Resource Allocation for Energy Efficient User Association in User-Centric Ultra-Dense Networks Integrating NOMA and Beamforming

Long Zhang[a], Guobin Zhang[b,*], Xiaofang Zhao[b,**], Yali Li[b], Chuntian Huang[b], Enchang Sun[c,d], Wei Huang[e]

[a] *School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China*
[b] *School of Electrical Engineering and Intelligentization, Dongguan University of Technology, Dongguan 523808, China*
[c] *Beijing Advanced Innovation Center for Future Internet Technology, Beijing University of Technology, Beijing 100124, China*
[d] *Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*
[e] *Institute of Power and Energy Efficiency, China Electric Power Research Institute, Beijing 100192, China*

Abstract

A coupling of wireless access via non-orthogonal multiple access (NOMA) and wireless backhaul via beamforming is a promising way for downlink user-centric ultra-dense networks (UDNs) to improve system performance. However, the ultra-dense deployment of radio access points in macrocell and the user-centric view of network design in UDNs raise important concerns about resource allocation and user association, among which notably is energy efficiency (EE) balance. To overcome this challenge, we develop a framework to investigate the resource allocation problem for energy efficient user association in such a scenario. The joint optimization framework aiming at the system EE maximization is formulated as a large-scale non-convex mixed-integer nonlinear programming problem, which is NP-hard to solve directly with lower complexity. Alternatively, taking advantages of the sum-of-ratios decoupling and successive convex approximation methods, we transform the original problem into a series of convex optimization subproblems. Furthermore, we solve each subproblem through the Lagrangian dual decomposition, and design an iterative algorithm in a distributed way that realizes the joint optimization of power allocation, sub-channel assignment, and user association simultaneously. Simulation results demonstrate the effectiveness and practicality of our proposed framework, which achieves the rapid convergence speed and ensures a beneficial improvement of system-wide EE.

Keywords:
Ultra-dense network (UDN), Resource allocation, User association, Non-orthogonal multiple access (NOMA), Beamforming, User-centric networking, Energy efficiency

1. Introduction

During the past few years, the rapid proliferation of massive wireless smart devices and the trend increase in emerging applications, e.g., eXtended reality (XR), super Hi-vision (8K) videos, ultra-immersive

*Principal corresponding author
**Corresponding author
Email addresses: zhanglong@hebeu.edu.cn (Long Zhang), guobinzh@163.com (Guobin Zhang), aozhy1119@126.com (Xiaofang Zhao), lyl0112@hotmail.com (Yali Li), hct12138@hotmail.com (Chuntian Huang), ecsun@bjut.edu.cn (Enchang Sun), huangwei2@epri.sgcc.com.cn (Wei Huang)

games, etc., have propelled the unprecedented growth in mobile data traffic. It is predicted that the total data traffic in global scale will reach 136 EB per month and 1000 times more until 2024 from the existing Long Term Evolution (LTE) system to the fifth generation (5G) mobile system [1]. Such a thousand-fold traffic growth necessitates the configuration of ultra-dense networks (UDNs) as a new evolution paradigm to meet the challenges of fulfilling network capacity and spectral efficiency (SE) enhancement requirements for 5G and beyond [2, 3]. Instead of relying on a tower-mounted macro base station (MBS) with high transmit power in macrocell sending signals to a large number of user equipments (UEs), e.g., 0.2 UEs/m^2, UDNs deploy tens or hundreds more of low-powered radio access points (APs) with smaller coverage areas to coherently provide wireless access service for those users. As such, the ultra-dense deployment of APs has potentials to bring multiple benefits, e.g., enlarged cell coverage, improved spatial reuse of wireless resources, enhanced performance gains, etc [4, 5].

In spite of being advantageous, such an increasing density of APs with dense cell coverage, e.g., 10^3 APs/km^2 or more, results in the complex distribution of APs in UDNs and even possible overlapped coverage for the users. Therefore, simply using the traditional cell-centric architecture poses extra challenges on network planning and design for UDNs, e.g., complicated resource management, severe inter-cell interference, large signalling overhead, etc. More seriously, irregular coverage of the cells may cause some users exist in the overlapped area with severe interference, while other users exist in the edge of the cells or the area without coverage, which seriously degrade the quality-of-service (QoS) performance of the users. Therefore, it is imperative to implement a transformation of the network architecture from cell-centric to user-centric by adopting the idea of "network serving user" and cell-free fashion [6]. In a user-centric UDN, each user is simultaneously and jointly served by its selected subset of APs, i.e., an AP group (APG), in which the density of the APs is comparable to or even higher than that of the users. Through the deconstruction of cellular structure, user-centric UDNs not only eliminate the cell boundaries with entirely suppressed inter-cell interference, but also achieve the dynamic configuration of APG and flexible resource allocation in a user-centric manner.

While the user-centric UDNs with ultra-densely deployed APs overlaid with a traditional MBS in macrocell enable multi-Gigabit-per-second user experience and SE increases in wireless access downlink, limited wireless resources bring about serious competitions among APs towards massive access opportunities for the users [7]. This drives the research community to design more resource-efficient wireless network paradigm that copes with the scarcity of wireless resources. Recently, non-orthogonal multiple access (NOMA) has been recognized as one of the enabling air-interface techniques for 5G and beyond due to its advantages in support of overloaded transmission with limited resources and higher SE [8]. The key idea of NOMA is to allow multiple signals multiplexed to transmit simultaneously on the same frequency/time resource block (RB) by differentiating the signals through distinct power levels or user-specific codes, i.e., power-domain or code-domain multiplexing. For power-domain NOMA, successive interference cancellation (SIC) is exploited at the receiver side to decode its own received signal and reduce the undesired interference effectively. In this regard, NOMA can be well tailored to the wireless access downlink scenario in user-centric UDNs, where massive connectivity and heavy data traffic for the users is required over limited wireless resources. From a user-centric point of view, multiple APs,

e.g., an APG, can cooperatively and concurrently serve every user on the same sub-channel in wireless access downlink via NOMA. By doing so, significant SE enhancement will be attained in comparison with conventional orthogonal multiple access (OMA) schemes.

On the other hand, tens or hundreds more of distributed APs in user-centric UDNs impose additional constraints on the design of backhaul connections. Unlike traditional macrocell, in which a dedicated, high-capacity wired backhaul exists, e.g., optical fiber and digital subscriber line connections, it is impractical and uneconomical for every AP to be connected via fiber backhauling to the core networks [9]. This is due to the dramatic increase in deployment cost and possible geographical limitations for placement, e.g., hard-to-reach locations of APs in urban areas. An alternative is to utilize the wireless backhauling, which allows low-cost plug-and-play APs to employ over-the-air links to the MBS for backhauling. To reduce computational complexity of wireless backhaul design and to improve system efficiency of wireless access, there is a need to apply the clustering scheme to classify all the APs into a specific disjoint part based on feasible policy, e.g., channel condition and spatial location. Given this context, it is critical to manage the interference in wireless backhaul connections, especially for the inter-cluster interference in downlink transmission [10, 11]. Recently, multiple-antenna techniques have been regarded as a promising solution to achieve both higher SE and powerful interference mitigation via transmit beamforming [12]. Thus, a natural idea is to link beamforming and wireless backhaul downlink together to manage the interference intelligently. In the wireless backhaul, with multiple antennas at the MBS, downlink beamforming can be used to simultaneously transmit the weighted signals to the APs in different clusters by concentrating the signal power to the intended AP while reducing the interference generated to the other APs.

Under such circumstances, the integration of the wireless access via NOMA and the wireless backhaul via beamforming into user-centric UDNs is not only an extension and branch of traditional UDNs, but also a practical application incentive promoted to provide significant performance gains in terms of coverage, rate, delay, capacity, SE, and energy efficiency (EE). Despite these potential advantages, such an integration also imposes additional challenges and revealed some serious concerns particularly with the ultra-dense and random deployment of APs and the user-centric view of network optimization design. Firstly, relying on the sub-channels, every user is capable of being simultaneously associated with multiple APs for wireless access, and every AP has to be wirelessly connected to the MBS for backhauling. Hence, an increased complexity incurred by the ultra-densely deployed nodes makes the user association along with the AP-MBS association a challenging problem. Secondly, due to the limited available resources shared by the high number of users and APs, flexible and efficient resource allocation schemes are essential and very crucial to alleviate competition, control interference, and optimize system performance. Thirdly, a large-scale deployment of APs inevitably triggers a enormous growth of energy consumption, causing global warming for our planet and more operational costs for network operators. As such, it is of paramount importance to take the EE into account in the fundamental design objective for user-centric UDNs from the green communication perspective. Furthermore, user association also shows significant influence on the overall system-level energy consumption [13, 14]. For instance, some of the APs are highly overloaded due to the excessive associations with users, resulting in the similar amount

of energy consumed by other lightly underutilized APs, which degrades the long-term EE performance. It is for this reason that energy efficient user association is a key issue in the field of EE in UDNs. Aiming to address the above problems, there are two key network bottlenecks that must be overcome, namely resource allocation for a large-scale node deployments over the shared radio resources and energy efficient user association for achieving a load balancing of APs and MBS. Admittedly, these bottlenecks and challenges motivate the need for better understanding of the interplay between resource allocation and energy efficient user association, which typically require a trade-off between them.

Motivated by the above observations, we can find that the exploration of resource allocation for energy efficient user association has become highly valuable. Correspondingly, our objective in this paper is to achieve the resource allocation for energy efficient user association for identifying such an interplay under the scenario of user-centric UDNs integrating wireless access and wireless backhaul. To the best of our knowledge, the problem of resource allocation for energy efficient user association through the efficient integration of user-centric UDNs with NOMA and beamforming has yet not been thoroughly studied in the literature, especially considering the joint coordination between the access downlink via NOMA and the backhaul downlink via beamforming. In this paper, we investigate a resource allocation problem for energy efficient user association for downlink user-centric UDNs integrating wireless access via NOMA and wireless backhaul via beamforming, aiming to maximize the system EE under the constraints of achievable rate for wireless access/backhaul connection, transmit power limit of the MBS and every AP, and user association relations. More specifically, the main contributions of this paper can be summarized as follows:

- We develop a novel resource allocation optimization framework to achieve the energy efficient user association in the downlink transmission of user-centric UDNs by jointly taking into account wireless access and wireless backhaul. This is a new approach to user-centric view of network optimization design in UDNs to capture the EE balance through a flexible paradigm of tightly integrating the access downlink via NOMA and the backhaul downlink via beamforming from a global standpoint. Our framework is the first time in the literature to identify a close coupling of NOMA based wireless access and beamforming based wireless backhaul in downlink user-centric UDNs.

- We formulate the resource allocation problem for energy efficient user association under such an integration of user-centric UDNs with NOMA and beamforming as a large-scale non-convex mixed-integer nonlinear programming problem, which is NP-hard to solve in reasonable time with the growing numbers of densely distributed users and APs. The objective of joint resource allocation and user association is to maximize the system EE of the downlink transmission subject to the constraints of achievable data rate for wireless access and backhaul connection, maximum transmit power for the MBS and each AP, and user association relations. The framework is shown to jointly optimize the transmit power allocated to the users and the APs, the sub-channel assignment for the access and backhaul downlink, and the association relations for both the user-AP and the AP-MBS simultaneously.

- To tackle this problem with a reduced complexity, we firstly conduct a series of reformulation based on the time-sharing relaxation strategy to relax the binary variables for user association. Then the sum-of-ratios decoupling method is used to construct the transformation of the fractional structure of the relaxed objective function into an equivalent parametric subtractive function. We accordingly employ the iterative successive convex approximation (SCA) to transform the original highly non-convex problem into a series of convex subproblems via the exponential-logarithmic approximation, and apply the Lagrangian dual decomposition approach to solve these optimization subproblems. To ensure rapid convergence speed of the update of optimal power, an effective algorithm in a fully distributed fashion is developed to determine a specific execution coordination between sub-channel assignment and power allocation.

- Through extensive simulations, we demonstrate the proposed algorithm in our framework is indeed an efficient and practical solution for joint resource allocation and user association in user-centric UDNs integrating NOMA and beamforming, and we obtain insights into how the various system parameters influence the convergence speed of the optimal power update and the system-wide EE. With regard to the same system parameters and requirements of data rate and power consumption for each user, each AP, and the MBS, we also show that the overall EE performance from a system point of view is always superior with the proposed framework when compared with the baseline schemes.

The rest of this paper is organized as follows. We first introduce the related work in Section 2. Section 3 describes the system model, followed by a construction of the optimization problem. In Section 4, we present the problem reformulation through the relaxation of binary variables, the sum-of-ratios decoupling, and the successive convex approximation technique. Section 5 provides the Lagrangian dual decomposition method to solve the convex subproblem and proposes a decentralized iterative algorithm to derive the optimal solutions. In Section 6, we present the simulation results to evaluate the proposed optimization framework. Finally, we conclude our paper in Section 7.

Notation: Throughout this paper, we use a, $\mathbf{a}$, $\mathbf{A}$, and $\mathcal{A}$ to denote a scalar variable, a vector, a matrix, and a set, respectively. The distribution of a circularly symmetric complex-valued Gaussian random variable x with mean ϱ and variance σ^2 is represented by $x \sim \mathcal{CN}\left(\varrho, \sigma^2\right)$, where $\sim$ stands for "distributed as". The identity matrix, or sometimes ambiguously called a unit matrix, is denoted as $\mathbf{I}$, and an $(n \times n)$-dimensional identity matrix is defined by $\mathbf{I}_n$. The superscript $[\cdot]^{\mathrm{T}}$ refers to the transpose of a matrix or a vector. In addition, we denote the statistical expectation of a random variable by the notation $\mathbb{E}\{\cdot\}$. Symbol $\mathbb{C}$ is used to indicate the complex number field. An n-dimensional complex vector is represented by $\mathbb{C}^{n\times 1}$, whereas $\mathbb{C}^{n\times m}$ corresponds to the generalization to an $(n\times m)$-dimensional complex matrix.

2. Related Work

Currently, many potential issues in the realization of user-centric UDNs have been identified and discussed separately [2, 3, 4, 6]. Among them, resource allocation in this scenario is a critical issue

that has gained widespread popularity. In [15], Lin *et al.* presented an optimization framework of the modularity-based user-centric clustering and resource allocation for UDNs to maximize the sum-rate per orthogonal RB. Based on a three-stage sequential method, a heuristic RB allocation solution was obtained by confining the maximum size of the clusters. In [16], Zhang *et al.* studied the joint sub-channel and power allocation problem in full-duplex user-centric UDNs to maximize the total capacity of system. Through the problem decomposition, the inter-cell and intra-cell resource allocation was proposed by using a tier-separate and variable-separate based approach. In [17], Cao *et al.* modeled the potential interference relationship of users in ultra-dense femtocell networks as conflict-graph according to the network partition state. Based on this graph model, a sub-channel allocation algorithm was devised by assigning the orthogonal sub-channels to the users that severely interfere each other or assigning the unavailable sub-channels in a profit-calculating method. However, all these works just concerned the resource allocation for the users in wireless access without capturing the potential benefits of designing wireless backhaul connections.

As mentioned in Section 1, the EE is an unneglected key performance metric in user-centric UDNs from the point of view of green communications. Some recent works have recognized the EE as one of important optimization criteria in resource allocation. In [18], Park *et al.* proposed a decentralized user-centric reverse association policy, which achieves the joint optimization of handover and power control to maximize the EE of AP. Both the spatial randomness of user movement and temporally correlated channel fading under a large number of APs were incorporated based on the spatio-temporal dynamics. From a secrecy EE perspective, Lin *et al.* [19] developed a user-centric clustering method to attain secure transmissions, i.e., user association, for both the dedicated and the embedded jamming. The proposed method can degrade the overheard signals of the eavesdroppers and guarantee secure transmission by considering the involvement status of each AP. However, the transmit power of each AP was treated as the same fixed value, whereas how to allocate the AP's power was not exploited. In addition, to increase both EE and SE, Zhang *et al.* [14] designed a joint optimization framework of load-aware user association and power allocation in mmWave-based UDNs with energy harvesting APs. These above studies are heuristic, although they only investigated the impact of either power allocation or user association on the EE maximization for system or AP in the scenario of wireless access. By contrast, we extensively consider the joint coordination between the access downlink and the backhaul downlink from a global standpoint.

Due to its appealing advantages such as enhanced SE, massive user connectivity, and low latency, etc, the combination of NOMA and user-centric UDNs has recently aroused enormous interests and attention from the research community. In [7], Liu *et al.* explored the efficient access framework for users in NOMA-based user-centric UDNs, in which both the access downlink and the backhaul downlink were designed by using power-domain NOMA, respectively. An optimization problem for user-centric access was formulated by aiming at maximizing the EE of system, and further transformed to an equivalent AP grouping problem along with cooperative resource allocation. Despite the joint consideration of the access and backhaul downlink, our work differs from the work in [7] due to the specific technology usage for the wireless backhaul downlink. We focus on the application of multiple-antenna technique

into the wireless backhaul downlink aiming to control the inter-cluster interference intelligently. To show the benefits of the EE and SE improvements of wireless backhaul in heterogeneous UDNs over the traditional OMA, Zhang *et al.* [20] developed a two-tiered hierarchical model, i.e., cooperative OMA and cooperative NOMA, for the cooperative wireless backhaul scheduling. A cooperative wireless backhaul optimization problem was devised by achieving the system EE maximization under the power and rate threshold constraints, and the greedy algorithm-based schemes were further proposed. In [21], Qin *et al.* presented a unified NOMA framework covering power-domain NOMA and code-domain NOMA in UDNs, and performed the uplink and downlink design under this framework. Particularly, the problem of resource allocation and user association for this scenario was studied by using the stochastic geometry model and the sequential convex programming method. To improve the fairness and resource efficiency among Internet of Things (IoT) users, the fairness factor was introduced by Wang and Zhou [22] into the design of utility function for resource allocation and computation offloading in an MEC-enabled ultra-dense IoT network with power-domain NOMA. The optimal trade-off between computation rate and power consumption for each user was obtained by adjusting the fairness factor. However, only a wireless access scenario with NOMA was considered in [20, 21, 22], which cannot capture the effect of the wireless backhauling design on the overall system performance. Meanwhile, none of these works dealt with the sub-channel assignment strategies under their proposed resource allocation frameworks.

Moreover, there is limited work available in the open literature applying multiple-antenna technique into user-centric UDNs. In [23], Kwon and Park explored the joint problem of time resource allocation, user association, and hybrid beamforming design in mmWave UDNs with wireless backhaul to maximize the weighted sum rate of the users with limited feedback. A two-stage approach was employed to solve this problem for reducing the complexity and overhead. However, the work in [23] focused on the hybrid beamforming design in both the backhaul links and the access links. By jointly considering the uplink feedback and downlink transmission process in UDNs, Teng *et al.* [24] analyzed the impact of delayed feedback and limited measure range on transmit beamforming performance. To resolve the interference problem caused by content-centric communications by cache-enabled UDNs, a collaborative multicast beamforming scheme was proposed by Nguyen *et al.* [25] aiming to maximize the cost efficiency in content delivery. By using the zero-forcing beamforming and generalized zero-forcing beamforming, the multi-content interference was forced to zero and mitigate it while amplifying the desired signals for the users, respectively. However, the above related works in [24, 25] applied the downlink beamforming only to the scenario of the access links and did not consider the wireless backhaul design.

To sum up, although a lot of works have been carried out on the resource allocation problem in user-centric UDNs, NOMA-aided UDNs, and beamforming-aided UDNs extensively, efficient integration of user-centric UDNs with NOMA and beamforming techniques has not been fully utilized. This research gap motivates us to pursue a solution for the problem of joint resource allocation and user association optimization to maximize the system-wide EE of the downlink transmission integrating both the access downlink via NOMA and the backhaul downlink via beamforming.

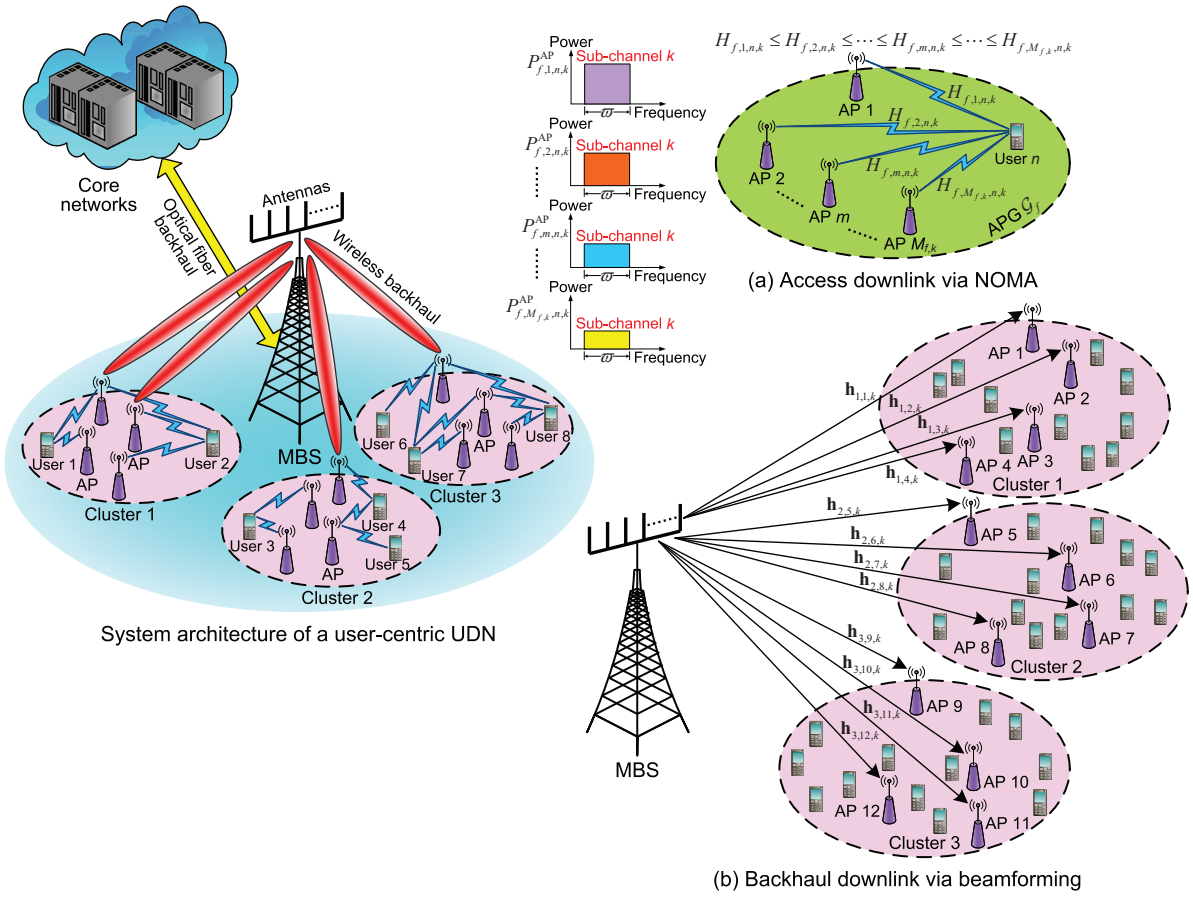

Figure 1: Illustration of a user-centric UDN integrating NOMA and beamforming for the downlink transmission.

3. System Model and Problem Formulation

In this section, we first introduce the network model of a typical user-centric UDN. Under this system configuration, we provide the transmission model from the downlink perspective, i.e., the access downlink via NOMA and the backhaul downlink via beamforming, and further describe the power consumption model for the downlink transmission. Then, the system EE maximization problem for the downlink transmission will be formulated.

3.1. Network Model

Consider a user-centric UDN as shown in Fig. 1, where an MBS with a large scale antenna array is located in the center with a large number of APs, denoted by a set $\mathcal{M} = \{1, 2, \cdots, M\}$, densely deployed within the macrocell coverage of that MBS. Particularly, the macrocell is connected to the core networks through optical fiber backhaul and the MBS is responsible for wireless backhaul connections for all the APs. The coverage radius of the macrocell is specified by r. There also exist N users randomly distributed in the overlapping macrocell coverage area, denoted by a set $\mathcal{N} = \{1, 2, \cdots, N\}$, sharing the same spectrum resource with the MBS and the APs. Note that each AP is equipped with one or more receive antenna(s) for backhaul connections, and also configured with multiple transmit antennas to serve more users simultaneously in a user-centric fashion. We assume that the locations of the APs are modeled by an independent homogeneous Poisson point process (PPP) Φ_{ρ_1} with density $\rho_1 = \frac{M}{\pi r^2}$

that is comparable to or even larger than user density $\rho_2 = \frac{N}{\pi r^2}$. For simplicity, we utilize a quasi-static deployment scenario for users, such that the location of each user remains unchanged within the considered time duration[1].

In this paper, we focus on joint resource allocation and user association in the downlink transmission of such a user-centric scenario by tightly integrating wireless access and wireless backhaul together. Specifically, the wireless downlink consists of two parts: (i) wireless access downlink from an AP to a user in the corresponding cluster, and (ii) wireless backhaul downlink from the MBS to an AP in the macrocell[2]. For the coordination between the MBS and the AP, we adopt a dynamic time division duplex (TDD) mode [7], in which both the MBS and the AP can independently transmit in the wireless backhaul connections and the wireless access, respectively.

The total available bandwidth W is equally divided to K orthogonal sub-channels, represented by a set $\mathcal{K} = \{1, 2, \cdots, K\}$. So each sub-channel has an equally-sized bandwidth of $\varpi = \frac{W}{K}$. Due to the densely deployment scenario, we consider the universal frequency reuse policy so that the sub-channels are available to all the users for the wireless access and all the APs for the wireless backhaul connections, respectively. To avoid the interference between the access downlink and the backhaul downlink, the sub-channel set $\mathcal{K}$ is separated into two subsets, i.e., $\mathcal{A} = \{1, 2, \cdots, \delta\}$ for the access downlink and $\mathcal{B} = \{\delta + 1, \delta + 2, \cdots, K\}$ for the backhaul downlink. In other words, the former δ sub-channels in $\mathcal{K}$ are used for the wireless access, and the other $K - \delta$ sub-channels in $\mathcal{K}$ are selected for the wireless backhaul connections.

Let us assume that the perfect knowledge of the channel side information (CSI) for every sub-channel is known at both the MBS and every AP. In accordance with the perfect CSI of every sub-channel, the APs allocate a subset of $\mathcal{A}$ to the users, and the MBS assigns a subset of $\mathcal{B}$ to the APs. To strike a balance between efficient user-centric wireless access and computational complexity, the ultra-densely distributed APs are initially separated into F disjoint clusters based on their spatial directions[3], denoted by a set $\mathcal{F} = \{1, 2, \cdots, F\}$, as displayed in Fig. 1. We suppose that an AP can only provide wireless access service exactly for one or more user(s) over a subset of $\mathcal{A}$ within the same cluster to avoid extra inter-cluster interference. More precisely, in every cluster f, user n can be simultaneously associated with at most M_f APs on one or more sub-channel(s) within the considered time duration, for $0 \leq M_f \ll M$, $f \in \mathcal{F}$, and $n \in \mathcal{N}$. As such, M_f APs in cluster f constitute a generalized APG, denoted by a set $\mathcal{G}_f$, to serve user n by concurrently transmitting independent signals in a user-centric way[4], for $\mathcal{G}_f \subset \mathcal{M}$. We wish to remark that the APs in generalized APG $\mathcal{G}_f$ also belong to cluster f.

[1] We would like to mention that our proposed optimization framework for joint resource allocation and user association is conducted within the considered time duration, which can be interpreted as a specific time slot or a period of time. However, the results about this framework will be easily extendable to the general case for multiple time slots.

[2] In what follows, unless otherwise stated, we use the terms "wireless access" and "access downlink" interchangeably. Furthermore, the terms "wireless backhaul" and "backhaul downlink" are all interchangeable.

[3] Noticing that a detailed discussion on the clustering method is beyond the scope of this work.

[4] We should pay more attention to the difference between the AP cluster and the APG in this work. An AP cluster is referred to as the result of the task of classifying all the APs into a specific disjoint part according to their spatial location relations. From a user perspective, an APG is a subset of APs in an AP cluster, and each AP in this subset is associated with that user in a user-centric fashion.

3.2. Transmission Model

3.2.1. Access Downlink via NOMA

In the access downlink, a user in each cluster can be simultaneously served by multiple APs in a user-centric fashion through an assigned sub-channel from the sub-channel set $\mathcal{A}$. Motivated by that, we assume that the considered system adopts the power-domain NOMA for the access downlink transmission, which enables that multiple signals from the APs in a cluster can multiplex on the same sub-channel at the same time. According to the NOMA principle, one user can receive from the APs in the same cluster via multiple sub-channels, and one sub-channel can be assigned to multiple users.

For convenience, let us define a binary variable as follows to indicate the association relationship between user n on sub-channel k and AP m in cluster f, for $f \in \mathcal{F}$, $m \in \mathcal{M}$, $n \in \mathcal{N}$, and $k \in \mathcal{A}$:

$$a_{f,m,n,k} = \begin{cases} 1 & \text{if user } n \text{ is associated with AP } m \text{ in cluster } f \text{ using sub-channel } k, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Let $P^{\mathrm{AP}}_{f,m,n,k}$ denote the allocated transmit power of AP m in cluster f to user n on sub-channel k. We further assume that all the sub-channels for the access downlink follow a quasi-static block fading, where the channel gains remain to be constant over the considered time duration, but may vary independently between different time duration. As such, we denote the downlink channel coefficient from AP m in cluster f to user n on sub-channel k as $h_{f,m,n,k} = g_{f,m,n,k} d_{f,m,n}^{-\vartheta_1}$, where $g_{f,m,n,k}$ is the flat Rayleigh fading channel gain, $d_{f,m,n}$ is the distance between AP m in cluster f and user n, and ϑ_1 is the path loss exponent. Let $N_{f,k}$ be the number of users using sub-channel k in cluster f, and $s_{f,m,n,k}$ be the transmitted symbol of AP m in cluster f to user n on sub-channel k. Thus, the received signal at user n on sub-channel k from AP m in cluster f can be expressed as:

$$y_{f,m,n,k} = h_{f,m,n,k} \sum_{i=1}^{N_{f,k}} \sqrt{P^{\mathrm{AP}}_{f,m,i,k}} s_{f,m,i,k} + z_{n,k}, \tag{2}$$

where $z_{n,k} \sim \mathcal{CN}\left(0, \sigma^2_{n,k}\right)$ is the additive white Gaussian noise (AWGN) at user n on sub-channel k with zero mean and variance $\sigma^2_{n,k}$. After receiving the superposed signals from $M_{f,k}$ APs on sub-channel k in generalized APG $\mathcal{G}_f$ with a user-centric approach, user n employs the SIC technique to decode its desired messages, for $0 \leq M_{f,k} < M_f$[5]. Let $H_{f,m,n,k} = \left|h_{f,m,n,k}\right|^2 / \sigma^2_{n,k}$ represent the channel to noise ratio (CNR) of sub-channel k from AP m in cluster f to user n. Without loss of generality, we assume that the CNRs of the received signals at user n on sub-channel k served by $M_{f,k}$ APs on sub-channel k in generalized APG $\mathcal{G}_f$ are sorted in the ascending order, i.e.:

$$H_{f,1,n,k} \leq H_{f,2,n,k} \leq \cdots \leq H_{f,m,n,k} \leq \cdots \leq H_{f,M_{f,k},n,k}. \tag{3}$$

Note that the received signals with lower CNRs from the APs in a generalized APG are allocated higher powers and can be recovered by treating the received signals with lower powers as the interference in the

[5]It should be pointed out that the group of $M_{f,k}$ APs on sub-channel k can be deemed to a subset of generalized APG $\mathcal{G}_f$ on the entire sub-channels.

SIC decoding [26, 27]. To be precise, for the received signal from AP m, user n on sub-channel k first decodes the message from AP j in generalized APG $\mathcal{G}_f$, for $j < m$, and then removes this message from its received signals, in the order of $j = 1, 2, \cdots, m-1$. Through the sequential decoding, the signals from AP j can be treated as the interference, for $j > m$. As a result, the received signal-to-interference-plus-noise ratio (SINR) at user n on sub-channel k served by AP m in generalized APG $\mathcal{G}_f$ by performing the SIC is given by:

$$\gamma_{f,m,n,k}^{\mathrm{AD}} = \frac{H_{f,m,n,k} P_{f,m,n,k}^{\mathrm{AP}}}{1 + \sum\limits_{j=m+1}^{M_{f,k}} H_{f,j,n,k} P_{f,j,n,k}^{\mathrm{AP}}}, \tag{4}$$

where $\sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k} P_{f,j,n,k}^{\mathrm{AP}}$ is the interference that user n on sub-channel k receives from other APs in generalized APG $\mathcal{G}_f$. Correspondingly, the achievable rate (in bit/s) of user n on sub-channel k served by AP m in generalized APG $\mathcal{G}_f$ can be written as:

$$R_{f,m,n,k} = \varpi \log_2 \left(1 + \frac{H_{f,m,n,k} P_{f,m,n,k}^{\mathrm{AP}}}{1 + \sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k} P_{f,j,n,k}^{\mathrm{AP}}} \right). \tag{5}$$

Recall that one or more user(s) in $\mathcal{N}$ over a subset of $\mathcal{A}$ can access to multiple APs in every cluster through a user-centric way. Let N_f denote the number of users that are associated with the APs in cluster f, for $0 \leq N_f \ll N$. Therefore, the achievable sum rate of the system for the access downlink via NOMA can be calculated by:

$$R_{\mathrm{Sum}}^{\mathrm{AD}} = \sum_{f=1}^{F} \sum_{n=1}^{N_f} \sum_{m=1}^{M_f} \sum_{k=1}^{\delta} a_{f,m,n,k} \varpi \log_2 \left(1 + \frac{H_{f,m,n,k} P_{f,m,n,k}^{\mathrm{AP}}}{1 + \sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k} P_{f,j,n,k}^{\mathrm{AP}}} \right). \tag{6}$$

3.2.2. Backhaul Downlink via Beamforming

In the backhaul downlink, the MBS concurrently transmits independent signals to the APs in different clusters over the sharing sub-channels. By exploiting multiple antennas at both the MBS and the APs, downlink beamforming is considered in the wireless backhaul not only to increase the SE, but also to combat the inter-cluster and intra-cluster interference.

Let Q be the number of the transmit antennas for beamforming in the antenna array of the MBS, for $Q \geqq M$. Denote $\phi_{f,k}$ as the number of APs on sub-channel k in cluster f, for $0 \leq \phi_{f,k} \ll M \leq Q$. The downlink channel between the MBS and $\phi_{f,k}$ APs on sub-channel k in cluster f is described by a matrix $\mathbf{H}_{f,k} = \left[\mathbf{h}_{f,1,k}, \mathbf{h}_{f,2,k}, \cdots, \mathbf{h}_{f,\phi_{f,k},k} \right]^{\mathrm{T}} \in \mathbb{C}^{\phi_{f,k} \times Q}$, and the row vector $\mathbf{h}_{f,m,k} \in \mathbb{C}^{1 \times Q}$ is the channel coefficient between the MBS and AP m on sub-channel k in cluster f. For ease of exposition, the channel coefficient vector is characterized by $\mathbf{h}_{f,m,k} = \tilde{\mathbf{h}}_{f,m,k} d_{f,m}^{-\vartheta_2}$, where $d_{f,m}$ is the distance between the MBS and AP m in cluster f, ϑ_2 is the path loss exponent, and $\tilde{\mathbf{h}}_{f,m,k}$ is the small scale fading (e.g., Rayleigh fading) channel coefficient vector that is assumed to be complex Gaussian distributed with zero mean and unit variance matrix, i.e., $\tilde{\mathbf{h}}_{f,m,k} \sim \mathcal{CN}(0, \mathbf{I}_Q)$. Thus, such kind of channel coefficient is time invariant over the considered time duration, but may still vary from different time duration. Moreover,

we suppose that the downlink channel coefficient vector of interest is available at the MBS by the aid of the CSI feedback information [12].

In order to represent the association relationship between the MBS and AP m on sub-channel k in cluster f, for $f \in \mathcal{F}$, $m \in \mathcal{M}$, and $k \in \mathcal{B}$, a binary variable is also introduced, which can be defined by:

$$b_{f,m,k} = \begin{cases} 1 & \text{if AP } m \text{ in cluster } f \text{ is associated with the MBS using sub-channel } k, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Let us utilize $\mathbf{s}_k = [s_{1,k}, s_{2,k}, \cdots, s_{F,k}]^{\mathrm{T}} \in \mathbb{C}^{F\times 1}$ to represent the transmitted symbol vector of the MBS on sub-channel k for F clusters. Assume that $P_{f,m,k}^{\mathrm{MBS}}$ is the allocated transmit power of the MBS to AP m on sub-channel k in cluster f. Thereby, the transmitted symbols for $\phi_{f,k}$ APs on sub-channel k in cluster f can be expressed as:

$$s_{f,k} = \sum_{m=1}^{\phi_{f,k}} \sqrt{P_{f,m,k}^{\mathrm{MBS}}} s_{f,m,k}, \tag{8}$$

where $s_{f,m,k}$ is the normalized transmitted symbol of the MBS to AP m on sub-channel k in cluster f, i.e., $\mathbb{E}\left\{|s_{f,m,k}|^2\right\} = 1$. To carry out the downlink beamforming, let $\mathbf{w}_{f,m,k}$ stand for the beamforming vector for AP m on sub-channel k in cluster f. Accordingly, the downlink beamforming matrix of the MBS on sub-channel k for F clusters is given by $\mathbf{W}_k = [\mathbf{w}_{1,k}, \mathbf{w}_{2,k}, \cdots, \mathbf{w}_{F,k}] \in \mathbb{C}^{\phi_{f,k}\times F}$, where $\mathbf{w}_{f,k} = \left[\mathbf{w}_{f,1,k}, \mathbf{w}_{f,2,k}, \cdots, \mathbf{w}_{f,\phi_{f,k},k}\right]^{\mathrm{T}} \in \mathbb{C}^{\phi_{f,k}\times 1}$ is the beamforming vector for $\phi_{f,k}$ APs on sub-channel k in cluster f. Note that the conventional beamforming approaches can be used in that the downlink channel coefficient vectors are known at the MBS as mentioned earlier. However, we do not discuss the issue of the beamforming vector optimization as it is beyond the scope of the paper.

By combining the transmitted symbol vector and the downlink beamforming matrix of the MBS on sub-channel k for F clusters, we can obtain the transmitted signals on sub-channel k, i.e., $\mathbf{X}_k = \mathbf{W}_k \mathbf{s}_k$. To simplify analysis, we consider that the number of the used transmit antennas for beamforming at the MBS is equal to the number of APs on sub-channel k in cluster f. As a result, the received signal at AP m on sub-channel k in cluster f can be modeled as:

$$\begin{aligned} y_{f,m,k} &= \mathbf{h}_{f,m,k}\mathbf{w}_{f,k}s_{f,k} + \mathbf{h}_{f,m,k}\sum_{\ell=1,\ell\neq f}^{F}\mathbf{w}_{\ell,k}s_{\ell,k} + z_{m,k} \\ &= \mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\sqrt{P_{f,m,k}^{\mathrm{MBS}}}s_{f,m,k} + \mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\sum_{j=1,j\neq m}^{\phi_{f,k}}\sqrt{P_{f,j,k}^{\mathrm{MBS}}}s_{f,j,k} \\ &\quad + \mathbf{h}_{f,m,k}\sum_{\ell=1,\ell\neq f}^{F}\mathbf{w}_{\ell,k}s_{\ell,k} + \wp_{f,m,k}z_{m,k}, \end{aligned} \tag{9}$$

where $z_{m,k} \sim \mathcal{CN}\left(0, \sigma_{m,k}^2\right)$ is the AWGN at AP m on sub-channel k with zero mean and variance $\sigma_{m,k}^2$. Thus, the SINR at AP m on sub-channel k in cluster f for the backhaul downlink via beamforming can be obtained as follows:

$$\gamma_{f,m,k}^{\mathrm{BD}} = \frac{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2 P_{f,m,k}^{\mathrm{MBS}}}{\underbrace{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2 \sum_{j=1,j\neq m}^{\phi_{f,k}} P_{f,j,k}^{\mathrm{MBS}}}_{\text{Intra-cluster interference}} + \underbrace{\sum_{\ell=1,\ell\neq f}^{F} \left|\mathbf{h}_{f,m,k}\mathbf{w}_{\ell,k}\right|^2 P_{\ell,k}^{\mathrm{MBS}}}_{\text{Inter-cluster interference}} + \underbrace{\sigma_{m,k}^2}_{\text{AWGN}}}, \tag{10}$$

where $P_{\ell,k}^{\mathrm{MBS}}$ is the total transmit power of the MBS to the APs on sub-channel k in cluster ℓ, for $\ell \in \mathcal{F} \setminus \{f\}$. It suffices to mention that the received signal at AP m on sub-channel k in cluster f is corrupted by the intra-cluster interference, the inter-cluster interference, and the AWGN. For analytical simplicity, we employ the zero-forcing beamforming to eliminate the inter-cluster interference [28]. As such, the achievable rate (in bit/s) of AP m on sub-channel k in cluster f is given by:

$$R_{f,m,k} = \varpi \log_2 \left(1 + \frac{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2 P_{f,m,k}^{\mathrm{MBS}}}{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2 \sum_{j=1,j\neq m}^{\phi_{f,k}} P_{f,j,k}^{\mathrm{MBS}} + \sigma_{m,k}^2}\right). \tag{11}$$

In consequence, the achievable sum rate of the system for the backhaul downlink via beamforming can be denoted as:

$$R_{\mathrm{Sum}}^{\mathrm{BD}} = \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K} b_{f,m,k}\varpi \log_2 \left(1 + \frac{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2 P_{f,m,k}^{\mathrm{MBS}}}{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2 \sum_{j=1,j\neq m}^{\phi_{f,k}} P_{f,j,k}^{\mathrm{MBS}} + \sigma_{m,k}^2}\right). \tag{12}$$

3.3. Power Consumption Model

Power consumption during the downlink transmission with the combination of wireless access via NOMA and wireless backhaul via beamforming is considered in this subsection. The total system power consumption of particular interest can be divided into the power consumed in the access downlink and the power consumed in the backhaul downlink.

For the access downlink, the power consumption is aimed at the power consumed at the users in receiving mode and at the APs in transmission mode, respectively. To be precise, the power consumption for user n in cluster f can be written as $P_{f,n}^{\mathrm{Con}} = P_{f,n}^{\mathrm{R}} + \psi_{\mathrm{A}} P_{f,n}^{\mathrm{D}}$, where $P_{f,n}^{\mathrm{R}}$ is the constant circuit power consumption for received signal processing, $P_{f,n}^{\mathrm{D}}$ is the dynamic circuit power consumption for signal decoding, and ψ_{A} is correlated with the number of APs in every APG on each sub-channel. Additionally, the power consumption for AP m in cluster f sending signal to user n on sub-channel k is determined by the transmitter circuit power consumption P_m^{C} and the transmit power $P_{f,m,n,k}^{\mathrm{AP}}$, i.e., $P_m^{\mathrm{Con}} = P_m^{\mathrm{C}} + P_{f,m,n,k}^{\mathrm{AP}}$. Thus, the sum power consumption in the access downlink can be expressed as:

$$\begin{aligned} P_{\mathrm{Sum}}^{\mathrm{AD}} &= \underbrace{\sum_{f=1}^{F}\sum_{n=1}^{N_f} \left(P_{f,n}^{\mathrm{R}} + \psi_{\mathrm{A}} P_{f,n}^{\mathrm{D}}\right)}_{\text{Receiving mode for users}} + \underbrace{\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\left(P_m^{\mathrm{C}} + P_{f,m,n,k}^{\mathrm{AP}}\right)}_{\text{Transmission mode for APs}} \\ &= \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\left(P_{f,n}^{\mathrm{Con}} + P_m^{\mathrm{C}} + P_{f,m,n,k}^{\mathrm{AP}}\right). \end{aligned} \tag{13}$$

For the backhaul downlink, the power consumption consists of the power consumed at the APs in receiving mode and at the MBS in transmission mode. Similarly, the power consumption for AP m in cluster f can be specifically defined as $P_{f,m}^{\mathrm{Con}} = P_{f,m}^{\mathrm{R}} + \psi_{\mathrm{B}} P_{f,m}^{\mathrm{D}}$, where $P_{f,m}^{\mathrm{R}}$ is the constant circuit power consumption for received signal processing, $P_{f,m}^{\mathrm{D}}$ is the dynamic circuit power consumption for signal decoding, and ψ_{B} is also correlated with the number of APs in every cluster on each sub-channel. In addition, the power consumption of the MBS for downlink beamforming mainly depends on the transmit power $P_{f,m,k}^{\mathrm{MBS}}$ of the MBS to to AP m on sub-channel k in cluster f. Accordingly, the sum power consumption in the backhaul downlink is given by:

$$\begin{aligned} P_{\mathrm{Sum}}^{\mathrm{BD}} &= \underbrace{\sum_{f=1}^{F}\sum_{m=1}^{M_f} \left(P_{f,m}^{\mathrm{R}} + \psi_{\mathrm{B}} P_{f,m}^{\mathrm{D}}\right)}_{\text{Receiving mode for APs}} + \underbrace{\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K} b_{f,m,k} P_{f,m,k}^{\mathrm{MBS}}}_{\text{Transmission mode for MBS}} \\ &= \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K} b_{f,m,k}\left(P_{f,m}^{\mathrm{Con}} + P_{f,m,k}^{\mathrm{MBS}}\right). \end{aligned} \tag{14}$$

Based on the sum power consumption in both the access downlink and the backhaul downlink, the total power consumption for the downlink transmission can be represented as:

$$P_{\mathrm{Tot}} = \underbrace{\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\left(P_{f,n}^{\mathrm{Con}} + P_{m}^{\mathrm{C}} + P_{f,m,n,k}^{\mathrm{AP}}\right)}_{\text{Access downlink}} + \underbrace{\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K} b_{f,m,k}\left(P_{f,m}^{\mathrm{Con}} + P_{f,m,k}^{\mathrm{MBS}}\right)}_{\text{Backhaul downlink}}. \tag{15}$$

3.4. Problem Formulation

In this paper, we investigate the resource allocation problem for energy efficient user association in the downlink transmission of the system with the emphasis on the EE metric. It has been shown that the system-wide EE metric of interest is generally described in terms of bit-per-Joule capacity, to indicate how efficiently one Joule power consumption is utilized for data transmission of the system. Considering the wireless access downlink from the APs to the users via NOMA and the wireless backhaul downlink from the MBS to the APs via beamforming, the actual total achievable rate (in bit/s) of the system for the downlink transmission is in general obtained by:

$$R_{\mathrm{Tot}} = \min\left\{R_{\mathrm{Sum}}^{\mathrm{AD}}, R_{\mathrm{Sum}}^{\mathrm{BD}}\right\}. \tag{16}$$

From the perspective of the wireless backhaul connections, the MBS in the macrocell must provide enough data rate for the APs to guarantee that all the users can obtain the wireless access from these APs in a user-centric way. To reach this goal, the achievable sum rate of the system for the backhaul downlink should not be less than that for the access downlink, i.e., $R_{\mathrm{Sum}}^{\mathrm{BD}} \geq R_{\mathrm{Sum}}^{\mathrm{AD}}$. Thus, the actual total achievable rate (in bit/s) for the downlink transmission, henceforth referred to as the sum of the data rate, on the wireless access downlink of the system for all the users, can be expressed by $R_{\mathrm{Tot}} = R_{\mathrm{Sum}}^{\mathrm{AD}}$. Therefore, the system EE of the downlink transmission, denoted by ξ_{EE} (in bit/Joule), can be formally defined as the ratio of the total achievable rate R_{Tot} (in bit/s) to the total power consumption P_{Tot} (in Watt), which is then calculated as follows:

$$\xi_{\mathrm{EE}} = \frac{R_{\mathrm{Tot}}}{P_{\mathrm{Tot}}} = \frac{R_{\mathrm{Sum}}^{\mathrm{AD}}}{P_{\mathrm{Tot}}}. \tag{17}$$

Under the above setup, our objective is to maximize the system EE of the downlink transmission while guaranteeing the data rate and power consumption requirements for the users, the APs, and the MBS, by the joint optimization of resource allocation and user association. Let $R_n^{\min}$ denote the minimum data rate for user n. We further employ $P^{\max}$ and $P_m^{\max}$ to stand for the maximum transmit power of the MBS and the maximum transmit power of AP m, respectively. Then the optimization problem can be mathematically formulated as:

$$(\text{P1}): \max_{\substack{\{a_{f,m,n,k},b_{f,m,k}\} \\ \left\{P^{\text{AP}}_{f,m,n,k},P^{\text{MBS}}_{f,m,k}\right\}}} \xi_{\text{EE}} = \frac{R^{\text{AD}}_{\text{Sum}}}{P_{\text{Tot}}} \tag{18a}$$

$$\text{s.t.} \quad \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi \log_2\left(1+\gamma^{\text{AD}}_{f,m,n,k}\right) \geq R_n^{\min}, \forall n, \tag{18b}$$

$$\sum_{f=1}^{F}\sum_{k=\delta+1}^{K} b_{f,m,k}\varpi \log_2\left(1+\gamma^{\text{BD}}_{f,m,k}\right) \geq \sum_{f=1}^{F}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi \log_2\left(1+\gamma^{\text{AD}}_{f,m,n,k}\right), \forall m, \tag{18c}$$

$$\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K} b_{f,m,k}P^{\text{MBS}}_{f,m,k} \leq P^{\max}, \forall f, \forall m, \forall k, \tag{18d}$$

$$\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}P^{\text{AP}}_{f,m,n,k} \leq P_m^{\max}, \forall f, \forall m, \tag{18e}$$

$$a_{f,m,n,k} \in \{0,1\}, \forall f, \forall m, \forall n, \forall k, \tag{18f}$$

$$b_{f,m,k} \in \{0,1\}, \forall f, \forall m, \forall k. \tag{18g}$$

With the constraint in (18b), the achievable rate of every user for wireless access via NOMA must satisfy its minimum data rate constraint. Constraint (18c) ensures that the achievable rate from the MBS to every AP for the backhaul connection via beamforming has to be greater than the wireless access rate from that AP to the users. Constraint (18d) is imposed to guarantee the maximum transmit power constraint for the MBS, and constraint (18e) indicates that the transmit power of every AP is restricted by its maximum power level. Finally, constraints (18f) and (18g) hold due to the definition of binary variable $a_{f,m,n,k}$ in the access downlink ($k \in \mathcal{A}$) and binary variable $b_{f,m,k}$ in the backhaul downlink ($k \in \mathcal{B}$), respectively, for $f \in \mathcal{F}$, $m \in \mathcal{M}$, and $n \in \mathcal{N}$.

4. Problem Analysis and Reformulation

In this section, we consider the solution to the optimization problem (P1) in (18) to find an optimal resource allocation and user association scheme for the system EE maximization of the downlink transmission. Clearly, the problem is a non-convex mixed-integer nonlinear programming problem due to the existence of the interference terms in the objective function in (18a), the nonlinear rate constraints in (18b) and (18c), and the binary-constrained variables in (18f) and (18g). Such kind of the problem

is NP-hard and computationally intractable. Especially, for the UDN scenario with larger numbers of densely distributed users and APs, it is extremely difficult to solve the problem directly with feasible lower computational complexity.

To efficiently solve the problem, we need to transform it into a more tractable convex optimization problem. Having this in mind, we first relax the binary variables into the continuous real variables to redesign some constraints for the problem reformulation. Then, we leverage the sum-of-ratios decoupling strategy to achieve the transformation of the fractional structure of the relaxed objective function into an equivalent parametric subtractive one. Lastly, we use the exponential-logarithmic transformation policy to construct a series of convex optimization subproblems, and further apply the method of the iterative successive convex approximation (SCA) to obtain the feasible lower-complexity solutions by iteratively tightening the lower bounds of the achievable sum rate functions.

4.1. Relaxation of Binary Variable

As noticed previously, binary variables $a_{f,m,n,k} \in \{0,1\}$ and $b_{f,m,k} \in \{0,1\}$ reflect the association relationship between the user and the AP on an assigned sub-channel k in the access downlink ($k \in \mathcal{A}$), and the association relationship between the MBS and the AP on an allocated sub-channel k in the backhaul downlink ($k \in \mathcal{B}$), respectively. That is, binary variable $a_{f,m,n,k}$ or $b_{f,m,k}$ can be interpreted as a user association-dependent indicator for assigning sub-channel k, i.e., the sub-channel allocation indicator. With the assigned sub-channel k, i.e., $a_{f,m,n,k} = 1$ or $b_{f,m,k} = 1$, the power can be allocated by AP m in cluster f to user n ($k \in \mathcal{A}$) or by the MBS to AP m in cluster f ($k \in \mathcal{B}$). Otherwise, the power will not allocated by the AP and the MBS over this sub-channel. Based on this insight along with the time-sharing relaxation mechanism [29], we turn to relax binary variables $a_{f,m,n,k}$ and $b_{f,m,k}$ to be two continuous real variables within the range of $[0,1]$, respectively. As a result, the actual power allocated by AP m in cluster f to user n on sub-channel k can be represented as $\widetilde{P}^{\mathrm{AP}}_{f,m,n,k} = a_{f,m,n,k} P^{\mathrm{AP}}_{f,m,n,k}$, for $f \in \mathcal{F}$, $m \in \mathcal{M}$, $n \in \mathcal{N}$, and $k \in \mathcal{A}$. Likewise, the actual power allocated by the MBS to AP m on sub-channel k in cluster f is expressed by $\widetilde{P}^{\mathrm{MBS}}_{f,m,k} = b_{f,m,k} P^{\mathrm{MBS}}_{f,m,k}$, for $f \in \mathcal{F}$, $m \in \mathcal{M}$, and $k \in \mathcal{B}$. In this case, the achievable sum rate of the system in (6) and (12) for the access downlink and the backhaul downlink can be respectively rewritten as:

$$\widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}} = \sum_{f=1}^{F}\sum_{n=1}^{N_f}\sum_{m=1}^{M_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi \log_2\left(1 + \frac{H_{f,m,n,k}\widetilde{P}^{\mathrm{AP}}_{f,m,n,k}}{1 + \sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\widetilde{P}^{\mathrm{AP}}_{f,j,n,k}}\right), \tag{19}$$

and

$$\widetilde{R}^{\mathrm{BD}}_{\mathrm{Sum}} = \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K} b_{f,m,k}\varpi \log_2\left(1 + \frac{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2 \widetilde{P}^{\mathrm{MBS}}_{f,m,k}}{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2 \sum_{j=1,j\neq m}^{\phi_{f,k}} \widetilde{P}^{\mathrm{MBS}}_{f,j,k} + \sigma^2_{m,k}}\right). \tag{20}$$

Accordingly, the total power consumption in (15) for the downlink transmission can be derived as:

$$\widetilde{P}_{\mathrm{Tot}} = \sum_{f=1}^{F}\sum_{m=1}^{M_f}\left(\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\left(P^{\mathrm{Con}}_{f,n} + P^{\mathrm{C}}_{m} + \widetilde{P}^{\mathrm{AP}}_{f,m,n,k}\right) + \sum_{k=\delta+1}^{K}\left(P^{\mathrm{Con}}_{f,m} + \widetilde{P}^{\mathrm{MBS}}_{f,m,k}\right)\right). \tag{21}$$

With such a relaxation process in mind, the original problem (P1) in (18) of particular interest can be reformulated as following problem:

$$(\text{P2}): \quad \max_{\substack{\{a_{f,m,n,k},b_{f,m,k}\} \\ \{\widetilde{P}^{\text{AP}}_{f,m,n,k},\widetilde{P}^{\text{MBS}}_{f,m,k}\}}} \quad \widetilde{\xi}_{\text{EE}} = \frac{\widetilde{R}^{\text{AD}}_{\text{Sum}}}{\widetilde{P}_{\text{Tot}}} \tag{22a}$$

$$\text{s.t.} \quad \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\log_2\left(1+\widetilde{\gamma}^{\text{AD}}_{f,m,n,k}\right) \geq \frac{R^{\min}_n}{\varpi}, \forall n, \tag{22b}$$

$$\sum_{f=1}^{F}\sum_{k=\delta+1}^{K} b_{f,m,k}\log_2\left(1+\widetilde{\gamma}^{\text{BD}}_{f,m,k}\right) \geq \sum_{f=1}^{F}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\log_2\left(1+\widetilde{\gamma}^{\text{AD}}_{f,m,n,k}\right), \forall m, \tag{22c}$$

$$\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K} \widetilde{P}^{\text{MBS}}_{f,m,k} \leq P^{\max}, \forall f, \forall m, \forall k, \tag{22d}$$

$$\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} \widetilde{P}^{\text{AP}}_{f,m,n,k} \leq P^{\max}_m, \forall f, \forall m, \tag{22e}$$

$$a_{f,m,n,k} \in [0,1], \forall f, \forall m, \forall n, \forall k, \tag{22f}$$

$$b_{f,m,k} \in [0,1], \forall f, \forall m, \forall k, \tag{22g}$$

where $\widetilde{\gamma}^{\text{AD}}_{f,m,n,k} = \frac{H_{f,m,n,k}\widetilde{P}^{\text{AP}}_{f,m,n,k}}{1+\sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\widetilde{P}^{\text{AP}}_{f,j,n,k}}$ and $\widetilde{\gamma}^{\text{BD}}_{f,m,k} = \frac{|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}|^2\widetilde{P}^{\text{MBS}}_{f,m,k}}{|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}|^2\sum_{j=1,j\neq m}^{\phi_{f,k}}\widetilde{P}^{\text{MBS}}_{f,j,k}+\sigma^2_{m,k}}$. We wish to remark that the optimal solution of the problem (P2) in (22) can be viewed as an upper bound of the solution to the original problem (P1) in (18) through the relaxed binary variables and constraints.

4.2. Equivalent Reformulation via Sum-of-Ratios Decoupling

Although the original problem (P1) in (18) has been transformed into a new one, we can easily find that the reformulated problem (P2) in (22) of particular interest is still not a convex problem. It is still rather challenging to derive an optimal solution for this problem due to the reasons: (i) the existence of the interference terms and the fractional component for the objective function in (22a), and (ii) the nonlinear and non-convex constraints in (22b) and (22c). Thus, we need to further convert this problem into an equivalent but more tractable one. Let us first recheck the structure of the objective function in (22a), which can be specifically rewritten by:

$$\widetilde{\xi}_{\text{EE}} = \sum_{f=1}^{F}\sum_{m=1}^{M_f} \frac{\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi\log_2\left(1+\frac{H_{f,m,n,k}\widetilde{P}^{\text{AP}}_{f,m,n,k}}{1+\sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\widetilde{P}^{\text{AP}}_{f,j,n,k}}\right)}{\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\left(P^{\text{Con}}_{f,n}+P^{\text{C}}_m+\widetilde{P}^{\text{AP}}_{f,m,n,k}\right)+\sum_{k=\delta+1}^{K}\left(P^{\text{Con}}_{f,m}+\widetilde{P}^{\text{MBS}}_{f,m,k}\right)}. \tag{23}$$

From (23), we can observe that the objective function holds the structure of a nonlinear sum of fractional functions. To maximize a sum of fractional functions subject to the non-convex constraints is

a sum-of-ratios fractional programming problem, which is difficult to solve by conventional optimization methods [30]. To address this problem, we attempt to adopt the sum-of-ratios algorithm by decoupling the numerators and denominators of the objective function with the fractional structure. More particularly, according to [30], the fractional form objective function of the problem (P2) in (22) is further reformulated into an equivalent parametric subtractive structure. Thereby, the optimization objective of the problem (P2) in (22) can be expressed as:

$$\max_{\left\{a_{f,m,n,k},b_{f,m,k},\widetilde{P}^{\mathrm{AP}}_{f,m,n,k},\widetilde{P}^{\mathrm{MBS}}_{f,m,k}\right\}} \widetilde{\xi}_{\mathrm{EE}} = \widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}} - \mu \widetilde{P}_{\mathrm{Tot}}, \tag{24}$$

where μ is an auxiliary parameter. So far, we break down the fractional structure of the objective function via the sum-of-ratios decoupling. Unfortunately, the objective function in (24) is still non-concave due to the interference terms in highly non-concave sum rate function $\widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}}$. To obtain the convex structure of the objective function, by the help of the feature of logarithmic structure, we can rewrite $\widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}}$ as the following difference of convex structures:

$$\begin{aligned}\widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}} = &\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi \log_2\left(1 + H_{f,m,n,k}\widetilde{P}^{\mathrm{AP}}_{f,m,n,k} + \sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\widetilde{P}^{\mathrm{AP}}_{f,j,n,k}\right) \\ &- \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi \log_2\left(1 + \sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\widetilde{P}^{\mathrm{AP}}_{f,j,n,k}\right).\end{aligned} \tag{25}$$

Through the above logarithmic operation, $\widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}}$ in the objective function of interest in (24) can be formulated as a sum of difference of convex functions. As a result, the reformulated problem (P2) in (22) can be further expressed by:

$$\begin{aligned}(\mathrm{P3}): \max_{\substack{\{a_{f,m,n,k},b_{f,m,k}\}\\ \left\{\widetilde{P}^{\mathrm{AP}}_{f,m,n,k},\widetilde{P}^{\mathrm{MBS}}_{f,m,k}\right\}}} \widetilde{\xi}_{\mathrm{EE}} = &\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi \log_2\left(1 + H_{f,m,n,k}\widetilde{P}^{\mathrm{AP}}_{f,m,n,k} + \sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\widetilde{P}^{\mathrm{AP}}_{f,j,n,k}\right) \\ &- \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi \log_2\left(1 + \sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\widetilde{P}^{\mathrm{AP}}_{f,j,n,k}\right) \\ &-\mu \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\left(P^{\mathrm{Con}}_{f,n} + P^{\mathrm{C}}_{m} + \widetilde{P}^{\mathrm{AP}}_{f,m,n,k}\right) \\ &-\mu \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K}\left(P^{\mathrm{Con}}_{f,m} + \widetilde{P}^{\mathrm{MBS}}_{f,m,k}\right) \qquad (26a)\end{aligned}$$

$$\text{s.t.} \quad (22b), (22c), (22d), (22e), (22f), (22g). \tag{26b}$$

4.3. Successive Convex Approximation

Apparently, the problem (P3) in (26) is not convex because the constraints in (22b) and (22c) is highly non-concave. To tackle such an issue, we resort to the iterative successive convex approximation (SCA) approach for solving the non-convex optimization problem, where, in each iteration, the original highly non-convex problem of particular interest is approximately transformed into a convex problem

[31]. According to [32, 33], by applying $\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k} = \frac{H_{f,m,n,k}\widetilde{P}^{\mathrm{AP}}_{f,m,n,k}}{1+\sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\widetilde{P}^{\mathrm{AP}}_{f,j,n,k}}$ into (19), a lower bound of $\widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}}$ can be characterized by:

$$\begin{aligned}\widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}} &= \sum_{f=1}^{F}\sum_{n=1}^{N_f}\sum_{m=1}^{M_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi \log_2\left(1+\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k}\right) \\ &\geq \sum_{f=1}^{F}\sum_{n=1}^{N_f}\sum_{m=1}^{M_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi \left(\alpha_{f,m,n,k}\log_2\left(\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k}\right)+\beta_{f,m,n,k}\right),\end{aligned} \tag{27}$$

where $\alpha_{f,m,n,k}$ and $\beta_{f,m,n,k}$ are the auxiliary approximation variables, respectively, for $f \in \mathcal{F}$, $m \in \mathcal{M}$, $n \in \mathcal{N}$, and $k \in \mathcal{A}$. When the following constants are satisfied, the approximation of $\widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}}$ is equivalent to or tight at the lower bound in (27)[6], i.e.:

$$\alpha_{f,m,n,k} = \frac{\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k}}{1+\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k}}, \tag{28}$$

$$\beta_{f,m,n,k} = \log_2\left(1+\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k}\right) - \frac{\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k}}{1+\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k}}\log_2\left(\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k}\right). \tag{29}$$

In the same way, by applying $\widetilde{\gamma}^{\mathrm{BD}}_{f,m,k} = \frac{|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}|^2\widetilde{P}^{\mathrm{MBS}}_{f,m,k}}{|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}|^2\sum_{j=1,j\neq m}^{\phi_{f,k}}\widetilde{P}^{\mathrm{MBS}}_{f,j,k}+\sigma^2_{m,k}}$ into (20), we can also obtain a lower bound of $\widetilde{R}^{\mathrm{BD}}_{\mathrm{Sum}}$, which is specified by:

$$\begin{aligned}\widetilde{R}^{\mathrm{BD}}_{\mathrm{Sum}} =& \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K} b_{f,m,k}\varpi\log_2\left(1+\widetilde{\gamma}^{\mathrm{BD}}_{f,m,k}\right) \\ &\geq \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K} b_{f,m,k}\varpi\left(\Lambda_{f,m,k}\log_2\left(\widetilde{\gamma}^{\mathrm{BD}}_{f,m,k}\right)+\Xi_{f,m,k}\right),\end{aligned} \tag{30}$$

where $\Lambda_{f,m,k}$ and $\Xi_{f,m,k}$ are the auxiliary approximation variables, respectively, for $f \in \mathcal{F}$, $m \in \mathcal{M}$, and $k \in \mathcal{B}$. When the following constants are satisfied, the approximation of $\widetilde{R}^{\mathrm{BD}}_{\mathrm{Sum}}$ is further achieved as the lower bound in (30), i.e.:

$$\Lambda_{f,m,k} = \frac{\widetilde{\gamma}^{\mathrm{BD}}_{f,m,k}}{1+\widetilde{\gamma}^{\mathrm{BD}}_{f,m,k}}, \tag{31}$$

$$\Xi_{f,m,k} = \log_2\left(1+\widetilde{\gamma}^{\mathrm{BD}}_{f,m,k}\right) - \frac{\widetilde{\gamma}^{\mathrm{BD}}_{f,m,k}}{1+\widetilde{\gamma}^{\mathrm{BD}}_{f,m,k}}\log_2\left(\widetilde{\gamma}^{\mathrm{BD}}_{f,m,k}\right). \tag{32}$$

For the given approximation variables $\alpha_{f,m,n,k}$, $\beta_{f,m,n,k}$, $\Lambda_{f,m,k}$, and $\Xi_{f,m,k}$, we then transform the problem (P3) in (26) into an approximated one, i.e.:

$$(\mathrm{P4}): \max_{\substack{\{a_{f,m,n,k},b_{f,m,k}\}\\ \left\{\widetilde{P}^{\mathrm{AP}}_{f,m,n,k},\widetilde{P}^{\mathrm{MBS}}_{f,m,k}\right\}}} \widetilde{\xi}_{\mathrm{EE}} = \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi\left(\alpha_{f,m,n,k}\log_2\left(\widetilde{\gamma}^{\mathrm{AD}}_{f,m,n,k}\right)+\beta_{f,m,n,k}\right)$$

[6]Note that the use of the logarithmic approximation makes a relaxation of highly non-concave sum rate function $\widetilde{R}^{\mathrm{AD}}_{\mathrm{Sum}}$ achieve the lower bound when both of the approximation constants are guaranteed. That is, the lower bound is said to be a tight lower bound.

$$-\mu\left(\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\left(P_{f,n}^{\mathrm{Con}}+P_{m}^{\mathrm{C}}+\widetilde{P}_{f,m,n,k}^{\mathrm{AP}}\right)\right.$$

$$\left.+\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K}\left(P_{f,m}^{\mathrm{Con}}+\widetilde{P}_{f,m,k}^{\mathrm{MBS}}\right)\right) \tag{33a}$$

$$\text{s.t.}\quad \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=1}^{\delta}a_{f,m,n,k}\left(\alpha_{f,m,n,k}\log_2\left(\widetilde{\gamma}_{f,m,n,k}^{\mathrm{AD}}\right)+\beta_{f,m,n,k}\right)\geq\frac{R_n^{\min}}{\varpi},\ \forall n, \tag{33b}$$

$$\begin{aligned}&\sum_{f=1}^{F}\sum_{k=\delta+1}^{K}b_{f,m,k}\left(\Lambda_{f,m,k}\log_2\left(\widetilde{\gamma}_{f,m,k}^{\mathrm{BD}}\right)+\Xi_{f,m,k}\right)\\&\geq\sum_{f=1}^{F}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}a_{f,m,n,k}\left(\alpha_{f,m,n,k}\log_2\left(\widetilde{\gamma}_{f,m,n,k}^{\mathrm{AD}}\right)+\beta_{f,m,n,k}\right),\ \forall m,\end{aligned} \tag{33c}$$

$$\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K}\widetilde{P}_{f,m,k}^{\mathrm{MBS}}\leq P^{\max},\ \forall f,\forall m,\forall k, \tag{33d}$$

$$\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\widetilde{P}_{f,m,n,k}^{\mathrm{AP}}\leq P_m^{\max},\ \forall f,\forall m, \tag{33e}$$

$$a_{f,m,n,k}\in[0,1],\ \forall f,\forall m,\forall n,\forall k, \tag{33f}$$

$$b_{f,m,k}\in[0,1],\ \forall f,\forall m,\forall k. \tag{33g}$$

Apparently, the problem (P4) in (33) is still non-concave. To address this issue, we intend to exploit the exponential-logarithmic transformation method to achieve the logarithmic change of variables, i.e., $\widehat{P}_{f,m,n,k}^{\mathrm{AP}}=\log_2\left(\widetilde{P}_{f,m,n,k}^{\mathrm{AP}}\right)$, for $f\in\mathcal{F}$, $m\in\mathcal{M}$, $n\in\mathcal{N}$, and $k\in\mathcal{A}$, and $\widehat{P}_{f,m,k}^{\mathrm{MBS}}=\log_2\left(\widetilde{P}_{f,m,k}^{\mathrm{MBS}}\right)$, for $f\in\mathcal{F}$, $m\in\mathcal{M}$, and $k\in\mathcal{B}$. For the exponential structure, we have $\widetilde{P}_{f,m,n,k}^{\mathrm{AP}}=\exp\left(\widehat{P}_{f,m,n,k}^{\mathrm{AP}}\right)$ and $\widetilde{P}_{f,m,k}^{\mathrm{MBS}}=\exp\left(\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right)$. To this end, by applying the logarithmic change of variables into a logarithmic transformation of the objective and constraint functions, we arrive at the following approximate parametric subproblem:

$$\begin{aligned}(\mathrm{P5}):\ \max_{\substack{\{a_{f,m,n,k},b_{f,m,k}\}\\ \left\{\widehat{P}_{f,m,n,k}^{\mathrm{AP}},\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right\}}}\ &\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}a_{f,m,n,k}\varpi\left(\alpha_{f,m,n,k}\log_2\left(\widehat{\gamma}_{f,m,n,k}^{\mathrm{AD}}\right)+\beta_{f,m,n,k}\right)\\&-\mu\left(\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\left(P_{f,n}^{\mathrm{Con}}+P_{m}^{\mathrm{C}}+\exp\left(\widehat{P}_{f,m,n,k}^{\mathrm{AP}}\right)\right)\right.\\&\left.+\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K}\left(P_{f,m}^{\mathrm{Con}}+\exp\left(\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right)\right)\right)\end{aligned} \tag{34a}$$

$$\text{s.t.}\quad \sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=1}^{\delta}a_{f,m,n,k}\left(\alpha_{f,m,n,k}\log_2\left(\widehat{\gamma}_{f,m,n,k}^{\mathrm{AD}}\right)+\beta_{f,m,n,k}\right)\geq\frac{R_n^{\min}}{\varpi},\ \forall n, \tag{34b}$$

$$\sum_{f=1}^{F}\sum_{k=\delta+1}^{K} b_{f,m,k}\left(\Lambda_{f,m,k}\log_2\left(\widehat{\gamma}_{f,m,k}^{\mathrm{BD}}\right)+\Xi_{f,m,k}\right) \geq \sum_{f=1}^{F}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\left(\alpha_{f,m,n,k}\log_2\left(\widehat{\gamma}_{f,m,n,k}^{\mathrm{AD}}\right)+\beta_{f,m,n,k}\right),\ \forall m, \tag{34c}$$

$$\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K}\exp\left(\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right)\leq P^{\max},\ \forall f,\forall m,\forall k, \tag{34d}$$

$$\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\exp\left(\widehat{P}_{f,m,n,k}^{\mathrm{AP}}\right)\leq P_m^{\max},\ \forall f,\forall m, \tag{34e}$$

$$a_{f,m,n,k}\in[0,1],\ \forall f,\forall m,\forall n,\forall k, \tag{34f}$$

$$b_{f,m,k}\in[0,1],\ \forall f,\forall m,\forall k, \tag{34g}$$

where $\log_2\left(\widehat{\gamma}_{f,m,n,k}^{\mathrm{AD}}\right) = \widehat{P}_{f,m,n,k}^{\mathrm{AP}} + \log_2\left(H_{f,m,n,k}\right) - \log_2\left(1+\sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\exp\left(\widehat{P}_{f,m,n,k}^{\mathrm{AP}}\right)\right)$ and $\log_2\left(\widehat{\gamma}_{f,m,k}^{\mathrm{BD}}\right)=\widehat{P}_{f,m,k}^{\mathrm{MBS}}+\log_2\left(\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2\right)-\log_2\left(\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2\sum_{j=1,j\neq m}^{\phi_{f,k}}\exp\left(\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right)+\sigma_{m,k}^2\right)$. It should be pointed out that the approximate subproblem (P5) in (34) follows the log-sum-exp function structure after the exponential-logarithmic transformation. Given the fact that the log-sum-exp function is strictly convex [34], we finally convert the original problem (P1) in (18) of particular interest into a standard convex maximization problem with logarithmic change variables.

For convex problem, lots of traditional convex optimization solutions can be used to solve it. In fact, we only maximize a lower bound of the objective function of (34a). To eventually solve the subproblem (P5) in (34), by help of the iterative SCA approach, we need to further tighten the bound in (27) by iteratively updating $\alpha_{f,m,n,k}$ in (28) and $\beta_{f,m,n,k}$ in (29), and meanwhile tighten the bound in (30) by iteratively updating $\Lambda_{f,m,k}$ in (31) and $\Xi_{f,m,k}$ in (32). After obtaining the optimal solution of (34), through the exponential transformation, we then derive the relaxed binary variables $\widetilde{P}_{f,m,n,k}^{\mathrm{AP}} = \exp\left(\widehat{P}_{f,m,n,k}^{\mathrm{AP}}\right)$ and $\widetilde{P}_{f,m,k}^{\mathrm{MBS}} = \exp\left(\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right)$, namely, the optimal power allocated by AP m in cluster f to user n on sub-channel k as well as the optimal power allocated by the MBS to AP m on sub-channel k in cluster f.

The detailed procedure of the adopted iterative algorithm via the SCA method to tighten the bounds in (27) and (30) is summarized in Algorithm 1. It is noteworthy that Algorithm 1 is implemented in an iterative way for each AP and the MBS, and is also distributed with guaranteed convergence and low complexity. For each iteration, approximation variables $\alpha_{f,m,n,k}^{(\tau+1)}$, $\beta_{f,m,n,k}^{(\tau+1)}$, $\Lambda_{f,m,k}^{(\tau+1)}$, and $\Xi_{f,m,k}^{(\tau+1)}$ are always better than the previous values $\alpha_{f,m,n,k}^{(\tau)}$, $\beta_{f,m,n,k}^{(\tau)}$, $\Lambda_{f,m,k}^{(\tau)}$, and $\Xi_{f,m,k}^{(\tau)}$. These bounds will be improved successively during each iteration, and the iterative process will terminate after finite iterations. So far, we have transformed the original problem (P1) in (18) into a sequence of convex maximization subproblems (P5) in (34) through the exponential-logarithmic approximation. In the following section, we will design an effective algorithm to solve the subproblem (P5) in (34) for obtaining the optimal solutions, aiming to achieve the joint power, sub-channel allocation, and user association in reasonable time complexity.

Algorithm 1 SCA-Based Iterative Algorithm for Approximation Variable Updating.

Initialization: Maximum number of iterations $I^{\max}$ and maximum tolerance $\varepsilon > 0$.
Set approximation variables $\alpha_{f,m,n,k}^{(0)} = 1$, $\beta_{f,m,n,k}^{(0)} = 0$, $\Lambda_{f,m,k}^{(0)} = 1$, $\Xi_{f,m,k}^{(0)} = 0$.
Set iteration index $\tau = 0$.
while $\left|\alpha_{f,m,n,k}^{(\tau+1)} - \alpha_{f,m,n,k}^{(\tau)}\right| > \varepsilon \parallel \left|\beta_{f,m,n,k}^{(\tau+1)} - \beta_{f,m,n,k}^{(\tau)}\right| > \varepsilon \parallel \left|\Lambda_{f,m,k}^{(\tau+1)} - \Lambda_{f,m,k}^{(\tau)}\right| > \varepsilon \parallel \left|\Xi_{f,m,k}^{(\tau+1)} - \Xi_{f,m,k}^{(\tau)}\right| > \varepsilon \parallel \tau < I^{\max}$ **do**
  Solve subproblem (P5) in (34) to obtain optimal solutions $\widetilde{P}_{f,m,n,k}^{\mathrm{AP}(\tau)} = \exp\left(\widehat{P}_{f,m,n,k}^{\mathrm{AP}(\tau)}\right)$ and $\widetilde{P}_{f,m,k}^{\mathrm{MBS}(\tau)} = \exp\left(\widehat{P}_{f,m,k}^{\mathrm{MBS}(\tau)}\right)$.
  Update $\alpha_{f,m,n,k}^{(\tau+1)}$ and $\beta_{f,m,n,k}^{(\tau+1)}$ to tighten the bound in (27) according to (28) and (29).
  Update $\Lambda_{f,m,k}^{(\tau+1)}$ and $\Xi_{f,m,k}^{(\tau+1)}$ to tighten the bound in (30) according to (31) and (32).
  $\tau = \tau + 1$.
end while
return $\alpha_{f,m,n,k}^{(\tau+1)}$, $\beta_{f,m,n,k}^{(\tau+1)}$, $\Lambda_{f,m,k}^{(\tau+1)}$, and $\Xi_{f,m,k}^{(\tau+1)}$.

5. Lagrangian Dual Decomposition and Optimal Solution

5.1. Lagrangian Dual Decomposition

Since the subproblem (P5) in (34) is a standard convex maximization problem after the SCA process, we can adopt the Lagrangian dual decomposition method to solve it to obtain the optimal sub-channel and power allocation for energy efficient user association. The detailed procedure is given in the following. The Lagrangian function corresponding to the subproblem (P5) in (34) can be expressed by:

$$
\begin{aligned}
&L\left(\{a_{f,m,n,k}\},\{b_{f,m,k}\}\left\{\widehat{P}_{f,m,n,k}^{\mathrm{AP}}\right\},\left\{\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right\},\boldsymbol{\lambda},\boldsymbol{\varphi},\eta,\boldsymbol{\chi}\right)\\
=&\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\varpi\left(\alpha_{f,m,n,k}\log_2\left(\widehat{\gamma}_{f,m,n,k}^{\mathrm{AD}}\right)+\beta_{f,m,n,k}\right)\\
&-\mu\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\left(P_{f,n}^{\mathrm{Con}}+P_{m}^{\mathrm{C}}+\exp\left(\widehat{P}_{f,m,n,k}^{\mathrm{AP}}\right)\right)\\
&-\mu\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K}\left(P_{f,m}^{\mathrm{Con}}+\exp\left(\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right)\right)\\
&+\sum_{n=1}^{N_f}\lambda_n\left(\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\left(\alpha_{f,m,n,k}\log_2\left(\widehat{\gamma}_{f,m,n,k}^{\mathrm{AD}}\right)+\beta_{f,m,n,k}\right)-\frac{R_n^{\min}}{\varpi}\right)\\
&+\sum_{m=1}^{M_f}\varphi_m\left(\sum_{f=1}^{F}\sum_{k=\delta+1}^{K} b_{f,m,k}\left(\Lambda_{f,m,k}\log_2\left(\widehat{\gamma}_{f,m,k}^{\mathrm{BD}}\right)+\Xi_{f,m,k}\right)\right.\\
&\qquad\left.-\sum_{f=1}^{F}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta} a_{f,m,n,k}\left(\alpha_{f,m,n,k}\log_2\left(\widehat{\gamma}_{f,m,n,k}^{\mathrm{AD}}\right)+\beta_{f,m,n,k}\right)\right)\\
&+\eta\left(P^{\max}-\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K}\exp\left(\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right)\right)\\
&+\sum_{f=1}^{F}\sum_{m=1}^{M_f}\chi_{f,m}\left(P_m^{\max}-\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\exp\left(\widehat{P}_{f,m,n,k}^{\mathrm{AP}}\right)\right),
\end{aligned}
\tag{35}
$$

where $\boldsymbol{\lambda}$ is the Lagrange multiplier (i.e., the dual variable) vector associated with constraint (34b) on the minimum data rate requirement for each user, $\boldsymbol{\varphi}$ is the Lagrange multiplier vector for constraint (34c) on the achievable rate between the backhaul connection and the wireless access of each AP, η is the Lagrange multiplier corresponding to constraint (34d) on the maximum transmit power for the MBS, and $\boldsymbol{\chi}$ is the Lagrange multiplier vector accounting for constraint (34e) on the maximum transmit power of each AP. The boundary constraints (34f) and (34g) will be absorbed in the Karush-Kuhn-Tucker (KKT) conditions [34]. Thereby, the Lagrange dual function is obtained as:

$$g\left(\boldsymbol{\lambda},\boldsymbol{\varphi},\eta,\boldsymbol{\chi}\right)=\max_{\substack{\{a_{f,m,n,k}\},\{b_{f,m,k}\}\\ \left\{\widehat{P}^{\mathrm{AP}}_{f,m,n,k}\right\},\left\{\widehat{P}^{\mathrm{MBS}}_{f,m,k}\right\}}} L\left(\{a_{f,m,n,k}\},\{b_{f,m,k}\}\left\{\widehat{P}^{\mathrm{AP}}_{f,m,n,k}\right\},\left\{\widehat{P}^{\mathrm{MBS}}_{f,m,k}\right\},\boldsymbol{\lambda},\boldsymbol{\varphi},\eta,\boldsymbol{\chi}\right). \tag{36}$$

Thus, the Lagrangian dual problem can be represented by:

$$\min_{\boldsymbol{\lambda},\boldsymbol{\varphi},\eta,\boldsymbol{\tau}} g\left(\boldsymbol{\lambda},\boldsymbol{\varphi},\eta,\boldsymbol{\chi}\right) \tag{37a}$$

$$\text{s.t.}\quad \boldsymbol{\lambda},\boldsymbol{\varphi},\eta,\boldsymbol{\chi}\geq 0. \tag{37b}$$

Due to the differentiability of the Lagrange dual function, we then perform the update process of the Lagrange dual multipliers in (37) based on the subgradient method to minimize the dual. Let l and $L^{\max}$ stand for the iteration index and the maximum number of iterations for the dual multiplier update process, respectively. Concretely, in the $(l+1)$-th iteration, for $l=1,2,\cdots,L^{\max}$, the dual multipliers can be independently updated by:

$$\lambda_n^{(l+1)}=\lambda_n^{(l)}-\zeta_\lambda^{(l)}\left(\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=1}^{\delta}a_{f,m,n,k}\left(\alpha_{f,m,n,k}\log_2\left(\widehat{\gamma}^{\mathrm{AD}}_{f,m,n,k}\right)+\beta_{f,m,n,k}\right)-\frac{R_n^{\min}}{\varpi}\right),\forall n, \tag{38}$$

$$\begin{aligned}\varphi_m^{(l+1)}=\varphi_m^{(l)}-\zeta_\varphi^{(l)}\Bigg(&\sum_{f=1}^{F}\sum_{k=\delta+1}^{K}b_{f,m,k}\left(\Lambda_{f,m,k}\log_2\left(\widehat{\gamma}^{\mathrm{BD}}_{f,m,k}\right)+\Xi_{f,m,k}\right)\\&-\sum_{f=1}^{F}\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}a_{f,m,n,k}\left(\alpha_{f,m,n,k}\log_2\left(\widehat{\gamma}^{\mathrm{AD}}_{f,m,n,k}\right)+\beta_{f,m,n,k}\right)\Bigg),\forall m,\end{aligned} \tag{39}$$

$$\eta^{(l+1)}=\eta^{(l)}-\zeta_\eta^{(l)}\left(P^{\max}-\sum_{f=1}^{F}\sum_{m=1}^{M_f}\sum_{k=\delta+1}^{K}\exp\left(\widehat{P}^{\mathrm{MBS}}_{f,m,k}\right)\right), \tag{40}$$

$$\chi_{f,m}^{(l+1)}=\chi_{f,m}^{(l)}-\zeta_\chi^{(l)}\left(P_m^{\max}-\sum_{n=1}^{N_f}\sum_{k=1}^{\delta}\exp\left(\widehat{P}^{\mathrm{AP}}_{f,m,n,k}\right)\right),\forall f,\forall m, \tag{41}$$

where $\zeta_\lambda^{(l)}$, $\zeta_\varphi^{(l)}$, $\zeta_\eta^{(l)}$, and $\zeta_\chi^{(l)}$ are the step sizes at the (l)-th iteration for dual multipliers λ_n, φ_m, η, and $\chi_{f,m}$, respectively. Additionally, the step size for each dual multiplier should satisfy the following conditions:

$$\sum_{l=1}^{\infty} \zeta_{\lambda}^{(l)} = \infty,\ \lim_{l\to\infty} \zeta_{\lambda}^{(l)} = 0,\ \text{for}\, \lambda_n, \forall n, \tag{42}$$

$$\sum_{l=1}^{\infty} \zeta_{\varphi}^{(l)} = \infty,\ \lim_{l\to\infty} \zeta_{\varphi}^{(l)} = 0,\ \text{for}\, \varphi_m, \forall m, \tag{43}$$

$$\sum_{l=1}^{\infty} \zeta_{\eta}^{(l)} = \infty,\ \lim_{l\to\infty} \zeta_{\eta}^{(l)} = 0,\ \text{for}\, \eta, \tag{44}$$

$$\sum_{l=1}^{\infty} \zeta_{\chi}^{(l)} = \infty,\ \lim_{l\to\infty} \zeta_{\chi}^{(l)} = 0,\ \text{for}\, \chi_{f,m}, \forall f, \forall m. \tag{45}$$

5.2. Optimal Solution for Joint Resource Allocation and User Association

We are now ready to enumerate the KKT conditions. Let us use $\left\{P_{f,m,n,k}^{*\mathrm{AP}}\right\}$, $\left\{P_{f,m,k}^{*\mathrm{MBS}}\right\}$, $\left\{a_{f,m,n,k}^{*}\right\}$, and $\left\{b_{f,m,k}^{*}\right\}$ to represent the optimal solutions to the subproblem (P5) in (34), respectively. According to the KKT conditions, upon taking the partial derivative of the Lagrangian function $L\left(\cdots\right)$ with respect to $\widehat{P}_{f,m,n,k}^{\mathrm{AP}}$ and $\widehat{P}_{f,m,k}^{\mathrm{MBS}}$ in (35), respectively, the optimal solutions $P_{f,m,n,k}^{*\mathrm{AP}}$ and $P_{f,m,k}^{*\mathrm{MBS}}$ to the subproblem (P5) in (34) can be respectively obtained as:

$$\frac{\partial L\left(\cdots\right)}{\partial P_{f,m,n,k}^{*\mathrm{AP}}} = a_{f,m,n,k}\alpha_{f,m,n,k}\left(\varpi + \lambda_n - \varphi_m\right)\left(1 - \frac{\sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\exp\left(P_{f,m,n,k}^{*\mathrm{AP}}\right)}{\left(1+\sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\exp\left(P_{f,m,n,k}^{*\mathrm{AP}}\right)\right)\ln 2}\right)$$
$$-\left(\mu + \chi_{f,m}\right)\exp\left(P_{f,m,n,k}^{*\mathrm{AP}}\right) = 0, \tag{46}$$

and

$$\frac{\partial L\left(\cdots\right)}{\partial P_{f,m,k}^{*\mathrm{MBS}}} = \varphi_m b_{f,m,k}\Lambda_{f,m,k}\left(1 - \frac{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2\sum_{j=1,j\neq m}^{\phi_{f,k}}\exp\left(P_{f,m,k}^{*\mathrm{MBS}}\right)}{\left(\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2\sum_{j=1,j\neq m}^{\phi_{f,k}}\exp\left(P_{f,m,k}^{*\mathrm{MBS}}\right)+\sigma_{m,k}^2\right)\ln 2}\right)$$
$$-\left(\mu+\eta\right)\exp\left(\widehat{P}_{f,m,k}^{\mathrm{MBS}}\right) = 0. \tag{47}$$

After some necessary algebraic manipulations, we then easily obtain the optimal power allocated by AP m in cluster f to user n on sub-channel k, for $f \in \mathcal{F}$, $m \in \mathcal{M}$, $n \in \mathcal{N}$, and $k \in \mathcal{A}$, and the optimal power allocated by the MBS to AP m on sub-channel k in cluster f, for $f \in \mathcal{F}$, $m \in \mathcal{M}$, and $k \in \mathcal{B}$, which can be given as follows:

$$P_{f,m,n,k}^{*\mathrm{AP}} = \ln\left(\frac{a_{f,m,n,k}\alpha_{f,m,n,k}\left(\varphi_m - \varpi - \lambda_n\right)}{\mu+\chi_{f,m}}\cdot\left(1 - \frac{\sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\exp\left(P_{f,m,n,k}^{*\mathrm{AP}}\right)}{\left(1+\sum_{j=m+1}^{M_{f,k}} H_{f,j,n,k}\exp\left(P_{f,m,n,k}^{*\mathrm{AP}}\right)\right)\ln 2}\right)\right), \tag{48}$$

and

$$P_{f,m,k}^{*\mathrm{MBS}} = \ln\left(\frac{\varphi_m b_{f,m,k}\Lambda_{f,m,k}}{\mu+\eta}\left(1 - \frac{\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2\sum_{j=1,j\neq m}^{\phi_{f,k}}\exp\left(P_{f,m,k}^{*\mathrm{MBS}}\right)}{\left(\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2\sum_{j=1,j\neq m}^{\phi_{f,k}}\exp\left(P_{f,m,k}^{*\mathrm{MBS}}\right)+\sigma_{m,k}^2\right)\ln 2}\right)\right). \tag{49}$$

It is noticeable that there does not exist a derived closed-form expression of the optimal power allocation

values from (48) and (49). However, the existence and uniqueness of the optimal power allocation $P^{*\mathrm{AP}}_{f,m,n,k}$ and $P^{*\mathrm{MBS}}_{f,m,k}$ are guaranteed according to [33]. Due to the space limitation, the specific detail about the strict mathematical proof of the existence and uniqueness of the optimal power allocation is omitted here, and readers can refer to [33] for more detailed description. Besides, we also would like to mention that the update of the optimal power allocation can be made locally by each AP and the MBS, respectively, via iteratively updating dual multipliers λ_n, φ_m, η, and $\chi_{f,m}$.

Meanwhile, according to the KKT conditions, upon taking the partial derivative of the Lagrangian function $L(\cdots)$ with respect to $a_{f,m,n,k}$ and $b_{f,m,k}$ in (35), respectively, the optimal solutions $a^*_{f,m,n,k}$ and $b^*_{f,m,k}$ to the subproblem (P5) in (34) can be respectively calculated by:

$$\begin{aligned}\frac{\partial L(\cdots)}{\partial a^*_{f,m,n,k}} =& (\varpi+\lambda_n-\varphi_m)\left(\alpha_{f,m,n,k}P^{*\mathrm{AP}}_{f,m,n,k}+\alpha_{f,m,n,k}\log_2\left(H_{f,m,n,k}\right)+\beta_{f,m,n,k}\right)\\ &-\alpha_{f,m,n,k}(\varpi+\lambda_n-\varphi_m)\log_2\left(1+\sum_{j=m+1}^{M_{f,k}}H_{f,j,n,k}\exp\left(P^{*\mathrm{AP}}_{f,m,n,k}\right)\right)\\ =&\begin{cases}<0 & a^*_{f,m,n,k}=0,\\ =0 & 0<a^*_{f,m,n,k}<1\\ >0 & a^*_{f,m,n,k}=1,\end{cases}\end{aligned} \tag{50}$$

and

$$\begin{aligned}\frac{\partial L(\cdots)}{\partial b^*_{f,m,k}} =& \varphi_m\left(\Lambda_{f,m,k}P^{*\mathrm{MBS}}_{f,m,k}+\Lambda_{f,m,k}\log_2\left(\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2\right)+\Xi_{f,m,k}\right)\\ &-\varphi_m\Lambda_{f,m,k}\log_2\left(\left|\mathbf{h}_{f,m,k}\mathbf{w}_{f,k}\right|^2\sum_{j=1,j\neq m}^{\phi_{f,k}}\exp\left(\widehat{P}^{\mathrm{MBS}}_{f,m,k}\right)+\sigma^2_{m,k}\right)\\ =&\begin{cases}<0 & b^*_{f,m,k}=0,\\ =0 & 0<b^*_{f,m,k}<1,\\ >0 & b^*_{f,m,k}=1.\end{cases}\end{aligned} \tag{51}$$

Therefore, sub-channel k^* is assigned to user n by AP m in cluster f via performing the maximization operation of $\frac{\partial L(\cdots)}{\partial a^*_{f,m,n,k}}$ in (50), for $f\in\mathcal{F}$, $m\in\mathcal{M}$, $n\in\mathcal{N}$, and $k\in\mathcal{A}$, such that we have $a^*_{f,m,n,k^*}=1$, which is further expressed as:

$$a^*_{f,m,n,k^*}\Big|_{k^*=\arg\max\limits_{k}\frac{\partial L(\cdots)}{\partial a^*_{f,m,n,k}}}=1. \tag{52}$$

Similarly, sub-channel k^* is also assigned to AP m in cluster f by the MBS via performing the maximization operation of $\frac{\partial L(\cdots)}{\partial b^*_{f,m,k}}$ in (51), for $f\in\mathcal{F}$, $m\in\mathcal{M}$, and $k\in\mathcal{B}$, such that we obtain $b^*_{f,m,k^*}=1$, which can be specified by:

$$b^*_{f,m,k^*}\Big|_{k^*=\arg\max\limits_{k}\frac{\partial L(\cdots)}{\partial b^*_{f,m,k}}}=1. \tag{53}$$

From (52) and (53), it suffices to mention that an assignment of 1 to either $a^*_{f,m,n,k}$ or $b^*_{f,m,k}$ not only achieves the optimal sub-channel allocation to each user or each AP, but also indicates the determination of user association index, namely, the association relation for the user-AP or the AP-MBS.

Algorithm 2 Joint Power Allocation, Sub-channel Assignment, and User Association Algorithm.

Initialization: User's minimum rate $R_n^{\min}$, MBS's maximum power $P^{\max}$, AP's maximum power $P_m^{\max}$, and maximum number of iterations $L^{\max}$.
Set Lagrange multipliers $\lambda_n^{(1)} = 0$, $\varphi_m^{(1)} = 0$, $\eta^{(1)} = 0$, and $\chi_{f,m}^{(1)} = 0$.
Set iteration index $l = 1$.
Obtain updated approximation variables $\alpha_{f,m,n,k}$, $\beta_{f,m,n,k}$, $\Lambda_{f,m,k}$, and $\Xi_{f,m,k}$ using **Algorithm 1**.
repeat
  Sub-channel Assignment and User Association:
  for $f = 1$ to F **do**
    for $m = 1$ to M **do**
      Calculate sub-channel k^* for the AP-MBS association b_{f,m,k^*}^* according to (53).
      Use sub-channel k^* in Step 9 to update b_{f,m,k^*}^*.
      for $n = 1$ to N **do**
        Calculate sub-channel k^* for the user-AP association a_{f,m,n,k^*}^* according to (52).
        Use sub-channel k^* in Step 12 to update a_{f,m,n,k^*}^*.
      end for
    end for
  end for
  Power Allocation:
  for $f = 1$ to F **do**
    for $m = 1$ to M **do**
      Calculate $\frac{\partial L(\cdots)}{\partial P_{f,m,k^*}^{*\mathrm{MBS}}}$ to update power allocation $P_{f,m,k^*}^{*\mathrm{MBS}}$ according to (49).
      if $\sum_{f=1}^{F}\sum_{m=1}^{M_f} b_{f,m,k^*}^* P_{f,m,k^*}^{*\mathrm{MBS}} > P^{\max}$ **then**
        $P_{f,m,k^*}^{*\mathrm{MBS}} = P^{\max}$.
      end if
      for $n = 1$ to N **do**
        Calculate $\frac{\partial L(\cdots)}{\partial P_{f,m,n,k^*}^{*\mathrm{AP}}}$ to update power allocation $P_{f,m,n,k^*}^{*\mathrm{AP}}$ according to (48).
        if $\sum_{n=1}^{N_f} a_{f,m,n,k^*}^* P_{f,m,n,k^*}^{*\mathrm{AP}} > P_m^{\max}$ **then**
          $P_{f,m,n,k^*}^{*\mathrm{AP}} = P_m^{\max}$.
        end if
      end for
    end for
  end for
  Update Lagrange multipliers $\lambda_n^{(l+1)}$, $\varphi_m^{(l+1)}$, $\eta^{(l+1)}$, and $\chi_{f,m}^{(l+1)}$ according to (38), (39), (40), and (41) under the step size constraint of (42), (43), (44), and (45).
  $l = l + 1$.
until onvergence **or** $l = L^{\max}$
return $P_{f,m,n,k^*}^{*\mathrm{AP}}$, $P_{f,m,k^*}^{*\mathrm{MBS}}$, a_{f,m,n,k^*}^*, and b_{f,m,k^*}^*.

So far, we have devised Algorithm 1 to generate the updated approximation variables used for tightening the bounds in (27) and (30), and also have given a solution for the joint resource allocation and user association problem by incorporating the approximation variables as well as the iteratively updated dual multipliers. By taking the advantage of the Lagrangian dual decomposition, we still need to devise an effective algorithm to identify a specific execution coordination between power allocation and sub-channel assignment and further to ensure fast convergence of the update of optimal power. As a result, we present a distributed iterative algorithm to realize the joint optimization of power allocation, sub-channel assignment, and user association simultaneously, which is sketched in the Algorithm 2.

In Algorithm 2, the Lagrange multipliers are firstly assumed to an fixed value after the setup of initialization. Then, the approximation variables are obtained by using Algorithms 1. Then, the algorithm undertakes the iterative process. In each iterative process, each user and each AP can distributvely update the corresponding user association index by using the assigned sub-channels. Based on the results of the optimal sub-channel assignment and user association, each AP and each MBS can also update their

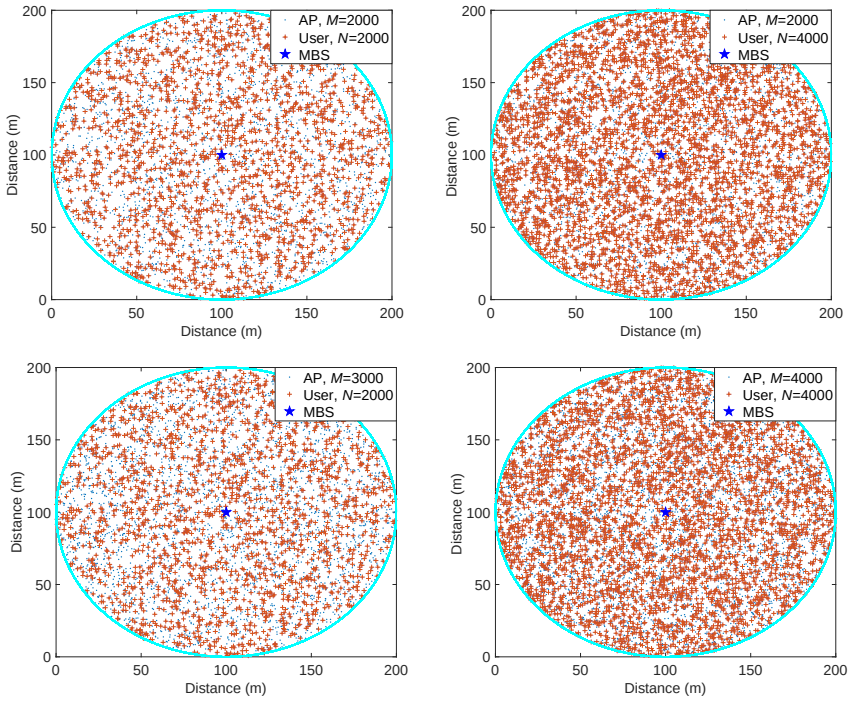

Figure 2: Simulation scenario of the user-centric UDN: M APs with independent homogeneous PPP Φ_{ρ_1} and N randomly generated users coexisting in a circular macrocell area with radius $r = 100$ m. The MBS is located in the center of the macrocell. Here, four typical deployment scenarios based on different combination relations between M and N are generated, respectively, i.e., (i) $M = 2000$, $N = 2000$, (ii) $M = 2000$, $N = 4000$, (iii) $M = 3000$, $N = 2000$, (iv) $M = 4000$, $N = 4000$.

transmit powers in a distributed manner. By updating the sub-channel assignment and user association as well as the power allocation alternatively, the iteration process is terminated when the convergence of the update of optimal power is guaranteed or the maximum number of iterations is reached.

6. Simulation Results

In this section, we conduct simulation experiments to evaluate the performance of our proposed resource allocation optimization framework, and to gain insights into how the various system parameters affect the achievable EE in a user-centric UDN integrating the access downlink via NOMA and the backhaul downlink via beamforming. The performance of the proposed algorithm in our framework is compared with three conventional baseline schemes [35, 14], including the equal-power based allocation strategy, the distance-based association algorithm, and the max-SINR association method. The equal-power based allocation strategy aims at the transmit power equally allocated by every AP to each associated user, and the transmit power equally allocated by the MBS to all the AP. For the distance-based association algorithm, each user associates with the nearest AP for access in a distributed manner. That is, user association at each user is determined by the distance metric between the user and the AP. Additionally, the max-SINR association method is a cellular based approach to achieve the user association based on the SINR level between the user and the AP. With this method, each user attempts to attach to the AP that provides the highest SINR.

Results are obtained with the following default system parameters. For our considered user-centric UDN scenario, the locations of the users are randomly generated with equal possibility in a circular

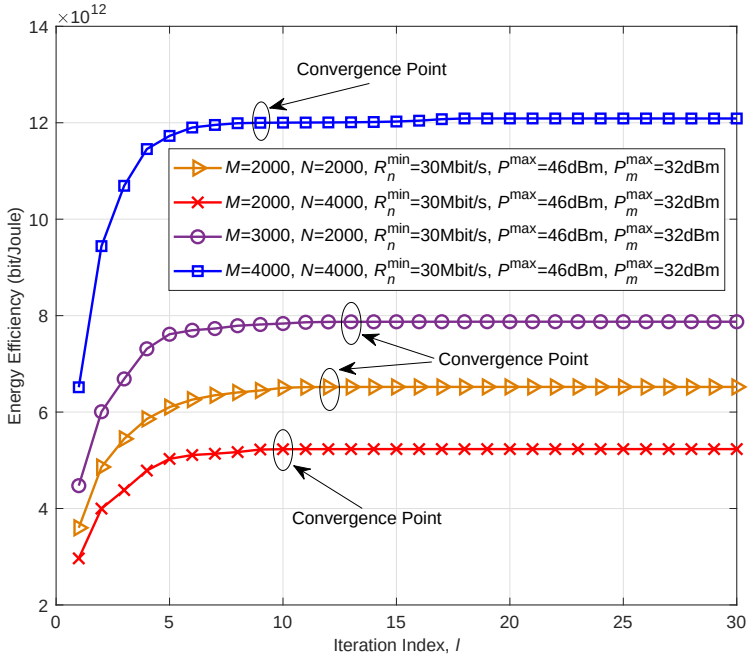

Figure 3: The convergence process of the proposed algorithm in terms of the EE over the number of iterations l under different settings of M and N according to four typical deployment scenarios in Fig. 2.

macrocell area with radius $r = 100$ m centered at the MBS. A large number of APs are also deployed within this area subject to an independent homogeneous PPP Φ_{ρ_1} to provide wireless access service for those users. Especially, the densities of the APs and the users are specified as $\rho_1 = 31.85M$ AP/km^2 and $\rho_2 = 31.85N$ user/km^2, respectively[7]. We set the minimum distance between the APs to be 2.5 m, and the minimum distance between the users is 0.8 m. As shown in Fig. 2, four typical deployment scenarios of the user-centric UDN are considered in the simulations with different combination relations between M and N. Unless otherwise mentioned, we specifically set the power consumption related parameters during the downlink transmission as: $P_{f,n}^{\mathrm{R}} = 5$ mW, $P_{f,n}^{\mathrm{D}} = 10$ mW, $P_m^{\mathrm{C}} = 50$ mW, $P_{f,m}^{\mathrm{R}} = 15$ mW, and $P_{f,m}^{\mathrm{D}} = 30$ mW. For the simplicity, we consider that the separation of all the APs into $F = 8$ disjoint clusters depends on their spatial direction to the MBS (i.e., the 45° direction angle interval), namely, $0° \sim 45°$, $45° \sim 90°$, $90° \sim 135°$, $135° \sim 180°$, $180° \sim 225°$, $270° \sim 315°$, and $315° \sim 360°$. For the sake of generality, each user in every cluster is assumed to be simultaneously associated with at most $M_f = 50$ APs on one or more sub-channel(s), the MBS is also assumed to be simultaneously associated with at most $\phi_{f,k} = 15$ APs on each sub-channel in every cluster.

In our simulations, the total number of sub-channels is $K = 5 \times 10^3$ with $\delta = 500$ for the access downlink and $K - \delta = 4.5 \times 10^3$ for the backhaul downlink to meet the resource management requirement for ultra-densely deployed nodes. The carrier center frequency is set to 2 GHz and the bandwidth of each sub-channel is set to $\varpi = 180$ kHz. For the access downlink via NOMA, we assume that each sub-channel is assigned to at most $N_{f,k} = 10$ users in every cluster to reduce the complexity of the SIC decoding.

[7]By choosing proper values of M and N, the densities of the APs and the users in our simulations are nearly close to the AP density and the user density, respectively, in the theoretical definition of UDNs.

In every generalized APG, each user can be simultaneously served by at most $M_{f,k} = 20$ APs on each sub-channel. The pass loss between the AP and the user in every cluster is obtained by a quasi-static block fading model with the small scale Rayleigh fading channel gain distributed as $g_{f,m,n,k} \sim \mathcal{CN}(0,1)$. For the backhaul downlink via beamforming, the small scale Rayleigh fading channel coefficient vector from the MBS to the AP in every cluster is assumed to satisfy the complex Gaussian model distributed as $\tilde{\mathbf{h}}_{f,m,k} \sim \mathcal{CN}(0, \mathbf{I}_Q)$. The beamforming vector for each AP on every sub-channel in every cluster is generated based on the channel coefficient vector between the MBS and that AP [36]. We assume that the number of the transmit antennas for beamforming in the antenna array of the MBS is equal to the number of APs on each sub-channel in every cluster for simplicity of simulations. Without loss of generality, the path loss exponents with respect to both the wireless access and backhaul downlink are set as the same value, i.e., $\vartheta_1 = \vartheta_2 = 2$. Unless otherwise stated, we set the noise powers at each user and each AP on the corresponding sub-channels to be the same ones with $\sigma_{n,k}^2 |_{k \in \mathcal{A}} = \sigma_{m,k}^2 |_{k \in \mathcal{B}} = \varpi N_0$, where the AWGN power spectral density is initialized by $N_0 = -174$ dBm/Hz.

Before validating the system performance through the above simulation settings, we first provide insight on the convergence behavior of the proposed algorithm. Fig. 3 displays the convergence process of the proposed algorithm in terms of the EE with different numbers of the APs M and the users N after using four typical deployment scenarios generated in Fig. 2. It can be observed that the proposed algorithm increases consistently and converges rapidly in less than 14 iterations to reach the optimal points for different values of M and N. In addition, we can find that the proposed algorithm maintains the best performance with respect to $M = 4000$ and $N = 4000$. That is because the overall EE performance of the system is not superior when $\frac{M}{N}$ is small or especially less than 1. Such behavior can be interpreted that enough number of the APs are required to host the comparable number of the users to guarantee the better performance of wireless access from a user-centric perspective, i.e., $\frac{M}{N} \geq 1$. When $\frac{M}{N}$ becomes smaller, the competition for wirelessly accessing to the limited number of APs prevents it obtaining the better solution. The results indicate that the choice of the numbers of the APs M and the users N has negligible effect on the system performance and the convergence speed of the proposed algorithm.

Fig. 4 presents the comparison of the system EE between the proposed algorithm and the baseline schemes with respect to two different numbers of the users, i.e., $N = 2000$ and $N = 4000$, respectively. It is immediately seen that the EE of the system using no matter the proposed algorithm or the baseline schemes will gradually increase with the growth of the number of the APs. That is, the larger number of the APs, the more obtained EE for the system. To explain, with the increasing number of the APs, more and more users can be hosted by allocating powers and assigning sub-channels properly, resulting in the alleviated resource competition and thereby improving the system performance. Referring to Fig. 4, we can also observe that the system EE with the number of the users $N = 2000$ outperforms that with the number of the users $N = 4000$ as for the same scheme with the increasing of M. The reason is that the comparability of the amount of the APs towards the number of the users usually plays an important role in improving the EE balance. Furthermore, under the constraint of the same number of the users N, our proposed algorithm can bring a beneficial performance gain in the EE compared to three other baseline

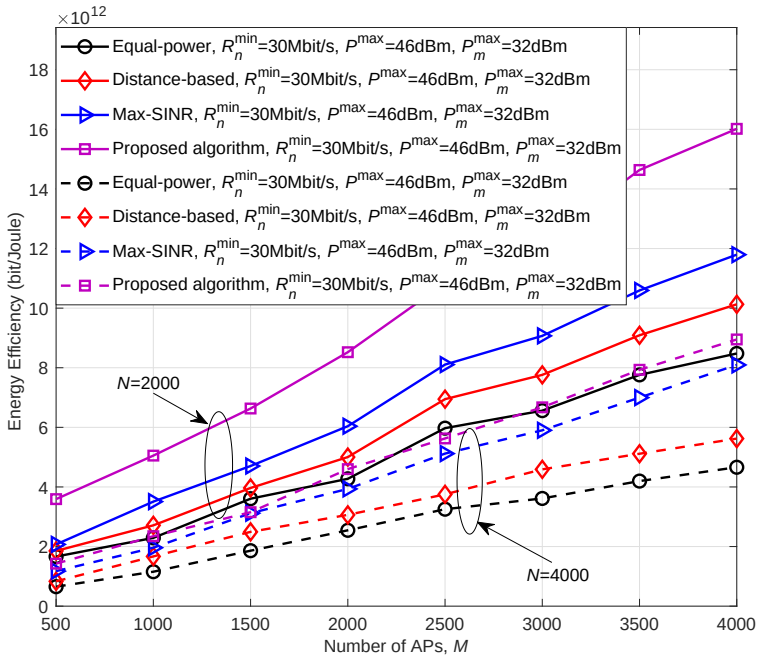

Figure 4: The comparison between the proposed algorithm and the baseline schemes in terms of the EE over the number of the APs with two kinds of setups for the number of the users, i.e., $N = 2000$ and $N = 4000$.

schemes especially for $N = 2000$. The EE gap between the proposed algorithm and the baseline schemes can be explained as follows: (i) The proposed algorithm obtains the better performance by achieving the joint optimization of power allocation, sub-channel assignment, and user association simultaneously. (ii) The baseline schemes only realize an optimization of a single criterion without a joint consideration of power, sub-channel, and user association. This result further provides a hint to choose appropriate joint optimization mechanisms to further improve the system performance.

In Fig. 5, we show the comparison between the proposed algorithm and the baseline schemes in terms of the system EE against the number of the users with respect to two different numbers of the APs, i.e., $M = 2000$ and $M = 4000$, respectively. From Fig. 5, it is evident that the simulated system EE markedly increases with the continuous evolution of the number of the users, i.e., higher user density. The reason for this is that larger densities of the users basically obtain more EE gains in spite of more resource competition and high interference. As a consequence the obtained performance gains in the EE are on an increasing trend gradually for more and more users in the system. Moreover, it can be also seen from this figure that our proposed algorithm greatly outperforms the baseline schemes in terms of the system EE no matter $M = 2000$ or $M = 4000$. This is due to the fact that the proposed algorithm fully takes the joint optimization of power allocation, sub-channel assignment, and user association into account and thereby achieves good performance. As can be seen from the result, the EE performance of the system no matter for the proposed algorithm or for the baseline schemes when $M = 4000$ is always much higher than that of $M = 2000$. This behavior is explained as follows: more APs or the increasing densities of the APs can actually host larger amount of the users under the same system configuration, which can reduce resource competition for the users and further enhance the system performance. This

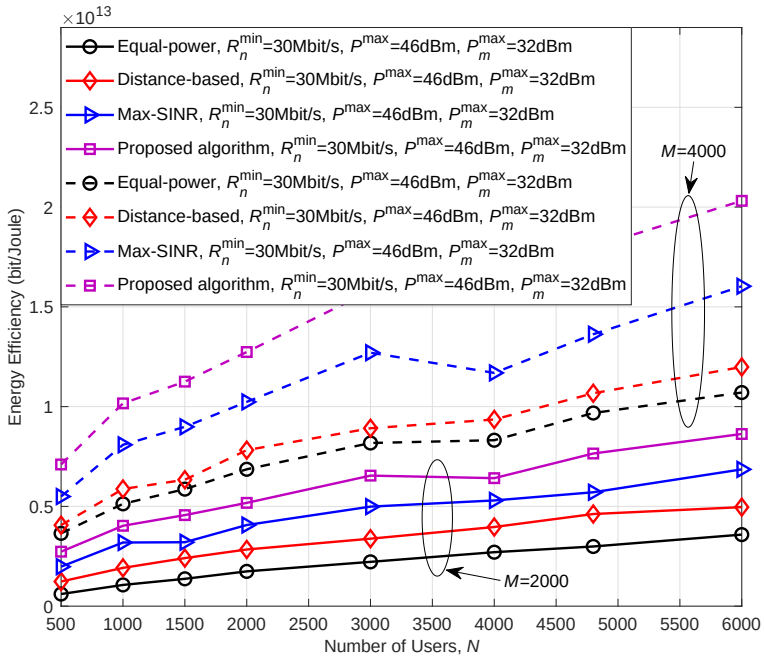

Figure 5: The comparison between the proposed algorithm and the baseline schemes in terms of the EE over the number of the users with two kinds of setups for the number of the APs, i.e., $M = 2000$ and $M = 4000$.

result manifests the importance of the selection of the density of the APs.

Finally, in Fig. 6, we analyze the system EE performance for different values of AP's maximum power $P_m^{\max}$ under the corresponding numbers of the users and the APs, i.e., $N = 2000$ and $M = 2000$. It can be easily seen from this figure that the proposed algorithm greatly outperforms the baseline schemes in terms of the system EE with the increasing maximum power of the AP, indicating that our proposed algorithm achieves a beneficial improvement of system-wide EE over other baseline schemes. Referring to this result, we also find that the increase of AP's maximum power from 1 W to 3.5 W results in the dramatic increase of the obtained system EE, while the simulated curves of the system EE finally all tend to different fixed values with the continuous evolution of AP's maximum power from 3.5 W to 5 W. This can be explained as follows: the transmit power of each user allocated by the AP is more likely to be updated when the AP's maximum power is in a small value below 3.5 W, thus resulting in the lower system EE. However, with the increase of AP's maximum power, the power of each user is properly allocated by the AP, thus satisfying the constraint of AP's maximum power. Furthermore, when AP's maximum power is enough larger, e.g., more than 3.5 W, the possibility of updating the power for each user is also very lower, which results in the nearly fixed values for the overall EE of the system. Such observations above demonstrate the benefit of the proposed algorithm in the maximum EE achievement and provide insightful guidelines for designing the practical user-centric UDNs.

7. Conclusion

In this paper, we proposed a resource allocation framework for energy efficient user association in downlink user-centric UDNs that closely integrate the wireless access via NOMA and wireless back-

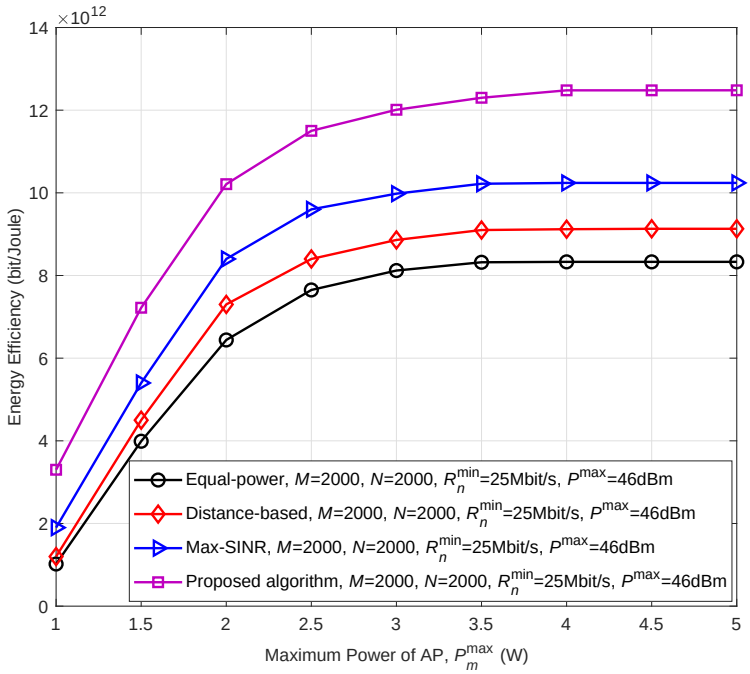

Figure 6: The comparison between the proposed algorithm and the baseline schemes in terms of the EE over the maximum power of the AP under the fixed numbers of the users and the APs, i.e., $N = 2000$ and $M = 2000$.

haul via beamforming. The framework was aimed at the realization of the maximization of the overall system-level EE by jointly optimizing user association index, sub-channel assignment, and transmit power allocation for every user and every AP. The aforementioned design problem was a large-scale non-convex mixed-integer nonlinear programming problem and thus difficult to be solved with affordable computational complexity, especially when the numbers of densely distributed users and APs were larger. Therefore, we conducted the problem reformulation through necessary variable relaxation and sum-of-ratios decoupling, and then converted this highly non-convex problem into the convex subproblem based on the SCA method. On this basis, a distributed iterative algorithm was further developed to achieve the joint optimization of power allocation, sub-channel assignment, and user association simultaneously. Simulation results demonstrate the convergence of this algorithm, and also show that this algorithm achieves good performance with beneficial increase on the system-wide EE compared with other baseline schemes, indicating its potential for a practical design.

Declaration of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61901115, 61971032 and 61802107, the Natural Science Foundation of Hebei Province under Grants F2019402206, F2017402068, F2018402198, and E2017402115, the Natural Science Foundation of Guangdong Province under Grant 2018A030313492, the Research Program for Top-notch Young Talents in

Higher Education Institutions of Hebei Province, China under Grant BJ2017037, and the Fundamental Research Funds for the Central Universities under Grants FRF-TP-18-008A3 and 328201911.

References

[1] K. M. S. Huq, S. A. Busari, J. Rodriguez, V. Frascolla, W. Bazzi, D. C. Sicker, Terahertz-enabled wireless system for beyond-5G ultra-fast networks: A brief survey, IEEE Network 33 (4) (2019) 89–95. doi:10.1109/mnet.2019.1800430.

[2] X. Ge, S. Tu, G. Mao, C.-X. Wang, T. Han, 5G ultra-dense cellular networks, IEEE Wireless Communications 23 (1) (2016) 72–79. doi:10.1109/mwc.2016.7422408.

[3] G. Chopra, R. K. Jha, S. Jain, A survey on ultra-dense network and emerging technologies: Security challenges and possible solutions, Journal of Network and Computer Applications 95 (2017) 54–78. doi:10.1016/j.jnca.2017.07.007.

[4] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song, Y. Zhang, Resource allocation for ultra-dense networks: A survey, some research issues and challenges, IEEE Communications Surveys & Tutorials 21 (3) (Third Quarter 2019) 2134–2168. doi:10.1109/comst.2018.2867268.

[5] M. Hawasli, S. A. Çolak, Toward green 5G heterogeneous small-cell networks: Power optimization using load balancing technique, AEU - International Journal of Electronics and Communications 82 (2017) 474–485. doi:10.1016/j.aeue.2017.09.012.

[6] S. Chen, F. Qin, B. Hu, X. Li, Z. Chen, User-centric ultra-dense networks for 5G: Challenges, methodologies, and directions, IEEE Wireless Communications 23 (2) (2016) 78–85. doi:10.1109/mwc.2016.7462488.

[7] Y. Liu, X. Li, F. R. Yu, H. Ji, H. Zhang, V. C. M. Leung, Grouping and cooperating among access points in user-centric ultra-dense networks with non-orthogonal multiple access, IEEE Journal on Selected Areas in Communications 35 (10) (2017) 2295–2311. doi:10.1109/jsac.2017.2724680.

[8] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C.-L. I, H. V. Poor, Application of non-orthogonal multiple access in LTE and 5G networks, IEEE Communications Magazine 55 (2) (2017) 185–191. doi:10.1109/mcom.2017.1500657cm.

[9] U. Siddique, H. Tabassum, E. Hossain, Downlink spectrum allocation for in-band and out-band wireless backhauling of full-duplex small cells, IEEE Transactions on Communications 65 (8) (2017) 3538–3554. doi:10.1109/tcomm.2017.2699183.

[10] M. S. Ali, E. Hossain, D. I. Kim, Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation, IEEE Access 5 (2017) 565–577. doi:10.1109/access.2016.2646183.

[11] G. Zhang, F. Ke, Y. Peng, C. Zhang, H. Zhang, User access and resource allocation in full-duplex user-centric ultra-dense heterogeneous networks, in: Proceedings of IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 2018. doi:10.1109/glocom.2018.8648065.

[12] J. Choi, On generalized downlink beamforming with NOMA, Journal of Communications and Networks 19 (4) (2017) 319–328. doi:10.1109/jcn.2017.000056.

[13] S. Moon, H. Kim, Y. Yi, BRUTE: Energy-efficient user association in cellular networks from population game perspective, IEEE Transactions on Wireless Communications 15 (1) (2016) 663–675. doi:10.1109/twc.2015.2477297.

[14] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, H. V. Poor, Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations, IEEE Journal on Selected Areas in Communications 35 (9) (2017) 1936–1947. doi:10.1109/jsac.2017.2720898.

[15] Y. Lin, R. Zhang, L. Yang, L. Hanzo, Modularity-based user-centric clustering and resource allocation for ultra dense networks, IEEE Transactions on Vehicular Technology 67 (12) (2018) 12457–12461. doi:10.1109/tvt.2018.2875547.

[16] G. Zhang, H. Zhang, Z. Han, G. K. Karagiannidis, Spectrum allocation and power control in full-duplex ultra-dense heterogeneous networks, IEEE Transactions on Communications 67 (6) (2019) 4365–4380. doi:10.1109/tcomm.2019.2897765.

[17] J. Cao, T. Peng, Z. Qi, R. Duan, Y. Yuan, W. Wang, Interference management in ultradense networks: A user-centric coalition formation game approach, IEEE Transactions on Vehicular Technology 67 (6) (2018) 5188–5202. doi:10.1109/tvt.2018.2799568.

[18] J. Park, S. Y. Jung, S.-L. Kim, M. Bennis, M. Debbah, User-centric mobility management in ultra-dense cellular networks under spatio-temporal dynamics, in: Proceedings of IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 2016. doi:10.1109/glocom.2016.7842367.

[19] Y. Lin, R. Zhang, L. Yang, L. Hanzo, Secure user-centric clustering for energy efficient ultra-dense networks: Design and optimization, IEEE Journal on Selected Areas in Communications 36 (7) (2018) 1609–1621. doi:10.1109/jsac.2018.2825178.

[20] Q. Zhang, K. Luo, W. Wang, T. Jiang, Joint C-OMA and C-NOMA wireless backhaul scheduling in heterogeneous ultra dense network, IEEE Transactions on Wireless Communications-doi:10.1109/twc.2019.2949791. Early Access.

[21] Z. Qin, X. Yue, Y. Liu, Z. Ding, A. Nallanathan, User association and resource allocation in unified NOMA enabled heterogeneous ultra dense networks, IEEE Communications Magazine 56 (6) (2018) 86–92. doi:10.1109/mcom.2018.1700497.

[22] Q. Wang, F. Zhou, Fair resource allocation in an MEC-enabled ultra-dense IoT network with NOMA, in: Proceedings of IEEE International Conference on Communications Workshops (ICC Workshops), Shanghai, China, 2019. doi:10.1109/iccw.2019.8757173.

[23] G. Kwon, H. Park, Joint user association and beamforming design for millimeter wave UDN with wireless backhaul, IEEE Journal on Selected Areas in Communications-doi:10.1109/jsac.2019.2947926. Early Access.

[24] Y. Teng, W. Sun, A. Liu, R. Yang, V. K. N. Lau, Mobility-aware transmit beamforming for ultra-dense networks with sparse feedback, IEEE Transactions on Vehicular Technology 68 (2) (2019) 1968–1972. doi:10.1109/tvt.2018.2886800.

[25] H. T. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, W.-J. Hwang, Collaborative multicast beamforming for content delivery by cache-enabled ultra dense networks, IEEE Transactions on Communications 67 (5) (2019) 3396–3406. doi:10.1109/tcomm.2019.2894797.

[26] B. Di, L. Song, Y. Li, Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks, IEEE Transactions on Wireless Communications 15 (11) (2016) 7686–7698. doi:10.1109/twc.2016.2606100.

[27] J. Zhu, J. Wang, Y. Huang, S. He, X. You, L. Yang, On optimal power allocation for downlink non-orthogonal multiple access systems, IEEE Journal on Selected Areas in Communications 35 (12) (2017) 2744–2757. doi:10.1109/jsac.2017.2725618.

[28] X. Sun, N. Yang, S. Yan, Z. Ding, D. W. K. Ng, C. Shen, Z. Zhong, Joint beamforming and power allocation in downlink NOMA multiuser MIMO networks, IEEE Transactions on Wireless Communications 17 (8) (2018) 5367–5381. doi:10.1109/twc.2018.2842725.

[29] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wen, M. Tao, Resource allocation in spectrum-sharing OFDMA femtocells with heterogeneous services, IEEE Transactions on Communications 62 (7) (2014) 2366–2377. doi:10.1109/tcomm.2014.2328574.

[30] S. Schaible, J. Shi, Fractional programming: The sum-of-ratios case, Optimization Methods and Software 18 (2) (2003) 219–229. doi:10.1080/1055678031000105242.

[31] B. R. Marks, G. P. Wright, A general inner approximation algorithm for nonconvex mathematical programs, Operations Research 26 (4) (1978) 681–683. doi:10.1287/opre.26.4.681.

[32] J. Papandriopoulos, J. S. Evans, SCALE: A low-complexity distributed protocol for spectrum balancing in multiuser DSL networks, IEEE Transactions on Information Theory 55 (8) (2009) 3711–3724. doi:10.1109/tit.2009.2023751.

[33] Q. Chen, G. Yu, R. Yin, G. Y. Li, Energy-efficient user association and resource allocation for multistream carrier aggregation, IEEE Transactions on Vehicular Technology 65 (8) (2016) 6366–6376. doi:10.1109/tvt.2015.2472558.

[34] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, UK, 2004. doi:10.1017/CBO9780511804441.

[35] G. Dong, H. Zhang, S. Jin, D. Yuan, Energy-efficiency-oriented joint user association and power allocation in distributed massive MIMO systems, IEEE Transactions on Vehicular Technology 68 (6) (2019) 5794–5808. doi:10.1109/tvt.2019.2912388.

[36] B. Kimy, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, D. Hong, Non-orthogonal multiple access in a downlink multiuser beamforming system, in: Proceedings of IEEE Military Communications Conference MILCOM, San Diego, CA, USA, 2013, pp. 1278–1283. doi:10.1109/milcom.2013.218.