QoE-Based Resource Allocation for Multi-Cell NOMA Networks

Jingjing Cui, *Student Member, IEEE*, Yuanwei Liu, *Member, IEEE*, Zhiguo Ding, *Senior Member, IEEE*, Pingzhi Fan, *Fellow, IEEE*, and Arumugam Nallanathan, *Fellow, IEEE*

***Abstract*— Quality of experience (QoE) is an important indicator in the fifth generation (5G) wireless communication systems. For characterizing user-base station (BS) association, subchannel assignment, and power allocation, we investigate the resource allocation problem in multi-cell multicarrier non-orthogonal multiple access (MC-NOMA) networks. An optimization problem is formulated with the objective of maximizing the sum mean opinion scores (MOSs) of users in the networks. To solve the challenging mixed integer programming problem, we first decompose it into two subproblems, which are characterized by combinational variables and continuous variables, respectively. For the combinational subproblem, a 3-D matching problem is proposed for modeling the relation among users, BSs, and subchannels. Then, a two-step approach is proposed to attain a suboptimal solution. For the continuous power allocation subproblem, the branch and bound approach is invoked to obtain the optimal solution. Furthermore, a low complexity suboptimal approach based on successive convex approximation techniques is developed for striking a good computational complexity-optimality tradeoff. Simulation results reveal that: 1) the proposed NOMA networks is capable of outperforming conventional orthogonal multiple access networks in terms of QoE and 2) the proposed algorithms for sum-MOS maximization can achieve significant fairness improvement against the sum-rate maximization scheme.**

Manuscript received July 21, 2017; revised February 12, 2018, April 26, 2018, and June 17, 2018; accepted July 7, 2018. Date of publication July 20, 2018; date of current version September 10, 2018. The work of J. Cui was supported in part by the National Science Foundation of China under Grant 61731017, in part by the 111 Project under Grant 111-2-14, and in part by the U.K. EPSRC under Grant EP/N029720/2. The work of Z. Ding was supported in part by the U.K. EPSRC under Grant EP/N005597/1 and in part by H2020-MSCA-RISE-2015 under Grant 690750. The work of P. Fan was supported in part by the National Science Foundation of China under Grant 61731017 and in part by the 111 Project under Grant 111-2-14. The work of A. Nallanathan was supported by the U.K. EPSRC under Grant EP/N029720/2. This paper was presented in part at the IEEE Global Communication Conference Workshop, Singapore, December, 2017 [1]. The associate editor coordinating the review of this paper and approving it for publication was S. Wang. *(Corresponding author: Yuanwei Liu.)*

J. Cui is with the Institute of Mobile Communications, Southwest Jiaotong University, Chengdu 610031, China, and also with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: cuijingj@foxmail.com).

Y. Liu and A. Nallanathan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: yuanwei.liu@qmul.ac.uk; a.nallanathan@qmul.ac.uk).

Z. Ding is with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: zhiguo.ding@manchester.ac.uk).

P. Fan is with the Institute of Mobile Communications, Southwest Jiaotong University, Chengdu 610031, China. (e-mail: p.fan@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2018.2855130

***Index Terms*— Multi-cell multicarrier non-orthogonal multiple access (MC-NOMA), quality of experience (QoE), resource allocation, three-dimensional (3D) matching, the branch and bound (BB) approach.**

I. Introduction

MULTICARRIER transmission techniques such as orthogonal frequency division multiple access (OFDMA), have been widely adopted in broadband wireless communication systems such as LTE and LTE-Advanced [2]. In conventional multicarrier systems, a given radio frequency band is divided into multiple orthogonal subcarriers and each subcarrier is allocated to at most one user to avoid multiuser interference (MUI). However, the fifth generation (5G) wireless communication system is expected to provide high data rates and massive connectivity to meet the rapid growth of wireless data services and requirements. Non-orthogonal multiple access (NOMA) is recognized as a promising candidate that provides an effective solution to address the challenging requirements of 5G mobile networks, such as massive connectivity, high data speed and low latency [3], [4]. Compared to the conventional orthogonal multiple access (OMA), NOMA allows multiple users to share the same orthogonal resources (e.g., time/frequency) by exploiting superposition coding in the power domain at transmitters and successive interference cancellation (SIC) techniques at receivers. The advantages behind this approach lies in the fact NOMA can opportunistically explore users' channel conditions [5].

Driven by the requirements of high quality video services such as embedded video contents in the webpages, video calls, online TVs, etc, an appropriate level of quality of experience (QoE) in 5G mobile networks is desired. QoE is a subjective assessment of media quality of users and has recently become an essential indicator in 5G wireless communication systems [6], [7]. Due to the various video characteristics, users may experience different QoE even if the data rates are same, which implies that effective QoE-based resource allocation is essential to provide better user satisfaction with limited radio resources. By taking advantage of NOMA features, this paper establishes the potential of QoE-based resource allocation in multi-cell NOMA networks.

A. Related Works

1) Studies on NOMA: Prior research contributions have studied the advantages of NOMA over OMA in different scenarios. In [8], the authors investigated the performance of

NOMA in a cellular downlink cell with randomly deployed users. The impact of user pairing on the sum rate performance was studied in NOMA systems [5]. Sparked by the characteristics of cogitative Radio (CR), the application of NOMA in large-scale CR networks was studied in [9] with carefully considering the channel ordering issue. To address the power allocation problem, a general power allocation scheme was studied in [10], which designed the power allocation coefficients based on the channel state information (CSI). In [11], the cooperative NOMA scheme was investigated by invoking simultaneous wireless information and power transfer (SWIPT) techniques, where a nearby user was regarded as an energy harvesting relay to assist a distant user. Driven by the partial CSI feedback, a power allocation strategy for downlink NOMA systems based on the average CSI was developed in [12] and an optimal decoding order was considered in [13], respectively. Furthermore, in [14], a dynamic user clustering and power allocation for uplink and downlink NOMA systems was investigated.

Regarding the resource allocation works in multicarrier NOMA (MC-NOMA), the authors developed a joint subcarrier and power allocation algorithm in [15], where a near optimal solution was developed based on Lagrangian duality and dynamic programming. In [16], the authors took the user-specific rate characteristics, which are calculated for each subchcannel, as the preference. Then, a many to many matching game with externalities model was proposed to solve the user scheduling and subchannel assignment problem. On the other hand, for full-duplex MC-NOMA systems, the authors in [17] exploited the monotonic optimization theory for the power allocation and user scheduling problem, and an optimal solution was developed to maximize the weighted sum system throughput. Furthermore, in [18], the energy efficiency of MC-NOMA was considered, where a low-complexity suboptimal algorithm based on matching theory was developed.

2) Studies on QoE-Based Resource Allocation: In [19], a QoE-based evaluation methodology was proposed to assess the LTE systems video capacity, where the proposed QoE-based radio resource allocation (RRA) scheme could enhance the video capacity. The authors in [20], proposed a user-oriented joint subcarrier and power allocation algorithms for the downlink of a heterogeneous OFDMA system, where the best possible QoE for each user is optimized. Sparked by the game theory, a QoE-oriented strategy for OFDMA RRA was studied in [21], where the goal was to achieve the best possible QoE by search a satisfactory equilibria through market-like resource exchanges. To satisfy the heterogeneous service requirements in multi-cell OFDMA networks, a QoE-based proportional fair (PF) scheduling was investigated in [22], which considered the network-wide users' QoE maximization as well as fairness among users. To mitigate the co-tiered interference, a joint matching-coalition game theoretical scheme was proposed to solve a QoE-based multichannel allocation problem in heterogeneous cellular networks in [23]. On the other hand, in [24], the QoE oriented resource allocation problem in OFDMA based multi-cell networks was investigated, where the multiple base stations (BSs) cooperated for interference mitigation. In addition, a game based joint spectrum sharing, power allocation and user scheduling approach was developed in [25], where the objective was to maximize the users' satisfaction across the network for providing better QoE.

B. Motivations and Contributions

As mentioned above, NOMA has received remarkable attention both in the world of academia and industry. However, so far few works consider the resource allocation for MC-NOMA in multi-cell networks. Moreover, there is still a paucity of research contributions on investigating the QoE issues of NOMA, which motivates this treatise. Note that the employment of NOMA on the BS, multiple users can be multiplexed on a specific subchannel, which makes the resource allocation problem of MC-NOMA different from that of OMA. The motivation and the challenges of this work is concluded as follows:

- Multi-cell MC-NOMA is not well investigated, especially for the problem both considering user association and resource allocation. In this treatise, we specifically consider the multi-cell networks, where the BS cooperation is performed to reduce the inter-cell interference.
- QoE is not considered for NOMA, especially for user-BS cooperation. Most existing work addresses fairness issue only by using network-level criteria like max-min but neglects the specific requirements of individual users. Note that the QoE is a user-centric measure demonstrating the user satisfaction, which has received many attentions from many enterprises and researchers. QoE-driven techniques will bring about the improvement of fairness and efficiency, but it does not add any cost of additional resource investment [24].

Therefore, we model the problem of resource allocation for MC-NOMA in multi-cell networks to improve the user QoE instead of throughput, which can provide potential performance gains satisfying the user demands [26]. Specifically, in this paper, we formulate the QoE-based resource allocation problem in the multi-cell MC-NOMA networks with allowing BS cooperations, which consists of user-BS association, subchannel assignment and power allocation. It involves a joint optimization decision by BSs. Furthermore, in the aggressive frequency reuse deployment, the co-channel interference makes the resource allocation problem among BSs coupled and correlated. In addition, the non-convexity of QoE makes the problem more complicated. The primary contributions of this paper are concluded as follows:

1) We investigate the application-oriented QoE in multi-cell MC-NOMA networks. We use the mean opinion score (MOS) to evaluate the QoE of users. With this aim, we formulate the sum MOS maximization problem by jointly designing user-BS association, subchannel assignment and power allocation, which is a combinatorial optimization problem.
2) To solve the challenging optimization problem, we decompose the joint resource allocation problem into two subproblems as: *i) the problem of user-BS association and subchannel assignment*; and *ii) power allocation optimization*. We construct a three-dimensional

(3D) matching problem to model the allocation among users, BSs and subchannels. Then, we also propose a two-step approach by solving two two-dimensional (2D) matching subproblems: UE-BS matching problems and (UE,BS)-SC matching problems, which provides a low-complexity solution of user-BS association and subchannel assignment.

3) For the non-convex power allocation problem, we propose a global optimal power allocation strategy based on the branch and bound (BB) approach, which provides an upper bound for power allocation. Moreover, we also propose a low-complexity suboptimal solution based on successive convex approximation (SCA) techniques.
4) We demonstrate that the proposed low-complexity solution by leveraging the matching theory based two-step approach and the SCA algorithm is capable of achieving a good performance comparing with the global optimal solution with exhaustive search and BB algorithms. Moreover, we demonstrate that the proposed multi-cell MC-NOMA framework outperforms the conventional multi-cell MC-OMA framework with the aid of both of the proposed algorithms.

C. Organization

The rest of the paper is organized as follows. In Section II, we present network model consists of the problem formulation for the QoE-based resource allocation. In Section III, we propose a low complexity algorithm for user-BS association and subchannel assignment using matching theory. Solutions to power allocation optimization problem are presented in Section IV, where a global optimal solution based on BB is provided and a low complexity power allocation based on SCA are proposed. Simulation results are presented in Section V, which is followed by conclusions in Section VI.

II. Network Model

A. System Description

Consider a multi-cell downlink NOMA transmission scenario as shown in Fig. 1, where multiple T base stations (BSs) communicate with K cellular users, denoted by $\mathcal{T} = \{1, \cdots, T\}$ and $\mathcal{K} = \{1, \cdots, K\}$, respectively. We assume that each cell is served by a BS and BSs are temporally synchronized.[1] Both BSs and users equip with one transmit and one receive antenna. The entire bandwidth W is partitioned into N subchannels, each with $\frac{W}{N}$. The index set of all subchannels is denoted by $\mathcal{N} = \{1, \cdots, N\}$. We consider the universal frequency reuse deployment in which every cell is available to the whole bandwidth. Invoked by the NOMA protocol, each subchannel can be shared by multiple users associated to the same BS. Considering the detection complexity of SIC receiver, we assume that the maximum number of users allocated in the n-th subchannel, denoted by SC_n, of BS t, denoted by BS_t, is L_t. Particularly, inspired by spectral aggregation, we consider that each cellular user $k \in \mathcal{K}$, denoted by UE_k, can potentially aggregate data from all available subchannels of the connected BS. Moreover, we assume that the BSs cooperate to jointly serve the users where the CSI of the direct link and the cross link channels are available at the BSs. This is employed to design the user scheduling and power allocation strategies, which enhances the reliability of data reception at each user by exploiting the multiple-BS diversity. In this paper, we consider a quasi-static channel, that is the channel condition remains constant within a time slot and varies independently from one to another. In the following, we introduce the following sets: $\mathcal{K}_n^t$, $\mathcal{N}_k^t$, and $\mathcal{T}_k^n$ denote the sets of users associated to BS_t on SC_n, the set of subchannels occupied by UE_k associated to BS_t and the set of BSs associated to UE_k on SC_n, respectively.

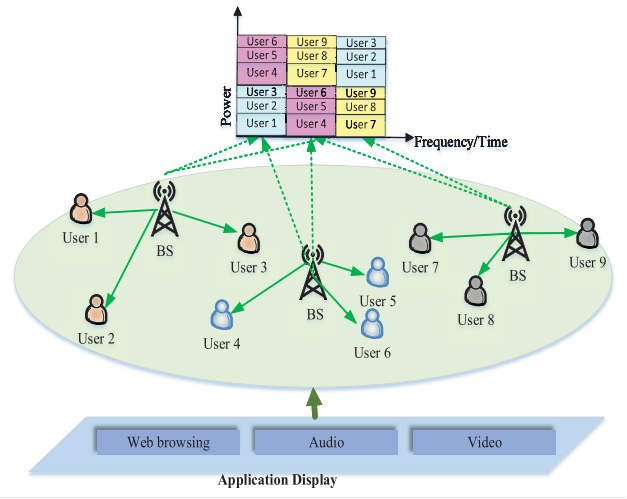


Fig. 1. An exemplary user-BS association and subchannel assignment in downlink multi-cell NOMA scenarios.

B. Signal Model

Denote $\nu_{t,k}$ and $\xi_{n,t}$ as the user-BS indicator and the subchannel-BS indicator, respectively. $\nu_{t,k} = 1$ indicates the k-user is served by the t-th BS, $\nu_{t,k} = 0$ if otherwise; $\xi_{n,t} = 1$ indicates that the n-th subchannel is allocated to the t-th BS; $\xi_{n,t} = 0$ if otherwise. Note that $\nu_{t,k}\xi_{n,t} = 1$ indicates UE_k is connected to BS_t and allocated with SC_n, and $\nu_{t,k}\xi_{n,t} = 0$ if otherwise. Thus, the superposition coded symbol x_n^t to be transmitted at BS t on channel n is given by

$$x_n^t = \sum_{k=1}^{K} \nu_{t,k}\xi_{n,t}\sqrt{P_{n,k}^t}\, x_{k,n}^t, \tag{1}$$

where $x_{n,k}^t$ is the transmit signal of BS_t in SC_n, $P_{n,k}^t$ is the allocated power of UE_k associated with BS_t on SC_n. In SC_n, UE_k, $k \in \mathcal{K}$ receives interference from other users in the same subchannel. As a consequence, the received signals of UE_k associated with BS_t on SC_n is given by

$$y_{n,k}^t = f_{n,k}^t x_n^t + I_{n,k}^t + \eta_{n,k}^t, \tag{2}$$

[1] It is assumed that a user can be associated to one BS. If a user can connect to multiple BS simultaneously, then the BSs can serve the user cooperatively. The cooperation between the BSs may further enhance the sum MOS of the system considered, hence our future research would consider investigating cooperative multi-cell NOMA systems.

where $\eta_{n,k}$ is the additive white Gaussian noise (AWGN) at UE_k on SC_n with variance σ^2, $f^t_{n,k}$ is the channel coefficients between BS_t and UE_k on SC_n. And $I^t_{n,k}$ is the accumulative interference to UE_k from other BSs on SC_n except BS_t, which is given by

$$I^t_{n,k} = \sum_{s=1, s\neq t}^{T} f^s_{n,k}\sqrt{P^s_n}x^s_n, \tag{3}$$

and P^s_n is the total power consumption of BS_s on SC_n,

$$P^s_n = \sum_{k=1}^{K} \nu_{s,k}\xi_{n,s}P^s_{n,k}, \tag{4}$$

To proceed further, we introduce an auxiliary term $g^t_{n,k}$ as

$$g^t_{n,k} = \frac{\nu_{t,k}\xi_{n,t}h^t_{n,k}}{\sum_{s=1,s\neq t}^{T} h^s_{n,k}P^s_n + \sigma^2}, \quad k, j \in \mathcal{K}^t_n, \tag{5}$$

where $h^t_{n,k} = |f^t_{n,k}|^2$ is the channel gain coefficient and $g^t_{n,k}$ can be viewed as an equivalent channel gain between BS_t and UE_k on SC_n.

In each subchannel, NOMA protocol is invoked. Specifically, consider a pair of two users (k, j) served by BS_t, in which UE_k wants to decode and remove UE_j's signal by SIC on SC_n, then the inequality holds: $g^t_{n,k} \geq g^t_{n,j}$.

In fact, in NOMA, SIC can be carried out at the users with stronger equivalent channel gains. Without loss of generality, it is assumed that all the channels on SC_n of BS_t follows the order as $g^t_{n,\pi(1)} \leq g^t_{n,\pi(2)} \leq \cdots \leq g^t_{\pi(|\mathcal{K}^t_n|)}$, where $\pi(k)$ denotes the k-th decoded user's index and $|\mathcal{K}^t_n|$ denotes the cardinality of $\mathcal{K}^t_n$. Therefore, $\mathrm{UE}_{\pi(k)}$ first decodes the messages of all the $(k-1)$ users, and then successively subtracts these messages to decode its own information. Following the principle above, the received signal-to-interference-plus-noise-ratio (SINR) for the k-th decoded user on SC_n is given by

$$\gamma^t_{n,\pi(k)} = \frac{\nu_{t,\pi(k)}\xi_{n,t}h^t_{n,\pi(k)}P^t_{n,\pi(k)}}{\sum\limits_{i=k+1}^{|\mathcal{K}^t_n|} \nu_{t,\pi(k)}\xi_{n,t}h^t_{n,\pi(k)}P^t_{n,\pi(i)} + \sum\limits_{s\neq t} h^s_{n,\pi(k)}P^s_n + \sigma^2}. \tag{6}$$

Then we focus on the data rate of $\mathrm{UE}_{\pi(k)}$ on SC_n at BS_t, which is given by $R^t_{n,\pi(k)} = \frac{W}{N}\log\left(1+\gamma^t_{n,\pi(k)}\right)$. Hence, the overall data rate of user k can be computed as

$$R_k = \sum_{t=1}^{T}\sum_{n=1}^{N} R^t_{n,\pi(k)}. \tag{7}$$

C. MOS Model for Web Browsing

Inspired by the widely used QoE metric, MOS model is used as a measure of the user's QoE for the services like video streaming, file download, or web browsing. As one of the most popular application in wireless networks, we focus on web browsing applications in this paper. It maps the subjective human perception of quality for web browsing applications to the objective metrics[2] In [27], the MOS model for web browsing applications is defined as follows:

$$\mathrm{MOS}_{web} = -C_1\ln(d(R_{web})) + C_2, \tag{8}$$

where R_{web} is the data rate. MOS_{web} represents the real score ranging from 1 to 5, which reflects the user perceived quality. The higher score means that the human perception quality is the better. C_1 and C_2 are constants determined by analyzing the experimental results of the web browsing applications, which are set to be 1.120 and 4.6746, respectively. $d(R_{web})$ is the delay time between a user sent a request for a web page and the entire web contents displayed. The delay time depends on multiple factors such as the web page size (e.g. the round trip time (RTT)) and the effects of the protocols (e.g., TCP and HTTP). In this paper, we adopt TCP and HTTP protocols for the multi-cell NOMA systems, where the function $d(R_{web})$ in [28] is modelled as

$$d(R_{web}) = 3\mathrm{RTT} + \frac{\mathrm{FS}}{R_{web}} + L\left(\frac{\mathrm{MSS}}{R_{web}} + \mathrm{RTT}\right) - \frac{2\mathrm{MSS}(2^L-1)}{R_{web}}, \tag{9}$$

where RTT $[s]$ is the round trip time, FS $[bit]$ is the web page size and MSS $[bit]$ is the maximum segment size. The parameter $L = \min\{L_1, L_2\}$ represents the number of slow start cycles with idle periods. L_1 denotes the number of cycles that the congestion window takes to reach the bandwidth-delay product and L_2 is the number of slow start cycles before the web page size is completely transferred, which are defined as follows [28].

$$L_1 = \log_2\left(\frac{R_{web}\mathrm{RTT}}{\mathrm{MSS}} + 1\right) - 1, \quad \text{and}$$
$$L_2 = \log_2\left(\frac{\mathrm{FS}}{2\mathrm{MSS}} + 1\right) - 1. \tag{10}$$

As discussed in [27], the impact of the RTT on the MOS function is minor compared to the data rate and the file size of web pages, especially for short ranges of RTT. In addition, as the 3GPP technical specification of the LTE release 8 proposed, it is expected that the future advanced LTE systems achieve even lower RTT [29] than the currently supported 10 ms. Thus, it is reasonable to assume $\mathrm{RTT} = 0$ ms.[3] Based on this assumption (9) is simplified as $d(R_{web}) = \frac{\mathrm{FS}}{R_{web}}$. Then the mapping for user $\pi(k)$ from the user data rate to the MOS function can be simplified as

$$\mathrm{MOS}^{\pi(k)}_{web} = C_1\ln\left(\sum_{t=1}^{T}\sum_{n=1}^{N} R^t_{n,\pi(k)}\right) + C^{\pi(k)}_3, \tag{11}$$

where $C^{\pi(k)}_3 = C_2 + C_1\ln\left(\frac{W}{N\times\mathrm{FS}_{\pi(k)}}\right)$ is a constant.

[2]Note that the relationship between the data rate and the QoE for different applications were modelled by different MOS models [20]–[22]. The proposed algorithm is capable of being extended to other applications with necessary modifications, which we may include in our future work.

[3]In this paper, we assume that the nodes are static, or slow moving in the considered networks. Therefore, the channels might stay the same for a quite long time period, and hence we can use the assumption that the channels are quasi-static such as in [24]–[27], or sometimes termed block-fading.

D. Problem Formulation

In this section, we formulate the problem to optimize the cross-layer QoE aware resource allocation based on designing user-BS association, subchannel assignment and power allocation. The optimization problem can be expressed as follows:

$$\max_{\substack{\{\{\nu_{t,\pi(k)}\},\\ \{\xi_{n,t}\},\{P^t_{n,\pi(k)}\}\}}} \quad U = \sum_{\pi(k)\in\mathcal{K}} \mathrm{MOS}^{\pi(k)}_{web} \tag{12a}$$

$$\text{s.t. } g^t_{n,\pi(k)} \geq g^t_{n,\pi(j)},\ k > j, \forall(k,j), \forall t, \forall n, \tag{12b}$$

$$\sum_{\pi(k)\in\mathcal{K}} \sum_{n\in\mathcal{N}} \nu_{t,\pi(k)} \xi_{n,t} P^t_{n,\pi(k)} \leq P^t, \forall t, \tag{12c}$$

$$2 \leq \sum_{\pi(k)\in\mathcal{K}} \nu_{t,\pi(k)} \leq L_t, \sum_{t\in\mathcal{T}} \nu_{t,\pi(k)} \leq 1,\ \forall t,\ \forall k, \tag{12d}$$

$$\sum_{n\in\mathcal{N}} \xi_{n,t} \leq S_t, \quad \sum_{t\in\mathcal{T}} \xi_{n,t} \leq T_n,\ \forall t,\ \forall n, \tag{12e}$$

$$P^t_{n,\pi(k)} \geq 0, \quad \pi \in \Pi,\ \nu_{t,\pi(k)},\ \xi_{n,t} \in \{0,1\}, \quad \forall n, \quad \forall k,\ \forall t, \tag{12f}$$

where Π represents the set of total possible decoding orders. Constraints (12b) are used to guarantee that SIC can be performed successfully for a specific order. Constraints (12c) denote the transmit power constraint for BS_t, $\forall t$, with the maximum power allowance P^t. Constraints (12d) are NOMA multiplexing constraints where L_t indicates the maximum number of multiplexed users connected to BS_t. Moreover, we consider each each user is capable of connecting one BS at each subchannel. Constraints (12e) represent each BS can occupy S_t subchannels at maximum and each subchannel can be shared by T_n BSs at maximum In this paper, we assume that the NOMA scheme is applied among the users in the same frequency band and time slot, which has been studied in [24] and [30]. The use of more sophisticated reuse schemes may further enhance the attainable performance of the systems considered, but this is beyond the scope of this treatise. In addition, regarding the case where we allow a BS sometimes serves only one user, the multi-cell network will work in a hybrid multiple access method. In this case, the performance of the hybrid network may be improved by optimizing resource allocation. However, the performance optimization of the hybrid network becomes more challenging especially in selecting the multiple access method. Our future work will investigate the resource allocation of the hybrid network, perhaps with the aid of the results derived in this work.

The sum MOS optimization problem (12) by jointly designing user-BS association, subchannel assignment and power allocation for a multi-cell NOMA network is a combinatorial optimization task, which generally yields unacceptable computation burden with brute-force search. Note that the optimization problem (12) includes the binary optimization variables for user-BS association and subchannel assignment and the continuous variables for the power allocation coefficients. To solve problem (12) effectively, we propose to decompose it into two subproblems: 1) the problem of user-BS association and subchannel assignment and; 2) the problem of power allocation among users.

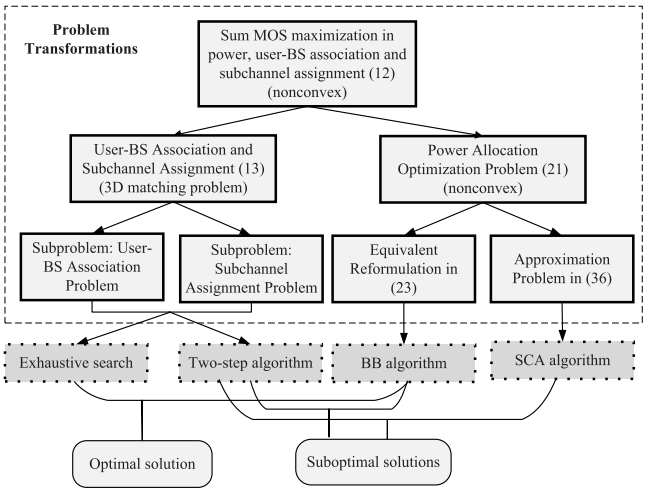


Fig. 2. Overview of the proposed approach to the sum MOS maximization problem. We obtain both the global optimal and suboptimal algorithms.

Fig. 2 gives an overview of the development in the paper, particularly the connections between key optimization problems and the algorithms. In Fig. 2, the key reformulated problems and the algorithm studied in this paper are illustrated in different boxes: The ones with solid boundaries are the reformulated problems, the ones with dotted boundaries are the designed algorithms, and the ones with rounded rectangle are the generated solutions. Due to the combinatorial features of user-BS association, subchannel assignment, exhaustive search provides a straightforward method to find the globally optimal combination for a small-scale network when the power allocation coefficients are fixed. In addition, we propose a low-complexity matching theory based algorithm which will be discussed in Section III. Furthermore, when the user-BS association and subchannel assignment scheme are fixed, finding the optimal solution is still nontrivial due to the non-convex property of problem (12) in terms of the power allocation coefficients. BB techniques provide an efficient approach to solve the non-convex optimization problem [31]–[33], which motivates us the application of the BB algorithm to obtain the optimal power allocation coefficients. Moreover, an low-complexity power allocation algorithm is also developed to avoid the huge complexity of the BB algorithm. The proposed optimal and suboptimal power allocation algorithms will be discussed in Section IV.

III. User-BS Association and Subchannel Assignment Using 3D Matching

In this section, we focus on solving the user-BS association and subchannel assignment problem in (12), which can be expressed as

$$\max_{\{\nu_{t,k}\},\{\xi_{n,t}\}} \sum_{k=1}^{K} \mathrm{MOS}^{k}_{web} \quad \text{s.t. (12d)} - \text{(12f)}. \tag{13}$$

Problem (13) is a combinational optimization problem among users, BSs and subchannels. From the point of the

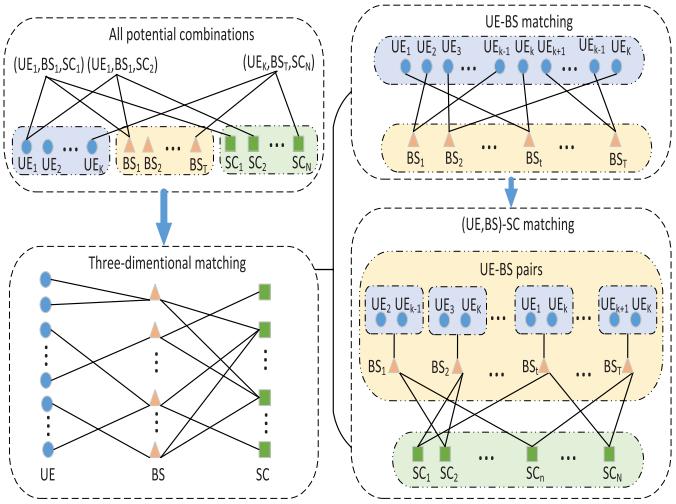


Fig. 3. Graphical expressions of 3D matching among users, BSs and subchannels.

graphical, the mutual relationship among users, BSs and subchannels can be represented in the left top part of Fig. 3. To further present the relationship among users, BSs and subchannels, a bi-partite graph based representation is shown in the left bottom part of Fig. 3. As illustrated in Fig. 3, UE_k is associated to BS_t, they compose an association unit $(\mathrm{UE}_k, \mathrm{BS}_t)$. When SC_n is allocated to the association unit $(\mathrm{UE}_k, \mathrm{BS}_t)$, we say UE_k, BS_t and SC_n are matched with each other, denoted by a matching triple $(\mathrm{UE}_k, \mathrm{BS}_t, \mathrm{SC}_n)$. Next, we first introduce the definition of 3D matching.

Definition 1: An instance of 3D matching involves three disjoint finite sets $\mathcal{K}$, $\mathcal{T}$ and $\mathcal{N}$, where the cardinalities are K, T and N, correspondingly, which are the size of the problem instance. A matching triple is denoted by $(\mathrm{UE}_k, \mathrm{BS}_t, \mathrm{SC}_n) \in \mathcal{K} \bigcup \mathcal{T} \bigcup \mathcal{N}$. A matching is a set of user-BS-subchannel assignment.

It is proved that 3D matching is NP-hard and there is no polynomial-complexity algorithm to find the optimal solution [34]. To solve the challenging problem, we propose a low-complexity suboptimal algorithm by decomposing the 3D matching problem into two 2D matching subproblems-UE-BS matching problems and (UE,BS)-SC matching problems. Then, we solve the two subproblems individually as shown in the right part of Fig. 3. Specifically, the first subproblem is to select a served BS for each user to transmit desired signals, which is a many-to-one matching problem between users and BSs, i.e, multiple users can be served by one BS using the NOMA protocol. Then, subchannel sharing for each BS is considered in the second subproblem, which is a many-to-many matching problem between BSs and subchannels, i.e., one BS to S_t subchannels and one subchannel to T_n BSs.

A. Preliminaries for Matching Theory

In a 2D matching, there are two finite and disjoint sets denoted by $\mathcal{M} = \{m_1, m_2, \cdots, m_n\}$ and $\mathcal{W} = \{w_1, w_2, \cdots, w_p\}$, respectively. Each $m_i \in \mathcal{M}$ has a preference list over the set of $\mathcal{W}$. In this paper, we build the preferences by the rate rather than the MOS value. Since the MOS value is a user-centric measure, it cannot be calculated in the formulated (UE,BS)-SC matching problem for subchannel assignment. In addition, in the formulated user-BS association matching problem, the preference built by the rate value is equivalent to that based on the MOS value, since a user's MOS value is the logarithm of the user's sum rate over all subchannels. Analogously, each $w_j \in \mathcal{W}$ has preferences over $\mathcal{M}$. The individual preferences represent the priorities of its selection among different alternatives. If m_i prefers w_1 to w_2, we express it as $w_1 \succ_{m_i} w_2$. In this paper, we assume that the preference list of each player has the following properties: 1) *complete ordering*: each player will never confront with an indeterminable choice, i.e., any two alternatives can be compared for an player to get a preferred one. 2) *transitive*: it can be express as if $w_1 \succ_{m_i} w_2$ and $w_2 \succ_{m_i} w_3$ then $w_1 \succ_{m_i} w_3$. Based on the above descriptions, we give the following definitions:

Definition 2: A many-to-many (one) matching φ is a function from the set $\mathcal{M} \bigcup \mathcal{W}$ into the set of unordered families of elements of $\mathcal{M} \bigcup \mathcal{W} \bigcup \{0\}$ such that

1) $|\varphi(m)| \leq q_w$ *for every* $m \in \mathcal{M}$;
2) $|\varphi(w)| = q_m$ *for every* $w \in \mathcal{W}$;
3) $\varphi(m) \in \mathcal{W}$ *if and only if* $\varphi(w) \in \mathcal{M}$;
4) $m = \varphi(w) \Leftrightarrow w = \varphi(m)$;

where q_w and q_m are positive integer quotas.

The notation φ has different meanings depending on the parameter. If the parameter is m, then $\varphi(m)$ maps to the matched $\mathcal{W}$ set. If the parameter is w, then $\varphi(w)$ gives the set of matched player of $\mathcal{M}$. Note that is $q_w = 1$, one can obtain the definition of many-to-one matching.

In a many-to-many (one) matching with externalities, it is not straightforward to define a stability concept because the gains from a matching pair depends on which players the other players have. Sparked by the definition of exchange stable stability, it is convenient to define a swap matching [35]. Specifically, a swap matching is defined as $\varphi_i^j = \{\varphi \setminus \{(i,m),(j,n)\} \bigcup \{(j,m),(i,n)\}\}$, where $\varphi(i) = m$ and $\varphi(j) = n$. Based on the swap operation, we introduce the two-sided exchange stability [35] as follows.

Definition 3: A matching φ is two-sided exchange-stable if and only if there does not exist a pair of players (i, j) with $m = \varphi(i)$ and $n = \varphi(j)$, such that

1) $\forall x \in \{i, j, m, n\}$, $U_m(\varphi_i^j) \geq U_m(\varphi)$;
2) $\exists x \in \{i, j, m, n\}$, *such that* $U_m(\varphi_i^j) > U_m(\varphi)$, *then the swap matching φ_i^j is approved, and (k, j) is called a swap-blocking pair in φ.*

where $U_m(\varphi)$ denotes the utility for player m under matching φ. In general, the pair of players satisfying condition 1) and condition 2) is called a swap-blocking pair.

The features of the swap-blocking pair ensure that if a swap matching is approved, the achievable rates of any user involved will not decrease and at least one of the user's rate will increase. Furthermore, the definition indicates that a swap matching is two-sided exchange-stable when all players are indifferent.

B. User-BS Association Problem

As discussed above, the user-BS association problem is a many-to-one matching problem. Due to the interference in (6), the SINR of user $\mathrm{UE}_{n,k}^{t}$ over each subchannel is related to the set of users sharing with the same subchannels. Furthermore, each BS not only considers which users to match with, but also that the inner-relationship among the subset of users due to the power domain multiplexing. Thus, more specifically, the formulated user-BS association problem is a many-to-one matching problem with externalities.

To model the externalities, the preference can be formulated as the rate of each user over all subchannels, where the rate of UE_k associated to BS_t can be expressed as

$$\mathcal{P}_k^t = \sum_{n\in\mathcal{N}} \log_2(1+\gamma_{n,\pi(k)}^t). \tag{14}$$

Then the preference of BS_t on a set of users $\varphi(t)$ can be defined as

$$\mathcal{P}^t = \textstyle\sum_{k\in\varphi(t)} \mathcal{P}_k^t. \tag{15}$$

Therefore, for a given UE_k, any two BS_t and $\mathrm{BS}_{t'}$, any two matchings φ and φ', we have the following relations,

$$(t,\varphi) \succ_{\mathrm{UE}_k} (t',\varphi') \Leftrightarrow \mathcal{P}_k^t(\varphi) > \mathcal{P}_k^t(\varphi'), \tag{16}$$

which indicates that UE_k prefers BS_t in φ to $\mathrm{BS}_{t'}$ in φ' only if UE_k can achieve a higher rate on BS_t than $\mathrm{BS}_{t'}$. Analogously, for any BS_t, $t \in \mathcal{T}$, its preference over the user set can be described as follows. For any two subsets of users $\mathcal{K}_1$ and $\mathcal{K}_2$ with $\mathcal{K}_1 \neq \mathcal{K}_2$, any two matchings φ and φ' with $\mathcal{K}_1 = \varphi(t)$ and $\mathcal{K}_2 = \varphi'(t)$ are defined as

$$(\mathcal{K}_1,\varphi) \succ_t (\mathcal{K}_2,\varphi') \Leftrightarrow \mathcal{P}^t(\varphi) > \mathcal{P}^{t'}(\varphi'). \tag{17}$$

It implies that BS_t prefers the set of users $\mathcal{K}_1$ to $\mathcal{K}_2$ only when BS_t can get a higher rate from $\mathcal{K}_1$.

Based on the established preference lists, we utilize a extend deferred acceptance (EDA) based algorithm proposed in [36] to construct a initial matching state between users and BSs. Then, the swap operation procedure is employed to further enhance the utility. In the EDA based initialization procedure, the BS first allocates the transmit power equally to the users. Hence, the users and the BSs can construct their own preference lists based on (16) and (17), respectively. Then each user proposes to the most preferred BS based on its preference list. At the BS acceptance phase, each BS accepts the users with prior preferences and rejects the others. The algorithm terminates when all users are matched to the BSs or every unmatched users has been rejected by every BS.

C. Subchannel Assignment

As discussed above, the problem to assign subchannels to (UE,BS) units is a many-to-many matching problem. Due to one subchannel can assign multiple BSs, the rate of each (UE,BS) unit over each subchannel is related to the other BSs sharing with the same subchannels. Thus, the formulated subchannel assignment problem is a many-to-many matching problem with externalities.

Similar to Subsection III-B, we formulate the preference as the sum rate of the users associated to the BSs on each subchannels. Specifically, the sum rate of users associated to BS_t on SC_n can be expressed as

$$\mathcal{P}_t^n = \sum_{k\in\mathcal{K}_t} \log_2(1+\gamma_{n,\pi(k)}^t). \tag{18}$$

Analogously, suppose ϕ and ϕ' are two different matchings, for a given BS_t, any two subchannels SC_n and $\mathrm{SC}_{n'}$, we have the following relations,

$$(n,\phi) \succ_{\mathrm{BS}_t} (n',\phi') \Leftrightarrow \mathcal{P}_t^n(\phi) > \mathcal{P}_t^n(\phi'), \tag{19}$$

which indicates that BS_t prefers SC_n in ϕ to $\mathrm{SC}_{n'}$ in ϕ' only if BS_k can achieve a higher rate on SC_n than $\mathrm{SC}_{n'}$. For any SC_n, $n \in \mathcal{N}$, its preference over the user set can be described as follows. For any two subsets of BSs $\mathcal{T}_1$ and $\mathcal{T}_2$ with $\mathcal{T}_1 \neq \mathcal{T}_2$, any two matchings ϕ and ϕ' with $\mathcal{T}_1 = \phi(n)$ and $\mathcal{T}_2 = \phi'(n)$ are defined as

$$(\mathcal{T}_1,\phi) \succ_t (\mathcal{T}_2,\phi') \Leftrightarrow \mathcal{P}^n(\phi) > \mathcal{P}^{n'}(\phi'). \tag{20}$$

It implies that SC_n prefers the set of BSs $\mathcal{T}_1$ to $\mathcal{T}_2$ only when SC_n can get a higher rate from $\mathcal{T}_1$.

Now based on the established preference lists, an initial matching state can be obtained by utilizing EDA based algorithm between (UE,BS) units and subchannels, where we assume that the (UE,BS) unit propose to subchannels. Specifically, similar to the process of EDA based user-BA association, the subchannels and the (UE,BS) units can construct their own preference lists based on (19) and (20), respectively. Then each (UE,BS) unit proposes to the most preferred subchannel based on its preference list. At the subchannel acceptance phase, each subchannel accepts the (UE,BS) unit with prior preferences and rejects the others. The algorithm terminates when all (UE,BS) units are matched to the subchannels or every unmatched users has been rejected by every subchannel.

Furthermore, we can conclude the complete procedure for solving user-BS association and subchannel assignment problem in **Algorithm 1**. In **Step-I**, user-BS association is performed, which consists of an initialization procedure in line 1 and a swap procedure in line 2 to line 9. EDA based algorithm is adopted for the initialization procedure. Then, all possible swap operations between users and BSs are checked to further enhance the system utility. A two-sided stable matching will be reached between users and BSs. In **Step-II**, the matching between (UE,BS) units and subchannels are performed. Similar to **Step-I**, the initialization algorithm can be realized by EDA based algorithm, where the preference lists for (UE,BS) units and subchannels are constructed from (18). We assume that (UE,BS) units propose to subchannels in the initialization algorithm. Then the swap procedure is conducted in line 12 to line 19 to further improve the utilities.

D. Analysis of the Proposed Two-Step Algorithm

1) Complexity: The computational complexity of the proposed two-step algorithm based user-BS association and subchannel assignment in **Algorithm 1** is relied on the

Algorithm 1 Two-Step Algorithm Based User-BS Association and Subchannel Assignment

Step-I: Many-to-one matching based UE-BS association

Construct the initial UE-BS matching set $\mathcal{A}$ by EDA based algorithm. Let $\mathcal{A}_I = \mathcal{A}$.
repeat
  For any user $k \in \mathcal{A}_I$, it searches for another user $j \in \mathcal{A}_I \setminus \mathcal{A}_I(\varphi(k))$.
  if k, j is a swap-blocking pair **then**
    $\varphi = \varphi_k^j$
  else
    Keep the current matching state
  end if
until No swap-blocking pair is found
Output the stable matching denoted as φ_I and the corresponding objective value $U_0 = U(\varphi_I)$.

Step-II: Many-to-many matching based SC assignment

Construct the initial (UE,BS)-SC matching set $\mathcal{A}_{II} = \mathcal{A}$ by EDA based algorithm.
repeat
  For any (UE,BS) unit $t \in \mathcal{A}_{II}$, it searches for another (UE,BS) unit s with $s \in \mathcal{A}_{II} \setminus \mathcal{A}_{II}(\varphi(t))$. Let $\mathcal{U} = \{U_0\}$.
  For a given t, calculate the candidate U_t^s for the swapping pair (t, s).
  if t, s is a swap-blocking pair **then**
    $\mathcal{U} = \mathcal{U} \cup \{U_t^s\}$.
  end if
  Keep the swapping-blocking pair with $t, s^* = \arg\max_{U_t^s} \mathcal{U}$, then $\varphi = \varphi_t^{s^*}$. Set $U_0 = U_t^{s^*}$.
until No swap-blocking pair is found.
Output the stable matching φ_{II}.

two 2D matching procedures. In the following, we will analyze the computational complexity of each 2D matching procedure.

- The initialization algorithm in **Step-I** requires each user to propose to one BSs and each BS can accept multiple users based on its preference list. Assume that the worst case that the proposing number of each user is T. The complexity is $\mathcal{O}(KT^2)$.
- For the swap procedure in **Step-I**, there are at most TL_t users can perform swap operation. In each iteration, for UE_k^t, the maximum swap operation number is $L_t(T-1)$, since each user associates with one BS. Therefore, a swap operation for K users in each iteration is $\frac{1}{2}KL_t(T-1)$. For a given number of total iteration V, the complexity can be presented as $\mathcal{O}(VKL_tT)$.
- The initialization algorithm in **Step-II** requires (UE,BS) units to propose to multiple subchannels and each subchannel makes a decision to accept multiple (UE,BS) units based on its preference list. The worst case that the proposing number of (UE,BS) unit is $N - S_t$. The complexity is $\mathcal{O}(N^2T^2)$.
- For the swap procedure in **Step-II**, there are T (UE,BS) units at N subchannels can perform swap operation. We consider the worst case that each (UE,BS) unit occupies S_t subchannels and each subchannel is shared by T_n (UE,BS) units. Therefore, in each iteration, for $(\mathrm{UE}, \mathrm{BS})_t$ the maximum swap operation number is $S_t(N - S_t)$. In each iteration, T (UE,BS) units require at most $\frac{1}{2}TT_nS_t(N - S_t)$ swap operations. For a given number of total iteration V', the complexity is approximated as $\mathcal{O}(V'TT_nS_tN)$.

As a result, the complexity of **Algorithm 1** can be calculated as $\mathcal{O}((K + N^2)T^2 + (VKL_t + V'T_nS_tN)T)$.

2) Stability and Convergence: After performing **Step-I** in **Algorithm 1**, any user UE_k with $k \in \mathcal{K}$ cannot find another BS BS_t, $t \in \mathcal{T}$ to form a swap-blocking pair under the current matching. Hence, a two-sided exchange-stable matching is formed between users and BSs. Then, by performing **Step-II** in **Algorithm 1** while treating the matched user and BS as a complete (UE,BS) unit, one can obtain a two-sided exchange-stable matching among (UE,BS) units and subchannels based on **Definition 2**. Since the utility function will increase monotonically by the swap operation in **Algorithm 1** and the utility function is bounded due to the transmit power constraint, **Algorithm 1** will terminate to a local solution after finite swap operation. Since the formulated two 2D matching subproblems are many-two-one matching with externalities and many-to-many matching with externalities, respectively, the proposed approaches in **Step-I** and **Step-II** converge to a two-sided exchange-stable status [35]. Note that not all two-sided exchange-stable matching are local optimal. The reason can be given by a simple example for **Step-I**: In a two-sided exchange-stable status, there exists possibility that user k associated to BS t refuses a swap as its utility would decrease, but user j associated to BS t' involved will benefit a lot from this swap operation and the utility of BS t and BS t' will increase. In this case, a forced swap will further increase the total utility compared to an approved swap matching. A similar case also exist for **Step-II**.

IV. Solutions for Power Allocation Optimization Problem

In this section, we try to solve the power allocation problem for given user-BS association and subchannel assignment. In this case, $\mathcal{K}_n^t$, $\mathcal{N}_k^t$, and $\mathcal{T}_k^n$ is known to BSs. For notation simplicity, we assume that $\pi(k) = k$ in the following. We first propose an optimal power allocation strategy based on BB algorithms.

Based on (12), the power allocation optimization problem can be formulated as:

$$\max_{\{P_{n,k}^t\}} \sum_{k\in\mathcal{K}} \mathrm{MOS}_{web}^k \tag{21a}$$

$$\text{s.t. } g_{n,k}^t \geq g_{n,j}^t, \quad k > j,\ \forall (k,j) \in \mathcal{K}_n^t,\ \forall t,\ \forall n, \tag{21b}$$

$$P_{n,k}^t \in \mathcal{P}, \quad \forall t,\ \forall n,\ \forall k, \tag{21c}$$

where $\mathcal{P} = \{P_{n,k}^t | \sum_{n\in\mathcal{N}^t} \sum_{k\in\mathcal{K}_n^t} P_{n,k}^t \leq P^t, P_{n,k}^t \geq 0, \forall n, \forall t\}$, $\mathcal{N}^t$ denotes the set of subchannels allocated to BS_t. Note that (21b) can be equivalently expressed as

$$\sum_{s\neq t} \left(h_{n,k}^t h_{n,j}^s - h_{n,j}^t h_{n,k}^s\right) P_n^s + \left(h_{n,k}^t - h_{n,j}^t\right)\sigma^2 \geq 0,\ k > j, \forall (k,j), \forall t, \forall n. \tag{22}$$

Though the constraints in (22) are linear and thus convex. However, problem (21) is still non-convex due to the non-convex objective function.

A. Optimal Power Allocation Strategy Using BB

In this subsection, we try to solve problem (21) over a M-dimensional simplex, where $M = \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{n=1}^{N} \nu_{t,k}\xi_{n,t}$ is the total number of variables. The key idea of BB approach can be described as follows: 1) transform the constraint sets into a multi-dimensional simplex; 2) compute upper and lower bounds.

First we introduce a set of variables $\Gamma = \{\Gamma^t_{n,k}, \forall t, \forall n, \forall k\}$ such that $\Gamma^t_{n,k} \leq \gamma^t_{n,k}$ in (6). Then, problem (21) can be transformed as

$$\max_{\{P^t_{n,k}\}\{\Gamma^t_{n,k}\}} \quad U(\Gamma) \quad \text{s.t. } \Gamma \in \mathcal{D}, \tag{23}$$

where $\mathrm{U}(\Gamma) = \sum_{k\in\mathcal{K}} C_1 \ln\left(\sum_{t\in\mathcal{T}_k}\sum_{n\in\mathcal{N}^t_k}\log_2\left(1+\Gamma^t_{n,k}\right)\right)$. In addition, the constraints $\mathcal{D}$ is defined as

$$\mathcal{D} = \left\{\Gamma \in \mathbb{R}^M \,\middle|\, \begin{array}{c}\Gamma^t_{n,k} \leq \gamma^t_{n,k}, \forall k, \forall n, \forall t, \\ \text{(21c) \& (22)}\end{array}\right\}. \tag{24}$$

Lemma 1: $U(\Gamma)$ *is an monotonically increasing function. More specifically,* $U(\Gamma)$ *a concave function.*

Proof: See Appendix A. ∎

Proposition 1: Problem (23) *have the same optimal solution to the optimization problem in* (21)*.*

Proof: See Appendix B. ∎

Let $\breve{\Gamma}$ be the largest possible SINR set. Note that $U(\Gamma) \in (-\infty, U(\breve{\Gamma})]$, where the minimum of $U(\Gamma)$ is unbounded. To tackle the difficulty, we introduce a new function $\tilde{U}(\Gamma)$

$$\tilde{U}(\Gamma) = \prod_{k\in\mathcal{K}}\Big(C_1\sum_{t\in\mathcal{T}_k}\sum_{n\in\mathcal{N}^t_k}\log_2\left(1+\Gamma^t_{n,k}\right)\Big), \tag{25}$$

which such that $\tilde{U}(\Gamma) \in [0, \tilde{U}(\breve{\Gamma}]$ and $U(\Gamma) = \ln\big(\tilde{U}(\Gamma)\big)$. Then, the transformation of problem (23) is

$$\max_{\{P^t_{n,k}\}\{\Gamma^t_{n,k}\}} \quad \tilde{U}(\Gamma) \quad \text{s.t. } \Gamma \in \mathcal{D}, \tag{26}$$

Problem (26) and problem (23) has the same optimal solution due to the monotonicity of logarithm function. Next we try to solve problem (26) using BB algorithms.

1) Construction of Multi-Dimensional Simplex $\mathcal{S}$*:* Let $\mathcal{S} = [v_1, v_2, \cdots, v_{M+1}]$ be an M-simplex in $\mathbb{R}^M$ satisfying $\mathcal{D}\bigcap\mathcal{S} \neq \emptyset$. The initial $\mathcal{S}$ should be a simple polytope tightly enclosing $\mathcal{D}$ with a small number of vertices. Because the feasible set $\mathcal{D}$ in (24) is an rectangular with removing some margins, a simple method to construct $\mathcal{S}$ is given as follows

$$\mathcal{S} = \{\Gamma \in \mathbb{R}^M | 0 \leq \Gamma^t_{n,k} \leq \breve{\Gamma}^t_{n,k}\} \tag{27}$$

with $\breve{\Gamma}^t_{n,k} = \frac{h^t_{n,k}P^t}{\sigma^2}$ denoting the largest possible SINR for UE_k on SC_n associated to BS_t. The vertex set of $\mathcal{S}$ is $V(\mathcal{S}) = \{v_1, v_2, \cdots, v_{M+1}\}$ with $v_1 = 0$, $v_j = \breve{\Gamma}^t_{n,k}e_j$, where e_j is the j-th basis vector of $\mathbb{R}^M$.

By constructing the multi-dimensional simplex, problem (23) has been transformed into a maximization of non-convex function $U(\Gamma)$ over an M-simplex $\mathcal{S}$.

2) Compute Lower and Upper Bounds: To compute lower and upper bounds, we first construct a bounding function, which is defined as

$$g(\Gamma) = \begin{cases} -\tilde{U}(\Gamma), & \text{if } \Gamma \in \mathcal{D}, \\ 0, & \text{otherwise.}\end{cases} \tag{28}$$

Note that for the feasible set $\mathcal{D}$ and the M-simplex $\mathcal{S}$ with $D \subseteq \mathcal{S}$, we have

$$\psi_{\min}(\mathcal{S}) = \inf_{\Gamma\in\mathcal{S}} g(\Gamma) = \inf_{\Gamma\in\mathcal{D}} -\tilde{U}(\Gamma). \tag{29}$$

which implies that the function $-\tilde{U}(\Gamma)$ is a lower bound of $g(\Gamma)$.

For $\mathcal{S}' = \{\hat{\Gamma} \leq \Gamma \leq \breve{\Gamma}\}$, we now have the lower bound and upper bound functions as

$$\begin{aligned}\psi_{\mathrm{lb}}(\mathcal{S}') &= \begin{cases} -\tilde{U}(\breve{\Gamma}), & \text{if } \hat{\Gamma} \in \mathcal{D}, \\ 0, & \text{otherwise.}\end{cases} \\ \psi_{\mathrm{ub}}(\mathcal{S}') &= \begin{cases} -\tilde{U}(\hat{\Gamma}), & \text{if } \hat{\Gamma} \in \mathcal{D}, \\ 0, & \text{otherwise.}\end{cases}\end{aligned} \tag{30}$$

From the definition of (28), one can know that $\psi_{\mathrm{lb}}(\mathcal{S}') = \psi_{\min}(\mathcal{S}') = \psi_{\mathrm{ub}}(\mathcal{S}') = 0$, if $\hat{\Gamma} \notin \mathcal{D}$. Consequently, for any $\mathcal{S}' \subseteq \mathcal{S}$, we have $\psi_{\mathrm{lb}}(\mathcal{S}') \leq \psi_{\min}(\mathcal{S}') \leq \psi_{\mathrm{ub}}(\mathcal{S}')$.

Based on the definition of lower and upper bounding functions in (30) and (30), the key step to compute the bounding functions is to check $\hat{\Gamma} \in \mathcal{D}$. It can be formulated as

$$\text{Find } \{P^t_{n,k}\} \tag{31a}$$

$$\text{s.t. } \gamma^t_{n,k} \geq \hat{\Gamma}^t_{n,k}, \tag{31b}$$

$$\sum_{s\neq t}\Big(h^t_{n,k}h^s_{n,j} - h^t_{n,j}h^s_{n,k}\Big)P^s_n + \big(h^t_{n,k} - h^t_{n,j}\big)\sigma^2 \geq 0, k > j, \quad \forall (k,j) \in \mathcal{K}^t_n, \ \forall t, \ \forall n, \tag{31c}$$

$$P^t_{n,k} \in \mathcal{P}, \quad \forall n, \ \forall t, \ \forall k. \tag{31d}$$

which is a convex problem on power allocation coefficients $\{P^t_{n,k}\}$.

Proposition 2: Problem (31) *can be transformed into a compact matrix form as follows:*

$$\text{Find } \mathbf{p}_n, n \in \mathcal{N} \tag{32a}$$

$$\text{s.t. } \mathbf{A}_n\mathbf{p}_n \succeq \mathbf{b}_n, \quad \bar{\mathbf{H}}_n\mathbf{p}_n \succeq \boldsymbol{\theta}_n, \ P^t_{n,k} \in \mathcal{P}, \ \forall n, \ \forall t, \ \forall k, \tag{32b}$$

where $\mathbf{A}_n = \mathbf{I} - (\boldsymbol{\Lambda}_n + \mathbf{D}_n\mathbf{G}_n)$ *and* $\boldsymbol{b}_n = \mathbf{D}_n\sigma^2$.

Proof: See Appendix C. ∎

Problem (32) is a linear programming (LP) feasibility, which can be described as follows. Define two sets $\mathbb{P}_1 = \{\mathbf{A}_n\mathbf{p}_n \succeq \mathbf{b}_n, P^t_{n,k} \in \mathcal{P}, \forall n, \forall t, \forall k\}$ and $\mathbb{P}_2 = \{\bar{\mathbf{H}}_n\mathbf{p}_n \succeq \boldsymbol{\theta}_n, P^t_{n,k} \in \mathcal{P}, \forall n, \forall t, \forall k\}$. The distance of the two sets is defined as

$$\mathbf{dist}(\mathbb{P}_1, \mathbb{P}_2) = \inf\{\|\mathbf{x}_1 - \mathbf{x}_2\| | \mathbf{x}_1 \in \mathbb{P}_1, \mathbf{x}_2 \in \mathbb{P}_2\}. \tag{33}$$

If the two sets intersect, the distance is zero. To find the distance between $\mathbb{P}_1$ and $\mathbb{P}_2$, we can solve the following QP

$$\min \ \|\mathbf{x}_1 - \mathbf{x}_2\| \quad \text{s.t. (32b).} \tag{34}$$

The optimal value is zero if and only if the two sets intersect. This problem is infeasible if and only if one of the sets is empty. Otherwise, the problem will return the optimal points $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively, that are close to each other.

One can verify that the matrix $\mathbf{\Lambda}_n + \mathbf{D}_n\mathbf{G}_n$ is irreducible nonnegative matrix. As in [37], a positive solution to $\mathbf{p}_n$ that satisfies $\mathbf{A}_n\mathbf{p}_n = \mathbf{b}_n$ exists if and only if the Perron-Frobeniuous eigenvalue of $\mathbf{\Lambda}_n + \mathbf{D}_n\mathbf{G}_n$, denoted as $\rho(\mathbf{\Lambda}_n + \mathbf{D}_n\mathbf{G}_n) < 1$. Therefore, we can check if $\hat{\Gamma} \in \mathcal{D}$ by the following proposition.

Proposition 3: For any $\hat{\Gamma}$, the following statements hold:

i) *If it exists one $n \in \mathcal{N}$ with $\rho(\mathbf{\Lambda}_n + \mathbf{D}_n\mathbf{G}_n) \geq 1$ or $\rho(\mathbf{I} + \bar{\mathbf{H}}_n) \geq 1$, we have $\hat{\Gamma} \notin \mathcal{D}$;*

ii) *If $\forall n \in \mathcal{N}$ with $\rho(\mathbf{\Lambda}_n + \mathbf{D}_n\mathbf{G}_n) < 1$ or $\rho(\mathbf{I} + \bar{\mathbf{H}}_n) < 1$, but $\exists n \in \mathcal{N}$ such that $\sum_{k\in\mathcal{K}_t}\sum_{n\in\mathcal{N}_t}\mathbf{p}_{n,i_k} > P^t$, we have $\hat{\Gamma} \notin \mathcal{D}$;*

iii) *If $\forall n \in \mathcal{N}$ with $\rho(\mathbf{\Lambda}_n + \mathbf{D}_n\mathbf{G}_n) < 1$, one need to check the LP feasibility problem in* (34)*.*

3) Optimal Power Allocation Based on BB Algorithms: Based on the above discussions, the procedures of the proposed BB algorithm for optimal power allocation is described as follows. Let $\mathcal{S}(v) = \{\mathcal{A}^1_{n,k}(v), \cdots, \cdots \mathcal{A}^T_{N,K}(v)\}$ denote the set of box subsets $\mathcal{A}^t_{n,k}(v) = \{\hat{\Gamma}^t_{n,k}(v) \leq \Gamma^t_{n,k} \leq \breve{\Gamma}^t_{n,k}(v)\}$ for all n, k and t at the v-th iteration. $\mathcal{S}(0)$ is the initial rectangular constraint set, on the root node of the binary tree, which is define (27). At the v-th iteration, we spilt $\mathcal{S}(v)$ into two subsets $\mathcal{Q}_I$ and $\mathcal{Q}_{II}$ along one of its longest edges, removing $\mathcal{S}(v)$ and adding the two new subsets to $\mathcal{R}(v)$. Next, we solve (31) based on **Proposition 3** over each subset $\mathcal{Q}_l,\ l \in \{I, II\}$. A lower bound and an upper bound can be obtained. Then, we choose the minimum over all upper bounds as $f_{ub}(v)$ and choose the minimum over all lower bounds as $f_{lb}(v)$, i.e., taking the minimum over all the upper and lower bounds at each leaf node across all the levels in the binary tree. Removing the leaf node $\mathcal{S}'$ such that $\psi_{lb}(\mathcal{S}') \geq f_{ub}(v)$, which will not affect the optimality of the BB tree. Repeat the above procedures until it satisfies the accuracy ϵ which is the difference between the global upper bound and the global lower bound. In the procedure of generating the BB tree, a sequence of subsets will be generated from $\mathcal{S}(0)$. The details are given in **Algorithm 2** that captures the global optimal solution of (12).

Remark 1: At the v-th iteration of **Algorithm 2***, $f_{ub}(v)$ and $f_{lb}(v)$ are the minimums over all the upper bounds and lower bounds at each leaf nodes in the BB tree, respectively, which give a global upper bound and lower bound on the optimal value of* (25)*. The stopping criterion for* **Algorithm 2** *can be $f_{ub}(v) - f_{lb}(v) \leq \epsilon$ for given a small ϵ. Accordingly, it means that $U^* - \ln\epsilon \leq \mathcal{U}^{\text{opt}}$.*

The overall complexity of **Algorithm 2** is determined by the complexity of each iteration and the number of iterations required for achieving the desired tolerance. During each iteration, it requires to solve a LP problem for the worst case. Since the formulated LP problem in (32) can be solved using an interior-point method, the computational complexity of LP is upper bounded by $\mathcal{O}((NKT)^2(NKT + T + NTL_t))$ [32],

Algorithm 2 The Optimal Power Allocation Algorithm Based on BB

1: Initialization for BB:
   1) Construct the initial simplex $\mathcal{S}(0)$ such that $\mathcal{D} \subseteq \mathcal{S}(0)$, which was described in Section IV-A.1.
   2) Compute $f_{lb}(1) = \psi_{\text{lb}}(\mathcal{S}_0)$ and $f_{ub}(1) = \psi_{\text{ub}}(\mathcal{S}(0))$, by (30) and (30), respectively.
   3) Set $\mathcal{R}(1) = \{\mathcal{S}_0\}$, optimal upper bound $U^* = U(1)$, tolerance $\epsilon > 0$ and $v = 1$.
2: **while** $f_{ub}(v) - f_{lb}(v) > \epsilon$ **do**
3: Pick $\mathcal{S}' \in \mathcal{R}(v)$ for which $\phi_{lb}(\mathcal{S}') = f_{lb}(v)$ and set $\mathcal{S}(v) = \mathcal{S}'$.
4: Subdivide $\mathcal{S}(v)$ along one of its longest edges into $\mathcal{Q}_I$ and $\mathcal{Q}_{II}$.
5: Compute $\psi_{\text{lb}}(\mathcal{Q}_I)$, $\psi_{\text{ub}}(\mathcal{Q}_{II})$ by solving problem (31).
6: Update the upper bound $f_{ub}(v)$ and the lower bound $f_{lb}(v)$ as follows:
$f_{lb}(v) = \min_{\mathcal{S}'\in\mathcal{R}(v+1)} \psi_{\text{lb}}(\mathcal{S}')$;
$f_{ub}(v) = \min_{\mathcal{S}'\in\mathcal{R}(v+1)} \psi_{\text{ub}}(\mathcal{S}')$;
update $f^*_{ub} = \min(f^*_{ub}, f_{ub}(v))$.
7: Update $\mathcal{R}(v+1)$ by removing all $\mathcal{S}'$ for which $\psi_{\text{lb}}(\mathcal{S}') \geq f_{ub}(v+1)$.
8: $v := v + 1$.
9: **end while**
10: Output the value $\tilde{U}^* = f^*_{ub}$ and the optimal power allocation P^*.

where term NKT denotes the number of variables and term $NKT + T + NTL_t$ is the number of constraints. In addition, the worst case computational complexity of **Algorithm 2** is exponential in the number of variables. Assume that 2^{KNT} is the total number of iterations required to obtain the ϵ-approximation solution. The complexity of BB algorithm can be approximated as $\mathcal{O}(2^{(KNT)^4})$.

B. Low-Complexity Power Allocation Strategies

Though BB approaches can find the optimal power allocation, the high computational complexity makes it difficult to realize. In this subsection, we proposed a suboptimal power allocation strategy based on the SCA techniques.

We first consider the MOS function in (12a), which can be rearranged as follows.

$$\begin{aligned}\sum_{k\in\mathcal{K}} C_1 \ln(\sum_{t=1}^{T}\sum_{n=1}^{N} R^t_{n,k}) + \sum_{k\in\mathcal{K}} C^k_3 \\ \overset{(a)}{\geq} \sum_{k\in\mathcal{K}} C_1 \sum_{t=1}^{T}\sum_{n=1}^{N} \ln(R^t_{n,k}) + \sum_{k\in\mathcal{K}} C^k_3 \\ \overset{(b)}{=} \sum_{n=1}^{N}\sum_{t\in\mathcal{T}^n}\sum_{k\in\mathcal{K}^t_n} C_1 \ln(R^t_{n,k}) + \sum_{k\in\mathcal{K}} C^k_3, \end{aligned} \tag{35}$$

where (a) follows the Jenssen's inequality [32] and (b) is based on the property of polynomials.

As a result, problem (21) can be equivalently reformulated as

$$\max_{\{P^t_{n,k}\}\{R^t_{n,k}\}} \sum_{n=1}^{N}\sum_{t\in\mathcal{T}^n}\sum_{k\in\mathcal{K}^t_n} C_1\ln(R^t_{n,k}) \tag{36a}$$

$$\text{s.t. } R^t_{n,k}\le\log\Big(1+\frac{h^t_{n,k}P^t_{n,k}}{\sum\limits_{i=k+1}^{|\mathcal{K}^t_n|} h^t_{n,k}P^t_{n,i}+\sum\limits_{s\neq t}h^s_{n,k}P^s_n+\sigma^2}\Big),$$

$$\text{(21b) \& (21c) \& (22).} \tag{36b}$$

where the relax rate constraint in (36b) will be strictly equal at the optimal solution. Problem (36) is non-convex due to the constraints in (36b). To solve it, we propose a convex approximation method in the following.

To illustrate the approximation, let us first consider the following change of variables:

$$e^{y^t_{n,k}}=2^{R^t_{n,k}}-1,\quad e^{z^t_{n,k}}=P^t_{n,k}, \tag{37}$$

for $\forall k\in\mathcal{K}$, $\forall n\in\mathcal{N}$, and $\forall t\in\mathcal{T}$, where $x^t_{n,k}$ and $y^t_{n,k}$ are slack variables. By substituting (37) into (36), one can obtain the following problem:

$$\max_{\substack{\{P^t_{n,k}\},\{R^t_{n,k}\},\\ \{x^t_{n,k}\},\{y^t_{n,k}\}}} \sum_{n=1}^{N}\sum_{t\in\mathcal{T}^n}\sum_{k\in\mathcal{K}^t_n} C_1\ln(R^t_{n,k}) \tag{38a}$$

$$\text{s.t. } e^{y^t_{n,k}-z^t_{n,k}}\Big(\sum_{i=k+1}^{|\mathcal{K}^t_{n,k}|}e^{z^t_{n,i}}+\sum_{s\neq t}\frac{h^s_{n,k}}{h^t_{n,k}}P^s_n(z)+\frac{\sigma^2}{h^t_{n,k}}\Big)\le 1, \tag{38b}$$

$$\sum_{n\in\mathcal{N}^t}\sum_{k\in\mathcal{K}^t_n}P^t_{n,k}(z)\le P^t, \tag{38c}$$

$$R^t_{n,k}\le\log_2(1+e^{y^t_{n,k}}), \tag{38d}$$

$$\sum_{s\neq t}h^t_{n,j}h^s_{n,k}P^s_n(z)+(h^t_{n,j}-h^t_{n,k})\sigma^2\le\sum_{s\neq t}h^t_{n,k}h^s_{n,j}P^s_n(z), \tag{38e}$$

$$\forall k\in\mathcal{K},\quad\forall n\in\mathcal{N},\ \forall t\in\mathcal{T}, \tag{38f}$$

where $P^s_n(z)=\sum_{i=1}^{|\mathcal{K}^s_n|}e^{z^s_{n,k}}$. Notice that we have replaced the equalities in (36b) and (37) with inequalities as in (38b) and (38d). Due to the monotonicity of the objective function, all inequalities in (38b) and (38d) would hold with equalities at the optimal solution. It is observed that the objective function is concave and the constraints in (38b) and (38c) are convex. Furthermore, constraints in (38d) and (38e) are not convex. Next, we use the first-order Taylor approximation to approximate the lower bound of the non-convex parts in (38d) and (38e). Let $\{\tilde{P}^t_{n,k}\}$ and $\{\tilde{R}^t_{n,k}\}$ be a set of feasible solution of (38). Then, we have

$$\tilde{y}^t_{n,k}=\ln 2^{\tilde{R}^t_{n,k}}-1,\quad \tilde{z}^t_{n,k}=\ln(P^t_{n,k}). \tag{39}$$

As a result, the lower-bound approximation for (38d) and (38e) can be given by

$$\log_2(1+e^{y^t_{n,k}})=\log_2(1+e^{\tilde{y}^t_{n,k}})+\frac{e^{\tilde{y}^t_{n,k}}(y^t_{n,k}-\tilde{y}^t_{n,k})}{\ln(2)\big(1+e^{y^t_{n,k}}\big)}, \tag{40a}$$

$$\sum_{s\neq t}h^t_{n,k}h^s_{n,j}P^s_n(z)=\sum_{s\neq t}h^t_{n,k}h^s_{n,j}P^s_n(\tilde{z})+\sum_{s\neq t}h^t_{n,k}h^s_{n,j}\sum_{i=1}^{\mathcal{K}^s_n}e^{\tilde{z}^s_{n,i}}(z^s_{n,i}-\tilde{z}^s_{n,i}), \tag{40b}$$

Consequently, by replacing (38d) and (38e) with (40a) and (40b), we obtain the following approximation of problem (38):

$$\max_{\substack{\{P^t_{n,k}\},\{R^t_{n,k}\},\\ \{x^t_{n,k}\},\{y^t_{n,k}\}}} \sum_{n=1}^{N}\sum_{t\in\mathcal{T}^n}\sum_{k\in\mathcal{K}^t_n} C_1\ln(R^t_{n,k}) \tag{41a}$$

$$\text{s.t. (38b) \& (38c) \& (38f),}$$

$$R^t_{n,k}\le\log_2(1+\tilde{y}^t_{n,k})+\frac{\tilde{y}^t_{n,k}(y^t_{n,k}-\tilde{y}^t_{n,k})}{\ln(2)\big(1+e^{y^t_{n,k}}\big)}, \tag{41b}$$

$$\sum_{s\neq t}h^t_{n,j}h^s_{n,k}P^s_n(z)+(h^t_{n,j}-h^t_{n,k})\sigma^2\le\sum_{s\neq t}h^t_{n,k}h^s_{n,j}P^s_n(\tilde{z})+\sum_{s\neq t}h^t_{n,k}h^s_{n,j}\sum_{i=1}^{\mathcal{K}^s_n}e^{\tilde{z}^s_{n,i}}(z^s_{n,i}-\tilde{z}^s_{n,i}). \tag{41c}$$

Problem (41) is a convex optimization problem; it can be efficiently solved by standard convex solvers such as CVX [38].

Problem (41) is formulated by approximating (38) at a feasible solution $(\{R^t_{n,k}\},\{P^t_{n,k}\})$, as described in (39). Note that for a fixed point $(\{R^t_{n,k}\},\{P^t_{n,k}\})$, the obtained objective of (41) is no less than that obtained in the fixed point. Therefore, the approximation can be improved by successively approximating problem (38) based on the optimal solution $(\{R^t_{n,k}\},\{P^t_{n,k}\})$ obtained by solving (41) in the previous approximation. The completed procedure of the proposed successive approximation approach is described in **Algorithm 3**. In particular, in each iteration of **Algorithm 3**, the objective function of problem (41) will be improved successively. However, due to the total power constraint, the generated sequence is bounded, which implies the convergence of **Algorithm 3**.

Due to the relaxation in (35), (38) provides a lower bound of problem (21). Furthermore, problem (41) is a lower bound approximation of problem (38) because of the approximation in (40). Consequently, the solution obtained is suboptimal. However, the complexity of **Algorithm 3** is lower than **Algorithm 2**. Assume that the number of iterations of **Algorithm 3** is $\bar{V}$, then $\bar{V}$ is less than $\mathcal{O}((NKT)^2)$ [39].

V. Simulation Results

In this section, the simulation and the performance results are evaluated to the performance of the multi-cell NOMA

Algorithm 3 SCA Algorithm for Solving (38)

1: Given a set of solution $(\{\tilde{R}^t_{n,k}[0]\},\{\tilde{P}^t_{n,k}[0]\})$, which is feasible to (38).
2: Compute the optimal objective value of problem (38), denoted as $\Phi(0)$. Set $v=1$.
3: **while** $\frac{|\Phi[v]-\Phi[v-1]|}{\Phi[v-1]} \leq \epsilon'$, where ϵ' is a given stopping criterion, **do**
4: $v := v+1$.
5: Obtain $\tilde{y}^t_{n,k}(v-1)$ and $\tilde{z}^t_{n,k}(v-1)$ by (39), and solve problem (41) to obtain the optimal solution $(\{\tilde{R}^t_{n,k}[v]\},\{\tilde{y}^t_{n,k}(v)\},\{\tilde{z}^t_{n,k}[v]\})$.
6: **end while**
7: Output the optimal $\{R^{t*}_{n,k}\}$ and $\{P^t_{n,k}=\ln(z^{t*}_{n,k})\}$.

system with proposed resource allocation scheme. In the simulations, the locations of the BSs are assumed to be fixed, and the locations of the users to be uniformly and independently distributed in a disk space with radius $R=500$ m, if it is not specified. The bandwidth of each subchannel is $\frac{W}{N}=75$ kHz. $h^t_{n,k}=|f^t_{n,k}|^2$ is the channel power gain from BS t to user UE_k on subchannel SC_n, which is expressed as $h^t_{n,k}=(d_{t,k})^{-\alpha}\chi^t_{n,k}$, where $d_{t,k}$ is the distance between BS t and user UE_k, α is the path loss exponent and $\chi^t_{n,k}$ is the fading coefficient with $\chi^t_{n,k}\sim\mathcal{CN}(0,1)$. We assume that the small scale fading parts of all channels follow from independent identically distributed (i.i.d) Rayleigh distribution. A total of 100 different channel realizations were used in the simulations, if it is not specified. The path loss exponent is 3.7 [24] and the noise experienced at each user is assumed identical. The noise power is $\sigma^2=-174+10\log 10(\frac{W}{N})$ dBm. Moreover, we consider that the system consists of $K=6$ users, $T=3$ BSs and $N=3$ subchannels, if it is not specified.

For a web browsing application, the web page sizes are determined according to the web traffic statistics collected and analyzed in the previous study [40]. In simulations, web users typically access the web page size with the average average web page size of 320 KB [27], if it is not specified.

To investigate the performance of the proposed multi-cell NOMA system, three different algorithms are simulated for the multi-cell NOMA system and the multi-cell OMA system, respectively. Firstly, we consider the global optimal resource allocation algorithm called 'Exhaust+BB'. In 'Exhaust+BB', for each combination among users, BSs and subchannels, the BB approach is invoked to attain the optimal power allocation scheme; Then, an exhaust search is exploited over all combinations of user-BS association and subchannel assignment. Furthermore, for the suboptimal algorithm-'Match+BB', the proposed matching approach is firstly invoked to obtain a suboptimal scheme of user-BS association and subchannel assignment; Then, the power allocation procedure is performed by applying BB approaches. In addition, for the low-complexity suboptimal algorithm-'Match+SCA', the proposed SCA-based power allocation scheme is applied after performing the proposed matching approach. Moreover, we also consider an multi-cell OMA system with TDMA, where an BS communicates with at most one user in one time slot. Since the BS applying NOMA can serve multiple users simultaneously in the same subchannel, the BS applying OMA requires multiple time slots to serve the same number of users in NOMA.

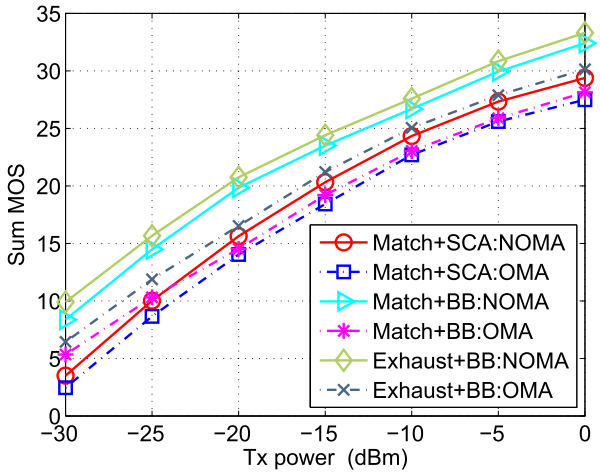


Fig. 4. Comparisons of sum MOS over different algorithms, where $K=6, T=3, N=3$, $L_t=2, T_n=2$, and $S_n=2$.

In Fig. 4, we investigate the sum MOS versus the maximum transmit power at each BS, P^t, for different algorithms mentioned above where $L_t=2, T_n=2$, and $S_n=2$. As it can be observed from Fig. 4, the sum MOS attained by different algorithms increases with the maximum transmit power of BSs P^t. This is because the received SINR at the users can be improved by optimally allocating the transmit power via the proposed algorithms which leads to an improvement of the system sum MOS. However, there is a diminishing trend in the sum MOS when P^t is higher than -10 dBm. In fact, as the P^t increases, the inter-cell interference becomes more severe, which degrades the received SINR at users. As a result the sum MOS of the systems will decrease. Besides, it can be observed from Fig. 4, the sum MOS of the global optimal algorithm-'Exhaust+BB' grows faster than 'Match+SCA' and 'Match+BB'. **Table I** compares various algorithms from the perspective of computational complexity. It shows that the complexity of 'Match+SCA' is greatly less than that of 'Exhaust+BB' and 'Match+BB'. From Fig. 4 and **Table I**, we can observe that though some performance is suffered in the proposed low-complexity suboptimal algorithm-'Match+SCA', its computational complexity will be reduced greatly compared to 'Exhaust+BB' and 'Match+BB', which indicates that the proposed 'Match+SCA' is efficient to solve the optimization problem (12). Furthermore, note that the proposed multi-cell NOMA system outperforms the conventional multi-cell OMA system in terms of the sum MOS.

In Fig. 5, we study the performance of the system sum MOS and the system sum rate over different P^t with $K=6$. We assume the 6 users run page applications with the $FS=[50,100,200,320,400,500]$ KB. For comparison, we consider the fixed power allocation scheme in NOMA, called as FNOMA, as a baseline, which has been widely adopted in [5] and [41]. Correspondingly, we term the

TABLE I
COMPARISON OF VARIOUS ALGORITHMS

Algorithm	Complexity	Optimality
Exhaust+BB	$\mathcal{O}(2^{(NKT)2^{(KNT)^2}})$	Global optimal
Match+BB	$\mathcal{O}((K+N^2)T^2+(VKL_t+V'T_nS_tN)T)+\mathcal{O}(2^{(KNT)^2}))$	Suboptimal
Match+SCA	$\mathcal{O}((K+N^2)T^2+(VKL_t+V'T_nS_tN)T)+\mathcal{O}(\bar{V}(NKT)^3)$	Suboptimal

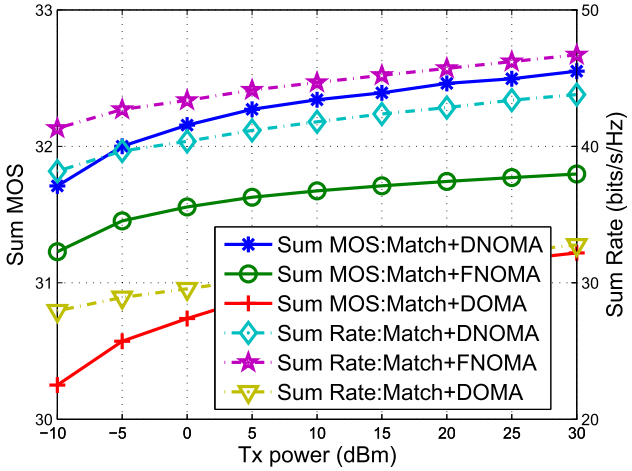


Fig. 5. Comparisons of sum MOS and sum rate with different transmit power, where $K=6, T=3, N=3$, $L_t=2, T_n=2, S_n=2$ and $[\bar{p}_1^t,\ \bar{p}_2^t]^T=[1/4,\ 3/4]^T$.

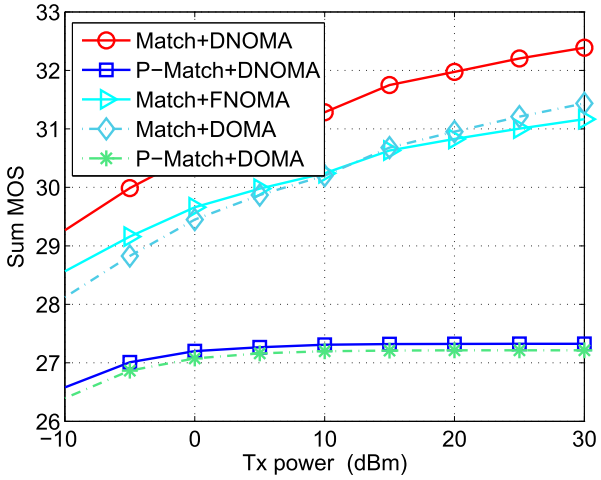


Fig. 6. Comparisons of sum MOS with different matching schemes, where $K=6, T=3, N=4$, $L_t=2, T_n=2$, and $S_n=2$.

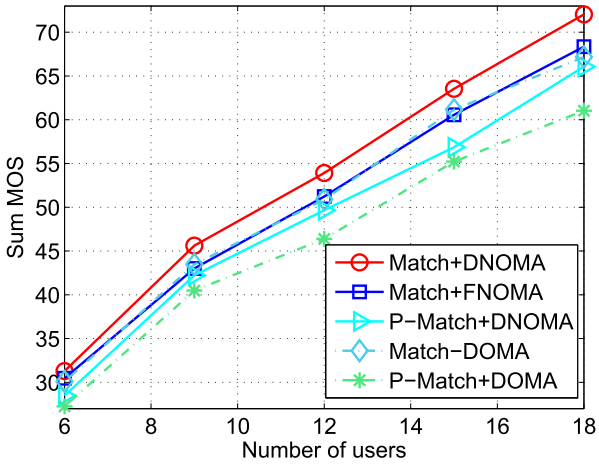


Fig. 7. Comparisons of sum MOS with different number of users, where $T=3, N=4$, $T_n=2$, and $S_n=3$.

proposed low-complexity suboptimal power allocation scheme based on SCA as DNOMA. To validate the effectiveness, we compare the proposed 'Match+DNOMA' with the corresponding OMA scheme, termed as 'Match+DOMA', and 'Match+FNOMA'. In 'Match+FNOMA', we assume that the BS allocates its power uniformly over the occupied subchannels. Besides, $L_t=2$ is assumed and the power allocation coefficients between the users associated the BS on a specific subchannel is assumed to be $\bar{p}_1^t$ and $\bar{p}_2^t$ for the users with the better equivalent channel gain and the poorer equivalent channel gain, respectively. As can be observed from Fig. 5, both the performance of sum MOS can be greatly enhanced by 'Match+DNOMA' compared with 'Match+FNOMA' and 'Match+DOMA'. Moreover, the curves of sum MOS and sum rate have similar increasing trends. It is because that one user's MOS function is a monotonically increasing function with its rate. In particular, as can be observed in (11), one user's MOS function is related with the logarithm of its rate and the applied web page size, which results in 'Match+FNOMA' can obtain a better sum-rate performance compare to 'Match+DNOMA'.

In Fig. 6, we investigate the sum-MOS performance of the proposed multi-cell NOMA networks for $K=6$, $T=3$ and $N=4$. Three different schemes are illustrated in Fig. 6: 'Match+DNOMA', 'Match+FNOMA' and 'P-Match+DNOMA'. The impact of the subchannel assignment is studied, where matching operation only performed for user-BS association and the subchannels randomly assigned to the BS satisfying the constraints in problem (12). It can be termed as 'P-Match+DNOMA' for simplicity. In addition, for completeness, the corresponding OMA schemes are also simulated. As can be observed from Fig. 6, 'Match+DNOMA' is capable of increasing the sum MOS compared to the other schemes. Moreover, the sum-MOS performance of 'P-Match+DNOMA' is worse than that of 'Match+DNOMA' and 'Match+FNOMA', which indicates that the subchannel assignment has important impact on the network utilities. Combined with the observations from Fig. 4, it can be concluded that 'Match+SCA' strikes a balance between the performance gain and the computational complexity.

In Fig. 7, we investigate the performance of the proposed multi-cell MC-NOMA networks versus different number of

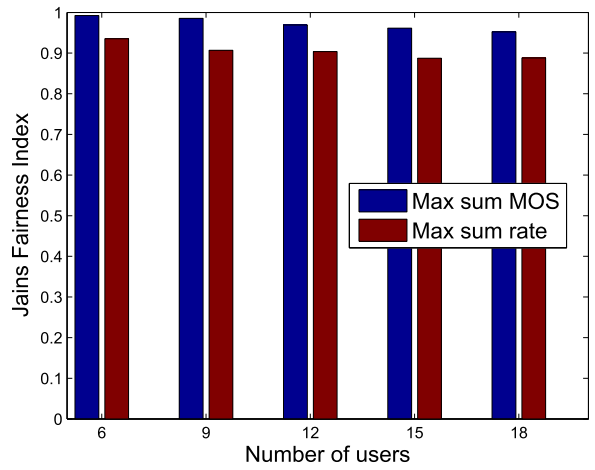


Fig. 8. Comparisons of the JFI between QoE-based and sum rate-based scheme with different number of users, where $T = 3, N = 4$, $T_n = 2$, and $S_n = 3$.

users in the system. Here the number of users associated with one BS is defined as the average number of the total users over the number of BSs, $L_t = \frac{K}{T}$. Moreover, three different NOMA-based schemes are compared: 'Match+DNOMA', 'Match+FNOMA', and 'P-MAtch+DNOMA'. For comparison, two OMA schemes are also compared in Fig. 7: 'Match+DOMA' and 'P-Match+DOMA'. As can be observed from Fig. 7, for all schemes, the sum-MOS performance will be enhanced with increasing the number of the number of users in the system. Besides, 'Match+DNOMA' is capable of outperforming the other schemes.

To illustrate the impact of sum MOS on the user fairness of multi-cell NOMA systems, we investigate the fairness of the proposed schemes and the baseline schemes based on Jain's fairness index (JFI) [42], which is an important indicator of measuring the performance metric. In particular, JFI is calculated as $J_{\mathrm{QOE}} = (\sum_{k\in\mathcal{K}} \mathrm{MOS}_{web}^{k})^2 / K \sum_{k\in\mathcal{K}} (\mathrm{MOS}_{web}^{k})^2$. Note that the JFI translates a set of MOS vector $\{\mathrm{MOS}_{web}^{1}, \cdots, \mathrm{MOS}_{web}^{K}\}$ into a score in the interval of $[\frac{1}{K}, 1]$ and higher JFI means the resource allocation is fairer.

Fig. 8 illustrates the evaluation of JFI versus the number of users in the network. Here, the JFIs in terms of the optimization objective with sum-MOS based maximization and sum-rate based maximization, where the proposed 'Match+DNOMA' scheme was employed. The JFI of sum-MOS based maximization is higher than that of sum-rate based, because the sum MOS function reduces the gap between users' rates. Moreover, the JFIs in terms of the two schemes decrease with the total number of users since the competition among users becomes more tenser when there are more users in the system.

VI. Conclusions

In this paper, we studied the QoE-based resource allocation algorithm design of an multi-cell MC-NOMA system in terms of user-BS association, subchannel assignment and power allocation. The algorithm design was formulated as a combinatorial non-convex optimization problem of maximizing the sum MOS of the system. By formulating user-BS association and subchannel assignment as a 3D matching problem, we proposed a low-complexity two-step approach based on 2D matching. Then, we developed an optimal power allocation strategy based on BB approaches to derive an upper bound for the sum MOS of the system. Besides, a suboptimal algorithm based on SCA was also developed to achieve a trade-off between computational complexity and performance. Simulation results has revealed that the proposed suboptimal algorithm obtain a good performance compared to the optimal algorithm. In addition, a substantial improvement of the sum MOS can be achieved by employing the proposed MC-NOMA scheme in multi-cell networks. Furthermore, the proposed QoE-based multi-cell MC-NOMA scheme was shown to provide a good balance between improving the sum MOS and maintaining fairness among users. In addition, it is promising direction to investigate a general algorithm to improve the user QoE for various services such as web browsing, voices, streaming audio and video, and so on. Therefore our future work will consider a general algorithm for MOS models with various services with the aid of the algorithms developed in this work.

Appendix A
Proof of Lemma 1

Since $\{\Gamma\}$ is continuous on $\mathbb{R}^M$, $U(\gamma)$ is differentiable on $\Gamma_{n,k}^{t}$, $\forall n \in \mathcal{N}_k^t$, $t \in \mathcal{T}_k$ and $\forall k \in \mathcal{K}$. The first-order derivatives on $\Gamma_{n',k'}^{t'}$ can be derived as

$$\frac{\partial U(\Gamma)}{\partial \Gamma_{n',k'}^{t'}} = \frac{C_1}{\sum_{t\in\mathcal{T}_{k'}} \sum_{n\in\mathcal{N}_{k'}^{t}} \log_2\left(1+\Gamma_{n,k'}^{t}\right)} \cdot \frac{1}{1+\Gamma_{n',k'}^{t'}}, \tag{A.1}$$

Note that $\sum_{t\in\mathcal{T}_{k'}} \sum_{n\in\mathcal{N}_{k'}^{t}} \log_2\left(1+\Gamma_{n,k'}^{t}\right)$ is the effective rate for UE_k. In the system, we assume that all users are scheduled where each user k such that $R_k > 0$. Note that the gradient of $\nabla U(\Gamma) \succeq 0$, $U(\Gamma)$ is an monotonically increasing function. Then we can compute the second-order derivatives of $U(\Gamma)$ as

$$\frac{\partial^2 U(\Gamma)}{\partial\left(\Gamma_{n',j'}^{t'}\right)^2} = \frac{-C_1\left(\sum_{t\in\mathcal{T}_{k'}} \sum_{n\in\mathcal{N}_{k'}^{t}} \log_2\left(1+\Gamma_{n,k'}^{t}\right)+1\right)}{\left(\left(1+\Gamma_{n',k'}^{t'}\right) \sum_{t\in\mathcal{T}_{k'}} \sum_{n\in\mathcal{N}_{k'}^{t}} \log_2\left(1+\Gamma_{n,k'}^{t}\right)\right)^2} \tag{A.2}$$

Obviously, $\frac{\partial^2 U(\Gamma)}{\partial(\Gamma)^2} \preceq 0$, which implies that $U(\Gamma)$ is concave [32].

APPENDIX B
PROOF OF PROPOSITION 1

The objective function in problem (21) can be written as

$$\begin{aligned}&\sum_{k\in\mathcal{K}}\mathrm{MOS}_{web}^{k}\\&=\sum_{k\in\mathcal{K}}\left(C_1\ln\left(\sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{N}}\log_2\left(1+\Gamma_{n,k}^{t}\right)\right)+C_3^k\right)\\&=\underbrace{\sum_{k\in\mathcal{K}}C_1\ln\left(\sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{N}}\log_2\left(1+\Gamma_{n,k}^{t}\right)\right)}_{U'(\Gamma)}+\underbrace{\sum_{k\in\mathcal{K}}C_3^k}_{\text{Constant}}\end{aligned}\quad\text{(B.1)}$$

Note that $\sum_{k\in\mathcal{K}}C_3^k$ is constant, which will not affect the optimal solution. In addition, based on **Lemma 1**, we have proved that $U'(\Gamma)$ is a monotonically increasing function of Γ. Based on the property, at the optimum the strict equality will be satisfied for $\Gamma_{n,k}^{t}\le\gamma_{n,k}^{t}$, $\forall n,\forall k$, and $\forall t$. Therefore, the relaxation is tight. Problem (23) will have the same optimal solution with problem (21).

APPENDIX C
PROOF OF PROPOSITION 2

By rearranging (31b) as

$$P_{n,k}^{t}-\Gamma_{n,k}^{t}\sum_{i=k+1}^{|\mathcal{K}_n^t|}P_{n,i}^{t}-\frac{\Gamma_{n,k}^{t}}{h_{n,k}^{t}}\sum_{s\neq t}|f_{n,k}^{s}|^2P_n^s\ge\frac{\Gamma_{n,k}^{t}}{h_{n,k}^{t}}\sigma^2.\quad\text{(C.1)}$$

For any subchannel SC_n, (C.2) can be expressed as

$$\left(\mathbf{I}-(\mathbf{\Lambda}_n+\mathbf{D}_n\mathbf{G}_n)\right)\mathbf{p}_n\succeq\mathbf{D}_n\sigma^2,\quad\text{(C.2)}$$

where

$$\begin{aligned}\mathbf{\Lambda}_n&=\mathrm{diag}\left(\left[\Lambda_n^1,\mathbf{\Lambda}_n^2,\cdots,\mathbf{\Lambda}_n^{T_n}\right]\right),\\\mathbf{D}_n&=\mathrm{diag}\left(\left[\boldsymbol{D}_n^1,\boldsymbol{D}_n^2,\cdots,\boldsymbol{D}_n^{T_n}\right]\right),\\\mathbf{p}_n&=[\mathbf{p}_n^1,\mathbf{p}_n^2,\cdots,\mathbf{p}_n^{T_n}]',\mathbf{G}_n=\left[\mathbf{G}_n^1,\mathbf{G}_n^2,\cdots,\mathbf{G}_n^{T_n}\right]'.\end{aligned}$$

and $\mathbf{\Lambda}_n^t$ is an upper triangle matrix with the element in the i-th row and the j-th column $\mathbf{\Lambda}_n^t[i,j]=\Gamma_{n,i}^{t}$ with $j>i$.

$$\mathbf{D}_n^t=\mathrm{diag}\left(\left[\tfrac{\Gamma_{n,1}^{t}}{h_{n,1}^{t}},\tfrac{\Gamma_{n,2}^{t}}{h_{n,2}^{t}},\cdots,\tfrac{\Gamma_{n,L_t}^{t}}{h_{n,L_t}^{t}}\right]\right),\quad\text{(C.4a)}$$

$$\mathbf{p}_n^t=\left[P_{n,1}^{t},P_{n,2}^{t},\cdots,P_{n,L_t}^{t}\right]',\quad\text{(C.4b)}$$

$$\mathbf{G}_n^t=\left[h_{n,1}^{1}\mathbf{1}'_{L_t},\cdots,h_{n,1}^{t-1}\mathbf{1}'_{L_t},\mathbf{0}'_{L_t},h_{n,1}^{t+1}\mathbf{1}'_{L_t},\cdots,h_{n,L_t}^{T_n}\mathbf{1}'_{L_t}\right]\quad\text{(C.4c)}$$

For example, we assume that $L_t=2$ and $T_n=3$, for $t=2$, we have

$$\begin{aligned}&\mathbf{\Lambda}_n^2=\begin{bmatrix}0&\Gamma_{n,1}^{2}\\0&0\end{bmatrix},\quad\mathbf{D}_n^2=\begin{bmatrix}\frac{\Gamma_{n,1}^{t}}{h_{n,1}^{t}}&0\\0&0\end{bmatrix},\ \mathbf{p}_n^2=\left[P_{n,1}^{t},P_{n,2}^{t}\right]',\\&\mathbf{G}_n^t=\left[h_{n,1}^{1},h_{n,2}^{1},0,0,h_{n,1}^{3},h_{n,2}^{3}\right]\end{aligned}$$

Analogously, for a pair of users UE_k and UE_j, with $k>j$, we can rewritten constraints in (31c) as follows.

$$\overline{\mathbf{H}}_n\bar{\mathbf{p}}_n\ge\boldsymbol{\theta}_n\quad\text{(C.6)}$$

where

$$\begin{aligned}\bar{\mathbf{H}}_n&=\left[\bar{\mathbf{H}}_n^1,\bar{\mathbf{H}}_n^2,\cdots,\bar{\mathbf{H}}_n^{T_n}\right],\quad\boldsymbol{\theta}_n=\left[\theta_n^1,\theta_n^2,\cdots,\theta_n^{T_n},\right],\\\bar{\mathbf{p}}_n&=\left[P_n^1,P_n^2,\cdots,P_n^{T_n}\right]',\\\mathbf{H}_n^t&=\left[\bar{h}_n^1,\cdots,\bar{h}_n^{t-1},0,\bar{h}_n^{t+1},\cdots,\bar{h}_n^{T_n}\right]\\\bar{h}_n^s&=|f_{n,k}^{t}|^2|f_{n,j}^{s}|^2-|f_{n,j}^{t}|^2|f_{n,k}^{s}|^2,\quad s\neq t,\\\theta_n^t&=(|f_{n,k}^{t}|^2-|f_{n,k}^{s}|^2)\sigma^2.\end{aligned}$$

Substituting (C.2) and (C.6) into optimization problem (31), we can attain problem (32).

REFERENCES

[1] J. Cui, Y. Liu, P. Fan, and A. Nallanathan, "A QoE-aware resource allocation strategy for multi-cell NOMA networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.

[2] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: Next-generation wireless broadband technology [Invited Paper]," *IEEE Wireless Commun.*, vol. 17, no. 3, pp. 10–22, Jun. 2010.

[3] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

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

[5] Z. Ding *et al.*, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[6] H. Shao, H. Zhao, Y. Sun, J. Zhang, and Y. Xu, "QoE-aware downlink user-cell association in small cell networks: A transfer-matching game theoretic solution with peer effects," *IEEE Access*, vol. 4, pp. 10029–10041, 2016.

[7] W. Wang, Y. Liu, Z. Luo, T. Jiang, Q. Zhang, and A. Nallanathan, "Toward cross-layer design for non-orthogonal multiple access: A quality-of-experience perspective," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 118–124, Apr. 2018.

[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] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.

[10] 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 Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov. 2016.

[11] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[12] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[13] J. Cui, Z. Ding, and P. Fan, "A novel power allocation scheme under outage constraints in NOMA systems," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1226–1230, Sep. 2016.

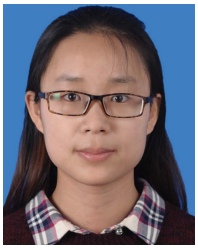
[14] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[15] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, Dec. 2016.

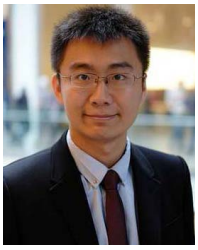
[16] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[17] 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 Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[18] 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.
[19] S. Singh, O. Oyman, A. Papathanassiou, D. Chatterjee, and J. G. Andrews, "Video capacity and QoE enhancements over LTE," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 7071–7076.
[20] M. Rugelj, M. Volk, U. Sedlar, J. Sterle, and A. Kos, "A novel user satisfaction prediction model for future network provisioning," *Telecommun. Syst.*, vol. 56, no. 3, pp. 417–425, 2014.
[21] C. Sacchi, F. Granelli, and C. Schlegel, "A QoE-oriented strategy for OFDMA radio resource allocation based on min-MOS maximization," *IEEE Commun. Lett.*, vol. 15, no. 5, pp. 494–496, May 2011.
[22] Y. H. Cho, H. Kim, S.-H. Lee, and H. S. Lee, "A QoE-aware proportional fair resource allocation for multi-cell OFDMA networks," *IEEE Commun. Lett.*, vol. 19, no. 1, pp. 82–85, Jan. 2015.
[23] D. Wu, Q. Wu, Y. Xu, J. Jing, and Z. Qin, "QoE-based distributed multichannel allocation in 5G heterogeneous cellular networks: A matching-coalitional game solution," *IEEE Access*, vol. 5, pp. 61–71, 2017.
[24] J. Zheng, Y. Cai, Y. Liu, Y. Xu, B. Duan, and X. Shen, "Optimal power allocation and user scheduling in multicell networks: Base station cooperation using a game-theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6928–6942, Dec. 2014.
[25] N. Zhang, S. Zhang, J. Zheng, X. Fang, J. W. Mark, and X. Shen, "QoE driven decentralized spectrum sharing in 5G networks: Potential game approach," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7797–7808, Sep. 2017.
[26] Z. Du, Q. Wu, P. Yang, Y. Xu, J. Wang, and Y.-D. Yao, "Exploiting user demand diversity in heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4142–4155, Aug. 2015.
[27] M. Rugelj, U. Sedlar, M. Volk, J. Sterle, M. Hajdinjak, and A. Kos, "Novel cross-layer QoE-aware radio resource allocation algorithms in multiuser OFDMA systems," *IEEE Trans. Commun.*, vol. 62, no. 9, pp. 3196–3208, Sep. 2014.
[28] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, P. Mogensen, and J. M. Lopez-Soler, "QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems," *Comput. Commun.*, vol. 33, no. 5, pp. 571–582, 2010.
[29] *Feasibility Study for Further Advancements for E-UTRA (LTE Advanced) Version 11.0.0*, document TR 36.912, Sophia-Antipolis, France, Sep. 2012.
[30] S. Tomida and K. Higuchi, "Non-orthogonal access with SIC in cellular downlink for user fairness enhancement," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Dec. 2011, pp. 1–6.
[31] R. Horst and H. Tuy, *Global Optimization: Deterministic Approaches*. Berlin, Germany: Springer, 2013.
[32] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[33] P. C. Weeraddana, M. Codreanu, M. Latva-Aho, and A. Ephremides, "Weighted sum-rate maximization for a set of interfering links via branch and bound," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3977–3996, Aug. 2011.
[34] C. Ng and D. S. Hirschberg, "Three-dimensional stable matching problems," *SIAM J. Discrete Math.*, vol. 4, no. 2, pp. 245–252, 1991.
[35] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Proc. Int. Symp. Algorithmic Game Theory*. Berlin, Germany: Springer, 2011, pp. 117–129.
[36] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1502–1517, Mar. 2018.
[37] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[38] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx
[39] W. C. Li, T. H. Chang, and C. Y. Chi, "Multicell coordinated beamforming with rate outage constraint—Part II: Efficient approximation algorithms," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2763–2778, Jun. 2015.
[40] M. Molina, P. Castelli, and G. Foddis, "Web traffic modeling exploiting TCP connections' temporal clustering through HTML-REDUCE," *IEEE Netw.*, vol. 14, no. 3, pp. 46–55, May/Jun. 2000.
[41] 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.
[42] R. Jain, D.-M. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Eastern Res. Lab., Digit. Equip. Corp., Hudson, MA, USA, Tech. Rep. DEC-TR-301, 1984, vol. 38.

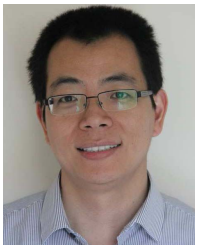


Jingjing Cui (S'14) received the B.S. degree in communication engineering from Tibet University, Lhasa, China, in 2012, and the Ph.D. degree in information and communications engineering from Southwest Jiaotong University, Chengdu, China, in 2018. She was a visiting Ph.D. student with the School of Computing and Communications, Lancaster University, U.K., from 2016 to 2017. She is currently a Research Assistant with the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. Her research interests include non-orthogonal multiple access, artificial intelligence for 5G networks, and convex optimization.



Yuanwei Liu (S'13–M'16) received the B.S. and M.S. degrees from the Beijing University of Posts and Telecommunications in 2011 and 2014, respectively, and the Ph.D. degree in electrical engineering from the Queen Mary University of London, U.K., in 2016. He was with the Department of Informatics, King's College London, from 2016 to 2017, where he was a Post-Doctoral Research Fellow. He has been a Lecturer (Assistant Professor) with the School of Electronic Engineering and Computer Science, Queen Mary University of London, since 2017.

His research interests include 5G wireless networks, Internet of Things, machine learning, stochastic geometry, and matching theory. He has served as a TPC member for many IEEE conferences, such as GLOBECOM and ICC. He received the Exemplary Reviewer Certificate of the IEEE WIRELESS COMMUNICATION LETTERS in 2015 and the IEEE TRANSACTIONS ON COMMUNICATIONS in 2016 and 2017. He currently serves as an Editor for the IEEE COMMUNICATIONS LETTERS and the IEEE ACCESS. He is also a Guest Editor of the IEEE JSTSP Special Issue on Signal Processing Advances for Non-Orthogonal Multiple Access in Next Generation Wireless Networks.



Zhiguo Ding (S'03–M'05–SM'15) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications in 2000, and the Ph.D. degree in electrical engineering from Imperial College London in 2005. From 2005 to 2018, he was with Queen's University Belfast, Imperial College London, Newcastle University, and Lancaster University. From 2012 to 2018, he was an Academic Visitor with Princeton University. Since 2018, he has been with The University of Manchester as a Professor of communications.

His research interests include 5G networks, game theory, cooperative and energy harvesting networks, and statistical signal processing. He received the Best Paper Award at IET ICWMC 2009 and the IEEE WCSP 2014, the EU Marie Curie Fellowship from 2012 to 2014, the Top IEEE TVT Editor 2017, the IEEE Heinrich Hertz Award 2018, and the IEEE Jack Neubauer Memorial Award 2018. He was an Editor for the IEEE WIRELESS COMMUNICATION LETTERS and the IEEE COMMUNICATION LETTERS from 2013 to 2016. He is serving as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the *Journal of Wireless Communications and Mobile Computing*.

Pingzhi Fan (M'93–SM'99–F'15) received the M.Sc. degree in computer science from the Southwest Jiaotong University, China, in 1987, and the Ph.D. degree in electronic engineering from the University of Hull, U.K., in 1994. He has been a Visiting Professor with the University of Leeds, U.K., since 1997, and a Guest Professor with Shanghai Jiao Tong University since 1999. He is currently a Professor and the Director of the Institute of Mobile Communications, Southwest Jiaotong University. He has over 290 research papers published in various international journals, and eight books (including edited), and is the inventor of 22 granted patents. His research interests include vehicular communications, wireless networks for big data, and signal design and coding. He is a fellow of IET, CIE, and CIC. He was a recipient of the U.K. ORS Award (1992), the NSFC Outstanding Young Scientist Award (1998), and the IEEE VTS Jack Neubauer Memorial Award (2018). He served as the general chair or TPC chair of a number of international conferences. He is the Founding Chair of the IEEE VTS BJ Chapter and the IEEE ComSoc CD Chapter and the Founding Chair of the IEEE Chengdu Section. He is a guest editor or an editorial member of several international journals. He also served as a Board Member of the IEEE Region 10, IET (IEE) Council, and IET Asia–Pacific Region. He is an IEEE VTS Distinguished Lecturer from 2015 to 2019.

Arumugam Nallanathan (S'97–M'00–SM'05–F'17) was an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, from 2000 to 2007. He was with the Department of Informatics, King's College London from 2007 to 2017, where he was a Professor of wireless communications from 2013 to 2017 and has been a Visiting Professor since 2017. He has been a Professor of wireless communications and the Head of the Communication Systems Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, since 2017. He published nearly 400 technical papers in scientific journals and international conferences. His research interests include 5G wireless networks, Internet of Things, and molecular communications. He was a co-recipient of the Best Paper Awards presented at the IEEE International Conference on Communications 2016, the IEEE Global Communications Conference 2017, and the IEEE Vehicular Technology Conference 2018. He is an IEEE Distinguished Lecturer. He was selected as a Web of Science Highly Cited Researcher in 2016.

Dr. Nallanathan received the IEEE Communications Society SPCE Outstanding Service Award 2012 and the IEEE Communications Society RCC Outstanding Service Award 2014. He served as the Chair for the Signal Processing and Communication Electronics Technical Committee of the IEEE Communications Society and the technical program chair and a member of technical program committees in numerous IEEE conferences. He was an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2006 to 2011, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2006 to 2017, the IEEE WIRELESS COMMUNICATIONS LETTERS, and the IEEE SIGNAL PROCESSING LETTERS. He is an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS.