IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 19, 2022, accepted March 21, 2022, date of publication April 7, 2022, date of current version April 15, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3165756

User Clustering and Resource Allocation in Hybrid NOMA-OMA Systems Under Nakagami-*m* Fading

ALI MAHMOUDI[1], BAHMAN ABOLHASSANI[1], S. MOHAMMAD RAZAVIZADEH[1], (Senior Member, IEEE), AND HA H. NGUYEN[2], (Senior Member, IEEE)

[1]School of Electrical Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran
[2]Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada

Corresponding author: S. Mohammad Razavizadeh (smrazavi@iust.ac.ir)

ABSTRACT In this paper, we tackle the problem of optimizing user clustering, power, and resource (time slot or bandwidth) allocation in the downlink of a hybrid non-orthogonal multiple access (NOMA)-orthogonal multiple access (OMA) system. In such a system, users are organized into several clusters under one of the following scenarios: (1) fixed cluster size, (2) fixed number of clusters, and (3) variable number of clusters and variable cluster size. A power domain NOMA (PD-NOMA) scheme is used in each cluster, while OMA is employed for allocating resources to different clusters. The goal is to maximize the minimum success probability (which is equivalent to minimizing the maximum outage probability) among all users to guarantee fairness. We prove that at the optimal solution, all users have the same success probability, which is called the common success probability (CSP). Then, we propose an efficient algorithm for finding the optimal CSP and cluster resource allocation factors simultaneously. The optimal power allocation factors and the optimal decoding order of users in each cluster are then derived in closed-form expressions based on the obtained optimal CSP. Simulation results show considerable performance gains by the proposed scheme, compared to existing schemes in terms of fairness, the minimum success probability of users, and the sum throughput.

INDEX TERMS Hybrid NOMA-OMA, user clustering, power allocation, resource allocation, fairness.

I. INTRODUCTION

Substantial growths in the number of users and emerging high data-rate applications with strict quality-of-service (QoS) requirements pose new challenges for the design/plan of future generations of cellular networks. It has been widely acknowledged that it is imperative to employ more efficient multiple access schemes and improve their performance to cope with such demands. Over the last few years, non-orthogonal multiple access (NOMA) has received a lot of attentions and regarded as a promising multiple access scheme due to its ability to serve multiple users in the same time/frequency resource block. In particular power-domain NOMA (PD-NOMA) is considered in various standardization activities since it can improve spectral efficiency, fairness and throughput of cell-edge users [1], [2]. In PD-NOMA, the base station (BS) combines the users' signals by superposition coding at the transmission side, whereas each user detects its own signal by successive interference cancellation (SIC). However, as the complexity and latency of the SIC method increase with the number of users [3], it is impractical when there is a large number of users in the network. To overcome this issue, it is possible to organize the users into several clusters and deploy orthogonal multiple access (OMA) techniques alongside NOMA.

In fact, the hybrid NOMA-OMA approach has been investigated in several works considering different design goals and under different assumptions [4]–[16]. In general, those existing works can be categorized based on different aspects such as performance metrics, optimization techniques, clustering methods and fading channel models. For example, some authors focus on maximizing the sum rate [4], [5], maximizing the energy efficiency (EE) [6], or minimizing the total power consumption [7]. Other authors consider establishing fairness among the users in terms of diversity order [8], data rate [9], outage [10],

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

and throughput [11]. In addition, user clustering algorithms in hybrid NOMA-OMA systems are considered in several works. For example, heuristic user clustering methods are proposed in [12], [17] based on the channel gains, while machine learning methods are studied in [13]. However, none of these methods are based on closed-form expressions that can quantify the resource demand of a cluster and hence can facilitate the clustering algorithm. In contrast, the user clustering algorithm developed in this paper will be based on closed-form expressions of the resource demand.

Another observation with regard to the existing user clustering methods for hybrid NOMA-OMA systems is that many of them use static algorithms, which require the total number of users to be fixed before running the algorithm [14]. There are other algorithms that consider dynamic scenarios in which some users can enter or exit the network during the running of the clustering algorithm [13]. Cluster size (N) is another important parameter in the clustering procedure. This parameter is fixed as $N = 2$ in some papers [14], [15], and as $N \geq 2$ in [16]. Moreover, a recent work considers the more general case of having a variable number of users in each cluster [5], whereas the work in [13] allows users dynamically leave their current cluster and join a better cluster based on some criteria.

A differentiating feature in the research works concerning the hybrid NOMA-OMA scheme is the assumption on the channel state information (CSI). Most of the works, such as [4]–[7], assume perfect instantaneous CSI, which is either impractical or imposes heavy signaling overhead to practically achieve it. In contrast, assuming and requiring statistical CSI only (which is also considered in this paper) can mitigate the overhead issue since the channels can be monitored over longer periods of time, and hence requiring less feedback to be sent to the transmitter. Furthermore, it is pointed out that most works on hybrid NOMA-OMA systems adopt the Rayleigh fading channel model [10], whereas a more general fading model, such as the Nakagami-m fading, has not been considered in the literature.

For clarity, Table 1 summarizes the key points in the above discussion and highlights the differences among existing works on user clustering in hybrid NOMA-OMA systems with respect to research objectives and assumptions.

Considering the above background, in this paper we investigate the problem of user clustering, resource allocation and decoding order selection in a hybrid NOMA-OMA system. In order to guarantee fairness among all the users, we maximize the minimum success probability among them, which is equivalent to minimizing the maximum outage probability. The channel model is Nakagami-m fading and only statistical CSI is available at the transmitter. This channel model presents a high complexity of the resource allocation problem under consideration and it affects all aspects of the solution, including optimal decoding order of the users, and resource allocation factors. For user clustering, we consider three different scenarios: (a) fixed number of users in each cluster, (b) fixed number of clusters, and (c) variable number of clusters and variable number of users in each cluster.

In order to solve the problem of maximizing the minimum success probability among all the users in a hybrid NOMA-OMA system, we first prove that at the optimal solution, all users have the same success probability, which is called a common success probability (CSP). Then, we propose an efficient algorithm to find the optimal CSP and optimal resource allocation factors simultaneously. Next, we derive the optimal inter-cluster power allocation factor for each cluster in a closed form, which is the sum of optimal power allocation factors of individual users in that cluster. We also derive closed-form expressions for the optimal decoding order and intra-cluster power allocation factors of individual users based on the optimal CSP and resource allocation factor of each cluster.

In summary, the contributions of this paper are as follows:

- Proposing a novel scheme for user clustering, resource allocation and decoding order selection in a hybrid NOMA-OMA system to guarantee fairness among all users in terms of success probability (or, equivalently, its complement outage probability).
- Proposing an efficient algorithm for finding both the optimal CSP of the users and optimal resource (time slot or bandwidth) allocation factors of the clusters in the system.
- Deriving closed-form expressions for the optimal decoding order, individual user power allocation factors and cluster power allocation factors.
- Proposing three efficient user clustering algorithms considering constraints such as fixed cluster sizes or fixed number of clusters.
- Showing that establishing fairness among all users in a hybrid NOMA-OMA system in terms of the success probability of users can also improve the sum throughput of the system.

The rest of the paper is organized as follows. Section II describes the system model. Section III studies the optimal intra-cluster power allocation and decoding order selection for one cluster. Section IV examines the problem of optimal inter-cluster power and resource allocation. Section V proposes user clustering algorithms. Section VI describes the complete proposed scheme. Section VII evaluates performance of the proposed scheme. Section VIII concludes the paper.

II. SYSTEM MODEL

We consider a hybrid NOMA-OMA downlink system with a single-antenna base station (BS) sending mutually-independent information to K single-antenna mobile users. With the hybrid NOMA-OMA, the BS arranges users into L clusters. An orthogonal multiple access scheme such as time division multiple access (TDMA) or orthogonal frequency-division multiple access (OFDMA) is used across different clusters, whereas a power domain NOMA (PD-NOMA) is used within a cluster. The choice for the

TABLE 1. Comparison of existing works on user clustering in hybrid NOMA-OMA systems.

Feature	Reference / Sub-Feature	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	this paper
Optimization objective	Max sum rate	✓	✓							✓	✓	✓	✓		
	Max sum throughput								✓					✓	
	Max energy efficiency			✓											
	Max min rate						✓								
	Max min success probability							✓							✓
	Max min diversity order					✓									
	Min total power consumption				✓										
Optimization variable	Power allocation		✓	✓	✓		✓	✓	✓			✓	✓	✓	✓
	Bandwidth/sub-channel allocation			✓		✓		✓					✓		✓
	Time-slot allocation		✓				✓	✓							✓
	User clustering	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Clustering method	Random user clustering							✓							
	Channel gain difference heuristic	✓		✓		✓	✓		✓			✓	✓	✓	
	Machine learning based										✓				
	Game theory based		✓							✓					
	Compressive sensing based				✓										
	Clustering cost minimization														✓
Clustering type	Static	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Dynamic		✓								✓				✓
Clustering size	Only 2 users (fixed)						✓		✓	✓		✓	✓		
	$N \geq 2$ users (fixed)	✓		✓	✓	✓		✓						✓	✓
	$N \geq 1$ user (variable)		✓								✓				✓
Link type	Uplink			✓									✓	✓	
	Downlink	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
CSI type	Perfect	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	
	Statistical							✓							✓
Fading model	Rayleigh	✓						✓	✓			✓			
	Nakagami-m														✓

inter-cluster orthogonal multiple access is irrelevant to the analysis in this paper. In particular, the resource allocation factor obtained for each cluster can be interpreted as a proportion of allocated time in the TDMA or as a proportion of allocated bandwidth in OFDMA. Therefore, in the rest of the paper we only refer to the time/bandwidth allocation factor as a resource allocation factor.

Denote the index set of users by $\mathcal{K} = \{1, 2, \ldots, K\}$, the kth user by U_k, index set of clusters by $\mathcal{C} = \{1, 2, \ldots, L\}$, the global index set of users in the ℓth cluster by $C_\ell = \{\nu_{\ell,1}, \nu_{\ell,2}, \ldots, \nu_{\ell,|C_\ell|}\}$ and number of users in C_ℓ by $|C_\ell|$. In fact, C_ℓ is a subset of $\mathcal{K}$, $C_\ell \subset \mathcal{K}$, and contains the global (inter-cluster) indices of the users. Clustering should be done such that each user is a member of exactly one cluster. Thus, we should have

$$\sum_{\ell \in \mathcal{C}} |C_\ell| = K, \quad C_i \cap C_j = \emptyset, \ \forall i, j \in \mathcal{C}, \ i \neq j. \tag{1}$$

We also define $\mathcal{I}_\ell = \{1, 2, \ldots, |C_\ell|\}$, which is intra-cluster index set of the users.

Let the total power of the transmitter be P_T and the total channel resource is W_T, which can be time or bandwidth. The power and resource allocation factor of cluster ℓ are denoted with δ_ℓ and ω_ℓ, respectively ($0 < \delta_\ell < 1$, $0 < \omega_\ell < 1$). Thus, the power and resource allocated to cluster ℓ are $\delta_\ell P_T$ and $\omega_\ell W_T$, respectively. For cluster C_ℓ, the BS combines $|C_\ell|$ independent signals of its users by superposition coding and sends the combined signal to them. Each user has to perform SIC to obtain its own signal. The transmitted signal for cluster C_ℓ, denoted by x_ℓ, is given as

$$x_\ell = \sum_{i \in \mathcal{I}_\ell} \sqrt{\alpha_{\nu_{\ell,i}} \delta_\ell P_T} x_{\nu_{\ell,i}}, \quad \ell \in \mathcal{C}. \tag{2}$$

In the above expression, $x_{\nu_{\ell,i}}$ is the transmitted signal of the ith user in the ℓth cluster, satisfying $E(|x_{\nu_{\ell,i}}|^2) = 1$, and $0 \leq \alpha_{\nu_{\ell,i}} \leq 1$ is the intra-cluster power allocation factor for $U_{\nu_{\ell,i}}$. Hence, $\alpha_{\nu_{\ell,i}} \delta_\ell$ specifies the proportion of the total power P_T that is allocated to user $U_{\nu_{\ell,i}}$. Denote the Nakagami-m fading channel coefficient between the BS and $U_{\nu_{\ell,i}}$ by $h_{\nu_{\ell,i}}$, and additive white Gaussian noise (AWGN) with zero mean and variance N_0 at $U_{\nu_{\ell,i}}$ by $z_{\nu_{\ell,i}}$. Then, the signal received by $U_{\nu_{\ell,i}}$ is

$$y_{\nu_{\ell,i}} = h_{\nu_{\ell,i}} x_\ell + z_{\nu_{\ell,i}}, \quad i \in \mathcal{I}_\ell, \ \ell \in \mathcal{C}. \tag{3}$$

It follows that the normalized instantaneous SNR of the received signal at $U_{\nu_{\ell,i}}$ in the ℓth cluster, $\psi_{\nu_{\ell,i}}$, is given as

$$\psi_{\nu_{\ell,i}} = \delta_\ell \gamma_{\nu_{\ell,i}} / \omega_\ell = (\delta_\ell P_T / \omega_\ell) |h_{\nu_{\ell,i}}|^2 / N_0, \tag{4}$$

where $\gamma_{\nu_{\ell,i}}$ is the normalized instantaneous SNR of $U_{\nu_{\ell,i}}$ when all the available power P_T and resource W_T are allocated to cluster C_ℓ (i.e., $\delta_\ell = 1, \omega_\ell = 1$). Thus, under the assumption of Nakagami-m fading, $\psi_{\nu_{\ell,i}}$ has a Gamma distribution [18]

$$f(\psi_{\nu_{\ell,i}}; m_{\nu_{\ell,i}}) = \frac{m_{\nu_{\ell,i}}^{m_{\nu_{\ell,i}}} \psi_{\nu_{\ell,i}}^{m_{\nu_{\ell,i}}-1}}{\bar{\psi}_{\nu_{\ell,i}}^{m_{\nu_{\ell,i}}} \Gamma(m_{\nu_{\ell,i}})} \exp\left(-\frac{m_{\nu_{\ell,i}} \psi_{\nu_{\ell,i}}}{\bar{\psi}_{\nu_{\ell,i}}}\right), \tag{5}$$

where $m_{\nu_{\ell,i}} \geq 1/2$ is the shape factor,

$$m_{\nu_{\ell,i}} = \frac{\bar{\psi}_{\nu_{\ell,i}}^2}{\sigma_{\psi_{\nu_{\ell,i}}}^2}. \tag{6}$$

The quantities $\bar{\psi}_{\nu_{\ell,i}}$ and $\sigma_{\psi_{\nu_{\ell,i}}}^2$, respectively are the mean and variance of the instantaneous SNR $\psi_{\nu_{\ell,i}}$, and Γ is the Gamma function, defined as

$$\Gamma(m) = \int_0^\infty x^{m-1} e^{-x} dx. \tag{7}$$

In this paper, to maximize the minimum success probability among all users in a hybrid NOMA-OMA system, we adopt a bottom-up problem solving approach. We first investigate the intra-cluster power allocation and decoding order optimization for one cluster. Then based on the obtained results, we solve the inter-cluster power and resource allocation problem. Finally we propose clustering algorithms and combine all the results into a unified scheme. For implementation, the BS follows these steps in the reverse order. First it organizes the users into clusters. Then it determines the inter-cluster power and resource allocation factors. Finally it calculates the optimal decoding order and intra-cluster power allocation factor of each user.

For each cluster, an optimization problem should be solved to maximize the minimum success probability of the cluster users by optimizing power allocation factor of each user and selecting the optimal decoding order of users within the cluster. In our previous work [19], we solved such a problem for a single NOMA cluster with K users. Specifically, we proved in [19] that at the optimal solution of the problem, all users have an equal success probability, which we called a common success probability (CSP) of the users. Then, the optimal decoding order and optimal power allocation factor of the users were derived based on their CSP in a closed form and an efficient algorithm was proposed for finding the optimal CSP of the users. The results in [19] thus lay a foundations for the analysis and optimization of the hybrid NOMA-OMA system operating over Nakagami-m fading channels wherein users are assigned into several clusters. As such, in the next section we briefly review the results in [19]. In Section VI we extend the results of [19] to the more general case of hybrid NOMA-OMA.

Given the large number of parameters and notations used throughout the paper, Table 2 summarizes the main system parameters to facilitate reading the paper.

TABLE 2. Definitions of system parameters.

Parameter	Definition
x_ℓ	The transmitted signal for users of cluster C_ℓ
$\mathcal{K}$	The global index set of all users
U_k	The kth user in $\mathcal{K}$
$\mathcal{C}$	The index set of clusters
C_ℓ	The global index set of users in the ℓth cluster ($C_\ell \subset \mathcal{K}$)
$\lvert C_\ell \rvert$	The total number of users in cluster C_ℓ
K	The total number of users
L	The total number of clusters
N	The number of users in each cluster (when it is fixed)
$\mathcal{I}_\ell$	The intra-cluster index set of users in cluster C_ℓ
W_T	The total available resource (time or bandwidth) at the BS
P_T	The total available power at the BS
δ_ℓ	The power allocation factor of cluster C_ℓ
ω_ℓ	The resource allocation factor of cluster C_ℓ
$x_{\nu_{\ell,i}}$	The transmitted signal of the ith user in the ℓth cluster satisfying $E(\lvert x_{\nu_{\ell,i}} \rvert^2) = 1$
$\alpha_{\nu_{\ell,i}}$	The intra-cluster power allocation factor of the ith user in the ℓth cluster
$h_{\nu_{\ell,i}}$	The Nakagami-m fading channel coefficient between BS and $U_{\nu_{\ell,i}}$
$z_{\nu_{\ell,i}}$	Additive white Gaussian noise (AWGN) at user $U_{\nu_{\ell,i}}$
$y_{\nu_{\ell,i}}$	The signal received by user $U_{\nu_{\ell,i}}$
$\psi_{\nu_{\ell,i}}$	The instantaneous SNR of the received signal at $U_{\nu_{\ell,i}}$ (normalized according to power and resources allocated to cluster C_ℓ)
$\bar{\psi}_{\nu_{\ell,i}}$	The mean of $\psi_{\nu_{\ell,i}}$
$\sigma^2_{\psi_{\nu_{\ell,i}}}$	The variance of $\psi_{\nu_{\ell,i}}$
$\gamma_{\nu_{\ell,i}}$	The instantaneous SNR of the received signal at $U_{\nu_{\ell,i}}$ (normalized according to total resource W_T and total power P_T)
$\bar{\gamma}_{\nu_{\ell,i}}$	The mean of $\gamma_{\nu_{\ell,i}}$
$\sigma^2_{\gamma_{\nu_{\ell,i}}}$	The variance of $\gamma_{\nu_{\ell,i}}$
$m_{\nu_{\ell,i}}$	Shape factor of the Nakagami-m fading model for $U_{\nu_{\ell,i}}$
$\nu_{\ell,i}$	The index of the ith user in the ℓth cluster (not sorted according to the optimal decoding order of that cluster)
$\pi_{\ell,i}$	The index of the ith user in the ℓth cluster (sorted according to an optimal decoding order of that cluster)
Γ	Gamma function defined in (7)
$\gamma^{\pi_{\ell,k}}_{\pi_{\ell,i}}$	The SNR relevant to decoding $x_{\pi_{\ell,i}}$ at $U_{\pi_{\ell,k}}$ by using SIC
$\alpha^{\pi_{\ell,i}}_{I_\ell}$	The sum of power allocation factors of users that are treated as noise in decoding $x_{\pi_{\ell,i}}$
$r_{\pi_{\ell,i}}$	The bit rate of user $U_{\pi_{\ell,i}}$ (normalized according to W_T)
$\zeta_{\pi_{\ell,i}}$	$\zeta_{\pi_{\ell,i}} = 2^{r_{\pi_{\ell,i}}} - 1$ is the parameter defined for simplifying inequalities in (21)
$\mathcal{O}^{\pi_{\ell,k}}_{\pi_{\ell,i}}$	The outage event for $U_{\pi_{\ell,k}}$ in decoding $x_{\pi_{\ell,i}}$ (defined in (11))
$p_{\pi_{\ell,i}}$	The success probability of user $U_{\pi_{\ell,i}}$ (derived in (15))
$\gamma^{\pi_{\ell,k}}_{th}$	The minimum SNR threshold for successful decoding of $x_{\pi_{\ell,k}}$ (derived in (13))
$Q(a,x)$	Regularized upper incomplete gamma function (defined in (16))
$\beta_{\pi_{\ell,k}}$	The metric derived in (17) for finding the optimal decoding order of users in each cluster
$\boldsymbol{\omega}$	The vector of all inter-cluster resource allocation factors $[\omega_1, \omega_2, \ldots, \omega_L]$
p	The common success probability (CSP) of all users in the optimal solution
ϵ	The precision of calculating output parameters in the proposed algorithms
$\mathcal{J}$	Jain's fairness index (defined in (47))

III. INTRA-CLUSTER POWER ALLOCATION AND DECODING ORDER SELECTION

Since this section focuses on power allocation and decoding order for users in one cluster, without loss of generality, we assume that all the power P_T and resource W_T are allocated to cluster C_ℓ ($\delta_\ell = 1$ and $\omega_\ell = 1$). Thus, according to (4), the instantaneous SNR of the ith user in cluster ℓ is given as

$$\psi_{\nu_{\ell,i}} = \gamma_{\nu_{\ell,i}}, \quad i \in \mathcal{I}_\ell. \tag{8}$$

Our objective is to maximize the minimum success probability by optimizing the intra-cluster power allocation factors and the decoding order among all users in the cluster.

With SIC decoding, each user decodes other "prior" user signals one by one, and cancels out their effects on the received signal until its own signal is obtained. In general, the decoding order is a permutation of users' indices, denoted by $\boldsymbol{\pi}_\ell = \{\pi_{\ell,1}, \pi_{\ell,2}, \ldots, \pi_{\ell,|C_\ell|}\}$. If $\pi_{\ell,i} = k$, then x_k is the ith signal to be decoded in cluster ℓ. The SNR at $U_{\pi_{\ell,k}}$ that is relevant to decoding $x_{\pi_{\ell,i}}$ can be calculated as follows

$$\gamma^{\pi_{\ell,k}}_{\pi_{\ell,i}} = \frac{\gamma_{\pi_{\ell,k}} \alpha_{\pi_{\ell,i}}}{\gamma_{\pi_{\ell,k}} \alpha^{\pi_{\ell,i}}_{I_\ell} + 1}, \quad k \in \mathcal{I}_\ell,\ i \le k, \ell \in \mathcal{C}, \tag{9}$$

where $\alpha^{\pi_{\ell,i}}_{I_\ell} = \sum_{j=i+1}^{|C_\ell|} \alpha_{\pi_{\ell,j}}$ is simply the sum of intra-cluster power allocation factors of the users whose signals are decoded after $x_{\pi_{\ell,i}}$ (those signals are treated as noise).

Therefore, based on Shannon's theorem, user $U_{\pi_{\ell,k}}$ cannot decode $x_{\pi_{\ell,i}}$ correctly, if

$$\gamma_{\pi_{\ell,i}}^{\pi_{\ell,k}} < 2^{r_{\pi_{\ell,i}}} - 1, \tag{10}$$

or if one of the prior signals was not decoded successfully, before decoding $x_{\pi_{\ell,i}}$. In (10), $r_{\pi_{\ell,i}}$ is the data rate of user $U_{\pi_{\ell,i}}$, normalized according to total resource W_T and $\gamma_{\pi_{\ell,i}}$ is the normalized SNR of the user (assuming that the total power P_T and resource W_T are allocated to one cluster C_ℓ). Thus, the outage event for user $U_{\pi_{\ell,k}}$ in decoding signal $x_{\pi_{\ell,i}}$ can be defined as

$$\mathcal{O}_{\pi_{\ell,i}}^{\pi_{\ell,k}} = \left\{ \bigcup_{j\in\mathcal{I}_\ell, j\le i} \gamma_{\pi_{\ell,j}}^{\pi_{\ell,k}} < 2^{r_{\pi_{\ell,j}}} - 1 \right\}, \; k \in \mathcal{I}_\ell, \; i \le k, \ell \in \mathcal{C}. \tag{11}$$

Note that for the notation $\mathcal{O}_{\pi_{\ell,i}}^{\pi_{\ell,k}}$ used for the outage event above, the superscript specifies the user who is performing the SIC, whereas the subscript specifies the signal that is being decoded.

Obviously, the outage event for user $U_{\pi_{\ell,k}}$ with respect to decoding $x_{\pi_{\ell,k}}$ is $\mathcal{O}_{\pi_{\ell,k}}^{\pi_{\ell,k}}$, which is simply the event that $U_{\pi_{\ell,k}}$ cannot decode $x_{\pi_{\ell,k}}$ (its own signal) correctly. Hence, the success probability of user $U_{\pi_{\ell,k}}$ can be written as

$$p_{\pi_{\ell,k}} = 1 - \Pr\{\mathcal{O}_{\pi_{\ell,k}}^{\pi_{\ell,k}}\}. \tag{12}$$

In [19], we show that for each user $U_{\pi_{\ell,k}}$, a minimum SNR threshold for successful decoding can be found as

$$\gamma_{th}^{\pi_{\ell,k}} = \frac{2^{r_{\pi_{\ell,k}}} - 1}{\alpha_{\pi_{\ell,k}} - (2^{r_{\pi_{\ell,k}}} - 1)\alpha_{I_\ell}^{\pi_{\ell,k}}}, \quad k \in \mathcal{I}_\ell. \tag{13}$$

Using the above expression simplifies the expression in (11) to

$$\mathcal{O}_{\pi_{\ell,i}}^{\pi_{\ell,k}} = \left\{ \bigcup_{j\in\mathcal{I}_\ell, j\le i} \gamma_{\pi_{\ell,k}} < \gamma_{th}^{\pi_{\ell,j}} \right\}, \quad k \in \mathcal{I}_\ell, \; i \le k. \tag{14}$$

Consequently, it is shown in [19] that the success probability of user $U_{\pi_{\ell,k}}$ can be calculated as

$$\begin{aligned} p_{\pi_{\ell,k}} &= 1 - Pr\{\mathcal{O}_{\pi_{\ell,k}}^{\pi_{\ell,k}}\} = 1 - Pr\left\{ \bigcup_{j\in\mathcal{I}_\ell, j\le k} \gamma_{\pi_{\ell,k}} < \gamma_{th}^{\pi_{\ell,j}} \right\} \\ &= 1 - Pr\left\{\gamma_{\pi_{\ell,k}} < \gamma_{th}^{\pi_{\ell,k}}\right\} = 1 - F_{\gamma_{\pi_{\ell,k}}}\left(\gamma_{th}^{\pi_{\ell,k}}\right), \\ &= Q\left(m_{\pi_{\ell,k}}, m_{\pi_{\ell,k}}\gamma_{th}^{\pi_{\ell,k}}/\bar{\gamma}_{\pi_{\ell,k}}\right), \end{aligned} \tag{15}$$

where $Q(\cdot,\cdot)$ is the regularized upper incomplete gamma function, defined as [20]

$$Q(a,x) = \frac{\int_x^\infty t^{a-1}e^{-t}dt}{\Gamma(a)}. \tag{16}$$

Then, we show that for maximizing the minimum success probability among users, all the users have an equal success probability, called the common success probability (CSP) (Theorem 2 in [19]). Subsequently, assuming that the optimal CSP of users in cluster ℓ is p_ℓ, we show that the optimal decoding order is given by the ascending order of parameter $\beta_{\pi_{\ell,k}}$, defined as

$$\beta_{\pi_{\ell,k}} = \frac{\bar{\gamma}_{\pi_{\ell,k}}}{m_{\pi_{\ell,k}}} Q^{-1}(m_{\pi_{\ell,k}}, p_\ell), \quad k \in \mathcal{I}_\ell, \tag{17}$$

where $Q^{-1}(\cdot,\cdot)$ is the inverse function of $Q(a,x)$ with respect to the second parameter x (see Lemma 3 in [19]). Note that the function $Q^{-1}(\cdot,\cdot)$ can be calculated using a numerical method.[1] In other words, the optimal decoding order $\boldsymbol{\pi}_\ell$ should be such that

$$\beta_{\pi_{\ell,1}} \le \beta_{\pi_{\ell,2}} \le \cdots \le \beta_{\pi_{\ell,|C_\ell|}}. \tag{18}$$

The parameter β actually represents the quality of the channel of each user. Thus, if a user has a lower β it should be given a higher power allocation factor and a higher priority in decoding order. Therefore, the optimal intra-cluster power allocation factors for users in each cluster can be calculated as (for more details, see Theorem 3 in [19])

$$\alpha_{\pi_{\ell,|C_\ell|}} = (2^{r_{\pi_{\ell,|C_\ell|}}} - 1)\frac{m_{\pi_{\ell,|C_\ell|}}}{\bar{\gamma}_{\pi_{\ell,|C_\ell|}} Q^{-1}(m_{\pi_{\ell,|C_\ell|}}, p_\ell)}, \tag{19a}$$

$$\begin{aligned} \alpha_{\pi_{\ell,|C_\ell|-1}} = (2^{r_{\pi_{\ell,|C_\ell|-1}}} - 1)\Bigg(&\frac{m_{\pi_{\ell,|C_\ell|}}(2^{r_{\pi_{\ell,|C_\ell|}}} - 1)}{\bar{\gamma}_{\pi_{\ell,|C_\ell|}} Q^{-1}(m_{\pi_{\ell,|C_\ell|}}, p_\ell)} \\ &+ \frac{m_{\pi_{\ell,|C_\ell|-1}}}{\bar{\gamma}_{\pi_{\ell,|C_\ell|-1}} Q^{-1}(m_{\pi_{\ell,|C_\ell|-1}}, p_\ell)}\Bigg), \end{aligned} \tag{19b}$$

$$\begin{aligned} \alpha_{\pi_{\ell,k}} = &(2^{r_{\pi_{\ell,k}}} - 1) \\ &\times \Bigg(\sum_{i=0}^{|C_\ell|-k-2} \left(\frac{m_{\pi_{\ell,|C_\ell|-i}}(2^{r_{\pi_{\ell,|C_\ell|-i}}} - 1)2^{\sum_{j=i+1}^{|C_\ell|-k-1} r_{\pi_{\ell,|C_\ell|-j}}}}{\bar{\gamma}_{\pi_{\ell,|C_\ell|-i}} Q^{-1}(m_{\pi_{\ell,|C_\ell|-i}}, p_\ell)} \right) \\ &+ \frac{m_{\pi_{\ell,k+1}}(2^{r_{\pi_{\ell,k+1}}} - 1)}{\bar{\gamma}_{\pi_{\ell,k+1}} Q^{-1}(m_{\pi_{\ell,k+1}}, p_\ell)} + \frac{m_{\pi_{\ell,k}}}{\bar{\gamma}_{\pi_{\ell,k}} Q^{-1}(m_{\pi_{\ell,k}}, p_\ell)} \Bigg), \\ &1 \le k \le |C_\ell| - 2. \end{aligned} \tag{19c}$$

In [22], necessary conditions are derived for power allocation factors of users in a NOMA system to prevent the signal constellations from overlapping in the superposition coding. It is assumed that each of $|C_\ell|$ users of the NOMA cluster employs a square quadrature amplitude modulation (QAM) constellation. We know the fact that the modulation order $M_{\pi_{\ell,i}}$ and bit rate $r_{\pi_{\ell,i}}$ of user $U_{\pi_{\ell,i}}$ are related as

$$\frac{r_{\pi_{\ell,i}}}{\log_2(M_{\pi_{\ell,i}})} = R_\ell, \tag{20}$$

where R_ℓ is the symbol rate of the transmitter for the ℓth cluster. Thus, we can restate the conditions derived in [22] for power allocation factors using the notations in this paper as

$$\sqrt{\frac{\alpha_{\pi_{\ell,i}}}{\zeta_{\pi_{\ell,i}}}} > \sum_{j=i+1}^{|C_\ell|} \sqrt{\frac{\alpha_{\pi_{\ell,j}}}{\zeta_{\pi_{\ell,j}}}} \left(\sqrt{\zeta_{\pi_{\ell,j}} + 1} - 1\right),$$

[1] This function is implemented in `SciPy` library of Python with the name `gammainccinv` [21].

$$1 \leq i \leq |C_\ell| - 1, \tag{21}$$

where $\zeta_{\pi_{\ell,i}} = 2^{r_{\pi_{\ell,i}}} - 1$, and without loss of generality, we set $R_\ell = 1$ (for more details, the reader is referred to Proposition 1 and Inequality (19a) in [22]). In the next theorem, we prove that our proposed power allocation scheme always satisfies those necessary conditions.

Theorem 1: The power allocation factors (19) for any number of users $|C_\ell|$ in the NOMA cluster and arbitrary modulation orders $M_{\pi_{\ell,i}}$ employed by the users satisfy the conditions given by (21).

Proof: See Appendix A. □

Furthermore, the sum of all power allocation factors as derived in (19) can be calculated in a closed form as

$$\begin{aligned} S(p_\ell, \boldsymbol{\pi}_\ell) &= \sum_{i=1}^{|C_\ell|} \alpha_{\pi_{\ell,i}} \\ &= \frac{m_{\pi_{\ell,1}}(2^{r_{\pi_{\ell,1}}} - 1)}{\bar{\gamma}_{\pi_{\ell,1}} Q^{-1}(m_{\pi_{\ell,1}}, p_\ell)} + \sum_{i=2}^{|C_\ell|} \frac{m_{\pi_{\ell,i}}(2^{r_{\pi_{\ell,i}}} - 1)2^{\sum_{j=1}^{i-1} r_{\pi_{\ell,j}}}}{\bar{\gamma}_{\pi_{\ell,i}} Q^{-1}(m_{\pi_{\ell,i}}, p_\ell)}, \end{aligned} \tag{22}$$

which is independent of individual intra-cluster power allocation factors. The sum of intra-cluster power allocation factors $S(p_\ell, \boldsymbol{\pi}_\ell)$ should be exactly one. A value less than one means some of the allocated resource remains unused and a value higher than one means that the cluster is using more resources than what has been allocated to it. Thus, in [19], we incorporated and proved the necessity of the constraint

$$S(p_\ell, \boldsymbol{\pi}_\ell) = 1, \tag{23}$$

to find the optimal CSP in an efficient way by performing a binary search on parameter p_ℓ. For completeness, the algorithm for finding the optimal CSP is included in Appendix B. In the next section, we extend that algorithm to simultaneously find both the optimal CSP and optimal inter-cluster resource allocation factors when users are grouped into several clusters in a hybrid NOMA-OMA system. We also generalize the obtained intra-cluster power allocation factors (19) to the case of hybrid NOMA-OMA in Section VI.

IV. INTER-CLUSTER POWER AND RESOURCE ALLOCATION

As explained in the previous section, within a cluster, maximizing the minimum success probability of all users can be done by the following steps:

1) Find the optimal CSP p_ℓ by running Algorithm 7 (see Appendix B).
2) Select the optimal decoding order of users in the cluster according to (18).
3) Calculate the optimal power allocation factors for users by (19).

Since all users in a cluster have the same success probability p_ℓ, the problem of maximizing the minimum success probability of users across all clusters can be formulated as

$$\max_{\delta_\ell, \omega_\ell, \ell \in \mathcal{C}} \min_{\ell \in \mathcal{C}} \quad p_\ell \tag{24a}$$

$$\text{s.t. } \delta_\ell \geq 0, \quad \ell \in \mathcal{C}, \tag{24b}$$

$$\sum_{\ell \in \mathcal{C}} \delta_\ell \leq 1, \tag{24c}$$

$$\omega_\ell \geq 0, \quad \ell \in \mathcal{C}, \tag{24d}$$

$$\sum_{\ell \in \mathcal{C}} \omega_\ell \leq 1, \tag{24e}$$

$$0 \leq p_\ell \leq 1. \tag{24f}$$

Similar to the intra-cluster optimization problem, we can also prove that at the optimal solution of the inter-cluster optimization problem in (24), the success probabilities of all users are equal. This result is summarized in the following lemma.

Lemma 1: At the optimal solution of problem (24), the success probabilities of all users across all the clusters are equal and we have

$$p_\ell = p_0, \quad \ell \in \mathcal{C}. \tag{25}$$

Proof: See Appendix C. □

Furthermore, we have the following results regarding the constraints of problem (24).

Lemma 2: At the optimal solution of problem (24), constraints (24b) and (24d) are satisfied with inequality, and constraint (24c) is satisfied with equality.

Proof: This lemma can be proved by contradiction. Suppose that for one of the clusters, either constraint (24b) or (24d) is satisfied with equality. Then the success probability of that cluster would be zero, which contradicts with the objective of maximizing the minimum success probability of all users. On the other hand, if constraint (24c) is satisfied with inequality, then all the cluster power allocation factors, $\delta_\ell, \ell \in \mathcal{C}$ can be multiplied by $1/\sum_{\ell \in \mathcal{C}} \delta_\ell$. Because the success probability is a strictly increasing function of power allocation factors, the increase of power allocation factors increases the success probabilities of users in all clusters, which is a contradiction. Thus, the lemma is proved. □

Recall that the results in the previous section were obtained when the total power P_T and resource W_T are allocated to a single cluster C_ℓ and the resulting data rates and SNRs of users in the cluster are normalized according to those parameters. In this section the power and resource allocated to cluster C_ℓ are $\delta_\ell P_T$ and $\omega_\ell W_T$, respectively. Thus, instead of parameters r and $\bar{\gamma}$, we need to use parameters r/ω_ℓ and $\delta_\ell \bar{\gamma}/\omega_\ell$, respectively, in the function $S(p_\ell, \boldsymbol{\pi}_\ell)$ defined in (22). On the other hand, from Lemma 1 we know that at the optimal solution of problem (24) the success probability of all users across all the clusters are the same. Thus, assuming that the CSP is p we can rewrite (23) for each cluster as follows:

$$S(p, \boldsymbol{\pi}_\ell) = \frac{m_{\pi_{\ell,1}}(2^{r_{\pi_{\ell,1}}/\omega_\ell} - 1)}{(\delta_\ell \bar{\gamma}_{\pi_{\ell,1}}/\omega_\ell) Q^{-1}(m_{\pi_{\ell,1}}, p)}$$

$$+\sum_{i=2}^{|C_\ell|}\frac{m_{\pi_{\ell,i}}(2^{r_{\pi_{\ell,i}}/\omega_\ell}-1)2^{\sum_{j=1}^{i-1}r_{\pi_{\ell,j}}/\omega_\ell}}{(\delta_\ell\bar{\gamma}_{\pi_{\ell,i}}/\omega_\ell)Q^{-1}(m_{\pi_{\ell,i}},p)}=1,\quad \ell\in\mathcal{C},\tag{26}$$

where $\pi_{\ell,i}$ is the index of the ith user in the optimal decoding order of cluster C_ℓ.

From (26) we can derive the power allocation factor of each cluster based on its resource allocation factor, CSP and statistical CSI in a closed form:

$$\delta_\ell=\omega_\ell\left(\frac{m_{\pi_{\ell,1}}(2^{r_{\pi_{\ell,1}}/\omega_\ell}-1)}{\bar{\gamma}_{\pi_{\ell,1}}Q^{-1}(m_{\pi_{\ell,1}},p)}+\sum_{i=2}^{|C_\ell|}\frac{m_{\pi_{\ell,i}}(2^{r_{\pi_{\ell,i}}/\omega_\ell}-1)2^{\sum_{j=1}^{i-1}r_{\pi_{\ell,j}}/\omega_\ell}}{\bar{\gamma}_{\pi_{\ell,i}}Q^{-1}(m_{\pi_{\ell,i}},p)}\right),\quad \ell\in\mathcal{C}.\tag{27}$$

In Lemma 2 we proved that at the optimal solution of problem (24), the sum of all inter-cluster power allocation factors δ_ℓ is equal to one. Thus, if we denote the vector of all inter-cluster resource allocation factors as $\boldsymbol{\omega}=[\omega_1,\omega_2,\ldots,\omega_L]$ and define $h(p,\boldsymbol{\omega})$ as

$$h(p,\boldsymbol{\omega})=\sum_{\ell\in\mathcal{C}}\omega_\ell\left(\frac{m_{\pi_{\ell,1}}(2^{r_{\pi_{\ell,1}}/\omega_\ell}-1)}{\bar{\gamma}_{\pi_{\ell,1}}Q^{-1}(m_{\pi_{\ell,1}},p)}+\sum_{i=2}^{|C_\ell|}\frac{m_{\pi_{\ell,i}}(2^{r_{\pi_{\ell,i}}/\omega_\ell}-1)2^{\sum_{j=1}^{i-1}r_{\pi_{\ell,j}}/\omega_\ell}}{\bar{\gamma}_{\pi_{\ell,i}}Q^{-1}(m_{\pi_{\ell,i}},p)}\right),\tag{28}$$

then according to (27) and (28) we should have

$$\sum_{\ell\in\mathcal{C}}\delta_\ell=1\Rightarrow h(p,\boldsymbol{\omega})=1.\tag{29}$$

Therefore, we can reformulate the problem in (24) as follows:

$$\max_{\omega_\ell,\ell\in\mathcal{C}}\ p\tag{30a}$$
$$\text{s.t. } h(p,\boldsymbol{\omega})=1,\tag{30b}$$
$$\sum_{\ell\in\mathcal{C}}\omega_\ell=1,\tag{30c}$$
$$\omega_\ell>0,\quad \ell\in\mathcal{C},\tag{30d}$$
$$0\le p\le 1.\tag{30e}$$

Under the normal and expected condition[2] that $0.5\le p\le 1$, we can prove that problem (30) is convex. As such, we propose an efficient algorithm for solving it by utilizing Karush-Kuhn-Tucker (KKT) conditions [23].

Lemma 3: Under the practical condition that $0.5\le p\le 1$, *problem (30) is convex.*

Proof: Refer to Appendix D. □

According to Lemma 3, after modifying last constraint of problem (30) to $0.5\le p\le 1$ problem is convex and the KKT conditions of the resulting convex optimization problem are as follows:

- Stationarity:
$$\mu_\ell-\lambda_1\frac{\partial h(p,\boldsymbol{\omega})}{\partial\omega_\ell}-\lambda_2=0,\tag{31a}$$
$$1-\lambda_1\frac{\partial h(p,\boldsymbol{\omega})}{\partial p}+\mu_{L+1}-\mu_{L+2}=0.\tag{31b}$$
- Primal feasibility:
$$h(p,\boldsymbol{\omega})=1,\tag{31c}$$
$$\sum_{\ell\in\mathcal{C}}\omega_\ell=1,\tag{31d}$$
$$\omega_\ell>0,\quad \ell\in\mathcal{C},\tag{31e}$$
$$0.5\le p\le 1.\tag{31f}$$
- Dual feasibility:
$$\mu_\ell\ge 0,\quad \ell\in\{1,2,\ldots,L+2\}.\tag{31g}$$
- Complementary slackness:
$$\mu_\ell(-\omega_\ell)=0,\quad \ell\in\{1,2,\ldots,L\},\tag{31h}$$
$$\mu_{L+1}(-p+0.5)=0,\tag{31i}$$
$$\mu_{L+2}(p-1)=0.\tag{31j}$$

Based on (31h), (31i) and (31j), it is straightforward to verify that all μ_ℓ values should be equal to zero. Otherwise it will result in special cases that are not practically feasible nor important. For instance, any of μ_ℓ, $\ell\in\{1,2,\ldots,L\}$ not being zero means that the resource allocation factor and consequently, the success probability of that cluster ℓ is zero. Thus, the KKT conditions (31) can be rewritten as

$$\frac{\partial h(p,\boldsymbol{\omega})}{\partial\omega_\ell}=-\frac{\lambda_2}{\lambda_1}=\lambda,\tag{32a}$$
$$\frac{\partial h(p,\boldsymbol{\omega})}{\partial p}=\frac{1}{\lambda_1},\tag{32b}$$
$$h(p,\boldsymbol{\omega})=1,\tag{32c}$$
$$\sum_{\ell\in\mathcal{C}}\omega_\ell=1,\tag{32d}$$
$$\omega_\ell>0,\quad \ell\in\mathcal{C},\tag{32e}$$
$$0.5\le p\le 1.\tag{32f}$$

To simplify the relations we define $\lambda=-\frac{\lambda_2}{\lambda_1}$ in (32a). Algorithm 1 is then proposed to find inter-cluster resource allocation factors ω_ℓ, $\ell\in\mathcal{C}$, and CSP p of all users simultaneously.

In this algorithm, the parameter ϵ specifies the precision of the output parameters and can be chosen arbitrarily as an input of the algorithm. As a default value we set it to $\epsilon=10^{-3}$ in the simulations. In lines 4 and 7 of this algorithm, it is necessary to find λ and $\boldsymbol{\omega}$ such that (32a) and (32d) are satisfied. These parameters can be found using Algorithms 2 and 3, respectively, and they are discussed further below. In Algorithm 1, the value of p is not restricted to the interval (0.5, 1) as stated by constraint (32f). For the case $p\in(0,0.5)$, the algorithm will converge to a solution that guarantees fairness among the users, however,

[2]It is pointed out that the case $0\le p<0.5$ is of no practical interest. However, our proposed algorithm for solving problem (30) still finds a solution, albeit the optimality of the solution is not guaranteed in this case.

Algorithm 1 Finding Resource Allocation Factors and CSP

Input: ϵ, users' statistical CSI and clustering.
Output: CSP p and resource allocation factors $\boldsymbol{\omega}$.
Initialize: $p_L = 0,\ p_H = 1,\ p = 0.5,\ \epsilon = 10^{-3}$
Find λ, ω_ℓ according to Algorithms 2 and 3, respectively.
while $|h(p, \boldsymbol{\omega}) - 1| > \epsilon$ and $p_H - p_L > \epsilon$ **do**
  $p = \frac{p_L + p_H}{2}$
  Find λ, ω_ℓ according to Algorithms 2 and 3, respectively.
  if $h(p, \boldsymbol{\omega}) < 1$ **then**
    $p_L = p$
  else if $h(p, \boldsymbol{\omega}) > 1$ **then**
    $p_H = p$
  else
    return $p,\ \boldsymbol{\omega}$
  end if
end while
return $p,\ \boldsymbol{\omega}$.

we cannot prove the optimality of such a solution by using the KKT conditions. As pointed before, the case that the success probabilities are less than 0.5 are not practically important.

Parameter λ can be found using (32a). In Appendix D (proof of Lemma 3), we obtain $\frac{\partial h(p,\boldsymbol{\omega})}{\partial \omega_\ell}$ in (62). Therefore, λ can be derived as follows:

$$\lambda = h'_{\omega_\ell}(p, \boldsymbol{\omega}) = \frac{-1}{\beta_{\pi_{\ell,1}}} - \sum_{i=1}^{|C_\ell|-1} \left(\left(\frac{1}{\beta_{\pi_{\ell,i}}} - \frac{1}{\beta_{\pi_{\ell,i+1}}} \right) \times f(2^{\sum_{j=1}^{i} r_{\pi_{\ell,j}}/\omega_\ell}) \right) - \frac{f(2^{\sum_{j=1}^{|C_\ell|} r_{\pi_{\ell,j}}/\omega_\ell})}{\beta_{\pi_{\ell,|C_\ell|}}}, \quad (33)$$

where $f(x) = x(\ln x - 1)$. It is clear from (33) that $\lambda < 0$ always holds true. On the other hand, from (65) in Appendix D we have

$$\frac{\partial \lambda}{\partial \omega_\ell} = \frac{\partial^2 h(p, \boldsymbol{\omega})}{\partial \omega_\ell^2} > 0, \quad (34)$$

which means that λ is a strictly increasing function of ω_ℓ, and also the converse function ω_ℓ is a strictly increasing function of λ. Thus, using the fact that ω_ℓ values should be such that (32d) is satisfied, we can derive boundaries for the acceptable range of λ values. We know that $\omega_\ell \in (0, 1)$ and based on (33), choosing ω_ℓ in the neighborhood of zero results in $\lambda \to -\infty$. This means that if λ is less than a threshold value, then all ω_ℓ values will be near to zero, and their sum would not add up to one to satisfy (32d). Therefore, an acceptable range for parameter λ is as follows:

$$\lambda_{\max} = \min_{\ell \in \mathcal{C}} \left. \frac{\partial h(p, \boldsymbol{\omega})}{\partial \omega_\ell} \right|_{\omega_\ell = 1}, \quad (35)$$

$$\lambda_{\min} = \min_{\ell \in \mathcal{C}} \left. \frac{\partial h(p, \boldsymbol{\omega})}{\partial \omega_\ell} \right|_{\omega_\ell = 1/L}, \quad (36)$$

$$\lambda \in (\lambda_{\min}, \lambda_{\max}). \quad (37)$$

Algorithm 2 Finding Parameter λ

1: **Input:** $p,\ \epsilon = 10^{-3}$, users' statistical CSI and clustering.
2: **Output:** λ such that (32a) and (32d) are satisfied.
3: **Initialize:** $\lambda_{\min} = \min_{\ell \in \mathcal{C}} \left. \frac{\partial h(p,\boldsymbol{\omega})}{\partial \omega_\ell} \right|_{\omega_\ell = 1/L}$, $\lambda_{\max} = \min_{\ell \in \mathcal{C}} \left. \frac{\partial h(p,\boldsymbol{\omega})}{\partial \omega_\ell} \right|_{\omega_\ell = 1}$, $\omega_\ell = 1,\ \forall \ell \in \mathcal{C}$
4: **while** $|\sum_{\ell \in \mathcal{C}} \omega_\ell - 1| > \epsilon$ and $\lambda_{\max} - \lambda_{\min} > \epsilon$ **do**
5: $\lambda = \frac{\lambda_{\min} + \lambda_{\max}}{2}$
6: Find $\omega_\ell,\ \ell \in \mathcal{C}$ according to Algorithm 3.
7: **if** $\sum_{\ell \in \mathcal{C}} \omega_\ell < 1$ **then**
8: $\lambda_{\min} = \lambda$
9: **else if** $\sum_{\ell \in \mathcal{C}} \omega_\ell > 1$ **then**
10: $\lambda_{\max} = \lambda$
11: **else**
12: **return** λ
13: **end if**
14: **end while**
15: **return** λ.

Algorithm 3 Finding Parameter ω_ℓ

1: **Input:** $\lambda,\ p,\ \epsilon = 10^{-3}$, statistical CSI of all users in a given cluster.
2: **Output:** ω_ℓ such that $e(\omega_\ell) = 0$ or $h'_{\omega_\ell}(p, \boldsymbol{\omega}) = \lambda$.
3: **Initialize:** $\omega_{\ell_{\min}} = 0,\quad \omega_{\ell_{\max}} = 1,\quad \omega_\ell = 0.5$
4: **while** $|e(\omega_\ell)| > \epsilon$ and $\omega_{\ell_{\max}} - \omega_{\ell_{\min}} > \epsilon$ **do**
5: $\omega_\ell = \frac{\omega_{\ell_{\min}} + \omega_{\ell_{\max}}}{2}$
6: **if** $e(\omega_\ell) > 0$ **then**
7: $\omega_{\ell_{\min}} = \omega_\ell$
8: **else if** $e(\omega_\ell) < 0$ **then**
9: $\omega_{\ell_{\max}} = \omega_\ell$
10: **else**
11: **return** ω_ℓ
12: **end if**
13: **end while**
14: **return** ω_ℓ.

Recall that ω_ℓ is a strictly increasing function of λ. Thus, if $\lambda > \lambda_{\max}$ the summation $\sum_{\ell \in \mathcal{C}} \omega_\ell > 1$ and if $\lambda < \lambda_{\min}$ the summation $\sum_{\ell \in \mathcal{C}} \omega_\ell < 1$. Now that the parameter λ is bounded, we can adopt a binary search for finding its value, as outlined in Algorithm 2. In line 6 of this algorithm, it is necessary to calculate ω_ℓ values, which are bounded to the interval (0, 1) and should satisfy (33) with the given value for λ. To this end, we consider the following function

$$e(\omega_\ell) = \lambda - h'_{\omega_\ell}(p, \boldsymbol{\omega}). \quad (38)$$

The root of $e(\omega_\ell)$ is the optimal value of ω_ℓ. Therefore, based on the fact that λ is a strictly increasing function of ω_ℓ, a binary search can be used to find the optimal ω_ℓ as proposed in Algorithm 3.

By using Algorithms 1, 2 and 3 we can obtain the optimal CSP p and the optimal inter-cluster resource allocation factors $\omega_\ell,\ \ell \in \mathcal{C}$. Then, the inter-cluster power allocation factors

δ_ℓ, $\ell \in \mathcal{C}$ can be readily found from the closed-form expression in (27).

It should be pointed out that Algorithms 1, 2 and 3 are operated jointly to find the optimal CSP and resource allocation factors. Specifically, Algorithm 1 performs a binary search on CSP p of the clusters, and finds the optimal value in $\log_2(1/\epsilon)$ iterations. In each iteration, it calls Algorithm 2, which also performs a binary search to find the proper value of λ in $\log_2(1/\epsilon)$ iterations. Likewise, Algorithm 2 in each iteration calls Algorithm 3 to find values of ω_ℓ by another binary search. These three nested binary search algorithms find the optimal values of ω_ℓ and CSP p of all L clusters in $L\left[\log_2(1/\epsilon)\right]^3$ iterations, which grows linearly with the number of clusters L. In contrast, the exhaustive search method would need to evaluate $(1/\epsilon)^{2L+1}$ states to find the optimal ω_ℓ values, optimal δ_ℓ values and optimal CSP of users, which grows exponentially with the number of clusters L. Thus, the computational complexity of our proposed method is much less than that of the exhaustive search method.

V. PROPOSED USER CLUSTERING ALGORITHMS

Building on the results given in the previous section, in this section we shall propose user clustering algorithms for the following three cases:

1) The number of users in each cluster $|C_\ell|$ is fixed.
2) The total number of clusters L is fixed, but the number of users in each cluster $|C_\ell|$ can be variable.
3) Both $|C_\ell|$ and L are variable.

All three algorithms are developed based on the same principle of minimizing the power consumption of all clusters according to the closed-form expression (27) for power allocation factor of each cluster. We consider constant values for the resource allocation factor ω_ℓ and target success probability p. In the clustering step, the goal is to find users who can cooperate the best in a NOMA setting, in the sense that they need the least power to achieve a given target success probability. After finding the clustering structure, the optimal power allocation factor, resource allocation factor, and optimal CSP of users are determined based on the total available power and resource at the transmitter, according to the results of the previous section. We also investigate the impact of selecting the initial value of CSP on the performance of user clustering by simulations and show that even without iterating over multiple initial values of CSP our proposed algorithms outperform existing algorithms (see Section VII). Therefore, in developing clustering algorithms we assume that resource is allocated equally to all clusters and consider $p = 0.95$ as a target success probability (but they can be chosen any other value arbitrarily). To derive the cost metric δ_ℓ for any cluster C_ℓ, it is necessary to select the optimal decoding order $\boldsymbol{\pi}_\ell$ according to (18).

A. CASE 1: EQUAL NUMBER OF USERS IN ALL CLUSTERS

Let K be the total number of users and N the number of users in each cluster. Then the number of clusters is $L = \lceil \frac{K}{N} \rceil$ (the number of users in the last cluster may be less than N if N does not divide K). For initialization of the clustering algorithm, we assume that the total available resource is divided equally among the clusters, i.e., $\omega_\ell = 1/L, \ell \in \mathcal{C}$. We also consider an arbitrarily given target success probability, for example $p = 0.95$.

We first sort users based on the ascending order of parameter β_{π_k}, defined in (17). The first user in the list is simply selected as the first user of the first cluster. To choose the second user of the first cluster, we examine every remaining user in the list together with the first user and form a two-user cluster. We calculate δ_ℓ for each of these two-user clusters according to (27) with $\omega_\ell = 1/L$, $p = 0.95$ and the optimal decoding order in (18). Then the user having the lowest δ_ℓ is chosen as the second user of the first cluster. The same procedure is then repeated in order to choose the 3rd, 4th, ..., and Nth users of the first cluster. After selecting the Nth user of the first cluster, we continue with the same procedure to create the next clusters until all users are clustered. Algorithm 4 provides pseudo-code for the proposed clustering scheme.

Ignoring the complexity in selecting the first user in each cluster, for selecting the second user in the first cluster δ_ℓ should be calculated $K-1$ times, and for selecting the third user, δ_ℓ needs to be calculated $K-2$ times, etc. Thus, the computational complexity of Algorithm 4 is at most

$$(K-1)+(K-2)+\cdots+1 = \frac{K(K-1)}{2} = O(K^2), \quad (39)$$

which increases polynomially in time with the total number of users K. It should also be pointed out that Algorithm 4 is a static algorithm since all the users should be available before running the algorithm.

B. CASE 2: FIXED NUMBER OF CLUSTERS L

Recall that Algorithm 4 assumes that the number of users in each cluster is fixed, which also means the number of clusters is fixed. For the case considered in this subsection, we relax that constraint and require that only the total number of clusters is fixed, whereas there is no constraint on the number of users in each cluster. To put K users into L clusters, we first sort the list of users based on the ascending order of parameter β_{π_k} in (17). Then we choose the first L users of the sorted list (who have the weakest channels) and put them into L clusters. Thus, after this step, each cluster has one user. For clustering the rest of users, based on the sorted list, we calculate δ_ℓ, $\ell \in \{1, 2, \ldots, L\}$ for each user assuming that it has joined cluster C_ℓ and select the cluster that results in the minimum value of δ_ℓ (after adding that user).

Algorithm 5 gives pseudo-code for this clustering scheme. It is pointed out that this algorithm can be deployed in a dynamic scenario as well. Since any newly arrived user can join one of the existing clusters based on the criterion of minimizing δ_ℓ without changing the whole clustering structure. Sorting the users based on β_{π_k} in advance has the

Algorithm 4 Clustering Algorithm With Fixed $|C_\ell|$

Input: Set of all users $\mathcal{K}$, their statistical CSI and N.
Output: Clustered sets of users with N users in each cluster.
$L = \lceil \frac{K}{N} \rceil$
$\mathcal{K} = \{U_1, U_2, \ldots, U_K\}$, sorted list of users based on ascending order of β_{π_k} defined in (17).
for $i \in \{1, 2, \ldots, L-1\}$ **do**
  U_{sel} is selected as the first user of $\mathcal{K}$, and remove it from $\mathcal{K}$.
  $C_i = \{U_{sel}\}$.
  for $j \in \{2, 3, \ldots, N\}$ **do**
    Select the jth user of C_i from $\mathcal{K}$ such that δ_i is minimum and remove that user from $\mathcal{K}$.
  end for
end for
Put all remaining users in the last cluster, $C_L = \mathcal{K}$.
return $\{C_1, C_2, \ldots, C_L\}$.

Algorithm 5 Clustering With a Fixed Number of Clusters L

Input: Set of all users $\mathcal{K}$, their statistical CSI and L.
Output: Clustered sets of users with the number of clusters equal to L.
Sort the users based on the ascending order of β_{π_k} in (17) and store them as $\mathcal{K} = \{U_1, U_2, \ldots, U_K\}$.
for $i = 1$ to L **do**
  $C_i = \{U_i\}$
end for
for $i = L+1$ to K **do**
  Put U_i in the cluster C_ℓ that results in minimum δ_ℓ, $1 \leq \ell \leq L$.
end for
return $\{C_1, C_2, \ldots, C_L\}$.

benefit of simplifying the calculation of δ_ℓ as explained next. In calculating δ_ℓ for a cluster, it is necessary to select the optimal decoding order for that cluster according to (18). But if we sort the users first, each user who joins a cluster will be the last user in the optimal decoding order of that cluster. However, for the newly arrived users in a dynamic scenario, the optimal decoding order should be calculated.

Ignoring the complexity in clustering the first L users, for clustering each of the remaining users, δ_ℓ should be calculated L times. Thus, the computational complexity of Algorithm 5 is proportional to $(K-L)L$, which increases polynomially in time with number of users K and number of clusters L.

C. CASE 3: CLUSTERING WITH VARIABLE CLUSTER SIZE $|C_\ell|$ AND VARIABLE CLUSTER COUNT L

In this case, we examine the most general scenario that the number of users in each cluster as well as total number of clusters are variable. Considering the latency and computational complexity of SIC, it is reasonable to set limits on the minimum and maximum numbers of clusters,

Algorithm 6 Clustering With Variable Number of Clusters L and Variable Number of Users in Each Cluster $|C_\ell|$

Input: Set of all users $\mathcal{K}$, their statistical CSI, initial target success probability p, $L_{\min}$ and $L_{\max}$.
Output: Clustered sets of users with the number of clusters in the range of $L_{\min}$ to $L_{\max}$.
Candidates = {}
for L in range $L_{\min}$ to $L_{\max}$ **do**
  Cluster users based on Algorithm 5 for L clusters.
  Set $\boldsymbol{\omega} = [1/L, 1/L, \ldots, 1/L]_{1\times L}$
  $S_L = h(p, \boldsymbol{\omega})$
  Add the clustering with its sum of power allocations S_L to Candidates.
end for
Best clustering = clustering in Candidates with the minimum S_L.
return Best clustering.

$L_{\min}$ and $L_{\max}$, respectively. In general, when the number of clusters decreases, more resource can be allocated to each cluster. On the other hand, as the number of users in each cluster increases, each cluster needs more power to achieve a target success probability. The computational complexity and latency of SIC also increase for a larger cluster. In this case, we employ Algorithm 5 to search over all numbers of clusters L in the range $\{L_{\min}, L_{\min}+1, \ldots, L_{\max}\}$. For each value of L, we cluster the users according to Algorithm 5 and by assuming a target common success probability (such as $p = 0.95$) and equal resource allocation ($\boldsymbol{\omega} = [1/L, 1/L, \ldots, 1/L]$), we derive the sum of power allocation factors of the clusters according to the closed-form expression $h(p, \boldsymbol{\omega})$ given in (28). Then, we choose the best clustering that results in the minimum sum of power allocation factors for all clusters.

Algorithm 6 provides pseudo-code for this clustering scheme. Since this algorithm runs Algorithm 5 in each iteration, its computational complexity is proportional to

$$(L_{\max} - L_{\min}) \left[(K - L_{avg}) L_{avg} \right], \tag{40}$$

where $L_{avg} = [L_{\min} + L_{\max}]/2$. Thus, the computational complexity of this algorithm still increases polynomially in time with the number of users K and the number of clusters L.

VI. THE COMPLETE USER CLUSTERING, POWER AND RESOURCE ALLOCATION SCHEME

In previous sections we developed and presented user clustering algorithms, inter-cluster power and resource allocation schemes, and intra-cluster power allocation and decoding order selection separately. In this section, we combine them in a unified procedure that can be implemented at the BS to organize users into clusters, and allocate power and resource to guarantee fairness among users. Recall that we require the statistical CSI, which contains mean and variance of SNR of users be reported to the BS via feedback channels once in

every coherence time interval. The user clustering algorithm and resource allocations can have separate update intervals. For instance, if the resource allocation update interval is T, then clustering can have an update interval of kT to reduce the computational complexity.

In all calculations we assume that all the rates and SNRs of the users are normalized according to the total available power P_T and total resource W_T. Thus, if a user reports $\bar{\psi}_{v_{\ell,i}}$ and $\sigma^2_{\psi_{v_{\ell,i}}}$ which are the mean and variance of its SNR, normalized according to $\delta_\ell P_T$ and $\omega_\ell W_T$ of its cluster, then the BS should replace them with $\bar{\gamma}_{v_{\ell,i}}$ and $\sigma^2_{\gamma_{v_{\ell,i}}}$, respectively, which according to (4) can be derived as

$$\bar{\gamma}_{v_{\ell,i}} = (\omega_\ell/\delta_\ell)\bar{\psi}_{v_{\ell,i}}, \tag{41}$$

$$\sigma^2_{\gamma_{v_{\ell,i}}} = (\omega_\ell/\delta_\ell)^2\sigma^2_{\psi_{v_{\ell,i}}}. \tag{42}$$

Likewise, for the downlink rates of users, the BS has to normalize them according to the total available resource W_T.

In Section III, we derived the optimal intra-cluster decoding order and power allocation factor of users assuming that the total power P_T and resource W_T of the transmitter are allocated to cluster C_ℓ ($\delta_\ell = 1$ and $\omega_\ell = 1$) in (18) and (19), respectively. To extend those results to the general case that δ_ℓ and ω_ℓ are not necessarily equal to one, we need to replace the rate r with r/ω_ℓ and the mean of SNR $\bar{\gamma}_{\pi_{\ell,k}}$ with $(\delta_\ell/\omega_\ell)\bar{\gamma}_{\pi_{\ell,k}}$ in the definition of parameter $\beta_{\pi_{\ell,k}}$ in (17) and the intra-cluster power allocation factors in (19). Thus, the optimal decoding order is based on the ascending order of parameter $\beta_{\pi_{\ell,k}}$, which is defined as

$$\beta_{\pi_{\ell,k}} = \frac{\delta_\ell\bar{\gamma}_{\pi_{\ell,k}}}{\omega_\ell m_{\pi_{\ell,k}}} Q^{-1}(m_{\pi_{\ell,k}}, p_\ell), \quad k \in \mathcal{I}_\ell. \tag{43}$$

However, since δ_ℓ and ω_ℓ do not change for users inside each cluster, deriving the optimal decoding order based on (17) or (43) gives the same result. Since (17) is more compact, we shall always use it for selecting the optimal decoding order.

Performing variable replacements in (19) for the generalized intra-cluster power allocation factors, we obtain

$$\alpha_{\pi_{\ell,|C_\ell|}} = \frac{(2^{r_{\pi_{\ell,|C_\ell|}}/\omega_\ell} - 1)}{\delta_\ell/\omega_\ell} \times \frac{m_{\pi_{\ell,|C_\ell|}}}{\bar{\gamma}_{\pi_{\ell,|C_\ell|}} Q^{-1}(m_{\pi_{\ell,|C_\ell|}}, p_\ell)}, \tag{44a}$$

$$\alpha_{\pi_{\ell,|C_\ell|-1}} = \frac{(2^{r_{\pi_{\ell,|C_\ell|-1}}/\omega_\ell} - 1)}{\delta_\ell/\omega_\ell} \left(\frac{m_{\pi_{\ell,|C_\ell|}}(2^{r_{\pi_{\ell,|C_\ell|}}/\omega_\ell} - 1)}{\bar{\gamma}_{\pi_{\ell,|C_\ell|}} Q^{-1}(m_{\pi_{\ell,|C_\ell|}}, p_\ell)} + \frac{m_{\pi_{\ell,|C_\ell|-1}}}{\bar{\gamma}_{\pi_{\ell,|C_\ell|-1}} Q^{-1}(m_{\pi_{\ell,|C_\ell|-1}}, p_\ell)} \right), \tag{44b}$$

$$\alpha_{\pi_{\ell,k}} = \frac{(2^{r_{\pi_{\ell,k}}/\omega_\ell} - 1)}{\delta_\ell/\omega_\ell} \times \left(\sum_{i=0}^{|C_\ell|-k-2} \left(\frac{m_{\pi_{\ell,|C_\ell|-i}}(2^{r_{\pi_{\ell,|C_\ell|-i}}/\omega_\ell} - 1)2^{\sum_{j=i+1}^{|C_\ell|-k-1} r_{\pi_{\ell,|C_\ell|-j}}/\omega_\ell}}{\bar{\gamma}_{\pi_{\ell,|C_\ell|-i}} Q^{-1}(m_{\pi_{\ell,|C_\ell|-i}}, p_\ell)} \right) + \frac{m_{\pi_{\ell,k+1}}(2^{r_{\pi_{\ell,k+1}}/\omega_\ell} - 1)}{\bar{\gamma}_{\pi_{\ell,k+1}} Q^{-1}(m_{\pi_{\ell,k+1}}, p_\ell)} + \frac{m_{\pi_{\ell,k}}}{\bar{\gamma}_{\pi_{\ell,k}} Q^{-1}(m_{\pi_{\ell,k}}, p_\ell)} \right), \quad 1 \le k \le |C_\ell| - 2. \tag{44c}$$

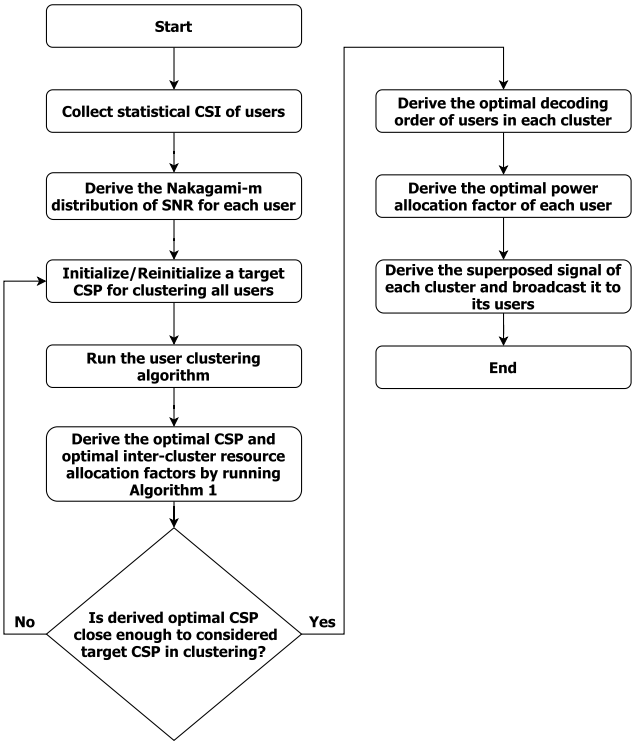

FIGURE 1. Flowchart of the complete proposed scheme for user clustering, power and resource allocation in the base station.

Finally, the complete procedure for user clustering, power and resource allocation is summarized in the flowchart of Figure 1 and elaborated further below.

1) Obtain the means and variances of SNRs of all users from the feedback channels.
2) Calculate the shape factor m of Nakagami-m fading channels for all users according to (6).
3) Initialize/Reinitialize a target common success probability (CSP) for user clustering algorithm.
4) Based on the predefined assumption about cluster size and total number of clusters (i.e., being fixed or variable) run one of Algorithms 4, 5 or 6 to cluster the users.
5) Run Algorithm 1 to obtain the optimal CSP p and optimal inter-cluster resource allocation factor ω_ℓ of all clusters (Algorithm 1 will call for Algorithms 2 and 3 inside itself).
6) If the obtained optimal CSP in Step 5 is good enough (e.g. the absolute difference is less than 0.05) as compared to the initial value of CSP considered,

continue to Step 7. Otherwise, go to Step 3 and reinitialize the CSP with the obtained CSP in Step 5.
7) Derive the optimal decoding order for users of each cluster based on the ascending order of parameter $\beta_{\pi_{\ell,k}}$ defined in (17).
8) Use equation (27) to compute the optimal inter-cluster power allocation factor δ_ℓ of each cluster and obtain the optimal intra-cluster power allocation factor of each user $\alpha_{\pi_{\ell,k}}$ according to (44). Then, the value $\delta_\ell \alpha_{\pi_{\ell,k}}$ is the proportion of the total power P_T that has been allocated to the kth user in the optimal decoding order of the ℓth cluster.
9) Obtain the signal to be transmitted to each cluster by superposition coding according to (2) and send it to the users of that cluster.

Note that each user has to perform SIC to obtain its own signal. If the BS follows the above procedure, fairness among the users will be guaranteed in terms of the outage or success probability of users, i.e., the minimum success probability among them will be maximized.

It is pointed out that according to (2), the BS does not need the values of the optimal inter-cluster power allocation factor δ_ℓ and optimal intra-cluster power allocation factor $\alpha_{\pi_{\ell,k}}$ separately to form the superimposed signal for each cluster. It only needs their product $\delta_\ell \alpha_{\pi_{\ell,k}}$, which specifies the proportion of the total power P_T that should be allocated to user $U_{\pi_{\ell,k}}$ and it can be derived directly from (44) by moving δ_ℓ to the other side of the equation. However, we obtain them separately to keep the logical flow, improve the modularity and readability of the paper, and also to emphasize the fact that the closed-form expression for the inter-cluster power allocation factor δ_ℓ can be used as a cost metric for user clustering algorithms.

A. COMPUTATIONAL COMPLEXITY ANALYSIS

To complete Section VI, we analyze the computational complexity of our proposed scheme for user clustering, power and resource allocation. It's noteworthy that the main loop of the proposed scheme for iterating over multiple initial target common success probabilities (CSPs) only affects performance of clustering algorithms, since the power and resource allocation algorithms establish fairness among the users for any given clustering. Besides, in Section VII-D, we show that without iterating over this loop and only with a fixed initial target CSP such as $p = 0.95$, our proposed scheme outperforms existing works. However, if the computing power at the BS and latency constraints of the system are flexible, performing a few iterations (less than 5) over the main loop will decrease the gap between the initial target CSP and the optimal CSP. Consequently, that results in a better performance of user clustering algorithm and in increasing the value of the optimal CSP (see Section VII-B for more details). Therefore, we only analyze the computational complexity of one iteration of the complete proposed scheme as depicted in flowchart of Figure 1.

The first step in the proposed scheme acquires the statistical CSI of users and should be done periodically once in the coherence time interval of the channels. If a user fails to send CSI feedback to the BS in the coherence time interval, it can be omitted from the set of users or served with the previously reported CSI (which may be outdated). Nevertheless, incorporating these details is out of scope of this paper. We assume that there are K users that have reported their statistical CSI to the BS and we derive efficient algorithms to cluster these users and allocate power and channel resources to them such that the minimum success probability among them is maximized. On the other hand, requiring only the statistical CSI is the most practical assumption as it has the minimum signaling overhead compared to other assumptions, especially the assumption of having perfect instantaneous CSI at the BS as considered in many other papers (see Table 1). Thus, we skip the computational complexity of collecting the statistical CSI of users, which can be performed periodically over the feedback channels.

Since, the derived equations for the Nakagami-m distribution of the SNR of users are in closed-form and initializing the target CSP is a constant parameter selection, they can be ignored in computational complexity of the proposed scheme. However, for the next major step which is clustering the users, one of Algorithms 4, 5 and 6 should be used. We showed that the computational complexity of these algorithms increase polynomially in time with increasing numbers of users and clusters. If the total number of users is K, then none of these clustering algorithms requires more than $O(K^2)$ iterations to perform the clustering. Thus, we consider $O(K^2)$ to be the computational complexity of the clustering step.

The next step is to find the optimal CSP and resource allocation factors according to Algorithm 1. As discussed in the last paragraph of Section IV, by considering the acceptable error in finding all the parameters to be ϵ, the computational complexity of Algorithm 1 is $L\left[\log_2(1/\epsilon)\right]^3$ which increases linearly in time with increasing number of clusters L. All the remaining steps of the scheme are to calculate some parameters such as decoding order and power allocation factors according to closed-form expressions. Thus, their computational complexity is negligible. Therefore, the overall computational complexity of our proposed scheme is proportional to

$$K^2 + L\left[\log_2(1/\epsilon)\right]^3, \tag{45}$$

operations. On the other hand, using an exhaustive search method for finding the K optimal user power allocation factors and L cluster resource allocation factors with precision ϵ requires investigating $\left(\frac{1}{\epsilon}\right)^{K+L}$ states that increases exponentially in time with the number of users K and number of clusters L. Moreover, considering all the possible clustering and decoding orders of users with fixed $N = K/L$ users in each cluster, the number of states in exhaustive

search is

$$\left(\frac{1}{\epsilon}\right)^{K+L} \times \frac{K!}{L!}, \tag{46}$$

which increases exponentially in time with the numbers of users and clusters. Hence, our proposed scheme significantly decreases the computational complexity of solving the problem. We will also evaluate the run time of the complete proposed scheme by simulations in section VII-D.

VII. SIMULATION RESULTS

In this section, performance of the proposed algorithms is evaluated by simulations and compared to those of existing algorithms. All simulations were executed on a laptop with Intel(R) Core(TM) i5-5200U CPU2.20 GHz and 8 GB of RAM.

A. PERFORMANCE OF POWER AND RESOURCE ALLOCATION SCHEME

In this subsection we investigate performance of our proposed scheme for power and resource allocation and compare it to the following power and resource allocation schemes:

1) *Equal allocation:* Power and resource are allocated equally to all clusters.
2) *Proportional allocation:* Power and resource are allocated to each cluster proportional to the ratio of the number of users in that cluster to the total number of users.
3) *Method of [24]:* Power is allocated to users according to the distributed power control method proposed in [24] (for more details see Equations (25), (26) and (28) in [24]). However, since no resource allocation scheme is proposed in that paper we use a proportional resource allocation scheme in this case.

For the first two of these inter-cluster power and resource allocation schemes, we employ our proposed intra-cluster power allocation to maximize the minimum success probability of users inside each cluster separately. However, for the third scheme we use the power allocation method proposed in [24]. The main goal of our proposed scheme is to establish fairness among all the users. Thus, we first compare performance of these schemes using Jain's index [25] in terms of the success probability of the users. This metric has been adopted in many works (e.g. [11], [26]) to evaluate fairness among users. The Jain's index for the success probability of K users is defined as follows:

$$\mathcal{J}(p_1, p_2, \ldots, p_K) = \frac{\left(\sum_{i=1}^{K} p_i\right)^2}{K \sum_{i=1}^{K} p_i^2} = \frac{\bar{\boldsymbol{p}}^2}{\bar{\boldsymbol{p}^2}}. \tag{47}$$

If the success probabilities of all the users are equal, then Jain's index is maximum and equal to one. In the worst case, where all the success probabilities are zero, except for one user, the index is minimum and equal to $1/K$.

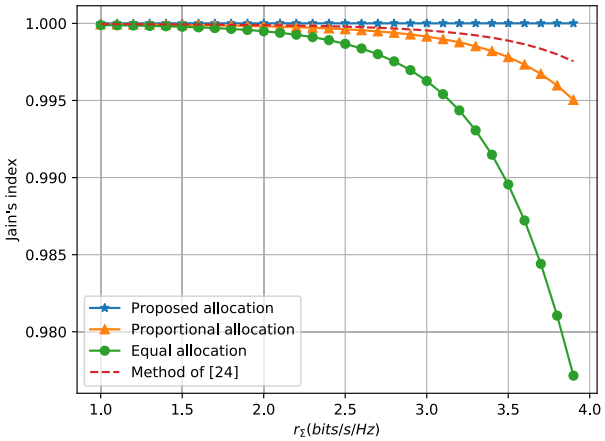

FIGURE 2. Jain's index versus the sum rate of users.

Figure 2 plots Jain's index versus the sum rate r_Σ by averaging the results over 100 simulation runs. For each simulation, five NOMA clusters are considered with a random number of users $N \in \{2, 3, 4\}$ in each cluster, random statistical CSI and rate for the users. Parameters m, $\bar{\gamma}$ and r of users are randomly generated from the following intervals with uniform distribution:

$$m \in (1, 3), \quad \bar{\gamma} \in (1, 3), \; r \in (0.01, 0.1). \tag{48}$$

Then for having different values of the sum rate, all the users' rates are multiplied by a proper constant factor. It is clear from Figure 2 that as r_Σ increases, the performance of our proposed scheme stays the same and fairness is established among all users. However for the other schemes, Jain's index quickly decreases as r_Σ increases.

Recall that the goal of our proposed scheme is to maximize the minimum success probability of users. Thus, we also compare the minimum success probability of the users among these schemes in the same simulations that we perform for Jain's index, and the results are plotted in Figure 3. The results show that by establishing fairness among all users in our proposed algorithm, the minimum success probability of users is significantly improved when compared to that of the other power and resource allocation schemes.

B. IMPACT OF ITERATION OVER MULTIPLE INITIAL CSP VALUES

In this section, we investigate the impact of the initial CSP value on the performance of our proposed scheme. To this end, 30 users are generated with random parameters as explained before and run Algorithms 4, 5 and 6 separately. We perform 8 iterations over the loop of the proposed scheme and reinitialize the CSP value of the clustering with the optimal CSP obtained in the last iteration as described in the flowchart of Figure 1.

Figure 4 plots the averages of the optimal CSP values over 100 simulation runs versus the number of iterations for different clustering algorithms. We set the first "initial CSP" value to be 0.8 (i.e., in iteration 0). It is clear that as the

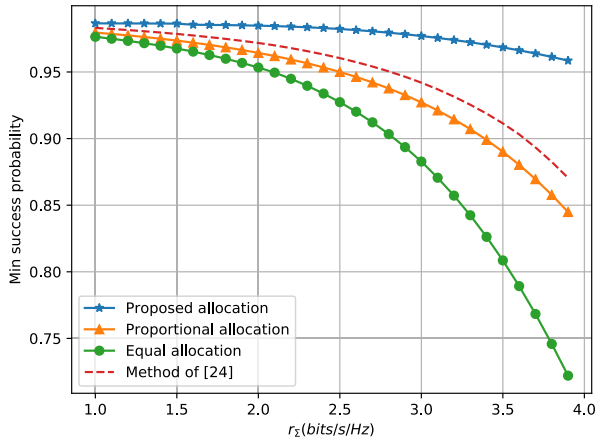

FIGURE 3. Minimum success probability of users versus their sum rate.

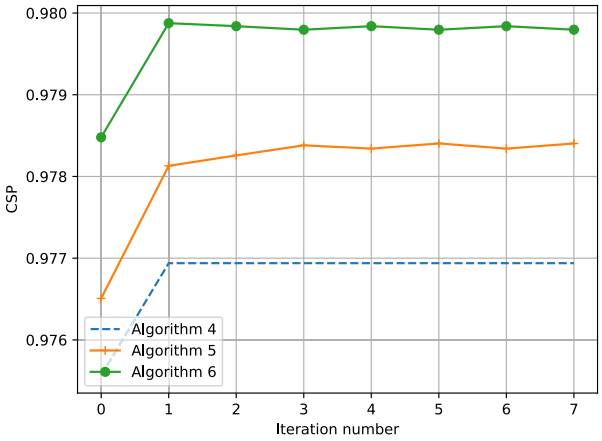

FIGURE 4. Optimal CSP of users achieved by the complete proposed scheme versus the iteration number.

initial CSP value gets closer to the optimal CSP value, the clustering algorithm performs better and the optimal CSP of users increases. This is because changing the initial CSP value affects both the optimal decoding order and δ_ℓ values for each cluster. Thus, by selecting an initial value of the CSP closer to the optimal CSP, the clustering algorithm determines the power demand and optimal decoding order of each cluster more accurately. In addition, this simulation shows that the proposed scheme converges very quickly, only after a few iterations, to the optimal CSP of users. Thus, to keep the computational complexity of our proposed scheme as low as possible, in the next two subsections we only consider a predefined CSP value of 0.95 and show that even without iterating over multiple CSPs, our proposed scheme still outperforms existing schemes.

C. PERFORMANCE OF USER CLUSTERING ALGORITHMS

In this section we evaluate the performance of Algorithms 4, 5 and 6 in terms of the minimum success probability of users. In addition to our proposed algorithms, we also consider two other algorithms for comparison. The

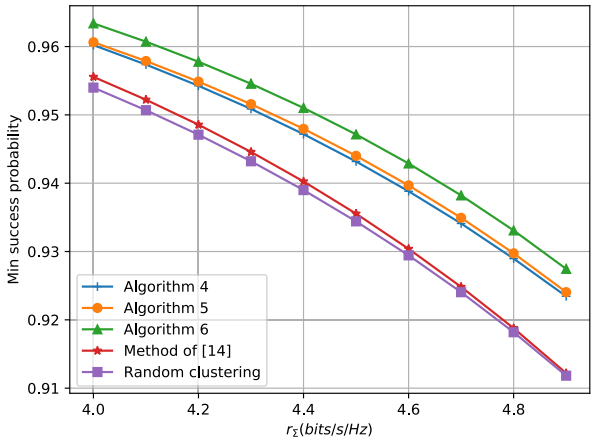

FIGURE 5. Minimum success probability of users versus their sum rate for different clustering methods and with our proposed power and resource allocation scheme.

first one is random user clustering, which does not utilize the statistical CSI of users for clustering and represents a lower bound of performance for other user clustering algorithms. In the simulation results this algorithm is labeled as "Random clustering". The second algorithm is the method proposed in [14], which is designed to cluster users into two-user clusters. In that method, users are sorted based on their average SNRs. Then, the first and last users are paired together, the second user and the one before the last user are paired, and in general, the kth user is paired with the $K - k + 1$th user, where the total number of users K is assumed to be even (refer to Theorem 3 in [14] for more details). In the simulation results this method is labeled as "Method of [14]". In order to focus on the impact of user clustering algorithms on the performance, we implement our proposed power and resource allocation scheme for all the aforementioned clustering algorithms.

As before, here we also consider 30 random users and scale up/down their sum rate by multiplying all the rates by a constant scale factor. Figure 5 plots the minimum success probability of the users against their sum rate. It is clear that our proposed algorithms outperform the two other reference algorithms. By comparing our proposed clustering algorithms, it can be seen that as we relax the constraints on the number of users in each cluster and the total number of clusters, the performance of the clustering algorithm improves. This is expected as a higher degree of freedom should help to form a better clustering structure.

D. PERFORMANCE OF THE COMPLETE PROPOSED SCHEME

In this section, we evaluate performance of our complete proposed scheme for user clustering and resource allocation and we compare it with the following reference methods:

1) Clustering method of [14] + equal resource allocation + power allocation of [24]
2) Random clustering + equal resource allocation + power allocation of [24]

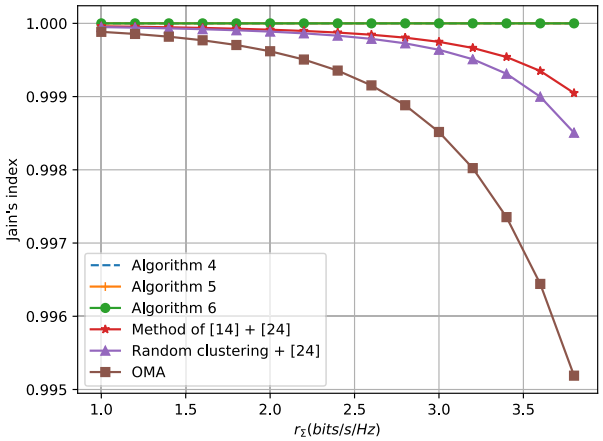

FIGURE 6. Jain's index versus the sum rate for different clustering algorithms.

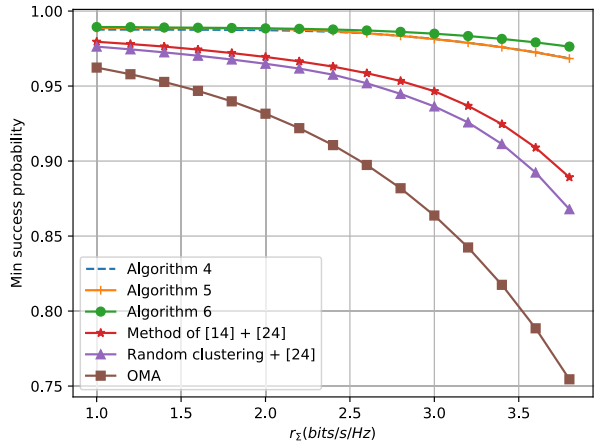

FIGURE 7. Minimum success probability of users versus the sum rate for different clustering algorithms.

3) OMA technique such as TDMA + our proposed power and resource allocation

It's noteworthy that our proposed scheme is capable of allocating resources to singleton clusters, which consist of only one user. By considering all the clusters to be singleton, our proposed hybrid NOMA-OMA scheme simplifies to the OMA since all the users will use orthogonal resources in that case. Thus, in the third reference method we consider singleton clusters to clarify the superiority of the hybrid NOMA-OMA scheme in comparison to the pure OMA. In the simulations, we compare the Jain's index, minimum success probability and sum throughput of users for the proposed and reference algorithms. The simulations are repeated 100 times and the averages of the obtained results are plotted in Figures 6, 7 and 8. At each repetition, $K = 30$ users with random CSI and rate parameters are simulated. Then, for each value of r_Σ, the rates of all users are scaled up or down with a proper constant. It is clear that as r_Σ increases, our proposed schemes outperform other reference algorithms in all the considered performance metrics. It's notable that the curves of our proposed clustering algorithms are very close to one another and they appear overlapped.

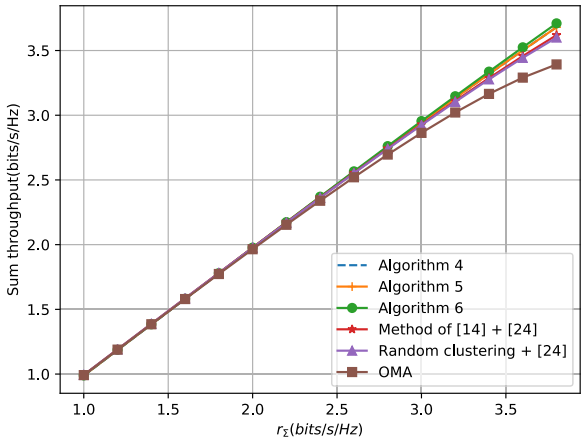

FIGURE 8. Sum throughput of users versus the sum rate for different clustering algorithms.

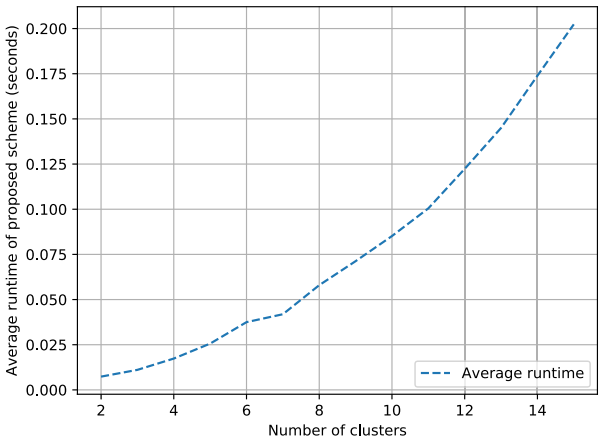

FIGURE 9. Average runtime of the proposed scheme versus the number of clusters.

Figure 9 depicts the runtime of our proposed scheme versus the number of clusters. From this figure, it is seen that the computational complexity of our proposed scheme almost linearly increases with the number of clusters, which is consistent with the complexity analysis given in Section VI-A. In this simulation for different number of random users, the proposed scheme is repeated 500 times and the average runtime of the whole scheme is calculated. The precision of calculating parameters such as resource allocation factors is set to $\epsilon = 10^{-3}$. It's noteworthy that we implemented the scheme in a single-thread mode. However, utilizing parallelism and multi-threading is possible for the implementation of binary searches of Algorithms 1, 2 and 3, which should reduce the runtime of the proposed scheme.

VIII. CONCLUSION

In this paper, we have tackled the problem of optimizing user clustering, power allocation to users, resource (time slot or bandwidth) allocation to clusters, and decoding order in each cluster for the downlink of a hybrid NOMA-OMA

system operating over Nakagami-m fading channels. In a hybrid NOMA-OMA system, users are organized into several clusters, where clusters use an orthogonal multiple access scheme to utilize channel resources while users in each cluster employ power-domain NOMA. The goal was to maximize the minimum success probability (or equivalently minimize the maximum outage probability) among all users. We first proved that at the optimal solution of the problem, all the users have a common success probability (CSP). We then proposed an efficient algorithm for finding the optimal CSP and resource allocation factors of clusters simultaneously. We also derived the inter-cluster power allocation factor for each cluster, intra-cluster power allocation factor for each user, and optimal decoding order of users inside each cluster in a closed-form expression based on the CSP, statistical CSI of users and resource allocation factor of each cluster. We proposed efficient algorithms for user clustering under three different scenarios where the number of users in each cluster and/or the total number of clusters are fixed or variable. All three algorithms were developed based on the same principle of minimizing the power consumption of each cluster while achieving a given target success probability. Simulation results show that our proposed schemes for user clustering, power and resource allocation outperform existing schemes not only in terms of fairness and the minimum success probability of users, but also in terms of the sum throughput. An interesting topic for a future work is to develop efficient user clustering and resource allocation methods for the uplink of a NOMA system operating over Nakagami-m fading channels in order to guarantee fairness among users.

APPENDIX A
PROOF OF THEOREM 1

Before proving Theorem 1, we restate the recursive equations for the power allocation factors of users, given in Equations (26) of [19]. Assuming that there are $|C_\ell| = N$ users in a NOMA cluster, their power allocation factors can be derived recursively as

$$\alpha_{\pi_{\ell,N}} = \zeta_{\pi_{\ell,N}} \times \frac{1}{\beta_{\pi_{\ell,N}}}, \tag{49a}$$

$$\alpha_{\pi_{\ell,N-1}} = \zeta_{\pi_{\ell,N-1}} \left(\alpha_{\pi_{\ell,N}} + \frac{1}{\beta_{\pi_{\ell,N-1}}} \right), \tag{49b}$$

$$\vdots$$

$$\alpha_{\pi_{\ell,N-i}} = \zeta_{\pi_{\ell,N-i}} \left(\alpha_{\pi_{\ell,N}} + \cdots + \alpha_{\pi_{\ell,N-i+1}} + \frac{1}{\beta_{\pi_{\ell,N-i}}} \right), \tag{49c}$$

$$\vdots$$

$$\alpha_{\pi_{\ell,1}} = \zeta_{\pi_{\ell,1}} \left(\alpha_{\pi_{\ell,N}} + \cdots + \alpha_{\pi_{\ell,2}} + \frac{1}{\beta_{\pi_{\ell,1}}} \right). \tag{49d}$$

The closed-form expressions for power allocation factors in (19) are derived based on these recursive equations.

We prove the theorem by induction. As the base of induction for $N = 2$ users we should prove

$$\sqrt{\frac{\alpha_{\pi_{\ell,1}}}{\zeta_{\pi_{\ell,1}}}} > \sqrt{\frac{\alpha_{\pi_{\ell,2}}}{\zeta_{\pi_{\ell,2}}}} \left(\sqrt{\zeta_{\pi_{\ell,2}} + 1} - 1 \right). \tag{50}$$

Since, both sides of the above inequality are non-negative we can raise them to the power of two. In addition, if we replace $\alpha_{\pi_{\ell,1}}$ according to (49b), and perform some algebraic manipulations, (50) converts to

$$\frac{\beta_{\pi_{\ell,2}}}{\beta_{\pi_{\ell,1}}} + 2\sqrt{\zeta_{\pi_{\ell,2}} + 1} > 2, \tag{51}$$

which is always true, since according to the optimal decoding order in (18) we know $\beta_{\pi_{\ell,2}}/\beta_{\pi_{\ell,1}} \geq 1$ and $\zeta_{\pi_{\ell,2}} \geq 0$. Then, assuming that for any number of users $|C_\ell| = N$ the inequalities in (21) hold, we prove that they are also true for $|C_\ell| = N + 1$. Thus, for $i = 1$ we have to prove

$$\sqrt{\frac{\alpha_{\pi_{\ell,1}}}{\zeta_{\pi_{\ell,1}}}} > \sum_{j=2}^{N+1} \sqrt{\frac{\alpha_{\pi_{\ell,j}}}{\zeta_{\pi_{\ell,j}}}} \left(\sqrt{\zeta_{\pi_{\ell,j}} + 1} - 1 \right). \tag{52}$$

According to the assumption of the induction, for the last N users in the optimal decoding order, i.e., for $U_{\pi_{\ell,2}}, U_{\pi_{\ell,3}}, \ldots, U_{\pi_{\ell,N+1}}$ we have

$$\sqrt{\frac{\alpha_{\pi_{\ell,2}}}{\zeta_{\pi_{\ell,2}}}} > \sum_{j=3}^{N+1} \sqrt{\frac{\alpha_{\pi_{\ell,j}}}{\zeta_{\pi_{\ell,j}}}} \left(\sqrt{\zeta_{\pi_{\ell,j}} + 1} - 1 \right). \tag{53}$$

Therefore, for the right hand side of (52) we have

$$\sum_{j=2}^{N+1} \sqrt{\frac{\alpha_{\pi_{\ell,j}}}{\zeta_{\pi_{\ell,j}}}} \left(\sqrt{\zeta_{\pi_{\ell,j}} + 1} - 1 \right) < \sqrt{\frac{\alpha_{\pi_{\ell,2}}}{\zeta_{\pi_{\ell,2}}}} \times \sqrt{\zeta_{\pi_{\ell,2}} + 1}. \tag{54}$$

Hence, by combining (52) and (54), we can complete the proof by showing that

$$\sqrt{\frac{\alpha_{\pi_{\ell,2}}}{\zeta_{\pi_{\ell,2}}}} \times \sqrt{\zeta_{\pi_{\ell,2}} + 1} < \sqrt{\frac{\alpha_{\pi_{\ell,1}}}{\zeta_{\pi_{\ell,1}}}}. \tag{55}$$

On the other hand, according to (49) we have

$$\frac{\alpha_{\pi_{\ell,1}}}{\zeta_{\pi_{\ell,1}}} = \frac{\alpha_{\pi_{\ell,2}}}{\zeta_{\pi_{\ell,2}}} + \alpha_{\pi_{\ell,2}} + \frac{1}{\beta_{\pi_{\ell,1}}} - \frac{1}{\beta_{\pi_{\ell,2}}}. \tag{56}$$

By using (56) the inequality (55) reduces to

$$\frac{1}{\beta_{\pi_{\ell,2}}} < \frac{1}{\beta_{\pi_{\ell,1}}}, \tag{57}$$

which is always true according to the optimal decoding order condition (18). Proving the conditions (21) for other values of $i = 2, \ldots, N$ is straightforward by following the same method. Therefore, the proof of theorem is complete.

Algorithm 7 Finding the Optimal CSP p_ℓ of Users in Cluster C_ℓ

Input: $m_{\pi_i}, \bar{\gamma}_{\pi_i}, r_{\pi_i} \forall i \in \mathcal{I}_\ell$ and ϵ.
Output: p_ℓ such that $S(p_\ell, \boldsymbol{\pi}_\ell) = 1$ (see (23)).
Initialization:
$p_l = 0, p_u = 1$
while $p_u - p_l > \epsilon$ **do**
  $p = \frac{p_l + p_u}{2}$
  $\boldsymbol{\pi}_\ell$ = sorted indices of users based on parameter β_{π_k} in the ascending order according to (18)
  if $S(p, \boldsymbol{\pi}_\ell) < 1$ **then**
    $p_l = p$
  else if $S(p, \boldsymbol{\pi}_\ell) > 1$ **then**
    $p_u = p$
  else
    return $p_\ell = p$
  end if
end while
return $p_\ell = \frac{p_l + p_u}{2}$.

APPENDIX B
ALGORITHM FOR FINDING THE OPTIMAL CSP OF USERS FOR ONE CLUSTER

In Algorithm 7 we recall the algorithm proposed in [19] as a reference to facilitate comparison with its extended version developed in this paper, namely Algorithm 1. Algorithm 7 is designed to find the optimal CSP of K users when all of them are grouped into one cluster. In this algorithm, ϵ is the precision of calculating the common success probability (CSP). Algorithm 1 finds both the optimal CSP and optimal inter-cluster resource allocation factors across clusters when users are grouped into several clusters.

APPENDIX C
PROOF OF LEMMA 1

In [10] a similar lemma is proved for the case of Rayleigh fading. For the case of Nakagami-m fading, the success probability is derived in (15) as

$$p_{\pi_{\ell,k}} = Q\left(m_{\pi_{\ell,k}}, m_{\pi_{\ell,k}} \gamma_{th}^{\pi_{\ell,k}} / \bar{\gamma}_{\pi_{\ell,k}}\right). \tag{58}$$

The function $Q(\cdot,\cdot)$ is a strictly decreasing function of the second parameter and according to (13) $\gamma_{th}^{\pi_{\ell,k}}$ is a strictly decreasing function of $\alpha_{\pi_{\ell,k}}$. Thus, the success probability $p_{\pi_{\ell,k}}$ is a strictly increasing function of power allocation factor $\alpha_{\pi_{\ell,k}}$. The lemma can be proved by contradiction. Suppose that at the optimal solution, the success probabilities of all clusters are not the same. Thus, some clusters have the minimum success probability. Denote those clusters by $\mathcal{C}' = \arg\min_{\ell \in \mathcal{C}} p_\ell$ and the rest of clusters by $\mathcal{C}''$. For these subsets we have $\mathcal{C}' \cup \mathcal{C}'' = \mathcal{C}$ and $\mathcal{C}' \cap \mathcal{C}'' = \emptyset$. Based on the fact that the success probability function of each cluster is a strictly increasing function of the power allocation factor of that cluster, we can find an appropriate positive value ϵ such that by subtracting ϵ from all the power allocation factors of clusters in $\mathcal{C}''$, and adding $\frac{\epsilon|\mathcal{C}''|}{|\mathcal{C}'|}$ to the power allocation factors of clusters in $\mathcal{C}'$, the minimum success probability of the clusters can be increased, which contradicts the optimality of the solution. This proves the lemma.

APPENDIX D
PROOF OF LEMMA 3

To prove Lemma 3, we know that the objective function and all the constraints of problem (30) are linear, except for (30b). Thus, to prove the convexity, it suffices to prove that in the constraint (30b), $h(p, \boldsymbol{\omega})$ is a convex function. To this end, we investigate the positiveness of the second order derivatives of $h(\cdot,\cdot)$ with respect to its parameters. To derive $h''_{\omega_\ell}(p, \boldsymbol{\omega}) = \frac{\partial^2 h(p,\boldsymbol{\omega})}{\partial \omega_\ell^2}$ we use the parameter:

$$\beta_{\pi_{\ell,i}} = \frac{\bar{\gamma}_{\pi_{\ell,i}} Q^{-1}(m_{\pi_{\ell,i}}, p)}{m_{\pi_{\ell,i}}}, \tag{59}$$

to simplify the algebraic relations and rewrite $h(p, \boldsymbol{\omega})$ as

$$h(p, \boldsymbol{\omega}) = \sum_{\ell \in \mathcal{C}} \left(\frac{\omega_\ell (2^{r_{\pi_{\ell,1}}/\omega_\ell} - 1)}{\beta_{\pi_{\ell,1}}} + \sum_{i=2}^{|C_\ell|} \frac{\omega_\ell 2^{\sum_{j=1}^{i} r_{\pi_{\ell,j}}/\omega_\ell} - \omega_\ell 2^{\sum_{j=1}^{i-1} r_{\pi_{\ell,j}}/\omega_\ell}}{\beta_{\pi_{\ell,i}}} \right). \tag{60}$$

Then we have

$$h'_{\omega_\ell}(p, \boldsymbol{\omega}) = \frac{\partial h(p, \boldsymbol{\omega})}{\partial \omega_\ell} = \frac{2^{r_{\pi_{\ell,1}}/\omega_\ell} - 1 - (\ln 2) \frac{r_{\pi_{\ell,1}}}{\omega_\ell} 2^{r_{\pi_{\ell,1}}/\omega_\ell}}{\beta_{\pi_{\ell,1}}} \tag{61a}$$

$$+ \sum_{i=2}^{|C_\ell|} \frac{2^{\sum_{j=1}^{i} r_{\pi_{\ell,j}}/\omega_\ell} - (\ln 2) \left(\sum_{j=1}^{i} \frac{r_{\pi_{\ell,j}}}{\omega_\ell} \right) 2^{\sum_{j=1}^{i} r_{\pi_{\ell,j}}/\omega_\ell}}{\beta_{\pi_{\ell,i}}} \tag{61b}$$

$$- \sum_{i=2}^{|C_\ell|} \frac{2^{\sum_{j=1}^{i-1} r_{\pi_{\ell,j}}/\omega_\ell} + (\ln 2) \left(\sum_{j=1}^{i-1} \frac{r_{\pi_{\ell,j}}}{\omega_\ell} \right) 2^{\sum_{j=1}^{i-1} r_{\pi_{\ell,j}}/\omega_\ell}}{\beta_{\pi_{\ell,i}}}. \tag{61c}$$

After combining a part of fraction (61a) with the summation in (61b), changing variable i to $i+1$ and using the function $f(x) = x(\ln x - 1)$ to rewrite equation (61) the first order derivative can be obtained as

$$h'_{\omega_\ell}(p, \boldsymbol{\omega}) = \frac{-1}{\beta_{\pi_{\ell,1}}} - \sum_{i=1}^{|C_\ell|-1} \left(\left(\frac{1}{\beta_{\pi_{\ell,i}}} - \frac{1}{\beta_{\pi_{\ell,i+1}}} \right) \times f\left(2^{\sum_{j=1}^{i} r_{\pi_{\ell,j}}/\omega_\ell}\right) \right) - \frac{f\left(2^{\sum_{j=1}^{|C_\ell|} r_{\pi_{\ell,j}}/\omega_\ell}\right)}{\beta_{\pi_{\ell,|C_\ell|}}}. \tag{62}$$

Subsequently, by using $f'(x) = \ln x$ and some straightforward algebraic manipulations the second order derivative can be calculated as

$$h''_{\omega_\ell}(p, \boldsymbol{\omega}) = \frac{\partial^2 h(p, \boldsymbol{\omega})}{\partial \omega_\ell^2} = \sum_{i=1}^{|C_\ell|-1} \left[\left(\frac{1}{\beta_{\pi_{\ell,i}}} - \frac{1}{\beta_{\pi_{\ell,i+1}}} \right) \times \left(\ln \left(2^{\sum_{j=1}^{i} r_{\pi_{\ell,j}}/\omega_\ell} \right) \right)^2 \left(2^{\sum_{j=1}^{i} r_{\pi_{\ell,j}}/\omega_\ell} \right) / \omega_\ell \right] + \frac{1}{\beta_{\pi_{\ell,|C_\ell|}}} \left(\ln \left(2^{\sum_{j=1}^{|C_\ell|} r_{\pi_{\ell,j}}/\omega_\ell} \right) \right)^2 \left(2^{\sum_{j=1}^{|C_\ell|} r_{\pi_{\ell,j}}/\omega_\ell} \right) / \omega_\ell. \tag{63}$$

According to (18), at the optimal solution of the problem, for each cluster the decoding order is selected such that:

$$\beta_{\pi_{\ell,1}} \leq \beta_{\pi_{\ell,2}} \leq \cdots \leq \beta_{\pi_{\ell,|C_\ell|}}, \quad \ell \in \mathcal{C}. \tag{64}$$

Thus, it is straightforward to verify that the following inequality always holds:

$$h''_{\omega_\ell}(p, \boldsymbol{\omega}) > 0, \quad \ell \in \mathcal{C}. \tag{65}$$

To calculate the second order derivative of $h(\cdot, \cdot)$ with respect to the first parameter p, we can rewrite it as

$$h(p, \boldsymbol{\omega}) = \sum_{\ell \in \mathcal{C}} \left(\frac{\eta_{\pi_{\ell,1}}}{Q^{-1}(m_{\pi_{\ell,1}}, p)} + \sum_{i=2}^{|C_\ell|} \frac{\eta_{\pi_{\ell,i}} 2^{\sum_{j=1}^{i-1} r_{\pi_{\ell,j}}/\omega_\ell}}{Q^{-1}(m_{\pi_{\ell,i}}, p)} \right), \tag{66}$$

where $\eta_{\pi_{\ell,i}}$ is defined to simplify the algebraic relations as

$$\eta_{\pi_{\ell,i}} = \frac{\omega_\ell m_{\pi_{\ell,i}} (2^{r_{\pi_{\ell,i}}/\omega_\ell} - 1)}{\bar{\gamma}_{\pi_{\ell,i}}}. \tag{67}$$

We see that $h(p, \boldsymbol{\omega})$ is comprised from a sum of sub-functions in the form of:

$$g_{\ell,i}(p) = \frac{\kappa_{\ell,i}}{Q^{-1}(m_{\pi_{\ell,i}}, p)}, \tag{68}$$

where $\kappa_{\ell,i}$ are some positive factors, defined as:

$$\kappa_{\ell,i} = \begin{cases} \eta_{\pi_{\ell,1}} & i = 1, \\ \eta_{\pi_{\ell,i}} 2^{\sum_{j=1}^{i-1} r_{\pi_{\ell,j}}/\omega_\ell} & 2 \leq i \leq |C_\ell|. \end{cases} \tag{69}$$

Thus, to evaluate the sign of $h''_p(p, \boldsymbol{\omega})$ which is the second order derivative of $h(\cdot, \cdot)$ with respect to the first parameter p, first we evaluate the sign of the derivatives of $g_{\ell,i}(p)$. If we denote $Q^{-1}(m_{\pi_{\ell,i}}, p) = x_{\ell,i}$, we have

$$Q^{-1}(m_{\pi_{\ell,i}}, p) = x_{\ell,i} \;\Rightarrow\; Q(m_{\pi_{\ell,i}}, x_{\ell,i}) = p, \tag{70}$$

and function $g_{\ell,i}(p)$ can be rewritten as

$$g_{\ell,i}(p) = \frac{\kappa_{\ell,i}}{x_{\ell,i}}. \tag{71}$$

For the second order derivative of this function with respect to p we have

$$g''_{\ell,i}(p) = \kappa_{\ell,i} \left[-\frac{\partial^2 x_{\ell,i}}{\partial p^2} x_{\ell,i}^{-2} + 2 \left(\frac{\partial x_{\ell,i}}{\partial p} \right)^2 x_{\ell,i}^{-3} \right]. \tag{72}$$

On the other hand, we know that:

$$\frac{\partial x_{\ell,i}}{\partial p} = \frac{\partial Q^{-1}(m_{\pi_{\ell,i}}, p)}{\partial p} = \frac{1}{\frac{\partial Q(m_{\pi_{\ell,i}}, x_{\ell,i})}{\partial x_{\ell,i}}}, \tag{73}$$

and

$$\frac{\partial Q(m_{\pi_{\ell,i}}, x_{\ell,i})}{\partial x_{\ell,i}} = \frac{-1}{\Gamma(m_{\pi_{\ell,i}})} x_{\ell,i}^{m_{\pi_{\ell,i}}-1} e^{-x_{\ell,i}}. \tag{74}$$

Thus, with some algebraic manipulations for the first and second order derivatives of $x_{\ell,i}$ with respect to p we have:

$$\frac{\partial x_{\ell,i}}{\partial p} = -\Gamma(m_{\pi_{\ell,i}}) e^{x_{\ell,i}} x_{\ell,i}^{-(m_{\pi_{\ell,i}}-1)}, \tag{75}$$

$$\frac{\partial^2 x_{\ell,i}}{\partial p^2} = \Gamma^2(m_{\pi_{\ell,i}}) e^{2x_{\ell,i}} x_{\ell,i}^{-2m_{\pi_{\ell,i}}+1} \left[x_{\ell,i} - (m_{\pi_{\ell,i}} - 1) \right]. \tag{76}$$

Consequently, by combining (72), (75) and (76) we have

$$g''_{\ell,i}(p) = \kappa_{\ell,i} \Gamma^2(m_{\pi_{\ell,i}}) e^{2x_{\ell,i}} x_{\ell,i}^{-2m_{\pi_{\ell,i}}-1} (1 - x_{\ell,i} + m_{\pi_{\ell,i}}). \tag{77}$$

Considering the second order derivative of $g_{\ell,i}(p)$ in (77) we see that if we can prove the following inequality

$$1 - x_{\ell,i} + m_{\pi_{\ell,i}} \geq 0, \tag{78}$$

then it is straightforward to prove that $g''_{\ell,i}(p)$ and $h''_p(p, \boldsymbol{\omega})$ are always positive, and hence proving the convexity of problem (30).

If we replace $x_{\ell,i}$ with its definition from (70) we have

$$1 - x_{\ell,i} + m_{\pi_{\ell,i}} \geq 0 \Rightarrow Q^{-1}(m_{\pi_{\ell,i}}, p) \leq m_{\pi_{\ell,i}} + 1 \overset{(*)}{\Rightarrow} p \geq Q(m_{\pi_{\ell,i}}, m_{\pi_{\ell,i}} + 1). \tag{79}$$

In $(*)$, we use the fact that the function $Q(a, b)$ is strictly decreasing with respect to its second parameter b. Besides, $Q(m, m+1)$ is a strictly increasing and bounded function of m, and its value is always less than 0.5. The plot of this function is shown in Figure 10.

The authors of [27] derive an approximation for the regularized lower incomplete gamma function $P(a + 1, a + \sqrt{2ay})$, which is also included in the NIST Digital Library of Mathematical Functions [28]:

$$P(a + 1, a + \sqrt{2ay}) = \frac{1}{2}\text{erfc}(-y) - \frac{1}{3}\sqrt{\frac{2}{\pi a}}(1 + y^2)e^{-y^2} + O(a^{-1}), \tag{80}$$

where the erfc(z) is defined as

$$\text{erfc}(z) = 1 - \text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-t^2} dt. \tag{81}$$

Moreover, the lower and upper regularized incomplete gamma functions are related to each other as

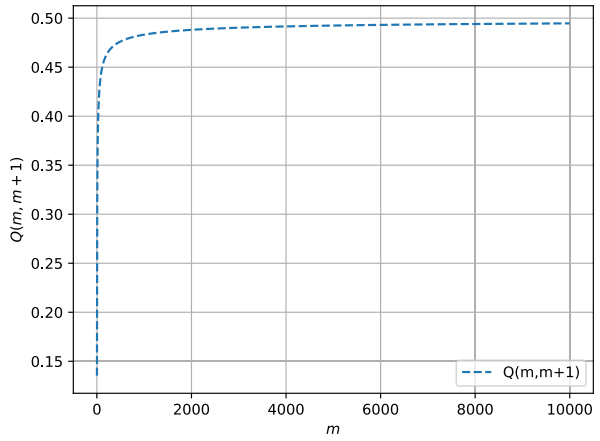

FIGURE 10. Plot of $Q(m, m+1)$ for different values of m.

$P(a,b) = 1 - Q(a,b)$. Thus, by utilizing (80) and considering $y = \sqrt{\frac{2}{a}}$ we have

$$\begin{aligned} Q(a+1, a+2) &= 1 - P(a+1, a+2) \\ &= 1 - \frac{1}{2}\text{erfc}\left(-\sqrt{\frac{2}{a}}\right) \\ &\quad + \frac{1}{3}\sqrt{\frac{2}{\pi a}}\left(1 + \frac{2}{a}\right)e^{-\frac{2}{a}} - O(a^{-1}), \end{aligned} \tag{82}$$

and the limit of $Q(m, m+1)$ at infinity can be calculated as

$$\begin{aligned} \lim_{m\to\infty} Q(m, m+1) &= \lim_{m\to\infty} Q(m+1, m+2) \\ &= 1 - \frac{1}{2}\text{erfc}(0) = \frac{1}{2}. \end{aligned} \tag{83}$$

Thus, we know that the inequality $Q(m_{\pi_{\ell,i}}, m_{\pi_{\ell,i}} + 1) \leq 0.5$ is always true and in the interval $0.5 \leq p \leq 1$ the inequality (79) and (78) are also always true. This proves that $h''_p(p, \boldsymbol{\omega}) \geq 0$ and completes the proof of Lemma 3.

REFERENCES

[1] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain NOMA with enabling technologies of future wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2192–2235, 4th Quart., 2020.

[2] M. Vaezi, G. A. Aruma Baduge, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 4, pp. 900–919, Dec. 2019.

[3] J. G. Andrews, "Interference cancellation for cellular systems: A contemporary overview," *IEEE Wireless Commun.*, vol. 12, no. 2, pp. 19–29, Apr. 2005.

[4] J.-M. Kang and I.-M. Kim, "Optimal user grouping for downlink NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 724–727, Oct. 2018.

[5] K. Wang, W. Liang, Y. Yuan, Y. Liu, Z. Ma, and Z. Ding, "User clustering and power allocation for hybrid non-orthogonal multiple access systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12052–12065, Oct. 2019.

[6] B. Rashid, A. Ahmad, S. Saleem, and A. Khan, "Joint energy efficient power and subchannel allocation for uplink MC-NOMA networks," *Int. J. Commun. Syst.*, vol. 33, no. 17, p. e4606, Sep. 2020.

[7] Z. Yang, C. Pan, W. Xu, and M. Chen, "Compressive sensing-based user clustering for downlink NOMA systems with decoding power," *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 660–664, May 2018.

[8] Y. Cheng, K. H. Li, K. C. Teh, S. Luo, and W. Wang, "Joint user clustering and subcarrier allocation for downlink non-orthogonal multiple access systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[9] X. Wei, H. Al-Obiedollah, K. Cumanan, M. Zhang, J. Tang, W. Wang, and O. A. Dobre, "Resource allocation technique for hybrid TDMA-NOMA system with opportunistic time assignment," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6.

[10] S. Shi, L. Yang, and H. Zhu, "Outage balancing in downlink nonorthogonal multiple access with statistical channel state information," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4718–4731, Jul. 2016.

[11] M. M. Al-Wani, A. Sali, B. M. Ali, A. A. Salah, K. Navaie, C. Y. Leow, N. K. Noordin, and S. J. Hashim, "On short term fairness and throughput of user clustering for downlink non-orthogonal multiple access system," in *Proc. IEEE 89th Veh. Technol. Conf. (VTC-Spring)*, Apr. 2019, pp. 1–6.

[12] S. M. A. Kazmi, A. Manzoor, and C. S. Hong, "User grouping for non-orthogonal multiple access (NOMA)," *Proc. Korean Soc. Inf. Sci. Acad. Presentation*, pp. 1337–1339, 2018. [Online]. Available: http://163.180.116.116/layouts/net/publications/data/KCC2018/17.Ahsan.pdf

[13] J. Ren, Z. Wang, M. Xu, F. Fang, and Z. Ding, "An EM-based user clustering method in non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8422–8434, Dec. 2019.

[14] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 328–331, Apr. 2019.

[15] H. T. H. Giang, T. N. K. Hoan, P. D. Thanh, and I. Koo, "Hybrid NOMA/OMA-based dynamic power allocation scheme using deep reinforcement learning in 5G networks," *Appl. Sci.*, vol. 10, no. 12, p. 4236, Jun. 2020.

[16] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[17] K. Shaukat Ali, M.-S. Alouini, E. Hossain, and M. Jahangir Hossain, "On clustering and channel disparity in non-orthogonal multiple access (NOMA)," 2019, *arXiv:1905.02337*.

[18] M. K. Simon and M.-S. Alouini, *Digital Communication Over Fading Channels*, vol. 95. Hoboken, NJ, USA: Wiley, 2005.

[19] A. Mahmoudi, B. Abolhassani, S. M. Razavizadeh, and H. H. Nguyen, "Outage balancing in downlink NOMA over Nakagami-*m* fading channels," *IEEE Access*, vol. 9, pp. 102886–102898, 2021.

[20] N. M. Temme, "Asymptotic inversion of incomplete gamma functions," *Math. Comput.*, vol. 58, no. 198, pp. 755–764, 1992.

[21] *Scipy.Special.Gammainccinv*. Accessed: Jun. 4, 2021. [Online]. Available: https://docs.scipy.org

[22] Y. Iraqi and A. Al-Dweik, "Power allocation for reliable SIC detection of rectangular QAM-based NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 8355–8360, Aug. 2021.

[23] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[24] Y. Fu, Y. Chen, and C. W. Sung, "Distributed power control for the downlink of multi-cell NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6207–6220, Sep. 2017.

[25] R. K. Jain *et al.*, "A quantitative measure of fairness and discrimination," Eastern Res. Lab., Digit. Equip. Corp., Hudson, MA, USA, Tech. Rep. 301, 1984, vol. 21.

[26] J. L. Jacob and T. Abrão, "Nonorthogonal multiple access systems optimization to ensure maximum fairness to users," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 4, p. e3875, Apr. 2020.

[27] F. Tricomi, "Asymptotische eigenschaften der unvollständigen gammafunktion," *Mathematische Zeitschrift*, vol. 53, no. 2, pp. 136–148, 1950.

[28] W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, Eds. *NIST Digital Library of Mathematical Functions*. Accessed: Jun. 15, 2021. [Online]. Available: http://dlmf.nist.gov/ and [Online]. Available: https://dlmf.nist.gov/8.11.E10

ALI MAHMOUDI received the B.Sc. degree from the University of Tehran, Tehran, Iran, and the M.Sc. degree from the Sharif University of Technology, Tehran. Currently, he is pursuing the Ph.D. degree with the School of Electrical Engineering, Iran University of Science and Technology, Tehran. He was a Researcher and a Software Developer with the Iran Telecommunication Research Center for two years. His research interests include primarily in the area of wireless communication systems, cellular networks, and optimization theory.

BAHMAN ABOLHASSANI was born in Tehran, Iran. He received the B.Sc. degree from the Iran University of Science and Technology (IUST), Tehran, and the M.Sc. and Ph.D. degrees from the University of Saskatchewan, Saskatoon, SK, Canada, all in electrical engineering. He was an Instrumentation Engineer with the College of Water and Power Technology, Iranian Ministry of Energy, for three years. Then, he worked as a Communication Systems Engineer in a number of private and government companies. He joined the School of Electrical Engineering, IUST, where he is currently an Associate Professor. He served as the Dean for the School of Electrical Engineering and an Associate Dean for research. He also served as a Sessional Lecturer for the University of Saskatchewan. His research interests include in the fields of wireless communication systems, networks planning, spread spectrum, cognitive radio networks, resource allocation, VANETs, and optimization of large systems.

S. MOHAMMAD RAZAVIZADEH (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Iran University of Science and Technology (IUST), Tehran, Iran, in 1997, 2000, and 2006, respectively. From June 2004 to April 2005, he was a Visiting Researcher with the Coding and Signal Transmission Laboratory, University of Waterloo, ON, Canada. From 2005 to 2011, he was with the Iran Telecommunication Research Center, as a Research Assistant Professor. Since 2011, he has been with the School of Electrical Engineering, IUST, where he is currently an Associate Professor. He was also a Visiting Professor with Korea University, South Korea; and Chalmers University, Sweden, during the summers of 2013 and 2015, respectively. His research interests include in the area of signal processing for wireless communication systems and cellular networks.

HA H. NGUYEN (Senior Member, IEEE) received the B.Eng. degree from the Hanoi University of Technology (HUT), Hanoi, Vietnam, in 1995, the M.Eng. degree from the Asian Institute of Technology (AIT), Bangkok, Thailand, in 1997, and the Ph.D. degree from the University of Manitoba, Winnipeg, MB, Canada, in 2001, all in electrical engineering. He joined the Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK, Canada, in 2001, and became a Full Professor, in 2007. He currently holds the position of NSERC/Cisco Industrial Research Chair in low-power wireless access for sensor networks. He is a coauthor, with Ed Shwedyk, of the textbook *A First Course in Digital Communications* (Cambridge University Press). His research interests include broad areas of communication theory, wireless communications, and statistical signal processing. He is a fellow of the Engineering Institute of Canada (EIC) and a registered member of the Association of Professional Engineers and Geoscientists of Saskatchewan (APEGS). He served as a technical program chair for numerous IEEE events. He was an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE WIRELESS COMMUNICATIONS LETTERS, from 2007 to 2011 and 2011 to 2016, respectively. He currently serves as an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.