Learning-Assisted User Clustering in Cell-Free Massive MIMO-NOMA Networks

Quang Nhat Le, Van-Dinh Nguyen, Octavia A. Dobre,
Nam-Phong Nguyen, Ruiqin Zhao, and Symeon Chatzinotas

***Abstract*—The superior spectral efficiency (SE) and user fairness feature of non-orthogonal multiple access (NOMA) systems are achieved by exploiting user clustering (UC) more efficiently. However, a random UC certainly results in a suboptimal solution while an exhaustive search method comes at the cost of high complexity, especially for systems of medium-to-large size. To address this problem, we develop two efficient unsupervised machine learning based UC algorithms, namely k-means++ and improved k-means++, to effectively cluster users into disjoint clusters in cell-free massive multiple-input multiple-output (CFmMIMO) system. Adopting full-pilot zero-forcing at access points (APs) to comprehensively assess the system performance, we formulate the sum SE optimization problem taking into account power constraints at APs, necessary conditions for implementing successive interference cancellation, and required SE constraints at user equipments. The formulated optimization problem is highly non-convex, and thus, it is difficult to obtain the global optimal solution. Therefore, we develop a simple yet efficient iterative algorithm for its solution. In addition, the performance of collocated massive MIMO-NOMA (COmMIMO-NOMA) system is also characterized. Numerical results are provided to show the superior performance of the proposed UC algorithms compared to baseline schemes. The effectiveness of applying NOMA in CFmMIMO and COmMIMO systems is also validated.**

***Index Terms*—Cell-free massive multiple-input multiple-output, full-pilot zero-forcing, k-means, machine learning, non-orthogonal multiple access, power allocation, user clustering.**

I. Introduction

The tremendous growth in the number of emerging applications will certainly pose enormous traffic demands with ultra-high connection density for next-generation wireless networks. It is approximated that more than 20 billion devices were connected to the Internet in 2020, and this number is predicted to exceed 35 billion devices by 2025 [1]. The global data traffic of mobile devices is expected to reach 226 exabytes per month by 2026 [2], and will further increase over the next decade. However, traditional orthogonal multiple-access (OMA) techniques seem to reach their fundamental limits in the near future, and therefore are no longer suitable to meet these requirements. Consequently, it calls for innovative techniques that utilize radio resources more efficiently to attain the optimal performance.

Non-orthogonal multiple-access (NOMA) has been envisaged as a key enabling technology that significantly enhances spectral efficiency (SE) and user fairness of traditional wireless communication systems [3]–[5]. In NOMA, multiple user equipments (UEs) are allowed to simultaneously transmit and receive their signals in the same time-frequency resource by using different signal signatures (i.e., code-domain NOMA) or power levels (i.e., power-domain NOMA) [6]–[8][1]. In particular, in a downlink system the key benefit of NOMA is attributed to the fact that UEs with better channel conditions are able to cancel the interference caused by UEs with poorer channel conditions using the successive interference cancellation (SIC) technique. User fairness is then achieved by allocating a large portion of the total power budget to weak UEs, which also guarantees the SIC's feasibility at strong UEs.

Recently, cell-free massive multiple-input multiple-output (CFmMIMO), which is a scalable version of massive MIMO networks, has been introduced to overcome the large propagation losses as well as provide better quality-of-experience services for cell-edge UEs [9]–[11]. CFmMIMO comprises a large number of access points (APs) that are spatially distributed over a wide area to coherently serve multiple UEs in the same time-frequency resources. All APs are coordinated by a central processing unit (CPU) through fronthaul links. Each AP performs beamforming based on its local channel state information (CSI) only, and this feature thus greatly reduces the complexity in terms of the fronthaul overhead. Since each UE is coherently served by all APs, the effect of cell boundaries can be effectively removed. It was shown in [9] and [12] that CFmMIMO is superior to small-cell and collocated massive MIMO (COmMIMO) in terms of SE and energy efficiency (EE), respectively. However, the key advantages of favorable propagation and channel hardening properties to multiplex numerous UEs are only achieved in the case of multiple antennas at APs and/or low propagation losses [13]. For the aforementioned reasons, it is of pivotal interest to study the combination of NOMA and CFmMIMO to reap all their benefits, towards fulfilling the conflicting demands on high SE, massive connectivity with low latency, and high reliability with user fairness of future wireless networks [14].

Q. N. Le and O. A. Dobre are with the Dept. of Electrical and Computer Engineering, Memorial University, St. John's, NL A1B 3X9, Canada (e-mail: {qnle, odobre}@mun.ca).

V.-D. Nguyen and S. Chatzinotas are with the Interdisciplinary Centre for Security, Reliability, and Trust (SnT) – University of Luxembourg, L-1855, Luxembourg (e-mail: {dinh.nguyen, symeon.chatzinotas}@uni.lu).

N-P. Nguyen is with the School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi, Vietnam (e-mail: phong.nguyennam@hust.edu.vn).

R. Zhao is with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: rqzhao@nwpu.edu.cn).

[1]This paper will focus on power-domain NOMA; henceforth, we refer to it as NOMA.

A. Related Work

Despite its potential, there are only a few research works investigating the benefit of NOMA in CFmMIMO systems in the literature. NOMA for downlink CFmMIMO was first studied in [15], where the closed-form expression of the achievable sum rate was derived. Numerical results showed the superior performance of NOMA compared to OMA. The authors in [16] investigated the impact of NOMA in the uplink CFmMIMO system and derived a closed-form approximation for the sum SE (SSE). Simulation results demonstrated that CFmMIMO-NOMA is capable of utilizing the scarce spectrum more efficiently. In [17], different types of precoding techniques such as maximum ratio transmission (MRT), full-pilot zero-forcing (fpZF), and modified regularized ZF (mRZF) at APs were considered in downlink CFmMIMO-NOMA. It was shown that downlink CFmMIMO-NOMA with mRZF and fpZF precoders significantly outperforms the OMA with MRT in terms of the achievable sum rate. These existing works mainly focused on characterizing the performance analysis in CFmMIMO-NOMA, but did not show how UEs are paired/grouped.

To be spectrally-efficient, it is crucial to group a sufficiently large number of UEs with distinct channel conditions that performs NOMA jointly [3]–[5], [18]. In the context of CFmMIMO-NOMA, the authors in [19] proposed three distance-based pairing schemes including near pairing, far pairing, and random pairing to group UEs into disjoint clusters. It is not surprising to see that the close pairing, where two UEs with the smallest distance between them are paired, provides the worst performance, which is also aligned with the NOMA principle [3], [4]. Another interesting study is to group a large number of UEs into one cluster [20], referred to as user clustering (UC), in which a low complexity suboptimal method based on the Jaccard distance coefficient was developed to find the most dissimilar UEs in the CFmMIMO-NOMA system. Nevertheless, the UC algorithms in the above-cited works were developed based on the distances among UEs only, without considering any learning features.

Recently, unsupervised machine learning (ML) techniques have been considered as an effective means for different optimization targets, which exploit adaptive learning features. In this regard, the authors in [21] proposed a kernel-power-density based algorithm to cluster multipath components of MIMO channels into disjoint groups. A cluster-based geometrical dynamic stochastic model was introduced in [22], where scattered nodes were grouped into different clusters according to the density of nodes in MIMO scenarios. In [23], a clustered sparse Bayesian learning algorithm was developed for channel estimation in a hybrid analog-digital massive MIMO system by using the sparsity characteristic of angular domain channel. The authors in [24] proposed a clustering scheme for machine-to-machine communications in a hybrid time-division multiple access-NOMA system in order to increase the battery lifetime of machines, using the popular k-means algorithm [25]. This work was extended in [26] to improve the network sum throughput by considering an enhanced k-means algorithm. Further, the k-means algorithm was used to cluster UEs in mmwave-NOMA [27] and CFmMIMO [28]. Although these works demonstrated the effectiveness of applying unsupervised ML to clustering tasks for various wireless communication systems, its application for UC in CFmMIMO-NOMA has not been previously studied.

On the other hand, the k-means has also been considered as the most well-known data clustering algorithm due to its simple implementation, that allows to provide more insight into the underlying nature and structure of the data. There are several variants of the k-means algorithm based on choosing different representative points for the clusters, including the k-medoids [29], k-medians [30], k-modes [31], and employing feature transformation techniques, including weighted k-means [32] and kernel k-means [33]. Different from the k-means algorithm where the representative point for each cluster is the mean of all the points within each cluster, the representative point for each cluster in the k-medoids, k-medians, and k-modes algorithms is the actual data point inside each cluster, the median of each cluster, and the mode of each cluster, respectively. Although the k-medoids and k-medians are more robust to outliers than the k-means, their computational complexity is much higher and therefore not suitable for large datasets. Moreover, the k-modes is designed to handle categorical data, and thus not appropriate for numerical data. Given that the k-means algorithm considers all features equally important, the weighted k-means introduces a feature weighting mechanism, where different features are assigned different weights [32]. In [33], the kernel functions are applied in the k-means in order to find non-linearly separable clusters. However, both the weighted and kernel k-means algorithms are computationally more expensive than the k-means.

B. Motivation and Main Contributions

In CFmMIMO-NOMA systems, the effect of network interference is increasingly abnormal and acute as the APs become denser. Most existing works on CFmMIMO-NOMA systems [15]–[17] focused on the performance analysis while they neglect the importance of UC, which has been shown to significantly improve the performance of NOMA-based systems [3], [4], [34]. A direct application of random UC schemes [4], [18] to CFmMIMO-NOMA systems would result in poor performance, even worse than the traditional linear beamforming without NOMA. In addition, a joint UC and beamforming design [5], which clusters UEs by means of the tensor model, is not very practical for CFmMIMO-NOMA due to excessively high complexity in terms of computational and signalling overhead. Although the k-means algorithm has been widely adopted for different clustering tasks [24]–[28], its main drawback is sensitivity to the initialization of centroids.

Taking into account all these issues, in this paper we devise novel UC algorithms along with an efficient transmission strategy such that the SSE of CFmMIMO-NOMA systems is remarkably enhanced. In particular, our main contributions are summarized as follows:

- We propose two efficient unsupervised ML-based UC algorithms, including k-means++ and improved k-means++, to effectively cluster UEs into disjoint clusters

in CFmMIMO-NOMA. The proposed k-means++ algorithms further address the limitation of k-means due to the randomness of initial centroids. In addition, they are able to ensure the maximum number of UEs per cluster, which can not be achieved by the conventional k-means.
- By adopting the fpZF precoding at APs, we formulate optimization problems for both CFmMIMO-NOMA and COmMIMO-NOMA systems by incorporating power constraints at APs, necessary conditions for implementing SIC at UEs, and the minimum SE requirement at UEs; these belong to the difficult class of nonconvex optimization problems. Towards appealing applications, two low-complexity iterative algorithms based on the inner approximation (IA) method [35] are developed for their solutions, which are guaranteed to converge to at least a locally optimal solution.
- Extensive numerical results are provided to confirm the effectiveness of the proposed UC algorithms on the SSE performance over the current state-of-the-art approaches, i.e., close-, far- and random-pairing schemes [19], and Jaccard-based UC scheme [20]. They also show the significantly achieved SSE gains of CFmMIMO-NOMA over COmMIMO-NOMA.

C. Paper Organization and Notations

The remainder of this paper is organized as follows. Section II describes the system model. In Section III, two unsupervised ML-based UC algorithms are presented. The proposed iterative algorithms for CFmMIMO-NOMA and COmMIMO-NOMA are introduced in Sections IV and V, respectively. Numerical results are given in Section VI, while Section VII concludes the paper.

Notations: Bold uppercase letters, bold lowercase letters, and lowercase characters stand for matrixes, vectors, and scalars, respectively. $|\cdot|$, $(\cdot)^H$, $(\cdot)^T$, $(\cdot)^*$, and $||\cdot||_2$ correspond to the cardinality, the Hermitian transpose, the transpose, the conjugate, and the $l2-$norm operators, respectively. $\mathbb{E}[\cdot]$ represents the expectation operation. $\mathcal{CN}(\mu, \sigma^2)$ stands for circularly symmetric complex Gaussian random variable (RV) with mean μ and variance σ^2.

II. System Model

A. System Description

We consider an CFmMIMO-NOMA system, where the set $\mathcal{M} \triangleq \{1, 2, \cdots, M\}$ of M APs is connected to the CPU through perfect wired fronthaul links to serve the set $\mathcal{N} \triangleq \{1, 2, \cdots, N\}$ of N UEs via a shared wireless medium, as shown in Fig. 1. Each AP is equipped with K antennas, while each UE has a single antenna. APs and UEs are assumed to be randomly distributed in a wide coverage area. The communication between APs and UEs follows the time division duplex (TDD) mode. Each coherence interval, denoted by τ_c, includes two phases: uplink training τ_p ($\tau_p < \tau_c$) and downlink data transmission ($\tau_c - \tau_p$). The total N UEs are grouped into L clusters and each UE belongs to one cluster only. We denote the set of L clusters by $\mathcal{L} \triangleq \{1, 2, \cdots L\}$. The set of UEs in the l-th cluster is defined as $\mathcal{N}_l \triangleq \{1_l, \cdots, n_l, \cdots, N_l\}$ with $|\mathcal{N}_l| = N_l$, where $\bigcup_{l\in\mathcal{L}} |\mathcal{N}_l| = N$ and $\mathcal{N}_l \bigcap \mathcal{N}_{l'} = \emptyset$ for $l \neq l'$.

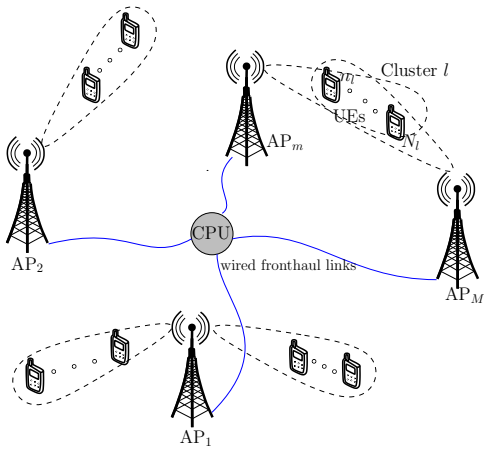

Fig. 1. An illustration of the CFmMIMO-NOMA system.

B. Signal Model and Sum Spectral Efficiency (SSE)

1) Uplink Training

In the uplink training phase, all UEs send their training pilots to APs for channel estimation. Then, downlink channels are achieved by leveraging the channel reciprocity property of the TDD mode. With the aim of minimizing the channel estimation overhead in CFmMIMO-NOMA, UEs in the same cluster share the same pilot sequence, and the pilot sequences among different clusters are pairwisely orthogonal [15], [19] which requires $\tau_p \geq L$. In this paper, we assume that $\tau_p = L$. Let us denote the pilot sequence sent from the UEs in the l-th cluster by $\boldsymbol{\phi}_l \in \mathbb{C}^{\tau_p \times 1}$ with $l \in \{1, 2, \ldots, \tau_p\}$, satisfying the orthogonality, i.e., $\|\boldsymbol{\phi}_l\|_2^2 = \tau_p$ and $\boldsymbol{\phi}_l^H \boldsymbol{\phi}_{l'} = 0$ if $l \neq l'$. The channel vector from UE n_l to AP_m is defined as $\mathbf{h}_{m,n_l} \in \mathbb{C}^{K\times 1}$. In this paper, we focus on slowly time-varying channels, and assume that the channel coefficients are static during the τ_c interval. The channel $\mathbf{h}_{m,n_l}$ is generally modeled as follows:

$$\mathbf{h}_{m,n_l} = \sqrt{\beta_{m,n_l}} \bar{\mathbf{h}}_{m,n_l}, \tag{1}$$

where β_{m,n_l} represents the large-scale fading coefficient accounting for path loss and shadowing, and $\bar{\mathbf{h}}_{m,n_l} \in \mathbb{C}^{K\times 1}$ is the small-scale fading vector in which the components are independent and identically distributed (i.i.d.) $\mathcal{CN}(0, 1)$ RVs. The training signals received at AP_m can be written as follows:

$$\mathbf{Y}_m^p = \sum_{l\in\mathcal{L}} \sum_{n_l\in\mathcal{N}_l} \sqrt{\rho_{n_l}} \mathbf{h}_{m,n_l} \boldsymbol{\phi}_l^H + \mathbf{W}_m^p, \tag{2}$$

where ρ_{n_l} and $\mathbf{W}_m^p \in \mathbb{C}^{K\times\tau_p}$ are the normalized transmit power of UE n_l and the additive noise matrix at AP_m whose elements follow $\mathcal{CN}(0, 1)$, respectively.

Given $\mathbf{Y}_m^p$, AP_m estimates $\mathbf{h}_{m,n_l}$ using the minimum mean square error (MMSE) criterion. The projection $\hat{\mathbf{y}}_m^p \in \mathbb{C}^{K\times 1}$ of $\mathbf{Y}_m^p$ at AP_m onto $\boldsymbol{\phi}_l$ can be derived as follows:

$$\hat{\mathbf{y}}_m^p = \mathbf{Y}_m^p \boldsymbol{\phi}_l = \tau_p \sum_{n_l\in\mathcal{N}_l} \sqrt{\rho_{n_l}} \mathbf{h}_{m,n_l} + \mathbf{W}_m^p \boldsymbol{\phi}_l. \tag{3}$$

Hence, the MMSE estimate of $\mathbf{h}_{m,n_l}$ is given as

$$\hat{\mathbf{h}}_{m,n_l} = \mathbb{E}\{\mathbf{h}_{m,n_l} (\hat{\mathbf{y}}_m^p)^H\} \left(\mathbb{E}\{\hat{\mathbf{y}}_m^p (\hat{\mathbf{y}}_m^p)^H\}\right)^{-1} \hat{\mathbf{y}}_m^p$$

$$= \upsilon_{m,n_l}\hat{\mathbf{y}}_m^p, \tag{4}$$

where $\upsilon_{m,n_l} = \dfrac{\sqrt{\rho_{n_l}}\beta_{m,n_l}}{\tau_p \sum\limits_{n'_l \in \mathcal{N}_l} \rho_{n'_l}\beta_{m,n'_l} + 1}$. The estimation error vector of $\mathbf{h}_{m,n_l}$ is given as

$$\mathbf{e}_{m,n_l} = \mathbf{h}_{m,n_l} - \hat{\mathbf{h}}_{m,n_l}, \tag{5}$$

where the elements of $\mathbf{e}_{m,n_l}$ and $\hat{\mathbf{h}}_{m,n_l}$ are i.i.d. RVs distributed as $\mathcal{CN}\left(\mathbf{0}, (\beta_{m,n_l} - \gamma_{m,n_l})\mathbf{I}_K\right)$ and $\mathcal{CN}\left(\mathbf{0}, \gamma_{m,n_l}\mathbf{I}_K\right)$, respectively, with $\gamma_{m,n_l} = \frac{\tau_p \rho_{n_l}\beta_{m,n_l}^2}{\tau_p \sum\limits_{n'_l \in \mathcal{N}_l} \rho_{n'_l}\beta_{m,n'_l}+1}$. Note that there is no cooperation among APs to exchange the channel estimate information.

Remark 1. *The so-called pilot contamination exists when APs estimate the channels of UEs belonging to the same cluster. The relationship of channel estimates of UE n_l and UE n'_l in the l-th cluster with $n_l \neq n'_l$ and $n_l, n'_l \in \mathcal{N}_l$, at AP$_m$ is expressed as follows:*

$$\hat{\boldsymbol{h}}_{m,n_l} = \frac{\sqrt{\rho_{n_l}}\beta_{m,n_l}}{\sqrt{\rho_{n'_l}}\beta_{m,n'_l}}\hat{\boldsymbol{h}}_{m,n'_l}. \tag{6}$$

2) Downlink Data Transmission

Under TDD operation, we consider the channel reciprocity to acquire CSI to precode the transmit signals in the downlink [9], [12]. In this paper, we adopt the fpZF precoding [36] to cancel inter-cluster interference, but still take into account intra-cluster interference. Compared with the pure ZF [37], each AP computes fpZF precoding using its local CSI only, leading to a distributed implementable algorithm. From (2), the full-rank matrix $\tilde{\mathbf{H}}_m \in \mathbb{C}^{K\times\tau_p}$ of fpZF precoder at AP$_m$ is given by [36]

$$\tilde{\mathbf{H}}_m = \mathbf{Y}_m^p\boldsymbol{\phi}, \tag{7}$$

where $\boldsymbol{\phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \cdots, \boldsymbol{\phi}_{\tau_p}] \in \mathbb{C}^{\tau_p\times\tau_p}$ denotes the collection of τ_p orthogonal pilot sequences. Hence, from (4) and (7), the channel estimate $\hat{\mathbf{h}}_{m,n_l}$ is rewritten as

$$\hat{\mathbf{h}}_{m,n_l} = \upsilon_{m,n_l}\tilde{\mathbf{H}}_m\boldsymbol{\varphi}_l, \tag{8}$$

where $\boldsymbol{\varphi}_l$ is the l-th column of the identity matrix $\mathbf{I}_{\tau_p}$. From (7) and (8), the beamforming vector $\boldsymbol{w}_{m,l} \in \mathbb{C}^{K\times 1}$ oriented to the l-th cluster at AP$_m$ can be expressed as follows:

$$\boldsymbol{w}_{m,l} = \frac{\tilde{\mathbf{H}}_m\left(\tilde{\mathbf{H}}_m^H\tilde{\mathbf{H}}_m\right)^{-1}\boldsymbol{\varphi}_l}{\sqrt{\mathbb{E}\left\{\left\|\tilde{\mathbf{H}}_m\left(\tilde{\mathbf{H}}_m^H\tilde{\mathbf{H}}_m\right)^{-1}\boldsymbol{\varphi}_l\right\|_2^2\right\}}}. \tag{9}$$

The transmitted signal $\boldsymbol{x}_m \in \mathbb{C}^{K\times 1}$ from AP$_m$ is given by

$$\boldsymbol{x}_m = \sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l}\sqrt{\rho_{n_l}^m}\boldsymbol{w}_{m,l}x_{n_l}, \tag{10}$$

where x_{n_l} is the symbol intended for UE n_l, and $\rho_{n_l}^m$ is the normalized transmit power (normalized by the noise power at AP$_m$) allocated to UE n_l at AP$_m$. Besides, x_{n_l} and $x_{n'_{l'}}$ for $l, l' \in \mathcal{L}$ and $n_l, n'_{l'} \in \mathcal{N}$ must satisfy the following condition

$$\mathbb{E}\{x_{n_l}(x_{n'_{l'}})^*\} = \begin{cases} 1, & \text{if } \ l = l' \ \text{ and } \ n = n', \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Then, the received signal at UE n_l in the l-th cluster can be written as

$$\begin{aligned} y_{n_l} &= \sum_{m\in\mathcal{M}}\mathbf{h}_{m,n_l}^H\boldsymbol{x}_m + z_{n_l} \\ &= \underbrace{\sum_{m\in\mathcal{M}}\sqrt{\rho_{n_l}^m}\mathbf{h}_{m,n_l}^H\boldsymbol{w}_{m,l}x_{n_l}}_{\text{Desired signal}} \\ &\quad + \underbrace{\sum_{m\in\mathcal{M}}\sum_{n'_l\in\mathcal{N}_l\setminus\{n_l\}}\sqrt{\rho_{n'_l}^m}\mathbf{h}_{m,n_l}^H\boldsymbol{w}_{m,l}x_{n'_l}}_{\text{Intra-cluster interference before SIC}} \\ &\quad + \underbrace{\sum_{m\in\mathcal{M}}\sum_{l'\in\mathcal{L}\setminus\{l\}}\sum_{n_{l'}\in\mathcal{N}_{l'}}\sqrt{\rho_{n_{l'}}^m}\mathbf{h}_{m,n_l}^H\boldsymbol{w}_{m,l'}x_{n_{l'}}}_{\text{Inter-cluster interference}} + z_{n_l}, \end{aligned} \tag{12}$$

where $z_{n_l} \sim \mathcal{CN}(0,1)$ is the additive white Gaussian noise (AWGN) at UE n_l.

Without loss of generality, in the l-th cluster we consider a descending order of channel gain, i.e., UEs 1_l and N_l are the users with strongest and weakest channel gains, respectively. By NOMA principle [3], [4], UE n_l in the l-th cluster first decodes the signals of UEs $n'_l > n_l$ with poorer channel conditions, and then its own signal is successively decoded after removing the interference from those UEs. Denote by $\text{SINR}_{n_l}^{n'_l}$ and $\text{SINR}_{n'_l}^{n'_l}$ the signal-to-interference-plus-noise ratios (SINRs) in decoding the signal of UE n'_l by UE n_l and itself, respectively. Towards an efficient and implementable SIC, the following necessary condition is considered [19]

$$\mathbb{E}\left\{\log_2\left(1+\text{SINR}_{n_l}^{n'_l}\right)\right\} \geq \mathbb{E}\left\{\log_2\left(1+\text{SINR}_{n'_l}^{n'_l}\right)\right\}, \tag{13}$$

$\forall n_l < n'_l, \forall l \in \mathcal{L}$.

Remark 2. *We note that perfect SIC is practically unattainable owing to the effects of intra-cluster pilot contamination and channel estimation errors. Consequently, the received signal at UE n_l in the l-th cluster after SIC processing can be written as follows:*

$$\begin{aligned} \bar{y}_{n_l} &= \underbrace{\sum_{m\in\mathcal{M}}\sqrt{\rho_{n_l}^m}\boldsymbol{h}_{m,n_l}^H\boldsymbol{w}_{m,l}x_{n_l}}_{\textit{Desired signal}} + \underbrace{\sum_{m\in\mathcal{M}}\sum_{n'_l=1}^{n_l-1}\sqrt{\rho_{n'_l}^m}\boldsymbol{h}_{m,n_l}^H\boldsymbol{w}_{m,l}x_{n'_l}}_{\textit{Intra-cluster interference after SIC}} \\ &\quad + \underbrace{\sqrt{\zeta_{n_l}}\sum_{m\in\mathcal{M}}\sum_{n''_l=n_l+1}^{N_l}\sqrt{\rho_{n''_l}^m}\boldsymbol{h}_{m,n_l}^H\boldsymbol{w}_{m,l}x_{n''_l}}_{\textit{Intra-cluster interference due to imperfect SIC}} \\ &\quad + \underbrace{\sum_{m\in\mathcal{M}}\sum_{l'\in\mathcal{L}\setminus\{l\}}\sum_{n_{l'}\in\mathcal{N}_{l'}}\sqrt{\rho_{n_{l'}}^m}\boldsymbol{h}_{m,n_l}^H\boldsymbol{w}_{m,l'}x_{n_{l'}}}_{\textit{Inter-cluster interference}} + z_{n_l}, \end{aligned} \tag{14}$$

where ζ_{n_l} is a general SIC performance coefficient at UE n_l in the l-th cluster. In particular, $\zeta_{n_l} = 1$ ($\zeta_{n_l} = 0$) indicates no SIC (perfect SIC), while $0 < \zeta_{n_l} < 1$ means imperfect SIC.

3) Downlink Performance Analysis

Given the UC algorithms that will be introduced in Section III, we first derive the SSE of CFmMIMO-NOMA. From (14), the SINR of UE n_l in the l-th cluster is given by (15) at the top of the next page, where DS $= \mathbb{E}\Big\{\sum\limits_{m\in\mathcal{M}}\sqrt{\rho_{n_l}^m}\mathbf{h}_{m,n_l}^H\boldsymbol{w}_{m,l}\Big\}$, BU $= \Big(\sum\limits_{m\in\mathcal{M}}\sqrt{\rho_{n_l}^m}\mathbf{h}_{m,n_l}^H\boldsymbol{w}_{m,l} -$

$$\text{SINR}_{n_l} = \frac{|\text{DS}|^2}{\mathbb{E}\{|\text{BU}|^2\} + \sum\limits_{n'_l=1}^{n_l-1} \mathbb{E}\{|\text{ICI}|^2\} + \sum\limits_{n''_l=n_l+1}^{N_l} \mathbb{E}\{|\text{RICI}|^2\} + \sum\limits_{l' \in \mathcal{L}\setminus\{l\}} \sum\limits_{n_{l'} \in \mathcal{N}_{l'}} \mathbb{E}\{|\text{UI}|^2\} + 1}, \tag{15}$$

$\mathbb{E}\Big\{\sum\limits_{m\in\mathcal{M}} \sqrt{\rho^m_{n_l}} \mathbf{h}^H_{m,n_l} \boldsymbol{w}_{m,l}\Big\}\Big)$, $\text{ICI} = \sum\limits_{m\in\mathcal{M}} \sqrt{\rho^m_{n'_l}} \mathbf{h}^H_{m,n_l} \boldsymbol{w}_{m,l}$, $\text{RICI} = \sqrt{\zeta_{n_l}} \sum\limits_{m\in\mathcal{M}} \sqrt{\rho^m_{n''_l}} \mathbf{h}^H_{m,n_l} \boldsymbol{w}_{m,l}$, and $\text{UI} = \sum\limits_{m\in\mathcal{M}} \sqrt{\rho^m_{n_{l'}}} \mathbf{h}^H_{m,n_l} \boldsymbol{w}_{m,l'}$ are the coherent beamforming gain (desired signal), beamforming gain uncertainty, intra-cluster interference after SIC, residual interference due to imperfect SIC, and inter-cluster interference, respectively.

To simplify (15), we first compute the expectation term in the denominator of (9) [38]:

$$\mathbb{E}\left\{\left\|\tilde{\mathbf{H}}_m \big(\tilde{\mathbf{H}}^H_m \tilde{\mathbf{H}}_m\big)^{-1} \boldsymbol{\varphi}_l\right\|^2_2\right\} = \frac{\upsilon^2_{m,n_l}}{\gamma_{m,n_l}(K-\tau_p)},\ \forall n_l \in \mathcal{N}_l. \tag{16}$$

From (8), (9), and (16), we have

$$\hat{\mathbf{h}}^H_{m,n_i} \boldsymbol{w}_{m,l} = \frac{\upsilon_{m,n_i}}{\upsilon_{m,n_l}} \boldsymbol{\varphi}^H_i \boldsymbol{\varphi}_l \sqrt{\gamma_{m,n_l}(K-\tau_p)} = \begin{cases} \sqrt{\gamma_{m,n_l}(K-\tau_p)}, & \text{if } i = l, \\ 0, & \text{if } i \neq l. \end{cases} \tag{17}$$

Lemma 1. *The closed-form expression for the SE of UE n_l in the l-th cluster is given by*

$$R_{n_l} = \Big(1 - \frac{\tau_p}{\tau_c}\Big) \log_2\Big(1 + \text{SINR}_{n_l}\Big) = \Big(1 - \frac{\tau_p}{\tau_c}\Big) \log_2\Big(1 + \min_{n'_l=1,\ldots,n_l} \text{SINR}^{n_l}_{n'_l}\Big),\ \forall n_l. \tag{18}$$

With $\boldsymbol{\rho} \triangleq \{\rho^m_{n_l}\}_{m\in\mathcal{M}, n_l\in\mathcal{N}_l, l\in\mathcal{L}}$, $\text{SINR}^{n_l}_{n_l}$ and $\text{SINR}^{n_l}_{n'_l}$, $\forall n'_l < n_l$, are derived as follows:

$$\text{SINR}^{n_l}_{n_l} = \frac{(K-\tau_p)\Big(\sum\limits_{m\in\mathcal{M}} \sqrt{\rho^m_{n_l}\gamma_{m,n_l}}\Big)^2}{\mathcal{I}^{n_l}_{n_l}(\boldsymbol{\rho}) + 1}, \tag{19}$$

$$\text{SINR}^{n_l}_{n'_l} = \frac{(K-\tau_p)\Big(\sum\limits_{m\in\mathcal{M}} \sqrt{\rho^m_{n_l}\gamma_{m,n'_l}}\Big)^2}{\mathcal{I}^{n_l}_{n'_l}(\boldsymbol{\rho}) + 1}, \tag{20}$$

where $\mathcal{I}^{n_l}_{n_l}(\boldsymbol{\rho})$ and $\mathcal{I}^{n_l}_{n'_l}(\boldsymbol{\rho})$ are defined as

$$\mathcal{I}^{n_l}_{n_l}(\boldsymbol{\rho}) \triangleq \sum_{n''_l\in\mathcal{N}_l\setminus\{n_l\}} \eta_{n''_{l'}} (K-\tau_p)\Big(\sum_{m\in\mathcal{M}} \sqrt{\rho^m_{n''_l}\gamma_{m,n_l}}\Big)^2 + \sum_{l'\in\mathcal{L}} \sum_{n''_{l'}\in\mathcal{N}_{l'}} \sum_{m\in\mathcal{M}} \eta_{n''_{l'}} \rho^m_{n''_{l'}} \left(\beta_{m,n_l} - \gamma_{m,n_l}\right),$$

$$\mathcal{I}^{n_l}_{n'_l}(\boldsymbol{\rho}) \triangleq \sum_{n''_l\in\mathcal{N}_l\setminus\{n_l\}} \eta_{n''_{l'}} (K-\tau_p)\Big(\sum_{m\in\mathcal{M}} \sqrt{\rho^m_{n''_l}\gamma_{m,n'_l}}\Big)^2 + \sum_{l'\in\mathcal{L}} \sum_{n''_{l'}\in\mathcal{N}_{l'}} \sum_{m\in\mathcal{M}} \eta_{n''_{l'}} \rho^m_{n''_{l'}} \left(\beta_{m,n'_l} - \gamma_{m,n'_l}\right),$$

with

$$\eta_{n''_{l'}} = \begin{cases} 1, & \text{if } l' \neq l \text{ or } l' = l \text{ and } n''_l \leq n_l, \\ \zeta_{n_l}, & \text{otherwise}. \end{cases}$$

Proof: We follow similar steps as in [17] to derive (19) and (20), by taking into account the residual interference due to imperfect SIC. ■

We define the virtual channel of UE n_l in the l-th cluster as $\mathbf{h}_{n_l} = [\gamma_{1,n_l}, \ldots, \gamma_{M,n_l}]^T$, $\forall n_l \in \mathcal{N}_l$. We assume that UEs in the l-th cluster are sorted based on their virtual channels, such as $\|\mathbf{h}_{1_l}\|_2 \geq \|\mathbf{h}_{2_l}\|_2 \geq \ldots \geq \|\mathbf{h}_{N_l}\|_2$, $\forall l \in \mathcal{L}$. From (18), the SSE of CFmMIMO-NOMA is expressed as

$$R_\Sigma = \sum_{l\in\mathcal{L}} \sum_{n_l\in\mathcal{N}_l} R_{n_l} = \Big(1 - \frac{\tau_p}{\tau_c}\Big) \sum_{l\in\mathcal{L}} \sum_{n_l\in\mathcal{N}_l} \log_2\Big(1 + \text{SINR}_{n_l}\Big). \tag{21}$$

III. Clustering Cell-Free Massive MIMO-NOMA System

In CFmMIMO systems, a large number of APs are deployed in the area, which leads to the heterogeneous locations between UEs and different APs. In this section, we propose two unsupervised ML-based UC algorithms to effectively divide all UEs into separate clusters, which are implemented at the CPU by exploiting the large-scale fading coefficients and considering all the APs. Similar to [19] and [28], large-scale fading coefficients of UEs are assumed to be collected and shared with the CPU before performing the UC algorithm. We note that it is only necessary to estimate the large-scale fading coefficients once every 40 τ_c intervals [15], and thus, conveying these coefficients via the fronthaul links occurs much less frequently than data transmission. Denote by $\boldsymbol{\beta}_n \triangleq [\beta_{1,n}, \beta_{2,n}, \ldots, \beta_{M,n}]^T \in \mathbb{R}^{M\times 1}$ the set of large-scale fading coefficients from all APs associated to UE $n, \forall n \in \mathcal{N}$. The vector $\boldsymbol{\beta}_n$ can be considered as an effective feature-vector denoting the location of UE n.

A. The k-means Algorithm

The k-means algorithm for UC studied in [27] and [28] is one of the simplest unsupervised ML algorithms to partition UEs in the coverage area into separate groups. The key idea is to find a user-specified number of clusters L, which are represented by L centroids, one for each cluster. The number of clusters L in the k-means algorithm is predetermined. The principle of k-means algorithm is given as follows. Firstly, L initial centroids are randomly selected. Secondly, each point is assigned to the nearest centroid, and each mass of points assigned to the same centroid creates a cluster. Then, the centroid of each cluster is updated according to the points associated to the cluster. The assignment and update processes of centroids are repeated until either there is no change in the clusters or centroids remain similarly.

In the context of NOMA systems, it should be noted that the k-means algorithm studied in [27] results in clusters that have unlimited number of UEs, which may not be applicable to NOMA systems because of the increase in SIC computational

complexity and the degradation of the decoding performance. As the number of UEs per group increases, it also becomes more challenging to achieve suitable receive power ratios among NOMA UEs, especially under practical setups where the SIC is imperfect and error propagation can be significant. Hence, we impose a constraint to limit the number of UEs per cluster. In the context of CFmMIMO-NOMA, the procedure of k-means can be summarized as follows:

- Step 1: L initial centroids are randomly selected from N UEs, where L is a predefined number. Let us define the set of L cluster centroids as follows:
$$\mathcal{C} = \{c_l, l \in \mathcal{L}\}, \tag{22}$$
where c_l represents the centroid of the l-th cluster.
- Step 2: Each UE $n \in \mathcal{N}$ is grouped to the nearest centroid, and hence, UEs assigned to the same centroid creates a cluster:
$$l' = \arg\min_{\forall l \in \mathcal{L}} f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_l}\right), \tag{23}$$
where $f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_l}\right) = \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{c_l}\|_2$ represents the Euclidean distance from UE n to centroid c_l [27]. As shown in (23), UE n is grouped to l'-th cluster (denoted by centroid $c_{l'}$) since the distance from UE n to centroid $c_{l'}$ is nearest.
- Step 3: The centroid of each cluster is recalculated under given UEs assigned to this cluster:
$$\boldsymbol{\beta}_{c_l} = \frac{1}{|\mathcal{N}_l|} \sum_{n \in \mathcal{N}_l} \boldsymbol{\beta}_n, \forall l \in \mathcal{L}, \tag{24}$$
where $\boldsymbol{\beta}_{c_l}$ represents the updated centroid for the l-th cluster, which can be calculated by the mean of all UEs belonging to the l-th cluster.
- Step 4: Steps 2-3 are repeated until convergence, i.e., there is no change in the clusters or the centroids remain the same.
- Step 5: If $\exists l'' \in \mathcal{L}$ such that $|\mathcal{N}_{l''}| > \iota$, where ι denotes the maximum number of UEs in each cluster, and with $\mathcal{L}''$ denoting the set of clusters with size exceeding ι, i.e., $\mathcal{L}'' = \{l'', l'' \in \mathcal{L} \text{ with } |\mathcal{N}_{l''}| > \iota\}$, the UEs from the oversized clusters in $\mathcal{L}''$ are pooled as:
$$\mathcal{N}' = \bigcup_{\forall l'' \in \mathcal{L} \text{ with } |\mathcal{N}_{l''}| > \iota} \mathcal{N}_{l''}. \tag{25}$$
Repeat Steps 1-4 to $\mathcal{N}'$ targeting $|\mathcal{L}''| + 1$ clusters. Update the number of clusters $\mathcal{L} \leftarrow \mathcal{L} + 1$.
- Step 6: Step 5 is repeated until $|\mathcal{N}_l| \leq \iota$, $\forall l \in \mathcal{L}$.

Note that Steps 5-6 are performed iteratively to ensure that all clusters are bounded above. The k-means algorithm for UC in CFmMIMO-NOMA is given in Algorithm 1. Note that k-means is a greedy algorithm, which can converge to a local minimum since its performance highly depends on the predefined number of clusters L and the centroid initialization process, i.e., how to select L initial centroids.

Algorithm 1 The k-means Algorithm for UC in CFmMIMO-NOMA

Input: L and $\boldsymbol{\beta}_n, \forall n \in \mathcal{N}$.
*//**Identify L cluster centroids at random c_l, $\forall l \in \mathcal{L}$ (Step 1)**//*
Set $\mathcal{C} = \emptyset$ and $l = 1$, where $\mathcal{C}$ denotes the set of cluster centroids.
while $l \leq L$ **do**
  $c_l = \mathsf{generateRandom}[1, N]$;
  if $c_l \notin \mathcal{C}$ **then**
    $\mathcal{C} \leftarrow c_l$;
    $l = l + 1$;
  end if
end while
*//**Main process (Steps 2-4)**//*
while $\mathcal{C}$ changes **do**
  *//**Identify $\mathcal{N}_{l'}$, $\forall l' \in \mathcal{L}$, containing the subset of UEs that are closer to $c_{l'}$ than c_l, with $l' \neq l$ (Step 2)**//*
  for $n \in \mathcal{N} \backslash \mathcal{C}$ **do**
    $l' = \arg\min_{\forall l \in \mathcal{L}} f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_l}\right)$, where $f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_l}\right) = ||\boldsymbol{\beta}_n - \boldsymbol{\beta}_{c_l}||_2$;
    $\mathcal{N}_{l'} \leftarrow n$;
  end for
  *//**Recalculate c_l of cluster $\mathcal{N}_l$, $\forall l \in \mathcal{L}$ (Step 3)**//*
  for $l = 1 : L$ **do**
    $\boldsymbol{\beta}_{c_l} = \frac{1}{|\mathcal{N}_l|} \sum_{n \in \mathcal{N}_l} \boldsymbol{\beta}_n$;
  end for
end while
*//**Ensure $|\mathcal{N}_l| \leq \iota$, $\forall l \in \mathcal{L}$ (Steps 5-6)**//*
$\mathcal{L}'' = \{l'', l'' \in \mathcal{L} \text{ with } |\mathcal{N}_{l''}| > \iota\}$;
$\mathcal{L} = \mathcal{L} \setminus \mathcal{L}''$;
while $\mathcal{L}'' \neq \emptyset$ **do**
  $\mathcal{N}' = \bigcup_{\forall l'' \in \mathcal{L}''} \mathcal{N}_{l''}$;
  Repeat Steps 2-22 to $\mathcal{N}'$ with $|\mathcal{L}'''| = |\mathcal{L}''| + 1$ clusters, where $\mathcal{L}'''$ denotes the set of $|\mathcal{L}''| + 1$ clusters;
  $\mathcal{L} = \mathcal{L} \cup \{l'', l'' \in \mathcal{L}''' \text{ with } |\mathcal{N}_{l''}| \leq \iota\}$;
  $\mathcal{L}'' = \{l'', l'' \in \mathcal{L}''' \text{ with } |\mathcal{N}_{l''}| > \iota\}$;
end while
Output: $\mathcal{N}_l$ and c_l, $\forall l \in \mathcal{L}$.

B. Proposed k-means++ Algorithm

One drawback of the k-means algorithm is that it is sensitive to the initialization of the centroids [39], [40]. If an initial centroid is a far point, it might not associate with any other points. Equivalently, more than one initial centroids might be created into the same cluster which leads to poor grouping. In this section, the k-means++ algorithm is developed to resolve this issue. It aims at providing a clever initialization of the centroids that improves the quality of the grouping process. Besides, the proposed k-means++ algorithm is able to control the maximum number of UEs per cluster. Except for the improvement in the centroid initialization process, the remainder of k-means++ algorithm is the same as in the k-means. In the context of CFmMIMO-NOMA, the proposed k-means++ can be summarized as follows:

- Step 1: The first initial centroid c_1 is randomly selected from N UEs.
- Step 2: For each UE n (with $n \in \mathcal{N}$ and $n \notin \mathcal{C}$), its distance from the nearest centroid is calculated as

follows:

$$f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_t}\right) = \left\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{c_t}\right\|_2, \tag{26}$$

where $c_t = \underset{\forall c_l \in \mathcal{C}}{\arg\min}\, f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_l}\right)$.

- Step 3: The next centroid is selected from UEs ($\forall n \in \mathcal{N}\backslash\mathcal{C}$) such that the probability of selecting a UE as a centroid is in direct proportion to its distance from the nearest and previously selected centroid, i.e., the UE having the maximum distance from the nearest centroid is virtually to be chosen next as a centroid:

$$c_l = \underset{\forall n \in \mathcal{N}\backslash\mathcal{C}}{\arg\max}\, f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_t}\right). \tag{27}$$

- Step 4: Steps 2-3 are repeated until $L-1$ centroids are selected.
- Step 5: The process continues following Steps 2-6 in the k-means algorithm.

The centroid initialization process of the proposed k-means++ (from step 1 to step 4) ensures that chosen centroids are far away from each other. This increases the opportunity of initially selecting centroids that are located in different clusters. The proposed k-means++ algorithm for UC in CFmMIMO-NOMA is described in Algorithm 2.

C. Proposed Improved k-means++ Algorithm

As shown in Sections III-A and III-B, the performance of the k-means algorithm can be enhanced by selecting the L initial centroids more effectively. Based on the characteristics of CFmMIMO-NOMA, we propose the improved k-means++ algorithm which includes a new approach to cleverly select L initial centroids. Since initial centroids are chosen as UEs that have highest large scale fading coefficients to the largest number of APs, the resulting clusters are served by more APs with better signal quality. The procedure of improved k-means++ is summarized as follows:

- Step 1: Each AP identifies an associated UE, denoted by Λ_m, which has the best connection, i.e., highest large-scale fading coefficient $\beta_{m,n}$:

$$\Lambda_m = \underset{\forall n \in \mathcal{N}}{\arg\max}\, \beta_{m,n}, \forall m \in \mathcal{M}. \tag{28}$$

- Step 2: The CPU then selects a subset of APs, denoted by Υ_n, which have best connections to UE n:

$$\Upsilon_n = \left\{\text{AP}_m : \text{UE } n == \Lambda_m\right\}, \forall n \in \mathcal{N}. \tag{29}$$

- Step 3: The CPU selects a UE having the highest number of serving APs as a centroid:

$$c_l = \underset{\forall n \in \mathcal{N}\backslash\mathcal{C}}{\arg\max}\, |\Upsilon_n|, \tag{30}$$

where $|\Upsilon_n|$ denotes the cardinality of Υ_n.
- Step 4: Step 3 is repeated until L centroids are chosen.
- Step 5: The process continues following Steps 2-6 in the k-means algorithm.

The centroid initialization process of the improved k-means++ for UC in CFmMIMO-NOMA (Steps 1-4 above) is described in Algorithm 3.

Algorithm 2 The k-means++ Algorithm for UC in CFmMIMO-NOMA

Input: L and $\boldsymbol{\beta}_n$, $\forall n \in \mathcal{N}$.
*//**Identify the first cluster centroid c_l (Step 1)**//*
Set $\mathcal{C} = \emptyset$ and $c_1 = \mathsf{generateRandom}[1, N]$;
$\mathcal{C} \leftarrow c_1$ and set $f = 0$;
*//**Identify $L-1$ cluster centroids, c_l, $l = 2, \ldots, L$ (Steps 2-4)**//*
for $l = 2 : L$ **do**
  for $n = 1 : N$ **do**
    for $t = 1 : l-1$ **do**
      if $n \neq c_t$ **then**
        $dis\,(1,t) = f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_t}\right)$, where $f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_t}\right) = \left\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{c_t}\right\|_2$;
      else
        $dis\,(1,t) = \text{NaN}$;
        $f = f + 1$;
      end if
    end for
    if $f == 0$ **then**
      $dist\,(1,n) = \max$ *dis*;
    else
      $dist\,(1,n) = \text{NaN}$;
      $f = 0$;
    end if
  end for
  $c_l = \underset{\forall n \in \mathcal{N}\backslash\mathcal{C}}{\arg\max}$ *dist*;
  $\mathcal{C} \leftarrow c_l$;
end for
*//**Main process (Step 5)**//*
while $\mathcal{C}$ changes **do**
  for $n \in \mathcal{N}\backslash\mathcal{C}$ **do**
    $l' = \underset{\forall l \in \mathcal{L}}{\arg\min}\, f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_l}\right)$, where $f_d\left(\boldsymbol{\beta}_n, \boldsymbol{\beta}_{c_l}\right) = \left\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{c_l}\right\|_2$;
    $\mathcal{N}_{l'} \leftarrow n$;
  end for
  for $l = 1 : L$ **do**
    $\boldsymbol{\beta}_{c_l} = \frac{1}{|\mathcal{N}_l|} \sum_{n \in \mathcal{N}_l} \boldsymbol{\beta}_n$;
  end for
end while
*//**Ensure $|\mathcal{N}_l| \leq \iota$, $\forall l \in \mathcal{L}$ (Step 5)**//*
$\mathcal{L}'' = \{l'', l'' \in \mathcal{L}$ with $|\mathcal{N}_{l''}| > \iota\}$;
$\mathcal{L} = \mathcal{L} \setminus \mathcal{L}''$;
while $\mathcal{L}'' \neq \emptyset$ **do**
  $\mathcal{N}' = \bigcup_{\forall l'' \in \mathcal{L}''} \mathcal{N}_{l''}$;
  Repeat Steps 2-22 to $\mathcal{N}'$ with $|\mathcal{L}'''| = |\mathcal{L}''| + 1$ clusters, where $\mathcal{L}'''$ denotes the set of $|\mathcal{L}''| + 1$ clusters;
  $\mathcal{L} = \mathcal{L} \cup \{l'', l'' \in \mathcal{L}'''$ with $|\mathcal{N}_{l''}| \leq \iota\}$;
  $\mathcal{L}'' = \{l'', l'' \in \mathcal{L}'''$ with $|\mathcal{N}_{l''}| > \iota\}$;
end while
Output: $\mathcal{N}_l$ and c_l, $\forall l \in \mathcal{L}$.

Algorithm 3 Centroid Initialization Process of the Improved k-means++ Algorithm for UC in CFmMIMO-NOMA

Input: L and $\boldsymbol{\beta}_n, \forall n \in \mathcal{N}$.
//**Identify UE that has the best connection to each AP (Step 1)**//
for $m = 1 : M$ **do**
  $\Lambda_m = \arg\max_{\forall n \in \mathcal{N}} \beta_{m,n}$;
end for
//**Identify the subset of APs that have best connections to each UE (Step 2)**//
for $n = 1 : N$ **do**
  for $m = 1 : M$ **do**
    if $n == \Lambda_m$ **then**
      $\Upsilon_n \leftarrow m$;
    end if
  end for
end for
//**Identify L cluster centroids, c_l, $\forall l \in \mathcal{L}$, that have large number of serving APs (Steps 3-4)**//
$\mathcal{C} = \emptyset$, where $\mathcal{C}$ denotes the set of cluster centroids.
for $l = 1 : L$ **do**
  $c_l = \arg\max_{\forall n \in \mathcal{N} \setminus \mathcal{C}} |\Upsilon_n|$;
  $\mathcal{C} \leftarrow c_l$;
end for
Output: $\mathcal{C}$.

IV. The Sum Spectral Efficiency Maximization

From (19) and (20), it is clear that the SSE of CFmMIMO-NOMA highly depends on the power allocation (PA) at all APs. Thus, it is necessary to optimize the transmit power at APs so that the SSE of CFmMIMO-NOMA can be enhanced. In this section, we aim at optimizing the normalized transmit power $\boldsymbol{\rho} \triangleq \{\rho_{n_l}^m\}_{m,n_l,l}$ to maximize the SSE under the constraints of the transmit power budget at the APs, SIC conditions, and minimum required SE at UEs. The optimization problem can be mathematically expressed as

$$\max_{\boldsymbol{\rho}} \quad \Big(1 - \frac{\tau_p}{\tau_c}\Big) \sum_{l \in \mathcal{L}} \sum_{n_l \in \mathcal{N}_l} \log_2\big(1 + \mathrm{SINR}_{n_l}\big) \tag{31a}$$

$$\text{s.t.} \quad \sum_{l \in \mathcal{L}} \sum_{n_l \in \mathcal{N}_l} \rho_{n_l}^m \le P_{\max}^m, \forall m \in \mathcal{M}, \tag{31b}$$

$$\rho_{n_l}^m \le \rho_{n_l+1}^m, n_l \in [1, N_l - 1], \forall m \in \mathcal{M}, l \in \mathcal{L}, \tag{31c}$$

$$\Big(1 - \frac{\tau_p}{\tau_c}\Big) \log_2\Big(1 + \mathrm{SINR}_{n_l}\Big) \ge \bar{R}_{n_l}, \forall n_l. \tag{31d}$$

Herein, constraint (31b) indicates that the total transmit power at AP_m is limited by the normalized maximum power $P_{\max}^m$, constraint (31c) is the necessary condition to implement SIC in the l-th cluster, $\forall l \in \mathcal{L}$, and constraint (31d) denotes the minimum SE requirement $\bar{R}_{n_l}$ of UE n_l, $\forall n_l$. We note that SINR_{n_l} in (31a) is a nonconvex and nonsmooth function with respect to $\boldsymbol{\rho}$, making problem (31) intractable. Therefore, it may not be possible to solve the problem directly. In addition, the globally optimal solution (e.g., exhaustive search) comes at the cost of high computational complexity, and may not be suitable for practical implementation. In what follows, we develop newly approximated functions using the IA framework [35], [41], and then propose a fast converging and low-complexity algorithm.

Equivalent Optimization Problem: To apply the IA method, several transformations are necessary to make (31) tractable. To do so, we introduce the auxiliary variables $\mathbf{r} \triangleq \{r_{n_l}\}_{\forall n_l}$ and $\boldsymbol{\varphi} \triangleq \{\varphi_{n_l}\}_{\forall n_l}$ to rewrite (31) equivalently as

$$\max_{\boldsymbol{\rho},\mathbf{r},\boldsymbol{\varphi}} \quad \Big(1 - \frac{\tau_p}{\tau_c}\Big) \sum_{l \in \mathcal{L}} \sum_{n_l \in \mathcal{N}_l} r_{n_l} \tag{32a}$$

$$\text{s.t.} \quad \ln\left(1 + \varphi_{n_l}\right) \ge r_{n_l} \ln 2, \ \forall n_l \in \mathcal{N}_l, \tag{32b}$$

$$\mathrm{SINR}_{n'_l}^{n_l} \ge \varphi_{n_l}, \ \forall n'_l < n_l, \ \forall n_l \in \mathcal{N}_l, \tag{32c}$$

$$\mathrm{SINR}_{n_l}^{n_l} \ge \varphi_{n_l}, \ \forall n_l \in \mathcal{N}_l, \tag{32d}$$

$$\Big(1 - \frac{\tau_p}{\tau_c}\Big) r_{n_l} \ge \bar{R}_{n_l}, \forall n_l, \tag{32e}$$

$$(31b), (31c). \tag{32f}$$

It is clear that the objective function becomes linear. The equivalence between (31) and (32) is verified by the following lemma.

Lemma 2. *Problems* (31) *and* (32) *share the same optimal solution set and the same optimal objective value. In particular, let* $(\boldsymbol{\rho}^\star, \mathbf{r}^\star, \boldsymbol{\varphi}^\star)$ *be the optimal solution to problem* (32)*, then* $\boldsymbol{\rho}^\star$ *is also the optimal solution to problem* (31) *and vice versa.*

Proof: The proof is done by showing the fact that constraints (32b)-(32d) will hold with equality at the optimum. We prove this statement by contradiction. Suppose that constraints (32c) and (32d) are inactive at the optimum for some users, i.e., there exists $\varphi'_{n_l} > 0$ such as $\min\big(\mathrm{SINR}_{n'_l}^{n_l}, \mathrm{SINR}_{n_l}^{n_l}\big) = \varphi'_{n_l} > \varphi^\star_{n_l}$. It is clear that φ'_{n_l} is also a feasible point to (32b), and $r'_{n_l} = \ln\left(1 + \varphi'_{n_l}\right)/\ln 2 > \ln\left(1 + \varphi^\star_{n_l}\right)/\ln 2 = r^\star_{n_l}$. As a consequence, this results in a strictly larger objective value, i.e., $\big(1 - \frac{\tau_p}{\tau_c}\big) \sum_{l \in \mathcal{L}} \sum_{n_l \in \mathcal{N}_l} r'_{n_l} > \big(1 - \frac{\tau_p}{\tau_c}\big) \sum_{l \in \mathcal{L}} \sum_{n_l \in \mathcal{N}_l} r^\star_{n_l}$, which contradicts the assumption that $(\boldsymbol{\rho}^\star, \mathbf{r}^\star, \boldsymbol{\varphi}^\star)$ represents the optimal solution to problem (32). ■

Inner Approximation (IA) for Problem (32)*:* The nonconvex parts include (32c) and (32d). The direct application of IA method is still not possible due to the complication of $\mathrm{SINR}_{n'_l}^{n_l}$ and $\mathrm{SINR}_{n_l}^{n_l}$. In the following, we make the change of variable as $\rho_{n_l}^m = (\hat{\rho}_{n_l}^m)^2, \forall n_l \in \mathcal{N}_l$. Let us handle (32c) first by rewriting $\mathrm{SINR}_{n'_l}^{n_l}$ as

$$\mathrm{SINR}_{n'_l}^{n_l} = \frac{(K - \tau_p)\big(\sum\limits_{m \in \mathcal{M}} \hat{\rho}_{n_l}^m \sqrt{\gamma_{m,n'_l}}\big)^2}{\mathcal{I}_{n'_l}^{n_l}(\hat{\boldsymbol{\rho}}) + 1}, \tag{33}$$

where $\hat{\boldsymbol{\rho}} \triangleq \{\hat{\rho}_{n_l}^m\}_{\forall n_l}$ and

$$\mathcal{I}_{n'_l}^{n_l}(\hat{\boldsymbol{\rho}}) \triangleq \sum_{n''_l \in \mathcal{N}_l \setminus \{n_l\}} \eta_{n''_l} (K - \tau_p)\big(\sum_{m \in \mathcal{M}} \hat{\rho}_{n''_l}^m \sqrt{\gamma_{m,n'_l}}\big)^2 + \sum_{l' \in \mathcal{L}} \sum_{n''_{l'} \in \mathcal{N}_{l'}} \sum_{m \in \mathcal{M}} \eta_{n''_{l'}} \Big(\hat{\rho}_{n''_{l'}}^m\Big)^2 \big(\beta_{m,n'_l} - \gamma_{m,n'_l}\big). \tag{34}$$

By introducing the slack variables $\boldsymbol{\varpi} \triangleq \{\varpi_{n'_l}^{n_l}\}_{\forall n_l}, \boldsymbol{\tau} \triangleq \{\tau_{n''_l}^{n'_l}\}_{\forall n_l}$, and $\boldsymbol{\theta} \triangleq \{\theta_{n'_l}^{n_l}\}_{\forall n_l}$, constraint (32c) can be equiv-

alently rewritten as (35) at the top of the next page, where

$$\mathcal{I}^{n_l}_{n'_l}(\hat{\boldsymbol{\rho}},\boldsymbol{\tau}) \triangleq \sum_{n''_l\in\mathcal{N}_l\setminus\{n_l\}} \eta_{n''_{l'}}(K-\tau_p)\big(\tau^{n'_l}_{n''_l}\big)^2 + \sum_{l'\in\mathcal{L}}\sum_{n''_{l'}\in\mathcal{N}_{l'}}\sum_{m\in\mathcal{M}} \eta_{n''_{l'}}\big(\hat{\rho}^m_{n''_{l'}}\big)^2\big(\beta_{m,n'_l}-\gamma_{m,n'_l}\big) \quad (36)$$

is a quadratic function. Here, constraint (35d) remains non-convex. We note that $(\varpi^{n_l}_{n'_l})^2/(\theta^{n_l}_{n'_l}+1)$ is the quadratic-over-linear function, which is convex with respect to $(\varpi^{n_l}_{n'_l},\theta^{n_l}_{n'_l})$. Let $(\varpi^{n_l,(\kappa)}_{n'_l},\theta^{n_l,(\kappa)}_{n'_l})$ be a feasible point of $(\varpi^{n_l}_{n'_l},\theta^{n_l}_{n'_l})$ at the κ-th iteration of an iterative algorithm and by the IA principle, constraint (35d) can be convexified as

$$(K-\tau_p)\Big(\frac{2\varpi^{n_l,(\kappa)}_{n'_l}}{\theta^{n_l,(\kappa)}_{n'_l}+1}\varpi^{n_l}_{n'_l} - \frac{\big(\varpi^{n_l,(\kappa)}_{n'_l}\big)^2}{\big(\theta^{n_l,(\kappa)}_{n'_l}+1\big)^2}(\theta^{n_l}_{n'_l}+1)\Big) \geq \varphi_{n_l}, \quad (37)$$

$\forall n'_l < n_l,\ \forall n_l \in \mathcal{N}_l$. Similarly, constraint (32d) can be iteratively approximated as (38) at the top of the next page, where

$$\mathcal{I}^{n_l}_{n_l}(\hat{\boldsymbol{\rho}},\boldsymbol{\tau}) \triangleq \sum_{n''_l\in\mathcal{N}_l\setminus\{n_l\}} \eta_{n''_{l'}}(K-\tau_p)\left(\tau^{n_l}_{n''_l}\right)^2 + \sum_{l'\in\mathcal{L}}\sum_{n''_{l'}\in\mathcal{N}_{l'}}\sum_{m\in\mathcal{M}} \eta_{n''_{l'}}\left(\hat{\rho}^m_{n''_{l'}}\right)^2(\beta_{m,n_l}-\gamma_{m,n_l}).$$

In summary, the convex approximate program of (32) solved at iteration $\kappa+1$ is given as

$$\max_{\hat{\boldsymbol{\rho}},\mathbf{r},\boldsymbol{\varphi},\boldsymbol{\varpi},\boldsymbol{\tau},\boldsymbol{\theta}} \quad \Big(1-\frac{\tau_p}{\tau_c}\Big)\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l} r_{n_l} \quad (39a)$$

$$\text{s.t.} \quad (32b),(32e),(35a)-(35c),(37),(38a)-(38d), \quad (39b)$$

$$\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l}(\hat{\rho}^m_{n_l})^2 \leq P^m_{\max}, \forall m\in\mathcal{M}, \quad (39c)$$

$$\hat{\rho}^m_{n_l} \leq \hat{\rho}^m_{n_l+1}, n_l\in[1,N_l-1], \forall m\in\mathcal{M}, l\in\mathcal{L}. \quad (39d)$$

Conic Quadratic Program: Problem (39) is a mix of exponential and quadratic constraints, resulting in a generic convex program. The major complexity in solving such a program is due to the logarithm function in (32b), making the use of convex solvers (e.g., SeDuMi [42] and MOSEK [43]) inefficient. To bypass this issue, we use a lower bound of $\ln\big(1+\varphi_{n_l}\big)$ as [4, Eq. (66)]

$$\ln\big(1+\varphi_{n_l}\big) \geq \ln(1+\varphi^{(\kappa)}_{n_l}) + \frac{\varphi^{(\kappa)}_{n_l}}{\varphi^{(\kappa)}_{n_l}+1} - \frac{(\varphi^{(\kappa)}_{n_l})^2}{\varphi^{(\kappa)}_{n_l}+1}\frac{1}{\varphi_{n_l}}, \quad (40)$$

$\forall \varphi^{(\kappa)}_{n_l}>0, \varphi_{n_l}>0$, which is a concave function. We note that (40) holds with equality at the optimum, i.e., $\varphi^{(\kappa)}_{n_l}=\varphi^{(\kappa+1)}_{n_l}$. Next, by introducing new variables $\bar{\boldsymbol{\varphi}} \triangleq \{\bar{\varphi}_{n_l}\}_{\forall n_l}$, the conic quadratic approximate program of (39) is given as

$$\max_{\hat{\boldsymbol{\rho}},\mathbf{r},\boldsymbol{\varphi},\bar{\boldsymbol{\varphi}},\boldsymbol{\varpi},\boldsymbol{\tau},\boldsymbol{\theta}} \quad \Big(1-\frac{\tau_p}{\tau_c}\Big)\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l} r_{n_l} \quad (41a)$$

$$\text{s.t.}\ (32e),(35a)-(35c),(37),(38a)-(38d),(39c),(39d), \quad (41b)$$

$$\mathcal{F}^{(\kappa)}(\varphi^{(\kappa)}_{n_l},\bar{\varphi}_{n_l}) \geq r_{n_l}\ln 2,\ \forall n_l\in\mathcal{N}_l, \quad (41c)$$

$$0.25\left(\varphi_{n_l}+\bar{\varphi}_{n_l}\right)^2 \geq 0.25\left(\varphi_{n_l}-\bar{\varphi}_{n_l}\right)^2+1,\ \forall n_l\in\mathcal{N}_l, \quad (41d)$$

where $\mathcal{F}^{(\kappa)}(\varphi^{(\kappa)}_{n_l},\bar{\varphi}_{n_l}) \triangleq \ln(1+\varphi^{(\kappa)}_{n_l}) + \frac{\varphi^{(\kappa)}_{n_l}}{\varphi^{(\kappa)}_{n_l}+1} - \frac{(\varphi^{(\kappa)}_{n_l})^2}{\varphi^{(\kappa)}_{n_l}+1}\bar{\varphi}_{n_l}$. We note that (41d) is a second-order cone constraint and must hold with equality at the optimum. The proposed IA-based iterative algorithm is summarized in Algorithm 4.

Algorithm 4 Proposed IA-based Iterative Algorithm to Solve Problem (31)

Initialization: Set $\kappa:=0$ and generate an initial feasible point $(\boldsymbol{\varpi}^{(0)},\boldsymbol{\theta}^{(0)},\boldsymbol{\varphi}^{(0)})$.

1: **repeat**
2: Solve the conic quadratic approximate program (41) to obtain the optimal solution, denoted by $(\hat{\boldsymbol{\rho}}^\star,\mathbf{r}^\star,\boldsymbol{\varphi}^\star,\bar{\boldsymbol{\varphi}}^\star,\boldsymbol{\varpi}^\star,\boldsymbol{\tau}^\star,\boldsymbol{\theta}^\star)$;
3: Update $(\boldsymbol{\varphi}^{(\kappa+1)},\boldsymbol{\varpi}^{(\kappa+1)},\boldsymbol{\theta}^{(\kappa+1)}) := (\boldsymbol{\varphi}^\star,\boldsymbol{\varpi}^\star,\boldsymbol{\theta}^\star)$;
4: Set $\kappa:=\kappa+1$;
5: **until** Convergence, i.e., $\Big(\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l} r^{(\kappa)}_{n_l} - \sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l} r^{(\kappa-1)}_{n_l}\Big)\Big/\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l} r^{(\kappa-1)}_{n_l} < \epsilon$
6: **Ouput**: $\boldsymbol{\rho}^\star$ with $\rho^{m,(\star)}_{n_l} = (\hat{\rho}^{m,(\star)}_{n_l})^2, \forall n_l\in\mathcal{N}_l$.

Convergence and Complexity Analysis: The proposed algorithm starts by randomly generating an initial feasible point for the updated variables $(\boldsymbol{\varpi}^{(0)},\boldsymbol{\theta}^{(0)},\boldsymbol{\varphi}^{(0)})$. In each iteration, we solve the convex program (41) to produce the next feasible point $(\boldsymbol{\varphi}^{(\kappa+1)},\boldsymbol{\varpi}^{(\kappa+1)},\boldsymbol{\theta}^{(\kappa+1)})$. This procedure is successively repeated until convergence, which is stated in the following proposition.

Proposition 1. *Initialized from a feasible point $(\boldsymbol{\varpi}^{(0)},\boldsymbol{\theta}^{(0)},\boldsymbol{\varphi}^{(0)})$, Algorithm 4 produces a sequence $\{\boldsymbol{\varphi}^{(\kappa)},\boldsymbol{\varpi}^{(\kappa)},\boldsymbol{\theta}^{(\kappa)}\}$ of improved solutions to problem (41), which satisfy the Karush-Kuhn-Tucker (KKT) conditions. In light of the IA principles, the sequence $\Big\{\big(1-\frac{\tau_p}{\tau_c}\big)\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l} r^{(\kappa)}_{n_l}\Big\}_{\kappa=1}^{\infty}$ is monotonically increasing and converges after a finite number of iterations for a given error tolerance $\epsilon>0$.*

Proof: Please see Appendix A. ∎

The computational complexity of Algorithm 4 mainly depends on solving the approximate problem (41), which is polynomial in the number of constraints and optimization variables. Problem (41) has $v = NM+3N+3\sum_{l=1}^{L}\frac{N_l(N_l-1)}{2}$ scalar real variables and $c = 8\sum_{l=1}^{L}\big(\frac{N_l(N_l-1)}{2}+M(N_l-1)\big)+M$ quadratic and linear constraints. As a result, the worst-case computational cost of Algorithm 4 in each iteration is $\mathcal{O}(v^2c^{2.5}+c^{3.5})$ [44].

V. Collocated massive MIMO-NOMA system

In this section, we consider a COmMIMO-NOMA system, which serves as a benchmark for CFmMIMO-NOMA. The main differences between CFmMIMO-NOMA and COmMIMO-NOMA systems are as follows: *i*) in CFmMIMO-NOMA, in general $\beta_{m,n_l} \neq \beta_{m',n_l}$, for $m\neq m'$, whereas in

$$
(32c) \Leftrightarrow \begin{cases} \sum\limits_{m\in\mathcal{M}} \hat{\rho}_{n_l}^m \sqrt{\gamma_{m,n'_l}} \geq \varpi_{n'_l}^{n_l},\ \forall n'_l < n_l, \forall n_l \in \mathcal{N}_l, & (35a) \\ \sum\limits_{m\in\mathcal{M}} \hat{\rho}_{n''_l}^m \sqrt{\gamma_{m,n'_l}} \leq \tau_{n''_l}^{n'_l},\ \forall n'_l < n_l,\ \forall n_l \in \mathcal{N}_l, & (35b) \\ \mathcal{I}_{n'_l}^{n_l}(\hat{\boldsymbol{\rho}}, \boldsymbol{\tau}) \leq \theta_{n'_l}^{n_l},\ \forall n'_l < n_l,\ \forall n_l \in \mathcal{N}_l, & (35c) \\ (K-\tau_p)\dfrac{\left(\varpi_{n'_l}^{n_l}\right)^2}{\theta_{n'_l}^{n_l}+1} \geq \varphi_{n_l},\ \forall n'_l < n_l,\ \forall n_l \in \mathcal{N}_l, & (35d) \end{cases}
$$

$$
(32d) \Leftrightarrow \begin{cases} \sum\limits_{m\in\mathcal{M}} \hat{\rho}_{n_l}^m \sqrt{\gamma_{m,n_l}} \geq \varpi_{n_l}^{n_l},\ \forall n_l \in \mathcal{N}_l, & (38a) \\ \sum\limits_{m\in\mathcal{M}} \hat{\rho}_{n''_l}^m \sqrt{\gamma_{m,n_l}} \leq \tau_{n''_l}^{n_l},\ \forall n_l \in \mathcal{N}_l, & (38b) \\ \mathcal{I}_{n_l}^{n_l}(\hat{\boldsymbol{\rho}}, \boldsymbol{\tau}) \leq \theta_{n_l}^{n_l},\ \forall n_l \in \mathcal{N}_l, & (38c) \\ (K-\tau_p)\Big(\dfrac{2\varpi_{n_l}^{n_l,(\kappa)}}{\theta_{n_l}^{n_l,(\kappa)}+1}\varpi_{n_l}^{n_l} - \dfrac{\left(\varpi_{n_l}^{n_l,(\kappa)}\right)^2}{\left(\theta_{n_l}^{n_l,(\kappa)}+1\right)^2}(\theta_{n_l}^{n_l}+1)\Big) \geq \varphi_{n_l},\ \forall n_l \in \mathcal{N}_l, & (38d) \end{cases}
$$

COmMIMO-NOMA, $\beta_{m,n_l} = \beta_{m',n_l}$; and ii) in CFmMIMO-NOMA, a power constraint is applied at each AP individually, whereas in COmMIMO-NOMA, a total power constraint is applied at the collocated AP equipped with MK antennas. Unless otherwise specified, all notations and symbols given in the previous sections will be reused in this section.

A. Performance Analysis

Similar to Lemma 1, the closed-form expression for the SE of UE n_l in the l-th cluster is given by

$$
\begin{aligned} R_{n_l}^{\mathtt{col}} &= \Big(1-\frac{\tau_p}{\tau_c}\Big)\log_2\big(1+\mathrm{SINR}_{n_l}^{\mathtt{col}}\big) \\ &= \Big(1-\frac{\tau_p}{\tau_c}\Big)\log_2\big(1+\min_{n'_l=1,\ldots,n_l}\mathrm{SINR}_{n'_l}^{n_l,\mathtt{col}}\big),\ \forall n_l \in \mathcal{N}_l. \end{aligned} \tag{42}
$$

By replacing $\rho_{n_l}^m$ with $\rho_{n_l}, \forall n_l$, $\mathrm{SINR}_{n_l}^{n_l,\mathtt{col}}$ and $\mathrm{SINR}_{n'_l}^{n_l,\mathtt{col}}$, $\forall n'_l < n_l$, are derived as follows:

$$
\mathrm{SINR}_{n_l}^{n_l,\mathtt{col}} = \frac{(K-\tau_p)\rho_{n_l}\gamma_{n_l}}{\mathcal{I}_{n_l}^{n_l}(\boldsymbol{\rho})+1}, \tag{43}
$$

$$
\mathrm{SINR}_{n'_l}^{n_l,\mathtt{col}} = \frac{(K-\tau_p)\rho_{n_l}\gamma_{n'_l}}{\mathcal{I}_{n'_l}^{n_l}(\boldsymbol{\rho})+1}, \tag{44}
$$

where

$$
\begin{aligned} \mathcal{I}_{n_l}^{n_l}(\boldsymbol{\rho}) \triangleq & \sum_{n''_l\in\mathcal{N}_l\setminus\{n_l\}} \eta_{n''_{l'}}(K-\tau_p)\rho_{n''_l}\gamma_{n_l} \\ & + \sum_{l'\in\mathcal{L}}\sum_{n''_{l'}\in\mathcal{N}_{l'}} \eta_{n''_{l'}}\rho_{n''_{l'}}\left(\beta_{n_l}-\gamma_{n_l}\right), \end{aligned} \tag{45}
$$

$$
\begin{aligned} \mathcal{I}_{n'_l}^{n_l}(\boldsymbol{\rho}) \triangleq & \sum_{n''_l\in\mathcal{N}_l\setminus\{n_l\}} \eta_{n''_{l'}}(K-\tau_p)\rho_{n''_l}\gamma_{n'_l} \\ & + \sum_{l'\in\mathcal{L}}\sum_{n''_{l'}\in\mathcal{N}_{l'}} \eta_{n''_{l'}}\rho_{n''_{l'}}\left(\beta_{n'_l}-\gamma_{n'_l}\right), \end{aligned} \tag{46}
$$

and $\gamma_{n_l} = \frac{\tau_p\rho_{n_l}\beta_{n_l}^2}{\tau_p\sum_{n'_l\in\mathcal{N}_l}\rho_{n'_l}\beta_{n'_l}+1}$; $\eta_{n''_{l'}}$ is defined as

$$
\eta_{n''_{l'}} = \begin{cases} 1, & \text{if } l' \neq l \text{ or } l' = l \text{ and } n''_l \leq n_l, \\ \zeta_{n_l}, & \text{otherwise.} \end{cases} \tag{47}
$$

The SSE of the COmMIMO-NOMA system is expressed as follows:

$$
R_{\Sigma}^{\mathtt{col}} = \sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l} R_{n_l}^{\mathtt{col}} = \Big(1-\frac{\tau_p}{\tau_c}\Big)\log_2\big(1+\mathrm{SINR}_{n_l}^{\mathtt{col}}\big). \tag{48}
$$

The SSE maximization problem for COmMIMO-NOMA is stated as

$$
\begin{aligned} \max_{\boldsymbol{\rho}} \quad & \Big(1-\frac{\tau_p}{\tau_c}\Big)\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l}\log_2\big(1+\mathrm{SINR}_{n_l}^{\mathtt{col}}\big) && (49a) \\ \mathsf{s.t.} \quad & \sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l}\rho_{n_l} \leq P_{\max}, && (49b) \\ & \rho_{n_l} \leq \rho_{n_l+1}, n_l \in [1, N_l-1], \forall l \in \mathcal{L}, && (49c) \\ & \Big(1-\frac{\tau_p}{\tau_c}\Big)\log_2\big(1+\mathrm{SINR}_{n_l}^{\mathtt{col}}\big) \geq \bar{R}_{n_l}, \forall n_l. && (49d) \end{aligned}
$$

B. Proposed Solution to Problem (49)

By making the change of variable as $\rho_{n_l} = (\hat{\rho}_{n_l})^2, \forall n_l \in \mathcal{N}_l$ and following similar steps from (32) to (39), problem (49) is equivalently transformed to the following tractable form

$$
\begin{aligned} \max_{\hat{\boldsymbol{\rho}},\mathbf{r},\boldsymbol{\varphi},\boldsymbol{\theta}} \quad & \Big(1-\frac{\tau_p}{\tau_c}\Big)\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l} r_{n_l} && (50a) \\ \mathsf{s.t.} \quad & \ln\left(1+\varphi_{n_l}\right) \geq r_{n_l}\ln 2,\ \forall n_l \in \mathcal{N}_l, && (50b) \\ & \mathcal{I}_{n'_l}^{n_l}(\hat{\boldsymbol{\rho}}) \leq \theta_{n'_l}^{n_l},\ \forall n'_l < n_l,\ \forall n_l \in \mathcal{N}_l, && (50c) \\ & \mathcal{I}_{n_l}^{n_l}(\hat{\boldsymbol{\rho}}) \leq \theta_{n_l}^{n_l},\ \forall n_l \in \mathcal{N}_l, && (50d) \\ & \frac{(K-\tau_p)(\hat{\rho}_{n_l})^2\gamma_{n'_l}}{\theta_{n'_l}^{n_l}+1} \geq \varphi_{n_l},\ \forall n'_l < n_l,\ \forall n_l \in \mathcal{N}_l, && (50e) \\ & \frac{(K-\tau_p)(\hat{\rho}_{n_l})^2\gamma_{n_l}}{\theta_{n_l}^{n_l}+1} \geq \varphi_{n_l},\ \forall n_l \in \mathcal{N}_l, && (50f) \end{aligned}
$$

TABLE I
SIMULATION PARAMETERS.

Parameter	Value
System bandwidth (B)	20 MHz
Number of APs (M)	32
Number of UEs (N)	20
Number of antennas per AP (K)	16
Total power budget for all APs	40 dBm
Power budget at UEs	23 dBm
Noise power at receivers	-104 dBm
SIC performance coefficient at UEs	0.05
Maximum number of UEs in each cluster (ι)	2
Minimum SE requirement of UE n_l ($\bar{R}_{n_l}$)	0.5 bps/Hz

$$\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l}(\hat{\rho}_{n_l})^2 \le P_{\max}, \tag{50g}$$

$$\hat{\rho}_{n_l} \le \hat{\rho}_{n_l+1}, n_l \in [1, N_l-1], \forall l \in \mathcal{L}, \tag{50h}$$

$$\left(1-\frac{\tau_p}{\tau_c}\right) r_{n_l} \ge \bar{R}_{n_l}, \forall n_l, \tag{50i}$$

where

$$\mathcal{I}^{n_l}_{n'_l}(\hat{\boldsymbol{\rho}}) \triangleq \sum_{n''_l\in\mathcal{N}_l\setminus\{n_l\}} \eta_{n''_{l'}}(K-\tau_p)(\hat{\rho}_{n''_l})^2\gamma_{n'_l} + \sum_{l'\in\mathcal{L}}\sum_{n''_{l'}\in\mathcal{N}_{l'}} \eta_{n''_{l'}}(\hat{\rho}_{n''_{l'}})^2(\beta_{n'_l}-\gamma_{n'_l}), \tag{51}$$

$$\mathcal{I}^{n_l}_{n_l}(\hat{\boldsymbol{\rho}}) \triangleq \sum_{n''_l\in\mathcal{N}_l\setminus\{n_l\}} \eta_{n''_{l'}}(K-\tau_p)(\hat{\rho}_{n''_l})^2\gamma_{n_l} + \sum_{l'\in\mathcal{L}}\sum_{n''_{l'}\in\mathcal{N}_{l'}} \eta_{n''_{l'}}(\hat{\rho}_{n''_{l'}})^2(\beta_{n_l}-\gamma_{n_l}). \tag{52}$$

The nonconvex constraints are (50e) and (50f). Let $(\hat{\rho}^{(\kappa)}_{n_l}, \theta^{n_l,(\kappa)}_{n_l})$ be a feasible point of $(\hat{\rho}_{n_l}, \theta^{n_l}_{n_l})$ at iteration κ. By (40), the conic quadratic approximate program for solving (50) is given as

$$\max_{\hat{\boldsymbol{\rho}},\mathbf{r},\boldsymbol{\varphi},\bar{\boldsymbol{\varphi}},\boldsymbol{\theta}} \quad \left(1-\frac{\tau_p}{\tau_c}\right)\sum_{l\in\mathcal{L}}\sum_{n_l\in\mathcal{N}_l} r_{n_l} \tag{53a}$$

$$\text{s.t.} \quad (41c), (41d), (50c), (50d), (50g)-(50i), \tag{53b}$$

$$(K-\tau_p)\gamma_{n'_l}\mathcal{G}^{(\kappa)}(\hat{\rho}_{n_l}, \theta^{n_l}_{n'_l}) \ge \varphi_{n_l}, \quad \forall n'_l < n_l, \ \forall n_l \in \mathcal{N}_l, \tag{53c}$$

$$(K-\tau_p)\gamma_{n_l}\mathcal{G}^{(\kappa)}(\hat{\rho}_{n_l}, \theta^{n_l}_{n_l}) \ge \varphi_{n_l}, \ \forall n_l \in \mathcal{N}_l, \tag{53d}$$

where $\mathcal{G}^{(\kappa)}(\hat{\rho}_{n_l}, \theta^{n_l}_{n'_l}) \triangleq \frac{2\hat{\rho}^{(\kappa)}_{n_l}}{\theta^{n_l,(\kappa)}_{n'_l}+1}\hat{\rho}_{n_l} - \frac{\left(\hat{\rho}^{(\kappa)}_{n_l}\right)^2}{\left(\theta^{n_l,(\kappa)}_{n'_l}+1\right)^2}(\theta^{n_l}_{n'_l}+1)$ and $\mathcal{G}^{(\kappa)}(\hat{\rho}_{n_l}, \theta^{n_l}_{n_l}) \triangleq \frac{2\hat{\rho}^{(\kappa)}_{n_l}}{\theta^{n_l,(\kappa)}_{n_l}+1}\hat{\rho}_{n_l} - \frac{\left(\hat{\rho}^{(\kappa)}_{n_l}\right)^2}{\left(\theta^{n_l,(\kappa)}_{n_l}+1\right)^2}(\theta^{n_l}_{n_l}+1)$. The solution to problem (49) can be found by using Algorithm 4, in which we replace problem (41) by problem (53) in Step 2. The worst-case computational complexity of solving (53) in each iteration is $\mathcal{O}(\bar{v}^2\bar{c}^{2.5}+\bar{c}^{3.5})$ [44], where $\bar{v} = 4N + \sum_{l=1}^{L}\frac{N_l(N_l-1)}{2}$ and $\bar{c} = \sum_{l=1}^{L}\left(N_l(N_l-1)+\frac{(N_l-1)^2}{2}\right)+2N+1$ are scalar real variables and constraints, respectively.

VI. NUMERICAL RESULTS

We now quantitatively assess the performance of the proposed unsupervised ML-based UC algorithms in CFmMIMO-NOMA system.

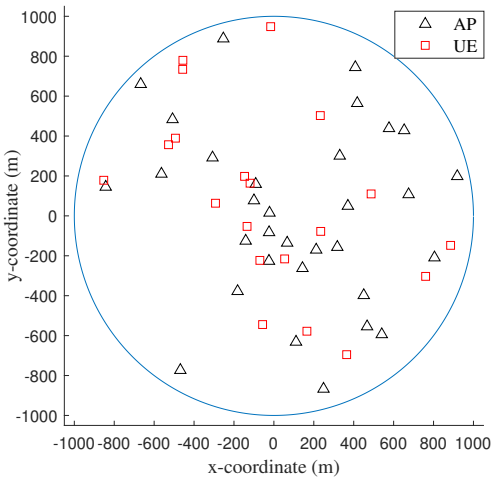

Fig. 2. A system topology with $M = 32$ APs and $N = 20$ UEs is used in numerical examples.

A. Simulation Parameters

A CFmMIMO-NOMA system including $M = 32$ APs and $N = 20$ UEs is considered as shown in Fig. 2, where all APs and UEs are uniformly distributed within a circular region with a radius of 1 km. The large-scale fading coefficient of all channels is modeled as [9] $\beta_{m,n_l} = 10^{\frac{\text{PL}(d_{m,n_l})+\sigma_{sh}z}{10}}$, $\forall m \in \mathcal{M}$, $n_l \in \mathcal{N}_l$, where d_{m,n_l} is the distance from AP$_m$ to UE n_l. The shadow fading is modeled as an RV z, which follows $\mathcal{CN}(0,1)$ with standard deviation σ_{sh} = 8 dB. The path loss $\text{PL}(d_{m,n_l})$ is calculated based on the three-slope path loss model in [9], [37], [45]. Unless otherwise stated, other key parameters are shown in Table I, where all APs are assumed to have the same power budget [9], [37]. The used convex solver is SeDuMi [42] in the MATLAB environment.

B. Selection of the Initial Number of Clusters L

The performance of the k-means based UC algorithms is highly affected by the initial value of number of clusters L [27], [28]. Thus, it is essential to investigate the particular feature of the UEs' distribution in CFmMIMO-NOMA system to choose a proper number of clusters, such that the SSE is maximized. A reliable and precise approach to validate the initial optimal number of clusters L is the silhouette score [46], which is the mean silhouette coefficient of all UEs. The silhouette coefficient of an UE is calculated as $\frac{c-b}{\max(c,b)}$, where b denotes the mean distance to other UEs in the same cluster (so-called the mean intra-cluster distance), and c represents the mean distance to UEs of the next closest cluster which is the one that minimizes b, excluding the UE's own cluster (so-called mean nearest-cluster distance). The value of the silhouette coefficient ranges from -1 to +1. A coefficient close to +1 means that the UE is well matched to its own cluster and far from other clusters. A coefficient close to 0 indicates that the UE is near a cluster boundary, whereas a coefficient close to -1 implies that the UE is assigned to the wrong cluster. Table II at the beginning of the next page

TABLE II
SILHOUETTE SCORE FOR CFMMIMO-NOMA AND COMMIMO-NOMA.

Number of clusters L		2	3	4	5	6	7	8	9	10	11	12	13	14	15
Silhouette Score	CFmMIMO-NOMA	0.72	0.15	0.23	0.31	0.35	0.37	0.63	0.78	**0.99**	0.25	0.40	0.47	0.53	0.64
	COmMIMO-NOMA	0.75	0.06	0.17	0.30	0.39	0.47	0.64	0.85	**0.98**	0.30	0.38	0.50	0.56	0.58

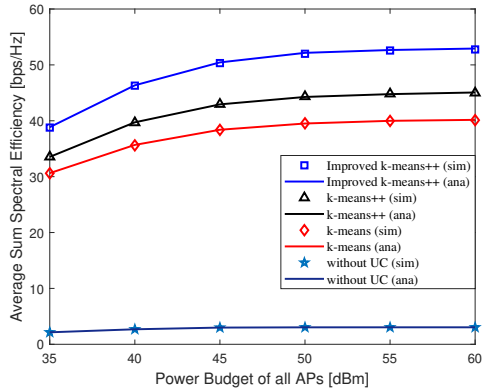

Fig. 3. The SSE of CFmMIMO-NOMA versus the total power budget of all APs for the k-means, k-means++, and improved k-means++ algorithms.

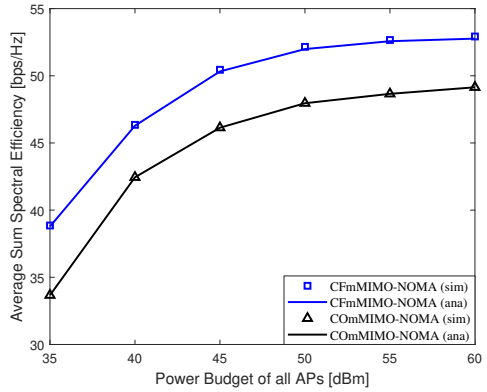

Fig. 4. The SSE of CFmMIMO-NOMA and COmMIMO-NOMA versus the total power budget of all APs.

shows the silhouette score versus the number of clusters L. It is observed that the inital optimal number of clusters for this setting is $L^\star = 10$. Note that this is the initial value of the number of clusters to execute the modified k-means and k-means based UC algorithms, and not the total number of clusters obtained after implementing the corresponding algorithms.

In what follows, we set $L = 10$ to verify the performance analysis in Section VI-C and to evaluate the performance of the proposed algorithms in Section VI-D.

C. Numerical Results for the Performance Analysis

We now investigate the performance of the two proposed unsupervised ML-based UC algorithms with fixed PA. The transmit power at each AP allocated to a specific UE follows the fixed PA scheme. Each AP allocates equal power to each cluster, and then, the fractional transmit PA [47] is used to allocate the power to a specific UE in each cluster based on the virtual channel gains presented in subsection II-B3. As a benchmark, we also consider the COmMIMO-NOMA system, which is presented in Section V.

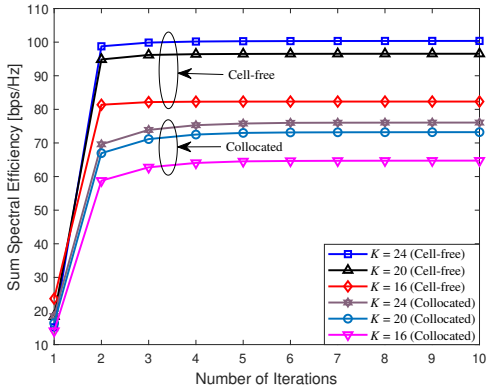

Fig. 5. Convergence behavior of Algorithm 4 with different number of AP antennas, K.

Fig. 3 illustrates the SSE performance of CFmMIMO-NOMA versus the total power budget of all APs for different UC algorithms. For comparison, the performance of the k-means algorithm and the CFmMIMO-NOMA without UC is also plotted. For the CFmMIMO-NOMA without UC, SIC is implemented at all UEs. It can be seen that the proposed UC algorithms significantly outperform the conventional k-means algorithm and without UC. This confirms the effectiveness of UC in CFmMIMO-NOMA systems. Furthermore, the improved k-means++ achieves the best SSE among all algorithms, which can be attributed to the fact that the effective initialization of centroids is capable of improving the quality of the clustering process, and thus, of NOMA for CFmMIMO.

Next, the SSE performance of the CFmMIMO-NOMA and COmMIMO-NOMA systems using the improved k-means++ algorithm versus the total power budget of all APs is shown in Fig. 4. We can observe that the performance of the CFmMIMO-NOMA system is better than that of COmMIMO-NOMA. This is attributed to the fact that CFmMIMO with many distributed APs brings the service antennas closer to UEs which not only reduces path losses but also provides higher degree of macro-diversity, compared to COmMIMO. In the following numerical results, unless otherwise specified, the improved k-means++ algorithm is used for UC.

D. Numerical Results for Optimal Power Allocation (Algorithm 4)

In Fig. 5, we evaluate the convergence speed of Algorithm 4 for CFmMIMO-NOMA and COmMIMO-NOMA with different values of K. The proposed algorithm converges within three iterations and the convergence speed of both systems is not sensitive to the number of AP antennas, K. As expected, the SSE is monotonically increasing after each iteration. Compared to the results in Figs. 3 and 4 with fixed PA at the power budget of 40 dBm, Algorithm 4 yields a significantly better performance in terms of SSE.

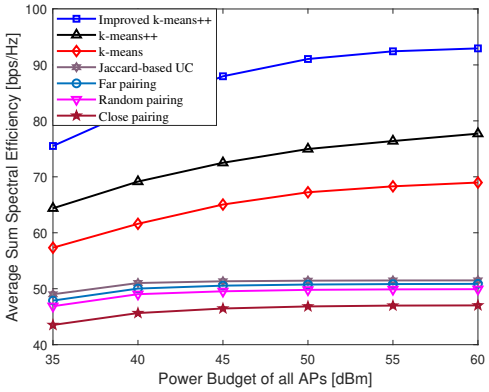

Fig. 6. The SSE of different UC algorithms.

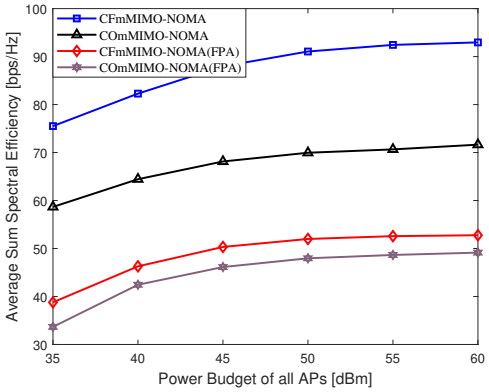

Fig. 7. SSE of CFmMIMO-NOMA and COmMIMO-NOMA: with and without PA.

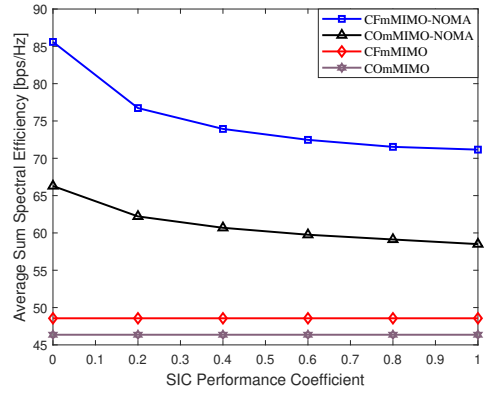

Fig. 8. The effect of SIC performance coefficient on the SSE of CFmMIMO-NOMA and COmMIMO-NOMA systems.

Fig. 6 shows the impact of the proposed k-means++ and improved k-means++ algorithms on the system performance of CFmMIMO-NOMA. For comparison, we also plot the SSE of the k-means (i.e., Algorithm 1) and the recently proposed UC approaches, including near pairing, far pairing, random pairing [19], and the Jaccard-based UC [20]. The main result observed from the figure is that the proposed unsupervised ML-based UC algorithms achieve better SSE performance compared to the baseline ones, and the performance gaps are wider when $P_{\max}$ increases. This implies that the two proposed UC schemes are capable of exploiting UC more effectively, so that the SSE is remarkably enhanced.

In Fig. 7, we demonstrate the benefit of optimizing PA for

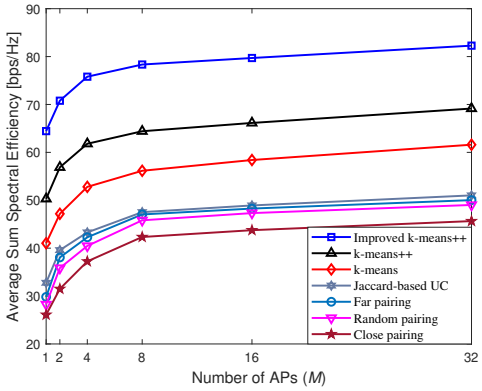

Fig. 9. The joint effect of the numbers of antennas K and APs M on the average SSE of different UC algorithms, for $MK = 512$.

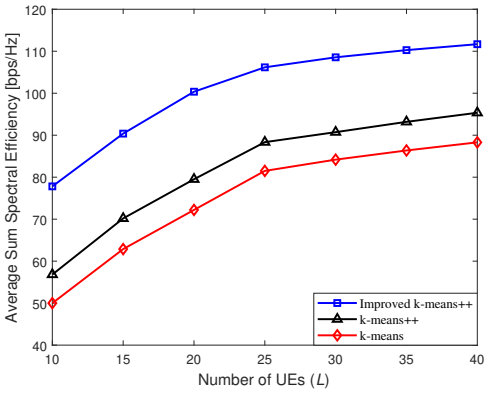

Fig. 10. Effect of the number of UEs L on the SSE for the k-means, k-means++, and improved k-means++ algorithms.

CFmMIMO-NOMA and COmMIMO-NOMA systems. The SSE of both systems is significantly enhanced with optimal PA compared to the fixed PA (FPA) scheme. Hence, this shows the necessity of optimizing PA for both systems, especially for CFmMIMO-NOMA.

Next, the effect of the SIC performance coefficient ζ_{n_l} on the SSE of CFmMIMO-NOMA and COmMIMO-NOMA is examined in Fig. 8. We note that $\zeta_{n_l} = 1$ ($\zeta_{n_l} = 0$) indicates no SIC (perfect SIC), while $0 < \zeta_{n_l} < 1$ means imperfect SIC. The system performance without NOMA/SIC is plotted. It is clear that the SSE of CFmMIMO-NOMA degrades when $\zeta_{n_l}, \forall n_l$ increases. It implies that the SIC performance coefficient is required to be small enough to exploit the full potential of NOMA in CFmMIMO. Nevertheless, the SSE achieved by CFmMIMO-NOMA and COmMIMO-NOMA systems is much higher than their counterparts without NOMA/SIC.

In Fig. 9, we show the joint effect of the numbers of antennas K and APs M on the average SSE of different UC algorithms. We fix $MK = 512$ and select M from the set $M \in [1, 2, 4, 8, 16, 32]$. When $M = 1$, then $K = 512$, which represents COmMIMO-NOMA. From the figure, we see that the SSE increases with the increase in M, which translates into a lower number of AP antennas, K. As such, this not only reduces path losses, but also increases the degree of macro-diversity.

Lastly, the impact of the number of UEs on the SSE of the proposed k-means++ and improved k-means++ algorithms in

CFmMIMO-NOMA system is illustrated in Fig. 10. It can be observed that the SSE significantly increases with the number of UEs.

VII. Conclusion

In this paper, we have investigated a downlink CFmMIMO-NOMA system, for which two efficient unsupervised ML-based UC algorithms have been proposed to effectively cluster the users. Using the fpZF precoding at APs, we have considered the problem of power allocation to maximize SSE. Since the formulated problem is intractable, we have developed a low-complexity iterative algorithm based on the IA framework for its solution. Numerical results have confirmed the effectiveness of the proposed UC algorithms, and show their superior performance compared to the baseline schemes. The proposed PA algorithm converges fast, and significantly outperforms CFmMIMO-NOMA without optimizing PA and COmMIMO-NOMA in terms of SSE.

Appendix A
Proof of Proposition 1

By contradiction and IA principles, we can easily prove that constraints (35a)-(35c), (37), (38a)-(38d) and (41d) must hold with equality at optimum. Let us define $\mathcal{F}(\varphi_{n_l}) \triangleq \ln(1+\varphi_{n_l})$. From (40), we have

$$\mathcal{F}(\varphi_{n_l}) \geq \mathcal{F}^{(\kappa)}(\varphi^{(\kappa)}, \bar{\varphi}_{n_l}), \tag{54}$$

and

$$\mathcal{F}(\varphi_{n_l}^{(\kappa)}) = \mathcal{F}^{(\kappa)}(\varphi^{(\kappa)}, \bar{\varphi}_{n_l}). \tag{55}$$

Thus, it is true that

$$\begin{aligned}\mathcal{F}(\varphi_{n_l}^{(\kappa)}) &\geq \mathcal{F}^{(\kappa-1)}(\varphi^{(\kappa)}, \bar{\varphi}_{n_l}) \\ &\geq \mathcal{F}^{(\kappa-1)}(\varphi^{(\kappa-1)}, \bar{\varphi}_{n_l}) = \mathcal{F}(\varphi_{n_l}^{(\kappa-1)}).\end{aligned} \tag{56}$$

These results imply that $(\boldsymbol{\varpi}^{(\kappa)}, \boldsymbol{\theta}^{(\kappa)}, \boldsymbol{\varphi}^{(\kappa)})$ is an improved solution to problem (41), compared to $(\boldsymbol{\varpi}^{(\kappa-1)}, \boldsymbol{\theta}^{(\kappa-1)}, \boldsymbol{\varphi}^{(\kappa-1)})$. By [35, Theorem 1], the sequence $\{\boldsymbol{\varpi}^{(\kappa)}, \boldsymbol{\theta}^{(\kappa)}, \boldsymbol{\varphi}^{(\kappa)}\}$ converges to at least local optima which satisfy the KKT conditions. As a result, the objective value of problem (41) is monotonically increasing, i.e., $\left(1 - \frac{\tau_p}{\tau_c}\right) \sum_{l \in \mathcal{L}} \sum_{n_l \in \mathcal{N}_l} r_{n_l}^{(\kappa)} \geq \left(1 - \frac{\tau_p}{\tau_c}\right) \sum_{l \in \mathcal{L}} \sum_{n_l \in \mathcal{N}_l} r_{n_l}^{(\kappa-1)}$. In addition, the sequence of the objective values is upper bounded due to power constraints (39c), which completes the proof.

References

[1] Iot-analytics.com, *State of the IoT 2020: 12 billion IoT Connections, Surpassing non-IoT for the First Time*, Nov. 2020. [Online]. Available: https://iot-analytics.com/state-of-the-iot-2020-12-billion-iot-connections-surpassing-non-iot-for-the-first-time

[2] *Ericsson Mobility Report*, Nov. 2020. [Online]. Available: https://ericsson.com/4adc87/assets/local/mobility-report/documents/2020/november-2020-ericsson-mobility-report.pdf

[3] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart. 2017.

[4] V.-D. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and O.-S. Shin, "Precoder design for signal superposition in MIMO-NOMA multicell networks," *IEEE J. Select. Areas Commun.*, vol. 35, no. 12, pp. 2681–2695, Dec. 2017.

[5] H. V. Nguyen, V.-D. Nguyen, O. A. Dobre, D. N. Nguyen, E. Dutkiewicz, and O.-S. Shin, "Joint power control and user association for NOMA-based full-duplex systems," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 8037–8055, Nov. 2019.

[6] Z. Shi, W. Gao, S. Zhang, J. Liu, and N. Kato, "AI-enhanced cooperative spectrum sensing for non-orthogonal multiple access," *IEEE Wirel. Commun.*, vol. 27, no. 2, pp. 173–179, Apr. 2020.

[7] Z. Shi, W. Gao, S. Zhang, J. Liu, and N. Kato, "Machine learning-enabled cooperative spectrum sensing for non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 5692–5702, Sep. 2020.

[8] T. K. Nguyen, H. H. Nguyen, and H. D. Tuan, "Max-min QoS power control in generalized cell-free massive MIMO-NOMA with optimal backhaul combining," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 10 949–10 964, Oct. 2020.

[9] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[10] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, July 2017.

[11] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, "On the uplink max–min SINR of cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2021–2036, Apr. 2019.

[12] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[13] Z. Chen and E. Björnson, "Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5205–5219, Nov. 2018.

[14] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Select. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[15] Y. Li and G. A. A. Baduge, "NOMA-aided cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 950–953, Dec. 2018.

[16] Y. Zhang, H. Cao, M. Zhou, and L. Yang, "Spectral efficiency maximization for uplink cell-free massive MIMO-NOMA networks," in *Proc. IEEE Inter. Conf. Commun. Works.*, May 2019, pp. 1–6.

[17] F. Rezaei, C. Tellambura, A. A. Tadaion, and A. R. Heidarpour, "Rate analysis of cell-free massive MIMO-NOMA with three linear precoders," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3480–3494, June 2020.

[18] S. M. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[19] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "On the performance of cell-free massive MIMO relying on adaptive NOMA/OMA mode-switching," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 792–810, Feb. 2020.

[20] F. Rezaei, A. R. Heidarpour, C. Tellambura, and A. A. Tadaion, "Underlaid spectrum sharing for cell-free massive MIMO-NOMA," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 907–911, Apr. 2020.

[21] R. He, Q. Li, B. Ai, Y. L.-A. Geng, A. F. Molisch, V. Kristem, Z. Zhong, and J. Yu, "A kernel-power-density-based algorithm for channel multipath components clustering," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7138–7151, Nov. 2017.

[22] X. Xie, Z. Zhang, H. Jiang, J. Dang, and L. Wu, "Cluster-based geometrical dynamic stochastic model for MIMO scattering channels," in *Proc. Inter. Conf. Wireless Commun. and Signal Process. (WCSP)*, Oct. 2017, pp. 1–5.

[23] Y. Wang, A. Liu, X. Xia, and K. Xu, "Exploiting the clustered sparsity for channel estimation in hybrid analog-digital massive MIMO systems," *IEEE Access*, vol. 7, pp. 4989–5000, Dec. 2018.

[24] A. B. Rozario and M. F. Hossain, "An architecture for M2M communications over cellular networks using clustering and hybrid TDMA-

NOMA," in *Proc. Inter. Conf. Infor. and Commun. Tech. (ICoICT)*, May 2018, pp. 18–23.

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

[26] E. Cabrera and R. Vesilo, "An enhanced k-means clustering algorithm with non-orthogonal multiple access (NOMA) for MMC networks," in *Proc. Inter. Telecommun Net. and App. Conf. (ITNAC)*, Nov. 2018, pp. 1–8.

[27] 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, Sep. 2018.

[28] F. Riera-Palou, G. Femenias, A. G. Armada, and A. Pérez-Neira, "Clustered cell-free massive MIMO," in *Proc. IEEE Globecom Workshops*, Dec. 2018, pp. 1–6.

[29] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, Mar. 2009.

[30] C. Whelan, G. Harrell, and J. Wang, "Understanding the k-medians problem," in *Proc. Int. Conf. Sci. Comput.*, 2015, pp. 219–222.

[31] F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the k-modes clustering algorithm," *Knowl. Based Syst.*, vol. 26, pp. 120–127, Feb. 2012.

[32] H. Liu, J. Wu, T. Liu, D. Tao, and Y. Fu, "Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1129–1143, May 2017.

[33] S. Yu, L. Tranchevent, X. Liu, W. Glanzel, J. A. K. Suykens, B. D. Moor, and Y. Moreau, "Optimized data fusion for kernel k-means clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1031–1039, May 2012.

[34] M. Morales-Céspedes, O. A. Dobre, and A. García-Armada, "Semi-blind interference aligned NOMA for downlink MU-MISO systems," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1852–1865, Mar. 2020.

[35] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Oper. Res.*, vol. 26, no. 4, pp. 681–683, July-Aug. 1978.

[47] A. Benjebbour, A. Li, Y. Kishiyama, H. Jiang, and T. Nakamura, "System-level performance of downlink NOMA combined with SUM-

[36] G. Interdonato, M. Karlsson, E. Bjornson, and E. G. Larsson, "Downlink spectral efficiency of cell-free massive MIMO with full-pilot zero-forcing," in *Proc. IEEE Global Conf. Signal and Infor. Process. (GlobalSIP)*, Nov. 2018, pp. 1003–1007.

[37] H. V. Nguyen, V.-D. Nguyen, O. A. Dobre, S. K. Sharma, S. Chatzinotas, B. Ottersten, and O.-S. Shin, "On the spectral and energy efficiencies of full-duplex cell-free massive MIMO," *IEEE J. Select. Areas Commun.*, vol. 38, no. 8, pp. 1698–1718, Aug. 2020.

[38] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Commun. Inf. Theory*, vol. 1, no. 1, pp. 1–182, June 2004.

[39] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. Symp. Discrete Algorithms*, Jan. 2007, pp. 1027–1035.

[40] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?" *Pattern Recognit.*, vol. 93, pp. 95–112, Sep. 2019.

[41] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, May 2010.

[42] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimiz. Methods and Softw.*, vol. 11-12, pp. 625–653, Sep. 1999.

[43] "I. MOSEK aps," 2014. [Online]. Available: http://www.mosek.com

[44] D. Peaucelle, D. Henrion, and Y. Labit, "Users guide for SeDuMi interface 1.04," 2002. [Online]. Available: http://homepages.laas.fr/peaucell/software/sdmguide.pdf

[45] A. Tang, J. Sun, and K. Gong, "Mobile propagation loss with a low base station antenna for NLOS street microcells in urban area," in *Proc. IEEE Veh. Tech. Conf. (VTC Spring)*, May 2001, pp. 333–336.

[46] A. Geron, *Hands-on Machine Learning with Scikit-learn and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media Inc., Sep. 2019.

IMO for future LTE enhancements," in *Proc. IEEE Globecom Workshops*, Dec. 2014, pp. 706–710.