# Optimal Power Allocation in Downlink Multicarrier NOMA Systems: Theory and Fast Algorithms

Sepehr Rezvani, *Student Member, IEEE*, Eduard A. Jorswieck, *Fellow, IEEE*, Roghayeh Joda, *Senior Member, IEEE*, and Halim Yanikomeroglu, *Fellow, IEEE*

***Abstract*— In this work, we propose globally optimal power allocation strategies to maximize the users sum-rate (SR), and system energy efficiency (EE) in the downlink of single-cell multicarrier non-orthogonal multiple access (MC-NOMA) systems. Each NOMA cluster includes a set of users in which the well-known superposition coding (SC) combined with successive interference cancellation (SIC) technique is applied among them. By obtaining the closed-form expression of intra-cluster power allocation, we show that MC-NOMA can be equivalently transformed to a virtual orthogonal multiple access (OMA) system, where the effective channel gain of these virtual OMA users is obtained in closed-form. Then, the SR and EE maximization problems are solved by using very fast water-filling and Dinkelbach algorithms, respectively. The equivalent transformation of MC-NOMA to the virtual OMA system brings new theoretical insights, which are discussed throughout the paper. The extensions of our analyses to other scenarios, such as considering users rate fairness, admission control, long-term performance, and a number of future next-generation multiple access (NGMA) schemes enabling recent advanced technologies, e.g., reconfigurable intelligent surfaces are discussed. Extensive numerical results are provided to demonstrate the performance gaps among single-carrier NOMA (SC-NOMA), OMA-NOMA, and OMA.**

***Index Terms*— Broadcast channel, NGMA, superposition coding, successive interference cancellation, multicarrier, NOMA, power allocation, water-filling, energy efficiency.**

## I. Introduction

### A. Evolution of NOMA: From Fully SC-SIC to Hybrid-NOMA

THE rapidly growing demands for high data rate services along with energy constrained networks necessitate the characterization and analysis of the next-generation multiple access (NGMA) techniques in wireless communication systems. It is proved that the capacity region of degraded single-input single-output (SISO) Gaussian broadcast channels (BCs) can be achieved by performing linear superposition coding (SC) at the transmitter side combined with coherent multiuser detection algorithms, like successive interference cancellation (SIC) at the receivers side [1]–[5]. The SC can be performed in code or power domain [6]. The SC-SIC technique is also called non-orthogonal multiple access (NOMA) [6]. Based on the adopted SC technique, NOMA can be divided into two main categories, namely code-domain NOMA, and power-domain NOMA [6]–[8]. In our work, we consider power-domain NOMA, and subsequently, the term NOMA is referred to as power-domain NOMA. In addition to the superior spectral efficiency of NOMA compared to orthogonal multiple access (OMA), i.e., frequency division multiple access (FDMA), and time division multiple access (TDMA) [4], [5], academic and industrial research has demonstrated that NOMA can support massive connectivity, which is important for ensuring that the fifth generation (5G) wireless networks can effectively support Internet of Things (IoT) functionalities [9], [10]. The concept of NOMA has been considered in the 3rd generation partnership project (3GPP) long-term evolution advanced (LTE-A) standard, where NOMA is referred to as multiuser superposition transmission (MUST) [11]. NOMA is also introduced on many existing as well as future wireless systems, because of its high compatibility with other communication technologies [9]. For example, a significant number of works addressed the integration of NOMA to simultaneous wireless information and power transfer [9], [12], cognitive radio networks [9], [12], cooperative communications [12]–[14], millimeter wave communications [12], [14], mobile edge computing networks [10], [12], [15], and reconfigurable intelligent surfaces (RISs) [16]–[18]. In [18], it is shown that the capacity region of the multiuser downlink SISO RIS system can be achieved by NOMA with time sharing. To this end, NOMA is a promising candidate solution for the beyond-5G (B5G)/sixth generation (6G) wireless networks [19].

The SIC complexity is cubic in the number of multiplexed users [20]. Another issue is error propagation, which increases with the number of multiplexed users [20]. Hence, single-carrier NOMA (SC-NOMA), where the signal of all the users is multiplexed, is still impractical for a large number of users. In this line, NOMA is introduced on multicarrier systems, called multicarrier NOMA (MC-NOMA), where the

Manuscript received July 15, 2021; revised November 14, 2021; accepted December 16, 2021. Date of publication January 14, 2022; date of current version March 17, 2022.The work of Sepehr Rezvani and Eduard A. Jorswieck was supported in part by the German Research Foundation (DFG) under Grant JO 801/24-1. The work of Halim Yanikomeroglu was supported by the Discovery Grant Program of the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2016-06275. *(Corresponding author: Sepehr Rezvani.)*

Sepehr Rezvani and Eduard A. Jorswieck are with the Department of Information Theory and Communication Systems, Technische Universität Braunschweig, 38106 Braunschweig, Germany (e-mail: rezvani@ifn.ing.tu-bs.de; jorswieck@ifn.ing.tu-bs.de).

Roghayeh Joda is with the Communication Department, ICT Research Institute, Tehran 143995571, Iran, and also with the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: r.joda@itrc.ac.ir).

Halim Yanikomeroglu is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: halim@sce.carleton.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2022.3143237.

Digital Object Identifier 10.1109/JSAC.2022.3143237

users are grouped into multiple clusters each operating in an isolated resource block, and SC-SIC is applied among users within each cluster [7]. Note that the space division multiple access (SDMA) can also be introduced on NOMA, where the clusters are isolated by zero-forcing beamforming [20]. MC-NOMA with disjoint clusters is based on SC-SIC and FDMA/TDMA, where each user occupies only one resource block, thus receives a single symbol. In FDMA-NOMA, no user benefits from the well-known multiplexing gain in the fading channels. To this end, NOMA is introduced on orthogonal frequency division multiple access (OFDMA), called OFDMA-NOMA or Hybrid-NOMA [6], [19]–[22]. Hybrid-NOMA is the general case of MC-NOMA, where each user can occupy more than one subchannel, and SC-SIC is applied to each isolated subchannel. Therefore, all the users can benefit from the multiplexing gain.

### *B. Related Works and Open Problems*

It is well-known that the dynamic resource allocation is necessary in downlink SC/MC-NOMA to achieve a preferable performance, as well as guaranteed quality of services (QoSs) for mission-critical applications [19]. Maximizing users sum-rate (SR) is one of the important objectives of resource allocation optimization, which is widely addressed not only for SC/MC-NOMA, but also for the other multiple access techniques. In downlink SC-NOMA, maximizing users SR leads to the full base station's (BS's) power consumption [23]. The energy consumption is becoming a social and economical issue due to the rapid increase of the data traffic and number of mobile devices [24]. Hence, minimizing the BSs power consumption while guaranteeing users minimum rate demands is another important objective of resource allocation optimization. To strike a balance between users SR and BS's power consumption, maximizing the well-known fractional system energy efficiency (EE) function, defined as $\frac{\text{Receivers Sum-Rate}}{\text{Transmitter's Total Power Consumption}}$, has attracted lots of attention [24], [25]. The EE is measured in bit/Joule, thus measuring the amount of data transmitted per Joule of the consumed transmitter's energy [24]. In the following, we review the related works which addressed resource allocation optimization for maximizing SR/EE in the downlink of single-cell SC/MC-NOMA systems.

*1) SC-NOMA:* In our previous work [23], we derived the closed-form expression of optimal powers to maximize the SR of $M$-user SC-NOMA system with minimum rate demands under the optimal channel-to-noise ratio (CNR)-based decoding order. The work in [26] addresses the problem of simultaneously maximizing users SR and minimizing total power consumption defined as a utility function for SC-NOMA. However, the analysis in [26] is affected by a detection constraint for successful SIC which is not necessary, since SISO Gaussian BCs are degraded. Hence, *the closed-form expression of optimal powers to maximize system EE in SC-NOMA is still an open problem*.

*2) MC-NOMA:* The joint power and subchannel allocation in MC-NOMA is proved to be strongly NP-hard [27]–[29]. In this way, these two problems are decoupled in most of the prior works. For any given set of clusters, the optimal power allocation for SR/EE maximization in MC-NOMA is more challenging compared to SC-NOMA. In MC-NOMA, there exists a competition among multiple clusters to get the cellular power. Actually, the optimal power allocation in MC-NOMA includes two components: 1) *Inter-cluster power allocation*: optimal power allocation among clusters to get the cellular power budget; 2) *Intra-cluster power allocation*: optimal power allocation among multiplexed users to get the clusters power budget. From the optimization perspective, the analysis in [23] is also valid for MC-NOMA with any predefined power budget for each cluster, e.g., the considered models in [30] and [31]. In this case, the intra-cluster power allocation can be equivalently decoupled into multiple SC-NOMA subproblems. There has been some efforts in finding the optimal joint intra- and inter-cluster power allocation, thus globally optimal power allocation, for MC-NOMA to maximize SR/EE [32]. In [32], FDMA-NOMA with 2 users per cluster is considered. The authors first obtain the closed-form expression of optimal intra-cluster power allocation for each 2-order cluster. Then, by substituting these closed-forms to the original problems, the optimal inter-cluster power allocation is obtained in efficient manners for various objectives. *In [32], all the analysis is based on allocating more power to each weaker user to guarantee successful SIC, which is not necessary, due to the degradation of SISO Gaussian BCs [3], [22]. Another concern is the generalization of the special FDMA-NOMA scheme with 2-order clusters to Hybrid-NOMA with arbitrary number of multiplexed users.*

The works on Hybrid-NOMA mainly focus on achieving the maximum multiplexing gain, where each user receives different symbols on the assigned subchannels. It is straightforward to show that Hybrid-NOMA with per-symbol/subchannel minimum rate constraints can be equivalently transformed to FDMA-NOMA, since a user on different assigned subchannels can be viewed as independent users with individual per-subchannel minimum rate demands. The fractional EE maximization problem for downlink FDMA/Hybrid-NOMA with per-symbol minimum rate demands is addressed by [33]–[37]. In this line, the EE maximization problem is solved by using the suboptimal difference-of-convex (DC) approximation method [33], Dinkelbach algorithm with Fmincon optimization software [34], and Dinkelbach algorithm with subgradient method [35]–[37]. Despite the potentials, there are some fundamental questions that are not yet solved in the literature for the SR/EE maximization problems of downlink Hybrid-NOMA with minimum rate constraints as follows:

1) *What are the closed-form of optimal powers for the SR/EE maximization problems?*
2) *Is the equal power allocation strategy a good solution for the SR/EE maximization problems?*
3) *Is there any users rate fairness guarantee in the SR/EE maximization problems?*
4) *When the full cellular power consumption is energy efficient?*
5) *How can we equivalently transform Hybrid-NOMA to a FDMA system?*

The answer of the first question brings new theoretical insights on the impact of minimum rate demands and channel gains on the optimal power coefficients among multiplexed users. Also, by analyzing the heterogeneity of optimal power allocation among multiplexed users/clusters, we can analytically observe which of the equal intra/inter-cluster power allocation strategies are mostly infeasible/near-optimal. The optimality conditions analysis for the SR/EE maximization problem shows us which users get additional rate rather than their individual minimum rate demands, which is important for guaranteeing users rate fairness. If we guarantee that the full power consumption leads to the maximum EE, the EE maximization problem can be reduced to the SR maximization problem, which subsequently decreases the complexity of the solution methods used in [35]–[37]. Finally, transforming a Hybrid-NOMA system with $N$ subchannels each having $K$ users to a FDMA system with $N$ subchannels will reduce the dimension of the SR/EE maximization problems of Hybrid-NOMA. This decreases the complexity of the solution algorithms, e.g., the pure convex solvers used in [35]–[37]. Moreover, Hybrid-NOMA-to-FDMA transformation facilitates the implementation of Hybrid-NOMA, since the optimization algorithms which are already developed for FDMA can be easily adopted to be used for Hybrid-NOMA.

In general, finding the optimal power allocation for SR/EE maximization problem in downlink Hybrid-NOMA with per-user minimum rate demands[1] is more challenging, due to the nonconvexity of minimum rate constraints. The works in [27]–[29], [38]–[43] address the problem of weighted SR/SR maximization for Hybrid-NOMA without guaranteeing users minimum rate demands. In Hybrid-NOMA with per-user minimum rate constraints, [44] proposes a suboptimal power allocation strategy for the EE maximization problem based on the combination of the DC approximation method and Dinkelbach algorithm. Also, a suboptimal penalty function method is proposed in [45]. We show that most of our analyses for Hybrid-NOMA with per-symbol minimum rate demands also hold for Hybrid-NOMA with per-user minimum rate demands by using the fundamental relations between these two schemes.

### C. Our Contributions

In this work, we address the problem of finding optimal power allocation for maximizing SR/EE of the downlink single-cell Hybrid-NOMA system including multiple clusters each having an arbitrary number of multiplexed users. We assume that each user has a predefined minimum rate demand on each assigned subchannel [32]–[37]. Our main contributions are listed as follows:

- We prove that for the three main objective functions as total power minimization, SR maximization and EE maximization, in each cluster, only the cluster-head[2] user deserves additional power while all the other users get power to only maintain their minimal rate demands.[3]
- We obtain the closed-form expression of intra-cluster power allocation within each cluster. We prove that the intra-cluster power allocation is mainly affected by the minimum rate demand of users with lower decoding order leading to high heterogeneity of intra-cluster power allocation. As a result, the equal intra-cluster power allocation will be infeasible in most of the cases. The users exact CNRs merely impact on the intra-cluster power allocation, specifically for high signal-to-interference-plus-noise ratio (SINR) regions.
- The feasible power allocation region of Hybrid-NOMA with per-symbol minimum rate demands is defined as the intersection of closed boxes along with affine maximum cellular power constraint. Then, the optimal value for the power minimization problem is obtained in closed form.
- For the SR/EE maximization problem, we show that Hybrid-NOMA can be transformed to an equivalent virtual FDMA system. Each cluster acts as a virtual OMA user whose effective CNR is obtained in closed form. Moreover, each virtual OMA user requires a minimum power to satisfy its multiplexed users minimum rate demands, which is obtained in closed form.
- A very fast water-filling algorithm is proposed to solve the SR maximization problem in Hybrid-NOMA. The EE maximization problem is solved by using the Dinkelbach algorithm with inner Lagrange dual with subgradient method or barrier algorithm with inner Newton's method. Different from [33]–[37], the closed-form of optimal powers among multiplexed users is applied to further reduce the dimension of the problems, thus reducing the complexity of the iterative algorithms, as well as increase the accuracy of the solutions, which is a win-win strategy.
- We propose a necessary and sufficient condition for the equal inter-cluster power allocation strategy to be optimal. We show that in the high SINR regions, the effective CNR of the virtual OMA users merely impacts on the inter-cluster power allocation showing the low heterogeneity of inter-cluster power allocation.
- We propose a sufficient condition to verify whether the full cellular power consumption is energy efficient or not. When this condition is fulfilled, we guarantee that at the optimal point of the EE maximization problem, the cellular power constraint is active, thus the EE maximization problem can be solved by using our proposed water-filling algorithm.

Our optimality conditions analysis shows that although *usually* more power will be allocated to the weaker user when all the multiplexed users have the same minimum rate demands, there still exists a critical users rate fairness issue in the SR/EE maximization problem. To this end, we propose a new rate fairness scheme for the downlink of Hybrid-NOMA systems which is a mixture of the well-known proportional fairness among cluster-head users, and weighted minimum rate fairness among

[1] Minimum rate constraint for each user over all the assigned subchannels.

[2] The user with the highest decoding order which cancels the signal of all the other multiplexed users.

[3] For the total power minimization problem, the cluster-head users also get power to only maintain their minimal rate demands.

non-cluster-head users. The extensions of our analyses for the pure Hybrid-NOMA system to other more general/complicated scenarios as well as the integration of Hybrid-NOMA to recent advanced technologies, e.g., reconfigurable intelligent surfaces are discussed in the paper. Extensive numerical results are provided to evaluate the performance of SC-NOMA, FDMA-NOMA with different maximum number of multiplexed users, and FDMA in terms of outage probability, minimum BS's power consumption, maximum SR and EE. The performance comparison between FDMA-NOMA and SC-NOMA brings new insights on the suboptimality-level of FDMA-NOMA due to user grouping based on FDMA. In this work, we answer the question *"How much performance gain can be achieved if we increase the order of NOMA clusters, and subsequently, decrease the number of user groups?"* for a wide range of the number of users and their minimum rate demands. The latter knowledge is highly necessary since multiplexing a large number of users would cause high complexity cost at the users' hardware. The complete source code of the simulations including a user guide is available in [46].

### D. Paper Organization

The rest of this paper is organized as follows: The system model is presented in Section II. The globally optimal power allocation strategies are presented in Section III. The possible extensions of our analyses and future research directions are presented in Section IV. The numerical results are presented in Section V. Our concluding remarks are presented in Section VI. The abbreviations used in the paper are summarized in Table I.

TABLE I
ABBREVIATIONS

| Abbreviation | Definition |
|---|---|
| 3GPP | Third generation partnership project |
| 5G | Fifth generation |
| 6G | Sixth generation |
| AWGN | Additive white Gaussian noise |
| B5G | beyond-5G |
| BC | Broadcast channel |
| BS | Base station |
| CNR | Channel-to-noise ratio |
| CSI | Channel state information |
| DC | Difference-of-convex |
| EE | Energy efficiency |
| FDMA | Frequency division multiple access |
| FD-NOMA | FDMA-NOMA |
| IoT | Internet of things |
| IPM | Interior point method |
| KKT | Karush–Kuhn–Tucker |
| LTE-A | Long-term evolution advanced |
| MC-NOMA | Multicarrier non-orthogonal multiple access |
| MUST | Multiuser superposition transmission |
| NGMA | Next-generation multiple access |
| NOMA | Non-orthogonal multiple access |
| OFDMA | Orthogonal frequency division multiple access |
| OMA | Orthogonal multiple access |
| QoS | Quality of service |
| RIS | Reconfigurable intelligent surface |
| SC | Superposition coding |
| SC-NOMA | Single-carrier non-orthogonal multiple access |
| SDMA | Space division multiple access |
| SIC | Successive interference cancellation |
| SINR | Signal-to-interference-plus-noise ratio |
| SISO | Single-input single-output |
| SR | Sum-rate |
| TDMA | Time division multiple access |

## II. HYBRID-NOMA: OFDMA-BASED SC-SIC

### A. Network Model and Achievable Rates

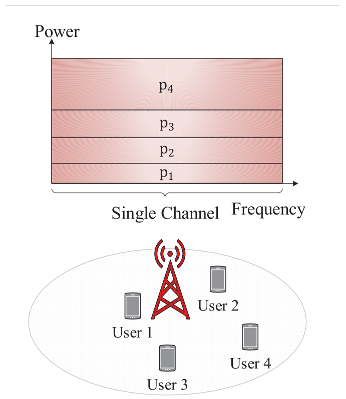
Consider the downlink channel of a multiuser system, where a BS serves $K$ users with limited processing capabilities in a unit time slot of a quasi-static channel. The set of users is denoted by $\mathcal{K} = \{1,\dots,K\}$. In this system, the total bandwidth $W$ (Hz) is equally divided into $N$ isolated subchannels with the set $\mathcal{N} = \{1,\dots,N\}$, where the bandwidth of each subchannel is $W_s = W/N$. NOMA is applied to each subchannel with maximum number of multiplexed users $U^{\max}$. Note that SC-NOMA is infeasible when $U^{\max} < K$. The set of multiplexed users on subchannel $n$ is denoted by $\mathcal{K}_n = \{k \in \mathcal{K} | \rho_k^n = 1\}$, in which $\rho_k^n$ is the binary channel allocation indicator, where if user $k$ occupies subchannel $n$, we set $\rho_k^n = 1$, and otherwise, $\rho_k^n = 0$. The set of subchannels occupied by user $k \in \mathcal{K}$, is indicated by $\mathcal{N}_k = \{n \in \mathcal{N} | \rho_k^n = 1\}$. In FDMA-NOMA, each user belongs to only one cluster [30]–[32], thus we have $\mathcal{K}_n \cap \mathcal{K}_m = \emptyset, \forall n,m \in \mathcal{N},\ n \neq m$, or equivalently, $|\mathcal{N}_k| = 1,\ \forall k \in \mathcal{K}$, where $|.|$ indicates the cardinality of a finite set. In the following, we consider the more general case Hybrid-NOMA with $|\mathcal{N}_k| \geq 1,\ \forall k \in \mathcal{K}$. The maximum number of multiplexed users $U^{\max}$ implies that $|\mathcal{K}_n| \leq U^{\max},\ \forall n \in \mathcal{N}$. The exemplary models of SC-NOMA, FDMA-NOMA, FDMA, Hybrid-NOMA with multiplexing all users in all the subchannels (Hybrid-NOMA with $K$ multiplexed users per cluster), Hybrid-NOMA with $U^{\max}$ multiplexed users per subchannel, and OFDMA are illustrated in Fig. 1.

Each subchannel can be modeled as a SISO Gaussian BC. The transmitted signal by the BS on subchannel $n$ is formulated by $x^n = \sum\limits_{k \in \mathcal{K}_n} \sqrt{p_k^n} s_k^n$, where $s_k^n \sim \mathcal{CN}(0,1)$ and $p_k^n \geq 0$ are the modulated symbol from Gaussian codebooks, and transmit power of user $k \in \mathcal{K}$ on subchannel $n \in \mathcal{N}$, respectively. Obviously, $p_k^n = 0, \forall n \in \mathcal{N},\ k \notin \mathcal{K}_n$. The received signal at user $k$ on subchannel $n$ is

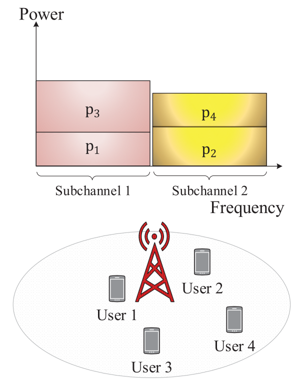
$$y_k^n = \underbrace{\sqrt{p_k^n} g_k^n s_k^n}_{\text{intended signal}} + \underbrace{\sum_{i \in \mathcal{K}_n \setminus \{k\}} \sqrt{p_i^n} g_k^n s_i^n}_{\text{co-channel interference}} + z_k^n, \tag{1}$$

where $g_k^n$ is the (generally complex) channel gain from the BS to user $k$ on subchannel $n$, and $z_k^n \sim \mathcal{CN}(0, \sigma_k^n)$ is the additive white Gaussian noise (AWGN). We assume that the perfect channel state information (CSI) is available at the BS as well as users.

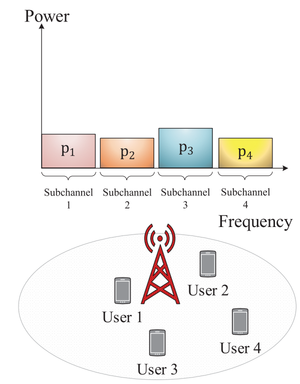
In Hybrid-NOMA, SC-SIC is applied to each multiuser subchannel according to the optimal CNR-based decoding order [1]–[5]. Let $h_k^n = |g_k^n|^2/\sigma_k^n,\ \forall n \in \mathcal{N}, k \in \mathcal{K}_n$. Then, the CNR-based decoding order is indicated by $h_i^n > h_j^n \Rightarrow i \rightarrow j, \forall i,j \in \mathcal{K}_n$, where $i \rightarrow j$ represents that user $i$ fully decodes (and then cancels) the signal of user $j$ before decoding its desired signal on subchannel $n$. Moreover, the signal of user $i$ is fully treated as noise at user $j$ on subchannel $n$.



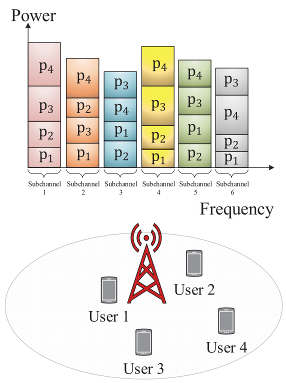

(a) SC-NOMA: Capacity-achieving, thus optimal, when $U^{\max} \geq K$, and infeasible when $U^{\max} < K$. No multiplexing gain is achieved.



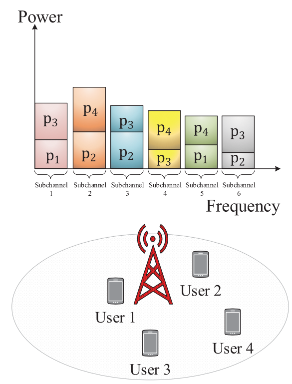

(b) FDMA-NOMA: Suboptimal but feasible. SC-SIC is applied to each cluster. No multiplexing gain is achieved.



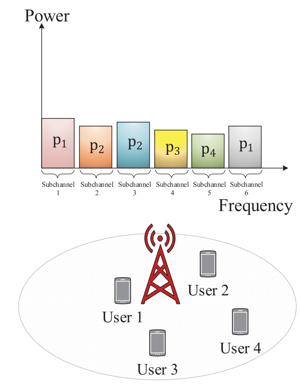

(c) FDMA: Suboptimal but feasible with no SC-SIC. No multiplexing gain is achieved.




(d) Hybrid-NOMA with $K$ multiplexed users: Capacity-achieving, thus optimal, when $U^{\max} \geq K$, and infeasible when $U^{\max} < K$. Multiplexing gain can be achieved.




(e) Hybrid-NOMA with 2 multiplexed users: Suboptimal but feasible. SC-SIC is applied among users within each cluster. Multiplexing gain can be achieved.




(f) OFDMA: Suboptimal but feasible with no SC-SIC. Multiplexing gain can be achieved.

Fig. 1. 1(a)-1(c): Exemplary models of SC-NOMA, FDMA-NOMA, and FDMA, respectively, where a single symbol is transmitted to each user. 1(d)-1(f): Exemplary models of Hybrid-NOMA with $K$ multiplexed users, Hybrid-NOMA with 2 multiplexed users, and OFDMA, respectively, where an independent symbol is transmitted to each user on each assigned subchannel. In these examples, we set $U^{\max} = 2$.

In summary, the SIC protocol in each isolated subchannel is the same as the SIC protocol of SC-NOMA. We call the stronger user $i$ as the user with higher decoding order in the user pair $i, j \in \mathcal{K}_n$. In each subchannel $n$, the index of the cluster-head user is denoted by $\Phi_n = \arg\max\limits_{k \in \mathcal{K}_n} h_k^n$. When $|\mathcal{K}_n| = 1$, the single user can be defined as the cluster-head user on subchannel $n$ since it does not experience any interference. The SINR of each user $i \in \mathcal{K}_n$ for decoding the desired signal of user $k$ on subchannel $n$ is $\gamma_{k,i}^n = \frac{p_k^n h_i^n}{\sum\limits_{\substack{j \in \mathcal{K}_n, \\ h_j^n > h_k^n}} p_j^n h_i^n + 1}$ [3]. User $i \in \mathcal{K}_n$ is able to fully decode the signal of user $k$ if and only if $\gamma_{k,i}^n \geq \gamma_{k,k}^n$, where $\gamma_{k,k}^n \equiv \gamma_k^n = \frac{p_k^n h_k^n}{\sum\limits_{\substack{j \in \mathcal{K}_n, \\ h_j^n > h_k^n}} p_j^n h_k^n + 1}$ is the SINR of user $k$ for decoding its own signal $s_k^n$. According to the Shannon's capacity formula, the achievable rate (in bps) of user $k \in \mathcal{K}_n$ on subchannel $n \in \mathcal{N}$ after successful SIC is given by [2], [3], [23]

$$R_k^n(\boldsymbol{p}^n) = \min_{\substack{i \in \mathcal{K}_n \\ h_i^n \geq h_k^n}} \left\{ W_s \log_2 \left(1 + \gamma_{k,i}^n(\boldsymbol{p}^n)\right) \right\},$$

where $\boldsymbol{p}^n = [p_k^n]_{1 \times K}$, is the vector of allocated powers to all the users on subchannel $n$. The matrix of power allocation among all the users and subchannels is denoted by $\boldsymbol{p} = [p_k^n]_{N \times K}$. Therefore, $\boldsymbol{p}^n$ is the $n$-th row of matrix $\boldsymbol{p}$. For the user pair $i, j \in \mathcal{K}_n$ with $h_i^n > h_k^n$, the condition $\gamma_{k,i}^n(\boldsymbol{p}^n) \geq \gamma_k^n(\boldsymbol{p}^n)$ or equivalently $p_k^n h_i^n \geq p_k^n h_k^n$ holds independent of $\boldsymbol{p}^n$. Accordingly, for any $\boldsymbol{p}^n$, the achievable rate of each user $k \in \mathcal{K}_n$ on subchannel $n$ is equal to its channel capacity formulated by [3]

$$R_k^n(\boldsymbol{p}^n) = W_s \log_2 \left(1 + \gamma_k^n(\boldsymbol{p}^n)\right). \tag{2}$$

The overall achievable rate of user $k \in \mathcal{K}$ can thus be obtained by $R_k(\boldsymbol{p}) = \sum\limits_{n \in \mathcal{N}_k} R_k^n(\boldsymbol{p}^n)$.

TABLE II
MAIN NOTATIONS

| Notation | Description |
|---|---|
| $\mathcal{K}$ | Set of all the users |
| $\mathcal{K}_n$ | Set of users on subchannel $n$ |
| $\mathcal{N}$ | Set of subchannels |
| $\mathcal{N}_k$ | Set of subchannels occupied by user $k$ |
| $W_s$ | Bandwidth of each subchannel |
| $U^{\max}$ | Maximum number of multiplexed users |
| $\rho_k^n$ | Channel allocation indicator for user $k$ and subchannel $n$ |
| $p_k^n$ | Allocated power to user $k$ on subchannel $n$ |
| $h_k^n$ | CNR of user $k$ on subchannel $n$ |
| $\Phi_n$ | Index of the cluster-head user on subchannel $n$ |
| $R_k^n$ | Achievable rate of user $k$ on subchannel $n$ |
| $\boldsymbol{p}^n$ | Vector of allocated powers on subchannel $n$ |
| $\boldsymbol{p}$ | Matrix of power allocation |
| $R_{k,n}^{\min}$ | Minimum rate demand of user $k$ on subchannel $n$ |
| $P^{\max}$ | Maximum transmit power of the BS |
| $P_n^{\text{mask}}$ | Maximum allowable power of subchannel $n$ |
| $P_{\text{C}}$ | BS's circuit power consumption |
| $E(\boldsymbol{p})$ | System EE |
| $q_n$ | Power consumption of cluster $n$ |
| $Q_n^{\min}$ | Lower-bound of $q_n$ |
| $H_n$ | Effective CNR of virtual OMA user $n$ |
| $P_{\text{EE}}$ | Power consumption of the BS |
| $P_{\min}$ | Lower-bound of $P_{\text{EE}}$ |
| $\lambda$ | Fractional parameter |

### B. Optimization Problem Formulations

Assume that the set of clusters, i.e., $\mathcal{K}_n,\ \forall n \in \mathcal{N}$, is predefined. The general power allocation problem for maximizing users SR in Hybrid-NOMA is formulated by

$$\max_{\boldsymbol{p}\geq 0} \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n) \tag{3a}$$

$$\text{s.t. } R_k^n(\boldsymbol{p}^n) \geq R_{k,n}^{\min}, \quad \forall k\in\mathcal{K},\ n\in\mathcal{N}_k, \tag{3b}$$

$$\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n} p_k^n \leq P^{\max}, \tag{3c}$$

$$\sum_{k\in\mathcal{K}_n} p_k^n \leq P_n^{\text{mask}}, \quad \forall n\in\mathcal{N}, \tag{3d}$$

where (3b) is the per-subchannel minimum rate constraint, in which $R_{k,n}^{\min}$ is the individual minimum rate demand of user $k$ on subchannel $n$ [32]–[37]. (3c) is the cellular power constraint, where $P^{\max}$ denotes the maximum available power of the BS. (3d) is the maximum per-subchannel power constraint, where $P_n^{\text{mask}}$ denotes the maximum allowable power on subchannel[4] $n$.

The overall system EE is formulated by $E(\boldsymbol{p}) = \frac{\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n)}{\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n} p_k^n + P_{\text{C}}}$, where constant $P_{\text{C}}$ is the circuit power consumption [24], [25]. The power allocation problem for maximizing system EE under the individual minimum rate demand of users in Hybrid-NOMA is formulated by

$$\max_{\boldsymbol{p}\geq 0}\ E(\boldsymbol{p}) \quad \text{s.t. (3b)-(3d).} \tag{4}$$

The main notations of the paper are summarized in Table II.

[4]We do not impose any specific condition on $P_n^{\text{mask}}$. We only take into account $P_n^{\text{mask}}$ in our analysis to keep the generality, such that $P_n^{\text{mask}} \geq P^{\max},\ \forall n\in\mathcal{N}$, as special case.

## III. Solution Algorithms

In this section, we propose globally optimal power allocation algorithms for the SR and EE maximization problems. The closed-form of optimal powers for the total power minimization problem is also derived to characterize the feasible set of our target problems.

### A. Sum-Rate Maximization Problem

Here, we propose a water-filling algorithm to find the globally optimal solution of (3). The SR of users in each cluster, i.e., $\sum_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n)$ is strictly concave in $\boldsymbol{p}^n$, since its Hessian is negative definite [47]. For more details, please see Appendix A in [48]. The overall SR in (3a) is thus strictly concave in $\boldsymbol{p}$, since it is the positive summation of strictly concave functions. Besides, the power constraints in (3c) and (3d) are affine, so are convex. The minimum rate constraint in (3b) can be equivalently transformed to the following affine form as $2^{(R_{k,n}^{\min}/W_s)}\Bigg(\sum_{\substack{j\in\mathcal{K}_n,\\ h_j^n>h_k^n}} p_j^n h_k^n + 1\Bigg) \leq \sum_{\substack{j\in\mathcal{K}_n,\\ h_j^n>h_k^n}} p_j^n h_k^n + 1 + p_k^n h_k^n,\ \forall k\in\mathcal{K},\ n\in\mathcal{N}_k$. Accordingly, the feasible set of (3) is convex. Summing up, problem (3) is convex in $\boldsymbol{p}$. Let us define $q_n = \sum_{k\in\mathcal{K}_n} p_k^n$ as the power consumption of cluster $n$. Problem (3) can be equivalently transformed to the following joint intra- and inter-cluster power allocation problem as

$$\max_{\boldsymbol{p}\geq 0,\boldsymbol{q}\geq 0} \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n) \tag{5a}$$

$$\text{s.t. } R_k^n(\boldsymbol{p}^n) \geq R_{k,n}^{\min},\ \forall k\in\mathcal{K},\ n\in\mathcal{N}_k, \tag{5b}$$

$$\sum_{n\in\mathcal{N}} q_n \leq P^{\max}, \tag{5c}$$

$$\sum_{k\in\mathcal{K}_n} p_k^n = q_n, \quad \forall n\in\mathcal{N}, \tag{5d}$$

$$0 \leq q_n \leq P_n^{\text{mask}}, \quad \forall n\in\mathcal{N}, \tag{5e}$$

where $\boldsymbol{q} = [q_n], \forall n\in\mathcal{N}$. In the following, we first convert the feasible set of (5) to the intersection of closed-boxes along with the affine cellular power constraint.

*Proposition 1: The feasible set of* (5) *is the intersection of* $q_n \in \left[Q_n^{\min}, P_n^{\text{mask}}\right],\ \forall n\in\mathcal{N}$*, and cellular power constraint* $\sum_{n\in\mathcal{N}} q_n \leq P^{\max}$*, where the lower-bound constant* $Q_n^{\min}$ *is*

$$Q_n^{\min} = \sum_{k\in\mathcal{K}_n} \beta_k^n \Bigg( \prod_{\substack{j\in\mathcal{K}_n\\ h_j^n>h_k^n}} \left(1+\beta_j^n\right) + \frac{1}{h_k^n} + \sum_{\substack{j\in\mathcal{K}_n\\ h_j^n>h_k^n}} \frac{\beta_j^n \prod_{\substack{l\in\mathcal{K}_n\\ h_k^n<h_l^n<h_j^n}} (1+\beta_l^n)}{h_j^n} \Bigg), \tag{6}$$

*in which* $\beta_k^n = 2^{(R_{k,n}^{\min}/W_s)} - 1,\ \forall n\in\mathcal{N},\ k\in\mathcal{K}_n$.

*Proof:* Please see Appendix A. □

The feasibility of problems (3) and (4) can be immediately determined as follows:

*Corollary 1: Problems* (3) *and* (4) *are feasible if and only if* $Q_n^{\min} \leq P_n^{\text{mask}},\ \forall n \in \mathcal{N}$, *and* $\sum_{n \in \mathcal{N}} Q_n^{\min} \leq P^{\max}$.

In the following, we find the closed-form of optimal intra-cluster power allocation as a linear function of any given feasible $\boldsymbol{q}$, thus satisfying Proposition 1.

*Proposition 2: For any given feasible* $\boldsymbol{q}$, *the optimal intra-cluster powers for each cluster* $n \in \mathcal{N}$ *can be obtained by*

$$p_k^{n*} = \Bigg( \beta_k^n \prod_{\substack{j \in \mathcal{K}_n \\ h_j^n < h_k^n}} \left(1 - \beta_j^n\right) \Bigg) q_n + c_k^n, \quad \forall k \in \mathcal{K}_n \setminus \{\Phi_n\}, \tag{7}$$

*and*

$$p_{\Phi_n}^{n}{}^{*} = \Bigg( 1 - \sum_{\substack{i \in \mathcal{K}_n \\ h_i^n < h_{\Phi_n}^n}} \beta_i^n \prod_{\substack{j \in \mathcal{K}_n \\ h_j^n < h_i^n}} \left(1 - \beta_j^n\right) \Bigg) q_n - \sum_{\substack{i \in \mathcal{K}_n \\ h_i^n < h_{\Phi_n}^n}} c_i^n, \tag{8}$$

*where* $\beta_k^n = \frac{2^{(R_{k,n}^{\min}/W_s)} - 1}{2^{(R_{k,n}^{\min}/W_s)}},\ \forall n \in \mathcal{N},\ k \in \mathcal{K}_n$, *and* $c_k^n = \beta_k^n \left( \frac{1}{h_k^n} - \sum_{\substack{j \in \mathcal{K}_n \\ h_j^n < h_k^n}} \frac{\prod_{\substack{l \in \mathcal{K}_n \\ h_j^n < h_l^n < h_k^n}} (1-\beta_l^n)\beta_j^n}{h_j^n} \right),\ \forall n \in \mathcal{N},\ k \in \mathcal{K}_n.$

*Proof:* Please see Appendix B. □

Since the closed-form expressions of optimal intra-cluster power allocation in Proposition 2 are valid for any given feasible $\boldsymbol{q}$, we can substitute (7) and (8) directly to problem (5). For convenience, we first rewrite (8) as

$$p_{\Phi_n}^{n}{}^{*} = \alpha_n q_n - c_n, \quad \forall n \in \mathcal{N}, \tag{9}$$

where $\alpha_n = \Bigg( 1 - \sum_{\substack{i \in \mathcal{K}_n \\ h_i^n < h_{\Phi_n}^n}} \beta_i^n \prod_{\substack{j \in \mathcal{K}_n \\ h_j^n < h_i^n}} \left(1 - \beta_j^n\right) \Bigg),\ \forall n \in \mathcal{N}$, and $c_n = \sum_{\substack{i \in \mathcal{K}_n \\ h_i^n < h_{\Phi_n}^n}} c_i^n,\ \forall n \in \mathcal{N}$, are nonnegative constants. According to the proof of Proposition 2 and (9), the SR function of each cluster $n \in \mathcal{N}$ at the optimal point can be formulated as a function of $q_n$ given by

$$\begin{aligned} R_{\text{opt}}^n(q_n) &= \sum_{k \in \mathcal{K}_n} R_k^n(\boldsymbol{p}^{*n}) = \sum_{\substack{k \in \mathcal{K}_n \\ k \neq \Phi_n}} (R_{k,n}^{\min}) + R_{\Phi_n}^n(q_n) \\ &= \sum_{\substack{k \in \mathcal{K}_n \\ k \neq \Phi_n}} (R_{k,n}^{\min}) + W_s \log_2\left(1 + \left(\alpha_n q_n - c_n\right) h_{\Phi_n}^n\right), \\ &\qquad \forall n \in \mathcal{N}. \end{aligned} \tag{10}$$

By utilizing Proposition 1 and (10), the joint intra- and inter-cluster power allocation problem (5) can be equivalently transformed to the following inter-cluster power allocation problem

$$\max_{\boldsymbol{q}} \sum_{n \in \mathcal{N}} W_s \log_2\left(1 + \left(\alpha_n q_n - c_n\right) h_{\Phi_n}^n\right) \tag{11a}$$

$$\text{s.t.} \sum_{n \in \mathcal{N}} q_n = P^{\max}, \tag{11b}$$

$$q_n \in [Q_n^{\min}, P_n^{\text{mask}}], \quad \forall n \in \mathcal{N}. \tag{11c}$$

**Algorithm 1** The Bisection Method for Finding $\nu^*$ in (13)

Initialize tolerance $\epsilon$, lower-bound $\nu_l$, upper-bound $\nu_h$, and maximum iteration $L$.
**for** $l = 1 : L$ **do**
  Set $\nu_m = \frac{\nu_l + \nu_h}{2}$.
  **if** $\sum_{n \in \mathcal{N}} \max\left\{\tilde{Q}_n^{\min}, \min\left\{\left(\frac{W_s/(\ln 2)}{\nu_m} - \frac{1}{H_n}\right), \tilde{P}_n^{\text{mask}}\right\}\right\} < \tilde{P}^{\max}$ **then**
    Set $\nu_h = \nu_m$.
  **else**
    Set $\nu_l = \nu_m$.
  **end if**
  **if** $\frac{\tilde{P}^{\max} - \sum_{n \in \mathcal{N}} \max\left\{\tilde{Q}_n^{\min}, \min\left\{\left(\frac{W_s/(\ln 2)}{\nu_m} - \frac{1}{H_n}\right), \tilde{P}_n^{\text{mask}}\right\}\right\}}{\tilde{P}^{\max}} \leq \epsilon$ **then**
    **break**.
  **end if**
**end for**

Let us define $\tilde{\boldsymbol{q}} = [\tilde{q}_n], \forall n \in \mathcal{N}$, where $\tilde{q}_n = q_n - \frac{c_n}{\alpha_n},\ \forall n \in \mathcal{N}$. Hence, (11) can be transformed to the following equivalent FDMA problem as

$$\max_{\tilde{\boldsymbol{q}}} \sum_{n \in \mathcal{N}} W_s \log_2\left(1 + \tilde{q}_n H_n\right) \tag{12a}$$

$$\text{s.t.} \sum_{n \in \mathcal{N}} \tilde{q}_n = \tilde{P}^{\max}, \tag{12b}$$

$$\tilde{q}_n \in [\tilde{Q}_n^{\min}, \tilde{P}_n^{\text{mask}}], \quad \forall n \in \mathcal{N}, \tag{12c}$$

where $H_n = \alpha_n h_{\Phi_n}^n,\ \forall n \in \mathcal{N}$, $\tilde{P}^{\max} = P^{\max} - \sum_{n \in \mathcal{N}} \frac{c_n}{\alpha_n}$, $\tilde{Q}_n^{\min} = Q_n^{\min} - \frac{c_n}{\alpha_n},\ \forall n \in \mathcal{N}$, and $\tilde{P}_n^{\text{mask}} = P_n^{\text{mask}} - \frac{c_n}{\alpha_n},\ \forall n \in \mathcal{N}$. Constraint (12b) is the affine cellular power constraint, and (12c) is derived based on Proposition 1. The objective function (12a) is strictly concave in $\tilde{\boldsymbol{q}}$, and the feasible set of (12) is affine, so is convex. Accordingly, problem (12) is convex. The equivalent FDMA problem (12) can be optimally solved by using the well-known water-filling algorithm [49]–[53]. After some mathematical manipulations, the optimal $\tilde{q}_n^*$ can be obtained as

$$\tilde{q}_n^* = \begin{cases} \frac{W_s/(\ln 2)}{\nu^*} - \frac{1}{H_n}, & \left(\frac{W_s/(\ln 2)}{\nu^*} - \frac{1}{H_n}\right) \in [\tilde{Q}_n^{\min}, \tilde{P}_n^{\text{mask}}], \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

such that $\tilde{\boldsymbol{q}}^* = [\tilde{q}_n^*],\ \forall n \in \mathcal{N}$, satisfies (12b). Moreover, $\nu^*$ is the dual optimal corresponding to constraint (12b). For more details, please see Appendix C. The pseudo-code of the bisection method for finding $\nu^*$ is presented in Alg. 1. After finding $\tilde{\boldsymbol{q}}^*$, we obtain $\boldsymbol{q}^*$ by using $q_n^* = (\tilde{q}_n^* + \frac{c_n}{\alpha_n})$, $\forall n \in \mathcal{N}$. Then, we find the optimal intra-cluster power allocation according to Proposition 2. Since problems (3) and (12) are equivalent, the obtained globally optimal solution for (12) is also globally optimal for (3).

### B. Energy Efficiency Maximization Problem

In this subsection, we find a globally optimal solution for problem (4). The feasible region of problem (4) is identical to the feasible region of problem (3). Hence, Proposition 1 can be used to characterize the feasible region of problem (4). Let us define $P_{\text{EE}} = \sum\limits_{n\in\mathcal{N}} q_n$ as the cellular power consumption in the EE maximization problem (4). For any given $P_{\text{EE}}$, problem (4) can be equivalently transformed to the SR maximization problem (3) in which $P^{\max} = P_{\text{EE}}$. As a result, the globally optimal solution of (4) can be obtained by exploring different values of $P_{\text{EE}} \in \left[\sum\limits_{n\in\mathcal{N}} Q_n^{\min}, P^{\max}\right]$, and applying the water-filling Alg. 1, in which $P^{\max} = P_{\text{EE}}$. Exploring $P_{\text{EE}} \in \left[\sum\limits_{n\in\mathcal{N}} Q_n^{\min}, P^{\max}\right]$ may be computationally prohibitive, specifically when the stepsize of exhaustive search is small and/or $\sum\limits_{n\in\mathcal{N}} Q_n^{\min} \to 0$, e.g., when the users have small minimum rate demands.

The SR function in the numerator of the EE function in (4) is strictly concave in $\boldsymbol{p}$. The denominator of the EE function is an affine function, so is convex. Therefore, problem (4) is a concave-convex fractional program with a pseudoconcave objective function [24], [25]. The pseudoconcavity of the objective function in (4) implies that any stationary point is indeed globally optimal and the Karush–Kuhn–Tucker (KKT) optimality conditions are sufficient if a constraint qualification is fulfilled [24], [25]. For more details, please see Appendix D. Hence, the globally optimal solution of (4) can be obtained by using the well-known Dinkelbach algorithm [24], [25]. In this algorithm, we iteratively solve the following problem

$$\begin{aligned} \max_{\boldsymbol{p}\geq 0}\ F(\lambda,\boldsymbol{p}) = & \left(\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n)\right) \\ & - \lambda\left(\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n} p_k^n + P_{\text{C}}\right) \\ \text{s.t. } & \text{(3b)-(3d)}, \end{aligned} \tag{14}$$

where $\lambda \geq 0$ is the fractional parameter, and $F(\lambda,\boldsymbol{p})$ is strictly concave in $\boldsymbol{p}$. This algorithm is described as follows: We first initialize parameter $\lambda_{(0)}$ such that $F\left(\lambda_{(0)},\boldsymbol{p}^*\right) \geq 0$. At each iteration $(t)$, we set $\lambda_{(t)} = E(\boldsymbol{p}^*_{(t-1)})$, where $\boldsymbol{p}^*_{(t-1)}$ is the optimal solution obtained from the prior iteration $(t-1)$. After that, we find $\boldsymbol{p}^*_{(t)}$ by solving (14) in which $\lambda = \lambda_{(t)}$. We repeat the iterations until $|F\left(\lambda_{(t)},\boldsymbol{p}^*_t\right)| \leq \Upsilon$, where $\Upsilon$ is a tolerance tuning the optimality gap. The pseudo-code of the Dinkelbach algorithm for solving (4) is presented in Alg. 2. Similar to the transformation of (3) to (5), we define $q_n = \sum\limits_{k\in\mathcal{K}_n} p_k^n$ as the power consumption of cluster $n$. The main problem (14) can be equivalently transformed to the following joint intra- and inter-cluster power allocation problem as

$$\begin{aligned} \max_{\boldsymbol{p}\geq 0,\boldsymbol{q}\geq 0}\ & \left(\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n)\right) - \lambda\left(\sum_{n\in\mathcal{N}} q_n + P_{\text{C}}\right) \\ \text{s.t. } & \text{(5b)-(5e)}. \end{aligned} \tag{15}$$

**Algorithm 2** The Dinkelbach Method for Solving the Energy Efficiency Maximization Problem

Initialize parameter $\lambda_{(0)}$ satisfying $F(\lambda_{(0)},\boldsymbol{p}^*) \geq 0$, tolerance $\Upsilon$ (sufficiently small), and $t = 0$.
**while** $F(\lambda,\boldsymbol{p}^*) > \Upsilon$ **do**
  Set $\lambda = E(\boldsymbol{p}^*)$. Then, solve (14) and find $\boldsymbol{p}^*$.
  **if** $|F(\lambda,\boldsymbol{p}^*)| \leq \Upsilon$ **then**
    **break**.
  **end if**
**end while**

The feasible set of problems (5) and (15) is identical, thus the feasibility of (15) can be characterized by Proposition 1.

*Proposition 3: For any given feasible $\boldsymbol{q}$, the optimal intra-cluster power allocation in problem* (15) *can be obtained by using* (7) *and* (8)*.*

*Proof:* When $\boldsymbol{q}$ is fixed, the second term $\lambda\left(\sum\limits_{n\in\mathcal{N}} q_n + P_{\text{C}}\right)$ in (15) is a constant. Hence, the objective function of (15) can be equivalently rewritten as maximizing users SR given by $\max\limits_{\boldsymbol{p}\geq 0}\left(\sum\limits_{n\in\mathcal{N}}\sum\limits_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n)\right)$, which is independent of $\lambda$. Hence, for any given feasible $\boldsymbol{q}$, problems (15) and (5) are identical. Accordingly, Proposition 2 also holds for any given feasible $\boldsymbol{q}$, and $\lambda$ in (15). □

Similar to the SR maximization problem (5), we substitute (7) and (8) to problem (15). By utilizing Proposition 1 and (10), the joint intra- and inter-cluster power allocation problem (15) can be equivalently transformed to the following inter-cluster power allocation problem

$$\begin{aligned} \max_{\boldsymbol{q}}\ \hat{F}(\boldsymbol{q}) = & \left(\sum_{n\in\mathcal{N}} W_s \log_2\left(1 + (\alpha_n q_n - c_n)\, h^n_{\Phi_n}\right)\right) \\ & - \lambda\left(\sum_{n\in\mathcal{N}} q_n\right) \end{aligned} \tag{16a}$$

$$\text{s.t. } \sum_{n\in\mathcal{N}} q_n \leq P^{\max}, \quad q_n \in [Q_n^{\min}, P_n^{\text{mask}}], \quad \forall n \in \mathcal{N}, \tag{16b}$$

where $\alpha_n$ and $c_n$ are defined in (9). Note that since $\lambda$ and $P_{\text{C}}$ are constants, the term $-\lambda P_{\text{C}}$ can be removed from (15), so is removed in (16a) during the equivalent transformation. The differences between problems (11) and (16) are the additional term $-\lambda\left(\sum\limits_{n\in\mathcal{N}} q_n\right)$ in $\hat{F}(\boldsymbol{q})$, and also inequality constraint (16b).

*Proposition 4: At the optimal point of the EE maximization problem* (14)*, if*

$$\frac{W_s/(\ln 2)}{\lambda} - \frac{1 - c_n h^n_{\Phi_n}}{\alpha_n h^n_{\Phi_n}} > P_n^{\text{mask}}, \quad \forall n \in \mathcal{N}, \tag{17}$$

*the cellular power constraint* (3c) *is active, meaning that* $\sum\limits_{n\in\mathcal{N}} q_n^* = P^{\max}$.

*Proof:* The optimal solution of (16) is unique if and only if the objective function (16a) is strictly concave. For the case that the concave function in (16a) is increasing in $\boldsymbol{q}$, we can guarantee that at the optimal point, the cellular power constraint (16b) is active. In other words, for the case that $\frac{\partial \hat{F}(\boldsymbol{q})}{\partial q_n} > 0,\ \forall n \in \mathcal{N}$, for any $q_n \in [Q_n^{\min}, P_n^{\text{mask}}],\ \forall n \in \mathcal{N}$, the optimal $\boldsymbol{q}^*$ satisfies $\sum\limits_{n\in\mathcal{N}} q_n^* = P^{\max}$. In this case, the cellular power constraint (16b) can be replaced with $\sum\limits_{n\in\mathcal{N}} q_n = P^{\max}$, thus the optimization problem (16) can be equivalently transformed to the SR maximization problem (11) whose globally optimal solution can be obtained via Alg. 1. In the following, we find a sufficient condition, where it is guaranteed that $\frac{\partial \hat{F}(\boldsymbol{q})}{\partial q_n} > 0,\ \forall n \in \mathcal{N}$, for any $q_n \in [Q_n^{\min}, P_n^{\text{mask}}],\ \forall n \in \mathcal{N}$. The condition $\frac{\partial \hat{F}(\boldsymbol{q})}{\partial q_n} > 0,\ \forall n \in \mathcal{N}$, can be rewritten as $\frac{W_s \alpha h^n_{\Phi_n}/(\ln 2)}{1+(\alpha_n q_n - c_n)h^n_{\Phi_n}} - \lambda > 0,\ \forall n \in \mathcal{N}$. After some mathematical manipulations, the latter inequality is rewritten as

$$q_n < \frac{W_s/(\ln 2)}{\lambda} - \frac{1 - c_n h^n_{\Phi_n}}{\alpha_n h^n_{\Phi_n}}, \quad \forall n \in \mathcal{N}. \tag{18}$$

The right-hand side of (18) is a constant providing an upper-bound for the region of $\boldsymbol{q}$ such that $\frac{\partial \hat{F}(\boldsymbol{q})}{\partial q_n} > 0,\ \forall n \in \mathcal{N}$. The inequality in (18) holds for any $q_n \in [Q_n^{\min}, P_n^{\text{mask}}],\ \forall n \in \mathcal{N}$, if and only if $\frac{W_s/(\ln 2)}{\lambda} - \frac{1-c_n h^n_{\Phi_n}}{\alpha_n h^n_{\Phi_n}} > P_n^{\text{mask}},\ \forall n \in \mathcal{N}$, and the proof is completed. □

If (17) holds for the given $\lambda$, we guarantee that $\sum\limits_{n\in\mathcal{N}} q_n^* = P^{\max}$, meaning that the EE problem (16) can be equivalently transformed to the SR maximization problem (11) whose globally optimal solution is obtained by using Alg. 1.

For the case that (17) does not hold, Alg. 1 may be suboptimal for (16). In this case, similar to the transformation of (11) to (12), we define $\tilde{\boldsymbol{q}} = [\tilde{q}_n],\ \forall n \in \mathcal{N}$, where $\tilde{q}_n = q_n - \frac{c_n}{\alpha_n},\ \forall n \in \mathcal{N}$. Problem (16) can thus be rewritten as

$$\max_{\tilde{\boldsymbol{q}}} \sum_{n\in\mathcal{N}} W_s \log_2(1 + \tilde{q}_n H_n) - \lambda \left(\sum_{n\in\mathcal{N}} \tilde{q}_n\right) \tag{19a}$$

$$\text{s.t.} \sum_{n\in\mathcal{N}} \tilde{q}_n \le \tilde{P}^{\max}, \quad q_n \in [\tilde{Q}_n^{\min}, \tilde{P}_n^{\text{mask}}],\ \forall n \in \mathcal{N}, \tag{19b}$$

where $H_n = \alpha_n h^n_{\Phi_n},\ \forall n \in \mathcal{N}$, $\tilde{P}^{\max} = P^{\max} - \sum\limits_{n\in\mathcal{N}} \frac{c_n}{\alpha_n}$, $\tilde{Q}_n^{\min} = Q_n^{\min} - \frac{c_n}{\alpha_n},\ \forall n \in \mathcal{N}$, and $\tilde{P}_n^{\text{mask}} = P_n^{\text{mask}} - \frac{c_n}{\alpha_n},\ \forall n \in \mathcal{N}$. The equivalent FDMA convex problem (19) can be solved by using the Lagrange dual method with subgradient algorithm or interior point methods (IPMs) [47], [54], [55]. The derivations of the subgradient algorithm for solving (19) is provided in Appendix E. Moreover, the derivations of the barrier algorithm with inner Newton's method for solving (19) is provided in Appendix F. According to the above, depending on the value of $\lambda$ at each Dinkelbach iteration, (14) can be solved by using Alg. 1 or subgradient/barrier method. The pseudo-codes of our proposed algorithms for solving (14) in Step 3 of Alg. 2 based on the subgradient and barrier methods are presented in Algs. 3 and 4, respectively. After finding $\boldsymbol{q}^*$ via Algs. 3 or 4, we find $\boldsymbol{p}^*$ by using (7) and (8).

**Algorithm 3** The Mixed Water-Filling/Subgradient Method for Solving Problem (14)

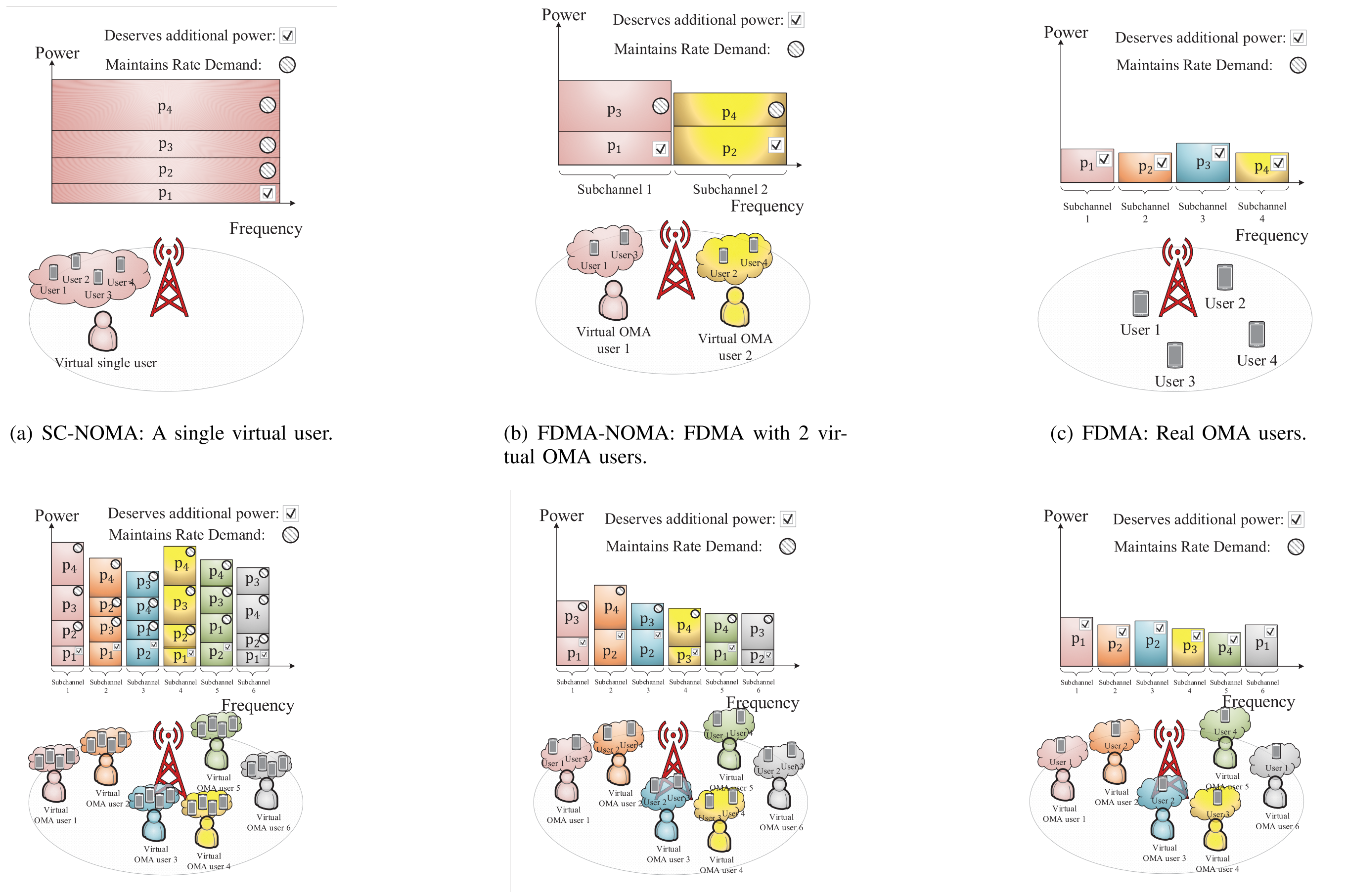
```
Calculate Ω_n = ( (W_s/(ln 2))/λ − (1 − c_n h^n_{Φ_n})/(α_n h^n_{Φ_n}) ) − P_n^mask, ∀n ∈ N.
if min_{n∈N}{Ω_n} > 0 then
    Find q* by using the water-filling Alg. 1.
else
    Initialize Lagrange multiplier ν^(0), step size ε_s, and iteration index t = 0.
    repeat
        Set t := t + 1.
        Find q̃^(t) by using q̃_n^(t) = [ (W_s/(ln 2))/(λ + ν^(t−1)) − 1/H_n ]_{Q̃_n^min}^{P̃_n^mask}, ∀n.
        Update ν^(t) = [ ν^(t−1) − ε_s ( P̃^max − Σ_{n∈N} q̃_n^(t) ) ]^+.
    until convergence of q̃^(t).
    Find q* by using q_n = q̃_n^* + c_n/α_n, ∀n ∈ N.
end if
```

**Algorithm 4** The Mixed Water-Filling/Barrier Method for Solving Problem (14)

```
Calculate Ω_n = ( (W_s/(ln 2))/λ − (1 − c_n h^n_{Φ_n})/(α_n h^n_{Φ_n}) ) − P_n^mask, ∀n ∈ N.
if min_{n∈N}{Ω_n} > 0 then
    Find q* by using the water-filling Alg. 1.
else
    Initialize q̃, 0 < α < 0.5, 0 < β < 1, μ > 1, t ≫ 1, 0 < ε_N ≪ 1, and 0 < ε_B ≪ 1.
    repeat
        Set Δq̃ = −∇U(q̃) (∇²U(q̃))^(−1).
        Set λ_B = −Δq̃.∇U(q̃)^T.
        if λ_B/2 ≤ ε_N then
            break
        end if
        Initialize l = 1.
        while (q̃_n + lΔq̃_n) ∉ [Q̃_n^min, P̃_n^mask], ∀n ∈ N, or Σ_{n∈N} (q̃_n + lΔq̃_n) > P̃^max do
            l := βl
        end while
        while U(q̃ + lΔq̃) > U(q̃) + αlΔq̃.∇U(q̃)^T do
            l := βl
        end while
        Set q̃ = q̃ + lΔq̃.
        if 1/t ≤ ε_B then
            break
        end if
        Set t := μt.
end if
```

### *C. Important Theoretical Insights of the Optimal Power Allocation for Maximizing SR/EE*

Here, we present the theoretical insights of optimal power allocation for the SR and EE maximization problems.




(a) SC-NOMA: A single virtual user.

(b) FDMA-NOMA: FDMA with 2 virtual OMA users.

(c) FDMA: Real OMA users.

(d) Hybrid-NOMA with $K$ multiplexed users: FDMA with 6 virtual OMA users.

(e) Hybrid-NOMA with 2 multiplexed users: FDMA with 6 virtual OMA users.

(f) OFDMA: FDMA with 6 virtual OMA users.

Fig. 2. The equivalent virtual FDMA models of Fig. 1 including virtual OMA users (see Remark 1).

*1) Sum-Rate Maximization:* In Hybrid-NOMA, it is guaranteed that at the optimal point, the cellular power constraint is active, meaning that all the available BS's power will be distributed among clusters. According to the proof of Proposition 2, it is guaranteed that at the optimal point, only the cluster-head users get additional power, and all the other users get power to only maintain their minimal rate demands on each subchannel. Hence, the remaining cellular power will be distributed among the cluster-head users. According to the analysis of KKT optimality conditions in Appendix C, it is shown that there is a competition among cluster-head users to get the rest of the cellular power.

*Remark 1: In the transformation of* (3) *to* (12)*, the Hybrid-NOMA system is equivalently transformed to a virtual FDMA system including a single virtual BS with maximum power* $\tilde{P}^{\max} = P^{\max} - \sum\limits_{n\in\mathcal{N}} \frac{c_n}{\alpha_n}$*, and* $N$ *virtual OMA users operating in* $N$ *subchannels with maximum allowable power* $\tilde{P}_n^{\text{mask}} = P_n^{\text{mask}} - \frac{c_n}{\alpha_n}$*,* $\forall n \in \mathcal{N}$*. Each cluster* $n \in \mathcal{N}$*, is indeed a virtual OMA user whose CNR is* $H_n = \alpha_n h^n_{\Phi_n}$*, which depends on* $\alpha_n$ *that is a function of the minimum rate demand of users with lower decoding order in cluster* $n$*, and the CNR of the cluster-head user, whose index is* $\Phi_n$*. The allocated power to the virtual OMA user* $n$ *is formulated by* $\tilde{q}_n = q_n - \frac{c_n}{\alpha_n}$*. Each virtual OMA user* $n$ *has also a minimum power demand* $\tilde{Q}_n^{\min} = Q_n^{\min} - \frac{c_n}{\alpha_n}$*, in order to guarantee the individual minimum rate demand of its multiplexed users in* $\mathcal{K}_n$ *on subchannel* $n$*. For any given virtual clusters power budget* $\tilde{\boldsymbol{q}} = [\tilde{q}_n], \forall n \in \mathcal{N}$*, the achievable rate of each virtual OMA user is the SR of its multiplexed users, which is the sum-capacity of subchannel* $n$*.*

Based on the definition of virtual OMA users for the SR maximization problem in Hybrid-NOMA and the KKT optimality conditions analysis, the exemplary models in Fig. 1 can be equivalently transformed to their corresponding virtual FDMA systems shown in Fig. 2. Note that FDMA/OFDMA is a special case of FDMA-NOMA/Hybrid-NOMA, where each subchannel is assigned to a single user. Hence, each OMA user acts as a cluster-head user, and subsequently, the virtual users are identical to the real OMA users, i.e., $\alpha_n = 1$, $H_n = h^n_{\Phi_n}$, and $c_n = 0$, for each $n \in \mathcal{N}$. As a result, each user in FDMA/OFDMA deserves additional power. In summary, the analysis for finding the optimal power allocation to maximize SR/EE of Hybrid-NOMA with per-symbol minimum rate constraints and FDMA is quite similar, and the only differences are $\alpha_n$ and $c_n$.

*Remark 2: Remark 1 shows that when $c_n \to 0$, the difference term $\frac{c_n}{\alpha_n} \to 0$ in $\tilde{q}_n$, $\tilde{Q}_n^{\min}$, and $\tilde{P}_n^{\text{mask}}$. Subsequently, when $c_n \to 0,\ \forall n \in \mathcal{N}$, we have $\sum\limits_{n\in\mathcal{N}} \frac{c_n}{\alpha_n} \to 0$ in $\tilde{P}^{\max}$. Accordingly, when $c_n \to 0,\ \forall n \in \mathcal{N}$, we guarantee that $\tilde{q}_n = q_n,\ \forall n \in \mathcal{N}$, $\tilde{Q}_n^{\min} = Q_n^{\min},\ \forall n \in \mathcal{N}$, $\tilde{P}_n^{\text{mask}} = P_n^{\text{mask}},\ \forall n \in \mathcal{N}$, and $\tilde{P}^{\max} = P^{\max}$. In other words, when $c_n \to 0,\ \forall n \in \mathcal{N}$, the network parameters of Hybrid-NOMA will be exactly the same as its virtual FDMA system.*

In each cluster $n$, the term $c_k^n$ tends to zero when $h_k^n \to \infty,\ k \in \mathcal{K}_n$. The numerical results verify that in most of the channel realizations, specifically high CNR regions, $c_k^n \approx 0,\ \forall n \in \mathcal{N}, k \in \mathcal{K}_n$ [23]. With assuming $c_k^n \approx 0,\ \forall n \in \mathcal{N}, k \in \mathcal{K}_n$, we have $c_n \approx 0, \forall n \in \mathcal{N}$ in (9). Hence, the results in Remark 2 are valid for the high CNR regions.

When $c_k^n \approx 0,\ \forall n \in \mathcal{N}, k \in \mathcal{K}_n$, the optimal intra-cluster powers in (7) and (8) can be approximated, respectively, as

$$p_k^{n*} \approx \left( \beta_k^n \prod_{\substack{j\in\mathcal{K}_n \\ h_j^n < h_k^n}} \left(1-\beta_j^n\right) \right) q_n, \quad \forall n \in \mathcal{N}, \quad k \in \mathcal{K}_n \setminus \{\Phi_n\}, \tag{20}$$

and

$$p_{\Phi_n}^{n}{}^* \approx \left( 1 - \sum_{\substack{i\in\mathcal{K}_n \\ h_i^n < h_{\Phi_n}^n}} \beta_i^n \prod_{\substack{j\in\mathcal{K}_n \\ h_j^n < h_i^n}} \left(1-\beta_j^n\right) \right) q_n, \quad \forall n \in \mathcal{N}. \tag{21}$$

For the case that the users in $\mathcal{K}_n$ have the same minimum rate demands $R_{k,n}^{\min}/W_s = R$ in bps/Hz, it is straightforward to show that (20) and (21) can be reformulated, respectively, by

$$p_k^{n*} \approx \frac{2^R - 1}{\left(2^R\right)^{\Theta_k}} q_n, \quad \forall n \in \mathcal{N},\ k \in \mathcal{K}_n \setminus \{\Phi_n\}, \tag{22}$$

and

$$p_{\Phi_n}^{n}{}^* \approx \frac{1}{\left(2^R\right)^{|\mathcal{K}_n|-1}} q_n, \quad \forall n \in \mathcal{N}, \tag{23}$$

where $\Theta_k = |\{i \in \mathcal{K}_n | h_i^n \leq h_k^n\}|$.

*Corollary 2: The approximated closed-form expressions* (22) *and* (23) *verify the high heterogeneity of optimal power coefficients among multiplexed users, thus the importance of finding optimal intra-cluster power allocation. For instance, the equal intra-cluster power allocation is infeasible in most of the cases, due to violating the minimum rate constraints in* (5b).

For the special case $|\mathcal{K}_n| = 2$, and $R = 1$ bps/Hz, we have $p_1^{n*} \approx p_2^{n*} \approx \frac{1}{2} q_n$, meaning that the equal intra-cluster power allocation is nearly optimal.

The inter-cluster power allocation is necessary when $\sum\limits_{n\in\mathcal{N}} P_n^{\text{mask}} > P^{\max}$, i.e., there is at least one cluster which is not allowed to operate at its maximum power $P_n^{\text{mask}}$. In this case, the distributed inter-cluster power allocation leads to violating the cellular power constraint (3c), since in the distributed power allocation among clusters, constraint (3d) will be active. Alternatively, when $\sum\limits_{n\in\mathcal{N}} P_n^{\text{mask}} \leq P^{\max}$, we guarantee that $q_n^* = P_n^{\text{mask}},\ \forall n \in \mathcal{N}$. There are a number of works, e.g., [30] and [31], assuming $P_n^{\text{mask}} = P^{\max}/N,\ \forall n \in \mathcal{N}$, i.e., equal inter-cluster power allocation while maintaining the cellular power constraint (3c). In this case, $q_n = P^{\max}/N,\ \forall n \in \mathcal{N}$, and the optimal intra-cluster power allocation can be obtained by using Proposition 2. In the following, we investigate the optimality condition for the equal inter-cluster power allocation.

*Proposition 5: When $c_k^n \approx 0,\ \forall n \in \mathcal{N}, k \in \mathcal{K}_n$, the equal inter-cluster power allocation, i.e., $q_n = P^{\max}/N,\ \forall n \in \mathcal{N}$, is optimal if and only if 1) $P^{\max}/N \in \left[Q_n^{\min}, P_n^{\text{mask}}\right],\ \forall n \in \mathcal{N}$; 2) $\frac{h_{\Phi_i}^i}{h_{\Phi_j}^j} = \frac{\alpha_j}{\alpha_i},\ \forall i,j \in \mathcal{N}$.*

*Proof:* The equal inter-cluster power allocation should be feasible to problem (12). According to Proposition 1, $q_n = P^{\max}/N,\ \forall n \in \mathcal{N}$, is feasible if and only if $P^{\max}/N \in [Q_n^{\min}, P_n^{\text{mask}}],\ \forall n \in \mathcal{N}$.

According to (13), two clusters/virtual OMA users $i,j \in \mathcal{N}$ get the same virtual powers, i.e., $\tilde{q}_i^* = \tilde{q}_j^*$, if and only if $H_i = H_j$. According to (9), for each cluster $n \in \mathcal{N}$, when $c_k^n \approx 0,\ \forall k \in \mathcal{K}_n$, we have $c_n \approx 0$. According to Remark 2, when $c_n \approx 0,\ \forall n \in \mathcal{N}$, we guarantee that $\tilde{q}_n = q_n,\ \forall n \in \mathcal{N}$. As a result, for two clusters $i,j \in \mathcal{N}$, we have $q_i^* = q_j^*$, if and only if $H_i = H_j$. By using $H_n = \alpha_n h_{\Phi_n}^n,\ \forall n \in \mathcal{N}$, defined in (12), $q_i^* = q_j^*$, if and only if $\frac{h_{\Phi_i}^i}{h_{\Phi_j}^j} = \frac{\alpha_j}{\alpha_i}$. Hence, $\tilde{q}_i^* = \tilde{q}_j^*,\ \forall i,j \in \mathcal{N}$, with $\sum\limits_{n\in\mathcal{N}} q_n^* = P^{\max}$, or equivalently $q_n^* = P^{\max}/N,\ \forall n \in \mathcal{N}$, if and only if $\frac{h_{\Phi_i}^i}{h_{\Phi_j}^j} = \frac{\alpha_j}{\alpha_i},\ \forall i,j \in \mathcal{N}$, and the proof is completed. □

According to Proposition 5, in Hybrid-NOMA, when $c_n \approx 0$, the equal inter-cluster power allocation is optimal if and only if all the virtual OMA users have exactly the same CNRs. These results also hold for FDMA, where $\alpha_n = 1,\ \forall n \in \mathcal{N}$, $H_n = h_{\Phi_n}^n,\ \forall n \in \mathcal{N}$, and $c_n = 0,\ \forall n \in \mathcal{N}$. According to Remark 1 and Proposition 5, the unique condition $\frac{h_{\Phi_i}^i}{h_{\Phi_j}^j} = \frac{\alpha_j}{\alpha_i},\ \forall i,j \in \mathcal{N}$, for the optimality of the equal inter-cluster power allocation states that when the cluster-head users have exactly the same CNRs, i.e., $h_{\Phi_i}^i = h_{\Phi_j}^j,\ \forall i,j \in \mathcal{N}$, the equal inter-cluster power allocation strategy is optimal if and only if $\alpha_i = \alpha_j,\ \forall i,j \in \mathcal{N}$. According to the definition of $\alpha_n$ in (9), one simple case that $\alpha_i \neq \alpha_j$ for some $i,j \in \mathcal{N}$, is considering different minimum rate demands for the users with lower decoding order.

*Corollary 3: In contrast to FDMA, the optimality condition of the equal inter-cluster power allocation strategy depends on the individual minimum rate demand of users with lower decoding order. This power allocation strategy can be suboptimal for Hybrid-NOMA even if the clusters have the same order and all the users in different clusters have the same CNRs. Moreover, the CNR of users with lower decoding order does not significantly affect the performance of the equal inter-cluster power allocation strategy.*

For the case that Proposition 5 holds, i.e., $H_i = H_j,\ \forall i,j \in \mathcal{N}$, thus $q_n^* = P^{\max}/N,\ \forall n \in \mathcal{N}$, the optimal $\nu^*$ in (13) can be obtained based on the quality $P^{\max}/N = \frac{W_s/(\ln 2)}{\nu^*} - \frac{1}{H_n}$. Hence, we have $\nu^* = \frac{\frac{P^{\max}}{N}+\frac{1}{H_n}}{W_s/(\ln 2)}$. In general, for the case that $H_n,\ \forall n \in \mathcal{N}$, is significantly large, i.e., high CNR regions of virtual OMA users, the second term $\frac{1}{H_n}$ in (13) tends to zero. In this case, we observe a low heterogeneity of inter-cluster power allocation among clusters, resulting in near-optimal performance for the equal inter-cluster power allocation strategy.

*2) EE Maximization:* Based on Proposition 3, we observe that the closed-form expressions of optimal intra-cluster power allocation are also valid for the EE maximization problem 4. Hence, Remark 1 and Fig. 2 are also valid for the EE maximization problem. Besides, Proposition 4 provides a sufficient condition during each Dinkelbach iteration in which the full cellular power consumption not only leads to the maximum SR, but also maximum EE. In other words, if (17) holds, the full cellular power consumption is energy efficient. The term $\frac{W_s/(\ln 2)}{\lambda} - \frac{1-c_n h^n_{\Phi_n}}{\alpha_n h^n_{\Phi_n}}$ in (17) is increasing in $\lambda = \frac{\sum\limits_{n\in\mathcal{N}}\sum\limits_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n)}{\sum\limits_{n\in\mathcal{N}}\sum\limits_{k\in\mathcal{K}_n} p_k^n + P_{\rm C}}$. The fractional parameter $\lambda$ is a decreasing function of $P_{\rm C}$. As a result, increasing $P_{\rm C}$ increases the term $\frac{W_s/(\ln 2)}{\lambda} - \frac{1-c_n h^n_{\Phi_n}}{\alpha_n h^n_{\Phi_n}}$ in (17). In other words, (17) holds when the circuit power consumption of the BS is significantly large.

*Corollary 4: In both the SR and EE maximization problems of Hybrid-NOMA with per-symbol minimum rate constraints, in each cluster, only the cluster-head user deserves additional power, and all the other users get power to only maintain their minimal rate demands. Our analysis proves that in the SR maximization problem, the BS operates at its maximum power budget. However, for the EE maximization problem, the BS may operate at lower power depending on the condition in Proposition 4.*

### D. Computational Complexity Analysis

In this subsection, we discuss about the computational complexity order of our proposed Algs. 1-4. To simplify the complexity analysis, we assume that $|\mathcal{K}_n| = K,\ \forall n \in \mathcal{N}$, in this subsection.

Alg. 1 belongs to the family of water-filling solutions which is comprehensively discussed in the literature [49]–[53]. The water-filling algorithms are mainly divided into two categories: 1) iterative algorithms, like bisection method, which stops until the error is below some tolerance threshold; 2) Exact algorithms based on hypothesis testing [49]. It is difficult to obtain the exact complexity of the bisection method to achieve an $\epsilon$-suboptimal performance, however we numerically observed that the error will be less than $10^{-6}$ mostly within 20 iterations. The exact algorithms have an exponential worst-case complexity on the order of $2^N$, however it is possible to obtain a linear worst-case complexity of $N$ [49], [51]. This linear complexity can be achieved by properly sorting the so-called sequences which is comprehensively discussed in [49] and [51]. Generally speaking, the number of water-filling iterations increases linearly with the number of subchannels $N$ [49], [51]. In each iteration, we obtain $\tilde{q}_n^*,\ \forall n \in \mathcal{N}$, by using (13), which needs $N$ operations. Therefore, the complexity of Alg. 1 is on the order of $N^2$. Note that the complexity of Alg. 1 is approximately independent of the number of multiplexed users $|\mathcal{K}_n|,\ \forall n \in \mathcal{N}$. This is due to the equivalent transformation of the Hybrid-NOMA problem (3) to its corresponding virtual FDMA problem (12). Increasing the number of multiplexed users $|\mathcal{K}_n|$ only increases the complexity of calculating $(Q_n^{\min}, \alpha_n)$ in the initialization step of Alg. 1 which is negligible.

Alg. 2 which is based on the Dinkelbach method converts the original problem (4) into a sequence of auxiliary problems, indexed by $\lambda$. The overall complexity of Alg. 2 mainly depends on both the convergence rate of the subproblems, as well as the computational complexity of each subproblem. By defining $E(\boldsymbol{p}) = \frac{f_1(\boldsymbol{p})}{f_2(\boldsymbol{p})}$, where $f_1(\boldsymbol{p}) = \sum\limits_{n\in\mathcal{N}}\sum\limits_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n)$, and $f_2(\boldsymbol{p}) = \sum\limits_{n\in\mathcal{N}}\sum\limits_{k\in\mathcal{K}_n} p_k^n + P_{\rm C}$, the convergence rate of Alg. 2 can be observed by formulating the update rule of the fractional parameter $\lambda$ as $\lambda_{(t+1)} = \frac{f_1\left(\boldsymbol{p}^*_{(t)}\right)}{f_2\left(\boldsymbol{p}^*_{(t)}\right)} = \lambda_{(t)} - \frac{f_1\left(\boldsymbol{p}^*_{(t)}\right)-\lambda_{(t)} f_2\left(\boldsymbol{p}^*_{(t)}\right)}{-f_2\left(\boldsymbol{p}^*_{(t)}\right)} = \lambda_{(t)} - \frac{F(\lambda_{(t)})}{F'(\lambda_{(t)})}$, where $t$ is the iteration index of Alg. 2, and $F(\lambda_{(t)}) = f_1\left(\boldsymbol{p}^*_{(t)}\right) - \lambda_{(t)} f_2\left(\boldsymbol{p}^*_{(t)}\right)$ [24]. It can be observed that Alg. 2 follows the Newton's method, meaning that the Newton's method is applied to the concave function $F(\lambda)$. Thus, Alg. 2 exhibits a super-linear convergence rate [24]. A detailed complexity analysis of the pure Newton's method can be found in Subsection 9.5.3 in [47]. For a general concave function $F(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^n$, if $F$ increases by at least $\Delta_F$ at each Newton's iteration, $\nabla^2 F(\boldsymbol{x}) \le -m$, and $\left\|\nabla^2 F(\boldsymbol{x}) - \nabla^2 F(\boldsymbol{y})\right\|_2 \le L\left\|\boldsymbol{x}-\boldsymbol{y}\right\|_2,\ \forall \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n$, the number of Newton's iterations to achieve an $\epsilon$-suboptimal solution is bounded above by $C_F = \frac{F(\boldsymbol{x}^*)-F(\boldsymbol{x}_{(0)})}{\Delta_F} + \log_2\log_2(\epsilon_0/\epsilon)$, where $\epsilon_0 = 2m^3/L^2$ [47]. For the accuracy around $\epsilon \approx 5.10^{-20}\epsilon_0$, we have $\log_2\log_2(\epsilon_0/\epsilon) \approx 6$ [47], thus in this case, the number of Newton's iterations is bounded above by $C_F \approx \frac{F(\boldsymbol{x}^*)-F(\boldsymbol{x}_{(0)})}{\Delta_F} + 6$.

In each iteration of Alg. 2, if Proposition 4 holds, we solve (14) for the given $\lambda$ by using the water-filling Alg. 1, whose overall complexity is $N^2$. For the case that Proposition 4 does not hold in each Dinkelbach iteration, we solve (14) for the given $\lambda$ by using the subgradient or barrier methods presented in Algs. 3 and 4, respectively. The duality gap of the barrier method in Alg. 4 after $L$ iterations is $1/(\mu^L t^0)$, where $t^0$ is the initial $t$, and $\mu$ is the stepsize for updating $t$ in the barrier method. Therefore, after exactly $\lceil \frac{\ln\left(\frac{1}{\epsilon_B \cdot \mu}\right)}{\ln(\mu)} \rceil$ barrier iterations, Alg. 4 achieves $\epsilon_B$-suboptimal solution [47]. In each barrier iteration, we apply the Newton's method. In general, it is difficult to obtain the exact complexity order of the pure Newton's method [47]. According to Subsection 11.5.3 in [47], when the self-concordance assumption holds, the total number of Newton's iterations over all the barrier iterations to achieve an $\epsilon_B$-suboptimal solution is bounded above by $\lceil \frac{\ln\left(\frac{1}{\epsilon_B \cdot \mu}\right)}{\ln(\mu)} \rceil \left(\frac{\mu-1-\ln\mu}{\Delta_F} + \log_2\log_2(1/\epsilon_N)\right)$, where

TABLE III
COMPUTATIONAL COMPLEXITY OF SOLVING THE SUM-RATE MAXIMIZATION PROBLEM (3)

| NOMA-to-OMA Transformation | Alg. 1 | Subgradient Method | Barrier Method |
|---|---|---|---|
| Complexity | $N^2$ | $C_S(1+N)$ | $\left\lceil \frac{\ln\left(\frac{1}{\epsilon_B \cdot \mu}\right)}{\ln(\mu)} \right\rceil C_N$ |
| **Pure Methods** | **Water-Filling Algorithm [56]** | **Subgradient Method** | **Barrier Method** |
| Complexity | $N^2K$ | $C_S(1+N+2KN)$ | $\left\lceil \frac{\ln\left(\frac{1+N+KN}{\epsilon_B \cdot \mu}\right)}{\ln(\mu)} \right\rceil C_N$ |

TABLE IV
COMPUTATIONAL COMPLEXITY OF SOLVING THE ENERGY EFFICIENCY MAXIMIZATION PROBLEM (4)

| NOMA-to-OMA Transformation | Alg. 2-inner Alg. 1 | Alg. 2-inner Alg. 3 | Alg. 2-inner Alg. 4 | Exploring $P_{\text{EE}}$-inner Alg. 1 |
|---|---|---|---|---|
| Complexity | $N^2$ | $C_F C_S(1+N)$ | $C_F \left\lceil \frac{\ln\left(\frac{1}{\epsilon_B \cdot \mu}\right)}{\ln(\mu)} \right\rceil C_N$ | $\left\lceil \frac{P^{\max} - \sum\limits_{n \in \mathcal{N}} Q_n^{\min}}{\delta} \right\rceil N^2$ |
| **Pure Methods** | **Alg. 2-inner Water-Filling [56]** | **Alg. 2-inner Subgradient Method [35]–[37]** | **Alg. 2-inner Barrier Method** | **Exploring $P_{\text{EE}}$-inner Water-Filling [56]** |
| Complexity | $N^2K$ | $C_F C_S(1+N+2KN)$ | $C_F \left\lceil \frac{\ln\left(\frac{1+N+KN}{\epsilon_B \cdot \mu}\right)}{\ln(\mu)} \right\rceil C_N$ | $\left\lceil \frac{P^{\max}}{\delta} \right\rceil N^2K$ |

$\epsilon_N$ is the tolerance of the Newton's method in each barrier iteration. The complexity of other operations in the centering step of each barrier iteration is negligible. As a result, when Proposition 4 does not hold, the overall worst-case complexity of Alg. 2 with inner Alg. 4 is approximately on the order of $C_F\left(\lceil \frac{\ln\left(\frac{1}{\epsilon_B \cdot \mu}\right)}{\ln(\mu)} \rceil \left(\frac{\mu-1-\ln\mu}{\Delta_F} + \log_2\log_2(1/\epsilon_N)\right)\right)$, where $C_F$ denotes the number of Dinkelbach iterations in Alg. 2.

The standard subgradient method produces a global optimum, however its exact computational complexity is still unknown in general [54], [55]. It is shown that the subgradient method converges with polynomial complexity in the number of optimization variables and constraints [54], [55]. In each subgradient iteration of Alg. 3, we need to calculate $\tilde{\boldsymbol{q}}^{(t)}$ in Step 8 which requires $N$ operations. Then, we update the Lagrange multiplier $\nu$ whose complexity order is 1. Thus, the overall complexity of Alg. 2 with inner Alg. 3 is $C_F C_S(N+1)$, where $C_S$ indicates the number of subgradient iterations.

The computational complexity order of our proposed as well as other existing globally optimal power allocation algorithms for solving the SR and EE maximization problems is summarized in Tables III and IV, respectively. In these tables, the term "pure" is referred to the case that we do not apply Propositions 1 and 2 (thus the equivalent transformation of Hybrid-NOMA to a FDMA system, denoted by "NOMA-to-OMA Transformation") in the convex solvers. The parameters $C_F$ and $C_S$ denote the number of Dinkelbach and subgradient iterations, respectively. Moreover, $C_N$ denotes the number of Newton's iteration in each barrier iteration. The parameter $\delta$ in Table IV indicates the stepsize of exhaustive search for finding $P_{\text{EE}}$. In Table III, the pure water-filling algorithm needs to update $p_k^n,\ \forall n \in \mathcal{N},\ k \in \mathcal{K}_n$, which requires $NK$ operations[5] [56]. Hence, the overall complexity of the pure water-filling algorithm is on the order of $N^2K$. Therefore, Alg. 1 reduces the complexity of the pure water-filling algorithm by $K$ times, where $K$ is the number of multiplexed users in each subchannel. It is also possible to solve problem (3) or its equivalent FDMA problem (12) by using the subgradient or barrier (with inner Newton's algorithm) methods. As can be seen, the equivalent NOMA-to-OMA transformation also reduces the complexity of these solvers. Besides, Alg. 1 has the lowest computational complexity compared to the other existing methods. The latter conclusions also hold for the EE maximization problem shown in Table IV. When Proposition 4 holds, we can use Alg. 1 with the lowest computational complexity compared to the other existing convex solvers. The superiority of the Dinkelbach algorithm can be observed by comparing it with a greedy search over all the possible power consumption of the BS, denoted by $P_{\text{EE}}$. Although Proposition 1 can reduce the search area, such that we can obtain the lower-bound of $P_{\text{EE}}$ as $\sum\limits_{n \in \mathcal{N}} Q_n^{\min}$ (see (6)), as well as reduce the complexity of the pure water-filling algorithm by using Proposition 2, the overall complexity of exploring $P_{\text{EE}}$ is still large, when the stepsize $\delta$ is significantly small.

The numerical experiments show that Alg. 2 converges in less than 6 iterations, meaning that $C_F \approx 6$. In each Dinkelbach iteration, the subgradient method in Alg. 3 converges within $C_S \approx 15$ iterations. Besides, Alg. 4 converges within 10 barrier iterations. For significantly large number of users around 100 to 200, the simulation codes in [46] verify that the convergence time of our proposed algorithms is on the order of milliseconds. Based on our numerical experiments, we observed that the convergence time of the subgradient method in Alg. 3 is less than that of the barrier method in Alg. 4.

### *E. Subchannel Allocation in MC-NOMA*

The optimal subchannel allocation problem, i.e., finding optimal $\boldsymbol{\rho} = [\rho_k^n]$ or equivalently cluster sets $\mathcal{K}_n$, in MC-NOMA is classified as integer nonlinear programming problem. The subchannel allocation is determined on the top of power allocation. Therefore, the exact closed form of inter-cluster power allocation is required for solving the subchannel allocation problem. Although Alg. 1 approaches the globally optimal solution with a fast convergence speed, the exact value of $\nu^*$ and subsequently, closed-form of $\boldsymbol{q}^*$ is still unknown in general. A similar issue exists for the water-filling algorithms for the FDMA problems [49]–[53]. The Dinkelbach

[5]The pure water-filling algorithm in [56] is for uplink MC-NOMA without considering users minimum rate constraints.

and subgradient methods also have similar issues, in which the exact value of optimal $\lambda$ and $\nu$ are unknown in general, respectively. The joint optimal user clustering and power allocation is known to be strongly NP-hard [27]–[29]. Although the latter problem is strongly NP-hard, the optimal number of clusters or subchannels in FDMA-NOMA can be obtained as follows:

*Proposition 6: In a $K$-user FDMA-NOMA system with limited number of multiplexed users $U^{\max}$, the optimal number of clusters is $N^* = \lceil K/U^{\max} \rceil$.*

*Proof:* Due to the degradation of SISO Gaussian BCs, it is proved that SC-NOMA is capacity achieving, meaning that the rate region of FDMA/TDMA is a subset of the rate region of SC-NOMA [1]–[5]. Hence, for the case that $K < U^{\max}$, the optimal user clustering is considering all the users in the same cluster, and apply SC-SIC among all the users, i.e., FDMA-NOMA turns into SC-NOMA. Now, consider $K = U^{\max} + C$, where $1 \leq C \leq U^{\max}$. In this case, SC-NOMA is infeasible, however, FDMA-NOMA divides $K$ users into two isolated clusters $\mathcal{K}_1$ and $\mathcal{K}_2$ satisfying $|\mathcal{K}_n| \leq U^{\max},\ n = 1, 2$, due to the existing limitation on the number of multiplexed users. Each cluster set $\mathcal{K}_n,\ n = 1, 2$ is a SISO Gaussian BC whose capacity region can be achieved by using SC-SIC. Hence, further dividing each user group $\mathcal{K}_n,\ n = 1, 2$, based on FDMA/TDMA would result lower achievable rate. The latter result holds for any possible 2 groups with $1 \leq C \leq U^{\max}$. Now, consider a general case $MU^{\max} + 1 \leq K \leq (M + 1)U^{\max}$ with nonnegative integer $M$. In this case, the lowest possible number of isolated clusters is $M + 1$. Further imposing FDMA/TDMA to any existing group would result in a suboptimal performance. Accordingly, the optimal number of clusters is exactly $\lceil K/U^{\max} \rceil$. □

Proposition 6 shows that the achievable rate of FDMA with the highest isolation among users is a subset of the achievable rate of FDMA-NOMA with any given user clustering. Since our globally optimal power allocation algorithms are valid for any given user clustering, the existing suboptimal user clustering algorithms, such as heuristic methods in [30], [31], [34]–[36], and matching-based algorithms in [32] and [33] can be applied. Another approach is the framework in [57] which is the joint optimization of power and subchannel allocation with the relaxed-and-rounding method. However, the output is still suboptimal without any mathematical performance improvement guarantee. Roughly speaking, there is still no mathematical understanding analysis for performance comparison among the existing suboptimal user clustering algorithms. The optimal user clustering is still unknown, and is considered as a future work.

## IV. Extensions and Future Research Directions

Here, we discuss about the possible extensions of our analyses to more general scenarios. For each case, the potential challenges are discussed in details.

### A. Users Maximum Rate Constraint

According to Propositions 2 and 3, we conclude that at the optimal point of the SR/EE maximization problems, only the cluster-head users get additional power. In practical systems, the achievable rate of users is also limited by a maximum value due to the discrete modulation and coding schemes [2], [3]. In the SR/EE maximization of Hybrid-NOMA with significantly large number of subchannels and/or multiplexed users, it merely happens that a cluster-head user's rate within a subchannel exceeds the truncated Shannon's bound. This is due to the fact that 1) The clusters power budget will be typically low, on the order of few Watts, or even mWatts; 2) Mostly, a large portion of the clusters power budget will be allocated to the non-cluster-head users. For the sake of completeness, we discuss about the impact of considering per-subchannel maximum rate constraints in the SR/EE maximization problems. To keep the generality, let us define $R_{k,n}^{\max}$ as the individual maximum allowable rate of user $k$ on subchannel $n$. The maximum rate constraint can thus be formulated as

$$R_k^n(\boldsymbol{p}^n) \leq R_{k,n}^{\max}, \quad \forall n \in \mathcal{N},\ k \in \mathcal{K}_n. \tag{24}$$

By adding (24) to the original SR maximization problem (3), constraints (3b) and (24) can be combined as

$$R_{k,n}^{\min} \leq R_k^n(\boldsymbol{p}^n) \leq R_{k,n}^{\max}, \quad \forall n \in \mathcal{N},\ k \in \mathcal{K}_n. \tag{25}$$

Obviously, the minimum rate demands should be chosen such that $R_{k,n}^{\min} \leq R_{k,n}^{\max},\ \forall n \in \mathcal{N}, k \in \mathcal{K}_n$, otherwise the feasible set of (25) will be empty. According to Proposition 1, we can guarantee that at the optimal point of the total power minimization problem, $R_k^n(\boldsymbol{p}^{*n}) = R_{k,n}^{\min} \leq R_{k,n}^{\max},\ \forall n \in \mathcal{N}, k \in \mathcal{K}_n$, meaning that the maximum rate constraint (24) has no impact on the lower-bound of $\boldsymbol{q}$. Thus, the lower-bound of $q_n$ can be obtained by (6). On the other hand, the upper-bound of $q_n$ can be achieved by solving the per-cluster total power maximization problem (when the cellular power constraint is eliminated). Let us denote $Q_n^{\max}$ as the power consumption of cluster $n$, where $R_k^n(\boldsymbol{p}^n) = R_{k,n}^{\max},\ \forall k \in \mathcal{K}_n$. Similar to Appendix A, it can be easily shown that $Q_n^{\max}$ can be obtained by (6) in which $\beta_k^n = 2^{(R_{k,n}^{\max}/W_s)} - 1,\ \forall n \in \mathcal{N},\ k \in \mathcal{K}_n$. In this way, the feasible set of problem (5) with maximum rate constraints in (24) can be characterized as the intersection of $q_n \in \left[Q_n^{\min}, \min\left\{Q_n^{\max}, P_n^{\text{mask}}\right\}\right],\ \forall n \in \mathcal{N}$, and cellular power constraint $\sum\limits_{n \in \mathcal{N}} q_n \leq P^{\max}$. It is straightforward to show that when $\sum\limits_{n \in \mathcal{N}} \min\left\{Q_n^{\max}, P_n^{\text{mask}}\right\} \leq P^{\max}$, the cellular power constraint (3c) will be always fulfilled, thus it can be removed from problems (3) and (4). In this case, the SR maximization problem (3) can be equivalently divided into $N$ SC-NOMA subproblems, since there is no longer the competition among clusters to get the cellular power budget. Subsequently, at the optimal point of the SR maximization problem, we guarantee that each cluster $n$ achieves its maximum allowable power, i.e., $q_n^* = \min\left\{Q_n^{\max}, P_n^{\text{mask}}\right\},\ n \in \mathcal{N}$. Hence, the inter-cluster power allocation is required if and only if $\sum\limits_{n \in \mathcal{N}} \min\left\{Q_n^{\max}, P_n^{\text{mask}}\right\} > P^{\max}$.

Consider a simple 2-user SC-NOMA system with $h_1 < h_2$, thus the optimal decoding order is $2 \rightarrow 1$. The SR maximization problem (3) with per-user maximum rate constraint can be formulated as follows:

$$\max_{\boldsymbol{p} \geq 0}\ R_1(\boldsymbol{p}) + R_2(\boldsymbol{p}) \tag{26a}$$

$$\text{s.t. } R_k(\boldsymbol{p}) \geq R_k^{\min}, \quad \forall k=1,2, \tag{26b}$$
$$R_k(\boldsymbol{p}) \leq R_k^{\max}, \quad \forall k=1,2, \tag{26c}$$
$$p_1+p_2 \leq \min\{P^{\max}, Q^{\max}\}, \tag{26d}$$

where $Q^{\max}$ is the maximum power consumption of the BS, due to constraint (26c). The problem (26) is convex with an affine feasible set. Assume that $R_k^{\min} < R_k^{\max}$, $k=1,2$. At the optimal point, Condition **C1** : $p_1^*+p_2^*=\min\{P^{\max}, Q^{\max}\}$ always holds. According to the proof of Proposition 2, when (26c) for user 2 is removed from (26), at the optimal point, the following properties hold

$$\textbf{C2}: \; p_1^*+p_2^*=P^{\max}, \quad \textbf{C3}: \; R_1(\boldsymbol{p}^*)=R_1^{\min}.$$

In this case, based on Proposition 2, the optimal powers can be obtained as

$$p_1^*=\beta_1\left(P^{\max}+\frac{1}{h_1}\right), \quad p_2^*=(1-\beta_1)\,P^{\max}-\frac{\beta_1}{h_1}, \tag{27}$$

where $\beta_1=\frac{2^{(R_1^{\min}/W_s)}-1}{2^{(R_1^{\min}/W_s)}}$. Constraint (26c) for user 2 can be rewritten as

$$p_2 \leq \left(2^{(R_2^{\max}/W_s)}-1\right)/h_2. \tag{28}$$

Hence, the maximum rate constraint of user 2 is indeed a maximum power consumption constraint for this user. Let us define

$$M_2=\left(2^{(R_2^{\max}/W_s)}-1\right)/h_2.$$

According to Condition **C1**, (27) and (28), the optimal powers with imposing (26c) for both the users can be obtained as

$$p_1^*=\min\{P^{\max}, Q^{\max}\}-\min\left\{(1-\beta_1)\,P^{\max}-\frac{\beta_1}{h_1}, M_2\right\},$$
$$p_2^*=\min\left\{(1-\beta_1)\,P^{\max}-\frac{\beta_1}{h_1}, M_2\right\}. \tag{29}$$

Hence, if $(1-\beta_1)\,P^{\max}-\frac{\beta_1}{h_1} \leq M_2$, we guarantee that $R_1(\boldsymbol{p}^*)=R_1^{\min}$, $R_2^{\min} \leq R_2(\boldsymbol{p}^*) \leq R_2^{\max}$, and $p_1^*+p_2^* \leq P^{\max} < Q^{\max}$. If $(1-\beta_1)\,P^{\max}-\frac{\beta_1}{h_1} > M_2$, and $P^{\max} \leq Q^{\max}$, we guarantee that $R_1^{\min} \leq R_1(\boldsymbol{p}^*) \leq R_1^{\max}$, $R_2(\boldsymbol{p}^*)=R_2^{\max}$, and $p_1^*+p_2^*=P^{\max} \leq Q^{\max}$. Finally, if $(1-\beta_1)\,P^{\max}-\frac{\beta_1}{h_1} > M_2$, and $P^{\max} > Q^{\max}$, we guarantee that $R_1(\boldsymbol{p}^*)=R_1^{\max}$, $R_2(\boldsymbol{p}^*)=R_2^{\max}$, and $p_1^*+p_2^*=Q^{\max} < P^{\max}$. According to the above, Proposition 2 holds if and only if $(1-\beta_1)\,P^{\max}-\frac{\beta_1}{h_1} \leq M_2$. When user 2 exceeds its maximum rate $R_2^{\max}$, we allocate power to user 2 until $R_2(\boldsymbol{p}^*)=R_2^{\max}$, and the rest of the cellular power will be allocated to user 1 until it achieves its maximum rate. The latter analysis can be generalized to the $K$-user SC-NOMA system. For more details, please see Appendix G. The analysis in Appendix G shows that there exists a closed-form of optimal power allocation for the general $K$-user SC-NOMA with per-user minimum and maximum rate constraints. During the power allocation, there exists a special user $i$, where all of the stronger users than user $i$ achieve their maximum rates, and all of the weaker users than user $i$ achieve their minimum rates. Due to the space limitations, obtaining the closed-form of optimal powers, and how to define the index of user $i$ for a given power budget is considered as a future work.[6] After obtaining the closed-form of optimal powers as a function of the clusters power budget $\boldsymbol{q}$ in Hybrid-NOMA with per-subchannel maximum and minimum rate constraints, it might be possible to transform the Hybrid-NOMA problem to a FDMA problem, which can be considered as a future work.

### *B. Hybrid-NOMA With Per-User Minimum Rate Constraints*

In our work, we considered a Hybrid-NOMA system, where the minimum rate demand of each user on each assigned subchannel is predefined, similar to [33]–[37]. This scheme is the generalized model of FDMA-NOMA considered in [30]–[32]. From the optimization perspective, the SR/EE maximization problem for Hybrid-NOMA with predefined minimum rate demand of each user on each assigned subchannel, and FDMA-NOMA has similar structures, and both of them are convex. A more general/complicated case is when we consider a per-user minimum rate constraint over all the assigned subchannels. The SR maximization problem for Hybrid NOMA with per-user minimum rate constraint can be formulated as

$$\max_{\boldsymbol{p}\geq 0} \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n) \tag{30a}$$
$$\text{s.t. (3c), (3d),}$$
$$R_k(\boldsymbol{p}) \geq R_k^{\min}, \quad \forall k \in \mathcal{K}, \tag{30b}$$

where $R_k(\boldsymbol{p})=\sum\limits_{n\in\mathcal{N}_k} R_k^n(\boldsymbol{p}^n)$ denotes the achievable rate of user $k$ over all the assigned subchannels in $\mathcal{N}_k$. The term $R_k^n(\boldsymbol{p}^n)$ for each user $k \in \mathcal{K}_n \setminus \{\Phi_n\}$ is nonconcave in $\boldsymbol{p}^n$, due to the co-channel interference term $\sum\limits_{\substack{j\in\mathcal{K}_n,\\ h_j^n>h_k^n}} p_j^n h_k^n$. Since each two terms $R_k^i(\boldsymbol{p}^i)$ and $R_k^j(\boldsymbol{p}^j)$ for subchannels $i,j \in \mathcal{N}_k$ includes disjoint set of powers, we can conclude that $R_k(\boldsymbol{p})=\sum\limits_{n\in\mathcal{N}_k} R_k^n(\boldsymbol{p}^n)$ is nonconcave when $|\mathcal{N}_k|>1$ and $\exists n \in \mathcal{N}_k$, $k \neq \Phi_n$, which makes (30b) nonconcave. It is still unknown how to equivalently transform (30b) to a convex form. To this end, the globally optimal solution of (30) with polynomial time complexity is not yet obtained in the literature. One suboptimal solution for (30) is to approximate each nonconcave rate function $R_k^n(\boldsymbol{p}^n)$ to its first order Taylor series, and then apply the sequential programming method [15], [44], [58]. A suboptimal penalty function method is also used in [45]. Let us define an auxiliary variable $r_k^n$ indicating the minimum rate demand of user $k \in \mathcal{K}_n$ on subchannel $n$ in bps. In this way, problem (30) can be equivalently transformed to the following joint power and minimum rate allocation problem as

$$\max_{\boldsymbol{p}\geq 0,\ \boldsymbol{r}\geq 0} \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{K}_n} R_k^n(\boldsymbol{p}^n) \tag{31a}$$
$$\text{s.t. (3c), (3d),}$$
$$R_k^n(\boldsymbol{p}^n) \geq r_k^n, \quad \forall n \in \mathcal{N},\ k \in \mathcal{K}_n, \tag{31b}$$

[6] In problem (3) without maximum rate constraints, the cluster-head user, whose index is $\Phi_n$, is the special user $i$, thus none of the other multiplexed users deserve additional power. This is the main reason that we define the special notation $\Phi_n$ for the cluster-head user of subchannel $n$ in Subsection II-A.

$$\sum_{n\in\mathcal{N}_k} r_k^n = R_k^{\min}, \quad \forall k\in\mathcal{K}, \tag{31c}$$

where $\boldsymbol{r} = [r_k^n],\ \forall n\in\mathcal{N}, k\in\mathcal{K}_n$. For any given feasible $\boldsymbol{r}$ satisfying constraints in (31c), problem (31) can be equivalently transformed to the convex problem (3) with minimum rate demands $r_k^n,\ \forall n\in\mathcal{N}, k\in\mathcal{K}_n$. Hence, our analyses and important theoretical insights hold for any given $\boldsymbol{r}$ in the SR/EE maximization problem of Hybrid-NOMA with per-user minimum rate constraints. According to the above, the only challenge which is not yet solved is how to find $\boldsymbol{r}^*$ in (31) or equivalently distribute $R_k^{\min}$ over the subchannels in $\mathcal{N}_k$.

*Corollary 5: In Hybrid-NOMA with per-user minimum rate demands over all the assigned subchannels, if user $k\in\mathcal{K}$ is a non-cluster-head user in all the assigned subchannels, e.g., a cell-edge user, at the optimal point of SR/EE maximization, it gets power to only maintain its minimum rate demand $R_k^{\min}$, meaning that $R_k(\boldsymbol{p}^*) = \sum_{n\in\mathcal{N}_k} R_k^n(\boldsymbol{p}^{*n}) = R_k^{\min}$.*

Accordingly, each user $k\in\mathcal{K}$ deserves additional power if and only if it is a cluster-head user in at least one of the assigned subchannels. As a result, when the minimum rate demand of users is zero, in both the SR and EE maximization problems, only the cluster-head users get positive power, thus Hybrid-NOMA will be identical to OFDMA (also see Lemma 8 in [27]). These results show that in both the SR and EE maximization problems of Hybrid-NOMA, there exists a critical fairness issue among users' achievable rate which is discussed in the following subsection.

### C. Users' Rate Fairness

According to (22) and (23), we observe that in the SR/EE maximization problems, a large portion of the clusters power budget will be allocated to the users with lower decoding order when all the multiplexed users have the same minimum rate demands within a cluster. It states that in contrast to FDMA, NOMA *usually* allocates more power to the weaker users when all the multiplexed users have the same minimum rate demands. This result shows that NOMA provides users fairness in terms of power allocation. However, according to Corollaries 4 and 5, we observe that this users' power fairness does not necessarily lead to the users' rate fairness, since only one user in each cluster gets additional rate. Accordingly, substantial works are required to guarantee users' rate fairness. There exist many fairness schemes which are recently considered for SC/MC-NOMA, as proportional fairness [27], [29], [32], [38], [39], [42], [58], max-min fairness [32], and etc. In the following, we first discuss about the advantages/challenges of the proportional fairness scheme, where our objective is to tune the users achievable rate at the optimal point by maximizing the weighted SR of users. Then, we propose a new fairness scheme which is a mixture of proportional fairness and users weighted minimum rate demands.

*1) Proportional Fairness:* In proportional fairness, we aim at maximizing the weighted SR of users formulated by $\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n}\omega_k R_k^n(\boldsymbol{p}^n)$, where $\omega_k$ is the weight of user $k\in\mathcal{K}$, that is a constant, and is determined on the top of resource allocation. The weighted SR maximization problem can thus be formulated by

$$\max_{\boldsymbol{p}\geq 0}\ \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n}\omega_k R_k^n(\boldsymbol{p}^n), \quad \text{s.t. (3b)-(3d).} \tag{32}$$

The feasible region of problem (32) can be characterized by using Proposition 1. For each cluster $n$, it can be shown that if $\omega_i \geq \omega_j,\ \forall i,j\in\mathcal{K}_n,\ h_i^n\geq h_j^n$, the weighted SR function $\sum_{k\in\mathcal{K}_n}\omega_k R_k^n(\boldsymbol{p}^n)$ is negative definite. In this case, the globally optimal powers can be obtained by using Proposition 2, meaning that the weights $\omega_k,\ \forall k\in\mathcal{K}_n$ do on affect the optimal intra-cluster power allocation policy, thus users achievable rate. Moreover, Alg. 1 finds the globally optimal solution of problem (32), such that based on (13), the optimal $\tilde{q}_n^*$ can be obtained as

$$\tilde{q}_n^* = \begin{cases} \dfrac{\omega_{\Phi_n} W_s}{\ln(2)\nu^*} - \dfrac{1}{H_n}, & \left(\dfrac{\omega_{\Phi_n} W_s}{\ln(2)\nu^*} - \dfrac{1}{H_n}\right) \in [\tilde{Q}_n^{\min}, \tilde{P}_n^{\text{mask}}], \\ 0, & \text{otherwise}. \end{cases} \tag{33}$$

The closed-form expression (33) states that when $\omega_i \geq \omega_j,\ \forall n\in\mathcal{N}, i,j\in\mathcal{K}_n,\ h_i^n\geq h_j^n$, we can only tune the fairness among cluster-head users. It corresponds to tuning the fairness among clusters/virtual OMA users defined in Remark 1. To tune fairness among the multiplexed users within each cluster in the proportional fairness scheme, we need to assign more weights to the weaker users. For the case that $\exists n\in\mathcal{N}, i\neq j\in\mathcal{K}_n,\ \omega_i<\omega_j,\ h_i^n\geq h_j^n$, the weighted SR function $\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n}\omega_k R_k^n(\boldsymbol{p}^n)$ could be nonconcave, which makes problem (32) nonconvex [27]. In this regard, the strong duality in (32) does not hold, thus there exists a certain duality gap in the Lagrange dual method [27]. Although there are some interesting approximation analyses for the weighted SR function [29], the globally optimal solution of problem (32) for the case that $\exists n\in\mathcal{N}, i\neq j\in\mathcal{K}_n,\ \omega_i<\omega_j,\ h_i^n\geq h_j^n$, is still an open problem. In this case, one suboptimal solution is to apply the well-known sequential programming method [58].

*2) Mixed Weighted Sum-Rate/Weighted Minimum Rate Fairness:* In contrast to FDMA, proportional fairness in SC/MC-NOMA leads to a nonconvex problem in general, which greatly increases the complexity of finding the globally optimal power allocation. Another issue in proportional fairness is properly determining users weights prior to resource allocation. It is still unknown how to properly choose the users weight in order to achieve the desired users data rates after the optimal power allocation optimization which is important to guarantee users rate fairness. According to (13), we can conclude that in FDMA (with $\alpha_n=1, H_n=h_{\Phi_n}^n$, and $c_n=0$, for each $n\in\mathcal{N}$), the users minimum rate demand merely impacts on the optimal power allocation policy, thus users achievable rate. It means that tuning the minimum rate demand of users in FDMA merely impacts on the users data rate at the optimal point. In contrast to FDMA, we observe that the non-cluster-head users minimum rate demands highly affect the optimal intra-cluster power allocation of SC/MC-NOMA formulated in Proposition 2. In particular, we observe that all the non-cluster-head users achieve their predefined minimum rate demands on each assigned subchannel at the optimal point of the SR/EE maximization problems. Hence, by properly increasing the target minimum rate demands of the non-cluster-head users, we not only guarantee the multiplexed users rate fairness, but also the exact achievable rate of the non-cluster-head users on each subchannel before power allocation optimization.

Let us define $\Lambda_{k,n}$ as the weight of the minimum rate demand of user $k$ on subchannel $n$. In our proposed fairness scheme, we define the minimum rate demand of each non-cluster-head user $k \in \mathcal{K}_n \setminus \{\Phi_n\}$ as $S^{\min}_{k,n} = \Lambda_{k,n} R^{\min}_{k,n}$ with $\Lambda_{k,n} \geq 1$. Based on Remark 1, the minimum rate demand of the cluster-head user $\Phi_n, \ \forall n \in \mathcal{N}$, merely impacts on the optimal power allocation formulated in Proposition 2, thus the cluster-head users achievable rate. To this end, we set $\Lambda_{\Phi_n,n} = 1, \ \forall n \in \mathcal{N}$. By using the fact that each cluster-head user acts as an OMA user (see the paragraph after Remark 1), we apply the proportional fairness scheme among the cluster-head users in which we define $\omega_{\Phi_n}$ as the weight of the cluster-head user $\Phi_n$. Finally, the power allocation problem for the mixed weighted SR/weighted minimum rate fairness can be formulated as

$$\max_{\boldsymbol{p} \geq 0} \sum_{n \in \mathcal{N}} \sum_{k \in \mathcal{K}_n \setminus \{\Phi_n\}} R^n_k(\boldsymbol{p}^n) + \omega_{\Phi_n} R^n_{\Phi_n}(\boldsymbol{p}^n) \quad (34a)$$

$$\text{s.t. (3c), (3d),}$$

$$R^n_k(\boldsymbol{p}^n) \geq S^{\min}_{k,n}, \quad \forall n \in \mathcal{N}, \ k \in \mathcal{K}_n. \quad (34b)$$

According to the discussions in Subsection IV-C.1, it is straightforward to show that the objective function (34a) is strictly concave if we set $\omega_{\Phi_n} \geq 1, \ \forall n \in \mathcal{N}$. For any given $\omega_{\Phi_n} \geq 1, \ \forall n \in \mathcal{N}$, the intra-cluster optimal powers of problem (34) can be obtained by using Proposition 2, in which we substitute $R^{\min}_{k,n}$ with $S^{\min}_{k,n}$. In this fairness scheme, $\omega_{\Phi_n}, \ \forall n \in \mathcal{N}$, are chosen to only tune the fairness among cluster-head users. Thus, we can set $\omega_{\Phi_n} \geq 1, \ \forall n \in \mathcal{N}$, such that the fairness of non-cluster-head users is guaranteed by parameter $\Lambda_{k,n}$ in $S^{\min}_{k,n} = \Lambda_{k,n} R^{\min}_{k,n}$ in constraint (34b). In summary, the fairness parameters in problem (34) satisfy $\Lambda_{k,n} \geq 1, \forall n \in \mathcal{N}, \ k \in \mathcal{K}_n \setminus \{\Phi_n\}, \Lambda_{k,\Phi_n} = 1, \forall n \in \mathcal{N}$, and $\omega_{\Phi_n} \geq 1, \ \forall n \in \mathcal{N}$. Note that in the objective function (34a), the weight of each non-cluster-head user within each cluster is one. The feasible region of problem (34) can be characterized by using Proposition 1, in which we substitute $R^{\min}_{k,n}$ with $S^{\min}_{k,n}$. Finally, the water-filling Alg. 1 can be applied to find the globally optimal solution of (34) in which the optimal $\tilde{q}^*_n$ is given by (33).

It can be shown that similar to proportional fairness, by properly choosing the fairness parameters $\Lambda_{k,n} \geq 1, \ \forall n \in \mathcal{N}, \ k \in \mathcal{K}_n \setminus \{\Phi_n\}$, and $\omega_{\Phi_n} \geq 1, \ \forall n \in \mathcal{N}$, our proposed fairness scheme can also achieve any feasible desired rates for all the users in Hybrid-NOMA, which is important to guarantee any users rate fairness level. Similar to the proportional fairness, it is still difficult to properly assign the weight of the cluster-head users in our proposed fairness scheme denoted by $\omega_{\Phi_n} \geq 1, \ \forall n \in \mathcal{N}$. Another challenge is properly setting $\Lambda_{k,n}$ of each non-cluster-head user $k \in \mathcal{K}_n \setminus \{\Phi_n\}$ prior to resource allocation optimization. This is because, the parameter $S^{\min}_{k,n}, \ \forall k \in \mathcal{K}_n \setminus \{\Phi_n\}$, significantly increases $Q^{\min}_n$ in (6). Hence, significantly large $\Lambda_{k,n}$ for user $k$ may lead to empty feasible region for each subchannel $n \in \mathcal{N}_k, \ k \neq \Phi_n$ (user $k$ is not cluster-head). It is worth noting that for any given $\Lambda_{k,n}$, Corollary 1 is useful to immediately verify whether the feasible region is empty or not. One interesting topic is how to achieve a preferable/absolute users rate fairness by properly choosing the fairness parameters $\Lambda_{k,n} \geq 1, \ \forall n \in \mathcal{N}, \ k \in \mathcal{K}_n \setminus \{\Phi_n\}$, and $\omega_{\Phi_n} \geq 1, \ \forall n \in \mathcal{N}$, in our proposed fairness scheme, which brings new theoretical insights on the fundamental relations between our proposed and the well-known proportional/max-min rate fairness schemes.

### *D. Imperfect Channel State Information*

Unfortunately, it is difficult to acquire the perfect CSI of users, due to the existence of channel estimation errors, feedback delay, and quantization error. In NOMA with imperfect CSI, the imperfect CSI may lead to incorrect user ordering for SIC within a cluster resulting in outage [59], namely SIC outage. By employing the stochastic method, the CNR of user $k \in \mathcal{K}$ on subchannel $n \in \mathcal{N}$ can be modeled as $h^n_k = \hat{h}^n_k + e^n_k$, where $e^n_k \sim \mathcal{CN}\left(0, \sigma^2_e\right)$ and $\hat{h}^n_k \sim \mathcal{CN}\left(0, 1 - \sigma^2_e\right)$ denote the estimation error normalized by noise and estimated CNR, respectively. Assume that the estimated CNR $\hat{h}^n_k$ and normalized estimation error $e^n_k$ are uncorrelated [35], [37], [59]. In each cluster $n$, by performing user ordering based on $\hat{h}^n_k, \ \forall k \in \mathcal{K}_n$, the SIC outage occurs if and only if there exists at least on user pair $i, j \in \mathcal{K}_n, \ \hat{h}^n_i > \hat{h}^n_j$, while $h^n_i < h^n_j$. Assume that $e^n_k \in [L^n_k, U^n_k], \ \forall n \in \mathcal{N}, \ k \in \mathcal{K}_n$. The SIC outage is thus zero if and only if $\min\limits_{e^n_i \in [L^n_i, U^n_i]} h^n_i \geq \max\limits_{e^n_j \in [L^n_j, U^n_j]} h^n_j, \ \forall n \in \mathcal{N}, \ k \in \mathcal{K}_n, \ \hat{h}^n_i > \hat{h}^n_j$. In the latter condition, $\min\limits_{e^n_i \in [L^n_i, U^n_i]} h^n_i = \hat{h}^n_i + L^n_i$, and $\max\limits_{e^n_j \in [L^n_j, U^n_j]} h^n_j = \hat{h}^n_j + U^n_j$. Therefore, the SIC outage is zero if and only if

$$\hat{h}^n_i + L^n_i \geq \hat{h}^n_j + U^n_j, \quad \forall n \in \mathcal{N}, \ k \in \mathcal{K}_n, \ \hat{h}^n_i > \hat{h}^n_j. \quad (35)$$

The condition (35) states that with imperfect CSI, there exists an additional outage, called SIC outage in SC/MC-NOMA, if for a multiplexed user pair, the best case of the CNR of the weaker user is greater than the worst case of the CNR of the stronger user. The SIC outage depends on the region of normalized estimation errors and estimated CNRs. Thus, the SIC outage cannot be tuned by means of power allocation optimization. The latter result is due to the fact that the SIC decoding order of users in SISO Gaussian BCs is independent of power allocation. The SIC outage probability of the 2-user SC-NOMA system is analyzed in [59]. Although when the condition in (35) is not fulfilled, we cannot achieve the zero-SIC outage by means of power allocation, the zero-SIC outage can be achieved by the user clustering of MC-NOMA, or in general, subchannel allocation. For example, when the lower-bound $L^n_i$ and upper-bound $U^n_j$ in (35) are available, we are able to impose the condition in (35) as a necessary constraint in user clustering problem to achieve the zero-SIC outage which increases the robustness of MC-NOMA. The impact of imposing the condition in (35) in user clustering can be considered as a future work.

The work in [35] provided new analyses for the EE maximization problem of Hybrid-NOMA with imperfect CSI, when the large-scale fading factors are slowly varying, thus they can be estimated perfectly at the BS. According to Subsection III in [35], it is straightforward to show that our analysis is also valid for Hybrid-NOMA with imperfect CSI, and considering per-symbol maximum outage probability and minimum rate constraints. It is worth noting that the imperfect CSI merely impacts on the intra-cluster power allocation, due to the high insensitivity of optimal intra-cluster power allocation to the users exact CNRs for the SR/EE maximization problems discussed in Subsection III-C.1. One important future direction of our work is to evaluate the optimality and robustness of the approximated closed-form of optimal powers in (20) and (21) for the imperfect CSI scenarios.

### *E. Admission Control*

One important application of NOMA is to support massive connectivity in the 5G networks, e.g., IoT use-cases [10]. When the number of users and/or their minimum rate demands increases, the parameter $Q_n^{\min}$ increases leading to tightening the feasible region of the formulated optimization problems characterized by Proposition 1. As a result, for significantly large number of users and/or their minimum rate demands, the feasible region will be empty, and subsequently, the problems will be infeasible. As such, the network cannot support all of the users simultaneously, thus an admission control policy is necessary to support the maximum possible number of users/transmitted symbols on subchannels. There are few works addressing the admission control for the SC/MC-NOMA systems [60]–[63]. The globally optimal admission control policy for the general SC/MC-NOMA systems with individual minimum rate demands is still an open problem. To admit more desired symbols in Hybrid-NOMA while reducing the cellular power consumption, one suboptimal solution is to first calculate the power consumption of each user on each subchannel given by (37). Then, eliminate the subchannel (thus transmitted symbol) for the user which consumes the highest power. After that, recalculate (37) for the updated $\mathcal{K}_n$. The latter steps will be continued until Corollary 1 is fulfilled. One future work can be how to incorporate the closed-form of optimal powers in Propositions 1 and 2 in the admission control policy to admit more users while minimizing the cellular power consumption, or maximizing the admitted users sum-rate, respectively.

### *F. Reconfigurable Intelligent Surfaces-Aided NOMA*

The NOMA technology has been recently integrated with RISs [16], [17]. In RIS-assisted NOMA, the joint power and phase shift allocation is shown to be necessary to achieve the optimal solution of the SR maximization problem [16], [17]. Unfortunately, the optimal joint power and phase shift optimization is intractable, thus many recent works applied the alternate optimization, where we find the optimal powers/phase shifts when the other is given. In general, for any given phase shifts, the RIS-NOMA system, such as the considered model in [18], can be equivalently transformed to a NOMA system with users equivalent channel gains. In this way, it is straightforward to show that all the analyses of power allocation for the SR/EE maximization problems of the pure SC/MC-NOMA system are also valid for an RIS-NOMA system with the given phase shifts, thus the users corresponding equivalent channel gains. For example, the closed-form of optimal powers for the RIS-assisted NOMA system in [18] can be obtained by using Proposition 2 with $N = 1$. From (20) and (21), it can be concluded that in the high-SINR regions of an RIS-NOMA system, the optimal powers are insensitive to the equivalent channel gains, thus phase shifts. Therefore, we expect that the alternate optimization approaches a near-optimal solution with a fast convergence speed in the high-SINR regions of an RIS-NOMA system. The extension of our analysis to an RIS-assisted MC-NOMA system can be considered as a future work.

### *G. Long-Term Resource Allocation*

Similar to most of the related works, we assume a dynamic resource allocation framework, where the allocated powers to the users will be readopted every time slot based on the arrival set of active users, and instantaneous CSI. It is shown that the short-term designs may lead to inferior system performance in a long-term perspective [64]. There are a number of works that addressed the long-term resource allocation optimization in NOMA, e.g., [64]–[66]. In [64], the authors developed the well-known Lyapunov optimization framework to convert the long-term sum-rate maximization problem of SC-NOMA with long-term average and short-term peak power constraints, and per-user maximum rate constraints into a series of online "weighted-sum-rate minus weighted-total power consumption" maximization problem in each time slot. The latter problem can be classified as the power allocation problem for SC-NOMA with proportional fairness. Although there has been some efforts in [64] to further reduce the searching space of optimal power allocation, the closed-form expression of optimal power allocation for the long-term optimization framework in [64] is still an open problem. The analysis will be more complicated if we consider the Hybrid-NOMA scheme with per-user/symbol minimum rate constraints, and optimal inter-cluster power allocation, which is still an open problem, and can be considered as a future work.

In [65], the long-term optimization is addressed by properly choosing the users weights in the proportional fairness scheme. In particular, the proportional fairness scheduler keeps track of the average rate of each user in the past time slots with limited length, and reflect these average rates to the users weights. A similar framework can also be applied to our proposed mixed weighted SR/weighted minimum rate fairness scheme in Subsection IV-C.2, where the fairness parameters $\Lambda_{k,n} \geq 1, \ \forall n \in \mathcal{N}, \ k \in \mathcal{K}_n \setminus \{\Phi_n\}$, and $\omega_{\Phi_n} \geq 1, \ \forall n \in \mathcal{N}$, are chosen in (34) based on the average users rate in the past time slots, which can be considered as a future work.

TABLE V
SYSTEM PARAMETERS

| Parameter | Value |
|---|---|
| BS maximum transmit power ($P^{\max}$) | 46 dBm |
| Circuit power consumption ($P_{\mathrm{C}}$) | 30 dBm |
| Coverage of BS | Circular with radii of 500 m |
| Wireless bandwidth ($W$) | 5 MHz |
| Number of users ($K$) | $\{5, 10, 15, \dots, 60\}$ |
| User distribution model | Uniform distribution |
| $U^{\max}$ in NOMA | $\{2, 4, 6\}$ |
| Minimum distance of users to BS | 20 m |
| Distance-depended path loss | $128.1 + 37.6 \log_{10}(d)$ dB, where $d$ is in Km |
| Lognormal shadowing standard deviation | 8 dB |
| Small-scale fading | Rayleigh flat fading |
| AWGN power density | -174 dBm/Hz |
| Minimum rate demand of each user ($R_k^{\min}$) | $\{0.25, 0.5, 0.75, 1, \dots, 5\}$ Mbps |

**Algorithm 5** Suboptimal User Clustering for FD-NOMA

Compute the number of clusters as $N = \lceil K/U^{\max} \rceil$.
Initialize $\rho_k^n = 0,\ \forall n \in \mathcal{N},\ k \in \mathcal{K}_n$, $n = 0$, and ranking vector $\mathbf{R} = [R_k],\ \forall k \in \mathcal{K}$.
**while** $\|\mathbf{R}\| > 0$ **do**
  Find $k^* = \arg\max_{k \in \mathcal{K}} \ \mathbf{R}$.
  Set $n := n + 1$.
  **if** $n > N$ **then**
    Set $n = 1$.
  **end if**
  Set $\rho_{k^*}^n = 1$, and $R_{k^*} = 0$.
**end while**

## V. SIMULATION RESULTS

In this section, we evaluate the performance of SC-NOMA, FDMA-NOMA (FD-NOMA) with different $U^{\max}$, and FDMA for different performance metrics as outage probability, BSs minimum power consumption to satisfy users minimum rate demands, maximum users SR, and maximum system EE. To reflect the randomness impact, we apply the Monte-Carlo simulations [27]–[37] by averaging over 50,000 channel realizations. The outage probability is calculated by dividing the number of infeasible solutions determined according to Corollary 1, by total number of channel realizations. According to Proposition 1, the minimum BS's power consumption can be obtained by $P_{\min} = \sum_{n \in \mathcal{N}} Q_n^{\min}$. All the algorithms in Table III can globally solve the SR maximization problem, however with different computational complexities. For our simulations, we select Alg. 1 with the lowest complexity compared to the others. Moreover, all the mentioned algorithms in Table IV can optimally solve the EE maximization problem with different computational complexities. For our simulations, we select Alg. 2 with inner Alg. 3 which has the lowest complexity compared to the others. Since SC-NOMA and FDMA are special cases of FDMA-NOMA, our selected algorithms are modified to optimally solve these problems. The simulation settings are shown in Table V. Without loss of generality, we set $P_n^{\mathrm{mask}} = P^{\max},\ \forall n \in \mathcal{N}$. In our simulations, we apply a fast suboptimal user clustering method for the flat fading channels of FD-NOMA presented in Alg. 5. In this method, we first obtain $N = \lceil K/U^{\max} \rceil$ according to Proposition 6. The ranking vector $\mathbf{R} = [R_k],\ \forall k \in \mathcal{K}$, is the vector of the ranking of users CNR, in which $R_k \in \{1, \dots, K\},\ \forall k \in \mathcal{K}$, such that $R_k > R_{k'}$ if $h_k^n > h_{k'}^n$. In Alg. 5, the first $N$ users with the highest CNRs are assigned to different clusters. The rest of the users with lower decoding orders are distributed over the subchannels based on their CNRs. The subchannel allocation of FDMA in flat fading channels is straightforward, since any subchannel-to-user allocation is optimal. The source code of the simulations including a user guide is available in [46]. In the following, the term '$X$-NOMA' is referred to as FD-NOMA with $U^{\max} = X$.

### A. System Outage Probability Performance

The impact of minimum rate demands and number of users on the system outage probability of different multiple access techniques is shown in Fig. 3. According to (6), $Q_n^{\min}$ is increasing in $R_k^{\min}$. For quite small $R_k^{\min}$ and/or $K$, the performance gap between different multiple access techniques is low. For larger $R_k^{\min}$ and/or $K$, we observe a significant performance gap between FDMA and $X$-NOMA ($X \geq 2$), and also between 2-NOMA and 4-NOMA. Moreover, it can be observed that the performance gap between 4-NOMA and 6-NOMA is low. Finally, for quite large $R_k^{\min}$ and/or $K$, the outage probability of all these techniques tends to 1. In summary, the outage probability follows: outage(SC-NOMA) $<$ outage(6-NOMA) $\approx$ outage(4-NOMA) $<$ outage(2-NOMA) $\ll$ outage(FDMA).

### B. Average Minimum BS's Power Consumption Performance

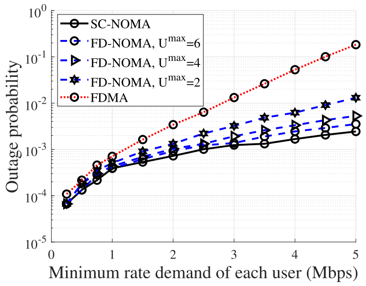
The impact of minimum rate demands and number of users on average total power consumption of different multiple access techniques is shown in Fig. 4. As can be seen, there exists a significant performance gap between FDMA and FD-NOMA for larger $R_k^{\min}$ and/or $K$. However, the performance gap between $X$-NOMA and $(X+1)$-NOMA is highly decreasing for $X \geq 4$. The latter performance gaps are highly increasing in $R_k^{\min}$ and $K$.

### C. Average Users Sum-Rate Performance

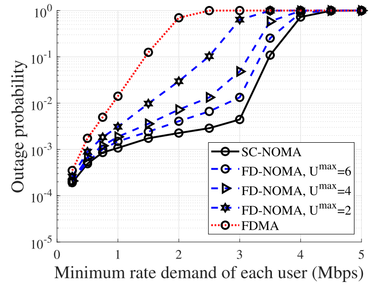
The impact of minimum rate demands and number of users on the average SR of different multiple access techniques is shown in Fig. 5. For the case that outage occurs, the SR is set to zero. The results in Figs. 5(a)-5(c) show that the SR of users is highly insensitive to the minimum rate demands when $R_k^{\min}$ and $K$ are significantly low, specifically for SC-NOMA and FD-NOMA. For significantly high $R_k^{\min}$ and/or $K$, we observe that the average SR decreases, due to increasing the outage probability shown in Fig. 3(a)-3(c). Besides, Figs. 5(d)-5(f) show that SC-NOMA well exploits the multiuser diversity, specifically for lower $R_k^{\min}$. In summary, the SR follows: SR(SC-NOMA) $>$ SR(6-NOMA) $\approx$ SR(4-NOMA) $>$ SR(2-NOMA) $\gg$ SR(FDMA).

### D. Average System Energy Efficiency Performance

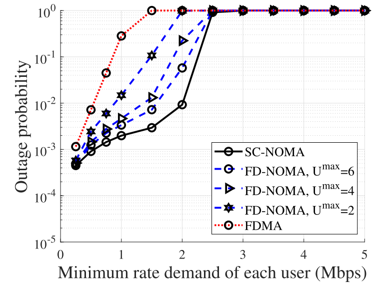
The impact of minimum rate demands and number of users on the average system EE of different multiple access



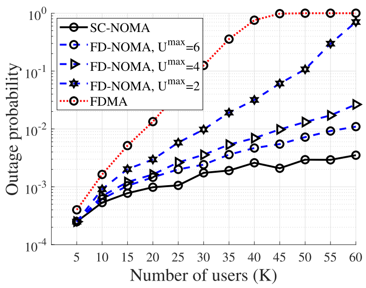

(a) Outage probability vs. users minimum rate demand for $K = 10$.



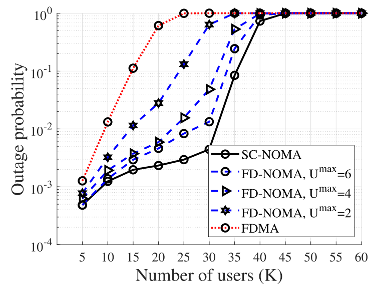

(b) Outage probability vs. users minimum rate demand for $K = 30$.



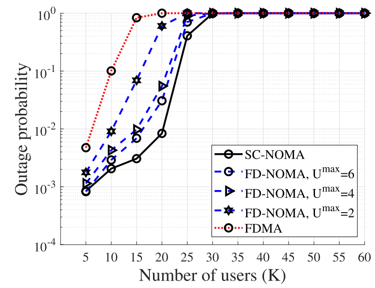

(c) Outage probability vs. users minimum rate demand for $K = 50$.




(d) Outage probability vs. number of users for $R_k^{\min} = 1.5$ Mbps, $\forall k \in \mathcal{K}$.

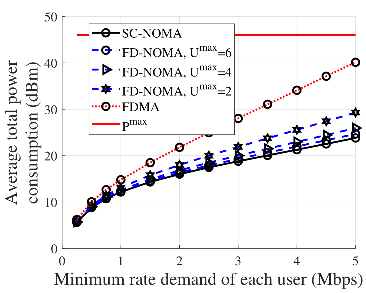



(e) Outage probability vs. number of users for $R_k^{\min} = 3$ Mbps, $\forall k \in \mathcal{K}$.

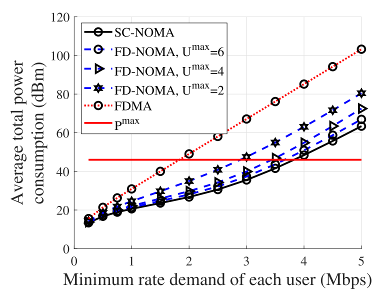



(f) Outage probability vs. number of users for $R_k^{\min} = 4.5$ Mbps, $\forall k \in \mathcal{K}$.

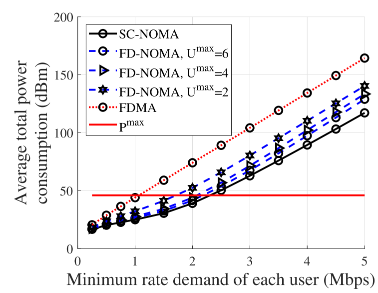
Fig. 3. Impact of the minimum rate demand and number of users on the outage probability of SC-NOMA, FD-NOMA, and FDMA.



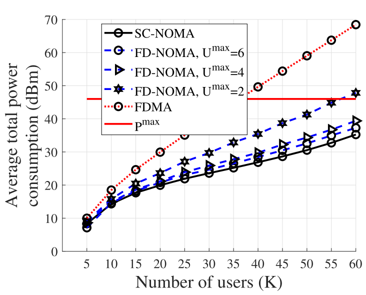

(a) Average total power consumption vs. users minimum rate demand for $K = 10$.



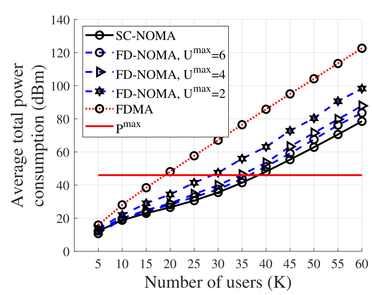

(b) Average total power consumption vs. users minimum rate demand for $K = 30$.



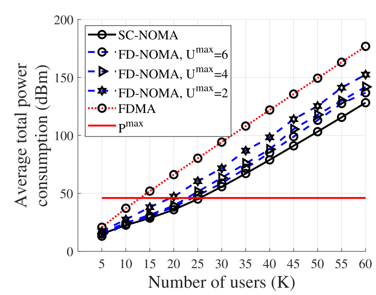

(c) Average total power consumption vs. users minimum rate demand for $K = 50$.




(d) Average total power consumption vs. number of users for $R_k^{\min} = 1.5$ Mbps, $\forall k \in \mathcal{K}$.



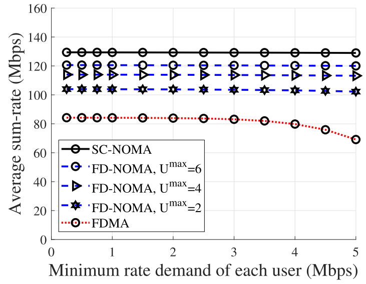

(e) Average total power consumption vs. number of users for $R_k^{\min} = 3$ Mbps, $\forall k \in \mathcal{K}$.



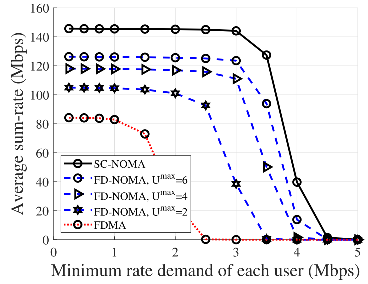

(f) Average total power consumption vs. number of users for $R_k^{\min} = 4.5$ Mbps, $\forall k \in \mathcal{K}$.

Fig. 4. Impact of the minimum rate demand and number of users on the average total power consumption of SC-NOMA, FD-NOMA, and FDMA.

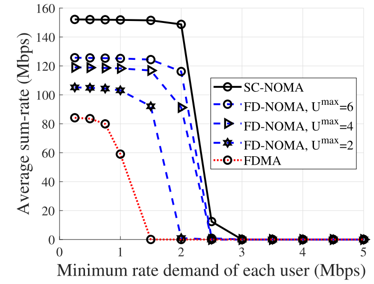
techniques is shown in Fig. 6. From Figs. 6(a)-6(c), we observe that the system EE is affected by $R_k^{\min}$ although the users SR are approximately insensitive to $R_k^{\min}$ shown in Figs. 5(a)-5(c). The main reason that EE is more affected by $R_k^{\min}$ compared to SR is the high sensitivity level of total power consumption to $R_k^{\min}$ shown in Figs. 4(a)-4(c). The impact of total power consumption on EE is highly affected by the circuit power consumption. It can be shown that when $P_{\mathrm{C}}$ increases, the system EE will be more insensitive to $R_k^{\min}$. From Figs. 6(d)-6(f), we observe that the system EE under



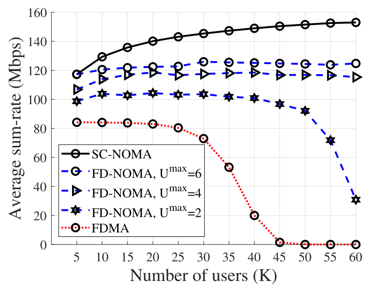

(a) Average sum-rate vs. users minimum rate demand for $K = 10$.



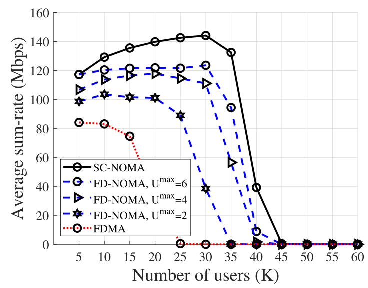

(b) Average sum-rate vs. users minimum rate demand for $K = 30$.



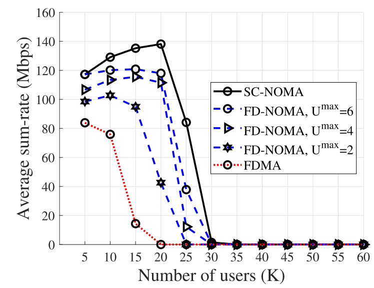

(c) Average sum-rate vs. users minimum rate demand for $K = 50$.




(d) Average sum-rate vs. number of users for $R_k^{\min} = 1.5$ Mbps, $\forall k \in \mathcal{K}$.

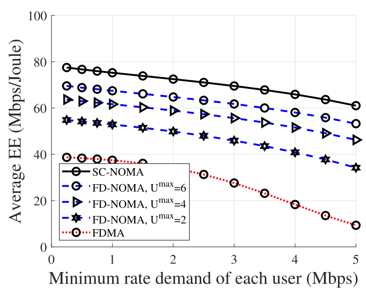



(e) Average sum-rate vs. number of users for $R_k^{\min} = 3$ Mbps, $\forall k \in \mathcal{K}$.

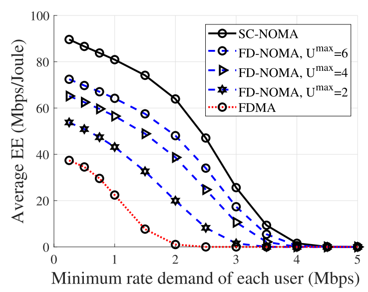



(f) Average sum-rate vs. number of users for $R_k^{\min} = 4.5$ Mbps, $\forall k \in \mathcal{K}$.

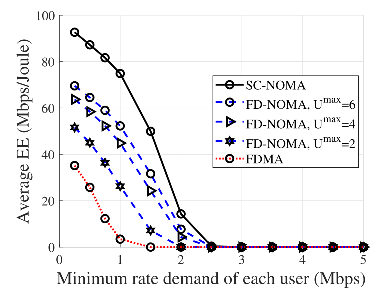
Fig. 5. Impact of the minimum rate demand and number of users on the average sum-rate of SC-NOMA, FD-NOMA, and FDMA.



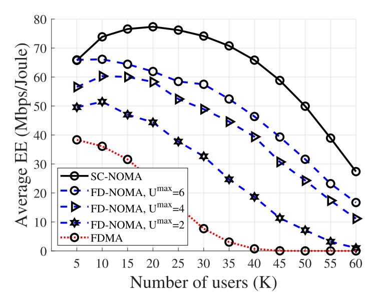

(a) Average EE vs. users minimum rate demand for $K = 10$.



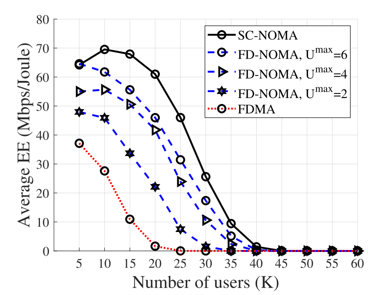

(b) Average EE vs. users minimum rate demand for $K = 30$.



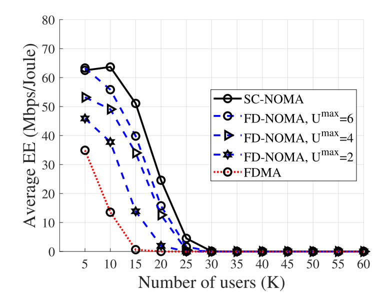

(c) Average EE vs. users minimum rate demand for $K = 50$.




(d) Average EE vs. number of users for $R_k^{\min} = 1.5$ Mbps, $\forall k \in \mathcal{K}$.




(e) Average EE vs. number of users for $R_k^{\min} = 3$ Mbps, $\forall k \in \mathcal{K}$.




(f) Average EE vs. number of users for $R_k^{\min} = 4.5$ Mbps, $\forall k \in \mathcal{K}$.

Fig. 6. Impact of the minimum rate demand and number of users on the average system EE of SC-NOMA, FD-NOMA, and FDMA.

minimum rate demands is increasing with $K$, when $K$ is small enough. In this situation, the system exploits the multiuser diversity, specifically for SC-NOMA. For significantly large $K$, the EE is decreasing with $K$ due to the existing minimum rate demands which highly affects the total power consumption. Following the results of Figs. 4 and 5, the average system EE follows: $\text{EE(SC-NOMA)} > \text{EE(6-NOMA)} \approx \text{EE(4-NOMA)} > \text{EE(2-NOMA)} \gg \text{EE(FDMA)}$.

## VI. Concluding Remarks

In this work, we addressed the problem of finding globally optimal power allocation algorithms to minimize the BSs power consumption, and maximize SR/EE of the general multiuser downlink single-cell Hybrid-NOMA system. For these objectives, we showed that Hybrid-NOMA with $N$ clusters can be equivalently transformed to $N$-user virtual FDMA system, where the effective CNR of each virtual OMA user is obtained in closed form. In this transformation, we exploited the closed-form of optimal powers among multiplexed users within each cluster to further reduce the dimension of our problems as well as increase the accuracy of the iterative convex solvers. In particular, we showed that the feasible region of power allocation in NOMA can be defined as the intersection of closed boxes along with cellular power constraint. Then, we proposed a fast water-filling algorithm for the SR maximization problem, as well as fast iterative algorithms for the EE maximization problem based on the Dinkelbach algorithm with inner Lagrange dual with subgradient method/barrier algorithm with inner Newton's method. The complexity of our proposed algorithms is also analyzed. The possible extensions of our analyses to more general cases with their corresponding new challenges are discussed in the paper. Numerical assessments show that there exist considerable performance gaps in terms of outage probability, BSs power consumption, users SR, and system EE between FDMA and 2-NOMA as well as between 2-NOMA and 4-NOMA. Moreover, we observed that the performance gaps between $X$-NOMA and $(X+1)$-NOMA highly decrease for $X \geq 4$, meaning that when $X \geq 4$, multiplexing more users merely improves the system performance.

## Appendix A
## Proof of Proposition 1

The feasibility of (5) can be determined by solving the power minimization problem as

$$\min_{\boldsymbol{p}\geq 0, \boldsymbol{q}\geq 0} \sum_{n\in\mathcal{N}} q_n \quad \text{s.t. (5b)-(5e).} \tag{36}$$

The problem (36) is also convex with affine objective function and constraints. Accordingly, the weak Slater's condition implies strong duality, thus (36) can be optimally solved by using the Lagrange dual method. For SC-NOMA, in Appendix C of [23], we proved that the maximum power budget does not have any effect on the optimal powers obtained in the power minimization problem when the feasible region is nonempty. Accordingly, problem (36) can be decoupled into $N$ SC-NOMA power minimization subproblems when the feasible region of problem (36) is nonempty. The total power minimization of $M$-user downlink SC-NOMA is solved in Appendix C of [23]. For convenience, consider cluster $n$ with $|\mathcal{K}_n| = K$ users whose CNRs are sorted as $h_1^n < h_2^n < \dots < h_K^n$ with optimal decoding order $K \to K-1 \to \dots \to 1$. As is proved in [23], at the optimal point $\boldsymbol{p}^{*n}$, all the multiplexed users in $\mathcal{K}_n$ get power to only maintain their individual minimal rate demands, meaning that

$$W_s \log_2\left(1 + \frac{p_k^{n*} h_k^n}{1 + \sum\limits_{j=k+1}^{K} p_j^{n*} h_k^n}\right) = R_{k,n}^{\min}, \quad \forall k = 1, \dots, K.$$

The optimal power of each user $k \in \mathcal{K}_n$ (in Watts) can thus be obtained by

$$p_k^{n*} = T_k^n \left(1 + \sum_{j=k+1}^{K} p_j^{n*} h_k^n\right), \quad \forall k = 1, \dots, K,$$

where $T_k^n = \frac{2^{(R_{k,n}^{\min}/W_s)}-1}{h_k^n}$, $\forall k = 1, \dots, K$. Let us rewrite the latter equation as

$$p_k^{n*} = \beta_k^n \left(\frac{1}{h_k^n} + \sum_{j=k+1}^{K} p_j^{n*}\right), \quad \forall k = 1, \dots, K,$$

where $\beta_k^n = 2^{(R_{k,n}^{\min}/W_s)} - 1$, $\forall k = 1, \dots, K$. The latter equation can be rewritten as

$$\begin{aligned}
p_k^{n*} &= \beta_k^n \left(\frac{1}{h_k^n} + \sum_{j=k+1}^{K} p_j^{n*}\right) \\
&= \beta_k^n \left(\frac{1}{h_k^n} + p_{k+1}^{n}{}^{*} + \sum_{j=k+2}^{K} p_j^{n*}\right) \\
&= \beta_k^n \left(\frac{1}{h_k^n} + \beta_{k+1}^n \left(\frac{1}{h_{k+1}^n} + \sum_{j=k+2}^{K} p_j^{n*}\right) + \sum_{j=k+2}^{K} p_j^{n*}\right) \\
&= \beta_k^n \left((1+\beta_{k+1}^n) \sum_{j=k+2}^{K} p_j^{n*} + \frac{1}{h_k^n} + \frac{\beta_{k+1}^n}{h_{k+1}^n}\right) \\
&= \beta_k^n \left((1+\beta_{k+1}^n) \left(p_{k+2}^{n}{}^{*} + \sum_{j=k+3}^{K} p_j^{n*}\right) + \frac{1}{h_k^n} + \frac{\beta_{k+1}^n}{h_{k+1}^n}\right) \\
&= \beta_k^n \Bigg((1+\beta_{k+1}^n) \Bigg(\beta_{k+2}^n \left(\frac{1}{h_{k+2}^n} + \sum_{j=k+3}^{K} p_j^{n*}\right) \\
&\qquad + \sum_{j=k+3}^{K} p_j^{n*}\Bigg) + \frac{1}{h_k^n} + \frac{\beta_{k+1}^n}{h_{k+1}^n}\Bigg) \\
&= \beta_k^n \Bigg((1+\beta_{k+1}^n)(1+\beta_{k+2}^n) \sum_{j=k+3}^{K} p_j^{n*} + \frac{1}{h_k^n} + \frac{\beta_{k+1}^n}{h_{k+1}^n} \\
&\qquad + \frac{\beta_{k+2}^n (1+\beta_{k+1}^n)}{h_{k+2}^n}\Bigg) \\
&\ \ \vdots \\
&= \beta_k^n \Bigg((1+\beta_{k+1}^n)(1+\beta_{k+2}^n)\dots(1+\beta_K^n) + \frac{1}{h_k^n} + \frac{\beta_{k+1}^n}{h_{k+1}^n} \\
&\qquad + \frac{\beta_{k+2}^n (1+\beta_{k+1}^n)}{h_{k+2}^n} \\
&\qquad + \dots + \frac{\beta_K^n (1+\beta_{K-1}^n)\dots(1+\beta_{k+1}^n)}{h_K^n}\Bigg).
\end{aligned}$$

As a result, we have

$$p_k^{n*} = \beta_k^n \left( \prod_{j=k+1}^{K} \left(1+\beta_j^n\right) + \frac{1}{h_k^n} + \sum_{j=k+1}^{K} \frac{\beta_j^n \prod_{i=k+1}^{j-1} \left(1+\beta_i^n\right)}{h_j^n} \right), \quad \forall k=1,\dots,K.$$

The latter equation can be rewritten as

$$p_k^{n*} = \beta_k^n \left( \prod_{\substack{j\in\mathcal{K}_n \\ h_j^n > h_k^n}} \left(1+\beta_j^n\right) + \frac{1}{h_k} + \sum_{\substack{j\in\mathcal{K}_n \\ h_j^n > h_k^n}} \frac{\beta_j^n \prod_{\substack{l\in\mathcal{K}_n \\ h_k^n < h_l^n < h_j^n}} \left(1+\beta_l^n\right)}{h_j} \right) \quad \forall n\in\mathcal{N}, \quad k\in\mathcal{K}_n. \tag{37}$$

Therefore, the minimum power consumption of cluster $n$ is given by

$$Q_n^{\min} = \sum_{k\in\mathcal{K}_n} p_k^{n*} = \sum_{k\in\mathcal{K}_n} \beta_k^n \left( \prod_{\substack{j\in\mathcal{K}_n \\ h_j^n > h_k^n}} \left(1+\beta_j^n\right) + \frac{1}{h_k} + \sum_{\substack{j\in\mathcal{K}_n \\ h_j^n > h_k^n}} \frac{\beta_j^n \prod_{\substack{l\in\mathcal{K}_n \\ h_k^n < h_l^n < h_j^n}} \left(1+\beta_l^n\right)}{h_j} \right), \quad \forall n\in\mathcal{N}. \tag{38}$$

The parameter $Q_n^{\min}$ is indeed the minimum power consumption of cluster $n$ to satisfy the minimum rate demand of users in $\mathcal{K}_n$. As a result, the feasible region of $q_n$ is lower-bounded by $Q_n^{\min}$ for each $n\in\mathcal{N}$. Accordingly, constraints (5b) and (5e) can be combined as $q_n \in \left[Q_n^{\min}, P_n^{\text{mask}}\right]$, $\forall n\in\mathcal{N}$, which is the intersection of affine closed boxes, so is convex. According to (5c), we guarantee that any $q_n \in \left[Q_n^{\min}, P_n^{\text{mask}}\right]$, $\forall n\in\mathcal{N}$, satisfying (5c) is feasible, and the proof is completed.

## Appendix B
## Proof of Proposition 2

For any given feasible $\boldsymbol{q}$, (5c) and (5e) can be removed from (5). Then, problem (5) can be equivalently divided into $N$ SC-NOMA subproblems. For each subproblem $n$, we find the intra-cluster power allocation among multiplexed users in $\mathcal{K}_n$ in closed-form. According to the KKT optimality conditions analysis in Appendix B of [23], it is proved that at the optimal point of SC-NOMA with CNR-based decoding order, only the cluster-head user gets additional power, and all the other users get power to only maintain their minimal rate demands.

For convenience, consider cluster $n\in\mathcal{N}$ with $|\mathcal{K}_n| = K$ users whose CNRs are sorted as $h_1^n < h_2^n < \dots < h_K^n$ with optimal decoding order $K \to K-1 \to \dots \to 1$. According to Appendix B of [23], the optimal powers in $\boldsymbol{p}^{*n}$ satisfy

$$W_s \log_2 \left( 1 + \frac{p_k^{n*} h_k^n}{1 + \sum_{j=k+1}^{K} p_j^{n*} h_k^n} \right) = R_{k,n}^{\min}, \quad \forall k=1,\dots,K-1, \tag{39}$$

and

$$p_K^{n*} = q_n - \sum_{k=1}^{K-1} p_k^{n*}. \tag{40}$$

Let us rewrite (39) as

$$p_k^{n*} = \frac{T_k^n \left( 1 + \left( q_n - \sum_{j=1}^{k-1} p_j^{n*} \right) h_k^n \right)}{1 + T_k^n h_k^n}, \quad \forall k=1,\dots,K-1,$$

where $T_k^n = \frac{2^{(R_{k,n}^{\min}/W_s)}-1}{h_k^n}$, $\forall k=1,\dots,K-1$. To find a closed-form expression for $\boldsymbol{p}^{*n}$, we rewire the latter equation as

$$p_k^{n*} = \beta_k^n \left( q_n - \sum_{j=1}^{k-1} p_j^{n*} + \frac{1}{h_k^n} \right), \quad \forall k=1,\dots,K-1,$$

where $\beta_k^n = \frac{2^{(R_{k,n}^{\min}/W_s)}-1}{2^{(R_{k,n}^{\min}/W_s)}}$, $\forall k=1,\dots,K-1$. The latter equation can also be rewritten as

$$\begin{aligned}
&p_k^{n*} \\
&= \beta_k^n \left( q_n - p_{k-1}^{n\,*} - \sum_{j=1}^{k-2} p_j^{n*} + \frac{1}{h_k^n} \right) \\
&= \beta_k^n \left( q_n - \beta_{k-1}^n \left( q_n - \sum_{j=1}^{k-2} p_j^{n*} + \frac{1}{h_{k-1}^n} \right) - \sum_{j=1}^{k-2} p_j^{n*} + \frac{1}{h_k^n} \right) \\
&= \beta_k^n \left( \left(1-\beta_{k-1}^n\right) q_n - \left(1-\beta_{k-1}^n\right) \sum_{j=1}^{k-2} p_j^{n*} + \frac{1}{h_k^n} - \frac{\beta_{k-1}^n}{h_{k-1}^n} \right) \\
&\;\;\vdots \\
&= \beta_k^n \Bigg( \left(1-\beta_{k-1}^n\right) \left(1-\beta_{k-2}^n\right) \dots \left(1-\beta_1^n\right) q_n + \frac{1}{h_k^n} - \frac{\beta_{k-1}^n}{h_{k-1}^n} \\
&\quad + \dots - \frac{\left(1-\beta_{k-1}^n\right) \beta_{k-2}^n}{h_2^n} \\
&\quad - \frac{\left(1-\beta_{k-1}^n\right) \left(1-\beta_{k-2}^n\right) \dots \left(1-\beta_2^n\right) \beta_1^n}{h_1^n} \Bigg).
\end{aligned}$$

According to the above, we have

$$p_k^{n*} = \beta_k^n \left( \prod_{j=1}^{k-1} \left(1-\beta_j^n\right) q_n + \frac{1}{h_k^n} - \sum_{j=1}^{k-1} \frac{\beta_j^n \prod_{i=j+1}^{k-1} \left(1-\beta_i^n\right)}{h_j^n} \right), \quad \forall k=1,\dots,K-1. \tag{41}$$

The optimal powers in (41) can be reformulated as

$$p_k^{n*} = \left( \beta_k^n \prod_{\substack{j \in \mathcal{K}_n \\ h_j^n < h_k^n}} \left(1 - \beta_j^n\right) \right) q_n + c_k^n,$$
$$\forall n \in \mathcal{N}, \quad k \in \mathcal{K}_n \setminus \{\Phi_n\},$$

where $\beta_k^n = \frac{2^{(R_{k,n}^{\min}/W_s)}-1}{2^{(R_{k,n}^{\min}/W_s)}}, \forall n \in \mathcal{N}, k \in \mathcal{K}_n$, $c_k^n = \beta_k^n \left( \frac{1}{h_k^n} - \sum\limits_{\substack{j \in \mathcal{K}_n \\ h_j^n < h_k^n}} \frac{\prod\limits_{\substack{l \in \mathcal{K}_n \\ h_j^n < h_l^n < h_k^n}} (1-\beta_l^n)\beta_j^n}{h_j^n} \right), \forall n \in \mathcal{N}, k \in \mathcal{K}_n$. Subsequently, based on (40), the optimal power of the cluster-head users can be formulated by

$$p_{\Phi_n}^n{}^* = \left( 1 - \sum_{\substack{i \in \mathcal{K}_n \\ h_i^n < h_{\Phi_n}^n}} \beta_i^n \prod_{\substack{j \in \mathcal{K}_n \\ h_j^n < h_i^n}} \left(1 - \beta_j^n\right) \right) q_n - \sum_{\substack{i \in \mathcal{K}_n \\ h_i^n < h_{\Phi_n}^n}} c_i^n,$$
$$\forall n \in \mathcal{N}.$$

## Appendix C
## Water-Filling Algorithm for Solving (12)

The Lagrange function of (12) is given by

$$L(\tilde{\boldsymbol{q}}, \nu) = \sum_{n \in \mathcal{N}} W_s \log_2\left(1 + \tilde{q}_n H_n\right) + \nu \left( \tilde{P}^{\max} - \sum_{n \in \mathcal{N}} \tilde{q}_n \right), \quad (42)$$

where $\nu$ is the Lagrange multiplier for the cellular power constraint (12b), and $q_n \in [\tilde{Q}_n^{\min}, \tilde{P}_n^{\text{mask}}]$, $\forall n \in \mathcal{N}$. The Lagrange dual function is

$$\begin{aligned} g(\nu) &= \sup_{\tilde{\boldsymbol{q}} \in \mathcal{P}} L(\tilde{\boldsymbol{q}}, \nu) \\ &= \sup_{\tilde{\boldsymbol{q}} \in \mathcal{P}} \Bigg\{ \sum_{n \in \mathcal{N}} W_s \log_2\left(1 + \tilde{q}_n H_n\right) \\ &\qquad + \nu \left( \tilde{P}^{\max} - \sum_{n \in \mathcal{N}} \tilde{q}_n \right) \Bigg\}, \end{aligned} \quad (43)$$

where $\mathcal{P}$ is the feasible set of problem (12). The Lagrange dual problem is formulated by

$$\min_{\nu} \; g(\nu), \quad \text{s.t. } \nu \in \mathbb{R}. \quad (44)$$

Assume that $\nu^*$ is the dual optimal. Moreover, $\tilde{\boldsymbol{q}}^* = [\tilde{q}_n^*], \forall n \in \mathcal{N}$, is primal. The KKT conditions are listed below

$$\begin{aligned} &\text{C1}: q_n \in [\tilde{Q}_n^{\min}, \tilde{P}_n^{\text{mask}}], \; \forall n \in \mathcal{N}, \\ &\text{C2}: \tilde{P}^{\max} - \sum_{n \in \mathcal{N}} \tilde{q}_n^* = 0, \\ &\text{C3}: \nabla_{\tilde{\boldsymbol{q}}^*} L(\tilde{\boldsymbol{q}}^*, \nu^*) = 0. \end{aligned}$$

Condition C3 can be rewritten as $\frac{W_s H_n / \ln 2}{1 + \tilde{q}_n^* H_n} - \nu^* = 0, \; \forall n \in \mathcal{N}$. Summing-up, for each $n \in \mathcal{N}$, we have

$$\tilde{q}_n^* = \begin{cases} \dfrac{W_s/(\ln 2)}{\nu^*} - \dfrac{1}{H_n}, & \left( \dfrac{W_s/(\ln 2)}{\nu^*} - \dfrac{1}{H_n} \right) \\ & \in [\tilde{Q}_n^{\min}, \tilde{P}_n^{\text{mask}}]; \\ 0, & \text{otherwise.} \end{cases} \quad (45)$$

To ease of convenience, we reformulate (45) as $\tilde{q}_n^* = \max\left\{ \tilde{Q}_n^{\min}, \min\left\{ \left( \frac{W_s/(\ln 2)}{\nu^*} - \frac{1}{H_n} \right), \tilde{P}_n^{\text{mask}} \right\} \right\}$. By substituting $\tilde{q}_n^*$ to the cellular power constraint (12b), we have

$$\sum_{n \in \mathcal{N}} \max\left\{ \tilde{Q}_n^{\min}, \min\left\{ \left( \frac{W_s/(\ln 2)}{\nu^*} - \frac{1}{H_n} \right), \tilde{P}_n^{\text{mask}} \right\} \right\} = \tilde{P}^{\max}. \quad (46)$$

The left-hand side is a piecewise-linear increasing function of $\frac{W_s/(\ln 2)}{\nu^*}$ with breakpoints at $\frac{1}{H_n}$, $\forall n \in \mathcal{N}$, so the equation has a unique solution which is readily determined. To find optimal $\nu^*$, we first initialize tolerance $\epsilon$, lower-bound $\nu_l$ and upper-bound $\nu_h$. The lower-bound $\nu_l$ should satisfy $\sum\limits_{n \in \mathcal{N}} \max\left\{ \tilde{Q}_n^{\min}, \min\left\{ \left( \frac{W_s/(\ln 2)}{\nu_l} - \frac{1}{H_n} \right), \tilde{P}_n^{\text{mask}} \right\} \right\} > \tilde{P}^{\max}$, and the upper-bound $\nu_h$ should satisfy $\sum\limits_{n \in \mathcal{N}} \max\left\{ \tilde{Q}_n^{\min}, \min\left\{ \left( \frac{W_s/(\ln 2)}{\nu_h} - \frac{1}{H_n} \right), \tilde{P}_n^{\text{mask}} \right\} \right\} < \tilde{P}^{\max}$. After the initialization step, we apply the bisection method to find $\nu^*$ presented in Alg. 1.

## Appendix D
## Proof of the Optimality of Alg. 2

Let us define the EE function as $E(\boldsymbol{p}) = \frac{f_1(\boldsymbol{p})}{f_2(\boldsymbol{p})}$, $\forall \boldsymbol{p} \in \mathcal{P}$, where $f_1(\boldsymbol{p}) = \sum\limits_{n \in \mathcal{N}} \sum\limits_{k \in \mathcal{K}_n} R_k^n(\boldsymbol{p}^n)$, $f_2(\boldsymbol{p}) = \sum\limits_{n \in \mathcal{N}} \sum\limits_{k \in \mathcal{K}_n} p_k^n + P_{\text{C}}$, and $\mathcal{P}$ denotes the feasible set of problem (4). In this formulation, $f_1(\boldsymbol{p})$ is concave, and $f_2(\boldsymbol{p})$ is affine, so is convex. Moreover, both $f_1$ and $f_2$ are differentiable. The feasible set $\mathcal{P}$ can be characterized by using Proposition 1 which is shown to be affine, so is convex. For any nonempty $\mathcal{P}$ (which can be determined by Corollary 1), the objective function $E(\boldsymbol{p}) = \frac{f_1(\boldsymbol{p})}{f_2(\boldsymbol{p})}$ is pseudoconcave, implying that any stationary point is indeed global maximum and the KKT conditions are sufficient if a constraint qualification is fulfilled [24], [25]. Therefore, the globally optimal solution of problem (4) can be obtained by using convex optimization algorithms [24], [25]. In particular, (4) can be equivalently transformed to the following problem as

$$\max_{\boldsymbol{p} \in \mathcal{P}, \lambda \in \mathbb{R}} \; \lambda \quad \text{s.t. } \frac{f_1(\boldsymbol{p})}{f_2(\boldsymbol{p})} - \lambda \geq 0,$$

which can be rewritten as

$$\max_{\boldsymbol{p} \in \mathcal{P}, \lambda \in \mathbb{R}} \; \lambda \quad \text{s.t. } f_1(\boldsymbol{p}) - \lambda f_2(\boldsymbol{p}) \geq 0.$$

It can be proved that solving the latter problem is equivalent to finding the root of the following nonlinear function [25]

$$F(\lambda) = \max_{\boldsymbol{p} \in \mathcal{P}} \; f_1(\boldsymbol{p}) - \lambda f_2(\boldsymbol{p}),$$

so the condition for the global optimality is

$$F(\lambda^*) = \max_{\boldsymbol{p} \in \mathcal{P}} \; f_1(\boldsymbol{p}) - \lambda^* f_2(\boldsymbol{p}) = 0.$$

Various methods can find the root of $F(\lambda)$, such as the Dinkelbach algorithm [67] which is based on the Newton's method. For more details, please see Proposition 3.2 in [24].

## APPENDIX E
## LAGRANGE DUAL WITH SUBGRADIENT METHOD FOR SOLVING (19)

The Lagrange function of (19) is formulated by

$$L(\tilde{\boldsymbol{q}},\nu)=\sum_{n\in\mathcal{N}} W_s\log_2\left(1+\tilde{q}_nH_n\right)-\lambda\left(\sum_{n\in\mathcal{N}}\tilde{q}_n\right)+\nu\left(\tilde{P}^{\max}-\sum_{n\in\mathcal{N}}\tilde{q}_n\right),\quad(47)$$

where $\nu$ is the Lagrange multiplier for the cellular power constraint (19b), and $q_n\in[\tilde{Q}_n^{\min},\tilde{P}_n^{\mathrm{mask}}],\ \forall n\in\mathcal{N}$. The dual function is given by

$$g(\nu)=\sup_{\tilde{\boldsymbol{q}}\in\mathcal{P}}L(\tilde{\boldsymbol{q}},\nu)=\sup_{\tilde{\boldsymbol{q}}\in\mathcal{P}}\left\{\sum_{n\in\mathcal{N}}W_s\log_2\left(1+\tilde{q}_nH_n\right)-\lambda\left(\sum_{n\in\mathcal{N}}\tilde{q}_n\right)+\nu\left(\tilde{P}^{\max}-\sum_{n\in\mathcal{N}}\tilde{q}_n\right)\right\},\quad(48)$$

where $\mathcal{P}$ is the feasible domain of problem (12). The Lagrange dual problem is formulated by

$$\min_{\nu}\ g(\nu),\quad \text{s.t. } \nu\in\mathbb{R}.\quad(49)$$

The optimal $\tilde{\boldsymbol{q}}^*$ can be obtained by $\nabla_{\tilde{\boldsymbol{q}}}L(\tilde{\boldsymbol{q}},\nu)=0$. Then, we have

$$\tilde{q}_n^*=\left[\frac{W_s/(\ln 2)}{\lambda+\nu^*}-\frac{1}{H_n}\right]_{\tilde{Q}_n^{\min}}^{\tilde{P}_n^{\mathrm{mask}}},\quad n\in\mathcal{N},\quad(50)$$

where $\nu^*$ is the dual optimal, which can be obtained by using the subgradient method [47]. In this algorithm, we iteratively update $\nu$ such that at iteration $(t+1)$

$$\nu^{(t+1)}=\left[\nu^{(t)}-\epsilon_{\mathrm{s}}\left(\tilde{P}^{\max}-\sum_{n\in\mathcal{N}}\tilde{q}_n^{(t)}\right)\right]^+,\quad(51)$$

where $\nu^{(t)}$ is the Lagrange multiplier $\nu$ at iteration $t$, and $\tilde{q}_n^{(t)}$ is the optimal solution obtained by (50) at iteration $t$. Moreover, $\epsilon_{\mathrm{s}}>0$ is the step size tuning the accuracy of the algorithm [68]. The iterations are repeated until the convergence is achieved. It is verified that the subgradient method will converge to the globally optimal solution after few iterations [68].

## APPENDIX F
## BARRIER ALGORITHM WITH INNER NEWTON'S METHOD FOR SOLVING (19)

Let us reformulate (19) as the following standard convex problem

$$\min_{\tilde{\boldsymbol{q}}}\ f_0(\tilde{\boldsymbol{q}})=-\sum_{n\in\mathcal{N}}W_s\log_2\left(1+\tilde{q}_nH_n\right)+\lambda\left(\sum_{n\in\mathcal{N}}\tilde{q}_n\right)\quad(52a)$$

$$\text{s.t. } f_1(\tilde{\boldsymbol{q}})=\sum_{n\in\mathcal{N}}\tilde{q}_n-\tilde{P}^{\max}\le 0,\quad(52b)$$

$$q_n\in[\tilde{Q}_n^{\min},\tilde{P}_n^{\mathrm{mask}}],\quad \forall n\in\mathcal{N}.\quad(52c)$$

Then, we approximate (52) to an unconstrained minimization problem as

$$\min_{\tilde{\boldsymbol{q}}}\ U(\tilde{\boldsymbol{q}})=tf_0(\tilde{\boldsymbol{q}})+\phi(\tilde{\boldsymbol{q}}),\quad(53)$$

in which

$$\phi(\tilde{\boldsymbol{q}})=-\log\left(-f_1(\tilde{\boldsymbol{q}})\right),$$

such that the domain of $\phi$ is

$$\textbf{dom}\ \phi=\{q_n\in[\tilde{Q}_n^{\min},\tilde{P}_n^{\mathrm{mask}}],\quad \forall n\in\mathcal{N}|f_1(\tilde{\boldsymbol{q}})<0\},$$

and $t\gg 1$ is a positive real constant. The problem (53) is convex since $tf_0(\tilde{\boldsymbol{q}})$ and $\phi(\tilde{\boldsymbol{q}})$ are convex. In each barrier iteration, we solve (53) by using the Newton's method. The gradient of $U(\tilde{\boldsymbol{q}})$ is formulated by

$$\nabla_{\tilde{\boldsymbol{q}}}U(\tilde{\boldsymbol{q}})=t\nabla_{\tilde{\boldsymbol{q}}}f_0(\tilde{\boldsymbol{q}})+\nabla_{\tilde{\boldsymbol{q}}}\phi(\tilde{\boldsymbol{q}}),$$

where $\nabla_{\tilde{\boldsymbol{q}}}f_0(\tilde{\boldsymbol{q}})=\left[\frac{\partial f_0}{\partial\tilde{q}_n}\right],\ \forall n\in\mathcal{N}$, in which

$$\frac{\partial f_0(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n}=-\frac{W_sH_n}{\ln(2)\left(1+\tilde{q}_nH_n\right)}+\lambda,\quad \forall n\in\mathcal{N}.$$

In addition, $\nabla_{\tilde{\boldsymbol{q}}}\phi(\tilde{\boldsymbol{q}})=\left[\frac{\partial\phi(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n}\right],\ \forall n\in\mathcal{N}$, in which $\frac{\partial\phi(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n}=-\frac{\frac{\partial f_1(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n}}{f_1(\tilde{\boldsymbol{q}})},\ \forall n\in\mathcal{N}$, such that $\frac{\partial f_1(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n}=1,\ \forall n\in\mathcal{N}$. Therefore, we have $\frac{\partial\phi(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n}=-\frac{1}{\sum_{n\in\mathcal{N}}\tilde{q}_n-\tilde{P}^{\max}}$. Summing up, the $n$-th element in the vector $\nabla_{\tilde{\boldsymbol{q}}}U(\tilde{\boldsymbol{q}})=\left[\frac{\partial U(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n}\right],\ \forall n\in\mathcal{N}$, is given by

$$\frac{\partial U(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n}=-t\left(\frac{W_sH_n}{\ln(2)\left(1+\tilde{q}_nH_n\right)}+\lambda\right)-\frac{1}{\sum_{n\in\mathcal{N}}\tilde{q}_n-\tilde{P}^{\max}},\quad \forall n\in\mathcal{N}.$$

The Hessian of $U(\tilde{\boldsymbol{q}})$ is formulated by

$$\nabla^2_{\tilde{\boldsymbol{q}}}U(\tilde{\boldsymbol{q}})=t\nabla^2_{\tilde{\boldsymbol{q}}}f_0(\tilde{\boldsymbol{q}})+\nabla^2_{\tilde{\boldsymbol{q}}}\phi(\tilde{\boldsymbol{q}}),$$

where $\nabla^2_{\tilde{\boldsymbol{q}}}f_0(\tilde{\boldsymbol{q}})=\left[\frac{\partial^2 f_0}{\partial\tilde{q}_i\partial\tilde{q}_j}\right],\ \forall i,j\in\mathcal{N}$, such that its entries are $\frac{\partial^2 f_0(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n^2}=\frac{W_s}{\ln(2)}\frac{H_n^2}{(1+\tilde{q}_nH_n)^2},\ \forall n\in\mathcal{N}$, and $\frac{\partial^2 f_0(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_i\tilde{q}_j}=0,\ \forall i,j\in\mathcal{N},\ i\neq j$. As a result, we have

$$\nabla^2_{\tilde{\boldsymbol{q}}}f_0(\tilde{\boldsymbol{q}})=\textbf{diag}\left(\left[\frac{\partial^2 f_0(\tilde{\boldsymbol{q}})}{\partial\tilde{q}_n^2}\right],\quad \forall n\in\mathcal{N}\right),$$

which is positive definite, since each element $\frac{\partial^2 f_0}{\partial\tilde{q}_n^2}$ in the main diagonal of $\nabla^2_{\tilde{\boldsymbol{q}}}f_0(\tilde{\boldsymbol{q}})$ is positive and the others are zero. The Hessian of $\phi(\tilde{\boldsymbol{q}})$ can be obtained by $\nabla^2_{\tilde{\boldsymbol{q}}}\phi(\tilde{\boldsymbol{q}})=\frac{1}{\left(\sum_{n\in\mathcal{N}}\tilde{q}_n-\tilde{P}^{\max}\right)^2}\mathbf{1}_{N\times N}$. The eigenvector of $\phi(\tilde{\boldsymbol{q}})$ is $\left[0,0,\ldots,0,\frac{1}{\left(\sum_{n\in\mathcal{N}}\tilde{q}_n-\tilde{P}^{\max}\right)^2}\right]_{1\times N}$. Then, we conclude that $\nabla^2_{\tilde{\boldsymbol{q}}}\phi(\tilde{\boldsymbol{q}})\succeq 0$. Finally, we have $\nabla^2_{\tilde{\boldsymbol{q}}}U(\tilde{\boldsymbol{q}})=t\nabla^2_{\tilde{\boldsymbol{q}}}f_0(\tilde{\boldsymbol{q}})+\nabla^2_{\tilde{\boldsymbol{q}}}\phi(\tilde{\boldsymbol{q}})\succ 0$ since $\nabla^2_{\tilde{\boldsymbol{q}}}f_0(\tilde{\boldsymbol{q}})\succ 0$ , $\nabla^2_{\tilde{\boldsymbol{q}}}\phi(\tilde{\boldsymbol{q}})\succeq 0$, and $t>0$. The latter result proves that $U(\tilde{\boldsymbol{q}})$ is strictly convex and its Hessian is nonsingular. Accordingly, $\left(\nabla^2_{\tilde{\boldsymbol{q}}}U(\tilde{\boldsymbol{q}})\right)^{-1}$ is positive and finite. Hence, the barrier method with inner Newton's method achieves an $\epsilon$-suboptimal solution [47].

## APPENDIX G
## OPTIMAL POWER ALLOCATION FOR MAXIMIZING SUM-RATE OF $K$-USER SC-NOMA WITH PER-USER MINIMUM AND MAXIMUM RATE CONSTRAINTS

Consider a $K$-user SC-NOMA system, whose users CNRs are sorted as $h_1 < h_2 < \cdots < h_K$ with optimal decoding order $K \to K-1 \to \cdots \to 1$. Let $M_k = \left(2^{(R_k^{\max}/W_s)} - 1\right)\left(\sum\limits_{j=k+1}^{K} p_j h_k + 1\right)/h_k, \ \forall k = 1,\ldots,K$. It can be shown that the optimal powers can be obtained by first calculating the optimal powers in Proposition 2. Then, we obtain $M_K$ of the strongest user. If $p_K^* \leq M_K$, the obtained powers are the optimal solution. If $p_K^* > M_K$, we set $p_K^* = M_K$. Then, we calculate $M_{K-1}$ and $p_{K-1} = P^{\max} - \left(\sum\limits_{\substack{j=1\\ j\neq K-1}}^{K} p_j\right)$ with the updated $p_K^* = M_K$. If $p_{K-1}^* \leq M_{K-1}$, the obtained powers are the optimal solution. If $p_{K-1}^* > M_{K-1}$, we set $p_{K-1}^* = M_{K-1}$. Then, we calculate $M_{K-2}$ and $p_{K-2} = P^{\max} - \left(\sum\limits_{\substack{j=1\\ j\neq K-2}}^{K} p_j\right)$ with the updated $p_K^* = M_K$, and $p_{K-1}^* = M_{K-1}$. If $p_{K-2}^* \leq M_{K-2}$, the obtained powers are the optimal solution. Otherwise, we continue these series until a user denoted by $i$ satisfies $p_i^* \leq M_i$. Accordingly, the achievable rate of users $K,\ldots,1$ can be obtained, respectively, as

$$\underbrace{R_K^{\max}, R_{K-1}^{\max}, \ldots, R_{i+1}^{\max}}_{\text{Maximum Rates}}, \underbrace{\left[R_i^{\min}, R_i^{\max}\right]}_{\text{Between}}, \underbrace{R_{i-1}^{\min}, \ldots, R_1^{\min}}_{\text{Minimum Rates}}.$$

## REFERENCES


[1] D. Hughes-Hartogs, "The capacity of a degraded spectral Gaussian broadcast channel," Ph.D. dissertation, Inform. Syst. Lab., Ctr. Syst Res., Stanford Univ., Stanford, CA, USA, Jul. 1975.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Hoboken, NJ, USA: Wiley, 2006.

[3] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[4] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels—Part I: Ergodic capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001.

[5] N. Jindal and A. Goldsmith, "Capacity and optimal power allocation for fading broadcast channels with minimum rates," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 2895–2909, Nov. 2003.

[6] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple Access Techniques for 5G Wireless Networks and Beyond*, 1st ed. Cham, Switzerland: Springer, 2018.

[7] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Jun. 2013, pp. 1–5.

[8] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.

[9] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2018.

[10] J. Du *et al.*, "When mobile-edge computing (MEC) meets nonorthogonal multiple access (NOMA) for the Internet of Things (IoT): System design and optimization," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 7849–7862, May 2021.

[11] *Study on Downlink Multiuser Superposition Transmission for LTE*, document 36.859, 3rd Generation Partnership Project (3GPP), Mar. 2015.

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

[13] S. Rezvani, N. M. Yamchi, M. R. Javan, and E. A. Jorswieck, "Resource allocation in virtualized CoMP-NOMA HetNets: Multi-connectivity for joint transmission," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4172–4185, Jun. 2021.

[14] M. Moltafet, R. Joda, N. Mokari, M. R. Sabagh, and M. Zorzi, "Joint access and fronthaul radio resource allocation in PD-NOMA-based 5G networks enabling dual connectivity and CoMP," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6463–6477, Dec. 2018.

[15] S. Rezvani, S. Parsaeefard, N. Mokari, M. R. Javan, and H. Yanikomeroglu, "Cooperative multi-bitrate video caching and transcoding in multicarrier NOMA-assisted heterogeneous virtualized MEC networks," *IEEE Access*, vol. 7, pp. 93511–93536, 2019.

[16] S. Gong *et al.*, "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2283–2314, 4th Quart., 2020.

[17] C. Pan *et al.*, "Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 14–20, Jun. 2021.

[18] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Capacity and optimal resource allocation for IRS-assisted multi-user communication systems," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3771–3786, Jun. 2021.

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

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

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

[22] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 174–180, Oct. 2019.

[23] S. Rezvani, E. A. Jorswieck, N. Mokari, and M. R. Javan, "Optimal SIC ordering and power allocation in downlink multi-cell NOMA systems," *IEEE Trans. Wireless Commun.*, early access, Oct. 2021, doi: 10.1109/TWC.2021.3120325.

[24] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Found. Trends Commun. Inf. Theory*, vol. 11, nos. 3–4, pp. 185–396, 2015.

[25] C. Isheden, Z. Chong, E. Jorswieck, and G. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2946–2957, Aug. 2012.

[26] W. U. Khan, F. Jameel, T. Ristaniemi, S. Khan, G. A. S. Sidhu, and J. Liu, "Joint spectral and energy efficiency optimization for downlink NOMA networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 2, pp. 645–656, Jun. 2020.

[27] 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.

[28] L. Salaun, C. S. Chen, and M. Coupechoux, "Optimal joint subcarrier and power allocation in NOMA is strongly NP-hard," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[29] L. Salaun, M. Coupechoux, and C. S. Chen, "Joint subcarrier and power allocation in NOMA: Optimal and approximate algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 2215–2230, 2020.

[30] 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.

[31] Z. Yang, C. Pan, W. Xu, Y. Pan, M. Chen, and M. Elkashlan, "Power control for multi-cell networks with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 927–942, Feb. 2018.

[32] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.

[33] 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.

[34] P. Gupta and D. Ghosh, "User fairness based energy efficient power allocation for downlink cellular NOMA system," in *Proc. 5th Int. Conf. Comput., Commun. Secur. (ICCCS)*, Oct. 2020, pp. 1–5.

[35] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. M. Leung, "Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2874–2885, Dec. 2017.

[36] F. Fang, J. Cheng, Z. Ding, and H. V. Poor, "Energy efficient resource optimization for a downlink NOMA heterogeneous small-cell network," in *Proc. IEEE 10th Sensor Array Multichannel Signal Process. Workshop (SAM)*, Jul. 2018, pp. 51–55.

[37] A. J. Muhammed, Z. Ma, Z. Zhang, P. Fan, and E. G. Larsson, "Energy-efficient resource allocation for NOMA based small cell networks with wireless backhauls," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3766–3781, Jun. 2020.

[38] 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.

[39] 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.

[40] Y. Fu, L. Salaun, C. W. Sung, C. S. Chen, and M. Coupechoux, "Double iterative waterfilling for sum rate maximization in multicarrier NOMA systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[41] Y. Fu, L. Salaün, C. W. Sung, and C. S. Chen, "Subcarrier and power allocation for the downlink of multicarrier NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11833–11847, Dec. 2018.

[42] L. Salaun, M. Coupechoux, and C. S. Chen, "Weighted sum-rate maximization in multi-carrier NOMA with cellular power constraint," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2019, pp. 451–459.

[43] E. C. Cejudo, H. Zhu, J. Wang, and O. Alluhaibi, "A fast algorithm for resource allocation in downlink multicarrier NOMA," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–5.

[44] A. B. M. Adam, X. Wan, and Z. Wang, "Energy efficiency maximization in downlink multi-cell multi-carrier NOMA networks with hardware impairments," *IEEE Access*, vol. 8, pp. 210054–210065, 2020.

[45] K. Huang, Z. Wang, H. Zhang, Z. Fan, X. Wan, and Y. Xu, "Energy efficient resource allocation algorithm in multi-carrier NOMA systems," in *Proc. IEEE 20th Int. Conf. High Perform. Switching Routing (HPSR)*, May 2019, pp. 1–5.

[46] S. Rezvani and E. A. Jorswieck. (Nov. 2021). *Optimal Power Allocation in Downlink Multicarrier NOMA Systems: Theory and Fast Algorithms*. [Online]. Available: https://gitlab.com/sepehrrezvani/MC-NOMA-SR-EE.git

[47] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[48] M. S. Ali, E. Hossain, A. Al-Dweik, and D. I. Kim, "Downlink power allocation for CoMP-NOMA in multi-cell networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3982–3998, Sep. 2018.

[49] D. P. Palomar and J. R. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 686–695, Feb. 2005.

[50] W. Yu and J. M. Cioffi, "Constant-power waterfilling: Performance bound and low-complexity implementation," *IEEE Trans. Commun.*, vol. 54, no. 1, pp. 23–28, Jan. 2006.

[51] P. He, L. Zhao, S. Zhou, and Z. Niu, "Water-filling: A geometric approach and its application to solve generalized radio resource allocation problems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3637–3647, Jul. 2013.

[52] N. Papandreou and T. Antonakopoulos, "Bit and power allocation in constrained multicarrier systems: The single-user case," *EURASIP J. Adv. Signal Process.*, vol. 2008, no. 1, pp. 1–14, Dec. 2007.

[53] C. Xing, Y. Jing, S. Wang, S. Ma, and H. V. Poor, "New viewpoint and algorithms for water-filling solutions in wireless communications," *IEEE Trans. Signal Process.*, vol. 68, pp. 1618–1634, 2020.

[54] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, K. C. Kiwiel and A. Ruszczynski, Eds. Cham, Switzerland: Springer, 1985.

[55] L. X. S. Boyd and A. Mutapcic, *Subgradient Methods (Lecture Notes)*. Stanford, CA, USA: Stanford Univ., Oct. 2003.

[56] R. Ruby, S. Zhong, H. Yang, and K. Wu, "Enhanced uplink resource allocation in non-orthogonal multiple access systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1432–1444, Mar. 2018.

[57] E. Che, H. D. Tuan, and H. H. Nguyen, "Joint optimization of cooperative beamforming and relay assignment in multi-user wireless relay networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5481–5495, Oct. 2014.

[58] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for non-orthogonal multiple access in HetNets," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5825–5837, May 2017.

[59] Z. Wei, D. W. K. Ng, J. Yuan, and H. M. Wang, "Optimal resource allocation for power-efficient MC-NOMA with imperfect channel state information," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3944–3961, Sep. 2017.

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

[61] D. Zhai and R. Zhang, "Joint admission control and resource allocation for multi-carrier uplink NOMA networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 922–925, Dec. 2018.

[62] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster," *IEEE Access*, vol. 6, pp. 5170–5181, 2018.

[63] R. Wang, W. Kang, G. Liu, R. Ma, and B. Li, "Admission control and power allocation for NOMA-based satellite multi-beam network," *IEEE Access*, vol. 8, pp. 33631–33643, 2020.

[64] W. Bao, H. Chen, Y. Li, and B. Vucetic, "Joint rate control and power allocation for non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2798–2811, Dec. 2017.

[65] M.-R. Hojeij, C. A. Nour, J. Farah, and C. Douillard, "Waterfilling-based proportional fairness scheduler for downlink non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 2, pp. 230–233, Apr. 2017.

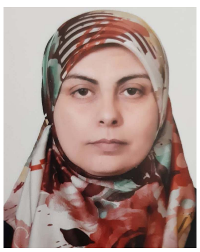
[66] D. Zhai, R. Zhang, L. Cai, B. Li, and Y. Jiang, "Energy-efficient user scheduling and power allocation for NOMA-based wireless networks with massive IoT devices," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1857–1868, Jun. 2018.

[67] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[68] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.

**Sepehr Rezvani** (Student Member, IEEE) received the M.Sc. degree in electrical and computer engineering from Tarbiat Modares University, Tehran, Iran. He is currently pursuing the Ph.D. degree with the Department of Information Theory and Communication Systems, Technische Universität Braunschweig, Braunschweig, Germany. His research interests include signal processing, analysis and applications of multiple access techniques, and optimization theory in wireless communication systems.

**Eduard A. Jorswieck** (Fellow, IEEE) was the Head of the Chair of Communications Theory and a Full Professor at the Dresden University of Technology, Germany, from 2008 to 2019. He is the Managing Director of the Institute of Communications Technology and the Head of the Chair of Communications Systems and a Full Professor with Technische Universität Braunschweig, Brunswick, Germany. He has published more than 150 journal articles, 15 book chapters, three monographs, and some 300 conference papers on his research topics. His main research interests are in the broad area of communications. He was a recipient of the IEEE Signal Processing Society Best Paper Award. His colleagues and he were recipients of the Best Paper and Best Student Paper Awards at the IEEE CAMSAP 2011, IEEE WCSP 2012, IEEE SPAWC 2012, IEEE ICUFN 2018, PETS 2019, and ISWCS 2019. Since 2017, he has been the Editor-in-Chief of the *EURASIP Journal on Wireless Communications and Networking*. He was on the Editorial Boards for the IEEE SIGNAL PROCESSING LETTERS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.



**Roghayeh Joda** (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 1998, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Tehran, Tehran, in 2001 and 2012, respectively. She was a visiting scholar with the Polytechnic Institute, New York University (NYU), Brooklyn, NY, USA. From September 2013 to August 2014, she was a Post-Doctoral Fellow at the University of Padua, Padua, Italy. In November 2014, she joined as a Research Assistant Professor at the ICT Research Institute, Tehran, where she was the Manager of two mega projects on 5G networks. She is currently a visiting researcher with the University of Ottawa, Ottawa, Canada. Her current research interests include communication theory, information theory, resource allocation, optimization, and machine learning with application to wireless networks, 5G and 6G networks.

**Halim Yanikomeroglu** (Fellow, IEEE) is a Professor with the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada. His research group has made substantial contributions to 4G and 5G wireless technologies; his group's current focus is the wireless infrastructure for 2030s and 2040s. His extensive collaboration with industry resulted in 39 granted patents. He is a fellow of the Engineering Institute of Canada (EIC) and the Canadian Academy of Engineering (CAE). He is a Distinguished Speaker for the IEEE Communications Society and the IEEE Vehicular Technology Society. He received several awards for his research, teaching, and service.