Energy-Efficient Resource Allocation and Subchannel Assignment for NOMA-Enabled Multiaccess Edge Computing

Lina Liu, *Student Member, IEEE*, Bo Sun, *Member, IEEE*, Xiaoqi Tan, *Member, IEEE*, and Danny H. K. Tsang, *Fellow, IEEE*

***Abstract*—In this article, we study an energy-efficient nonorthogonal multiple access (NOMA) enabled multiaccess edge computing (MEC) system with strict latency requirements. We aim to minimize the energy consumption of all users by optimizing the resource allocation (including power and computation resources) and subchannel assignment, subject to the given latency constraint. The formulated problem, however, is a nonconvex combinatorial optimization problem. Nevertheless, we decompose the problem into a resource allocation subproblem and a subchannel assignment subproblem, and then solve the two subproblems iteratively. On one hand, we investigate the hidden convexity of the resource allocation subproblem under the optimal conditions, and propose an efficient algorithm to optimally allocate the resources by dual decomposition methods. On the other hand, we formulate the subchannel assignment subproblem into an integer linear programming problem and strictly prove that the problem is nondeterministic polynomial-time hard. We then solve it optimally by branch-and-bound methods, which is shown to be efficient in extensive simulations. Moreover, through considerable simulation results, we show that our proposed algorithm helps greatly reduce users' energy consumption when communication resources (e.g., bandwidth) are limited. Additionally, it is verified that NOMA outperforms orthogonal multiple access in multiuser latency-sensitive MEC systems.**

***Index Terms*—Energy minimization, multiaccess edge computing (MEC), nonorthogonal multiple access (NOMA), resource allocation, subchannel assignment.**

I. Introduction

The Internet of Things (IoT) has had explosive growth in recent years. Emerging IoT applications, such as autonomous driving, virtual reality/augmented reality and so forth, are latency-sensitive and computation-intensive [1], [2]. However, IoT devices are typically characterized by restricted computation resource. Hence, it is challenging to implement these latency-sensitive applications on such devices, and a significant surge in demand for computation resource is induced [3]. Recently, multiaccess edge computing (MEC) has been considered as a promising solution to tackle these issues [4]. MEC extends the cloud-computation capabilities and IT service environment to the edge of the network. Compared to faraway centralized data centers, MEC servers can be deployed at various access points [e.g., mobile base stations (BSs)] close to IoT devices. By offloading the computation tasks to MEC servers, devices can finish the data delivery and computation with a short latency. MEC systems have been investigated with different optimization objectives and in multifarious situations [5]–[7]. In [5], the joint computation offloading and resource allocation strategy was studied to minimize users' energy consumption. In [6], latency-minimization problems were formulated and studied in a scenario where users may opt for partial offloading to the edge server via an access point with the aid of an intelligent reflecting surface. In [7], a dynamic spectrum management framework was proposed to improve spectrum resource utilization in MEC systems in autonomous vehicular networks.

Meanwhile, the idea of the IoT is now connecting a massive collection of smart objects to the Internet [8]. To this end, nonorthogonal multiple access (NOMA) has been recognized as a promising technology to improve the network capacity and spectral efficiency. Different from conventional orthogonal multiple access (OMA) schemes, where different users are served with orthogonal resources in either the time, frequency, or code domain, NOMA allows controllable interferences by allocating nonorthogonal resources, such as different power levels or low-density spreading codes [9]. NOMA is accordingly classified into power-domain NOMA and code-domain NOMA. For power-domain NOMA, which is the technology mainly discussed in this article, successive interference cancellation (SIC) is often adopted at the receiver to eliminate the multiuser interference and decode the superimposed signals. Specifically, the successfully decoded signal is removed from the superimposed signal before processing the subsequent users' decoding. By allowing multiple users to share time and frequency resources, NOMA helps improve the spectral efficiency and network capacity, and enhance the system throughput [10]. NOMA has attracted huge research interests in both downlink and uplink scenarios. In [11] and [12], UAV-aided downlink NOMA scheme was proposed to guarantee the secure transmissions. In [13], a joint beamforming and power allocation design for a NOMA-based satellite terrestrial integrated network was proposed to maximize the network's sum rate. In [14], a UAV-enabled uplink

Manuscript received August 24, 2020; revised January 21, 2021; accepted March 5, 2021. Date of publication April 2, 2021; date of current version March 24, 2022. *(Corresponding author: Lina Liu.)*

Lina Liu, Bo Sun, and Danny H. K. Tsang are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: lliuaw@ust.hk; bsunaa@ust.hk; eetsang@ust.hk).

Xiaoqi Tan is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 2E4, Canada (e-mail: xiaoqi.tan@utoronto.ca).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSYST.2021.3064919, provided by the authors.

Digital Object Identifier 10.1109/JSYST.2021.3064919

NOMA network was investigated with the aim of maximizing the users' sum rate.

Motivated by the encouraging abilities of the MEC framework and NOMA technology, NOMA-enabled MEC systems have been increasingly investigated with different optimization objectives, e.g., latency minimization [15], [16], energy minimization [17]–[20], joint latency and energy minimization [21]–[24], and some other objectives, including maximizing the energy efficiency defined as the ratio of the system's sum rate over devices' sum power consumption [25], maximizing the successful computation probability defined as the probability that the targeted tasks can be successfully executed within a given delay budget [26] and maximizing the secure computation efficiency defined as the secure computing rate [27]. Since we study the energy minimization problem in this article, we list related literature in detail below. Both partial and binary offloading with NOMA transmission were studied in [17]. However, users were clustered into one NOMA group, which introduced too much transmission interference and greatly increased the SIC complexity. To overcome this problem, Pham *et al.* [21] studied a system with multiple NOMA groups, where different NOMA groups offloaded in orthogonal subchannels. However, Pham *et al.* [21] only considered the communication resources. The joint communication resource and users' computation resource allocation problem was studied in [18]. Nevertheless, the computation resource allocation at the BS, which might help further reduce users' energy consumption, was not investigated. In [22], the computation resource allocation at the BS was taken into account and optimized, along with users' power allocation, but the user clustering was not studied. In [19], the computation resource allocation at the BS, communication resource allocation, power allocation, and user clustering were optimized based on a greedy algorithm. In [20], a similar system setting to that in [19] was studied, and a decomposition-driven algorithm was proposed to solve the formulated problem heuristically. While the work in [18]–[22] all adopted frequency-division multiple access (FDMA) scheme for the transmission of different NOMA groups, time-division multiple access based transmission was applied in [23] and [24]. In particular, Yang *et al.* [23] assumed that NOMA technology was applied throughout users' transmission, while Zhu *et al.* [24] proposed a hybrid NOMA transmission scheme, which incorporated NOMA and OMA during the offloading.

In this article, we devise an MEC system with multiple NOMA groups, where different groups of users are assigned to different subchannels. We intend to study the energy minimization problem subject to users' latency requirements. In order to solve the formulated nonconvex problem, we decompose the original problem into two subproblems. Our article is most related to prior works Kiani and Ansari [19] and Zeng and Fodor [20]. However, in those works, the optimality of either one or two of the decomposed subproblems was not guaranteed by their proposed heuristic algorithms. In contrast, we solve both subproblems optimally and then iteratively deal with the two subproblems in an efficient way. In particular, we make the following detailed contributions.

1) We formulate an energy minimization problem by optimizing the power and computation resource allocation and subchannel assignment. We decompose the original nonconvex problem and address the resource allocation subproblem and subchannel assignment subproblem iteratively.
2) Given a fixed subchannel assignment, we formulate the resource allocation subproblem and identify the hidden convexity by reformulation. We apply dual decomposition methods to optimally determine the computation resource allocation at the BS and the user's power allocation.
3) As an extension of our previous work [28], we make some significant improvements for the subchannel assignment update. With the obtained computation resource allocation, we reformulate the subchannel assignment subproblem into an integer linear programming (ILP) problem and prove its nondeterministic polynomial-time (NP) hardness. We then update the subchannel assignment by branch-and-bound (BNB) methods optimally, which is shown to be efficient in extensive simulations.
4) By considerable numerical tests, we show that our proposed algorithm greatly reduces users' energy consumption for computation offloading with limited communication resources, and can be applied to enable computation offloading of more users. Furthermore, we substantiate that NOMA technology helps reduce users' energy consumption in latency-sensitive MEC systems.

The remainder of this article is organized as follows. The system model and problem formulation for a special case are described in Section II. We propose the decomposition-driven algorithm in Section III. In Section IV, we extend the special case to general cases and propose the solutions accordingly. Simulation results and performance analysis are presented in Section V. Finally, Section VI concludes this article.

At this stage, it is worth reviewing the notation we adopt throughout this article. We use uppercase calligraphic symbols to denote sets, e.g., $\mathcal{K}$ denotes the set of users, except that we denote by $\mathcal{N}(0, \sigma^2)$ the normal distribution. We use lowercase bold symbols to denote matrices or vectors, e.g., $\boldsymbol{x} := \{x_{k,m} \mid k \in \mathcal{K}, m \in \mathcal{M}\}$, which is a $K \times M$ matrix, denotes the subchannel assignment decision variables.

II. Energy Minimization Problem in the NOMA-Enabled MEC System: A Special Case

A. System Model

In the proposed NOMA-enabled MEC system, there are one BS and K users, who all have a single antenna for transmitting and receiving. The BS is mounted with an edge server, whose computation resource is shared by the K users to execute their offloaded computation workloads. Users follow the BS's coordination to offload their computation-intensive tasks over a frequency band with a total bandwidth B, which is divided into N orthogonal subchannels with equal bandwidth. One user can only transmit in one subchannel, whereas one subchannel can hold up to two users who offload concurrently by adopting power-domain NOMA technology. The number of users who can share the same subchannel to transmit data is constrained to be two to control the SIC complexity and error propagation [29]. We study a time-slotted system with a given slot duration τ. Each user has a computation task and requires it to be offloaded, executed, and downloaded within τ. Assume a quasi-static Rayleigh fading channel situation [21], [26], [30], [31], so that users have constant channel gains over one slot and time-varying channel gains from slot to slot.

In the following, we start from the special case $K = 2N$. In Section IV, we show that we can easily generalize the system to cases $N \leq K \leq 2N$ and the algorithm proposed for the special case in Section III is still applicable for general cases. Let $\mathcal{K} := \{1, \ldots, 2N\}$ denote the set of users. Considering that the two users in the same subchannel have distinct channel gains, we suppose each subchannel has two corresponding distinct positions. Therefore, N subchannels have $M = 2N$ positions in total and each subchannel position holds one user ($K = M$). Note that we introduce the concept of subchannel position in order to better differentiate the two users on the same subchannel, who play different roles during NOMA transmission. In particular, the subchannel position has the physical meaning of distinguishing the first and second decoded users in the NOMA scheme. Denote by $\mathcal{N} := \{1, \ldots, N\}$ and $\mathcal{M} := \{1, \ldots, 2N\}$ the set of all subchannels and subchannel positions. Furthermore, we differentiate the subchannel positions by letting $m = 2n - 1$ and $m = 2n$ represent the subchannel positions with large and small channel gains of subchannel n, respectively. The subsets of subchannel positions with large and small channel gains can then be denoted by $\mathcal{M}^l := \{m \mid m = 2n - 1, n = 1, \ldots, N\}$ and $\mathcal{M}^s := \{m \mid m = 2n, n = 1, \ldots, N\}$ accordingly.

It is worth noting that although we establish a time-slotted system, our system model can be accommodated to circumstances where users have diverse latency requirements. Denote by δ_k user k's latency requirement. Users with $\delta_k \geq \tau$ will be admitted to the system. Since the most energy-efficient mechanism for each user is to transmit with maximum allowable time (according to Lemma 1, which we derive in Section III-A), each user can offload continuously throughout $\lfloor \frac{\delta_k}{\tau} \rfloor$ time slots to save energy consumption. Therefore, user k can separate its workloads into $\lfloor \frac{\delta_k}{\tau} \rfloor$ parts and offload one part during each time slot. Each part will be offloaded, executed, and returned within each time slot, and the whole task will be completed after $\lfloor \frac{\delta_k}{\tau} \rfloor$ time slots. Considering that the procedures are implemented in a continuous manner, the BS can serve users with various latency requirements simultaneously and we only need to focus on one time slot to model the system.

B. Problem Formulation

Denote by $\boldsymbol{x} := \{x_{k,m} \mid k \in \mathcal{K}, m \in \mathcal{M}\}$ the subchannel assignment decision variables. Specifically, $x_{k,m} = 1$ represents that user k is assigned to the subchannel position m, and $x_{k,m} = 0$ otherwise. It is demanded that each user is assigned to one subchannel position for data offloading and each subchannel position holds one user. In addition, the user in subchannel position $m \in \mathcal{M}^l$ should have a no smaller channel gain than the user who transmits over the same subchannel. Thus, the feasible set of $x_{k,m}$ is represented as

$$\left.\begin{aligned} \mathcal{X} = \{\boldsymbol{x} \mid & x_{k,m} \in \{0,1\} \quad \forall k \in \mathcal{K} \quad \forall m \in \mathcal{M} \\ & \sum_{k=1}^{K} x_{k,m} = 1 \quad \forall m \in \mathcal{M}; \sum_{m=1}^{M} x_{k,m} = 1 \quad \forall k \in \mathcal{K} \\ & \sum_{k \in \mathcal{K}} x_{k,m} |h_{k,m}|^2 \geq \sum_{k \in \mathcal{K}} x_{k,m+1} |h_{k,m+1}|^2 \quad \forall m \in \mathcal{M}^l \end{aligned}\right\}$$

where $h_{k,m}$ and $h_{k,m+1}$ ($m \in \mathcal{M}^l$) are, respectively, large and small uplink channel gains. From the view of the BS, the received signal in subchannel n is a superimposed signal from the two users who transmit over subchannel positions $2n - 1$ and $2n$. Denote by y_n the received signal over subchannel n. It can be represented as $y_n = \sum_k^K x_{k,2n-1} \sqrt{p_k} h_{k,2n-1} s_{k,2n-1} + \sum_k^K x_{k,2n} \sqrt{p_k} h_{k,2n} s_{k,2n} + z_n$, where $s_{k,2n-1}$ and $s_{k,2n}$ are the modulated symbols, p_k is the transmission power, and $z_n \sim \mathcal{N}(0, \sigma^2)$ is the additive white Gaussian noise. Similar to the work in [18]–[23], we suppose the BS has perfect knowledge of the channel state information (CSI) and, hence, can apply SIC based on the users' channel gains to decode the signals superimposed on each subchannel. The signal of the user on subchannel position $2n - 1$ with a large channel gain will be decoded first by treating that of the user on subchannel position $2n$ as interference [32]. After subtracting the decoded signal of the user with a large channel gain, the remaining user on the same subchannel does not suffer from any interference. The data rate of the two users can be expressed as

$$\begin{aligned} r_k = & \frac{B}{N} \sum_{m \in \mathcal{M}^l} x_{k,m} \log_2 \left(1 + \frac{p_k |h_{k,m}|^2}{\sigma^2 + i_m} \right) \\ & + \frac{B}{N} \sum_{m \in \mathcal{M}^s} x_{k,m} \log_2 \left(1 + \frac{p_k |h_{k,m}|^2}{\sigma^2} \right) \end{aligned} \tag{1}$$

where $i_m = \sum_{j \in \mathcal{K}} x_{j,m+1} p_j |h_{j,m+1}|^2$. We denote the data size of user k to be offloaded by d_k (bit), and then derive the offloading time of user k as

$$t_k^{\mathrm{o}} = d_k / r_k. \tag{2}$$

Energy consumed by user k to perform the offloading is

$$e_k = \sum_{m=1}^{M} x_{k,m} p_k t_k^{\mathrm{o}}. \tag{3}$$

Let C_k (cycle/b) denote the CPU resource required per task bit of user k. User k's execution time can be represented as

$$t_k^{\mathrm{e}} = d_k C_k / f_k \tag{4}$$

where f_k (cycle/s) is the computation resource allocated to user k at the BS. The feasible set of f_k can be described as $\mathcal{F} := \{\boldsymbol{f} \mid f_k \geq 0 \quad \forall k \in \mathcal{K}; \sum_{k=1}^{K} f_k \leq F\}$, where $\boldsymbol{f} := \{f_k \mid k \in \mathcal{K}\}$ and F is the total computation resource available at the BS. Following the work in [18]–[23], the latency for users to download the processed data from the BS is assumed negligible because the processed data has a much smaller size compared to the offloaded raw data and the BS has more power to transmit with a higher data rate. Thus, the total latency consists of two main parts, namely, data offloading latency and computation execution latency, which can be written as

$$t_k = t_k^{\mathrm{o}} + t_k^{\mathrm{e}}. \tag{5}$$

We aim at minimizing the total energy consumption of all users, which is spent on task offloading. Therefore, we formulate a weighted sum energy minimization problem with weighting factors $\boldsymbol{w} := [w_1, ..., w_K]^T \in \mathbb{R}_+^K$ as follows (here "WSEM" refers to "weighted sum energy minimization"):

$$\text{(WSEM)} \quad \min_{\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{f}} \quad E = \sum_{k=1}^{K} w_k e_k$$

$$\text{subject to:} \quad t_k^{\mathrm{o}} + t_k^{\mathrm{e}} \leq \tau \quad \forall k \in \mathcal{K} \tag{6a}$$

$$\sum_{m=1}^{M} x_{k,m} p_k \leq P_k \quad \forall k \in \mathcal{K} \tag{6b}$$

$$\boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{p} \in \mathcal{P}, \quad \boldsymbol{f} \in \mathcal{F}$$
$$\text{constraints (1)–(5)} \tag{6c}$$

where $\boldsymbol{p} := \{p_k \mid k \in \mathcal{K}\}, \mathcal{P} := \{\boldsymbol{p} \mid p_k \geq 0 \quad \forall k \in \mathcal{K}\}$ and P_k is the maximum transmission power of user k. Constraint (6a) guarantees that the offloaded data can be processed within the time slot. Constraint (6b) limits the transmission power to the power budget. It can be observed from the simulation results that, after the optimization, the decoding order is in accordance with the order of the received power channel gain, i.e., the resultant solution of the formulated problem accords with $p_k|h_{k,n}|^2 \geq p_j|h_{j,n}|^2$, which promotes an energy-efficient NOMA transmission since user j can cause less interference compared to user k during NOMA transmission.

Problem (WSEM) intends to minimize users' energy consumption subject to a strict latency constraint by jointly optimizing the power and computation resource allocations, as well as the subchannel assignment. Due to the combinatorial property, problem (WSEM) is a mixed-integer nonconvex problem, for which it is difficult to find a computationally efficient solution approach [33]. In the sequel, we design an efficient algorithm to address problem (WSEM) by iteratively settling the resource allocation and subchannel assignment.

III. Decomposition-Driven Algorithm Design

In this section, we propose the decomposition-driven algorithm to solve (WSEM). Specifically, we first propose to solve power and computation resource allocation subproblem (Sub-RA) by dual decomposition based on the hidden convexity (here "RA" refers to "resource allocation"). After that, subchannel assignment subproblem (Sub-SA) is dealt with by BNB methods (here "SA" refers to "subchannel assignment"). Finally, this section ends with summarization of the overall algorithm.

A. Power and Computation Resource Allocation Subproblem (Sub-RA)

We first formulate the power and computation resource allocation subproblem with a fixed subchannel assignment. Under this condition, there is a deterministic one-to-one mapping between users and subchannel positions. We can hence exploit the correspondence to represent either the subchannel position or user by index m. Note that here user m in fact represents user k who transmits over subchannel position m according to the deterministic mapping between user k and its allocated subchannel position m. We use the same index for simplicity. Particularly, user $m = 2n - 1$ offloads with a large channel gain over subchannel n and user $m = 2n$ offloads with a small channel gain over subchannel n. According to (1), the data rates of the two users in subchannel n are

$$r_m = \begin{cases} \frac{B}{N}\log_2\left(1 + \frac{p_{2n-1}|h_{2n-1}|^2}{\sigma^2 + p_{2n}|h_{2n}|^2}\right), & m = 2n-1 \\ \frac{B}{N}\log_2\left(1 + \frac{p_{2n}|h_{2n}|^2}{\sigma^2}\right), & m = 2n \end{cases} \tag{7}$$

where $p_{2n-1}(p_{2n})$ and $h_{2n-1}(h_{2n})$ are the transmission power and channel gain of user $2n-1(2n)$. User k's offloading time, execution time, and energy consumption can be derived as

$$t_m^{\mathrm{o}} = d_m/r_m, \quad t_m^{\mathrm{e}} = d_m C_m/f_m, \quad e_m = p_m t_m^{\mathrm{o}} \tag{8}$$

where $m \in \{2n-1, 2n\}$. Note that the different data rate expressions for the two users $2n-1$ and $2n$ result in the different energy consumption expressions in (8). Therefore, for a better structural interpretation, we regard the two users transmitting over subchannel n as a group and represent their total energy consumption as

$$E_n(p_{2n-1}, p_{2n}) = w_{2n-1}p_{2n-1}\frac{d_{2n-1}}{r_{2n-1}} + w_{2n}p_{2n}\frac{d_{2n}}{r_{2n}} \tag{9}$$

where r_{2n-1} and r_{2n} are functions of p_{2n-1} and p_{2n}, as shown in (7). We exploit (9) to transform the problem of minimizing the total energy consumption of all users into the equivalent problem of minimizing the total energy consumption of all groups in all subchannels. The consequent resource allocation problem can be written as

$$\text{(Sub-RA)} \quad \min_{p_m, f_m} \quad E = \sum_{n=1}^{N} E_n(p_{2n-1}, p_{2n})$$
$$\text{subject to:} \quad t_m^{\mathrm{o}} + t_m^{\mathrm{e}} \leq \tau \quad \forall m \in \mathcal{M} \tag{10a}$$
$$0 < p_m \leq P_m \quad \forall m \in \mathcal{M} \tag{10b}$$
$$f_m > 0 \quad \forall m \in \mathcal{M} \tag{10c}$$
$$\sum_{m=1}^{M} f_m \leq F$$
$$\text{constraints (7)–(9).} \tag{10d}$$

We now derive the following lemma.

Lemma 1: Constraints (10a) are binding for all users with the optimal power and computation resource allocations.

Proof: Please refer to Appendix A. ■

Under the binding conditions, the power allocation decisions p_{2n-1} and p_{2n}, as well as the data rates r_{2n-1} and r_{2n}, can be uniquely determined by the computation resource allocation decisions. To be specific, we rewrite the power allocation as functions of the transmission rates according to (7)

$$p_m = \begin{cases} \frac{\sigma^2}{|h_{2n-1}|^2}\exp(ar_{2n})[\exp(ar_{2n-1}) - 1], & m = 2n-1 \\ \frac{\sigma^2}{|h_{2n}|^2}[\exp(ar_{2n}) - 1], & m = 2n \end{cases} \tag{11}$$

where $a := \frac{N\ln 2}{B}$. The relationship between the transmission rates and the computation resource allocation decisions can be derived from $t_m^{\mathrm{o}} + t_m^{\mathrm{e}} = \tau$ as

$$r_m = d_m/[\tau - d_m C_m/f_m], m \in \{2n-1, 2n\}. \tag{12}$$

By slightly abusing the notation, we redefine the energy consumption of the two users in subchannel n as $E_n(p_{2n-1}, p_{2n}) := E_n(f_{2n-1}, f_{2n})$ and reformulate problem (Sub-RA) as (here "E-RA" refers to "equivalent resource allocation" subproblem)

$$\text{(Sub-E-RA)} \min_{f_{2n-1}, f_{2n}} \quad E = \sum_{n=1}^{N} E_n(f_{2n-1}, f_{2n})$$
$$\text{subject to:} \quad \sum_{n=1}^{N}(f_{2n-1} + f_{2n}) \leq F \tag{13a}$$
$$(f_{2n-1}, f_{2n}) \in \mathcal{F}_n$$
$$\text{constraints (9), (11), and (12)} \tag{13b}$$

where $\mathcal{F}_n := \{(f_{2n-1},\ f_{2n}) \mid f_{2n-1},\ f_{2n} > 0;\ r_{2n-1}(f_{2n-1}),$ $r_{2n}(f_{2n}) > 0; \quad 0 < p_{2n-1}(f_{2n-1}, \quad f_{2n}) \leq P_{2n-1}, \quad 0 <$

$p_{2n}(f_{2n}) \leq P_{2n}\}$ defines the feasible set of $(f_{2n-1},\ f_{2n})$. Note that the computation resource allocation is the only optimization variable in problem (Sub-E-RA), and this facilitates the development of the following theorem.

Theorem 1: Problem (Sub-E-RA) is a convex problem.

Proof: Please refer to Appendix B. ■

To handle the coupling constraint (13a), we apply dual decomposition methods to solve problem (Sub-E-RA) [34]. This approach is in fact solving the dual problem instead of the original primal one. Because of the convexity of problem (Sub-E-RA), strong duality holds between the primal and dual problems. Finally, given a subchannel assignment, we can obtain the optimal computation resource allocation. The relevant power allocation decisions can also be deduced.

B. *Subchannel Assignment Subproblem (Sub-SA)*

Given the optimal computation resource allocation f_m^* obtained from problem (Sub-E-RA), we formulate the subchannel assignment subproblem. Recall that each subchannel position m corresponds to a user k in problem (Sub-E-RA). Therefore, we slightly abuse the notation and denote the computation resource allocated to user k by f_k^*. In the following subchannel assignment problem, we fix f_k^*. Afterward, user k's offloading time and data rate can be derived as constants as

$$t_k^{\mathrm{o}} = \tau - d_k C_k / f_k^*, \quad r_k = d_k / [\tau - d_k C_k / f_k^*]. \tag{14}$$

User k's offloading time and data rate are determined regardless of which subchannel position it is assigned to. However, user k's power allocation is based on the subchannel assignment according to (11). In the following, we use $p_{k,m}$ to further specify user k's transmission power in subchannel position m. In particular

$$p_{k,m} = \begin{cases} \frac{\sigma^2}{|h_{k,m}|^2} \exp(a \sum\limits_{j \in \mathcal{K}} x_{j,m+1} r_j)[\exp(a r_k) - 1] \\ \qquad\qquad\qquad\qquad m = 2n - 1 \\ \frac{\sigma^2}{|h_{k,m}|^2} [\exp(a r_k) - 1], \qquad m = 2n. \end{cases} \tag{15}$$

Therefore, we need to decide the subchannel assignment and the corresponding power allocation together, given the computation resource allocation f_k^*.

We observe that the transmission power (and hence the energy consumption) of users on subchannel positions $m \in \mathcal{M}^l$ is related to the other user who shares the same subchannel for transmission, according to (15). So dependence between the two users on the same subchannel should be taken into account. Inspired by the structural property shown in (9), we introduce a new variable, $\hat{x}_{k,j,n}$, to denote the subchannel assignment in order to capture the interplay between the two users on the same subchannel. In particular, when $\hat{x}_{k,j,n} = 1$, user k is assigned to subchannel position $m = 2n - 1$ and user j ($j \in \mathcal{K}$) is assigned to subchannel position $m = 2n$. Otherwise, $\hat{x}_{k,j,n} = 0$. Denote by $\hat{e}_{k,j,n}$ the total energy consumption of users k and j on subchannel n corresponding to $\hat{x}_{k,j,n} = 1$. According to (9) and (15), the energy consumption of the two users can be expressed as

$$\begin{aligned} \hat{e}_{k,j,n} = {} & w_k \sigma^2 / |h_{k,2n-1}|^2 \exp(a r_j)[\exp(a r_k) - 1] t_k^{\mathrm{o}} \\ & + w_j \sigma^2 / |h_{j,2n}|^2 [\exp(a r_j) - 1] t_j^{\mathrm{o}}, \quad k \neq j. \end{aligned} \tag{16}$$

Therefore, $\hat{e}_{k,j,n}$ can be calculated under the current optimal computation resource allocation f_k^* with the help of (14) in calculating $r_k, t_k^{\mathrm{o}}, r_j$, and t_j^{o}. Moreover, $\hat{e}_{k,j,n} = 100$ (i.e., an arbitrarily large energy consumption value) $\quad \forall k = j \in \mathcal{K} \quad \forall n \in \mathcal{N}$ is assumed to prevent such assignment.

We formulate the subchannel assignment subproblem as

$$\text{(Sub-SA)} \ \min_{\hat{x}_{k,j,n}} E = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{K} \hat{x}_{k,j,n} \hat{e}_{k,j,n}$$

$$\text{subject to: } \hat{x}_{k,j,n} \in \{0, 1\} \quad \forall k, j \in \mathcal{K} \quad \forall n \in \mathcal{N} \tag{17a}$$

$$\sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{K}} \hat{x}_{k,j,n} = 1 \quad \forall n \in \mathcal{N} \tag{17b}$$

$$\sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{K}} \hat{x}_{k,j,n} + \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{K}} \hat{x}_{j,k,n} = 1 \quad \forall k \in \mathcal{K} \tag{17c}$$

$$(|h_{k,n}| - |h_{j,n}|) \hat{x}_{k,j,n} \geq 0 \quad \forall k, j \in \mathcal{K} \quad \forall n \in \mathcal{N}. \tag{17d}$$

Problem (Sub-SA) is an ILP problem. Constraint (17b) indicates that each subchannel holds two different users since the case that a subchannel holds only one user can be prevented by setting $\hat{e}_{k,j,n}$ to a large number, e.g., $\hat{e}_{k,j,n} = 100$. Constraint (17c) requires that each user can only choose one subchannel position, either a subchannel position with a large channel gain or a small channel gain. Constraint (17d) shows that if $\hat{x}_{k,j,n} = 1$, user k should have a no smaller channel gain than user j on subchannel n. Note that the solution of subchannel assignment subproblem can also be interpreted as a NOMA grouping decision and a subchannel assignment of different NOMA groups. More specifically, if $\hat{x}_{k,j,n} = 1$, user k and user j form a NOMA group and transmit over subchannel n. The decoding order is further decided according to their channel gains. Thus, if the subchannel assignment subproblem is solved optimally, we can attain the optimal NOMA grouping and the optimal subchannel assignment of different NOMA groups simultaneously. Unfortunately, it turns out that problem (Sub-SA) can be extremely challenging.

Theorem 2: Problem (Sub-SA) is NP-hard.

Proof: Please refer to Appendix C. ■

Due to the NP-hardness, we cannot expect exact and polynomial-time deterministic solution algorithms. We then apply BNB methods to solve problem (Sub-SA) [35]. Although BNB methods are often slow and may have exponential worst-case performance, we observe from extensive simulations that BNB methods produce the optimal solutions efficiently for our problem. Note that after obtaining the optimal solution for problem (Sub-SA), we can transform $\hat{x}_{k,j,n} = 1$ to $x_{k,2n-1} = 1, x_{j,2n} = 1$ in the original problem (WSEM) and then use the updated subchannel assignment to calculate the new power and computation resource allocations.

C. *Overall Algorithm and Complexity Analysis*

Both subproblems can be solved efficiently, which enables an effective solution to problem (WSEM). We summarize the overall algorithm to compute the resource allocation and subchannel assignment in Algorithm 1. By iteratively solving problem (Sub-RA) in step (5) and problem (Sub-SA) in step (6), the optimal solution of problem (Sub-RA) or problem (Sub-SA) is obtained.

Algorithm 1: Overall Algorithm.

```
1: Input: Number of maximum iterations $I$, system parameters $K, N, B, \tau$, base station information $F$, user information $w_k, d_k, C_k, P_k, \forall k \in \mathcal{K}$, subchannel information $h_{k,m}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}$.
2: Output: The optimal subchannel assignment $\boldsymbol{x}^*$, computation resource allocation $\boldsymbol{f}^*$, power allocation $\boldsymbol{p}^*$, and minimal energy consumption $E^*$.
3: Initialization: Randomly choose a subchannel assignment $\boldsymbol{x}^{(0)} \in \mathcal{X}$; define the set of considered subchannel assignments as $\mathcal{X}_r = \boldsymbol{x}^{(0)}$; set $E^*$ to some arbitrarily large value.
4: for $i = 1 : I$ do
5:   Calculate the optimal energy consumption $E^{(i)}$, computation resource allocation $\boldsymbol{f}^{(i)}$ and power allocation $\boldsymbol{p}^{(i)}$ by solving problem (Sub-E-RA), given the subchannel assignment $\boldsymbol{x}^{(i-1)}$;
6:   Based on $\boldsymbol{f}^{(i)}$, find the optimal subchannel assignment $\hat{\boldsymbol{x}}^{(i)}$ by solving problem (Sub-SA) and transfer it into $\boldsymbol{x}^{(i)}$;
7:   if $\boldsymbol{x}^{(i)} \in \mathcal{X}_r$ then
8:     Randomly choose a subchannel assignment decision $\boldsymbol{x}^{(i)} \in \mathcal{X} \backslash \mathcal{X}_r$
9:   end if
10:  $\mathcal{X}_r = \mathcal{X}_r \bigcup \boldsymbol{x}^{(i)}$
11:  if $E^{(i)} < E^*$ then
12:    $\boldsymbol{x}^* = \boldsymbol{x}^{(i-1)}, \boldsymbol{f}^* = \boldsymbol{f}^{(i)}, \boldsymbol{p}^* = \boldsymbol{p}^{(i)}, E^* = E^{(i)}$;
13:  end if
14: end for
```

The objective value of problem (WSEM) is nonincreasing in each of these two steps. Moreover, we design steps (7)–(9) to avoid repeatedly calculating the considered subchannel assignments. In other words, when the newly determined subchannel assignment has already been calculated previously, the algorithm will randomly choose another subchannel assignment that has not been considered before.

The computational complexity of Algorithm 1 lies in *line 5–9*. In *line 5*, the dual decomposition method is applied to solve the power and computation resource allocations. In particular, we first take the Lagrangian form of problem (Sub-E-RA) to relax the coupling constraint (13a)

$$\begin{aligned} \min_{f_{2n-1}, f_{2n}} \quad & \sum_{n=1}^{N} E_n(f_{2n-1}, f_{2n}) \\ & \quad + \lambda \left[\sum_{n=1}^{N} (f_{2n-1} + f_{2n}) - F \right] \\ \text{s.t.} \quad & (f_{2n-1}, f_{2n}) \in \mathcal{F}_n. \end{aligned} \tag{18a}$$

Constraints (9), (11), and (12) are involved in the derivations of problem (18) and are hence omitted in the following explanations for dual decomposition methods and subsequent analysis. Afterward, we can solve the optimization problem in two tiers. For the lower tier problem, we solve the computation resource allocation problem for each subchannel n independently

$$\begin{aligned} \min_{f_{2n-1}, f_{2n}} \quad & E_n(f_{2n-1}, f_{2n}) + \lambda \left(f_{2n-1} + f_{2n} \right) \\ \text{s.t.} \quad & (f_{2n-1}, f_{2n}) \in \mathcal{F}_n. \end{aligned} \tag{19a}$$

Define the minimal value obtained from problem (19) for each subchannel as $g_n(\lambda)$. For the higher tier problem, we update the dual variable by solving the dual problem

$$\begin{aligned} & \max_{\lambda} g(\lambda) = \sum_{n=1}^{N} g_n(\lambda) - \lambda F \\ & \text{s.t. } \lambda \geq 0. \end{aligned} \tag{20a}$$

We further apply subgradient methods to update the auxiliary variable λ. Finally, given a subchannel assignment, we can obtain the optimal computation resource allocation efficiently. The relevant power allocation decisions can also be deduced. Let T_D be the number of updates required for the subgradient methods to ensure the convergence of the dual problem (20). After each update of λ, N lower tier problems (19) are solved. If we apply the gradient descent methods to solve problems (19), let T_P be the number of iterations required to guarantee the convergence of each of the N lower tier problems. Then, the complexity is $\mathcal{O}(T_P T_D N)$. In *line 6*, BNB methods are applied to decide the subchannel assignment, which is claimed to be quite efficient in our problem and is justified by the numerical results in terms of analysis on BNB ratio (see Section V-D). We denote the complexity of solving the subchannel assignment subproblem by O_{BNB}. In *line 7–9*, at most I comparisons are implemented. Hence, the complexity of the overall scheme is $\mathcal{O}(T_P T_D N I + O_{\text{BNB}} I + I))$. For one thing, problem (19) is convex and problem (20) is concave, and hence T_P and T_D can be tuned to be acceptable. For another thing, O_{BNB} is shown to be acceptable by extensive simulation results. However, the complexity of an exhaustive search for subchannel assignment when $K = 2N - 1$ and $K = 2N$ is at least exponential in N (i.e., $\mathcal{O}((2N)!/2^N) \geq \mathcal{O}(N^N)$) and is much larger than N^N when $N \geq 5$. Even in the case when $K = N + 1$ (the case we will cover in the extension to general cases in Section IV), the exhaustive search for subchannel assignment also has factorial complexity $\mathcal{O}((N+1)!N/2)$. Hence, the proposed scheme is acceptable to the system, which implies a major reduction in the computational complexity.

IV. Extension to General Cases

In this section, we show that we can generalize the special case described in Section II to cases $N \leq K \leq 2N$ and our algorithm proposed in Section III is applicable for general cases.

In particular, when $N \leq K < 2N$, we define a set $\mathcal{K}_v = \{K+1, \ldots, 2N\}$ of virtual users. Users $k \in \mathcal{K}$ can hence be differentiated as real users. In particular, virtual users have the following properties.

1) Virtual users occupy different subchannels so that each subchannel holds at least one real user.
2) $h_{k,m} > \max_{k' \in \mathcal{K}, m' \in \mathcal{M}} h_{k,'m'} \quad \forall k \in \mathcal{K}_v \quad \forall m \in \mathcal{M}$. Therefore, all virtual users will be assigned to odd positions of subchannels and real users on the same subchannels with virtual users have OMA transmission data rate according to (1).
3) Virtual users will not be allocated with any computation resource so that the computation resource at the BS is utilized by all real users.
4) Virtual users have no power allocation and hence no energy consumption.

Then, the energy minimization problem under cases $N \leq K < 2N$ can be formulated according to problem (WSEM) by changing the feasible set of $\boldsymbol{x}$ as

$$\mathcal{X} = \Bigg\{ \boldsymbol{x} \mid x_{k,m} \in \{0,1\} \quad \forall k \in \mathcal{K} \cup \mathcal{K}_v \quad \forall m \in \mathcal{M}$$
$$\sum_{k \in \mathcal{K} \cup \mathcal{K}_v} x_{k,m} = 1 \forall m \in \mathcal{M}^l; \sum_{k \in \mathcal{K}} x_{k,m} = 1 \quad \forall m \in \mathcal{M}^s$$
$$\sum_{m=1}^{M} x_{k,m} = 1 \forall k \in \mathcal{K} \cup \mathcal{K}_v; \sum_{k \in \mathcal{K} \cup \mathcal{K}_v} x_{k,m} |h_{k,m}|^2$$
$$\geq \sum_{k \in \mathcal{K}} x_{k,m+1} |h_{k,m+1}|^2 \quad \forall m \in \mathcal{M}^l \Bigg\}.$$

Compared to the feasible set of $\boldsymbol{x}$ presented in Section II-B, apart from changing the feasible set of k from $\mathcal{K}$ to $\mathcal{K} \cup \mathcal{K}_v$, we specify in the second line that each odd subchannel position will be allocated one real or virtual user, whereas each even subchannel position will be allocated one real user. Moreover, $p_k = f_k = 0 \quad \forall k \in \mathcal{K}_v$, which are not optimization variables.

We then illustrate the implementation of our algorithm with virtual users. The joint optimization problem can still be decomposed into two subproblems as problem (Sub-RA) and problem (Sub-SA). We can follow (7)–(13) to formulate the resource allocation subproblem. For subchannels holding two real users, the expressions and constraints are exactly the same as (7)–(13). For subchannels holding one real user and one virtual user, modifications are needed. Specifically, the data rate, offloading time, execution time, and energy consumption of the virtual users are all set to 0, and constraints related to these terms can be eliminated. Meanwhile, the resource allocation subproblem only optimizes the power allocation and computation resource allocation for real users, i.e., p_m and f_m with index m corresponding to real users. The newly formulated resource allocation subproblem is still convex and dual decomposition methods can be applied to solve it.

After obtaining the optimal computation resource, we follow (14)–(17) to formulate the new subchannel assignment subproblem by extending the feasible set of $k, j \in \mathcal{K}$ to $k, j \in \mathcal{K} \cup \mathcal{K}_v$ and specifying that $\sum_{n \in \mathcal{N}} \sum_{k \in \mathcal{K} \cup \mathcal{K}_v, k \neq j} \hat{x}_{k,j,n} = 0 \quad \forall j \in \mathcal{K}_v$. Other differences are similar to those when formulating the new resource allocation subproblem. The data rate, offloading time, power allocation, and energy consumption of all virtual users are 0. Therefore, when calculating $\hat{e}_{k,j,n}$ for a virtual user k and a real user j according to (16), we only calculate the second term that accounts for the real user's energy consumption. The newly formulated subchannel assignment subproblem can be solved by BNB methods efficiently.

Then, we follow the logic of Algorithm 1 to obtain the final solutions. In fact, $K = N$ is OMA (FDMA) transmission and we use this case as a comparison scheme to NOMA transmission in Section V-B. Cases $K < N$ are for OMA transmission with sufficient communication resources and are out of the scope of this article. For cases $K > 2N$, admission control should be applied since the system has the maximum capacity of $2N$. Under this situation, the BS will provide MEC service to $2N$ users only so as to entertain as many users as possible. We may provide higher priority to users who have partitioned their workloads to fit the time-slotted system but have not finished offloading the whole task, as described in the last paragraph of Section II-A, or users who first request for the MEC service.

V. Simulation Results and Performance Analysis

In this section, we examine the performance of our proposed algorithm and analyze the influence of the NOMA scheme on users' energy consumption in MEC systems.

A. System Settings

The system settings are as follows unless expressly stated. Each user has a random data size d_k, which follows a uniform distribution $d_k \in [0.05, 0.5]$ Kb. $C_k = 1000$ cycles/b is assumed identical for all users. Users are uniformly distributed with a random distance to the BS $a_k \in [5100]$m (parameter settings are scaled according to the system settings in [26]). Without loss of generality, $P_k = 30$ dBm and $w_k = 1 \quad \forall k \in \mathcal{K}$ are assumed. The BS has the total computation resource $F = 20 \times 10^9$ cycles/s. The noise power spectrum density at the BS is $N_0 = -174$ dBm/Hz. The length of time slot $\tau = 0.5$ ms. The total bandwidth $B = 10$ MHz. We consider the path loss as $\sqrt{G_0 (a_k / a_0)^{-\theta}}$, where $G_0 = -40$ dB corresponds to the path loss at a reference distance $a_0 = 1$ m, and $\theta = 3.7$ is the path loss exponent [17]. Let $g_{k,m} \sim \mathcal{CN}(0,1)$ denote the small-scale Rayleigh fading between the BS and user k over subchannel position m and then user k's channel gain on subchannel position m can be represented as $h_{k,m} = \sqrt{G_0 (a_k / a_0)^{-\theta}} g_{k,m} \in \mathbb{C}$. $K = 2N$ is assumed in NOMA simulations for better interpretation and better comparison, considering that $N < K < 2N$ refers to a combination of NOMA and OMA transmission.

B. Comparison Schemes

Denote by NOMA-J our proposed joint optimization over resource allocation and subchannel assignment in the NOMA-enabled MEC system. To evaluate the performance of our proposed algorithm, we consider several schemes for comparison. We first consider two other NOMA-enabled schemes: first, NOMA-CH: Users adopt NOMA for computation offloading and are allocated with equal computation resource at the BS. The subchannel assignment is optimized. Second, NOMA-COMP: NOMA users are randomly assigned to subchannels with an assignment decision $\boldsymbol{x} \in \mathcal{X}$. Users' power and computation resource allocations are optimized. Then, we further consider FDMA-based systems to appraise the influence of NOMA technology on computation offloading, where the total bandwidth is divided into N orthogonal subchannels with equal bandwidth. Each subchannel can hold one user, and $K = N$ is assumed. The FDMA-based comparison schemes are described as follows.

1) *FDMA-CH:* Each user is allocated with equal computation resource, and the subchannel assignment is optimized.
2) *FDMA-COMP:* Each user is randomly assigned to one subchannel, followed by the power and computation resource allocation optimization.
3) *FDMA-J:* The resource allocation and subchannel assignment are both optimized.

These FDMA-based methods are designed according to our proposed NOMA-based system for easy comparison and interpretation. Related NOMA-based and FDMA-based systems have the same settings for the total bandwidth, the total computation resource, the length of time slots, users' data sizes, and

TABLE I
FAILURE PROBABILITIES OF NOMA-CH AND FDMA-CH

No. of users	4	6	8	10	12	14	16	18	20	22
NOMA-CH	0	0	0	0	0	0	0	0	0.18	0.84
FDMA-CH	0	0	0	0	0	0	0	0.16	0.65	1

TABLE II
COMPARISON OF ENERGY CONSUMPTION RATIOS BETWEEN NOMA-J AND NOMA-CH AND FDMA-J AND FDMA-CH

No. of users	4	6	8	10	12	14	16	18	20	22
NOMA-J/CH	0.9389	0.9503	0.9009	0.8797	0.7678	0.7419	0.5648	0.4191	0.1062	0.0175
FDMA-J/CH	0.9709	0.8929	0.9082	0.8893	0.7459	0.6710	0.4750	0.0652	0.0926	-

TABLE III
BNB EFFICIENCY FOR DIFFERENT SYSTEM SIZES

No. of users	4	6	8	10	12	14	16	18	20	22
BNB ratio	0.21	0.65	0.64	0.89	0.91	0.93	0.98	0.96	0.84	0.93

distances to the BS except for channel states. We set the maximum number of iterations $I = 10$, the parameter $L_\eta = 100$, and study different scales of the MEC systems, ranging from a 4-user system to a 22-user system. Note that for the 4-user system, there are only six subchannel assignments in total, so we set $I = 6$ for this case. For each setting, we run 100 realizations independently with different Rayleigh fading channels and take the average over all realizations for analysis.

C. Performance Evaluation for the Sub-RA Approach

To begin with, we evaluate the performance of our power and computation resource allocation approach. We show that it enables our system to serve more users with limited communication resources and helps reduce users' energy consumption.

We notice that while NOMA-COMP, NOMA-J, FDMA-COMP, and FDMA-J (all with power and computation resource allocations) can always have feasible solutions to the offloading problem, NOMA-CH and FDMA-CH (without power and computation resource allocations) may fail to find feasible solutions. More quantitatively, we define the failure probability as the ratio of the number of infeasible realizations over the number of all realizations, and present the failure probability for NOMA-CH and FDMA-CH in Table I.

At the same time, the failure probabilities of NOMA-COMP, NOMA-J, FDMA-COMP, and FDMA-J are all 0. We then compare users' energy consumption of NOMA-J and NOMA-CH, as well as that of FDMA-J and FDMA-CH in Table II. The energy consumption ratio between NOMA-J and NOMA-CH is named NOMA-J/CH. The ratio FDMA-J/CH is defined similarly.

By combining the two tables, we can better analyze the performance of our power and computation resource allocation approach.

When $K \leq 16$, the failure probabilities of NOMA-CH and FDMA-CH are all 0. It is also shown in Table II that users can greatly mitigate the influence of nonoptimal power and computation resource allocations and achieve a similar performance to NOMA-J (FDMA-J) in NOMA-CH (FDMA-CH) schemes when $K \leq 12$. Even for 16, NOMA-CH and FDMA-CH still consumes only twice the energy compared to the corresponding joint optimization schemes. This indicates that in these cases, the communication resources are still sufficient. Although users are allocated with equal computation resource, users can reduce the total energy consumption by applying an appropriate subchannel assignment. However, as the number of users increases, communication resources become scarce in terms of bandwidth sharing. Users with too much data to process may suffer from a long offloading latency. Without flexible computation resource allocation to compensate for such long offloading latency, the BS may fail to find a suitable subchannel assignment solution to allow all users to finish data processing within the desired latency. The latency requirement hence may be violated in NOMA-CH (FDMA-CH) schemes. Therefore, NOMA-CH's (FDMA-CH's) failure probability becomes positive when $K \geq 20$ ($K \geq 18$). It is also shown that FDMA-CH does not work at all when $K = 22$. This indicates that NOMA-CH and FDMA-CH cannot be applied in a large-scale MEC system with a crowd of users. Even for realizations with feasible solutions, NOMA-CH and FDMA-CH can have high energy consumption. This is revealed by the reduced energy consumption comparison ratio. The reason is that in order to satisfy the latency requirement, users with large data sizes to process have to transmit with more power to have a desirable offloading latency.

D. Performance Evaluation for the Sub-SA Approach

Next, we assess the performance of our subchannel assignment approach. We first show that the subchannel assignment subproblem can be solved by BNB methods efficiently and then show that the subchannel assignment updates help greatly save users' energy consumption.

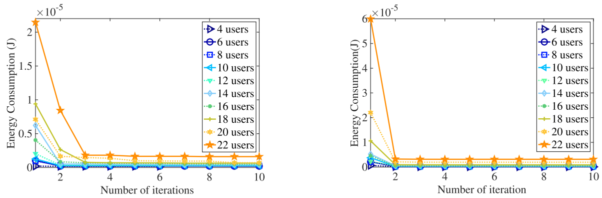
Define one partition as a division of feasible set into two convex sets when we obtain a decimal solution in BNB methods. Therefore, the total number of partitions needed to obtain an integral solution in BNB methods can be used to reflect the BNB efficiency. The less partitions we need, the faster we obtain the integral solutions. We observe that when updating the subchannel assignment by BNB methods, no partitions are needed in some realizations, whereas only a few partitions are needed to obtain integral solutions in other realizations. We record the number of partitions conducted in each subchannel assignment update, and present the ratio between the total number of partitions and the total number of subchannel assignment updates for all realizations in Table III. We name the ratio as BNB ratio for simplicity.

As is shown in the table, the BNB ratio is always smaller than 1 even for $K = 22$. This indicates that we can directly obtain the integral solutions after eliminating the integral constraint in some realizations regardless of the system size, and the BNB-based subchannel assignment algorithm works quite well in other realizations (considering that the BNB ratio can be larger than 1, e.g., 20 if each channel assignment requires 20 partitions in BNB methods on average).

We further show users' energy consumption ratio between NOMA-J (which has the same initial subchannel assignment as NOMA-COMP but can perform subchannel assignment updates) and NOMA-COMP, as well as that between FDMA-J

TABLE IV
COMPARISON OF ENERGY CONSUMPTION RATIOS BETWEEN NOMA-J AND NOMA-COMP AND FDMA-J AND FDMA-COMP

No. of users	4	6	8	10	12
NOMA-J/COMP	0.8075	0.1988	0.2639	0.2129	0.1629
FDMA-J/COMP	0.1313	0.1105	0. 1332	0.0830	0.1597
No. of users	14	16	18	20	22
NOMA-J/COMP	0.0657	0.1219	0.0685	0.0998	0.0750
FDMA-J/COMP	0.0857	0.1104	0.0848	0.0887	0.0524



(a) (b)

Fig. 1. Dynamics of users' energy consumption with increasing number of iterations when $K = 2N$. (a) NOMA-J system. (b) FDMA-J system.

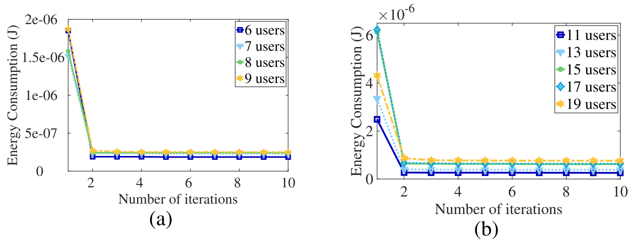
and FDMA-COMP in Table IV. The ratios are named NOMA-J/COMP and FDMA-J/COMP, respectively.

It is worth noting that since the power and computation resource allocation subproblem is convex, the optimal resource allocation solutions are always available given the subchannel assignment. Therefore, the energy consumption saving achieved by NOMA-J (FDMA-J) compared to NOMA-COMP (FDMA-COMP) is due to subchannel assignment updates. Although there are no clear trends for the energy ratios (the variation of the ratios is because NOMA-COMP and FDMA-COMP highly depend on the random initial subchannel assignment), it is easy to show that the subchannel assignment works in all sizes of systems tested and helps reduce users' energy consumption notably.

E. Performance Evaluation for the Overall Algorithm

After analyzing the two subproblems separately, we examine the overall algorithm for the joint optimization of power and computation resource allocations and subchannel assignment. We present the dynamics of users' total energy consumption of NOMA-J and FDMA-J in Fig. 1. We take the average of the minimal energy consumption E^* over 100 settings in each iteration, and then plot it versus the number of iterations.

It can be seen that our algorithm can efficiently find a good joint subchannel assignment and resource allocation. Only a few iterations are needed to have a great energy consumption reduction compared to the initial state. When brute force in the subchannel assignment is tractable ($K = 6, 8$ in NOMA-based systems, and $K = 4, 6$ in FDMA-based systems), we compare the results of our NOMA-J and FDMA-J algorithms with those of brute force, and observe that our algorithm finds exactly the same optimal solutions as brute force. (The brute force of the NOMA-based and FDMA-based systems find global optima by exhaustive search over all subchannel assignments followed by optimal resource allocation, and are named as NOMA-B and FDMA-B, respectively. The energy consumptions of both schemes are shown in Fig. 3.) We also present other global optima in Table V of cases that are not shown in Fig. 3, which correspond to the extended general cases discussed in Section IV. We find that our algorithm finds the global optimal solutions as brute force.



(a) (b)

Fig. 2. Dynamics of users' energy consumption with increasing number of iterations when $N < K < 2N$ for our proposed scheme (hybrid NOMA and OMA transmission). (a) $N = 5$. (b) $N = 10$.

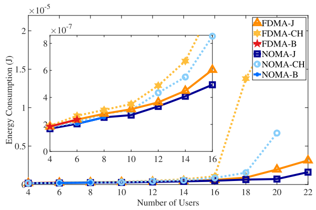


Fig. 3. Energy consumption of different schemes with an increasing number of users.

TABLE V
COMPARISON OF ENERGY CONSUMPTION IN JOULE BETWEEN NOMA-J AND NOMA-B

System scale	$N = 3, K = 4$	$N = 3, K = 5$	$N = 4, K = 5$
NOMA-J	7.07×10^{-8}	1.95×10^{-7}	1.50×10^{-7}
NOMA-B	7.07×10^{-8}	1.95×10^{-7}	1.50×10^{-7}
System scale	$N = 4, K = 6$	$N = 4, K = 7$	$N = 5, K = 6$
NOMA-J	1.85×10^{-7}	2.17×10^{-7}	1.68×10^{-8}
NOMA-B	1.85×10^{-7}	2.17×10^{-7}	1.68×10^{-7}

When brute force is not viable, we can still observe a considerable energy reduction compared to the initial state achieved by our joint optimization.

Additionally, to show that our algorithm works well for cases $N < K < 2N$ (cases $K = N$ have been shown in FDMA-J scheme), we present the dynamics of users' energy consumption with increasing number of iterations when $N = 5$ and $N = 10$ in Fig. 2. Different values of K are considered in these cases. Note that when $N < K < 2N$, users apply hybrid NOMA and OMA transmission. It is verified that our algorithm also achieves good performance for general cases.

F. Impact Analysis of NOMA in MEC Systems

Furthermore, we analyze the impact of NOMA on energy-efficient offloading in MEC systems by comparing different schemes. We do not show the plots for NOMA-COMP and FDMA-COMP, considering that these two schemes have been shown to be inefficient even for a small system scale and cannot reflect any trend of energy consumption with an increasing number of users. The energy consumption curves of the other schemes, including NOMA-B and FDMA-B, are plotted in Fig. 3. (We also neglect NOMA-CH and FDMA-CH when $K = 22$ since they have large failure probability in this case.)

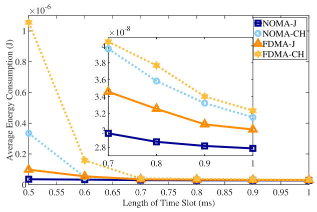


Fig. 4. Average energy consumption for each user when $K = 20$.

It is noted that NOMA-based schemes always perform better than FDMA-based schemes when $K \leq 10$. When $K \geq 12$, the resource allocation becomes important in reducing users' energy consumption. Hence, NOMA-CH cannot outperform FDMA-J in terms of energy minimization. However, we observe that NOMA-J always outperforms FDMA-J, and NOMA-CH always outperforms FDMA-CH in terms of energy minimization. This indicates that, when multiple users request offloading with constrained communication resources, NOMA's efficient use of bandwidth will have positive effect on reducing users' energy consumption, and we can conclude that it is better to apply NOMA technology for energy-efficient computation offloading in latency-stringent MEC systems.

Finally, we change τ from 0.5 to 1 ms and plot the average energy consumption for each user when $K = 20$ in Fig. 4.

When the length of a time slot becomes larger and data sizes remain unchanged, it is equivalent to having looser latency requirements. We can observe that NOMA-based scheme can always achieve lower energy consumption compared to corresponding FDMA-based scheme in latency-intensive cases. Also, NOMA-based MEC systems are less sensitive to the latency requirement compared to FDMA-based MEC systems in terms of energy consumption.

VI. Conclusion

In this article, we have investigated the joint resource allocation and subchannel assignment problem in NOMA-enabled MEC systems, where users apply NOMA technology to perform computation offloading. An energy-efficient problem aimed at minimizing the total energy consumption of all users has been formulated as a nonconvex combinatorial problem. An algorithm has been devised to solve the problem by dealing with resource allocation and subchannel assignment subproblems in an iterative way. For the power and computation resource allocation subproblem, we have studied the hidden convexity under the optimal conditions, and efficiently solved the problem by dual decomposition methods. For the subchannel assignment subproblem, we have proved the NP-hardness of the reformulated problem and solved it exactly by branch and bound methods. The overall algorithm has been demonstrated to be computationally efficient, with only a small number of iterations required to produce a decent performance. Meanwhile, numerical results have shown that compared to FDMA-based MEC systems, NOMA-based MEC systems have great advantages in reducing energy consumption, and our proposed algorithm can enable MEC service for multiple users with constrained communication resources. Regarding the future work, the combination with massive multiple-input multiple-output can be considered, which may further improve the performance of NOMA-enabled offloading. Moreover, the imperfect CSI situations can also be taken into account to make the system model more practical.

Appendix A
Proof of Lemma 1

The main idea is to prove that the energy consumption of each user is monotonically increasing with its transmission power. Hence, all users tend to transmit with the minimal allowable transmission power to save energy, which leads to the longest allowable offloading time and the binding latency constraints. To show the monotonicity, we first consider the energy consumption of user $2n$ on subchannel n, which can be easily proved to be monotonically increasing with transmission p_{2n}. For user $2n-1$, we can treat the interference from user $2n$ as noise. The same conclusion can be drawn that user $2n-1$ prefers to transmit with minimal transmission power. It is obvious that low-power transmission is better for both users in a NOMA group because when user $2n$ transmits with minimal power, it introduces less interference during user $2n-1$'s transmission, which also helps reduce user $2n-1$'s transmission power. As different NOMA groups of users transmit on different subchannels and incur no tradeoff, all users prefer to transmit with minimal transmission power, resulting in the longest allowable transmission time. Finally, under the optimal conditions, constraints (10a) are binding for all users regardless of the computation resource allocation. □

Appendix B
Proof of Theorem 1

The convexity of the feasible region is proved by showing that the feasible set of (f_{2n-1}, f_{2n}) is an intersection of the convex sets defined by all constraints. The convexity of the objective is proved by showing that its Hessian matrix with respect to (f_{2n-1}, f_{2n}) is positive semidefinite. The derivations follow the fundamental mathematics, and hence are omitted for brevity. For the complete proof, please see the supplementary material.

Appendix C
Proof of Theorem 2

We first transfer problem (Sub-SA) into an equivalent form as follows (here, "E-SA" refers to "equivalent subchannel assignment" subproblem):

$$\text{(Sub-E-SA)} \quad \min_{x_{k,j,n}} \quad E = \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{K} \hat{x}_{k,j,n}\eta_{k,j,n}$$

$$\text{subject to:} \quad \text{Constraints} \quad (17a)\text{–}(17c)$$

where $\eta_{k,j,n} = \hat{e}_{k,j,n} + L_\eta \mathcal{I}\{|h_{k,n}| < |h_{j,n}|\}$. To relax constraint (17d), an indicator function $\mathcal{I}\{|h_{k,n}| < |h_{j,n}|\}$ is added into the objective with a large weight L_η. Undesirable subchannel assignments that violate the SIC decoding requirement are prevented by introducing a heavy penalty. To prove the NP-hardness of problem (Sub-SA), it is equivalent to prove that problem (Sub-E-SA) is NP-hard. The proof is further made by proving the NP-completeness of the corresponding recognition version problem of problem (Sub-E-SA) [36]. Denote by $\hat{\boldsymbol{x}} := \{\hat{x}_{k,j,n} \mid k, j \in \mathcal{K}, n \in \mathcal{N}\}$ the new subchannel assignment variables. The recognition problem can be described as Problem (1).

Problem 1: Is there a feasible solution formed by $\hat{\boldsymbol{x}}$, where constraints (17a) to (17c) are satisfied?

Problem (1) follows the recognition version description of an ILP problem [36]. To show the NP-completeness of Problem (1), we will show that Problem (1) $\in$ NP and the exact cover problem is reducible to Problem (1) in polynomial time, where the exact cover problem is a known NP-complete problem [37].

1) Problem (1) $\in$ NP. It is easy to see that we can nondeterministically guess a solution $\hat{x}_{k,j,n}$ and deterministically verify whether constraints (17a) to (17c) are satisfied.

2) The exact cover problem is reducible to Problem (1) in polynomial time. To proceed, we first introduce the exact cover problem [37]. We are given a finite nonempty set U, which is called the universe, and a collection $\mathcal{S} = \{S_1, \ldots, S_q, \ldots, S_Q\}$ $(Q \geq 1)$ of nonempty subsets of U, where $S_q \quad \forall q \in \{1, \ldots, Q\}$, are the subsets of U. The problem is whether there exists an exact cover, that is, a subcollection $\mathcal{C} \subseteq \mathcal{S}$ of subsets of $\mathcal{S}$ such that the sets in $\mathcal{C}$ are disjoint and their union is equal to U. In other words, every element of U belongs to exactly one set in $\mathcal{C}$. For example, let $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$ and $\mathcal{S} = \{\{u_1, u_4\}, \{u_2, u_4, u_6\}, \{u_2, u_5\}, \{u_3, u_6\}, \{u_4, u_5, u_6\}\}$. The subcollection $\mathcal{C} = \{\{u_1, u_4\}, \{u_2, u_5\}, \{u_3, u_6\}\}$ is an exact cover. We now reduce the exact cover problem to Problem (1). We construct our universe U and the collection $\mathcal{S}$ as follows. First, we map the subchannel index $n \in \{1, \ldots, N\}$ to $n' \in \{K+1, \ldots, K+N\}$ and obtain a new subchannel assignment variable $\tilde{x}_{k,j,n'}$. Now, $\tilde{x}_{k,j,n'} = 1$ indicates user k is assigned to subchannel position $2(n'-K)-1$, whereas user j is assigned to subchannel position $2(n'-K)$. This step aims to differentiate the indexes for users and subchannels. We hence have the universe $U = \{1, \ldots, K, K+1, \ldots, K+N\}$, which is a union of all users and subchannels. Then, the construction of collection $\mathcal{S}$ is described in the following. For each subchannel that is indexed by n', we choose two different users k' and j'. We have $k,' j' \in \mathcal{K}$ and $k' \neq j'$. Unlike k and j indexes in variables $\tilde{x}_{k,j,n'}$, indexes k' and j' do not relate to any subchannel position information. Let $\{k,' j,' n'\}$ be an element in $\mathcal{S}$, and $\mathcal{S}$ is a collection of all possible assignments to every subchannel, i.e., $\mathcal{S} = \{\{1, 2, K+1\}, \ldots, \{K-1, K, K+N\}\}$. Recall that the problem is whether there is a feasible solution formed by $\hat{\boldsymbol{x}}$, where constraints (17a) to (17c) are satisfied, which is exactly whether there is a subcollection $\mathcal{C} \subseteq \mathcal{S}$ to be an exact cover of U. To be specific, if an exact cover of U exists, for each element in $\mathcal{C}$, which is in the form of $\{k,' j,' n'\}$, we can extract the maximum number n' and identify the subchannel $n' - K$. The remaining two numbers k' and j' are indexes of users. We further compare $h_{k,'n'-K}$ and $h_{j,'n'-K}$. If $h_{k,'n'-K} > h_{j,'n'-K}$, then $\hat{x}_{k,'j,'n'-K} = 1, \hat{x}_{j,'k,'n'-K} = 0$; if $h_{k,'n'-K} < h_{j,'n'-K}$, then $\hat{x}_{k,'j,'n'-K} = 0, \hat{x}_{j,'k,'n'-K} = 1$; and if $h_{k,'n'-K} = h_{j,'n'-K}$, we can randomly assign one of $\hat{x}_{k,'j,'n'-K}$ and $\hat{x}_{j,'k,'n'-K}$ to be zero and the other to be one. It is obvious that the transformation is in polynomial time, since the size of $\mathcal{S}$ is $|\mathcal{S}| = K(K-1)N/2$. Thus, we guarantee that if we have an algorithm that can find the solution of the exact cover problem in polynomial time, we can have an efficient algorithm to solve Problem (1).

Therefore, Problem (1) is NP-complete and problem (Sub-SA) is NP-hard. □

References

[1] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[2] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[3] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on multi-access edge computing for Internet of Things realization," *IEEE Commun. Surv. Tut.*, vol. 20, no. 4, pp. 2961–2991, Oct.–Dec. 2018.

[4] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surv. Tut.*, vol. 19, no. 3, pp. 1657–1681, Jul.–Sep. 2017.

[5] H. Li, H. Xu, C. Zhou, X. Lü, and Z. Han, "Joint optimization strategy of computation offloading and resource allocation in multi-access edge computing environment," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10214–10226, Sep. 2020.

[6] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency minimization for intelligent reflecting surface aided mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2666–2682, Jul. 2020.

[7] H. Peng, Q. Ye, and X. Shen, "Spectrum management for multi-access edge computing in autonomous vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 7, pp. 3001–3012, Jul. 2020.

[8] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surv. Tut.*, vol. 17, no. 4, pp. 2347–2376, Oct.–Dec. 2015.

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

[10] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[11] W. Wang *et al.*, "Joint precoding optimization for secure SWIPT in UAV-aided NOMA networks," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5028–5040, Aug. 2020.

[12] N. Zhao *et al.*, "Security enhancement for NOMA-UAV networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 3994–4005, Apr. 2020.

[13] Z. Lin, M. Lin, J.-B. Wang, T. de Cola, and J. Wang, "Joint beamforming and power allocation for satellite-terrestrial integrated networks with non-orthogonal multiple access," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 657–670, Jun. 2019.

[14] J. Lu *et al.*, "UAV-enabled uplink non-orthogonal multiple access system: Joint deployment and power control," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10090–10102, Sep. 2020.

[15] Y. Wu, K. Ni, C. Zhang, L. P. Qian, and D. H. Tsang, "NOMA assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12244–12258, Dec. 2018.

[16] L. P. Qian, A. Feng, Y. Huang, Y. Wu, B. Ji, and Z. Shi, "Optimal SIC ordering and computation resource allocation in MEC-aware NOMA NB-IoT networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2806–2816, Apr. 2019.

[17] F. Wang, J. Xu, and Z. Ding, "Multi-antenna NOMA for computation offloading in multiuser mobile edge computing systems," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2450–2463, Mar. 2018.

[18] Z. Song, Y. Liu, and X. Sun, "Joint radio and computational resource allocation for NOMA-based mobile edge computing in heterogeneous networks," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2559–2562, Dec. 2018.

[19] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1299–1306, Apr. 2018.

[20] M. Zeng and V. Fodor, "Energy minimization for delay constrained mobile edge computing with orthogonal and non-orthogonal multiple access," *Ad Hoc Netw.*, vol. 98, 2020, Art. no. 102060.

[21] Q.-V. Pham, T. H. Nguyen, Z. Han, and W.-J. Hwang, "Coalitional games for computation offloading in NOMA-enabled multi-access edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1982–1993, Feb. 2020.

[22] X. Diao, J. Zheng, Y. Wu, and Y. Cai, "Joint computing resource, power, and channel allocations for D2D-assisted and NOMA-based mobile edge computing," *IEEE Access*, vol. 7, pp. 9243–9257, 2019.

[23] Z. Yang, C. Pan, J. Hou, and M. Shikh-Bahaei, "Efficient resource allocation for mobile-edge computing networks with NOMA: Completion time and energy minimization," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7771–7784, Nov. 2019.

[24] J. Zhu, J. Wang, Y. Huang, F. Fang, K. Navaie, and Z. Ding, "Resource allocation for hybrid NOMA MEC offloading," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4964–4977, Jul. 2020.

[25] B. Liu, C. Liu, and M. Peng, "Resource allocation for energy-efficient MEC in NOMA-enabled massive IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1015–1027, Apr. 2021.

[26] Y. Ye, R. Q. Hu, G. Lu, and L. Shi, "Enhance latency-constrained computation in MEC networks using uplink NOMA," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2409–2425, Apr. 2020.

[27] H. Lin, Y. Cao, Y. Zhong, and P. Liu, "Secure computation efficiency maximization in NOMA-enabled mobile edge computing networks," *IEEE Access*, vol. 7, pp. 87504–87512, 2019.

[28] L. Liu, B. Sun, X. Tan, Y. S. Xiao, and D. H. K. Tsang, "Resource allocation and channel assignment for NOMA-based mobile edge computing," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2019, pp. 1–6.

[29] Z. Zhang, H. Sun, and R. Q. Hu, "Downlink and uplink non-orthogonal multiple access in a dense wireless network," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2771–2784, Dec. 2017.

[30] T. Bai, J. Wang, Y. Ren, and L. Hanzo, "Energy-efficient computation offloading for secure UAV-edge-computing systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6074–6087, Jun. 2019.

[31] Z. Ning *et al.*, "When deep reinforcement learning meets 5G-enabled vehicular networks: A distributed offloading framework for traffic big data," *IEEE Trans. Ind. Inform.*, vol. 16, no. 2, pp. 1352–1361, Feb. 2019.

[32] S. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tut.*, vol. 19, no. 2, pp. 721–742, Apr.–Jun. 2016.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[34] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[35] J. Clausen, "Branch and bound algorithms—Principles and examples," Dept. Comput. Sci., Univ. Copenhagen, Copenhagen, Denmark, pp. 1–30, 1999.

[36] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Chelmsford, MA, USA: Courier Corp., 1998.

[37] M. R. Garey and D. S. Johnson, *Computers and Intractability*, vol. 29. New York, NY, USA: Freeman, 2002.