)\, \mathbf{H}_{k+1}(n) \dots \mathbf{H}_K(n)] \tag{40}$$

where $k \in \mathcal{K}$ and $\mathbf{H}_{-k}(n) \in \mathbb{C}^{LM\times(J-J_k)}$ is the matrix containing all interfering channels for the kth cluster. We seek $\mathbf{B}_k(n)$ orthogonal to the column span of $\mathbf{H}_{-k}(n)$, i.e., $\mathbf{H}_{-k}(n)^T\mathbf{B}_k(n) = \mathbf{0}$. Here, it is assumed that the total number of antennas LM is larger than the total number of users J.

The singular value decomposition (SVD) can be employed to calculate the cluster beamformers. Applying the SVD to $\mathbf{H}_{-k}(n)$ yields,

$$\mathbf{H}_{-k}(n) = \mathbf{U}_k(n)\Sigma_k(n)\mathbf{V}_k(n)^H \tag{41}$$

where $\mathbf{U}_k(n) \in \mathbb{C}^{LM\times LM}$ and $\mathbf{V}_k(n) \in \mathbb{C}^{(J-J_k)\times(J-J_k)}$ are unitary matrices and $\Sigma_k(n) \in \mathbb{R}^{LM\times(J-J_k)}$ is the rectangular diagonal matrix of singular values. Let r denote the rank of matrix $\mathbf{H}_{-k}(n)$, which corresponds to the number of non-zero diagonal entries in $\Sigma_k(n)$. The null-space of the interfering channel matrix $\mathbf{H}_{-k}(n)$ is spanned by the left singular vectors (i.e. columns of matrix $\mathbf{U}_k(n)$) associated to the zero singular values of $\mathbf{H}_{-k}(n)$. We can express the kth cluster beamformer as,

$$\mathbf{B}_k(n) = [\mathbf{u}_{r+1,k}(n)\, \mathbf{u}_{r+2,k}(n)\, \dots \mathbf{u}_{LM,k}(n)] \tag{42}$$

where $\mathbf{u}_{i,k}(n)$ denotes the ith column of $\mathbf{U}_k(n)$.

As mentioned before, constraint (25e) implies that $\|\mathbf{w}_{jk}^{(l)}(n)\|_2 = 0$ if $s_{j,k}^{(l)}(n) = 0$. Without loss of optimality, the binary RRH selection variable $s_{j,k}^{(l)}(n)$ can be replaced by $\| \|\mathbf{w}_{jk}^{(l)}(n)\|_2^2 \|_0$, as in [42], [43]. Therefore, upon substitution of (38) and ℓ_0-norm, problem (25) can be rewritten as,

$$\min_{\mathbf{v}_{j,k}(n)} \sum_n \sum_k \sum_j \|\mathbf{w}_{jk}(n)\|^2 \tag{43a}$$

$$\text{s.t. } \mathbf{w}_{jk}(n) = \mathbf{B}_k(n)\mathbf{v}_{jk}(n), \quad \forall j,k,n \tag{43b}$$

$$\gamma_{jk}(n) \geq \gamma_{\min}, \quad \forall n \in \mathcal{N}_{jk} \tag{43c}$$

$$\sum_k \sum_j \| \|\mathbf{w}_{jk}^{(l)}(n)\|_2^2 \|_0 R_{jk}(n) \leq C_{\max}, \quad \forall l,n \tag{43d}$$

It should be noted that the fronthaul capacity constraint (43d) which is expressed in the form of ℓ_0-norm, indicates the inherently dynamic RRH selection. That is, owing to this fronthaul constraint, the network-wide beamforming vectors $\mathbf{w}_{jk}(n)$ may have a sparse structure. Although the number of constraints is reduced and the binary RRH selection variable is removed, problem (43) is still non-convex due to constraints (43c) and (43d).

As mentioned before, using cluster beamformer $\mathbf{B}_k(n)$ obtained from BD can remove inter-cluster interference. Hence, the SINR of the jth user in the kth cluster over the nth subcarrier can be expressed as,

$$\gamma_{jk}(n) = \frac{|\mathbf{h}_{jk}(n)\mathbf{w}_{jk}(n)|^2}{\sum_{j'\neq j}^{J_k} |\mathbf{h}_{jk}(n)\mathbf{w}_{j'k}(n)|^2 + \sigma_{jk}^2} \tag{44}$$

where the inter-cluster interference term in the denominator is removed. Consequently, SINR constraint (43c) can be rewritten as follows,

$$\sqrt{\sum_{j'\neq j}^{J_k} |\mathbf{h}_{jk}(n)\mathbf{w}_{j'k}(n)|^2 + \sigma_{jk}^2} \leq \frac{\mathbf{h}_{jk}(n)\mathbf{w}_{jk}(n)}{\sqrt{\gamma_{\min}}} \tag{45}$$

which is a SOC constraint.

To address the non-convexity of constraint (43d), we first introduce the auxiliary variables $u_{j,k}(n)$, $v_{j,k}(n)$, and $t_{j,k}^{(l)}(n)$ as the upper bounds of the SINR, transmission rate, and ℓ_0-norm for the jth user in the kth cluster over the nth subcarrier. Hence, constraint (43d) can be rewritten as follows,

$$\gamma_{jk}(n) \leq u_{jk}(n), \tag{46}$$

$$\log_2(1+u_{jk}(n)) \leq v_{jk}(n), \tag{47}$$

$$\| \|\mathbf{w}_{jk}^{(l)}(n)\|_2^2 \|_0 \leq t_{j,k}^{(l)}(n), \tag{48}$$

$$\sum_k \sum_j t_{j,k}^{(l)}(n) v_{jk}(n) \leq C_{\max}. \tag{49}$$

We then propose to approximate the non-convex ℓ_0-norm by a reweighted ℓ_1-norm as follows [44],

$$\| \|\mathbf{w}_{jk}^{(l)}(n)\|_2^2 \|_0 \approx \beta_{jk}^{(l)}(n) \|\mathbf{w}_{jk}^{(l)}(n)\|_2^2. \tag{50}$$

$\beta_{jk}^{(l)}(n)$ is a constant weight which is updated in each iteration according to,

$$\beta_{jk}^{(l)}(n) = \frac{1}{\|\hat{\mathbf{w}}_{jk}^{(l)}(n)\|_2^2 + \tau} \tag{51}$$

where $\hat{\mathbf{w}}_{jk}^{(l)}(n)$ is obtained from previous iteration and τ is a small constant regularization factor controlling the smoothness of the approximation. Based on the updating rule (51), $\beta_{jk}^{(l)}(n)$ is inversely proportional to the transmit power level $\|\hat{\mathbf{w}}_{jk}^{(l)}(n)\|_2^2$. Hence, the RRHs with lower transmit power for the jth user in the kth cluster would have higher weights and hence would be forced to further reduce its transmit power and eventually be dropped out of the group of participating RRHs for that user.

We can employ the approach mentioned in IV.B to deal with the non-convexity of the constraints and use CCCP

Algorithm 3 Proposed CCCP-Based Iterative Algorithm for Beamforming

Initialization: Initialize $\hat{\mathbf{v}}_{jk}(n)$ randomly. Calculate $\hat{\mathbf{w}}_{jk}(n)$, $\hat{t}_{j,k}^{(l)}(n)$ and $\beta_{jk}^{(l)}(n)$. Set iteration index $t = 0$ and termination threshold $\epsilon > 0$.

Repeat

1) Update $\hat{\mathbf{w}}_{jk}(n)$, $\hat{t}_{j,k}^{(l)}(n)$, and $\hat{v}_{jk}(n)$ via solving problem (52).

2) Calculate $\beta_{jk}^{(l)}(n)$.

3) Set $t = t + 1$.

Until Termination criterion is met: $\Delta P_T < \epsilon$.

to solve the optimization problem. Hence, based on CCCP, we solve the following subproblem at each iteration,

$$\min_{\mathbf{v}_{jk}(n)} \sum_n \sum_k \sum_j \|\mathbf{w}_{jk}(n)\|^2 \tag{52a}$$

$$\text{s.t. (30c),(34c),(34d),(36),(41b),(43)} \tag{52b}$$

$$\beta_{jk}^{(l)}(n)\|\mathbf{w}_{jk}^{(l)}(n)\|_2^2 \le t_{j,k}^{(l)}(n) \tag{52c}$$

$$\sum_k \sum_j (t_{j,k}^{(l)}(n) + v_{jk}(n))^2 - 4C_{\max} \le \Omega(t_{j,k}^{(l)}(n), v_{jk}(n); \hat{t}_{j,k}^{(l)}(n), \hat{v}_{jk}(n)). \tag{52d}$$

Problem (52) is convex and can be solved via any general-purpose solver using interior-point methods [45]. The proposed CCCP-based iterative algorithm is summarized in Algorithm 3.

Initialization: In this case, an initial point for Algorithm 3 is obtained by generating $\hat{\mathbf{v}}_{jk}(n)$ randomly. Then, $\hat{\mathbf{w}}_{jk}(n)$ and $\beta_{jk}^{(l)}(n)$ are calculated as in (38) and (51) respectively. $\hat{t}_{j,k}^{(l)}(n)$ is set to $\| \|\hat{\mathbf{w}}_{jk}^{(l)}(n)\|_2^2 \|_0$, and $\hat{v}_{jk}(n)$ is set to the transmission rate calculated using $\hat{\mathbf{w}}_{jk}(n)$.

D. CONVERGENCE AND COMPLEXITY ANALYSIS

With a feasible initial point, repeated application of the CCCP iteration is guaranteed to converge to a stationary solution of the problem with DC constraints. It can be seen that the optimal solution obtained from the previous iteration, i.e., $\hat{\mathbf{w}}_{jk}(n)$, is feasible for the convex subproblem at the next iteration for both algorithms. The achieved objective at the current iteration cannot be greater than the one at the previous iteration. Since, the objective function is non-increasing and bounded below by zero, it follows that both algorithms will converge to a point that according to [46] is locally optimal. We refer the interested reader to [46] for a rigorous proof of the convergence.

For Algorithm 2, the overall complexity is dominated by solving the MI-SOCP problem in (36). In particular, there are JLN binary variables $s_{j,k}^{(l)}(n)$, resulting in 2^{JLN} combinations for all the binary variables. Thus, assuming that MI-SOCP algorithm terminates after T_M iterations, the worst-case complexity can be written as

$$C_M \triangleq \mathcal{O}(T_M\, 2^{JLN}(JCLM)^3). \tag{53}$$

At each iteration, the CCCP-based algorithm solves the convex subproblem (52) which can be approximated by a sequence of SOCPs via the successive approximation method. Each SOCP can then be solved via a general-purpose solver, e.g., SDPT3 in CVX [47] with a complexity of $\mathcal{O}((JCLM)^3)$. Assuming that the CCCP terminates after T_C iterations, the worst-case computational complexity is therefore given by,

$$C_B \triangleq \mathcal{O}(T_C(JCLM)^3). \tag{54}$$

V. SIMULATION RESULTS

In this section, numerical experiments are carried out to illustrate the performance of the proposed energy-efficient user clustering and downlink beamforming for MIMO-SCMA in C-RAN.

A. METHODOLOGY

In our simulations, unless otherwise specified, we consider a network with $L = 3$ RRHs, each equipped with $M = 5$ antennas and serving $J = 12$ single-antenna users. The RRHs and the users are independently distributed in a square area $[-50, 50] \times [-50, 50]$ meters. The RRHs are connected to the central processor via a limited-capacity fronthaul link with maximum capacity $C_{\max} = 50$ bps/Hz. The maximum available total transmit power is $P_{\max} = 50$ dBm.

We consider the channel model as described in Section II.B with bandwidth of $W = 2$ GHz and carrier frequency of 28 GHz. The AoDs are assumed to follow a uniform distribution in $[0, 2\pi]$. The inter-antenna spacing is $d = \lambda/2$ to reduce the effect of mutual coupling and correlation among neighbouring antenna elements. The noise figure is $N_f = 40$ dBm, hence, the noise power is $\sigma_{jk}^2 = -174 + 10\log_{10}(W) + N_f$ dBm [23]. The pathloss exponent of the LOS and NLOS paths in (4) are $\alpha^{(0)} = 2$ and $\alpha^{(p)} = 3$, respectively. For SCMA encoder, the number of subcarriers is $N = 4$, and the number of non-zero elements for each codeword is $C = 2$. The corresponding factor graph matrix is,

$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}. \tag{55}$$

It should be noted that the structure of the factor graph matrix with fixed N and C does not affect system performance significantly. Table 2 summarizes the simulation setting parameters.

We use Monte Carlo experiments to evaluate the performance of the proposed algorithms for user clustering and downlink beamforming. The total transmit power and sum rate are measured for different parameter configurations and the results are compared with benchmark approaches in the literature.

TABLE 2. Simulation setting parameters.

Notation	Description	Value
L	Number of RRHs	3
M	Number of antennas per RRH	5
J	Total number of users	12
N	Number of SCs	5
C	Number of non-zero elements for a codeword	2
$\alpha^{(0)}$	path loss exponent of the LOS path	2
$\alpha^{(p)}$	path loss exponent of the NLOS path	3
$C_{\max}$	Maximum capacity constraint for each RRH	50 bps/Hz
$P_{\max}$	Maximum available total transmit power	50 dBm
ϵ	Termination threshold	10^{-6}

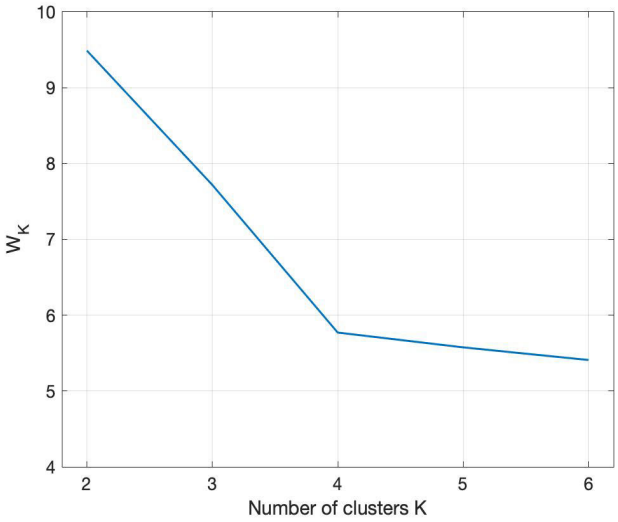
B. RESULTS AND DISCUSSION

Figures 2a and 2b find the optimal number of clusters K and evaluate its impact on the performance of the proposed scheme. In Figure 2a, we plot the sum of the normalized within-cluster SSE distance which serves as clustering criterion in the elbow method described in Section III.B. It can be seen W_K decreases when K increases and the elbow point can be found at $K = 4$.[1] To gain further insight into the impact of the number of clusters, we investigate the transmit power performance versus target SINR, γ_{min} in Figure 2b, where the number of clusters increases from 2 to 6. It is observed that the total transmit power increases monotonically as γ_{min} increases. Moreover, the best performance is achieved when the number of clusters is $K = 4$. On one hand, for $K < 4$, an increase in the number of users in a cluster results in larger intra-cluster interference which results in higher transmit power. On the other hand, increasing the number of clusters intensifies inter-cluster interference which increases power consumption in the first stage beamforming for interference cancellation. We thereby observe that better user clustering with lower total transmit power can be found at $K = 4$ and the elbow method can efficiently find the optimal number of clusters in this case.

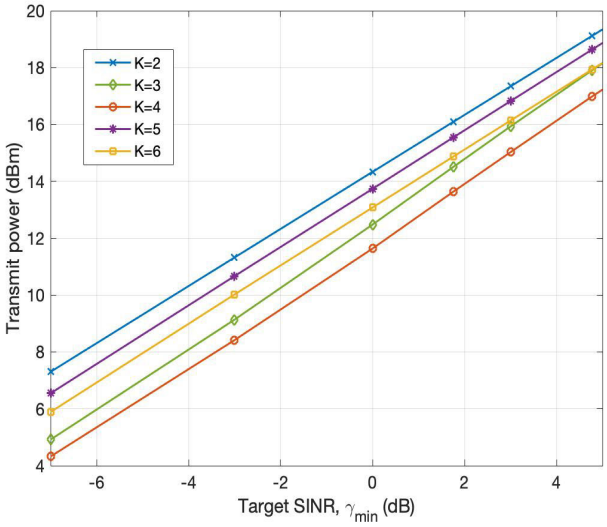
Figures 3a and 3b present the convergence behaviour of the proposed constrained K-means and CCCP-based algorithms. Figure 3a shows the objective value achieved by the constrained K-means algorithm with three different initial points. In this regards, $J = 12$ users are grouped into $K = 4$ non-overlapping clusters of size less than 6, i.e., $q = 6$. We observe that the algorithm converges rapidly in a few steps and the gap between final results of different initial points is small. In Figure 3b, the convergence performance of the CCCP-based algorithms is investigated for the case of $\gamma_{\min} = 3$ dB. It can be seen that the algorithm converges in less than 15 iterations monotonically to a same value for different initial points.

In Table 3, we present the run-time comparison between the proposed two-stage approach and the MI-SOCP

[1]Due to space limitation, we omit the results on the Silhouette and gap statistic methods here for brevity However, it should be noted that using each of these methods for determining the optimal number clusters gives the same result while elbow method entails lower computational complexity.



(a) Elbow method.



(b) Transmit power versus target SINR $\gamma_{\min}$.

FIGURE 2. The impact of the number of clusters.

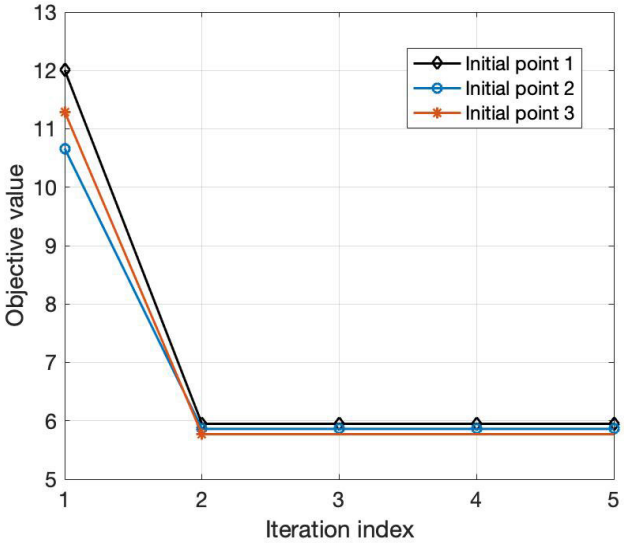
TABLE 3. Run-time of different beamforming approaches.

	γ_{min}			
	0	2	4	6
Two-Stage approach	3.6108	4.3852	5.7032	7.0414
MI-SOCP approach	182.3750	238.3348	323.9511	397.0151

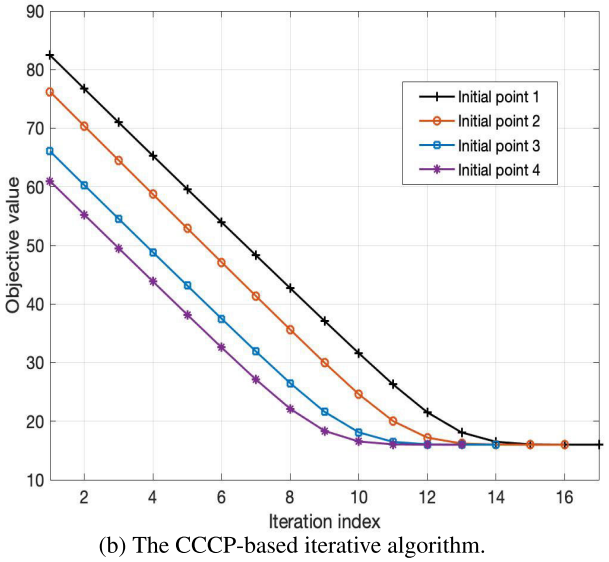
beamforming for different values of $\gamma_{\min}$. The results,[2] show that the complexity of the proposed two-stage beamforming algorithm is much less than that of the MI-SOCP approaches, owing to the use of the smoothed ℓ_0-norm approximation.

In Figure 4, we compare the transmit power versus sum rate among different clustering and beamforming algorithms. The cluster-head approach proposed in [48] is used as a benchmark for user clustering which selects the K users

[2]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.



(a) The constrained K-means algorithm.



(b) The CCCP-based iterative algorithm.

FIGURE 3. **The convergence of the proposed algorithms.**

with the highest channel gains as the cluster centers. The cluster assignment is then used to group users into clusters. We also consider the performance of the MIMO-SCMA system with exhaustive and random clustering. The combination of exhaustive search for user clustering and MI-SOCP for beamforming is shown to attain the best performance among all the algorithms. However, this comes at the cost of high computational complexity. As it can be seen from Figure 4, the proposed constrained K-means clustering algorithm exhibits better performance compared to random search and cluster-head approach and can partition users more efficiently. Regarding the beamforming algorithms, it is shown in Figure 4 that the suboptimal solution achieved by the proposed two-stage beamforming is very close to the high-quality solution obtained by MI-SOCP.

To better appreciate the benefits of the proposed MIMO-SCMA scheme in terms of spectral efficiency,

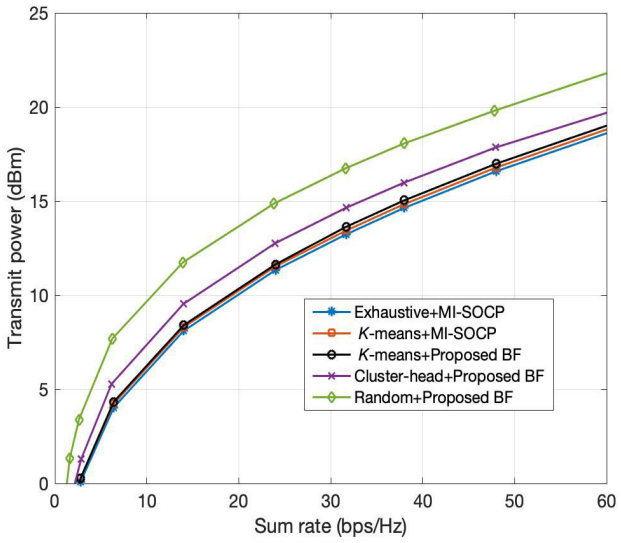


FIGURE 4. **Transmit power versus sum rate for different clustering and beamforming approaches.**

we examine the achievable sum rate of the users within the network. We consider orthogonal multiple access (OMA) and power domain NOMA as benchmarks with similar parameters except the number of multiplexed signals over each subcarrier. Specifically, in OMA, each user is assigned only one subcarrier such that no interference occurs with other user signals. Hence, the maximum number of users in each cluster is equal to the number of subcarriers, i.e., $q = N$. In power domain NOMA, all users have access to all the subcarriers and no constraint is applied to the maximum number of users in a cluster. In this section, we refer to power domain NOMA simply by NOMA.

Figure 5 compares the total transmit power versus achievable sum rate among the proposed SCMA, power domain NOMA and OMA schemes. Two different channel models are considered for this purpose, one with no NLOS path, i.e., $P = 0$, and the other with $P = 3$ NLOS components. It is observed that in both cases, the results of the proposed SCMA scheme outperforms other schemes in terms of sum rate and the performance gap gets larger as the transmit power increases. Moreover, we observe that the transmit power for NOMA is more than that of OMA. However, as the sum rate increases, the results for OMA exhibits a noticeable increase in transmit power compared to NOMA.

To investigate the impact of imperfect CSI on the proposed user clustering and downlink beamforming, we model the estimated channel vector as follows,

$$\hat{\mathbf{h}}_{jk}(n) = \mathbf{h}_{jk}(n) + \mathbf{\Delta}_{jk}(n) \tag{56}$$

where $\mathbf{h}_{jk}(n)$ is the actual channel vector and $\mathbf{\Delta}_{jk}(n)$ is CSI error with i.i.d. entries following a complex Gaussian distribution, i.e., $\mathbf{\Delta}_{jk}(n) \sim \mathcal{CN}(\mathbf{0}, \sigma_e^2\mathbf{I})$.

Figure 6 shows the transmit power comparison for the channel model with $P = 3$ NLOS paths, where the perfect

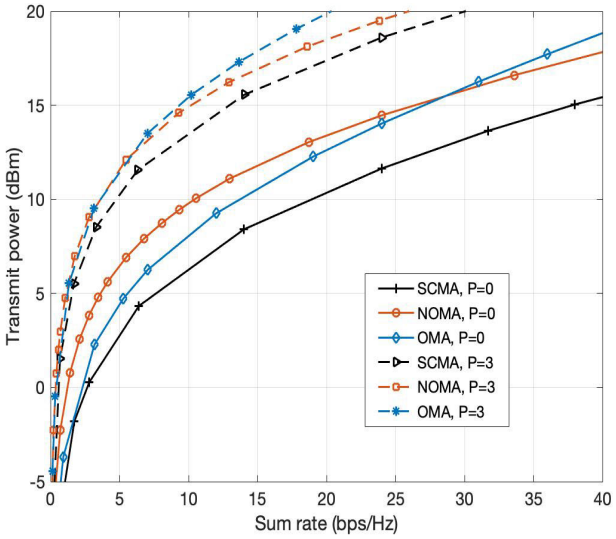


FIGURE 5. Transmit power versus sum rate for different transmission schemes.

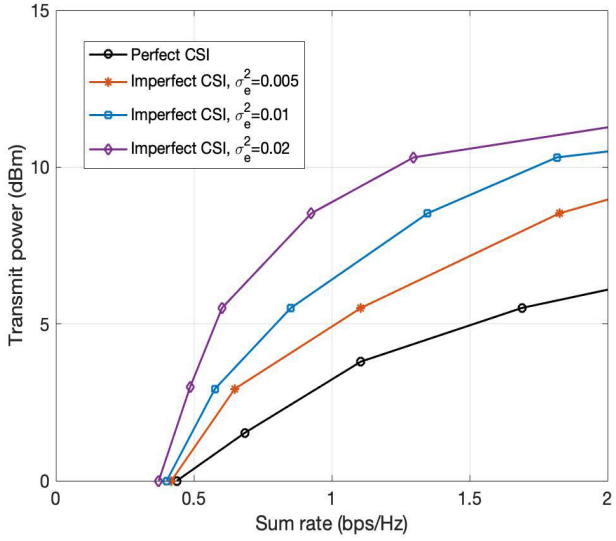


FIGURE 6. Transmit power versus sum rate for perfect and imperfect CSI.

and imperfect CSI with different σ_e scenarios are considered. It can be seen that the system performance is sensitive to the CSI accuracy. This is due to the fact that the channel correlation is used as the similarity metric for the proposed constrained K-means algorithm, which largely depends on the obtained CSI at the central processor. Moreover, the BD algorithm does not work well in the presence of imperfect CSI and can not remove inter-cluster interference totally. In order to enhance the performance of the proposed user clustering and downlink beamforming in the presence of imperfect CSI, one can use a more sophisticated similarity metric in the clustering algorithm or robust beamforming in the second stage of the proposed approach [49]–[51]. However, these considerations are beyond of the scope of this work.

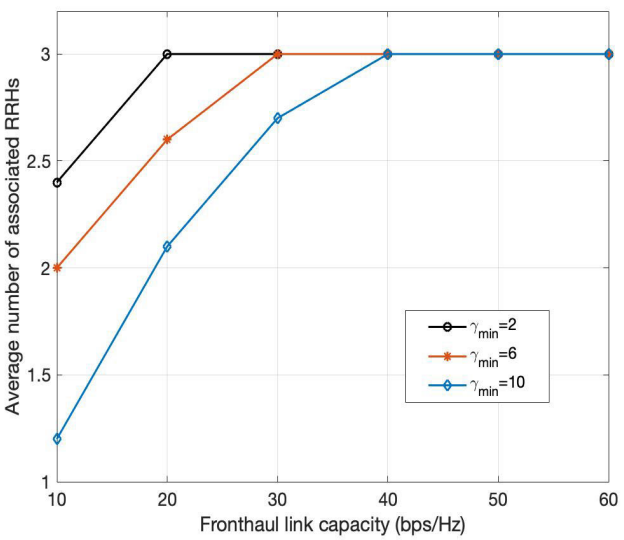


FIGURE 7. Average number of associated RRHs versus fronthaul link capacity for $M = 20$.

Figure 7 presents the average number of associated RRHs per user versus the fronthaul link capacity for different target SINRs, γ_{min}. It can be seen that due to the limitation on the capacity of the fronthaul links, each user can be only served by a small group of RRHs. For fixed γ_{min}, the number of RRHs associated with each user will increase as the fronthaul link capacity grows. Moreover, for a fixed fronthaul link capacity, the group of associated RRHs will increase as γ_{min} gets smaller. In fact, the data rate of each user will become smaller for a lower γ_{min}. Thus, each RRH can serve more users with lower data rate.

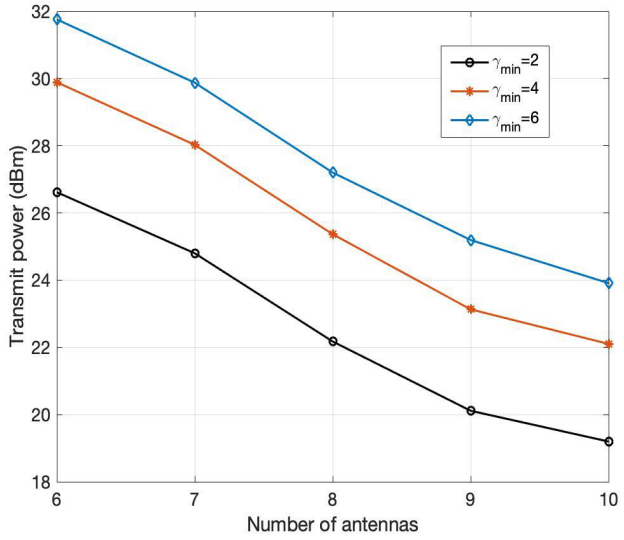


FIGURE 8. Transmit power versus number of antennas.

Figure 8 shows the total transmit power versus number of antennas M for different target SINRs, γ_{min}. In this regard, the channel model with $P = 3$ NLOS paths is considered and

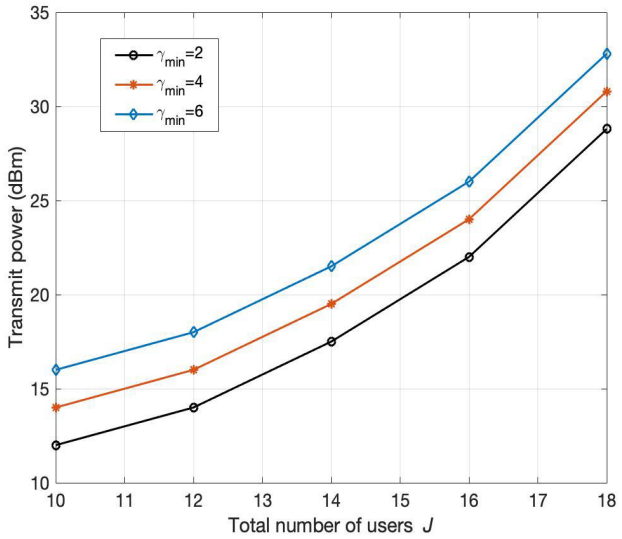


FIGURE 9. Transmit power versus total number of users.

$J = 18$ users are grouped into $K = 4$ clusters. It is worth noting that the number of clusters are determined through the elbow method for this case. For larger M, better beamforming results are expected as more degrees of freedom will be left for inter-cluster interference cancellation. We can observe that the total transmission power decreases as the number of antennas increases which is a consequence of narrower beamforming.

Figure 9 shows the total transmit power versus total number of users J for different target SINRs, γ_{min}. In this regard, the channel model with $P = 3$ NLOS paths is considered and the users are grouped into $K = 4$ clusters. As the results indicate, the total transmission power depends on the number of users and γ_{min}. As the number of users or the target SINR increases, larger total transmission power is needed to satisfy QoS and fronthaul capacity requirements.

VI. CONCLUSION

In this work, the design of user clustering and beamforming approach was investigated for MIMO-SCMA in C-RAN. We proposed a constrained K-means algorithm for user clustering. By taking advantage of CSI available at the central processor, this algorithm was applied to partition users into non-overlapping clusters based on the correlation between channel vectors. The beamforming design was formulated as an optimization problem, with the aim to minimize the total transmit power under the SINR and fronthaul capacity constraints, and two iterative algorithms were proposed for its solution. In the first approach, the high-quality solution was achieved by solving a MI-SOCP in each iteration via dedicated solvers. In the second approach, a two-stage low-complexity beamforming design was proposed where in the first stage, the BD was employed to obtain the cluster beamformers, while in the second stage, the design of user-specific beamformers was formulated as an optimization problem. Through simulations, it was shown that the proposed user clustering and beamforming approaches for MIMO-SCMA systems can effectively decrease total transmit power, eliminate inter-cluster interference, and improve spectral efficiency compared to the benchmark approaches.

APPENDIX A
PROOF OF PROPOSITION 2

In order to prove proposition 3.2, we first transform the cluster assignment subproblem in Algorithm 1 into its equivalent form as a minimum cost flow (MCF) linear network optimization problem. We then show that the optimal selection variable $\iota_{j,k}$ is binary, which can be found using fast network simplex algorithms instead of complex mixed integer linear programming [52].

In general, a MCF problem has an underlying directed graph structure $G = (\mathcal{V}, \mathcal{E})$ defined by a set of vertices (nodes), $\mathcal{V}$, and a set of edges (arcs), $\mathcal{E}$. For each node $\nu \in \mathcal{V}$, we associate a value $b(\nu)$ indicating whether it is a supply node ($b(\nu) > 0$), a demand node ($b(\nu) < 0$), or a transshipment node ($b(\nu) = 0$). For each edge $(\nu, \omega) \in \mathcal{E}$, we associate a flow of $f(\nu, \omega)$ on the edge with cost of $c(\nu, \omega)$ per unit flow. The optimization model for the MCF problem can be formulated as,

$$\min \sum_{(\nu,\omega)\in\mathcal{E}} f(\nu,\omega)c(\nu,\omega) \tag{57a}$$

$$\text{s.t.} \sum_{\omega} f(\nu,\omega) - \sum_{\nu} f(\omega,\nu) = b(\nu), \quad \forall \nu \in \mathcal{V} \tag{57b}$$

$$0 \le f(\nu,\omega) \le u(\nu,\omega), \quad \forall (\nu,\omega) \in \mathcal{E} \tag{57c}$$

where $u(\nu, \omega)$ is the maximum capacity of flow on the edge $(\nu, \omega) \in \mathcal{E}$. The problem is feasible if the sum of the supplies equals the sum of the demands, i.e.,

$$\sum_{\nu\in\mathcal{V}} b(\nu) = 0. \tag{58}$$

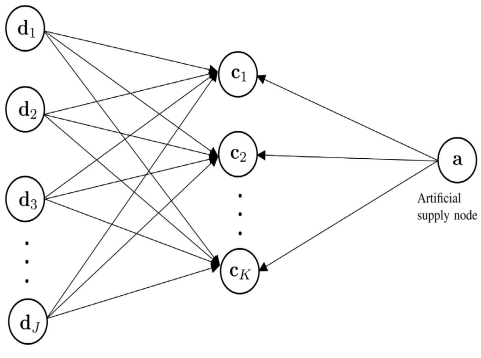


FIGURE 10. The MCF equivalent directed graph structure.

Let each data point $\mathbf{d}_j$ correspond to a supply node with $b(\mathbf{d}_j) = 1$ and each cluster center $\mathbf{c}_k$ correspond to a demand node with $b(\mathbf{c}_k) = -q$. The cost of the edge $(\mathbf{d}_j, \mathbf{c}_k)$ can be expressed as,

$$c(\mathbf{d}_j, \mathbf{c}_k) = \|\mathbf{d}_j - \mathbf{c}_k\|_2^2. \tag{59}$$

To satisfy the feasibility constraint of the problem, we consider an articial supply node, $\mathbf{a}$, such that,

$$b(\mathbf{a}) = -J + Kq. \tag{60}$$

This artificial node has no edge to or from data points, while the cost of edge form node $\mathbf{a}$ to cluster center $\mathbf{c}_k$ is zero, i.e. $c(\mathbf{a}, \mathbf{c}_k) = 0\ \forall k \in \mathcal{K}$. These identifications establish the equivalence between the MCF and the cluster assignment subproblem in Algorithm 1 in which the selection variable $\boldsymbol{\iota}_{j,k}$ corresponds to flow $f(\mathbf{d}_j, \mathbf{c}_k)$. The MCF equivalent directed graph structure is shown in Figure 9.

According to [52, Proposition 5.4], since $b(\mathbf{d}_j)$, $b(\mathbf{c}_k)$, and $b(\mathbf{a})$ are all integers, the optimal flow solution is integer-valued. Since the selection variable $\boldsymbol{\iota}_{j,k}$ corresponds to flow $f(\mathbf{d}_j, \mathbf{c}_k)$, and since $\sum_k f(\mathbf{d}_j, \mathbf{c}_k) = 1$, the optimal $\boldsymbol{\iota}_{j,k}$ is integer with maximum value equal to 1, i.e. $\boldsymbol{\iota}_{j,k} \in \{0, 1\}$.

The MCF formulation allows one to solve the cluster assignment subproblem via network simplex algorithm which is faster than general linear programming codes. Specifically, the complexity of solving cluster assignment subproblem via network simplex algorithm is given by [52],

$$O(|\mathcal{V}||\mathcal{E}|^2(\log(|\mathcal{V}|))^2) \tag{61}$$

where the number of vertices $|\mathcal{V}|$, and number of edges $|\mathcal{E}|$ in our case are,

$$|\mathcal{V}| = J + K + 1, \tag{62}$$
$$|\mathcal{E}| = JK + K. \tag{63}$$

It is of interest to investigate the asymptotic complexity of the algorithms when J and K are large, i.e., when we let $J > K \to \infty$. Under this condition, we can obtain the asymptotic complexity as,

$$C \triangleq O(J^3K^2(\log(J))^2). \tag{64}$$

REFERENCES

[1] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[2] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[3] Y. Cai, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann, "Modulation and multiple access for 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 629–646, 1st Quart., 2018.

[4] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE Veh. Technol. Conf.*, Las Vegas, NV, USA, Sep. 2014, pp. 1–5.

[5] M.-M. Zhao, Y. Cai, M.-J. Zhao, B. Champagne, and T. A. Tsiftsis, "Improving caching efficiency in content-aware C-RAN-based cooperative beamforming: A joint design approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4125–4140, Jun. 2020.

[6] M. Alam and Q. Zhang, "Designing optimum mother constellation and codebooks for SCMA," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017, pp. 1–6.

[7] L. Yu, P. Fan, X. Lei, and P. T. Mathiopoulos, "Ber analysis of SCMA systems with codebooks based on star-QAM signaling constellations," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 1925–1928, Sep. 2017.

[8] J. Bao, Z. Ma, M. Xiao, T. A. Tsiftsis, and Z. Zhu, "Bit-interleaved coded SCMA with iterative multiuser detection: Multidimensional constellations design," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5292–5304, Nov. 2018.

[9] L. Yu, P. Fan, D. Cai, and Z. Ma, "Design and analysis of SCMA codebook based on star-QAM signaling constellations," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10543–10553, Nov. 2018.

[10] Q. Wang, T. Li, R. Feng, and C. Yang, "An efficient large resource-user scale SCMA codebook design method," *IEEE Commun. Lett.*, vol. 23, no. 10, pp. 1787–17904, Jul. 2019.

[11] Y. Du, B. Dong, Z. Chen, P. Gao, and J. Fang, "Joint sparse graph-detector design for downlink MIMO-SCMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 1, pp. 14–17, Feb. 2016.

[12] Z. Wu, C. Zhang, X. Shen, and H. Jiao, "Low complexity uplink SFBC-based MIMO-SCMA joint decoding algorithm," in *Proc. 3rd IEEE Int. Conf. Comput. Commun.*, Chengdu, China, Dec. 2017, pp. 968–972.

[13] C. Yan, N. Zhang, and G. Kang, "Downlink multiple input multiple output mixed sparse code multiple access for 5G system," *IEEE Access*, vol. 6, pp. 20837–20847, 2018.

[14] W. Yuan, N. Wu, Q. Guo, Y. Li, C. Xing, and J. Kuang, "Iterative receivers for downlink MIMO-SCMA: Message passing and distributed cooperative detection," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3444–3458, May 2018.

[15] F. J. Martin-Vega, Y. Liu, G. Gomez, M. C. Aguayo-Torres, and M. Elkashlan, "Modeling and analysis of NOMA enabled CRAN with cluster point process," in *Proc. Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–6.

[16] X. Gu, X. Ji, Z. Ding, W. Wu, and M. Peng, "Outage probability analysis of non-orthogonal multiple access in cloud radio access networks," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 149–152, Jan. 2018.

[17] J. Zhao, Y. Liu, T. Mahmoodi, K. K. Chai, Y. Chen, and Z. Han, "Resource allocation in cache-enabled CRAN with non-orthogonal multiple access," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[18] M. Moltafet, S. Parsaeefard, M. R. Javan, and N. Mokari, "Robust radio resource allocation in MISO-SCMA assisted C-RAN in 5G networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5758–5768, Jun. 2019.

[19] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.

[20] S. Norouzi, A. Morsali, and B. Champagne, "Optimizing transmission rate in NOMA via block diagonalization beamforming and power allocation," in *Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process. (PACRIM)*, Victoria, BC, Canada, Aug. 2019, pp. 1–5.

[21] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, Apr. 2003.

[22] M. Vameghestahbanati, I. D. Marsland, R. H. Gohary, and H. Yanikomeroglu, "Multidimensional constellations for uplink SCMA systems—A comparative study," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2169–2194, 3rd Quart., 2019.

[23] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7425–7440, Nov. 2018.

[24] J. Cui, Z. Ding, and P. Fan, "The application of machine learning in mmWave-NOMA systems," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Porto, Portugal, Jun. 2018, pp. 1–6.

[25] G. Lee, Y. Sung, and J. Seo, "Randomly-directional beamforming in millimeter-wave multiuser MISO downlink," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1086–1100, Feb. 2016.

[26] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[27] T. Niknam, E. T. Fard, N. Pourjafarian, and A. Rousta, "An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering," *Eng. Appl. Artif. Intell.*, vol. 24, no. 2, pp. 306–317, Mar. 2011.

[28] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured massive access for scalable cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1086–1100, Apr. 2021.

[29] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[30] P. S. Bradley, O. L. Mangasarian, and W. N. Street, "Clustering via concave minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 368–374.

[31] P. S. Bradley, K. P. Bennett, and A. Demiriz, "Constrained K-means clustering," Microsoft Res., Redmond, WA, USA, Tech. Rep., May 2000.
[32] M. Inaba, N. Katoh, and H. Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering," in *Proc. 10th Annu. Symp. Comput. Geometry (SCG)*, New York, NY, USA, 1994, pp. 332–339.
[33] N. G. Andrew, "Clustering with K-means algorithm," *Mach. Learn.*, pp. 1–5, Sep. 2012.
[34] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Jan. 1987.
[35] R. Tibshirani, G. Walther, and T. Hastie, "Stimating the number of clusters in a data set via the gap statistic," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 63, no. 2, pp. 42–411, 2001.
[36] P. Luong, F. Gagnon, C. Despins, and L.-N. Tran, "Optimal joint remote radio head selection and beamforming design for limited fronthaul C-RAN," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5605–5620, Nov. 2017.
[37] Y. Cheng, M. Pesavento, and A. Philipp, "Joint network optimization and downlink beamforming for CoMP transmissions using mixed integer conic programming," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3972–3987, Aug. 2013.
[38] M. Tawarmalani and N. V. Sahinidis, "Global optimization of mixed-integer nonlinear programs: A theoretical and computational study," *Math. Program.*, vol. 99, no. 3, pp. 563–591, 2004.
[39] MOSEK ApS. (2019). *The MOSEK Optimization Toolbox for MATLAB Manual. Version 9.0.* [Online]. Available: http://docs.mosek.com/9.0/toolbox/index.html
[40] Gurobi Optimization, LLC. (2021). *Gurobi Optimizer Reference Manual*. [Online]. Available: https://www.gurobi.com
[41] X. Zhu, Z. Wang, C. Qian, L. Dai, J.i Chen, S. Chen, and L. Hanzo, "Soft pilot reuse and multicell block diagonalization precoding for massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3285–3298, May 2016.
[42] 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.
[43] E. Chen, M. Tao, and Y.-F. Liu, "Joint base station clustering and beamforming for non-orthogonal multicast and unicast transmission with backhaul constraints," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6265–6279, Sep. 2018.
[44] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.
[45] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[46] G. R. Lanckriet and B. K. Sriperumbudur, "On the convergence of the concave-convex procedure," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1759–1767.
[47] M. Grant and S. Boyd, "CVX: MATLAB software for disciplined convex programming, version 2.1," CVX Res., Stanford, CA, USA, Tech. Rep., Mar. 2014.
[48] S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE Access*, vol. 5, pp. 565–577, 2016.
[49] Y. Teng and W. Zhao, "Robust group sparse beamforming for dense C-RANs with probabilistic SINR constraints," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.
[50] Y. Wang, L. Ma, and Y. Xu, "Robust beamforming and base station activation for energy efficient downlink C-RAN," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Toronto, ON, Canada, Sep. 2017, pp. 1–5.
[51] D. Yan, R. Wang, E. Liu, and Q. Hou, "Robust beamforming optimization for downlink cloud radio access networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.
[52] D. P. Bertsekas, *Network Optimization: Continuous and Discrete Models*. Belmont, MA, USA: Athena Scientific, 1998.

SARA NOROUZI received the B.Sc. degree in electrical engineering from Shiraz University of Technology, Shiraz, Iran, in 2012, and the M.Sc. degree in electrical engineering from Shiraz University, Shiraz, in 2016. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada. From 2016 to 2017, she was with Iranian Huawei Technologies Company, Tehran, Iran, where she worked as an RF Engineer. Her research interests include signal processing, wireless communications, optimization, and artificial intelligence. She was a recipient of the McGill Engineering Doctoral Award.

YUNLONG CAI (Senior Member, IEEE) received the M.Sc. degree in electronic engineering from the University of Surrey, Guildford, U.K., in 2006, and the Ph.D. degree in electronic engineering from the University of York, York, U.K., in 2010. From 2010 to 2011, he was a Postdoctoral Fellow with the Electronics and Communications Laboratory, Conservatoire National des Arts et Metiers, Paris, France. Since February 2011, he has been with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, where he is currently a Full Professor. From August 2016 to January 2017, he was a Visiting Scholar with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. His research interests include transceiver design for multiple-antenna systems, cooperative and relay communications, UAV communications, and machine learning for communications. He is also an Associate Editor of IEEE Signal Processing Letters.

BENOIT CHAMPAGNE (Senior Member, IEEE) received the B.Ing. degree in engineering physics from the École Polytechnique de Montréal, in 1983, the M.Sc. degree in physics from the Université de Montréal, 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 Professor and then an Associate Professor with the INRS-Telecommunications, Université du Quebec, Montreal. He joined McGill University, Montreal, in 1999, where he is currently a Full Professor with the Department of Electrical and Computer Engineering. From 2004 to 2007, he worked as an Associate Chair of Graduate Studies with the Department of Electrical and Computer Engineering. His research focuses on the study of advanced algorithms for the processing of information bearing signals by digital means. He has coauthored more than 300 refereed publications in these areas. His research has been funded by the Natural Sciences and Engineering Research Council of Canada, the Fonds de Recherche sur la Nature et les Technologies from the Government of Quebec, and some major industrial sponsors, including Nortel Networks, Bell Canada, InterDigital, and Microsemi. His research interests include areas of statistical signal processing, including detection and estimation, sensor array processing, adaptive filtering, and applications thereof to broadband communications and audio processing. He has been an Associate Editor of the IEEE Signal Processing Letters, the IEEE Transactions on Signal Processing, and the *EURASIP Journal on Applied Signal Processing*. He has also served on the technical committees for several international conferences in the fields of communications and signal processing.