IEEE Access
The journal for rapid open access publishing

IEEE Xplore Digital Library
http://ieeexplore.ieee.org/Xplore

Energy-Efficient Power Allocation for MIMO-NOMA with Multiple Users in a Cluster

Ming Zeng, *Student Member, IEEE*, Animesh Yadav, *Member, IEEE*, Octavia A. Dobre, *Senior Member*, *IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—In this paper, energy-efficient power allocation (PA) is investigated for a multiple-input multiple-output non-orthogonal multiple access (MIMO-NOMA) system with multiple users in a cluster. To ensure the quality of service (QoS) for the users, a minimum rate requirement is pre-defined for each user. Because of the QoS requirement, it is first necessary to determine whether the considered energy efficiency (EE) maximization problem is feasible or not, by comparing the total transmit power with the required power for satisfying the QoS of the users. If feasible, a closed-form solution is provided for the corresponding sum rate maximization problem, and on this basis, the EE maximization problem is solved by applying non-convex fractional programming. Otherwise, a low complexity user admission scheme is proposed, which admits users one by one following the ascending order of the required power for satisfying the QoS. Numerical results are presented to validate the effectiveness of the proposed energy-efficient PA strategy and user admission scheme.**

***Index Terms*—Non-orthogonal multiple access (NOMA), multiple-input multiple-output (MIMO), energy efficiency (EE), user admission, power allocation (PA), quality-of-service (QoS).**

I. Introduction

Power-domain non-orthogonal multiple access (NOMA) has been widely considered as a promising candidate for the next generation of wireless communication systems [1]–[5]. By applying superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, NOMA multiplexes multiple users in the power domain, to access the same time-frequency resource. When compared with conventional orthogonal multiple access (OMA) scheme, NOMA achieves higher spectral efficiency (SE) [6]–[8]. The authors in [6] show via simulation that NOMA provides a larger sum rate than OMA, while in [7], the authors prove the dominance of NOMA over OMA by comparing their achievable rate regions. Furthermore, the authors in [8] validate that NOMA achieves higher ergodic sum rate than OMA for a cellular downlink scenario with randomly deployed users.

However, the above works only consider single-input single-output (SISO) channels. Recently, multiple-input multiple-output (MIMO) has also been integrated into NOMA to further enhance the SE [9]–[13]. For MIMO-NOMA systems, users are usually paired into clusters to reduce the complexity of SIC at the receiver, with users in the same cluster sharing a common beamformer. The authors in [9] and [10] show that MIMO-NOMA achieves larger sum rates than MIMO-OMA for a two-user multi-cluster system, while [11] and [12] further validate that this performance advantage still holds for a multi-user per cluster system. Note that the above works only consider power allocation (PA) within each cluster, by allocating equal power to each cluster. In [13], the authors propose a beamspace MIMO-NOMA scheme for a millimeter wave system, which allows power to be distributed among clusters. Simulation results illustrate that the proposed beamspace MIMO-NOMA achieves higher SE when compared with existing beamspace MIMO-OMA. In [14], the authors extend the study of MIMO-NOMA from single cell to multicell and investigate the precoder design. Numerical results show that the proposed NOMA design improves both edge and sum throughput compared with conventional OMA.

Nevertheless, the studies above mainly focus on the SE of NOMA systems. As energy efficiency (EE) becomes one of the major concerns for 5G, it is of interest to investigate the EE for NOMA [15]–[17]. In [15], the authors study the joint subchannel assignment and PA to maximize the EE for a multi-carrier NOMA system. The obtained simulation results show that NOMA achieves higher SE and EE than OMA. However, this work is only applicable to systems with two users per cluster. In [16], EE is studied under a single-carrier multi-user NOMA system with quality-of-service (QoS) requirement for each user. A PA algorithm is proposed based on non-convex fractional programming, and numerical results validate that NOMA exhibits better EE performance than OMA. Note that both [15] and [16] consider SISO systems. Since current and future communication systems rely on the multiple antenna (MIMO) structure, the EE under MIMO-NOMA is of interest. In [17], the authors investigate the EE in a millimeter wave massive MIMO-NOMA system with a low-complexity radio frequency (RF) chain structure at the base station (BS). A hybrid analog/digital precoding scheme is proposed first. Based on this, a PA problem aiming to maximize the EE is formulated under users' QoS requirements and per-cluster equal power constraint, and an iterative algorithm is proposed to obtain an optimal PA.

To the best of our knowledge, none of the existing works has studied the EE for a multi-cluster MIMO-NOMA with multiple users per cluster. Moreover, most existing studies assume that the total transmit power is large enough to satisfy the QoS requirements for all users, without considering

M. Zeng, A. Yadav, and O. A. Dobre are with Memorial University, St. John's, NL A1B 3X9, Canada (e-mail: mzeng, animeshy, odobre@mun.ca).

H. V. Poor is with Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), though its Discovery program, and the U.S. National Science Foundation under Grants CCF-1420575 and ECCS-1647198.

Digital Object Identifier: 10.1109/ACCESS.2017.2779855

the situation when this assumption does not hold [15]–[17]. Toward filling this research gap, the contributions of this paper are summarized as follows:

- We study the EE for a multi-cluster multi-user MIMO-NOMA system with pre-defined QoS for each user in a systematic way: we first determine whether all users can be admitted or not by comparing the total transmit power with the power required to satisfy the QoS for all users; when all users can be admitted, we aim to maximize the EE of the system; otherwise, we aim to maximize the number of admitted users;
- For the EE maximization problem, global PA is considered: we first determine how to allocate power within each cluster; on this basis, we derive the relationship between the sum rate increment and required extra power for each cluster; by exploiting this relationship, a water-filling-like optimal PA is proposed to maximize the sum rate of the system under any given total power; lastly, it is proved that the EE function is pseudo-concave over the final "water" level, and can be solved accordingly;
- For the user admission problem, a low complexity algorithm is proposed, which admits the users one by one following the ascending order of the required power to satisfy their QoS; further analysis on its optimality and complexity is provided, which validates the effectiveness of the proposed algorithm.

The rest of the paper is organized as follows. The system model and problem formulation are introduced in Section II. The proposed energy-efficient PA strategy and user admission scheme are elaborated in Section III. Simulation results are shown in Section IV, while conclusions are finally drawn in Section V.

II. System Model and Problem Formulation

A. System model

We consider a downlink multi-user MIMO system in this paper, in which the BS equipped with M antennas sends data to multiple receivers, each equipped with N antennas. The total number of users in the system is $M \times L$, which are grouped into M clusters randomly with L ($L \geq 2$) users per cluster. NOMA is applied among the users in the same cluster. The channel matrix between the BS and the lth user in the mth cluster, i.e., user (m,l) ($m \in \{1,\ldots,M\}, l \in \{1,\ldots,L\}$) is denoted as $\mathbf{H}_{m,l} \in \mathbb{C}^{N\times M}$, which is assumed to be quasi-static independent and identically distributed (i.i.d.). In addition, the precoding matrix used by the BS is denoted as $\mathbf{P} \in \mathbb{C}^{M\times M}$, whereas the detection vector for user (m,l) is represented by $\mathbf{v}_{m,l} \in \mathbb{C}^{N\times 1}$. They should satisfy: a) $\mathbf{P} = \mathbf{I}_M$, with $\mathbf{I}_M$ denoting the $M \times M$ identity matrix; b) $|\mathbf{v}_{m,l}|^2 = 1$ and $\mathbf{v}_{m,l}^H \mathbf{H}_{m,l}\mathbf{p}_k = 0$ for any $k \neq m$, where $\mathbf{p}_k$ is the kth column of $\mathbf{P}$ [11]. Note that the number of antennas should satisfy $N \geq M$ to make this feasible. Because of the zero-forcing (ZF) based detection design, the inter-cluster interference can be removed even when there exist multiple users in a cluster. Note that only a scalar value $|\mathbf{v}_{m,l}^H \mathbf{H}_{m,l}\mathbf{p}_m|^2$ needs to be fed back to the BS from user (m,l).

For the considered MIMO-NOMA scheme, the BS multiplexes the intended signals for all users at the same frequency and time resource. Therefore, the corresponding transmitted signals from the BS can be expressed as

$$\mathbf{x} = \mathbf{P}\mathbf{s}, \tag{1}$$

where the information-bearing vector $\mathbf{s} \in \mathbb{C}^{M\times 1}$ can be further written as

$$\mathbf{s} = \begin{bmatrix} \sqrt{P_{\max}\Omega_{1,1}}s_{1,1} + \cdots + \sqrt{P_{\max}\Omega_{1,L}}s_{1,L} \\ \vdots \\ \sqrt{P_{\max}\Omega_{M,1}}s_{M,1} + \cdots + \sqrt{P_{\max}\Omega_{M,L}}s_{M,L} \end{bmatrix}, \tag{2}$$

where $s_{m,l}$ and $\Omega_{m,l}$ denote the signal and power allocation coefficient for user (m,l), satisfying $\sum_{m=1}^{M}\sum_{l=1}^{L}\Omega_{m,l} \leq 1$. $P_{\max}$ denotes the total transmit power for the BS.

Accordingly, at user (m,l), the observed signal is given by

$$\mathbf{y}_{m,l} = \mathbf{H}_{m,l}\mathbf{P}\mathbf{s} + \mathbf{n}_{m,l}, \tag{3}$$

where $\mathbf{n}_{m,l}$ is the independent and identically distributed (i.i.d.) additive white Gaussian (AWGN) noise vector, $\mathcal{CN}(0,\sigma^2\mathbf{I})$.

By applying the detection vector $\mathbf{v}_{m,l}$ on the observed signal, (3) can be expressed as

$$\mathbf{v}_{m,l}^H\mathbf{y}_{m,l} = \mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m\sum_{l=1}^{L}\sqrt{P_{\max}\Omega_{m,l}}s_{m,l} + \underbrace{\sum_{k=1,k\neq m}^{M}\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_k\mathbf{s}_k}_{\text{interference from other clusters}} + \mathbf{v}_{m,l}^H\mathbf{n}_{m,l}, \tag{4}$$

where $\mathbf{s}_k$ denotes the kth row of $\mathbf{s}$.

Owing to the constraint[1] on the detection vector, i.e., $\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_k = 0$ for any $k \neq m$, (4) can be simplified as

$$\mathbf{v}_{m,l}^H\mathbf{y}_{m,l} = \mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m\sum_{l=1}^{L}\sqrt{P_{\max}\Omega_{m,l}}s_{m,l} + \mathbf{v}_{m,l}^H\mathbf{n}_{m,l}. \tag{5}$$

Without loss of generality, the effective channel gains are ordered as [11]

$$|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2 \geq \cdots \geq |\mathbf{v}_{m,L}^H\mathbf{H}_{m,L}\mathbf{p}_m|^2. \tag{6}$$

At the receiver, each user conducts SIC to remove the interference from the users with worse channel gains, i.e., the interfence from user $(m,l+1),\ldots,(m,L)$ is removed by user (m,l).[2] As a result, the achieved data rate at user (m,l) is given by [11]

$$R_{m,l} = \log_2\left(1 + \frac{\rho\Omega_{m,l}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}{1+\rho\sum_{k=1}^{l-1}\Omega_{m,k}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}\right), \tag{7}$$

where $\rho = P_{\max}/\sigma^2$ denotes the transmit signal-to-noise ratio (SNR).

[1] Due to the specific selection of $\mathbf{P}$, this constraint is further reduced to $\mathbf{v}_{m,l}^H\tilde{\mathbf{H}}_{m,l} = 0$, where $\tilde{\mathbf{H}}_{m,l} = [\mathbf{h}_{1,ml}\cdots\mathbf{h}_{m-1,ml}\,\mathbf{h}_{m+1,ml}\cdots\mathbf{h}_{M,ml}]$ and $\mathbf{h}_{i,ml}$ is the ith column of $\mathbf{H}_{m,l}$ [11]. Hence, $\mathbf{v}_{m,l}$ can be expressed as $\mathbf{U}_{m,l}\mathbf{w}_{m,l}$, where $\mathbf{U}_{m,l}$ is the matrix consisting of the left singular vectors of $\tilde{\mathbf{H}}_{m,l}$ corresponding to the non-zero singular values, and $\mathbf{w}_{m,l}$ is the maximum ratio combining vector expressed as $\mathbf{U}_{m,l}^H\mathbf{h}_{m,ml}/|\mathbf{U}_{m,l}^H\mathbf{h}_{m,ml}|$.

[2] [12] proves that SIC is guaranteed to be successful.

B. Problem formulation

The total power consumption is comprised of two parts: the fixed circuit power consumption $P_{\rm c}$, and the flexible transmit power $P_{\rm t} = P_{\max}\sum_{m=1}^{M}\sum_{l=1}^{L}\Omega_{m,l}$. Similar to [16], we define the EE of the system as

$$\eta_{\rm EE} = \frac{R^{\rm sum}}{P_{\rm c}+P_{\rm t}}, \tag{8}$$

where $R^{\rm sum} = \sum_{m=1}^{M}\sum_{l=1}^{L} R_{m,l}$ denotes the achievable sum rate.

We aim to maximize the EE of the system when each user has a pre-defined minimum rate. The considered problem can be formulated as:

$$\begin{aligned} \max_{\Omega_{m,l}} \ & \eta_{\rm EE}* && \text{(9a)}\\ \text{s.t.} \ & R_{m,l} \geq R_{m,l}^{\min}, m \in \{1,\cdots,M\}, l \in \{1,\cdots,L\} \\ & \sum_{m=1}^{M}\sum_{l=1}^{L}\Omega_{m,l} \leq 1, \end{aligned}$$

where (10) and (10) represent the users' minimum rate requirements and the transmit power constraint, respectively.

III. Proposed Solution

Owing to the existence of the minimum rate requirements, i.e., (10), the formulated problem (9) may be infeasible when the transmit power is not large enough. In this case, instead of EE maximization, maximizing the number of admitted users makes more sense. As such, it is of importance to determine the feasibility of problem (9), which can be done by comparing the total transmit power constraint with the minimum power required to satisfy the minimum rate requirements of all users. The minimum required power can be expressed as

$$P_{\rm req} = P_{\max}\sum_{m=1}^{M}\sum_{l=1}^{L}\Omega_{m,l}^{\min}, \tag{10}$$

where $\Omega_{m,l}^{\min} = (2^{R_{m,l}^{\min}}-1)(\sum_{k=1}^{l-1}\Omega_{m,k}^{\min} + \frac{1}{\rho|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2})$ is the minimum required power to satisfy the QoS requirement of user (m,l) [18, (14)]. As a result, if

$$P_{\rm req} \leq P_{\max} \Longleftrightarrow \sum_{m=1}^{M}\sum_{l=1}^{L}\Omega_{m,l}^{\min} \leq 1, \tag{11}$$

problem (9) is feasible and vice versa.

A. EE maximization when (9) is feasible

The objective function in (9) is of fractional form; hence (9) is a non-convex problem and obtaining an optimal solution is non-trival. To solve it in a tractable way, we first turn to the corresponding SE maximization problem. According to the definition of EE in (8), to maximize the EE, we need to maximize the corresponding SE under any given power of $P_{\rm f}, P_{\rm f} \in [P_{\rm req}, P_{\max}]$, and then select the appropriate value of $P_{\rm f}$.

The SE maximization problem can be formulated as

$$\begin{aligned} \max_{\Omega_{m,l}} \ & R^{\rm sum}* && \text{(12a)}\\ \text{s.t.} \ & R_{m,l} \geq R_{m,l}^{\min}, m \in \{1,\cdots,M\}, \\ & l \in \{1,\cdots,L\} \\ & P_{\max}\sum_{m=1}^{M}\sum_{l=1}^{L}\Omega_{m,l} \leq P_{\rm f}. \end{aligned}$$

Note that problem (12) is still non-convex due to the non-concavity involved in the objective function. In order to proceed towards an optimal solution, we first determine the PA within each cluster and then across clusters. For PA within each cluster, the following lemma provides some insight:

Lemma 1: Under any given total power constraint for a cluster,[3] in order to maximize the cluster sum rate, PA in the cluster should be conducted such that each user (except the first user) receives the amount of power such that its QoS requirement is just satisfied, while the first user receives the remaining power.

Proof: To prove the lemma, we first prove that transferring power from any other user to the first user leads to an increased sum rate. Assume that the power transfer happens between the nth user and the 1st user, and denote the extra power coefficient as $\Delta P_{\rm tr}$. According to (7), the rates of the users with worse channel gains than the nth user remain unchanged, since the total interference does not change. Thus, when comparing the two cluster sum rates, we only need to compare the first n users.

The sum rate of the first n users before power transfer can be expressed as

$$\begin{aligned} \sum_{l=1}^{n} R_{m,l} &= \sum_{l=1}^{n}\log_2\left(\frac{1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}{1+\rho\sum_{k=1}^{l-1}\Omega_{m,k}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}\right) \\ &= \log_2\Bigg(\prod_{l=1}^{n-1}\frac{1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}{1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,l+1}^H\mathbf{H}_{m,l+1}\mathbf{p}_m|^2} \\ &\quad \times (1+\rho\sum_{k=1}^{n}\Omega_{m,k}|\mathbf{v}_{m,n}^H\mathbf{H}_{m,n}\mathbf{p}_m|^2)\Bigg). \end{aligned}$$

Likewise, the sum rate of the first n users after power transfer can be expressed as

$$\begin{aligned} &\sum_{l=1}^{n} R'_{m,l} \\ &= \log_2\Bigg(\prod_{l=1}^{n-1}\frac{1+\rho(\Delta P_{\rm tr}+\sum_{k=1}^{l}\Omega_{m,k})|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}{1+\rho(\Delta P_{\rm tr}+\sum_{k=1}^{l}\Omega_{m,k})|\mathbf{v}_{m,l+1}^H\mathbf{H}_{m,l+1}\mathbf{p}_m|^2} \\ &\quad \times (1+\rho\sum_{k=1}^{n}\Omega_{m,k}|\mathbf{v}_{m,n}^H\mathbf{H}_{m,n}\mathbf{p}_m|^2)\Bigg). \end{aligned}$$

Since $|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2 \geq |\mathbf{v}_{m,l+1}^H\mathbf{H}_{m,l+1}\mathbf{p}_m|^2$, it can be

[3]The total power is large enough to satisfy the QoS requirements of all users in the cluster.

easily verified that

$$\frac{1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,l}^{H}\mathbf{H}_{m,l}\mathbf{p}_m|^2}{1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,l+1}^{H}\mathbf{H}_{m,l+1}\mathbf{p}_m|^2} < \frac{1+\rho(\Delta P_{\mathrm{tr}}+\sum_{k=1}^{l}\Omega_{m,k})|\mathbf{v}_{m,l}^{H}\mathbf{H}_{m,l}\mathbf{p}_m|^2}{1+\rho(\Delta P_{\mathrm{tr}}+\sum_{k=1}^{l}\Omega_{m,k})|\mathbf{v}_{m,l+1}^{H}\mathbf{H}_{m,l+1}\mathbf{p}_m|^2}. \quad (13)$$

Therefore, we can prove that $\sum_{l=1}^{n} R_{m,l} < \sum_{l=1}^{n} R'_{m,l}$, which demonstrates that transferring power from other users to the first user yields a larger sum rate. On the other hand, each user should also satisfy its QoS constraint. Combining these two facts, we can conclude that Lemma 1 holds. ■

Remark: The above analysis can be extended to show that transferring power from any user to another user with better channel gains leads to an increased sum rate. This implies that the users with better channel gains have a higher priority than their counterparts. In the user admission section, this property of NOMA is further exploited.

The above lemma shows how to allocate power within a cluster. Now we consider PA across clusters. We first allocate the power such that each user's QoS requirement is just satisfied, which requires the power of P_{req}. Correspondingly, the remaining power is denoted as $P_{\mathrm{rem}} = P_{\mathrm{f}} - P_{\mathrm{req}}$. Then, we allocate the remaining power across clusters to maximize the system sum rate. To determine how to allocate power across clusters, the intuition is to compare how much additional power is needed for each cluster when increasing its sum rate by the same unit. The following lemma provides the details:

Lemma 2: Denote the achieved rate of user (m,l) as $\hat{R}_{m,l}$ ($\hat{R}_{m,l} \geq R_{m,l}^{\min}$), the additional power required for increasing the sum rate of the mth cluster by ΔR is given by

$$\Delta P_m = (2^{\Delta R}-1)\frac{P_{\max}2^{\sum_{l=1}^{L}\hat{R}_{m,l}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}. \quad (14)$$

Proof: We prove Lemma 2 by mathematical induction. Starting with two users per cluster, according to (7), we have the following:

$$\hat{\Omega}_{m,1} = \frac{2^{\hat{R}_{m,1}}-1}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2} \quad (15a)$$

$$\hat{\Omega}_{m,2} = \frac{(2^{\hat{R}_{m,2}}-1)(1+\rho\Omega_{m,1}|\mathbf{v}_{m,2}^{H}\mathbf{H}_{m,2}\mathbf{p}_m|^2)}{\rho|\mathbf{v}_{m,2}^{H}\mathbf{H}_{m,2}\mathbf{p}_m|^2} \quad (15b)$$

$$= \frac{2^{\hat{R}_{m,2}}-1}{\rho|\mathbf{v}_{m,2}^{H}\mathbf{H}_{m,2}\mathbf{p}_m|^2} + \frac{(2^{\hat{R}_{m,1}}-1)(2^{\hat{R}_{m,2}}-1)}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}.$$

According to Lemma 1, when some additional power is added to the mth cluster, only the rate of the first user will change, while others remain fixed. Thus, when there is ΔR sum rate increment for the mth cluster, it is only added to $R_{m,1}$, resulting in the change from $\hat{R}_{m,1}$ to $\hat{R}_{m,1}+\Delta R$. Update $R_{m,1}$ in (15), and after some algebraic manipulations, the additional power required is given by

$$\Delta P_m^{(2)} = (2^{\Delta R}-1)\frac{P_{\max}2^{\sum_{l=1}^{2}\hat{R}_{m,l}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}. \quad (16)$$

This completes the proof for the two user per cluster case.

Assume that (14) holds for n users per cluster, i.e., $\Delta P_m^{(n)} = P_{\max}\sum_{l=1}^{n}\Delta\hat{\Omega}_{m,l} = (2^{\Delta R}-1)\frac{P_{\max}2^{\sum_{l=1}^{n}\hat{R}_{m,l}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$. On this basis, we consider the case with $n+1$ users. According to (7), the power coefficient for user $(m,n+1)$ before increasing the sum rate is given by

$$\hat{\Omega}_{m,n+1} = \frac{(2^{\hat{R}_{m,n+1}}-1)(1+\rho|\mathbf{v}_{m,n+1}^{H}\mathbf{H}_{m,n+1}\mathbf{p}_m|^2\sum_{l=1}^{n}\hat{\Omega}_{m,l})}{\rho|\mathbf{v}_{m,n+1}^{H}\mathbf{H}_{m,n+1}\mathbf{p}_m|^2}. \quad (17)$$

After the ΔR sum rate increment, the rate of user $(m,n+1)$ remains unchanged according to Lemma 1. Thus, the power coefficient increment for user $(m,n+1)$ is $\Delta\hat{\Omega}_{m,n+1} = (2^{\hat{R}_{m,n+1}}-1)\sum_{k=1}^{n}\Delta\hat{\Omega}_{m,k}$.

Accordingly, the total required extra power for the $n+1$ users can be expressed as

$$\begin{aligned}\Delta P_m^{(n+1)} &= \Delta P_m^{(n)} + P_{\max}\Delta\hat{\Omega}_{m,n+1}\\ &= P_{\max}\sum_{k=1}^{n}\Delta\hat{\Omega}_{m,k} + P_{\max}(2^{\hat{R}_{m,n+1}}-1)\sum_{k=1}^{n}\Delta\hat{\Omega}_{m,k}\\ &= P_{\max}2^{\hat{R}_{m,n+1}}\sum_{k=1}^{n}\Delta\hat{\Omega}_{m,k}\\ &= 2^{\hat{R}_{m,n+1}}(2^{\Delta R}-1)\frac{P_{\max}2^{\sum_{l=1}^{n}\hat{R}_{m,l}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}\\ &= (2^{\Delta R}-1)\frac{P_{\max}2^{\sum_{l=1}^{n+1}\hat{R}_{m,l}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}, \end{aligned} \quad (18)$$

which completes the proof. ■

We observe that only the channel gains of the first user and the minimum rate requirement of all users affect the power increment for each cluster. Moreover, for smaller $\frac{P_{\max}2^{\sum_{l=1}^{L}\hat{R}_{m,l}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$, less additional power is needed for increasing the sum rate by the same unit. This observation can be used for designing an iterative PA algorithm. Specifically, during each iteration, the cluster with the smallest $\frac{P_{\max}2^{\sum_{l=1}^{L}\hat{R}_{m,l}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$ is selected for receiving the additional power. On the other hand, after this cluster receives a certain amount of additional power, its sum rate $\sum_{l=1}^{L}\hat{R}_{m,l}$ increases, and it may no longer be the one with the smallest $\frac{P_{\max}2^{\sum_{l=1}^{L}\hat{R}_{m,l}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$. This process repeats until P_{f} is fully used. This iterative algorithm is similar to the classical water-filling technique, and a closed-form solution can be obtained accordingly.

More precisely, after the initial feasible PA, we obtain $R_{m,l} = R_{m,l}^{\min}, m \in \{1,\cdots,M\}, l \in \{1,\cdots,L\}$. We consider $\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$ as the initial "water" level. Furthermore, we introduce an axillary variable λ as the final "water" level. If $\lambda \leq \frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$, the mth cluster receives no power and remains unchanged. Otherwise, the mth cluster receives some extra power to reach the final "water" level, i.e., $\lambda = \frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}+\Delta R_m}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2} = 2^{\Delta R_m}\times\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$, where ΔR_m

is the rate increment. In this case, according to (14), the required extra power can be expressed as

$$\Delta P_m = (2^{\Delta R_m}-1)\frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2} = \lambda - \frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2}. \quad (19)$$

Considering both cases, the required power for the mth cluster can be further expressed as

$$\Delta P_m = \left[\lambda - \frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right]^+, \quad (20)$$

where $x^+ = \max(x;0)$. This provides a closed-form solution for the SE maximization, once λ is known. To attain the value of λ, we refer to the total power constraint, which should satisfy

$$\sum_{m=1}^{M}\left[\lambda - \frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right]^+ = P_{\text{rem}}. \quad (21)$$

The left side of the above equation is piecewise and monotonically increasing over λ. Thus, a unique value of λ exists and can be obtained by solving (21). Note that there is a point-to-point mapping between λ and P_{f}, and further, λ increases with P_{f}.

Moreover, the sum rate increment for the mth cluster can be expressed as

$$\Delta R_m = \left[\log_2(\lambda) - \log_2\left(\frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right)\right]^+. \quad (22)$$

The following lemma shows the optimality of the proposed SE maximization PA strategy.

Lemma 3: The proposed SE maximization PA strategy maximizes the sum rate of the system.

Proof: Assume that we have obtained the solution via the proposed PA algorithm, i.e., λ is known and so are other values, e.g., the extra power for each cluster. Now, we shift Δp power between two clusters whose final "water" level is λ. Denote the two clusters as the qth and nth cluster, respectively. For the proposed PA strategy, the sum rate increment for the qth cluster after the initial PA can be expressed as

$$\Delta R_q = \log_2(\lambda) - \log_2\left(\frac{P_{\max}2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho|\mathbf{v}_{q,1}^H\mathbf{H}_{q,1}\mathbf{p}_q|^2}\right) = \log_2\left(\Delta P_q + \frac{P_{\max}2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho|\mathbf{v}_{q,1}^H\mathbf{H}_{q,1}\mathbf{p}_q|^2}\right) - \log_2\left(\frac{P_{\max}2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho|\mathbf{v}_{q,1}^H\mathbf{H}_{q,1}\mathbf{p}_q|^2}\right). \quad (23)$$

For ΔR_n, a similar expression can be written.

After shifting some power between two clusters, we have

$$\Delta R_q' = \log_2(\Delta P_q + \Delta p + \frac{P_{\max}2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho|\mathbf{v}_{q,1}^H\mathbf{H}_{q,1}\mathbf{p}_q|^2}) - \log_2\left(\frac{P_{\max}2^{R_{q,2}^{\min}}}{\rho|\mathbf{v}_{q,1}^H\mathbf{H}_{q,1}\mathbf{p}_q|^2}\right) = \log_2(\lambda+\Delta p) - \log_2\left(\frac{P_{\max}2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho|\mathbf{v}_{q,1}^H\mathbf{H}_{q,1}\mathbf{p}_q|^2}\right). \quad (24)$$

Likewise, a similar equation can be written for $\Delta R_n'$.

Accordingly, we calculate the sum rate difference as follows:

$$\begin{aligned}\Delta R_{\text{sum}} &= \Delta R_q + \Delta R_n - \Delta R_q' - \Delta R_n' \\ &= \log_2(\lambda) + \log_2(\lambda) - \log_2(\lambda+\Delta p) - \log_2(\lambda-\Delta p) \\ &= \log_2(\lambda^2) - \log_2[\lambda^2 - (\Delta p)^2] > 0. \end{aligned} \quad (25)$$

The above equation clearly shows that shifting power between two clusters whose final "water" level is λ leads to a lower sum rate. Following the same procedure, we can also prove that this holds when shifting power from the cluster whose final "water" level equals to λ to another cluster whose final "water" level exceeds λ. This validates the optimality of the proposed scheme. ■

Now we have solved the SE maximization problem (12) under P_{f}. On this basis, we need to select the appropriate P_{f} to maximize the EE of the system. Consider P_{f} as the variable here, but replace it with λ owing to the point-to-point mapping between them.

Accordingly, the consumed transmit power can be rewritten as

$$\begin{aligned} P_{\text{t}} &= P_{\text{f}} \\ &= P_{\text{req}} + \sum_{m=1}^{M}\Delta P_m \\ &= P_{\text{req}} + \sum_{m=1}^{M}\left[\lambda - \frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right]^+. \end{aligned} \quad (26)$$

Similarly, the sum rate can be rewritten as

$$\begin{aligned} R^{\text{sum}} &= \sum_{m=1}^{M}\sum_{l=1}^{L} R_{m,l}^{\min} + \sum_{m=1}^{M}\Delta R_m \\ &= \sum_{m=1}^{M}\sum_{l=1}^{L} R_{m,l}^{\min} + \sum_{m=1}^{M}\left[\log_2(\lambda) - \log_2\left(\frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right)\right]^+. \end{aligned} \quad (27)$$

As a result, the expression of η_{EE} can be written; this is provided on the top of next page.

Clearly, in (28), the only variable is λ, as other parameters are known. Moreover, (28) is a piecewise function, and its specific form depends on the interval λ lies in. To find the intervals, we arrange $\frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2}$ in an ascending order and use h_t to denote the tth value after ordering for simplicity of notation. Since the total transmit power $P_{\max}$ is known, we can calculate the maximum index of the interval that λ can lie

$$\eta_{\mathrm{EE}}=\frac{\sum_{m=1}^{M}\sum_{l=1}^{L}R_{m,l}^{\min}+\sum_{m=1}^{M}\left[\log_2(\lambda)-\log_2\left(\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right)\right]^{+}}{P_{\mathrm{c}}+P_{\mathrm{req}}+\sum_{m=1}^{M}\left[\lambda-\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right]^{+}}. \tag{28}$$

in according to (21), by setting $P_{\mathrm{rem}}=P_{\max}-P_{\mathrm{req}}$. Denote the maximum index as T, then λ can only lie in $[h_t,h_{t+1}],t=1,\cdots,T$. Moreover, we have the following theorem:

Theorem 1: For each interval $[h_t,h_{t+1}],t=1,\cdots,T$, η_{EE} is a strictly pseudo-concave function with respect to (w.r.t.) λ.

Proof: Once t is known, η_{EE} can be turned into

$$\eta_{\mathrm{EE}}=\frac{\sum_{m=1}^{M}\sum_{l=1}^{L}R_{m,l}^{\min}+t\log_2(\lambda)-\sum_{k=1}^{t}\log_2(h_k)}{P_{\mathrm{c}}+P_{\mathrm{req}}+t\lambda-\sum_{k=1}^{t}h_k}. \tag{29}$$

It can be seen that the numerator is a strictly concave function over λ, while the denominator is an affine mapping over λ. Thus, η_{EE} is a strictly pseudo-concave function w.r.t. λ [19, Proposition 6]. ∎

For $\lambda\in[h_t,h_{t+1}]$, as η_{EE} is a strictly pseudo-concave function w.r.t. λ, η_{EE} admits a unique maximizer, which is obtained either at the unique root of the equation $\frac{\partial\eta_{\mathrm{EE}}}{\partial\lambda}=0$ or at the two boundary points h_t or h_{t+1} [19, Proposition 5]. Denote this maximizer as η_{EE}^{t}. Likewise, when λ lies in $[h_k,h_{k+1}],k\neq t$, denote the unique maximizer as η_{EE}^{k}. As η_{EE} belongs to two different functions for these two intervals, we cannot determine the comparative values of these two maximizers analytically. Instead, an explicit comparison has to be done, i.e., $\max\left\{\eta_{EE}^{t},\eta_{EE}^{k}\right\}$. As the total number of intervals is T, we need to obtain the maximizer in each interval and select the maximum for η_{EE}, which can be expressed as

$$\eta_{\mathrm{EE}}^{\max}=\max\left\{\eta_{\mathrm{EE}}^{1},\cdots,\eta_{\mathrm{EE}}^{T}\right\}. \tag{30}$$

So far, we have derived the solution for maximizing the EE of the system. We summarize the procedures in Algorithm 1. Moreover, the following theorem proves its optimality.

Theorem 2: The derived solution achieves the maximum EE for the system.

Proof: According to Lemma 3, for any given total power, the proposed solution maximizes the SE of the system by appropriately allocating power across clusters and inside each cluster. Then, Theorem 1 guarantees that the EE is maximized for each feasible interval. As (30) selects the maximum value from all these maximizers, this selected maximum value is the global optimum. ∎

Algorithm 1 Proposed EE Maximization PA Algorithm

Initialize parameters.
  $P_{\max}$, $R_{m,l}^{\min}$, $\rho|\mathbf{v}_{m,l}^{H}\mathbf{H}_{m,l}\mathbf{p}_m|^2,l\in\{1,\cdots,L\}$
Calculate:
  $\mathbf{H}\leftarrow sort\left(\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right)$;
  $h_t\leftarrow\mathbf{H}(t)$;
  $\lambda^{\max}\leftarrow\sum_{t=1}^{M}[\lambda-h_t]^{+}=P_{\max}-P_{\mathrm{req}}$;
  $T\leftarrow\lambda^{\max}\in[h_T,h_{T+1}]$;
  $\eta_{\mathrm{EE}}^{t}\leftarrow\max\left\{\eta_{\mathrm{EE}}(\frac{\partial\eta_{\mathrm{EE}}}{\partial\lambda}=0),\eta_{\mathrm{EE}}(h_t),\eta_{\mathrm{EE}}(h_{t+1})\right\},t\in\{1,\cdots,T-1\}$;
  $\eta_{\mathrm{EE}}^{T}\leftarrow\max\left\{\eta_{\mathrm{EE}}(\frac{\partial\eta_{\mathrm{EE}}}{\partial\lambda}=0),\eta_{\mathrm{EE}}(h_T),\eta_{\mathrm{EE}}(\lambda^{\max})\right\}$;
  $\eta_{\mathrm{EE}}^{\max}\leftarrow\max\left\{\eta_{\mathrm{EE}}^{1},\cdots,\eta_{\mathrm{EE}}^{T}\right\}$;
end

B. User admission when problem (9) *is infeasible*

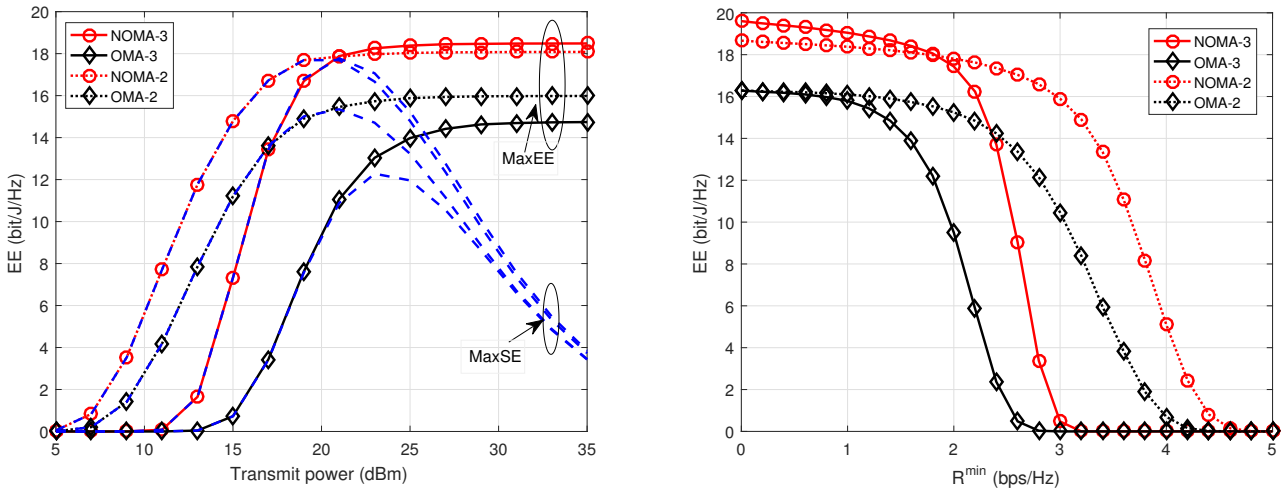
When (9) is infeasible, admitting as many users as possible is a more reasonable goal, when compared with EE maximization. The user admission problem can be formulated as

$$\begin{aligned}\max_{\Omega_{m,l}}\quad&\sum_{m=1}^{M}\sum_{l=1}^{L}x_{m,l}* &\text{(31a)}\\ \text{s.t.}\quad&R_{m,l}\geq R_{m,l}^{\min}x_{m,l},\\&\sum_{m=1}^{M}\sum_{l=1}^{L}\Omega_{m,l}\leq 1,\\&x_{m,l}\in\{0,1\},\end{aligned}$$

where $x_{m,l}$ is the binary decision variable indicating whether user (m,l) is admitted or not.

In [12], under the assumption of equal power for each cluster, we propose a greedy user admission algorithm, which admits users one by one following the descending order of their channel gains within each cluster. In this paper, as power can be transferred among clusters, user admission should be conducted globally. Based on the observation that the users with better channel gains own higher priority than their counterparts in each cluster due to SIC, we still admit users within each cluster following the descending order of their channel gains. Furthermore, with multi-clusters in the system, we also need to determine the order for admitting users across clusters. This can be done by comparing the required power for satisfying the QoS of each user in different clusters, and select the one with the minimum power consumption during each user admission process.

More exactly, the user admission is conducted iteratively as follows: during each iteration, we first select the user with the best channel gain from each cluster; among these selected users, the required power is calculated with considering the interference from the already admitted users; then, the user with the minimum required power is chosen to be admitted; if the total remaining power exceeds the required power for



(a) How EE varies with the transmit power: $R^{\min} = 2$ bps/Hz (b) How EE varies with the minimum rate requirement: $P_{\mathrm{t}} = 20$ dBm

Fig. 1: Scenario 1: $d_1 = d_2 = d_3 = 80$ m.

admitting this user, the selected user is admitted and eliminated from the candidates; besides, the total remaining power is updated; otherwise, the process terminates; the process repeats until no further user can be admitted.

Theorem 3: The proposed scheme maximizes the number of admitted users when the users' QoS requirements in each cluster satisfy the following conditions:

$$R_{m,k}^{\min} \leq R_{m,n}^{\min}, \forall k \in \{1, \ldots, l\}, n \in \{l+1, \ldots, L\}, \quad (32)$$

where l represents the total number of admitted users in the mth cluster under the proposed scheme.

Proof: Refer to Appendix I. ■

Corollary 1: The proposed user admission scheme maximizes the number of admitted users when the SINR thresholds of the users in each cluster satisfy the following conditions:

$$R_{m,1}^{\min} \leq \cdots \leq R_{m,L}^{\min}. \quad (33)$$

Particularly, when the QoS requirements of the users are equal, the proposed user admission scheme is optimal in terms of both sum rate and number of admitted users.

Proof: When (33) is satisfied, it can be easily proved that (32) holds for any l. Thus, the proposed scheme maximizes the number of admitted users. If the QoS requirements of the users are equal, it can be easily inferred that maximizing the number of admitted users also leads to the maximization of the sum rate. ■

Lemma 4: The complexity of the proposed user admission algorithm is $O(M^2 L)$.

Proof: The proposed user admission algorithm admits users one by one following the ascending order of the required power for satisfying their QoS requirements, which requires $O(ML)$ operations. During each user admission, the main complexity comes from the operation of selecting the minimum value across all clusters, which requires $O(M)$ operations. In all, the complexity of the proposed user admission algorithm is $O(M^2 L)$. ■

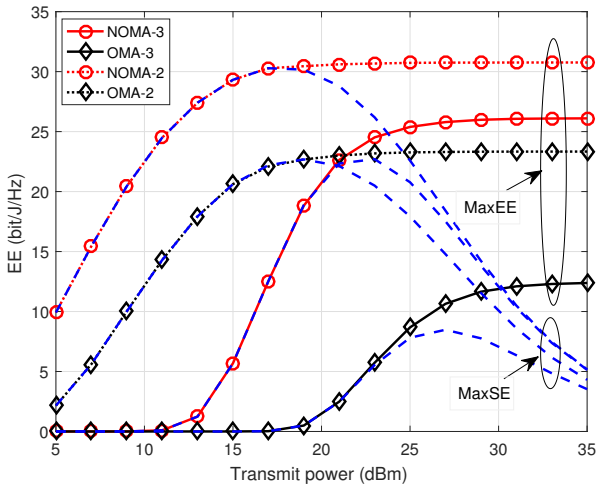
TABLE I: Simulation Parameters.

Parameters	Value
Number of antennas	$M = 3, N = 3$
Channel bandwidth	10 [MHz]
Thermal noise density	-174 [dBm/Hz]
Path-loss model	$120 + 30 \log_{10}(d)$, d in kilometer

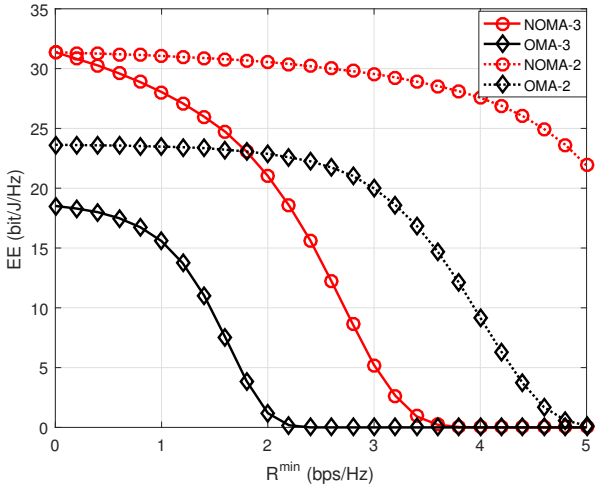
IV. Simulation Results

In this section, simulations are conducted to verify the performance of the proposed PA strategy and user admission scheme. The specific values of the adopted simulation parameters are summarized in Table I [12]. All results are obtained by averaging over 10^4 random trials, unless mentioned otherwise. Particularly, in the case when the total transmit power cannot support the QoS requirements for all users, the EE of these trials is set to zero since the objective is not EE maximization.

First, we evaluate the effectiveness of the proposed energy-efficient PA strategy. To compare NOMA with conventional OMA, we adopt OMA with equal degrees of freedom for each user as the baseline algorithm. Note that OMA can be considered as a special case of NOMA, with one user in each cluster. The energy-efficient PA for OMA can be attained by employing the proposed energy-efficient PA strategy for NOMA with some minor adjustment, e.g., now the cluster number becomes $M \times L$. The above energy-efficient PA strategies are denoted as "MaxEE". As a baseline algorithm, the PA strategy that consumes full power to maximize the SE of the system is also presented, which is denoted as "MaxSE". This "MaxSE" PA for NOMA can be obtained by employing the proposed water-filling sum rate maximization algorithm. In terms of the "MaxSE" PA for OMA, it can be achieved by employing the classical water-filling algorithm.

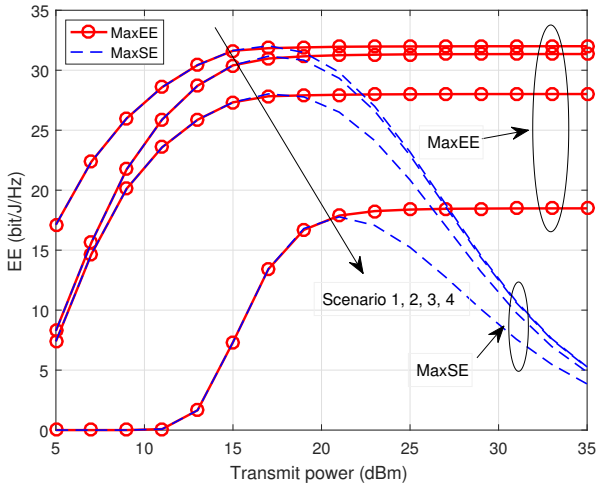


(a) How EE varies with the transmit power: $R^{\min} = 2$ bps/Hz

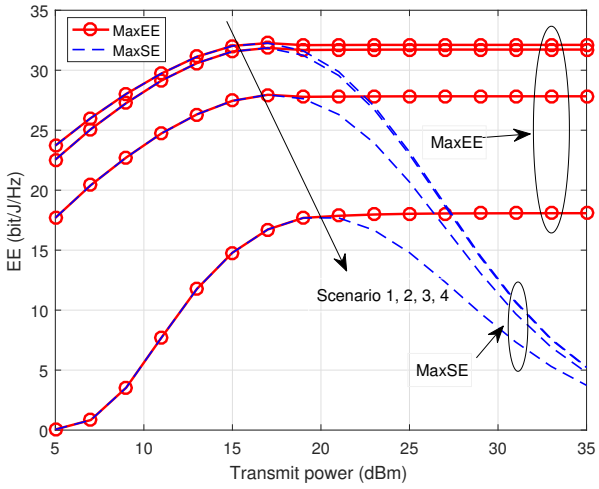


(b) How EE varies with the minimum rate requirement: $P_{\mathrm{t}} = 20$ dBm

Fig. 2: Scenario 2: $d_1 = 40\ m, d_2 = 80\ m, d_3 = 120\ m$.



(a) Three users per cluster



(b) Two users per cluster

Fig. 3: EE versus total power available at the BS, for different cases of user locations.
Scenario 1: $d_1 = 60$ m, $d_2 = 50$ m, $d_3 = 40$ m, $(d_1 + d_2 + d_3)/3 = 50$ m. Scenario 2: $d_1 = 70$ m, $d_2 = 55$ m, $d_3 = 40$ m, $(d_1 + d_2 + d_3)/3 = 55$ m.
Scenario 3: $d_1 = 60$ m, $d_2 = 55$ m, $d_3 = 50$ m, $(d_1 + d_2 + d_3)/3 = 55$ m. Scenario 4: $d_1 = 80$ m, $d_2 = 80$ m, $d_3 = 80$ m, $(d_1 + d_2 + d_3)/3 = 80$ m.

To show how EE varies as the number of users in each cluster increases, two scenarios with different distances are presented in Figs. 1 and 2, in which "-3" and "-2" mean three and two users per cluster, respectively. For each scenario, we show how EE varies with the total transmit power and minimum rate requirement. According to Figs. 1 and 2, NOMA achieves higher EE than OMA for both two and three user cases, respectively.

Specifically, subfigures 1(a) and 2(a) show how EE varies with the transmit power, in which the dashed lines in both figures denote the "MaxSE", while all other lines represent the "MaxEE". Clearly, under low transmit power, "MaxSE" equals "MaxEE", and both grow with the transmit power. As the transmit power reaches a certain threshold, further increase in the transmit power does not yield a higher EE, and thus, "MaxEE" remains stable, while "MaxSE" decreases. This indicates the necessity of employing energy-efficient PA, especially under high transmit power. In scenario 1, under low transmit power, NOMA-2 achieves higher EE compared with NOMA-3. However, under high transmit power, an opposite result can be observed. This can be explained by the fact that under low transmit power, it is more difficult to satisfy the QoS for three users. On the other hand, under high power, more users lead to a higher diversity, which increases the EE. In contrast, in scenario 2, NOMA-2 always attains higher EE than its counterpart. This is due to the fact as the distance

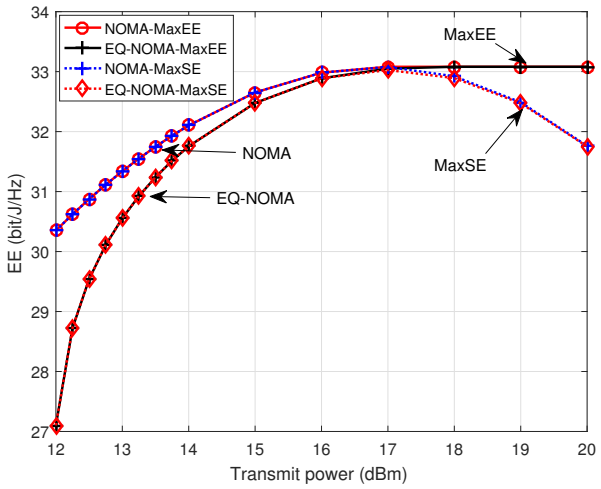


Fig. 4: EE versus total power available at the BS, for NOMA and EQ-NOMA.

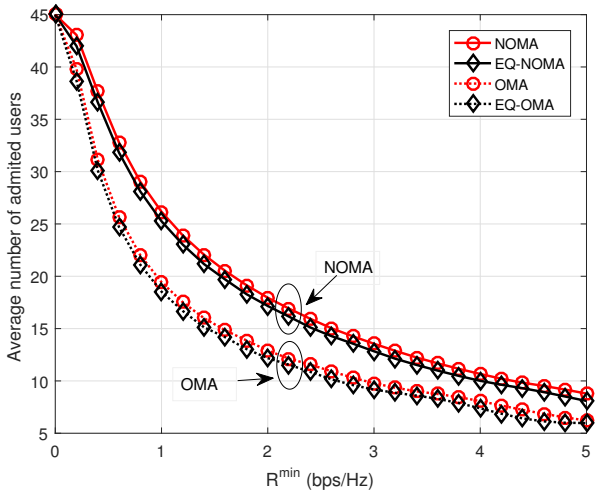


Fig. 6: Average number of admitted users versus $R^{\min}$: number of requesting users per cluster is 15; $P_{\mathrm{t}} = 20$ dBm.

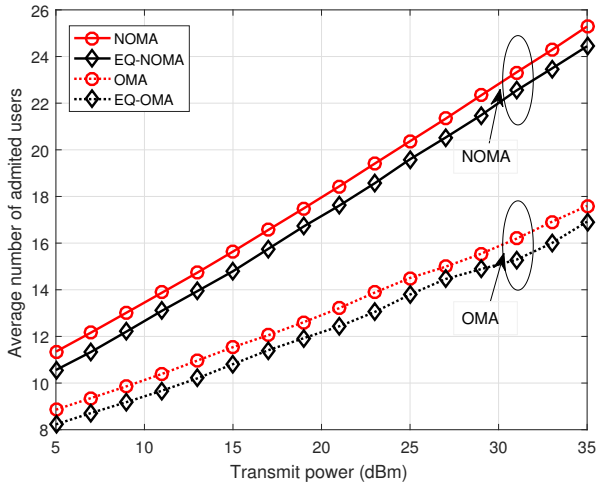


Fig. 5: Average number of admitted users versus transmit power: number of requesting users per cluster is 15; $R^{\min} = 2$ bps/Hz.

difference between the users increases, it costs more energy to admit an extra user. Thus, even under high transmit power, the benefit introduced by the diversity is not enough to compensate the energy required for admitting the extra user. Combining the two scenarios, we can conclude that whether admitting more users yields a higher EE depends on the transmit power level and the distance difference between the users.

Subfigures 1(b) and 2(b) show how EE varies with $R^{\min}$. It can be seen that EE decreases with $R^{\min}$. More exactly, in scenario 1, NOMA-3 achieves higher EE than NOMA-2 under low $R^{\min}$, and vice versa. This can be explained by connecting $R^{\min}$ with the transmit power, i.e., lower $R^{\min}$ has the same impact on EE as higher transmit power. In contrast, in scenario 2, NOMA-2 always achieves higher EE than NOMA-3, which agrees with subfigure (a).

Results in Figs. 1 and 2 indicate that the distance has an impact on EE; accordingly, in Fig. 3, further analysis on this is provided. Obviously, the larger the distance, the lower the achieved EE. Furthermore, comparing scenarios 2 and 3, we can conclude that the channel gain of the strongest user plays a vital role in EE, which fits our observation in Lemma 2. On the other hand, by comparing three and two user cases for scenario 2, it implies that the distance difference between users has a larger impact on the multi-user case, especially under lower transmit power. To conclude, not only the average distance, but also the distance of the strongest user plays an important role in EE. Moreover, the distance difference affects EE more for the three user case under low transmit power.

In Fig. 4, we compare EE achieved by the proposed PA strategy with that achieved by the algorithm in [17], in which equal power is assigned to each cluster, and thus is denoted as "EQ-NOMA". Further, for both algorithms, both "MaxEE" and "MaxSE" are plotted. It can be seen that under low transmit power, the proposed PA strategy achieves higher EE than the one in [17], which validates the necessity of applying global PA. On the other hand, under high transmit power, their performance is the same. This can be explained by the fact that under high transmit power, the equally divided power is enough for EE maximization, and thus, allowing power to be transferred among clusters brings no benefit.

Figs. 5-7 show the performance of the proposed user admission scheme, which is denoted as "NOMA". As a baseline algorithm, we consider the NOMA scheme in [12], which assigns equal power to each cluster, and is denoted as "EQ-NOMA". To compare NOMA with conventional OMA, OMA with PA across clusters and OMA with equal power per cluster are presented, denoted as "OMA" and "EQ-OMA", respectively. According to Figs. 5-7, it can be seen that NOMA outperforms OMA in terms of the number of admitted users versus transmit power, minimum rate requirement, and number of requesting users. Moreover, for both NOMA and OMA, allowing power to be transferred among clusters leads to a larger number of admitted users. In addition, it is clear that the average number of admitted users grows with the transmit

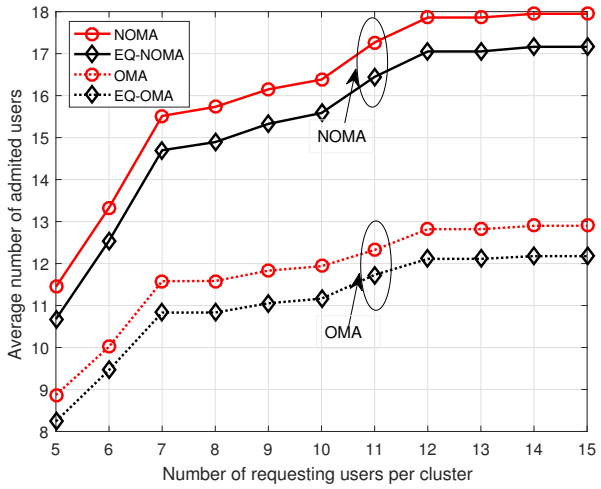


Fig. 7: Average number of admitted users versus number of requesting users per cluster: $R^{\min} = 2$ bps/Hz; $P_{\rm t} = 20$ dBm.

power, but decreases with $R^{\min}$. Furthermore, it also increases with the number of requesting users per cluster. This is due to the fact that when more users are requesting the service, it is more likely that more users will have a better channel gains, yielding a lower power to satisfy their minimum rate requirements. As the total power is fixed, more users can be admitted accordingly.

V. Conclusion

In this paper, we have studied the EE maximization problem for a multi-cluster multi-user MIMO-NOMA system under a QoS constraint for each user. An optimal PA strategy has been proposed to solve the considered EE maximization problem when it is feasible. A low complexity user admission protocol has been proposed otherwise, which admits users one by one following the ascending order of the required power for satisfying the QoS requirements. Numerical results show that the proposed PA strategies outperform OMA and equal power NOMA in terms of both EE and the number of admitted users, which verifies their effectiveness. In addition, the EE of the NOMA system mainly depends on the channel condition of the first user, and it is necessary to apply an energy-efficient PA strategy, especially at high transmit power. On the other hand, whether more users leads to increased EE depends on the transmit power level and users' channel gain difference.

Appendix A
Proof of Theorem 3

Proof: We first consider the user admission in each cluster. In the following, we will prove through contradiction that the proposed scheme maximizes the number of admitted users in each cluster.

Consider the case in which only l users can be admitted to the mth cluster when employing the proposed user admission scheme. Suppose there exists an alternate scheme, which also admits l users, but replaces the kth user with the nth user as one admitted user, $k \in \{1,\ldots,l\}, n \in \{l+1,\ldots,L\}$. In this case, it seems that the alternate scheme transfers the power of the kth user to the nth user. Moreover, from the $(k+1)$th user to the lth user, the required power for satisfying their QoS requirements decreases, as the interference from the kth user is removed. This reduced power can also be considered to be transferred to the nth user. According to the remark from Lemma 1, a lower sum rate is achieved by transferring power from the strong users to the weak users. Since all other users' rates remain the same, the achievable rate of the nth user must be lower than that of the kth user, $R_{m,n} \leq R^{\min}_{m,k}$. On the other hand, $R^{\min}_{m,k} \leq R^{\min}_{m,n}$. Therefore, $R_{m,n} \leq R^{\min}_{m,n}$, which indicates that more power is needed to satisfy the QoS requirement of the nth user. This shows that the proposed scheme requires the minimum power when there is one replacement between the users. Following the same procedure, the conclusion can be easily extended to the case in which there exist multiple replacements of the users, which means that the proposed scheme requires the minimum power for admitting l users, i.e., $\Omega_{\rm sum} \leq \Omega^{\rm alt}_{\rm sum}$, where $\Omega_{\rm sum}$ and $\Omega^{\rm alt}_{\rm sum}$ are the total power coefficients of admitting l users for the proposed scheme and the alternate one, respectively.

Suppose the alternate scheme can admit an extra user, denoted as a_{l+1}. Without loss of generality, the channel gain of this user is assumed to be the lowest. Note that this assumption does not add an extra constraint since we can simply exchange it with the one of the lowest channel gain, and consider the latter as the extra admitted user. According to (7), $\Omega^{\rm alt}_{m,a_{l+1}} \geq (2^{R^{\min}_{m,a_{l+1}}}-1)\left(\Omega^{\rm alt}_{\rm sum} + \frac{1}{\rho|\mathbf{v}^H_{m,a_{l+1}}\mathbf{H}_{m,a_{l+1}}\mathbf{p}_m|^2}\right)$. In addition, admitting the a_{l+1} to the proposed scheme requires $\Omega_{m,a_{l+1}} = (2^{R^{\min}_{m,a_{l+1}}}-1)\left(\Omega_{\rm sum} + \frac{1}{\rho|\mathbf{v}^H_{m,a_{l+1}}\mathbf{H}_{m,a_{l+1}}\mathbf{p}_m|^2}\right)$. As $\Omega_{\rm sum} \leq \Omega^{\rm alt}_{\rm sum}$, we have $\Omega_{m,a_{l+1}} + \Omega_{\rm sum} \leq \Omega^{\rm alt}_{m,a_{l+1}} + \Omega^{\rm alt}_{\rm sum}$. Thus, this extra user can also be admitted to the proposed scheme, which conflicts with the proposition that only l users can be admitted by the proposed scheme.

With multi-clusters, since the proposed scheme selects the user with the minimum required power across clusters during each iteration, this clearly yields the maximum number of admitted users. ■

References

[1] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* vol. 1, no. 2, May 2017. [Online]. Available: http://5g.ieee.org/tech-focus.

[2] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. pp, no. 99, pp. 1–1, Oct. 2016.

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

[4] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[5] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in non-orthogonal multiple access networks for 5G and beyond," *IEEE Network*, vol. 31, no. 4, pp. 8–14, Jul. 2017.

[6] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5.

[7] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, UK: Cambridge University Press, 2005.

[8] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[9] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[10] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, Aug. 2017.

[11] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[12] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[13] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," *IEEE J. Select. Areas Commun.*, to appear, 2017.

[14] V. Nguyen, H. Tuan, T. Duong, H.V., Poor, and O. Shin, "Precoder design for signal superposition in MIMO-NOMA multicell networks," *IEEE J. Select. Areas Commun.*, vol. PP, no. 99, pp. 1–1, Jul. 2017.

[15] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[16] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[17] W. M. Hao, et al., "Energy-efficient power allocation in millimeter wave massive mimo with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. PP, no. 99, pp. 1–1, Jun. 2017.

[18] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Global Commun. Conf.*, Washington DC, USA, Dec. 2016.

[19] A. Zappone, P. Lin, and E. Jorswieck, "Energy efficiency in secure multi-antenna systems," *IEEE Trans. Signal Process.,* submitted for publication. [Online]. Available: http://arxiv.org/abs/1505.02385.