Energy Efficient Resource Allocation in Hybrid Non-Orthogonal Multiple Access Systems

Jia Shi, *Member, IEEE*, Wenjuan Yu, Qiang Ni, *Senior Member, IEEE*, Wei Liang,
Zan Li, *Senior Member, IEEE*, and Pei Xiao, *Senior Member, IEEE*

Abstract—By blending the concepts of non-orthogonal multiple access (NOMA) and orthogonal frequency division multiplexing (OFDM), in this paper a novel hybrid scheme is conceived for supporting diverse services in future wireless systems. Motivating to maximize energy efficiency (EE), the joint resource management of user clustering (UC) and power allocation is investigated for the downlink hybrid NOMA systems. Under two different power consumption cases, the optimal resource allocation (Opt-RA) algorithm is developed with the help of converting the original mixed integer non-linear programming (MINLP) problem to the tractable decoupled problems. For practical implementation, the heuristic resource allocation (Heur-RA) algorithm is also proposed, in which including a low-complexity UC algorithm based on the candidate search-and-allocation approach. Our simulation results show that, both the Opt-RA and Heur-RA algorithms achieve significantly higher EE performance than other existing algorithms. Further, the results also prove that, the hybrid NOMA conceived is able to exploit the advantages of NOMA scheme, and is superior to conventional orthogonal multiple access (OMA) in terms of EE, as well as achieving higher flexibility for system configuration than NOMA.

I. INTRODUCTION

Current 4G system is reaching maturity, and will be evolved to 5G system with very limited amount of new spectrum that can be exploited. However, by the time 5G comes to fruition, the key challenges lie on the explosive growth for the number of devices and huge increment of data volume, which boosts the utilization of NOMA technique for exploiting the valuable spectrum more efficiently [1].

A. Related Works

NOMA technique allows multiple users to share time and spectrum resources in the same spatial layer via power-domain multiplexing in contrast to conventional orthogonal multiple access (OMA) techniques including time-division multiple access (TDMA), frequency-division multiple access (FDMA)

J. Shi and Zan Li are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, 710071, China (email: {jiashi,zanli}@xidian.edu.cn).

Wenjuan Yu and Pei Xiao are with 5GIC, University of Surrey, Guildford, Surrey, GU1 2XH, UK. Email: ({w.yu,p.xiao@surrey.ac.uk}).

Q. Ni is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, UK (email: q.ni@lancaster.ac.uk).

Wei Liang is with Northwestern Polytechnical University, Xi'an, 710072, China. (Email: liangwei@nwpu.edu.cn.)

[2]. The efficient resource allocation (RA) approaches are originally investigated for optimizing the OMA systems [3]–[5], and they will will provide a new degree of freedom for NOMA systems. The studies for RA in NOMA systems have drawn a lot research attentions on the improvement of spectrum efficiency (SE) or sum rate [6]–[12]. In [8], the low-complexity polynomial algorithm has been proposed to approximately solve the non-convex power allocation problem. While, the authors in [8] have studied the decoupled problem of the user clustering and power allocation in the NOMA systems, where the proposed UC is based on exhaustive search approach with very high complexity required. Recently, motivating to maximizing sum rate, Liang *et al* [9] have proposed one-to-one matching based user pairing and investigated power allocation problem in the cognitive radio NOMA systems. Further, [10] has proposed a novel subchannel allocation approach for the downlink NOMA systems by leveraging many-to-many matching theory. More recently, research efforts have been devoted to resource management in various NOMA systems, for example, the authors in [12] have developed a joint power allocation and subcarrier partitioning scheme for a low density spreading (LDS) multiple access systems, i.e. a code-domain NOMA system.

Energy consumption is becoming a major social and economical issue for future wireless communications especially with explosive data traffic. Hence, it is important to address energy efficiency (EE) when designing NOMA systems. In [4], energy efficiency of the single-carrier FDMA system was studied, where the energy efficiency is quantified in terms of the energy consumption gain and the energy reduction gain [13]. Nevertheless, limited research efforts, such as [14]–[17], have been devoted to the study of energy efficient RA in NOMA systems. Fang *et al* have investigated energy efficient RA in [15] for the downlink NOMA systems, where, however, the proposed gradient based binary search power allocation algorithm requires relatively high complexity. Furthermore, in [16], it has proposed the novel power allocation algorithms which are EE-optimal and SE-optimal respectively. Recent studies have applied NOMA to future machine-to-machine communications in [18] and to millimeter wave (mmWave) networks in [18]. Interestingly, the authors of [14] have investigated the application of NOMA technique in mmWave networks by jointly studying the beamforming, user scheduling, and power allocation.

Future 5G communications will demand diverse service types, and need to support massive connectivity of users and/or devices, as well as meeting the requirements for low latency,

low-cost devices. Whereas, neither conventional orthogonal multiple access (OMA) nor NOMA is fully perfect for accommodating various services, which thereby stimulates the study of hybrid NOMA technique. Particularly, in 5G systems, a hybrid NOMA technique can utilize available spectrum to perform the mixture of NOMA and OFDM schemes, in which the NOMA-SC can accommodate the UEs for massive machine type communications (mMTC) service while the OFDM-SC can support the UEs for enhanced mobile broadband (eMBB) service. However, the main challenges for applying hybrid NOMA lie in resource management, such that: 1) how to efficiently coordinate various radio frequency resource, energy resource, and spatial resource, etc.; 2) how to effectively mitigate the potential interference among intra-SC links as well as that among inter-SC links; 3) how to find the best trade-off between obtaining optimal key performance indicators (KPIs) and minimizing implementation complexity. So far, current studies on hybrid NOMA systems are very limited, and hardly address the resource management issues. Such that, a hybrid NOMA system, by blending the concepts of NOMA and TDMA, was first proposed in [19]. The study of [19] focused on performance analysis only, which showed that the hybrid NOMA can achieve a better fairness performance than NOMA. The authors of [20] commenced the first work about RA in hybrid NOMA systems, where a coalition based user grouping algorithm was developed. Furthermore, in [21], the hybrid NOMA scheme of a combination between NOMA and OFDM has been proposed and claimed to be applicable to 5G communications. Nevertheless, the very limited current studies on RA in hybrid NOMA, such as [20], [21], only focused on channel allocation, while aiming to maximize the sum rate.

B. Motivations and Contributions

Compared to traditional NOMA or OMA technique, hybrid NOMA exploits a range of advantages, including supporting diverse services, being more spectrum efficient than OMA, less susceptible to interference than NOMA, as well as requiring less successive interference cancellation (SIC) complexity than NOMA. To the best of our knowledge, very limited studies are devoted to resource allocation in hybrid NOMA systems, especially the investigation of EE motivated resource allocation in hybrid NOMA systems. Against this background, in this paper, we focus on efficiently managing spectrum and power resources in the downlink hybrid NOMA systems. For clarify, the contributions of this paper are summarized as follows.

- By blending the concepts of NOMA and OFDM, we conceive a novel hybrid NOMA scheme, which can exploit the advantages of both NOMA and OFDM techniques. To jointly consider user clustering (UC) and power allocation in the hybrid systems, we formulate and analyze the optimization problem, which is a mixed integer non-linear programming (MINLP) problem. For addressing practical issues, two power consumption cases are studied: 1) Limited power consumption (LPC), which adjusts the total transmit power adaptively; 2) Full power consumption (FPC), which uses the fixed total transmit power available.

- The optimal RA algorithm, referred to as Opt-RA, is proposed to find the optimal UC and power allocation solutions, by means of converting the original MINLP problem to the equivalent decoupled problems which are tractable. In particular, given a UC result, the optimal power allocation solutions can be found for both the LPC and FPC cases, by iteratively solving the equivalent concave problems which are derived with the help of novel transformations and approximations, such as Charnes-Cooper transformation, and first-order Taylor series approximation, etc. The proposed power allocation algorithms are of low complexity, and are proved to converge within very few iterations.

- Inspired by the analysis of the Opt-RA, we propose a low-complexity RA algorithm, namely heuristic RA (Heur-RA), for the hybrid NOMA systems. To avoid the exhaustive search based UC in the Opt-RA, the Heur-RA proposes a novel UC algorithm, that assigns users to subcarriers (SCs) based on the candidate search-and-allocation approach for minimize the implementation complexity. Further, the complexity analysis is carried out for the proposed algorithms including Opt-RA and Heur-RA as well as the other widely-used RA algorithms. The results show that our Heur-RA algorithm requires very low complexity and, hence can be seen as a promising option for practical implementation.

- Comprehensive performance analysis is carried out for the proposed Opt-RA and Heur-RA algorithms in the context of the hybrid NOMA systems. It is shown that, the proposed Heur-RA algorithm achieves significantly higher EE performance than the other existing algorithms, and its performance is close to that obtained by the Opt-RA under certain scenarios. In addition, our simulation results also show that, the hybrid NOMA conceived is able to exploit the advantages of NOMA scheme, and is superior to conventional OMA in terms of EE, and achieves higher flexibility for system configuration than NOMA. Therefore, the hybrid NOMA associated with the proposed RA schemes may constitute a promising candidate that facilitates practical implementation in future communication systems.

II. System Model and Problem Formulation

A. System Model

Motivated to boost the diverse use cases in future wireless communications, such as 5G scenarios, we conceive the downlink hybrid NOMA system, which can be depicted by Fig. 1. As the example shown in the figure, different number of users can be accommodated on different SCs with the aid of the hybrid NOMA technique. In the system conceived, let us assume M orthogonal SCs, denoted by $\mathcal{M} = \{1, \ldots, M\}$, are employed to serve K users, denoted by $\mathcal{K} = \{1, \ldots, K\}$, which are grouped into $\boldsymbol{L} = [L_1, \ldots, L_M]$ clusters supported by the hybrid NOMA scheme combining the concepts of both the NOMA and OFDM schemes. Specifically, L_m, $m \in \mathcal{M}$, is the number of users which can be accommodated on SC m, and SC m is referred to as a NOMA-SC if $L_m > 1$, while
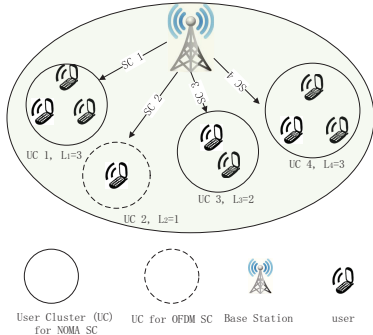
Fig. 1.   System Model for a downlink hybrid NOMA, where the example has $\boldsymbol{L} = [3, 1, 2, 3]$.

it belongs to a OFDM-SC if $L_m = 1$. Let us define, $\mathcal{M}_n = \{m \mid m \in \mathcal{M}, L_m > 1\}$ and $\mathcal{M}_o = \{m \mid m \in \mathcal{M}, L_m = 1\}$ contain the indexes of the NOMA-SCs and OFDM-SCs and, hence, we have $\mathcal{M} = \mathcal{M}_n \cup \mathcal{M}_o$. The BS is assumed to have full knowledge of channel state information (CSI), and carries out UC to assign each SC different number of users according to the setup $\boldsymbol{L}$. To avoid considering various trivial cases, in this paper we study the most challenging case for UC, which assumes that each user has to be assigned to only one SC, and assumes $K = \sum_{m=1}^{M} L_m$. To characterize diverse service requirements in future wireless networks, we assume that the power constraints for each SC can be different, which can be expressed by $\{P_m, \forall m \in \mathcal{M}\}$.

On the basis of NOMA principles, the BS transmits information to the users superposed on a NOMA-SC via power domain division. At the receiver sides, the users decode its message and eliminate the multi-user interference with the aid of successive interference cancellation (SIC). Specifically, each user on a NOMA-SC first decodes the users on the same SC with poorer channel conditions, and removes their messages one by one. Further, the messages from the users having higher channel conditions are treated as noise. By contrast, each user on OFDM-SCs is free of multi-user interference. Furthermore, let us assume that the decoding order of the SIC at a receiver does not change as long as each user's minimum SINR requirement (i.e. minimum decoding threshold) is met, which is a commonly-used assumption in the existing works such as [7], [16], [22]. The signal-to-interference-plus-noise rate (SINR) of a user in the hybrid NOMA system can be expressed by

$$\gamma_{k_n, m_n} = \frac{p_{k_n, m_n} |h_{k_n, m_n}|^2}{|h_{k_n, m_n}|^2 \sum_{j \in \mathcal{A}_{m_n, [k_n]}} p_{j, m_n} + \sigma^2},$$

$$\gamma_{k_o, m_o} = \frac{p_{k_o, m_o} |h_{k_o, m_o}|^2}{\sigma^2} \qquad (1)$$

where $k_n \in \mathcal{A}_{m_n}$, $m_n \in \mathcal{M}_n$ and $k_o \in \mathcal{A}_{m_o}$, $m_o \in \mathcal{M}_o$. Assume that, $\mathcal{A}_{m_n}$ and $\mathcal{A}_{m_o}$ contain the indexes of the users assigned to NOMA-SC m_n and OFDM-SC m_o. Note that, $\mathcal{A}_{m_n, [k_n]} = \{\cup j \mid |h_{j, m_n}| > |h_{k_n, m_n}|, \ k_n, j \in \mathcal{A}_{m_n}\}$ is the set of indexes for the users imposing interference on user k_n. In (1), h_{k_n, m_n} and h_{k_o, m_o} are the channel gains following Rayleigh fading, p_{k_n, m_n} and p_{k_o, m_o} are the transmit power for user k_n and user k_o. Further, $\sigma^2 = 1/\gamma_s$ is the variance of

additive Gaussian noise, and γ_s is the average transmit signal-to-noise rate (SNR).

B. Problem Formulation

In this paper, we consider the resource allocation, including UC and power allocation, in the downlink hybrid NOMA systems. In contrast to the existing studies, our resource allocation aims to maximize the EE of the system, and the optimization problem is given by

$$(\mathbb{P}_0): \ \{\mathcal{S}^*, \mathcal{P}^*\} = \arg \max_{\mathcal{S}, \mathcal{P}} \{\eta_{EE}\}$$

$$\text{subject to: } (5) - (8), \ (10) \ (\text{or } (11))$$

where η_{EE} is the EE of the system, and is given by

$$\eta_{EE} = \frac{\sum_{k \in \mathcal{K}} \sum_{m_n \in \mathcal{M}_n} R_{k, m_n} + \sum_{k \in \mathcal{K}} \sum_{m_o \in \mathcal{M}_o} R_{k, m_o}}{\sum_{k \in \mathcal{K}} \sum_{m \in \mathcal{M}} p_{k, m} + P_c}. \qquad (2)$$

The rates R_{k_n, m_n} and R_{k_o, m_o} are

$$R_{k, m_n} = \log_2 \left(1 + \frac{s_{k, m_n} p_{k, m_n} |h_{k, m_n}|^2}{I_{k, m_n} + \sigma^2}\right), \qquad (3)$$

$$R_{k, m_o} = \log_2 \left(1 + \frac{s_{k, m_o} p_{k, m_o} |h_{k, m_o}|^2}{\sigma^2}\right) \qquad (4)$$

where

$$I_{k, m_n} = \sum_{m_n \in \mathcal{M}_n} \sum_{j \in \mathcal{K}, |h_{j, m_n}| > |h_{k, m_n}|} s_{k, m_n} s_{j, m_n} p_{j, m_n} |h_{k, m_n}|^2.$$

The constraints for the optimization problem are expressed as

$$s_{k, m} = \{0, 1\}, \ \forall k \in \mathcal{K}, \ \forall m \in \mathcal{M}, \qquad (5)$$

$$\sum_{k \in \mathcal{K}} s_{k, m} = L_m, \ \forall m \in \mathcal{M}, \qquad (6)$$

$$\sum_{m \in \mathcal{M}} s_{k, m} = 1, \ \forall k \in \mathcal{K}, \qquad (7)$$

$$\sum_{m \in \mathcal{M}} s_{k, m} p_{k, m} \geq \tilde{P}_k^{min}, \ \forall k \in \mathcal{K}, \qquad (8)$$

$$\sum_{k \in \mathcal{K}} s_{k, m} p_{k, m} \leq \hat{P}_m^{max}, \ \forall m \in \mathcal{M}. \qquad (9)$$

Above, $s_{k, m}$ is the UC indicator, $s_{k, m} = 1$ if SC m is allocated to user k, otherwise $s_{k, m} = 0$. Define that, the UC and power allocation variables are contained in the sets of $\mathcal{S}$ and $\mathcal{P}$. In (6), each OFDM-SC is constrained to accommodate one user only, i.e. $L_m = 1$, while each NOMA-SC can be assigned to more than one user, i.e. $L_m > 1$.

In Problem $(\mathbb{P}_0)$, $p_{k, m}$ is the transmit power allocated to user k on SC m. Note that, (8) constrains the minimum transmit power for each user, which actually guarantees the minimum QoS requirement for the user. Further, (9) addresses fairness among SCs, preventing assigning too much transmit power to certain SCs, such as OFDM-SCs. In this paper, we conceive two cases for the total power consuming, which include all the possible scenarios, so that the EE performance of our hybrid NOMA system can be analyzed comprehensively. In

TABLE I
DEFINITIONS OF KEY NOTATIONS & ABBREVIATIONS.

Notation	Definition	Abbreviation	Definition
K	No. of Users	SC	Subcarrier
$\mathcal{K}$	Set of User Indexes	UC	User Clustering
$\boldsymbol{L}$	No. of Users on SCs	LPC	Limited Power Consumption
M	No. of SCs	FPC	Full Power Consumption
$\mathcal{M}$	Set of SC Indexes	EE	Energy Efficiency
$\mathcal{S}$	Set of UC variables	SE	Spectrum Efficiency
$\mathcal{P}$	Set of Power Allocation Variables	Opt-RA	Optimal Resource Allocation
$\mathcal{A}_m$	Set of Users Assigned to SC m	Heur-RA	Heuristic Resource Allocation

the first case, namely limited power consumption (LPC), the total power available is upper bounded by

$$\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} s_{k,m}p_{k,m} \leq P. \qquad (10)$$

The similar assumptions can be found in the existing works of [7], [23]. The LPC case is considered for evaluating the performance of green communication motivated hybrid NOMA systems, in which case, the total transmit power is not always fully used and is optimized according to varying communication environment. By contrast, under the second case, namely full power consumption (FPC), the total power available are fully used, given by

$$\sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} s_{k,m}p_{k,m} = P. \qquad (11)$$

The LPC case is also a widely used assumption for EE maximization, such as [8], [16]. The motivation for studying the FPC case is to evaluate the EE performance of the hybrid NOMA systems in the extreme scenario of transmitting signals with all the transmit power available at all times, which thereby plays as a benchmark for the LPC case. Furthermore, by considering the FPC case, our EE maximization becomes equivalent to maximizing the spectrum efficiency (SE) of the system, since the denominator of the objective function in Problem ($\mathbb{P}_0$) is a fixed value. For practical implementation, we have the assumption that $P \leq \sum_{m\in\mathcal{M}} \hat{P}_m^{max}$.

III. OPTIMAL RESOURCE ALLOCATION ALGORITHMS

In this section, we analyze the general theory of our resource allocation. Then, we propose the optimal resource allocation algorithms for both the LPC and FPC cases.

A. Analysis for Optimal Resource Allocation

Owing to the existence of integer variables and non-linearity of the objective function, Problem ($\mathbb{P}_0$) is a mixed integer nonlinear programming (MINLP) problem, which is NP hard and extremely difficult to solve. In order to solve the problem efficiently, we need to decouple the UC and power allocation, which is similar to the methods in [7]. Hence, Problem ($\mathbb{P}_0$) can be formulated as

$$(\mathbb{P}_1): \quad \{\mathcal{S}^*, \mathcal{P}^*\} = \arg \max_P \left\{ \max_{\mathcal{S}} \{\eta_{EE}\} \right\} \qquad (12)$$

subject to: (5) − (9), and (10) or (11)

where η_{EE} is defined in (2).

For clarity, Problem ($\mathbb{P}_1$) can be further expressed by Problems ($\mathbb{P}_2$) and ($\mathbb{P}_3$), which indicate that the UC and the power allocation are carried out separately. First, for the UC, the problem is given as

$$(\mathbb{P}_2): \quad \{\mathcal{S}^*\} = \arg \max_{\mathcal{S}} \left\{ \frac{\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} R_{k,m}}{\sum_{k\in\mathcal{K}} p_k + P_c} \right\} \qquad (13)$$

subject to (5) − (7).

Above, $R_{k,m}$ is given by (3) or (4). The UC is first carried upon fixing the power allocation, which should be feasible according to (9) and (10) (or (11)).

Based on the UC results, the power allocation is implemented, and expressed as

$$(\mathbb{P}_3): \quad \{\mathcal{P}^*\} = \arg \max_{\mathcal{P}} \{\eta(\mathcal{P}) \mid \mathcal{S}^*\}, \qquad (14)$$

subject to (20) (or (21)) and

$$p_{k,m} \geq \tilde{P}_k^{min}, \ \forall k \in \mathcal{A}_m, \forall m \in \mathcal{M}, \quad (15)$$

$$\sum_{k\in\mathcal{A}_m} p_{k,m} \leq \hat{P}_m^{max}, \ \forall m \in \mathcal{M}, \qquad (16)$$

In Problem ($\mathbb{P}_3$), we have

$$\eta(\mathcal{P}) = \frac{\sum_{m_n\in\mathcal{M}_n, k_n\in\mathcal{A}_{m_n}} R_{k_n,m_n} + \sum_{m_o\in\mathcal{M}_o, k_o\in\mathcal{A}_{m_o}} R_{k_o,m_o}}{\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{A}_m} p_{k,m} + P_c}. \qquad (17)$$

The rates of R_{k_n,m_n} and R_{k_o,m_o} now become

$$R_{k_n,m_n} = \log_2(1 + \gamma_{k_n,m_n})$$
$$= \log_2\left(1 + \frac{p_{k_n,m_n}|h_{k_n,m_n}|^2}{\sum_{j\in\mathcal{A}_{m_n,[k_n]}} p_{j,m_n}|h_{k_n,m_n}|^2 + \sigma^2}\right), \qquad (18)$$

$$R_{k_o,m_o} = \log_2(1 + \gamma_{k_o,m_o})$$
$$= \log_2\left(1 + \frac{p_{k_o,m_o}|h_{k_o,m_o}|^2}{\sigma^2}\right), \qquad (19)$$

where $k_n \in \mathcal{A}_{m_n}$, $m_n \in \mathcal{M}_n$, $k_o \in \mathcal{A}_{m_o}$, $m_o \in \mathcal{M}_o$. Note that in (18), the user, having the highest channel gain on a NOMA-SC, does not suffer any interference. When the LPC case is considered, Problem ($\mathbb{P}_3$) is further constrained by

$$\sum_{k\in\mathcal{A}_m}\sum_{m\in\mathcal{M}} p_{k,m} \leq P, \qquad (20)$$

By contrast, for the FPC case, Problem ($\mathbb{P}_3$) is also constrained by

$$\sum_{k\in\mathcal{A}_m}\sum_{m\in\mathcal{M}} p_{k,m} = P. \qquad (21)$$

In comparison to Problem $(\mathbb{P}_0)$, the decoupled problems $(\mathbb{P}_2)$ and $(\mathbb{P}_3)$ require relatively low complexity to solve, though they are still non-convex problems. Therefore, we propose the optimal resource allocation, given by Algorithm 1, which can find the optimal resource allocation for Problem $(\mathbb{P}_0)$. As in [7], solving the original problem $(\mathbb{P}_0)$ is equivalent to finding the best one among all possible optimal solutions for the decoupled problems $(\mathbb{P}_2)$ and $(\mathbb{P}_3)$. However, we note that the decoupled problem $(\mathbb{P}_2)$ for UC is still an MINLP problem. Therefore, in Algorithm 1, it first finds all the possible UC derived by exhaustive search, as shown by line 1 of Algorithm 1, where N_{perm} is the total number of the possible UC. Then, in the lines 2-11, based on each clustering result it derives the optimal power allocation, which follows Algorithm 2 in Section III-B or Algorithm 3 in Section III-C, respectively, for the LPC or FPC case. In the end, the optimal resource allocation can be found by the one achieving the highest EE value among all possible results.

B. Optimal Power Allocation for Limited Power Consumption case

Upon deriving the UC result $\{\mathcal{S}^*\}$, our power allocation needs to solve Problem $(\mathbb{P}_3)$ with the constraints of (15), (16), and (20) for the LPC case. Due to the existence of interference, we note that the achievable rate of the users on NOMA-SCs, given by (18), is a quasi-concave function. Hence, in the objective function of (17), the numerator of the sum rate becomes a summation of concave functions and quasi-concave functions. Although the non-negative summation of concave functions preserves concavity, this property does not hold for quasi-concave functions, which makes Problem $(\mathbb{P}_3)$ non-concave and is very difficult to solve. In order to solve the non-concave problem of $(\mathbb{P}_3)$, we resort to an efficient approximation method to convert the objective function of (17). Let us first introduce the following theorem.

Theorem 1: For any $\gamma \geq 0$ and $\tilde{\gamma} \geq 0$, the following inequality holds: $\log_2(1+\gamma) \geq a \log_2 \gamma + b$, where a and b are defined as $a = \dfrac{\tilde{\gamma}}{1+\tilde{\gamma}}$, $b = \log_2(1+\tilde{\gamma}) - \dfrac{\tilde{\gamma}}{1+\tilde{\gamma}} \log_2 \tilde{\gamma}$, and the bound is tight for $\gamma = \tilde{\gamma}$.

Proof: This can be found in [24]. ∎

By using the approximation method in Theorem 1, the lower bound for the achievable rate of user k_n on NOMA-SC m_n, and that of user k_o on OFDM-SC m_o, are given by

$$
\begin{aligned}
R_{k_n,m_n} &\geq \widetilde{R}_{k_n,m_n} \\
&= a_{k_n,m_n} \log_2 \left(\frac{p_{k_n,m_n}|h_{k_n,m_n}|^2}{\sum_{j \in \mathcal{A}_{m_n,[k_n]}} p_{j,m_n}|h_{k_n,m_n}|^2 + \sigma^2} \right) \\
&\quad + b_{k_n,m_n},
\end{aligned}
\tag{22}
$$

$$
R_{k_o,m_o} \geq \widetilde{R}_{k_o,m_o} = a_{k_o,m_o} \log_2 \left(\frac{p_{k_o,m_o}|h_{k_o,m_o}|^2}{\sigma^2} \right) + b_{k_o,m_o}.
\tag{23}
$$

In (22) and (23), a_{k_n,m_n}, b_{k_n,m_n}, a_{k_o,m_o} and b_{k_o,m_o} are approximation constants for some SNRs/SINRs of the users to be specified as Theorem 1.

Algorithm 1: Optimal Resource Allocation Algorithm

1:	**UC searching:** With the aid of exhaustive search approach, find all the possible UC results, which are included in $\{\mathcal{S}^{(n)}, n = 1, \ldots, N_{perm}\}$;
2:	**for** $n = 1, \ldots, N_{perm}$
3:	Solve Problem $(\mathbb{P}_3)$ based on the UC of $\mathcal{S}^{(n)}$;
4:	**if** *LPC scenario*
5:	Run Algorithm 2;
6:	**else**
7:	Run Algorithm 3;
8:	**end**
9:	**Output:** the optimal power allocation result $\mathcal{P}^{(n)}$;
10:	Compute the EE value $\eta_{EE}(\mathcal{S}^{(n)}, \mathcal{P}^{(n)})$ according to (2);
11:	**end**
12:	**Output:** the optimal resource allocation result $\{\mathcal{S}^*, \mathcal{P}^*\} = \arg\max_{n=1,\ldots,N_{perm}} \left\{ \eta_{EE}(\mathcal{S}^{(n)}, \mathcal{P}^{(n)}) \right\}$;

Furthermore, to convert Problem $(\mathbb{P}_3)$ into concave, it also needs to introduce the new variables:

$$
q_{k_n,m_n} = \ln p_{k_n,m_n}, \ k_n \in \mathcal{A}_{m_n}, m_n \in \mathcal{M}_o, \text{ and}
$$
$$
q_{k_o,m_o} = \ln p_{k_o,m_o}, \ k_o \in \mathcal{A}_{m_o}, m_o \in \mathcal{M}_o.
\tag{24}
$$

Let us define $\mathcal{Q}_{m_n} = \{q_{k_n,m_n}, \forall k_n \in \mathcal{A}_{m_n}\}$, $\mathcal{Q}_{m_o} = \{q_{k_o,m_o}, \forall k_o \in \mathcal{A}_{m_o}\}$. In that case, the transformed variables can be collected in the vector $\mathcal{Q} = \{\mathcal{Q}_1, \ldots, \mathcal{Q}_M\}$. Upon substituting the new variables $\mathcal{Q}$, the power allocation problem is transformed into

$$
(\mathbb{P}_4): \quad \{\mathcal{Q}^*\} = \arg \max_{\mathcal{Q}} \{\tilde{\eta} \mid \mathcal{S}^*\}
$$

$$
\text{subject to:} \quad e^{q_{k,m}} \geq \tilde{P}_k^{min}, \ \forall k \in \mathcal{A}_m, \forall m \in \mathcal{M}
\tag{25}
$$

$$
\sum_{k \in \mathcal{A}_m} e^{q_{k,m}} \leq \hat{P}_m^{max}, \ \forall m \in \mathcal{M},
\tag{26}
$$

$$
\sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{A}_m} e^{q_{k,m}} \leq P,
\tag{27}
$$

In Problem $(\mathbb{P}_4)$, the objective function can be given by

$$
\tilde{\eta} = \frac{\sum_{m_n \in \mathcal{M}_n} \sum_{k_n \in \mathcal{A}_{m_n}} \widetilde{R}_{k_n,m_n} + \sum_{m_o \in \mathcal{M}_o} \sum_{k_o \in \mathcal{A}_{m_o}} \widetilde{R}_{k_o,m_o}}{\sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{K}} e^{q_{k,m}} + P_c}
\tag{28}
$$

where R_{k_n,m_n} for the user on NOMA-SC and R_{k_o,m_o} for the user on OFDM-SC become

$$
\begin{aligned}
\widetilde{R}_{k_n,m_n} &= a_{k_n,m_n} \log_2 \left(\frac{e^{q_{k_n,m_n}}|h_{k_n,m_n}|^2}{\sum_{j \in \mathcal{A}_{m_n,[k_n]}} e^{q_{j,m_n}}|h_{k_n,m_n}|^2 + \sigma^2} \right) \\
&\quad + b_{k_n,m_n},
\end{aligned}
\tag{29}
$$

$$
\widetilde{R}_{k_o,m_o} = a_{k_o,m_o} \log_2 \left(\frac{e^{q_{k_o,m_o}}|h_{k_o,m_o}|^2}{\sigma^2} \right) + b_{k_o,m_o}.
\tag{30}
$$

When analyzing the above problem, we derive the following important findings in lemma 1.

Lemma 1: Problem $(\mathbb{P}_4)$ for the LPC case is a concave-convex fractional problem, and any of its local maximum is a global maximum and KKT conditions are sufficient to maximum.

Proof: The proof is given by Appendix A. ∎

Furthermore, we apply the Charnes-Cooper transformation [25] to Problem $(\mathbb{P}_4)$, which can be converted to an equivalent concave program. Let us introduce the following lemma.

Lemma 2: Given a ratio Problem $(\mathbb{P}') : \min_{x \in S} \dfrac{f(x)}{g(x)}$, where f is concave and g is convex, $f, g : S \to R$, $S \subseteq R^n$, it can be transformed into a concave program

$$(\mathbb{P}'') : \min_{y/\phi \in S} \quad \phi f(y/\phi), \qquad \text{subject to:} \quad \phi g(y/\phi) \le 1,$$

by using the Charnes-Cooper transformation $y = \dfrac{1}{g(x)}x$, $\phi = \dfrac{1}{g(x)}$, where $\phi > 0$.

Proof: Consider $\phi f(y/\phi)$ and $\phi g(y/\phi)$ as the perspective of the functions f, g, the concavity and convexity preserve, respectively. Therefore, Problem $(\mathbb{P}')$ is a concave problem in (y, ϕ). ∎

Upon leveraging the transformation in lemma 2 for Problem $(\mathbb{P}_4)$, our optimization problem for power allocation under the LPC scenario reduces to the following equivalent problem, as

$$(\mathbb{P}_5) : \{\tilde{\mathcal{Q}}^*, \phi^*\} = \arg \max_{\tilde{\mathcal{Q}}, \phi} \left\{ \phi \left[\sum_{m_n \in \mathcal{M}_n, k_n \in \mathcal{A}_{m_n}} \widetilde{R}_{k_n, m_n} \left(\tilde{\mathcal{Q}}, \phi \right) \right. \right.$$
$$\left. \left. + \sum_{m_o \in \mathcal{M}_o, k_n \in \mathcal{A}_{m_o}} \widetilde{R}_{k_o, m_o} \left(\tilde{\mathcal{Q}}, \phi \right) \right] \right\} \quad (32)$$

$$\text{subject to:} \quad \phi \left(\sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{A}_m} e^{\tilde{q}_{k,m}/\phi} + P_c \right) \le 1, \quad (33)$$

$$\phi \left(\tilde{P}_k^{min} - e^{q_{k,m}} \right) \le 0, \ \forall k \in \mathcal{A}_m, \ \forall m \in \mathcal{M}, \quad (34)$$

$$\phi \left(\sum_{k \in \mathcal{A}_m} e^{\tilde{q}_{k,m}/\phi} - \hat{P}_m^{max} \right) \le 0, \ \forall m \in \mathcal{M}, \quad (35)$$

$$\phi \left(\sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{A}_m} e^{\tilde{q}_{k,m}/\phi} - P \right) \le 0, \quad (36)$$

where

$$\widetilde{R}_{k_o, m_o} \left(\tilde{\mathcal{Q}}, \phi \right) = a_{k_o, m_o} \log_2 \left(\frac{e^{\tilde{q}_{k_o, m_o}/\phi} |h_{k_o, m_o}|^2}{\sigma^2} \right) + b_{k_o, m_o}, \quad (37)$$

$$\widetilde{R}_{k_n, m_n} \left(\tilde{\mathcal{Q}}, \phi \right) = b_{k_n, m_n} +$$
$$a_{k_n, m_n} \log_2 \left(\frac{e^{\tilde{q}_{k_n, m_n}/\phi} |h_{k_n, m_n}|^2}{\sum\limits_{j \in \mathcal{A}_{m_n}, |h_{j,m_n}| > |h_{k,m_n}|} e^{\tilde{q}_{j, m_n}/\phi} |h_{k_n, m_n}|^2 + \sigma^2} \right). \quad (38)$$

Note that, with the aid of the Charnes-Cooper transformation, in Problem $(\mathbb{P}_5)$ we introduce the new variables $\tilde{\mathcal{Q}} = \{\tilde{q}_{k,m}, \forall m \in \mathcal{M}, \forall k \in \mathcal{A}_m\}$ and ϕ, where defining $\tilde{q}_{k,m} = q_{k,m}\phi$. Hence, if the optimal solution $\tilde{\mathbf{q}}$ and ϕ are found, then $\mathcal{Q} = \tilde{\mathcal{Q}}/\phi$ is optimal for Problem $(\mathbb{P}_4)$.

Based on the above analysis, we are now able to propose the optimal power allocation algorithm for the LPC scenario, summarized in Algorithm 2, which finds the optimal solution to the original problem of $(\mathbb{P}_3)$ by iteratively solving Problem $(\mathbb{P}_5)$. Moreover, referring to Algorithm 2, the conclusions in Theorem 2 can be derived.

Theorem 2: Algorithm 2 can derive a converged EE result for the LPC case, and the power allocation solution satisfies the KKT conditions for Problem $(\mathbb{P}_3)$ conditioned on (20).

Proof: The proof is given by Appendix B. ∎

Algorithm 2: Optimal Power Allocation Algorithm for LPC

1:    **Initialization:** (1) Set iteration index $i = 0$, (2) Set maximum iteration number $I_{max}^{(LPC)}$, (3) Set $\epsilon > 0$, (4) Initialize power allocation variables $\mathcal{P}^{(0)}$;
2:    **while** $i <= I_{max}^{(LPC)}$ & $\tilde{\eta}^{(i)} - \tilde{\eta}^{(i-1)} > \epsilon$
3:    Compute $\tilde{\gamma}_{k_n, m_n}^{(i)} = \gamma_{k_n, m_n}(p_{k_n, m_n}^{(i-1)}), \forall k_n, m_n$; $\tilde{\gamma}_{k_o, m_o}^{(i)} = \gamma_{k_o, m_o}(p_{k_o, m_o}^{(i-1)}), \forall k_o, m_o$;
4:    Compute the variables $\{a_{k_n, m_n}^{(i)}, a_{k_o, m_o}^{(i)}, b_{k_n, m_n}^{(i)}, b_{k_o, m_o}^{(i)}, \forall k_n, m_n, k_o, m_o\}$;
5:    Solve the concave optimization problem $(\mathbb{P}_5)$;
6:    **Output:** the optimal variables in $\tilde{\mathcal{Q}}^{(i)} = \{\tilde{q}_{k_n, m_n}^{(i)}, \tilde{q}_{k_o, m_o}^{(i)}, \forall k_n, m_n, k_o, m_o\}$ and the optimization variable $\phi^{(i)}$;
7:    Update the optimization variables in $\mathcal{Q}^{(i)}(\phi^{(i)}, \tilde{\mathcal{Q}}) = \{q_{k_n, m_n}^{(i)}, q_{k_o, m_o}^{(i)}, \forall k_n, m_n, k_o, m_o\}$;
8:    Update the optimization variables in $\mathcal{P}^{(i)}(\mathbf{q}^{(i)})$;
9:    Compute the value of the objective function in (28): $\tilde{\eta}^{(i)}(\mathcal{P}^{(i)})$;
10:   **end**
12:   **Output:** the optimal variables $\mathcal{P}^*$;

C. Optimal Power Allocation for Full Power Consumption case

Let us now consider the power allocation for FPC case, described by Problem $(\mathbb{P}_3)$ constrained by (21), in which the BS is required to transmit signal with full power. In that case, the denominator of the objective function in Problem $(\mathbb{P}_3)$ is a fixed value and, hence, the EE maximization for FPC is equivalent to the maximization of sum rate, yielding

$$(\mathbb{P}_6) : \ \{\mathcal{P}^*\} = \arg \max_{\mathcal{P}} \left\{ R_{\Sigma} = \sum_{m_n \in \mathcal{M}_n, k_n \in \mathcal{A}_{m_n}} R_{k_n, m_n} + \right.$$
$$\left. \sum_{m_o \in \mathcal{M}_o, k_o \in \mathcal{A}_{m_o}} R_{k_o, m_o} \mid \mathcal{S}^* \right\}, \quad (39)$$

subject to: $(15), (16)$ and (21).

Above, R_{k_n, m_n}, expressed as (18), is a quasi-concave function, although R_{k_o, m_o}, defined in (19), is a concave function. Hence, the objective function in (39) is non-concave due to the fact that, a summation of concave functions and quasi-concave functions is non-concave. Hence, Problem $(\mathbb{P}_6)$ is still difficult to solve, and for this kind of problems, the global optimal solution mainly plays as a theoretical benchmark rather than a practical solution. Henceforth, we motivate to solve the problem locally with relatively low complexity.

To obtain a tractable solution, we note that, upon taking the exponential of the objective function in (39), the problem can be transformed to

$$(\mathbb{P}_7) : \{\mathcal{P}^*\} = \arg \max_{\mathcal{P}} \left\{ \prod_{m_n \in \mathcal{M}_n} \prod_{k_n \in \mathcal{A}_{m_n}} (1 + \gamma_{k_n, m_n}) \times \right.$$
$$\left. \prod_{m_o \in \mathcal{M}_o} \prod_{k_o \in \mathcal{A}_{m_o}} (1 + \gamma_{k_o, m_o}) \right\} \quad (40)$$

subject to: $(15), (16)$ and (21),

where γ_{k_n, m_n}, γ_{k_o, m_o} can be found in (18) and (19). Problem

$(\mathbb{P}_7)$ can be equivalently cast as

$$(\mathbb{P}_8) : \{\mathcal{P}^*, \mathcal{T}^*\} = \arg \max_{\mathcal{P}, \mathcal{T}} \left\{ \prod_{m_n \in \mathcal{M}_n} \prod_{k_n \in \mathcal{A}_{m_n}} t_{k_n, m_n} \times \right.$$
$$\left. \prod_{m_o \in \mathcal{M}_o} \prod_{k_o \in \mathcal{A}_{m_o}} t_{k_o, m_o} \right\} \quad (41)$$

subject to: $(15), (16), (21)$ and $(42), (43)$.

$$p_{k_n, m_n} |h_{k_n, m_n}|^2 \geq (t_{k_n, m_n} - 1) \left(\sum_{q \in \mathcal{A}_{m_n, [k_n]}} p_{q, m_n} |h_{k_n, m_n}|^2 \right.$$
$$\left. + \sigma^2 \right), \ \forall k_n \in \mathcal{A}_{m_n}, \forall m_n \in \mathcal{M}_n, \quad (42)$$
$$p_{k_o, m_o} |h_{k_o, m_o}|^2 \geq (t_{k_o, m_o} - 1) \sigma^2, \quad \forall k_o \in \mathcal{A}_{m_o}, \forall m_o \in \mathcal{M}_o. \quad (43)$$

Let us collect the new variables in set $\mathcal{T} = \mathcal{T}_n \cup \mathcal{T}_o$, in which $\mathcal{T}_n = \{t_{k_n, m_n}, \forall k_n \in \mathcal{A}_{m_n}, \forall m_n \in \mathcal{M}_n\}$ and $\mathcal{T}_o = \{t_{k_o, m_o}, \forall k_o \in \mathcal{A}_{m_o}, \forall m_o \in \mathcal{M}_o\}$.

The equivalence of Problem $(\mathbb{P}_8)$ and Problem $(\mathbb{P}_7)$ can be easily recognized by noting the fact that constraint (42) is active at the optimum. Otherwise, it can obtain a strictly larger objective by increasing $\{t_{k_n, m_n}, t_{k_o, m_o}\}$ without violating the constraint. Further, the objective function in Problem $(\mathbb{P}_8)$ can be replaced by its geometric mean, so that it becomes a concave function. Moreover, it is observed that, constraint (42) is non-convex, although constraints (15), (16), (21) and (43) are all convex. For converting (42) to convex, it introduces the slack variables collected in set $\Theta_n = \{\theta_{k_n, m_n}, \forall k_n \in \mathcal{A}_{m_n}, \forall m_n \in \mathcal{M}_n\}$. Therefore, Problem $(\mathbb{P}_8)$ can be reformulated as

$$(\mathbb{P}_9) : \{\mathcal{P}^*, \mathcal{T}^*, \Theta_n^*\} = \arg \max_{\mathcal{P}, \mathcal{T}, \Theta_n} \left(\prod_{m_n \in \mathcal{M}_n} \prod_{k_n \in \mathcal{A}_{m_n}} t_{k_n, m_n} \right.$$
$$\left. \prod_{m_o \in \mathcal{M}_o} \prod_{k_o \in \mathcal{A}_{m_o}} t_{k_o, m_o} \right)^{\frac{1}{K}} \quad (44)$$

subject to: $(15), (16), (21), (43)$ and

$$p_{k_n, m_n} |h_{k_n, m_n}|^2 \geq t_{k_n, m_n} \theta_{k_n, m_n} - \theta_{k_n, m_n},$$
$$\forall k_n \in \mathcal{A}_{m_n}, \forall m_n \in \mathcal{M}_n \quad (45)$$
$$\theta_{k_n, m_n} \geq \sum_{q \in \mathcal{A}_{m_n, [k_n]}} p_{q, m_n} |h_{k_n, m_n}|^2 + \sigma^2,$$
$$\forall k_n \in \mathcal{A}_{m_n}, \forall m_n \in \mathcal{M}_n. \quad (46)$$

In Problem $(\mathbb{P}_9)$, constraint (46) is convex, while constraint (45) is not, due to the bilinear product on the right-hand side. To tackle it, we first use the following transformation

$$t_{k_n, m_n} \theta_{k_n, m_n} = \frac{1}{4} \left[(t_{k_n, m_n} + \theta_{k_n, m_n})^2 - (t_{k_n, m_n} - \theta_{k_n, m_n})^2 \right] \quad (47)$$

which holds for non-negative $\{t_{k_n, m_n}, \theta_{k_n, m_n}\}$. Then, as introduced by [26], it applies the sequential parametric convex approximation (SPCA) method to (47), so that the left-hand side of (47) can be replaced by an upper convex

approximation function. In particular, the upper bound of $-(t_{k_n, m_n} - \theta_{k_n, m_n})^2$ can be obtained by the first-order Taylor series around the points $(\tilde{t}_{k_n, m_n}^{(i)}, \tilde{\theta}_{k_n, m_n}^{(i)})$, due to its concavity. In other words, we have

$$t_{k_n, m_n} \theta_{k_n, m_n} \leq \frac{1}{4} \left(\tilde{t}_{k_n, m_n}^{(i)} + \tilde{\theta}_{k_n, m_n}^{(i)} \right)^2$$
$$- \frac{1}{4} \left[\left(\tilde{t}_{k_n, m_n}^{(i)} - \tilde{\theta}_{k_n, m_n}^{(i)} \right)^2 + 2 \left(\tilde{t}_{k_n, m_n}^{(i)} - \tilde{\theta}_{k_n, m_n}^{(i)} \right) \times \right.$$
$$\left. \left(t_{k_n, m_n} - \tilde{t}_{k_n, m_n}^{(i)} - \theta_{k_n, m_n} + \tilde{\theta}_{k_n, m_n}^{(i)} \right) \right]. \quad (48)$$

Let us define $f(t_{k_n, m_n}, \theta_{k_n, m_n})$ and $g(t_{k_n, m_n}, \theta_{k_n, m_n})$ as the left- and right-hand side of the above inequality. Then, the following results hold

$$f(\tilde{t}_{k_n, m_n}^{(i)}, \tilde{\theta}_{k_n, m_n}^{(i)}) = g(\tilde{t}_{k_n, m_n}^{(i)}, \tilde{\theta}_{k_n, m_n}^{(i)}), \ \nabla f(\tilde{t}_{k_n, m_n}^{(i)}, \tilde{\theta}_{k_n, m_n}^{(i)})$$
$$= \nabla g(\tilde{t}_{k_n, m_n}^{(i)}, \tilde{\theta}_{k_n, m_n}^{(i)}), \quad (49)$$

where ∇f is the gradient of f. Thus, constraint (45) can be approximated as

$$p_{k_n, m_n} |h_{k_n, m_n}|^2 \geq \frac{1}{4} (t_{k_n, m_n} + \theta_{k_n, m_n})^2 - \theta_{k_n, m_n} -$$
$$\frac{1}{4} \left[\left(\tilde{t}_{k_n, m_n}^{(i)} - \tilde{\theta}_{k_n, m_n}^{(i)} \right)^2 + 2 \left(\tilde{t}_{k_n, m_n}^{(i)} - \tilde{\theta}_{k_n, m_n}^{(i)} \right) \times \right.$$
$$\left. \left(t_{k_n, m_n} - \tilde{t}_{k_n, m_n}^{(i)} - \theta_{k_n, m_n} + \tilde{\theta}_{k_n, m_n}^{(i)} \right) \right],$$
$$\forall k \in \mathcal{A}_{m_n}, \forall m_n \in \mathcal{M}_n, \quad (50)$$

where $\tilde{t}_{k_n, m_n}^{(i)}$ and $\tilde{\theta}_{k_n, m_n}^{(i)}$ are constants, and can be given by the optimum solutions of t_{k_n, m_n} and θ_{k_n, m_n} in $(i-1)$th iteration. Hence, (50) becomes convex in the variables of interest. Upon leveraging the above approximations, we can reformulate our power allocation for FPC case as

$$(\mathbb{P}_{10}) : \ \{\mathcal{P}^*, \mathcal{T}^*, \Theta_n^*\} = \arg \max_{\mathcal{P}, \mathcal{T}, \Theta_n} \left\{ \Psi(\mathcal{T}) \right\}, \quad (51)$$

subject to: $(15), (16), (21), (43), (46)$ and (50).

In Problem $(\mathbb{P}_{10})$, the objective function can be given by

$$\Psi(\mathcal{T}) = \left(\prod_{m_n \in \mathcal{M}_n} \prod_{k_n \in \mathcal{A}_{m_n}} t_{k_n, m_n} \prod_{m_o \in \mathcal{M}_o} \prod_{k_o \in \mathcal{A}_{m_o}} t_{k_o, m_o} \right)^{\frac{1}{K}}. \quad (52)$$

Now, the above problem becomes concave, and can be iteratively solved to find the local optimal solution to Problem $(\mathbb{P}_3)$ constrained by (21). Therefore, we propose the power allocation algorithm for FPC case, summarized in Algorithm 3.

In Algorithm 3, let us denote the objective value derived in iteration i by $\Psi^{(i)}(\mathcal{T}^{(i)})$, where $\mathcal{T}^{(i)}$ is the optimal solution to Problem $(\mathbb{P}_{10})$ in iteration i. Further, in order to guarantee the convergence of Problem $(\mathbb{P}_{10})$, the algorithm updates parameters $\tilde{t}_{k_n, m_n}^{(i+1)}, \tilde{\theta}_{k_n, m_n}^{(i+1)}$ by the optimal solutions during for each iteration. Accordingly, we have the following propositions.

Proposition: The proposed power allocation in Algorithm 3 returns a non-decreasing sequence of objective values, i.e., $\Psi^{(i+1)} \geq \Psi^{(i)}, \forall i$, and hence it converges.

Proof: The proof is provided in Appendix C. ∎

Algorithm 3: Optimal Power Allocation Algorithm for FPC

1: **Initialization:** (1) Set iteration index $i = 0$; (2) Set maximum iteration number $I_{max}^{(FPC)}$; (3) Set $\epsilon > 0$; (4) Initialize parameters $\tilde{t}_{k_n,m_n}^{(0)}, \tilde{\theta}_{k_n,m_n}^{(0)}, \forall k_n, \forall m_n$; (5) Initialize variables $\{\mathcal{P}^{(0)}, \mathcal{T}_n^{(0)}, \mathcal{T}_o^{(0)}, \Theta_n^{(0)}\}$;
2: **while** $i <= I_{max}^{(FPC)}$ & $\Psi^{(i)} - \Psi^{(i-1)} > \epsilon$
3:    Solve the optimization problem $(\mathbb{P}_{10})$;
4:    **Output:** the solution in sets $\{\mathcal{P}^{(i)}, \mathcal{T}_n^{(i)}, \mathcal{T}_o^{(i)}, \Theta_n^{(i)}\}$;
5:    Compute the objective value $\Psi(\mathcal{T}^{(i)})$ according to (52);
6:    Update the parameters for problem $(\mathbb{P}_{10})$: $\tilde{t}_{k_n,m_n}^{(i+1)} \leftarrow t_{k_n,m_n}^{(i)}, \tilde{\theta}_{k_n,m_n}^{(i+1)} \leftarrow \theta_{k_n,m_n}^{(i)}, \forall k_n \in \mathcal{A}_{m_n}, \forall m_n \in \mathcal{M}_n$;
7:    Update: $i \rightarrow i + 1$;
8: **end**
9: **Output:** the optimal solutions $\mathcal{P}^*$;

Algorithm 4: Heur-RA Algorithm

1: **Initialization:** Set $n = 1$; Set $\mathcal{C}_m = \emptyset, \check{\mathcal{K}}_m = \mathcal{K}, \forall m \in \mathcal{M}$; Set $\check{\mathcal{M}}_k = \mathcal{M}, \forall k \in \mathcal{K}$;
2: **while** $n \leq N_{max}$
3:    Identify the best candidate from SC's perspective:
   $\mathcal{C}_m \leftarrow \mathcal{C}_m \cup \{\check{k}\}$ where $\check{k} = \arg\max_{k \in \check{\mathcal{K}}_m}\{|h_{m,k}|^2\}$, $\forall m \in \mathcal{M}$;
4:    Identify the best candidate from user's perspective:
5:    $\mathcal{C}_{\check{m}} \leftarrow \mathcal{C}_{\check{m}} \cup \{k\}$ where $\check{m} = \arg\max_{m \in \check{\mathcal{M}}_k}\{|h_{m,k}|^2\}$, $\forall k \in \mathcal{K}$;
6:    **if** $|\mathcal{C}_m| \geq L_m, \forall m \in \mathcal{M}$
7:      **User Allocation**: *Run* Algorithm 5;
8:      **if**
9:        *Run* Algorithm 2;
10:     **else**
11:       *Run* Algorithm 3;
12:     **end**
13:     **Output:** the optimal power allocation result $\mathcal{P}^{(n)}$
14:     Compute the EE value $\eta_{EE}(\mathcal{S}^{(n)}, \mathcal{P}^{(n)})$ according to (2);
15:     Update: $n \leftarrow n + 1$;
16:    **end**
17: **end**
18: **Output:** the final resource allocation result $\{\mathcal{S}^*, \mathcal{P}^*\} = \arg\max_{n=1...N_{max}}\{\eta_{EE}(\mathcal{S}^{(n)}, \mathcal{P}^{(n)})\}$;

Furthermore, although the above proposition proves that the proposed algorithm converges to a finite objective value, it is not possible to prove convergence to a global maximum due to the non-convex property of the original problem of $(\mathbb{P}_6)$ (i.e. Problem $(\mathbb{P}_3)$ constrained by (21)). However, we can prove that under some regularity conditions, the accumulation point of the sequence of solutions satisfies the KKT conditions, as summarized in the following proposition.

Proposition: As the iteration number i tends to infinity, the solutions given by Algorithm 3 converges to the KKT points of the original problem of $(\mathbb{P}_6)$.

Proof: The proof is provided in Appendix D. ∎

IV. HEURISTIC RESOURCE ALLOCATION

In this section, we will develop a low-complexity heuristic resource allocation algorithm, namely Heur-RA, which aims for practical implementation. Then, comprehensive complexity analysis will be carried out for the proposed algorithm, as well as for other existing algorithms.

A. Proposed Heur-RA Algorithm

The Opt-RA algorithm in Algorithm I demands very high complexity, which is mainly caused by using the exhaustive search approach of UC. Specifically, when supporting a relative high number of users the system, the implementation of the optimal algorithm will become very challenging. Hence, we propose the suboptimal Heur-RA algorithm with low complexity required. The Heur-RA aims to find promising suboptimal solutions to Problem $(\mathbb{P}_1)$ by separately solving the UC and power allocation problems. In particular, our Heur-RA algorithm is first to find an efficient approach for UC, based on which, the optimal power allocation is carried out.

Following the analysis in Section III, our Heur-RA algorithm aims to develop the user allocation that can find promising suboptimal solutions to the MINLP problem of $(\mathbb{P}_2)$. As suggested by [27], solving Problem $(\mathbb{P}_2)$ can be approximated by solving the following problem

$$(\mathbb{P}_{11}): \{\mathcal{S}^*\} = \arg\max_{\mathcal{S}}\left\{\frac{\sum_{m \in \mathcal{M}} s_{k,m}|h_{k,m}|^2}{2\sigma^2}, \forall k \in \mathcal{K}\right\} \tag{53}$$

$$\text{subject to } (5) - (7)$$

which is though a MINLP problem. Hence, the suboptimal UC can be designed to maximize the SNR of each user. Our heuristic resource allocation is now motivated to find solutions to Problem $(\mathbb{P}_1)$ by solving the decoupled problems of $(\mathbb{P}_{11})$ and $(\mathbb{P}_3)$.

Let us now introduce the Heur-RA algorithm, and the principles are summarized in Algorithm 4. During each iteration, given by lines 3-7 of Algorithm 4, the UC follows the candidate search-and-allocation approach, which motivates to maximize the SNR of each user. Then, based on the derived UC result, the power allocation is implemented according to Algorithm 2 or 3 (under the LPC or FPC case) for maximizing the EE of the system, which follows the steps in lines 8-12. After that, it computes the EE value in the end of each iteration for UC and power allocation. The algorithm outputs the final resource allocation that maximizes the EE, which is shown in line 18. In order to keep the implementation complexity for searching candidates as low as possible, we can set up the maximum number of iterations N_{max}, according to practical system requirements.

Our UC is motivated to maximize the SINR of the worst user as well as to maximize the SINR of the best user, which is equivalent to maximize the SINR of each user. For the UC, the relating notations in Algorithm 4 are defined as follows. Note that n is the iteration index for UC. Accordingly, $\mathcal{C}_m$ includes the indexes of the candidate users for SC m. By contrast, $\check{\mathcal{K}}_m$ is the set of the unselected candidates for SC m, and $\check{\mathcal{M}}_k$ is the set of the unselected candidates for user k. Let us explain the principles of the UC by focusing on one iteration of Algorithm 4. The UC first searches candidate searching given by lines 3-5 of Algorithm 4, which is then followed by the user allocation process given by Algorithm 5. In particular, the candidate search is for finding the best potential options for our UC, as well as for avoiding the unwanted ones, so that the

---

Algorithm 5: User Allocation Stage of Heur-RA Algorithm

1:     **Initialization**: Set $\mathcal{C}_m^{(n)} = \mathcal{C}_m, \mathcal{A}_m^{(n)} = \emptyset, \forall m \in \mathcal{M}$;
       Set $\tilde{\mathcal{M}} = \mathcal{M}, \tilde{\mathcal{K}} = \mathcal{K}$;

2:     **for** $k = 1 : K$

3:       **if** $\mathcal{C}_m^{(n)} = \emptyset, \forall m \in \mathcal{M}$

4:          $\Phi = \tilde{\mathcal{M}}$;

5:       **else**

6:          Find SC/SCs with least number of candidates:
           $\Phi = \{m \mid m = \arg\min_{i \in \mathcal{M}} \{|\mathcal{C}_i^{(n)}|\}\}$;

7:       **end**

8:       Search the SC/SCs, in Φ, having minimum number
        of users: $\tilde{\Phi} = \{m \mid m = \arg\min_{i \in \Phi}\{L_i - |\mathcal{A}_i^{(n)}|\}\}$;

9:       Identify the SC and its best user:
        $\{m^*, k^*\} = \arg\max_{m \in \tilde{\Phi}, k \in \mathcal{V}} \{|h_{m,k}|^2\}$,
        where $\mathcal{V} = \mathcal{C}_m^{(n)}$ if $\mathcal{C}_m^{(n)} \neq \emptyset$; Otherwise, $\mathcal{V} = \tilde{\mathcal{K}}$;

10:     Allocate user for the identified SC: $\mathcal{A}_{m^*}^{(n)} \leftarrow \mathcal{A}_{m^*}^{(n)} \cup \{k^*\}$;

11:     Update: $\mathcal{C}_m^{(n)} \leftarrow \mathcal{C}_m^{(n)} - \{k^*\}, \forall m \in \tilde{\mathcal{M}}$; $\tilde{\mathcal{K}} \leftarrow \tilde{\mathcal{K}} - \{k^*\}$;
       $\tilde{\mathcal{M}} \leftarrow \tilde{\mathcal{M}} - \{m^*\}$;

12:    **end**

---

allocation process can be completed with minimum number of iterations. In lines 3 and 4 of Algorithm 4, it searches the best candidates having the highest channel qualities from both SC's and user's perspectives, which is for the sake of maximizing selecting diversity. Then, in line 5, it updates the relating sets $\tilde{\mathcal{M}}_k, \tilde{\mathcal{K}}_m$ for the unselected candidates. After that, it checks if the condition of $|\mathcal{C}_m| \geq L_m, \forall m \in \mathcal{M}$ is met. If yes, meaning that each SC has the number of candidates required by allocation, it then proceeds to the user allocation process based on Algorithm 5.

The principles of the proposed user allocation are summarized in Algorithm 5. Based on the candidate identification, there are K user allocation iterations of each allocating one candidate or non-candidate user to a SC. To maximize the SNR of each user, the user allocation stage is to first assign the candidate users. When candidate allocation is available, by following line 6 of Algorithm 5, it identifies the SCs having the least number of candidates unassigned, which are included in set Φ with the highest allocation priority. This is due to that, the number of candidate allocations can be maximized. On the contrary, to maximize the minimum channel qualities assigned, in line 8 it finds the SCs collected in $\tilde{\Phi}$ that can accommodate the least number of users. Then, shown by lines 9 and 10, the algorithm allocates the best candidate to the SC in $\tilde{\Phi}$, which aims to maximize the best channel quality of each SC. When all candidates have been assigned, Algorithm 5 proceeds to non-candidate allocation if the allocation is not completed. In this case, the algorithm proceeds to line 4, i.e. $\Phi = \tilde{\mathcal{M}}$, containing indexes of the unassigned SCs without candidates available. Correspondingly, by following lines 8-11, the non-candidate allocation process can be carried out in a similar way as that of candidate allocation.

B. Complexity Analysis

In this section, let us analyze the complexity required by the Heur-RA algorithm, which is compared with those required by the optimal algorithms and the other existing algorithms. For clarity, the complexity analyzed in this paper is reflected by the number of comparisons required by the RA algorithms.

In the context of the Opt-RA algorithm, the UC is based on exhaustive search approach. Therefore, the total number of all possible UC is

$$N_{perm} = \binom{L_1}{K} \prod_{m=2}^{M} \binom{L_m}{K - \sum_{i=1}^{m} L_i}. \tag{54}$$

For each of the N_{perm} UC results, the Opt-RA algorithm needs to operate the power allocation (Algorithm 2 or 3). In the context of the LPC, shown by Algorithm 2, the optimal power allocation is obtained by iteratively solving the convex problem of ($\mathbb{P}_5$) the aid of the interior point method, which requires the complexity of $\mathcal{O}(r^4\delta)$ [28], [29]. Note that, r is the number of variables and δ is the number of bits for representing the coefficients in the optimization problem. Similarly, when considering the FPC scenario, the required complexity is $\mathcal{O}(I_{max}^{(FPC)}(3K)^4\delta^{(FPC)})$ for obtaining the optimal power allocation solution. In summary, the complexity of the Opt-RA algorithm can be given by $\mathcal{O}(N_{perm}I_{max}^{(i)}(K+1)^4\delta^{(i)})$, $i \in \{LPC, FPC\}$. Note that, $I_{max}^{(LPC)}$ and $I_{max}^{(FPC)}$ are respectively the maximum iterations for the power allocation algorithms.

Let us now analyze the complexity of UC under the Heur-RA algorithm. The proposed UC in Algorithm 4 needs to identify candidates from both SC's and user's perspectives, which is equivalent to ordering the K channels of each SC as well as ordering the M channels of each user. Hence, the number of comparisons required by candidate searching is upper bounded by $2KM(\ln K + \ln M)$ [30]. Further, the UC needs to check if each SC has enough candidates before user allocation, which, at most, requires $N_{max}M$ comparisons. Known from Algorithm 5, it identifies the sets of Φ and $\tilde{\Phi}$ during the user allocation, which requires $2K(M-1)$ for each iteration. In total, the number of comparisons required by the UC of the heuristic algorithm is upper bounded by

$$\Lambda^{(\text{Heur-RA-UC})} \leq 2KM(\ln K + \ln M) + 2N_{max}K(M-1) \\ + N_{max}M \tag{55}$$

which results in a complexity of $\mathcal{O}(KM\ln K)$. In addition to UC, the heuristic algorithm operates the power allocation for each of the N_{max} UC results identified. Upon summing up the above, the Heur-RA requires a complexity of $\mathcal{O}(\max\{KM\ln K, N_{max}I_{max}^{(i)}(K+1)^4\delta^{(i)}\})$, $i \in \{LPC, FPC\}$.

For comparison, we consider two widely-used algorithms: Greedy-RA and Hungarian-RA algorithms, which employ the greedy algorithm [31] and Hungarian algorithm [32] for UC (from both user and SC perspectives), respectively. Based on the UC, they both use the optimal power allocation of Algorithms 2 or 3. Specifically, the complexity of Greedy-RA is $\mathcal{O}(\max\{K^2, N_d I_{max}^{(i)}(K+1)^4\delta^{(i)}\})$, $i \in \{LPC, FPC\}$, while the complexity of Hungarian-RA is $\mathcal{O}(\max\{K^3, N_d I_{max}^{(i)}(K+1)^4\delta^{(i)}\})$, $i \in \{LPC, FPC\}$. Note that, we have $N_d = 2$, since both them operate the UC form the user and SC perspectives. Known from the simulation results in Section V, the maximum number of iterations N_{max} for Heur-RA to operate the UC is very small, and $N_{max} < 3$ for most scenarios. In that case, it implies that, the complexity

(a) EE versus γ_s
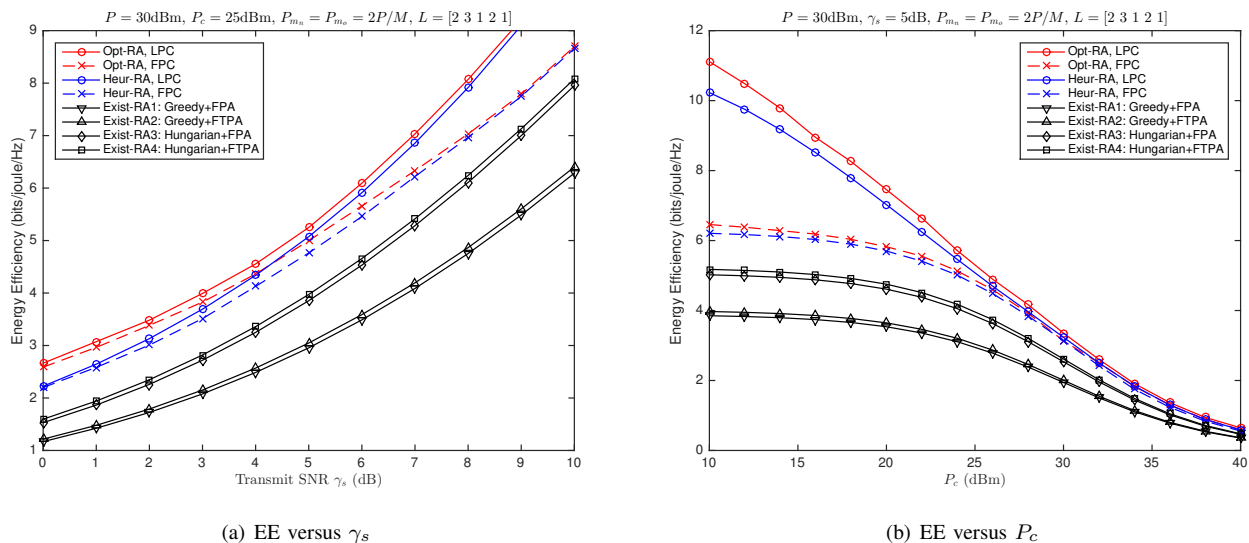
(b) EE versus P_c

Fig. 2. EE of downlink hybrid NOMA systems employing various RA schemes, when different γ_s and P_c values are applied.

TABLE II
KEY SIMULATION PARAMETERS.

Parameter	Value	Parameter	Value
K	9	$\boldsymbol{L}$	[2 3 1 2 1]
M	5	P	30 dBm
$\bar{P}_k^{min}$	10 dBm	$\hat{P}_m^{max}$	40 dBm
P_c	25 dBm	α	0.3

required by the Heur-RA is much lower than that of the Opt-RA and the Hungarian-RA, and it is similar to that for low-complexity Greedy-RA algorithm. Hence, we may conclude that the Heur-RA is a low complexity algorithm, which is a promising solution for practical implementation.

V. NUMERICAL RESULTS

In this section, we provide a range of simulation results for demonstrating the achievable EE performance of the down-link hybrid NOMA systems employing the proposed Opt-RA, Heur-RA algorithms as well as the other existing RA schemes. In our simulations, we assume that each user on a SC experiences independent flat Rayleigh fast fading, and the noise variance on each user is the same. For comparison, we consider four existing RA schemes, which employs the Greedy and Hungarian algorithms for UC. Moreover, fixed power allocation (FPA) and fractional transmit power allocation (FTPA) are used by the existing RA schemes. In our simulations, if it is not specified, we have the key simulation parameters summarized in Table II.

In Fig. 2, it depicts the average EE versus transmit SNR and constant power consumption. First of all, our proposed RA algorithms including Heur-RA and Opti-RA can significantly outperform the existing RA schemes, and the performance gain gets larger at low γ_s and low P_c regions. Second, from the figure, it shows, for both the LPC and FPC scenarios, the performance gap between the Heur-RA and the optimal benchmark (i.e. Opti-RA) is minimized at high γ_s and high P_c. Third, we observe that, the LPC scenario always outperforms

the FPC scenario. Further, Fig. 2 (a) shows that the gaps between LPC and FPC increase when γ_s increases. This implies that the system should use less transmit power to achieve the best EE when channels become more noisy. At last, seen from Fig. 2 (b), the EE performance achieved by FPC is significantly outperformed by LPC. This reveals, for achieving energy efficient systems, it requires to adaptively adjust the total transmit power P if constant power consumption P_c does not play a dominating role.

Let us now discuss the EE versus P performance of the hybrid NOMA systems in Fig. 3. We once again observe that the proposed RA algorithms can hugely outperform the existing algorithms regardless of the two scenarios. It should be note that, the FPC scenario also reflects the throughput performance of the system and hence, our proposed algorithms are capable of finding a good EE and throughput trade-off. In conclusion, observed from both Fig. 2 and Fig. 3, the Opt-RA provides the benchmark for the EE of the hybrid NOMA systems, and Heur-RA algorithm with low complexity required could be one promising solution to the future hybrid NOMA systems. Let us first discuss the results in Fig. 3 (a). First, we observe that the proposed algorithms under FPC and the existing algorithms have bell-shape EE performance. This implies that increasing the transmit power will not always resulting an energy efficient system. Second, the EE for LPC scenario almost coincide with that for FPC if $P \leq 28$dBm, which indicates the hybrid NOMA system of energy efficient demands full transmit power before reaching the best EE. Third, it observes that the algorithms under the LPC have the performance floors as P increases, which accounts for our hybrid NOMA system could gain the best EE as long as the total transmit power is greater than the certain values. The reasons behind can be explained as follows. When the total transmit power P used exceeds a relatively large value, the SE (or sum rate) can not be improved due to the fact that interference signals become the dominant factor of the SINR of a user. Observed from the optimization problem of $(\mathbb{P}_0)$,
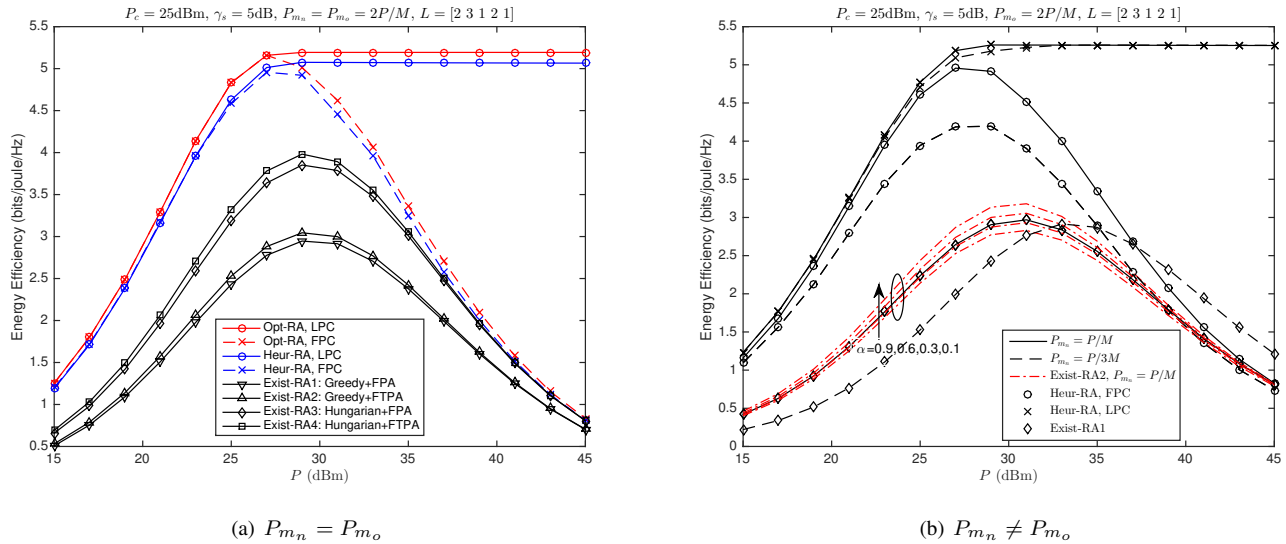
Fig. 3. EE of downlink hybrid NOMA systems using various RA schemes, when different total power P are available.

it will stop increasing the use of the total transmit power in order to obtain the maximum objective EE value, even if more and more transmit power is available.

In contrast to Fig. 3 (a), Fig. 3 (b) evaluates the EE performance of the hybrid NOMA systems, where varying the per-NOMA-SC power constraint P_{m_n} and fixing per-OFDM-SC power constraint P_{m_o}. In general, seen from Fig. 3 (b), when each NOMA-SC has more transmit power available, the EE performance increases regardless of the RA schemes employed by the systems. Further, as P_{m_n} increases, the performance improvement for FPC is huge, while a slight improvement for LPC can only be observed at the optimal P values. This observation implies, the users on NOMA-SCs require more power as the total transmit power increases. In this figure, the EE achieved by the FTPA increases as α becomes smaller. Hence, the practical hybrid NOMA system needs to find a good balance between fairness and EE performance. From Fig. 3, we may conclude that, it is crucial to find the best trade-off between EE and total transmit power when designing practical hybrid NOMA systems in future. Furthermore, as mentioned in Section II.B, by considering the FPC case, it can directly reflect the SE performance of the hybrid NOMA systems. By comparing the performance of the FPC and LPC cases in Figs. 2 and 3, we may further conclude that, 1) upon fixing the total transmit power P, it will have a positive impact on the EE performance when monotonically improving the SE performance of the systems; by contrast, 2) the SE and EE may become two conflicting performance metrics if the total transmit power P keep increasing after a certain threshold, which is a vital parameter and should be optimized in practical hybrid NOMA systems.

Fig. 4 shows the convergence results of the proposed optimal power allocation algorithms for the LPC and FPC scenarios. Note that, the results are derived by averaging the Monte-Carlo simulations with 10^4 realizations at each iteration. First of all, we observe that, for the specific simulation cases, our proposed power allocation algorithms are capable of converging to the optimums within very few number of

iterations, which is smaller than 10. Fig. 4 (a) depicts that, for LPC, the proposed algorithm requires more iterations to converge when P increases or P_c decreases. By contrast, Fig. 4 (b) shows that, the algorithm for FPC demands slightly more iterations only when P_c increases, which is because the full power is always used.

To demonstrate the advantages of our Heur-UC algorithm, Fig. 5 compares the EE of the systems employing various UC algorithms. Without loss of generality, in the figure, we apply the FPA approach to the systems, where considering three cases with varying UC requirements of $\tilde{L}$. It is observed from Fig. fig-EE-UC (a) that our Heur-UC algorithm approaches the high-complexity Hungarian algorithm and outperforms the famous greedy algorithm. The performance gap between Heur-UC and Hungarian decreases, while that between Heur-UC and greedy becomes larger, when the system supports more number of users and SCs, i.e. larger size of $\tilde{L}$. Further, the EE of Heur-UC coincides with that of Hungarian algorithm when P_{m_n} reduces to $P/3M$. This is because, when the available power on each NOMA-SC becomes smaller, the EE performance is dominated by the user with the best channel condition on every SC. Observed from Fig. 5 (b), the EE achieved by Heur-UC algorithm is approaching that for Hungarian algorithm as noise variance becomes smaller. The other observations are similar to those in Fig. 5 (a). In conclusion, the above observations imply that the proposed Heur-UC with low complexity can achieve a promising selecting diversity for UC in the hybrid NOMA systems.

At last, Fig. 6 compares the EE obtained by the hybrid NOMA with that for the OMA systems. First, we clearly observe that the hybrid NOMA is superior to conventional OMA when the same RA scheme is used. Moreover, by comparing the three cases of L, we observe that the EE performance is slightly degraded as a compromise for supporting more number of high-SE users on OFDM-SCs. Once again, the figure confirms the performance gain achieved by our Heur-UC over the existing greedy algorithm. From the above observations, we may conclude that, the hybrid NOMA
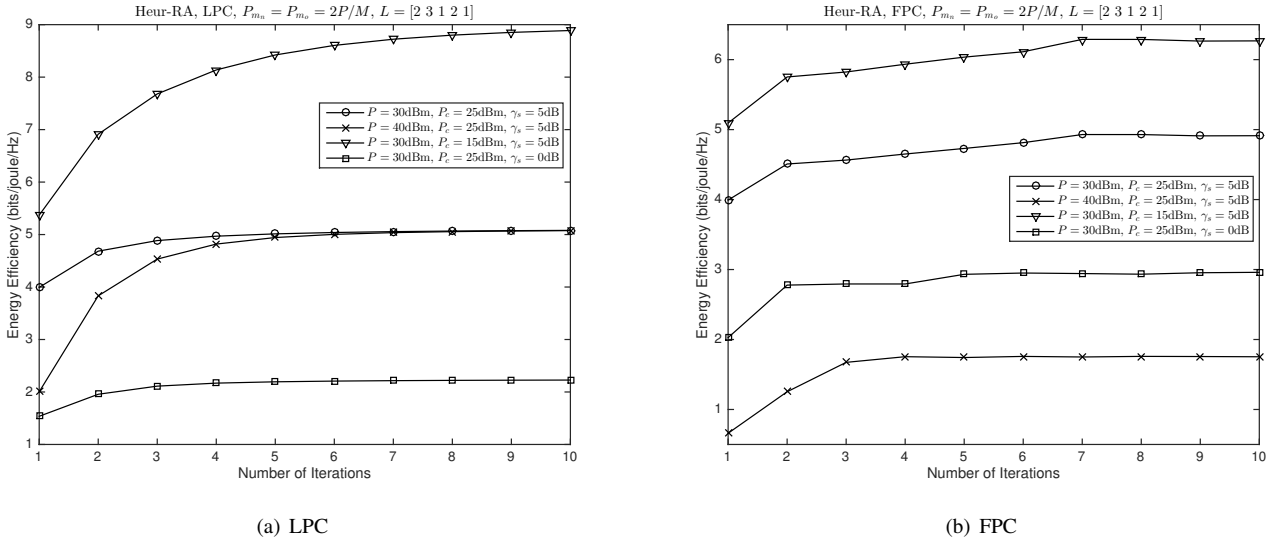
(a) LPC

(b) FPC

Fig. 4. Convergence of the proposed optimal power allocation algorithms under the LPC and FPC scenarios.
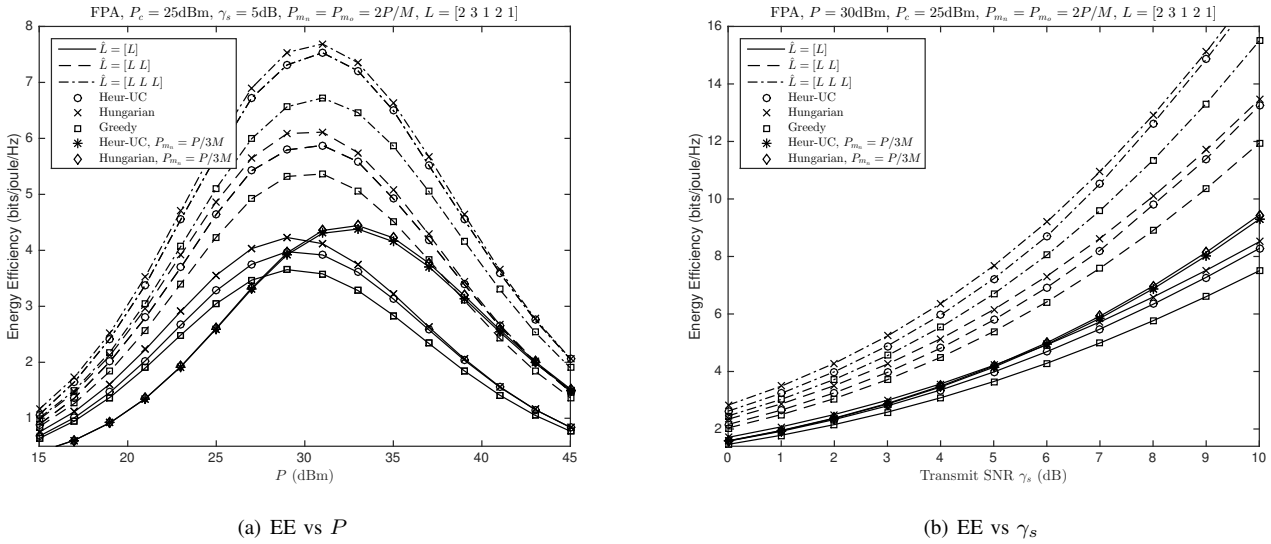

(a) EE vs P

(b) EE vs γ_s

Fig. 5. EE of downlink hybrid NOMA systems using FPA and various UC, when different P and γ_s.

is a promising solution to future communications, owing to the facts including: exploiting the advantages of NOMA technique, high flexibility for system configuration, as well as making compromise between massive connectivity and high throughput services.

VI. CONCLUSION

In this paper, we have investigated EE maximizing motivated resource allocation, including jointly UC and power allocation, in the downlink hybrid NOMA systems. The Opt-RA algorithm has been proposed to find the optimal solutions for both the LPC and FPC scenarios. Further, we have proposed the Heur-RA algorithm to find suboptimal solutions with relatively low complexity required. Our simulation results have shown that, the Heur-RA algorithm achieves significantly higher EE than other existing algorithms, and its performance is close to that obtained by the Opt-RA. Moreover, our results have also implied that, the hybrid NOMA conceived is able to

exploit the advantages of NOMA scheme, and is superior to conventional OMA, and achieves higher flexibility for system configuration than NOMA. In conclusion, the hybrid NOMA associated with the proposed RA schemes may constitute a promising candidate for future wireless systems.

APPENDIX A: PROOF OF LEMMA 2

After some simplification steps, (29) can be equivalently transformed to

$$
\widetilde{R}_{k_n,m_n} = a_{k_n,m_n} \frac{1}{\ln 2} q_{k_n,m_n} + \log_2 |h_{k_n,m_n}|^2 - a_{k_n,m_n} \times
$$

$$
\log_2 \left(\sum_{j \in \mathcal{A}_{m_n}, |h_{j,m_n}| > |h_{k_n,m_n}|} e^{q_{j,m_n}} |h_{k_n,m_n}|^2 + 2\sigma^2 \right)
$$

$$
+ \beta_{k_n,m_n}, \tag{56}
$$

which can be proved to be concave in $\mathbf{q}$, due to the fact that the log-sum-exp function is convex. Therefore, the sum rate
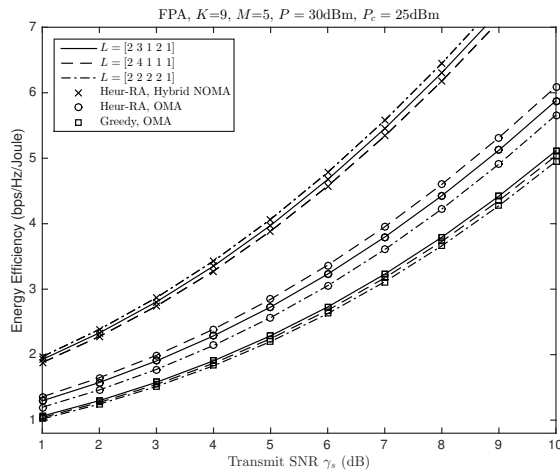
Fig. 6. Comparison of EE of hybrid NOMA systems with that of OMA systems, when using various transmit SNR values.

of the users on NOMA-SCs, i.e. the first sum term of the numerator in the objective function (28), is concave in $\mathbf{q}$. We also note that (30) is actually an affine function in $\{\mathbf{q}_{k_o,m_o}\}$. Hence, the numerator of the objective function (28) is concave in $\mathbf{q}$. Further, we can easily prove that the denominator of (28) is convex in $\mathbf{q}$. Let us now define the feasible sets of the solutions as $\Omega = \{\mathbf{q} \in \mathbb{R}^{K \times 1} \mid \sum_{k \in \mathcal{A}_m} e^{q_{k,m}} \leq P_m, \forall m, \sum_m \sum_k e^{q_{k,m}} \leq P, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}\}$, which can be found to be a convex set. Therefore, we conclude that, the transformed problem $\mathbb{P}_4$ is a concave-convex fractional programming problem, because that, the numerator and denominator of the objective function are concave and convex respectively, and the feasible set Ω is a convex set. Moreover, we find that, the objective function in (28) is pseudo-concave, since its denominator and the numerator are differentiable. This implies that any local maximum is a global maximum and that KKT conditions are sufficient to the maximum solution.

APPENDIX B: PROOF OF THEOREM 2

Let us first explain Algorithm 2 can derive a converged EE value. In iteration i, we find the optimal solutions $\mathcal{Q}^{(i)}$ to Problem ($\mathbb{P}_4$) by solving Problem ($\mathbb{P}_5$), thereby deriving the EE value of $\tilde{\eta}(\mathcal{Q}^{(i)})$ according to (28). As $p_{k,m}^{(i)} = e^{q_{k,m}^{(i)}}, \forall k, m$, it finds out the EE value $\eta(\mathcal{P}^{(i)})$ according to (17). When applying Theorem 1, we have the following result

$$\eta(\mathcal{P}^{(i)}) = \tilde{\eta}(\mathcal{Q}^{(i)}), \; i \leq I_{max} \qquad (57)$$

which holds for $\gamma_{k,m}^{(i)} = \tilde{\gamma}_{k,m}^{(i)}, \forall k, m$.

Further, with the aid of interior point method, the global optimal solution to the concave problem of ($\mathbb{P}_5$) can be found. Therefore, the following result can be derived

$$\tilde{\eta}(\mathbf{q}^{(i)}) \leq \tilde{\eta}(\mathbf{q}^{(i+1)}), \; i \leq I_{max}. \qquad (58)$$

Known from (57) and (57), Algorithm 2 iteratively solves Problem ($\mathbb{P}_5$) and updates the EE values, which can be improved at each iteration and finally converges.

Since Problem ($\mathbb{P}_5$) is concave in $(\mathcal{Q}, \phi)$, KKT conditions are sufficient and necessary for the optimal solution. Set ν,

$\lambda_{k,m}$, where $k \in \mathcal{K}, m \in \mathcal{M}$, and μ as the Lagrange multiplier, which satisfy $\{\nu, \mu, \hat{\lambda}_m, \tilde{\lambda}_{k,m}, \forall k, m\} \in \mathbb{R}_+, \mathbb{R}_+ \equiv [0, \infty]$, the Lagrangian function can be expressed as

$$\mathcal{L} = -\sum_{k \in \mathcal{A}_m} \sum_{m \in \mathcal{M}} \tilde{\lambda}_{k,m} \left[\phi \left(\tilde{P}_k^{min} - e^{q_{k,m}} \right) \right] +$$

$$\phi \left[\sum_{m_n \in \mathcal{M}_n, k_n \in \mathcal{A}_{m_n}} \widetilde{R}_{k_n, m_n} \left(\tilde{\mathcal{Q}}, \phi \right) + \sum_{m_o \in \mathcal{M}_o, k_o \in \mathcal{A}_{m_o}} \widetilde{R}_{k_o, m_o} \left(\tilde{\mathcal{Q}}, \phi \right) \right]$$

$$-\mu \left[\phi \left(\sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{A}_m} e^{\frac{\tilde{q}_{k,m}}{\phi}} - P \right) \right] - \nu \left[\phi \left(\sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{A}_m} \exp(\frac{\tilde{q}_{k,m}}{\phi}) \right) \right.$$

$$\left. +P_c \right) - 1] - \sum_{m \in \mathcal{M}} \hat{\lambda}_m \left[\phi \left(\sum_{k \in \mathcal{A}_m} e^{\frac{\tilde{q}_{k,m}}{\phi}} - \hat{P}_m^{max} \right) \right]. \qquad (59)$$

According to (59), we can derive the KKT conditions for Problem ($\mathbb{P}_5$), which are omitted here due to the lack of space. As calculated by Algorithm 2, the optimal solutions of $\{\tilde{\mathcal{Q}}^*, \phi^*\}$ can be obtained when solving Problem ($\mathbb{P}_5$) at convergence, along with which, it finds the multipliers $\{\nu^*, \mu^*, \hat{\lambda}_m^*, \tilde{\lambda}_{k,m}^*, \forall k, m\}$. Note that, the results of $\{\tilde{\mathcal{Q}}^*, \phi^*, \nu^*, \mu^*, \hat{\lambda}_m^*, \tilde{\lambda}_{k,m}^*, \forall k, m\}$ can satisfy the KKT conditions for Problem ($\mathbb{P}_5$). Upon substituting $\mathcal{P}^* = \exp\{\mathcal{Q}^*/\phi^*\}$, we can deduce the KKT conditions for Problem ($\mathbb{P}_3$). The approximation in Theorem 1 is exact at convergence, it knows $\{\tilde{\mathcal{Q}}^*, \phi^*, \nu^*, \mu^*, \hat{\lambda}_m^*, \tilde{\lambda}_{k,m}^*, \forall k, m\}$ satisfies Slater conditions for Problem ($\mathbb{P}_3$). Known from [33], KKT conditions are the first-order necessary conditions for any maximum of Problem ($\mathbb{P}_3$) if Slater's constraint qualification satisfies. Hence, the proof is completed.

APPENDIX C: PROOF OF PROPOSITION 1

To show the sequence of the objective values is non-decreasing, it needs to first prove that, the optimal solutions to Problem ($\mathbb{P}_{10}$) derived in iteration i of Algorithm 3 is a feasible point of the problem in iteration $i + 1$.

Let us assume the optimal solutions to Problem ($\mathbb{P}_{10}$) are $\mathcal{P}^{(i)} = \{p_{k,m}^*, \forall k, m\}, \mathcal{T}^{(i)} = \{t_{k,m}^*, \forall k, m\}, \Theta_n^{(i)} = \{\theta_{k_n,m_n}^*, \forall k_n, m_n\}$ for iteration i. The constraints in (50) obtained at the optimum point of iteration i is written as

$$\frac{1}{4} \left(t_{k_n,m_n}^* + \theta_{k_n,m_n}^* \right)^2 - \theta_{k_n,m_n}^* - p_{k_n,m_n}^* |h_{k_n,m_n}|^2$$

$$-\frac{1}{4} \left[\left(t_{k_n,m_n}^{(i)} - \theta_{k_n,m_n}^{(i)} \right)^2 + 2 \left(t_{k_n,m_n}^{(i)} - \theta_{k_n,m_n}^{(i)} \right) \right.$$

$$\left. \left(t_{k_n,m_n}^* - t_{k_n,m_n}^{(i)} - \theta_{k_n,m_n}^* + \theta_{k_n,m_n}^{(i)} \right) \right] \leq 0,$$

$$\forall k_n \in \mathcal{A}_{m_n}, \forall m_n \in \mathcal{M}_n. \qquad (60)$$

Before carrying out iteration $i + 1$, the parameters are updated as $\tilde{t}_{k_n,m_n}^{(i+1)} = t_{k_n,m_n}^*, \tilde{\theta}_{k_n,m_n}^{(i+1)} = \theta_{k_n,m_n}^*, \forall k_n, m_n$. Let us now proceed to iteration $i+1$. When replacing the variables by the optimal solutions in iteration i, it is straightforward that the constraints of (15), (16), (21), (43), and (46) can be satisfied. Further, for constraint (50) in iteration $i + 1$, upon substituting the updated parameters $\{\tilde{t}_{k_n,m_n}^{(i+1)}, \tilde{\theta}_{k_n,m_n}^{(i+1)}\}$,

it results in

$$\frac{1}{4}\left(t^*_{k_n,m_n} + \theta^*_{k_n,m_n}\right)^2 - \theta^*_{k_n,m_n} - p^*_{k_n,m_n}|h_{k_n,m_n}|^2$$
$$- \frac{1}{4}\Big[\left(t^*_{k_n,m_n} - \theta^*_{k_n,m_n}\right)^2 +$$
$$2\left(t^*_{k_n,m_n} - \theta^*_{k_n,m_n}\right)\left(t^*_{k_n,m_n} - t^*_{k_n,m_n} - \theta^*_{k_n,m_n} + \theta^*_{k_n,m_n}\right)\Big] \tag{61a}$$

$$= \frac{1}{4}\left(t^*_{k_n,m_n} + \theta^*_{k_n,m_n}\right)^2 - \theta^*_{k_n,m_n} - p^*_{k_n,m_n}|h_{k_n,m_n}|^2$$
$$- \frac{1}{4}\left(t^*_{k_n,m_n} - \theta^*_{k_n,m_n}\right)^2 \tag{61b}$$

$$\leq \frac{1}{4}\left(t^*_{k_n,m_n} + \theta^*_{k_n,m_n}\right)^2 - \theta^*_{k_n,m_n} - p^*_{k_n,m_n}|h_{k_n,m_n}|^2$$
$$- \frac{1}{4}\Big[\left(t^{(i)}_{k_n,m_n} - \theta^{(i)}_{k_n,m_n}\right)^2 + 2\left(t^{(i)}_{k_n,m_n} - \theta^{(i)}_{k_n,m_n}\right) \times$$
$$\left(t^*_{k_n,m_n} - t^{(i)}_{k_n,m_n} - \theta^*_{k_n,m_n} + \theta^{(i)}_{k_n,m_n}\right)\Big] \tag{61c}$$

$$\leq 0. \tag{61d}$$

Above, (61b) is derived by replacing the parameters with the solutions in iteration i. Then, in (61c), it employs the first-order Taylor series approximation for $\frac{1}{4}\left(t^*_{k_n,m_n} - \theta^*_{k_n,m_n}\right)^2$ near $(\tilde{t}^{(i)}_{k_n,m_n}, \tilde{\theta}^{(i)}_{k_n,m_n})$ which is its upper bound. Finally, (61d) is obtained by using the result in (60).

Henceforth, we prove that the optimal solutions obtained in the ith iteration is a feasible point of Problem $(\mathbb{P}_{10})$ in the $(i+1)$st iteration. Thus, since Problem $(\mathbb{P}_{10})$ is a concave problem, its objective value obtained in the $(i+1)$st iteration is larger than or equal to the one derived in the ith iteration. As a result, proposition 1 is proved.

APPENDIX D: PROOF OF PROPOSITION 2

Let $\mathcal{V}^{(i)} = \{\mathcal{P}^{(i)}, \mathcal{T}^{(i)}, \Theta^{(i)}\}$ be the accumulation point of the optimal solutions to Problem $(\mathbb{P}_{10})$ derived in iteration i of Algorithm 3. Indicated by Proposition 1, we have that $\mathcal{V}^{(i)} \to \mathcal{V}^*$ as $i \to \infty$. It is noted that the original problem of $(\mathbb{P}_6)$ can be equivalently converted to problem $(\mathbb{P}_9)$. Nevertheless, when converting Problem $(\mathbb{P}_9)$ to Problem $(\mathbb{P}_{10})$, it employs the first-order taylor series to transform the non-convex constraint of (45) to the convex constraint of (50). Note that, in comparison with the non-convex function in (45), the introduced convex upper bound in (48) has the same value and gradient value at the point $\{\mathcal{V}^{(i)}\}$ for any iteration of Algorithm 3. Clearly, this property still holds as $\mathcal{V}^{(i)} \to \mathcal{V}^*$. Henceforth, based on the above property, it can be shown that this accumulation point $\mathcal{V}^*$ also satisfies the KKT conditions for Problem $(\mathbb{P}_6)$.

REFERENCES

[1] Y. Liu, H. Xing, C. Pan, A. Nallanathan, M. Elkashlan and L. Hanzo, "Nonorthogonal Multiple Access for 5G and Beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp 2347-2381, Sept. 2017.

[2] W. Yu, L. Musavian and Q. Ni, "Link-Layer Capacity of NOMA Under Statistical Delay QoS Guarantees," *IEEE Transactions on Communications*, vol. 66, no. 10, pp 4907 ? 4922, Oct. 2018.

[3] J. Zhang, L. Yang, L. Hanzo and H. Gharavi, "Advances in Cooperative Single-Carrier FDMA Communications: Beyond LTE-Advanced," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp 730 ? 756, Second Quarter 2015.

[4] J. Zhang, L. Yang and L. Hanzo, "Energy-Efficient Dynamic Resource Allocation for Opportunistic-Relaying-Assisted SC-FDMA Using Turbo-Equalizer-Aided Soft Decode-and-Forward," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 1, pp. 235-246, Jan. 2013.

[5] L. Li, Y. Xu, Z. Zhang, J. Yin, W. Chen and Z. Han,"A Prediction-Based Charging Policy and Interference Mitigation Approach in the Wireless Powered Internet of Things," *IEEE J. Sel. Areas Commun.*, vol.37, no. 3, pp. 1-13, 2018.

[6] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I and H. V. Poor, "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp 185-191, Feb. 2017.

[7] Y. Liu, M. Elkashlan, Z. Ding and G. K. Karagiannidis, "Fairness of User Clustering in MIMO Non-Orthogonal Multiple Access Systems," *IEEE Communications Letters*, vol. 20, no. 7, pp 1465-1468, July 2016.

[8] S. Timotheou and I. Krikidis, "Fairness for Non-Orthogonal Multiple Access in 5G Systems," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp 1647-1651, Dec. 2015.

[9] Wei Liang, Zhiguo Ding, Yonghui Li and Lingyang Song, "User Pairing for Downlink Non-Orthogonal Multiple Access Networks using Matching Algorithm," *IEEE Transactions on Communications*, vol. 65, no. 12, pp 5319-5332, Dec. 2017.

[10] B. Di, S. Bayat, L. Song and Y. Li, "Radio Resource Allocation for Downlink Non-Orthogonal Multiple Access (NOMA) Networks Using Matching Theory," *in Proc. of IEEE Global Communications Conference (GLOBECOM)*, pp 1-6, Dec. 2015.

[11] L. Lei, D. Yuan, C. K. Ho and S. Sun, "Joint Optimization of Power and Channel Allocation with Non-Orthogonal Multiple Access for 5G Cellular Systems," *in Proc. of IEEE Global Communications Conference (GLOBECOM)*, pp 1-6, Dec. 2015.

[12] M. Al-Imari, M. A. Imran and P. Xiao, "Radio Resource Allocation for Multicarrier Low-Density-Spreading Multiple Access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp 2382-2393, March 2017.

[13] C. Han, J. Zhang, L. Hanzo, et al., "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 46-54, June 2011.

[14] J. Cui, Y. Liu, Z. Ding, P. Fan, A. Nallanathan, "Optimal User Scheduling and Power Allocation for Millimeter Wave NOMA Systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp 1502-1517, Feb. 2018.

[15] F. Fang, H. Zhang, J. Cheng and V. C. M. Leung, "Energy-Efficient Resource Allocation for Downlink Non-Orthogonal Multiple Access Network," *IEEE Transactions on Communications*, vol. 64, no. 9, pp 3722-3732, Sept. 2016.

[16] Y. Zhang, H. M. Wang, T. X. Zheng and Q. Yang, "Energy-Efficient Transmission Design in Non-orthogonal Multiple Access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp 2852-2857, March 2017.

[17] W. Hao, M. Zeng, Z. Chu and S. Yang, "Energy-Efficient Power Allocation in Millimeter Wave Massive MIMO with Non-Orthogonal Multiple Access," *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp 782-785, Aug. 2017.

[18] Z. Yang, W. Xu, H. Xu, J. Shi and M. Chen, "Energy Efficient Non-Orthogonal Multiple Access for Machine-to-Machine Communications," *IEEE Communications Letters*, vol. 21, no. 4, pp 817-820, April 2017.

[19] Z. Yang, Z. Ding, P. Fan and Z. Ma, "Outage Performance for Dynamic Power Allocation in Hybrid Non-Orthogonal Multiple Access Systems," *IEEE Communications Letters*, vol. 20, no. 8, pp 1695-1698, Aug. 2016.

[20] Kaidi Wang, Z. Ding and Wei Liang, "A game theory approach for user grouping in hybrid non-orthogonal multiple access systems," *in Proc. of International Symposium on Wireless Communication Systems (ISWCS)*, pp 643-647, Sept. 2016.

[21] Z. Q. Al-Abbasi and D. K. C. So, "Resource Allocation in Non-Orthogonal and Hybrid Multiple Access System with Proportional Rate Constraint," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp 6309 - 6320, Oct. 2017.

[22] Z. Yang, Z. Ding, P. Fan and N. Al-Dhahir, "A General Power Allocation Scheme to Guarantee Quality of Service in Downlink and Uplink NOMA Systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp 7244-7257, Nov. 2016.

[23] Y. Sun, D. W. K. Ng, Z. Ding and R. Schober, "Optimal Joint Power and Subcarrier Allocation for Full-Duplex Multicarrier Non-Orthogonal Multiple Access Systems," *IEEE Transactions on Communications*, vol. 65, no. 3, pp 1077-1091, March 2017.

[24] J. Papandriopoulos and J. S. Evans, "Low-Complexity Distributed Algorithms for Spectrum Balancing in Multi-User DSL Networks," *in Proc. of*

IEEE International Conference on Communications (ICC), pp 3270-3275, June 2006.

[25] F. Tan, T. Lv and S. Yang, "Power Allocation Optimization for Energy-Efficient Massive MIMO Aided Multi-Pair Decode-and-Forward Relay Systems," *IEEE Transactions on Communications*, vol. 65, no. 6, pp 2368-2381, June 2017.

[26] M. F. Hanif, Z. Ding, T. Ratnarajah and G. K. Karagiannidis, "A Minorization-Maximization Method for Optimizing Sum Rate in the Downlink of NOMA Systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp 76-88, June 2017.

[27] Jia Shi and L. L. Yang, "Novel subcarrier-allocation schemes for downlink MC DS-CDMA systems," *IEEE Transactions on Communications*, vol. 13, no. 10, pp 5716-5728, Oct. 2014.

[28] N. K. Karmarkar, "A New Polynomial-Time Algorithm for Linear Programming," *Combinatorica 4*, pp 373-395, 1984.

[29] I. J. Lustig and R. E. Marsten and D. F. Shanno, "Feature Article-Interior Point Methods for Linear Programming: Computational State of the Art," *ORSA Journal on Computing*, vol. 6, no. 1, pp 1-14, 1994.

[30] T. Liu, C. Yang and L. L. Yang, "A low-complexity subcarrier-power allocation scheme for frequency-division multiple-access systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 5, pp 1564-1570, May 2010.

[31] O. Nwamadi, X. Zhu and A. Nandi, "Dynamic subcarrier allocation for single carrier - FDMA systems," *in Proc. of EUSIPCO*, Aug. 2008.

[32] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp 83-97, 1995.

[33] Stephen boyd, Lieven Vandenberghe, "Convex Optimization," *Cambridge University Press*, 2004.
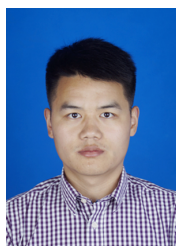
Jia Shi received both his MSc. and Ph.D degrees from University of Southampton, UK, in 2010 and 2015, respectively. He was a research associate with Lancaster University, UK, during 2015-2017. Then, he became a research fellow with 5GIC, University of Surrey, UK, from 2017 to 2018. Since 2018, he joined Xidian University, China, and now is an Associate Professor in the State Key Lab. of Integrated Services Networks (ISN). His current research interests include mmWave communications, artificial intelligence, resource allocation in wireless systems, covert communications, physical layer security, cooperative communication, etc. He is now an associate editor of Electronics Letters, and is also serving as a guest editor of China Communications.

Wenjuan Yu is currently working as a Research Fellow at 5G Innovation Centre, University of Surrey. She received the B.Sc. degree in Electronic and Information Engineering, from Shandong University of Technology, China, in 2010, the M.Sc. degree from School of Telecommunications Engineering, Xidian University, China, in 2013, and the Ph.D. degree from School of Computing and Communications, Lancaster University, UK, in 2018. Her research interests include low-latency communications towards 5G & Beyond, green communications, cross-layer optimization toward delay QoS provisioning.

Qiang Ni (M'04–SM'08) received the B.Sc., M.Sc., and Ph.D. degrees from the Huazhong University of Science and Technology, China, all in engineering. He is currently a Professor and the Head of the Communication Systems Group, School of Computing and Communications, Lancaster University, Lancaster, U.K. His research interests include the area of future generation communications and networking, including green communications and networking, millimeter-wave wireless communications, cognitive radio network systems, non-orthogonal multiple access (NOMA), heterogeneous networks, 5G and 6G, SDN, cloud networks, energy harvesting, wireless information and power transfer, IoTs, cyber physical systems, machine learning, big data analytics, and vehicular networks. He has authored or co-authored over 200 papers in these areas. He was an IEEE 802.11 Wireless Standard Working Group Voting Member and a contributor to the IEEE Wireless Standards.
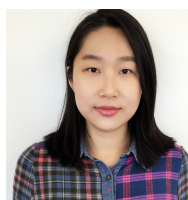
Wei Liang received her M.Sc. and Ph.D. degree in wireless communication at University of Southampton, Southampton, U.K in 2010 and 2015, respectively. She was a Postdoctoral Research Fellow in Lancaster University from 2015 to 2018. She is a associate professor in Northwestern Polytechnical University now. Her research interests include adaptive coded modulation, network coding, matching theory, game theory, cooperative communication, cognitive radio network, Non-orthogonal multiple access and Machine Learning.
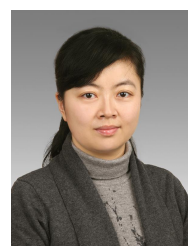
Zan Li received the B.S. degree in Communications engineering from Xidian University, China, in 1998, and the M.S. and Ph.D. degrees in Communication & information systems from Xidian University, China, in 2001 and in 2004, respectively. She is currently a Professor of the State Key Laboratory of Integrated Services Networks (ISN) (Xidian University). Her research interests focus on topics in wireless communications and signal processing, including weak signal detection, spectrum sensing, and cooperative communication.

Pei Xiao (SM'11) is a professor of Wireless Communications at the Institute for Communication Systems, home of 5G Innovation Centre (5GIC) at the University of Surrey. He is the technical manager of 5GIC, leading the research team in the new physical layer work area, and coordinating/supervising research activities across all the work areas within 5GIC (www.surrey.ac.uk/5gic/research). Prior to this, he worked at Newcastle University and Queen?s University Belfast. He also held positions at Nokia Networks in Finland. He has published extensively in the fields of communication theory and signal processing for wireless communications.