IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 23, 2021, accepted May 9, 2021, date of publication May 14, 2021, date of current version May 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3080283

A Q-Learning-Based Resource Allocation for Downlink Non-Orthogonal Multiple Access Systems Considering QoS

QI ZHAI[1], MIODRAG BOLIĆ[2], YONG LI[1], WEI CHENG[1], AND CHENXI LIU[1]
[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China
[2]School of Electrical Engineering Computer Science, University of Ottawa, Ottawa, CO K1N6N5, Canada

Corresponding authors: Miodrag Bolić (mbolic@uOttawa.ca) and Yong Li (ruikel@nwpu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61401360, and in part by the Fundamental Research Funds for the Central Universities under Grant 3102017zy026.

ABSTRACT As a technology that can accommodate more users and significantly improve spectral efficiency, non-orthogonal multiple access (NOMA) has attracted the attention of many scholars in recent years. The basic idea of NOMA is to implement multiple access in the power domain and decode the desired signal via successive interference cancellation (SIC). However, the resource allocation problem in such NOMA system is non-convex. It is difficult to directly solve this optimization problem through conventional methods. As such, we propose to apply a reinforcement learning (RL) approach based on cooperative Q-learning to solve the resource allocation problem in multi-antenna downlink NOMA systems. First, we formulate the resource allocation process as a sum rate maximization problem, subject to the power budget constraints and quality of service (QoS) condition. Second, we design a reward function to improve the sum rate while meeting the power and capacity constraints. Multiple Q-tables are created and cooperatively updated to get the optimal beamforming matrix. Then, we analyze the convergence of our proposed RL based power allocation method. Our simulations show that the proposed power allocation scheme yields excellent performance in terms of sum rate, energy efficiency, and spectral efficiency.

INDEX TERMS NOMA, power allocation, reinforcement learning, sum rate, spectral efficiency.

I. INTRODUCTION

The development of mobile internet and internet of things (IoT) has put forward challenging requirements for the fifth-generation wireless communication system (5G), which is expected to achieve higher spectral efficiency and lower latency [1], [2]. In addition, with the tremendous growth of the number of mobile users, next generation wireless networks need to achieve a greater number of user connections and higher data rates [3]. In traditional orthogonal multiple access (OMA), the number of users accommodated in the system is limited by the number of available orthogonal resources. Therefore, it is a challenge for traditional OMA to meet 5G's spectral efficiency and massive connectivity requirements. Based on this situation, many new technologies have emerged, such as non-orthogonal multiple access (NOMA), multiple-input multiple-output (MIMO), millimeter wave (mm Wave) communication, etc. [4]. As an effective way to improve spectral efficiency, NOMA has gained wide attention from scholars in recent years.

The associate editor coordinating the review of this manuscript and approving it for publication was Young Jin Chun.

A. MOTIVATION

With the rapid development of cellular networks, the number and types of new services of mobile terminals in the network have exploded, which has led to an increasing demand for long term data from users. On the other hand, people's pursuit of a fast life has promoted the emergence of new communication devices in the network, resulting in an increase of the sudden access requests [5], [6]. These demands strain the limited spectrum resources in cellular networks. And usually the overall performance of the network depends on how to efficiently and dynamically manage the resources in the network, such as time slots, power, and frequency band. In order to improve the spectral efficiency and throughput, and provide users with better quality of service (QoS), it is

very important to allocate resources in the network efficiently and reasonably.

The basic idea of NOMA is to realize non-orthogonal resource allocation among users by increasing complexity at receivers. With non-orthogonal resource allocation, NOMA can get massive connectivity and achieve higher spectral efficiency. Current research on the NOMA system mainly focuses on the code domain [7], [8] and power domain [9], [10]. In the power domain, the transmitter superposes signals with different power to be sent to multiple users on the shared spectrum. At the receiver, different users can decode the desired signal through successive interference cancellation (SIC) [11], [12]. In code domain NOMA, different spread-spectrum codes are assigned to different users and are then multiplexed over the same time-frequency resources [13]. Compared with traditional OMA that the number of users supported is limited by the available orthogonal resources, NOMA greatly improves the spectral efficiency [14].

The emergence of artificial intelligence (AI) brings the new opportunities for wireless communication. In particular, reinforcement learning, as a technology which aims to maximize the expected long-term rewards, has attracted the attention of scholars. Q-learning is a typical model-independent reinforcement learning algorithm based on value iteration. In Q-learning, the agent makes observations and takes actions within an environment, and in turn, receives rewards. During the learning process, the errors that occurred can be corrected. Therefore, Q-learning is intended to dynamically improve the performance in the process of interacting with the environment. In addition, Q-learning achieves the balance between exploration and exploitation. Therefore, in order to focus on long-term reward in the process of dynamically allocating resources of the system, we investigate the resource allocation problem in the NOMA system by applying Q-learning scheme.

B. RELATED WORKS

At present, the investigation on resource allocation in NOMA communication system has made certain achievements in many aspects, such as user pairing [15], [16], channel assignment, and power allocation [17], [18]. The problem of user association and channel assignment in downlink multi-cell NOMA networks is solved in [19], the authors propose a low-complexity iterative solution to obtain the optimal power allocation, while accounting for inter-user interference and maintaining QoS per user. In [20], authors predefine some power allocation schemes and investigate the user pairing problem in NOMA uplink communication system. They analyze the combinatorial problem of user pairing to achieve the maximum sum rate and propose optimum and sub-optimum algorithms with a polynomial-time complexity. Different from existing works [11], [21], the works in [22] consider a NOMA amplify-and-forward two-way relay wireless network with eavesdroppers. The authors aim to maximize the achievable secrecy energy efficiency by jointly designing the subcarrier assignment, user pair scheduling and power allocation.

On the other hand, machine learning (ML) brings new opportunities for wireless communication. There are many studies on solving the optimization problem through ML [23]–[26]. Reinforcement learning (RL), as a machine learning method, can solve many complex problems by interacting with the system. So far, RL has been applied in many fields such as power allocation and user association [27], [28]. Researchers have proposed many ML-based approaches to solve various problem of NOMA system [29]–[32]. In [24], the authors propose a long short-term memory (LSTM) network which can automatically detect the channel characteristics. In [29], the authors use deep learning (DL) based predictions to accelerate the optimization process in conventional optimization methods for tackling the NOMA resource scheduling problems. In addition, there have been some existing works on the power allocation of NOMA systems. The works in [30] exploit an attention-based neural network to allocate resources to users in a near optimal way. The works in [31] train a deep neural network (DNN) which is used as a predictor to predict the best power allocated to users' signal. The DL based long-term power allocation scheme proposed in [32] can efficiently derive a more accurate decoding order than the conventional solution.

C. CONTRIBUTION

The power allocation problem of downlink NOMA communication systems was discussed in several papers [19], [20]. In addition, there have been several cases of using artificial intelligence to solve resource allocation problem in wireless communication systems. However, most of the current research on resource allocation of the NOMA system based on machine learning is realized through deep learning [29], [31]. To the best of our knowledge, there is currently no research on solving the power allocation problem of multi-antenna NOMA system through cooperative Q-learning algorithm. Our contributions can be summarized as follows:

- We formulate the power allocation in multi-antenna downlink NOMA communication system as a non-convex optimization problem, which aims to maximize the sum rate of the system and considers the constraints of the power allocated to all users, the total power budget of the system, and the QoS for each user.
- By introducing the reinforcement learning, we propose a cooperative Q-learning based algorithm to solve the aforementioned non-convex problem. We define a reward function, which improves the sum rate while meeting the power conditions and capacity constraint. The optimal beamforming matrix can be obtained by cooperatively updating the Q-table in each iteration.
- To evaluate the performance of the proposed method, we verify the proposed cooperative Q-learning based power allocation approach. The simulation results show that the proposed scheme finally converges after some

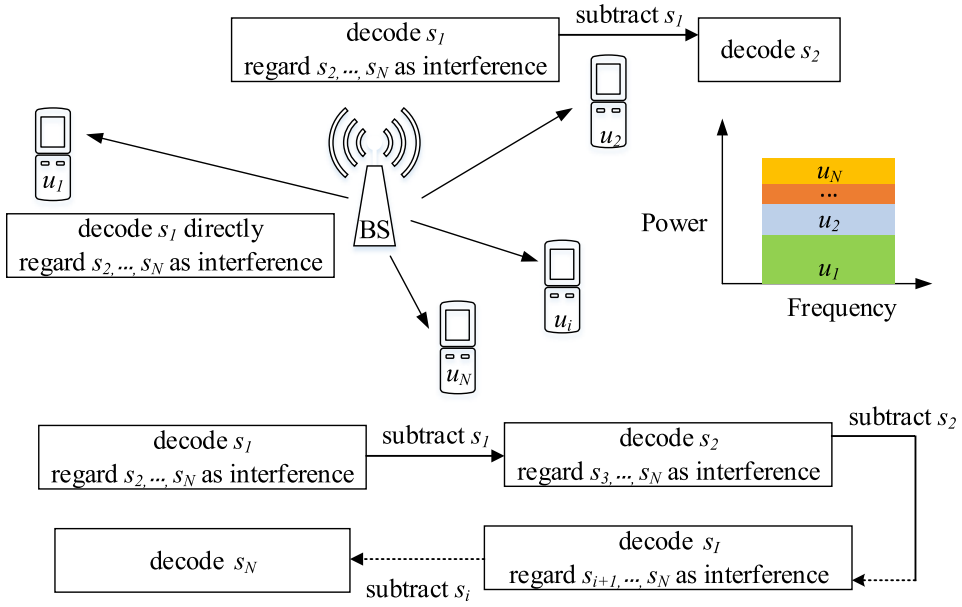

FIGURE 1. An downlink SISO NOMA system with N users.

iterations. Comparing with the conventional power allocation methods, the developed Q-learning based method has better performance under different power budget and different number of users.

D. ORGANIZATION

The rest of this paper is organized as follows. In Section II, we construct the model of the multi-antenna downlink NOMA communication system and formulate the optimization problem. In order to solve the formulated problem, we propose a cooperative Q-learning based solution in Section III. After that, we verify our proposed method and the simulation results are presented in Section IV. Finally, we conclude this paper in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. POWER ALLOCATION FOR NOMA SYSTEMS

In NOMA systems, the signals of different users are multiplexed on the shared wireless channel through power allocation. Then at the decoder, users recover the specific signals via SIC. The key problem is how to allocate limited power to all users to maximize the system performance.

In traditional single-input single-output (SISO) NOMA scenarios, the base station (BS) and users are all equipped with single antenna. According to the principle of NOMA, users with better channel condition will be allocated lower power, while users with worse channel gain will be allocated higher power. At the receiver, SIC is applied to decode all desired signals. To be specific, the signal of the weakest channel user (who is allocated the highest power and regards signals of all the other users as interferences) will be decoded first and subtracted, whereas the signal of the highest channel gain user (who is allocated the lowest power) can be decoded directly.

For example, in a SISO NOMA system with N users shown in Fig. 1, we denote the channel gain of the user as h. Without loss of generality, we assume that channel gains are sorted in a descending order: $|h_1|^2 > |h_2|^2 > \ldots > |h_N|^2$. Then SIC can be used to decode the desired signals. To be specific, u_i needs to decode signals of $u_j (i < j)$ that are allocated with more power and subtracts them from the received signal until it can decode its own signal. During the decoding process, u_i will treat signals of $u_m (m < i)$ that are allocated with less power as interference.

The power allocation problem in a multiple-input single-output (MISO) NOMA system is different from that in a SISO NOMA systems. In MISO NOMA scenarios, the BS is equipped with multiple antennas and users are equipped with single antenna. Unlike that in single-antenna NOMA systems, channel gains in the MISO NOMA systems are jointly determined by multiple antennas between the BS and users.

B. SYSTEM MODEL

In this paper we consider a downlink MISO NOMA communication scenario consisting of one base station and N users, where all the users are randomly distributed at different distances from the BS and equipped with single antenna. The BS is equipped with M antennas and non-orthogonally transmits all the signals over the shared frequency resources.

We denote the user set as $U = \{u_1, u_2, \ldots, u_N\}$. In this scenario, the BS transmits a signal, s_i, for $u_i \in U$ with zero mean and unit variance. All signals are superposed at the BS:

$$\mathbf{x} = \mathbf{w}_1 s_1 + \mathbf{w}_2 s_2 + \ldots + \mathbf{w}_N s_N, \quad (1)$$

where $\mathbf{w}_i \in \mathbb{C}^{M\times 1}$ represents the beamforming vector from the BS to u_i. Then, the received signal at u_i can be formulated as:

$$y_i = \mathbf{h}_i^H \mathbf{x} + n_i. \tag{2}$$

where $\mathbf{h}_i \in \mathbb{C}^{M\times 1}$ represents the channel response from the BS to u_i, and n_i denotes the additive white Gaussian noise (AWGN) with zero mean and variance σ^2. $\mathbf{h}_i$ is characterized by $\mathbf{h}_i = d_i^{-\mu}\mathbf{g}_i$ where d_i is the distance between the BS and u_i, and μ denotes the path loss exponent. Each element in $\mathbf{g}_i$ follows the Rayleigh distribution.

Without loss of generality, we assume that $||\mathbf{h}_1||_2 > ||\mathbf{h}_2||_2 > \ldots > ||\mathbf{h}_N||_2$. Equation (2) can be further written as:

$$y_i = \mathbf{h}_i^H \mathbf{w}_i s_i + \sum_{j=1}^{i-1} \mathbf{h}_j^H \mathbf{w}_j s_j + \sum_{k=i+1}^{n} \mathbf{h}_k^H \mathbf{w}_k s_k + n_i. \tag{3}$$

In order to decode s_i, u_i needs to decode signals of u_k and subtracts them from y_i. When decoding s_i, u_i regards signals of u_j as interference. Then, the signal-to-interference-plus-noise ratio (SINR) of each user is given by

$$\begin{aligned}
\gamma_1 &= \frac{|\mathbf{h}_1^H \mathbf{w}_1|^2}{\sigma^2}, \\
\gamma_2 &= \frac{|\mathbf{h}_2^H \mathbf{w}_2|^2}{|\mathbf{h}_2^H \mathbf{w}_1|^2 + \sigma^2}, \\
&\cdots, \\
\gamma_N &= \frac{|\mathbf{h}_N^H \mathbf{w}_N|^2}{\sum_{i=1}^{N-1} |\mathbf{h}_N^H \mathbf{w}_i|^2 + \sigma^2}.
\end{aligned} \tag{4}$$

The corresponding data rate of u_i can be written as

$$C_i = Blog_2(1 + \frac{|\mathbf{h}_i^H \mathbf{w}_i|^2}{\sum_{j=1}^{i-1} |\mathbf{h}_i^H \mathbf{w}_j|^2 + \sigma^2}), \tag{5}$$

where B denotes the bandwidth of the channel.

C. PROBLEM FORMULATION

Let $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N]$ denotes the beamforming matrix at the BS, the goal of the optimization problem is to maximize the sum rate of all users by finding the optimal $\mathbf{W}$, while meeting the power constraints and QoS for each user. The optimization problem of the NOMA system could be generally formulated as

$$\max_{\mathbf{W}} \sum_{i=1}^{N} Blog_2(1 + \gamma_i), \tag{6a}$$

$$s.t.\ ||\mathbf{w}_i||^2 \le p_{max} \quad \forall i = 1, 2, \ldots, N, \tag{6b}$$

$$\sum_{i=1}^{N} ||\mathbf{w}_i||^2 \le P_{tot}, \tag{6c}$$

$$C_i \ge C_{th} \quad \forall i = 1, 2, \ldots, N. \tag{6d}$$

where p_{max} is the power limitation that the BS can allocate to each user and P_{tot} is the total power constraint. C_{th} represents the minimum required capacity for users.

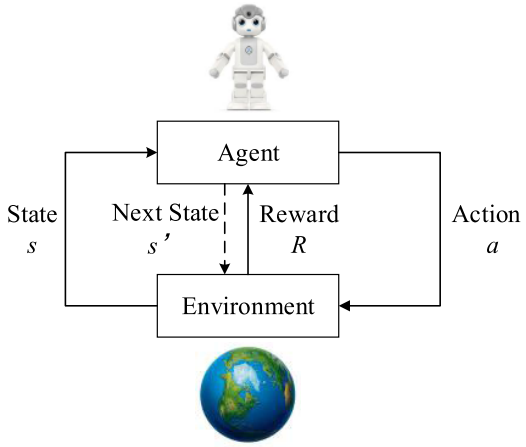

FIGURE 2. Reinforcement learning diagram.

Considering (4), (5) and (6), it can be concluded that the formulated optimization problem is non-convex. Next, we will propose a cooperative Q-learning based approach to solve this problem.

III. A COOPERATIVE Q-LEARNING BASED POWER ALLOCATION SOLUTION

In this section, we first introduce the basics of reinforcement learning. Then, we define state space, action space, and propose a cooperative Q-learning based approach to solve the power allocation problem. Finally, we present the pseudo code and specific steps of the algorithm.

A. BASICS OF REINFORCEMENT LEARNING

Reinforcement learning is a process in which a single or multiple agents take actions in the process of interacting with the environment to obtain rewards and change their state. It can be solved by policy iteration and value iteration. As shown in Fig. 2, in this reinforcement learning procedure, the agent observes the environment and its own state s, then decides the action, a, to be taken. After that, the agent will receive a feedback (reward R) from the environment, and transit to the next state s'. The goal of this reinforcement learning approach is to maximize the cumulative discounted rewards during the interaction.

Q-learning is a reinforcement learning algorithm based on value iteration. It is mainly composed of states, actions, and rewards. Specifically, in Q-learning, each agent creates and maintains a Q-table. The rows of the table represent states, and the columns represent actions. The values, named Q-value, represent different expected future rewards at different states and different actions. Using the Bellman equation, we can calculate Q-value as follows

$$Q(s, a) = E[R_t + \gamma Q(s_{t+1}, a_{t+1} | s_t = s, a_t = a)], \tag{7}$$

where s_t and s_{t+1} represent the states at time step t and $t+1$, respectively. a_t and a_{t+1} represent the actions at time step t and $t+1$. E denotes the expectation operator and R_t is the

received reward at time step t. γ is the discount factor, which determines how much the agent cares about rewards in the distant future. When γ is set to 0, only the current reward is considered. The greater the discount factor is, the more important future rewards are.

The Q-table's updating rule is given by

$$Q(s,a) = Q(s,a) + \alpha(R_t + \gamma \underbrace{maxQ(s_{t+1}, a_{t+1})}_{(a)} - \underbrace{Q(s,a)}_{(b)}) \tag{8}$$

where α is the learning rate, which defines the proportion of newly learned information to the current Q-value. A value of 0 means that the agent will not learn anything, i.e., old information is important, and a value of 1 means that the newly discovered information is the only important information. Terms (a) and (b) are optimal future Q-value and current Q-value, respectively.

B. PROPOSED SOLUTION

In the Q-learning algorithm, the agent interacts with the environment by sensing states and taking actions. The environment includes everything in the NOMA system except the agent. In the context of the multi-antenna NOMA system mentioned before, the solution of the formulated power allocation problem can be equivalent to a cooperative reinforcement learning process where the BS generates multiple Q-tables for antennas. Multiple antennas cooperatively update their Q-tables while interacting with the environment.

The Q-learning approach consists of three main parts: states, actions, and reward. To solve the power allocation problem, we define state space, action space, and reward function for the cooperative reinforcement learning process as follows:

1) State space: The agent observes the environment and senses the states. All possible states form the state space, denoted as S, which is characterized by distances from users to the BS. In this paper, the state space S contains N states, i.e., $S = \{s_1, s_2, \ldots, s_N\}$.
We set the BS in center of the cell and its position (x_{BS}, y_{BS}) can be denoted as (0,0). Therefore, distances from users to the BS are only determined by the coordinates of users. We represent the coordinate of u_i as $s_i = (x_i, y_i)$. Then, the state space S can be written as $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$.
2) Action space: All the actions can be taken by the agent form the action space, denoted as A. In this paper, we discretize the power into different power levels. Different actions in the action space represent different power levels that the BS can allocate to users at each antenna. Denote the number of power levels as K, then $A = \{a_1, a_2, \ldots, a_K\}$. The first element a_1 represents the minimum power level, while the last element a_K denotes the maximum power level that can be allocated to users at each antenna. To be specific, the action space can be calculated as $A = \{p_{min}, p_{min} + \frac{p_{max}-p_{min}}{K-1}, \ldots, p_{min} + \frac{(K-2)(p_{max}-p_{min})}{K-1}\}$.
In this paper, we set the number of power levels as $K = 30$ and $p_{min} = 0.001W$. Each element $a_i \in A$ represents the power level allocated by the BS at each antenna. Since there is no other detailed information about the environment, the agent takes action form A with equal probability.
3) Reward function: For reinforcement learning, reward function plays an important role. The definition of the reward function should ensure that the agent gets the appropriate reward when it takes good actions and pays the proper penalty when it takes bad actions. Good actions here refer to actions that can help achieve the objective function. On the contrary, bad actions refer to actions that make the objective function more difficult to achieve. Hence, we take the following points into consideration.
   - For the formulated optimization problem, we want to maximize the sum rate of the NOMA system. Therefore, a higher sum rate results in a higher reward. In addition, higher data rate for each user helps to increase the sum rate of the system and the reward.
   - In order to satisfy the QoS constraint of each user, the data rate that meets the required threshold helps to obtain the reward, while the data rate that does not meet the requirement results in a negative reward.
   - When the transmission power meets the total power constraint, the system will be rewarded, and when the transmission power exceeds the power budget, it will cause a negative reward.

By considering the above points, we define the reward function for u_i at time step t as follows:

$$R_t^i = \underbrace{C_{i,t}C_{sum,t}^2}_{(a)} + \underbrace{(C_{i,t} - C_{th})}_{(b)} - \underbrace{(P_{tr,t} - P_{tot})}_{(c)} \tag{9}$$

where $C_{i,t}$ and $C_{sum,t}$ are the data rate of u_i and the sum rate of the system at time step t, respectively. $P_{tr,t}$ is the total transmission power at time step t.
According to (9), we can find that the reward function consists of three terms. Term (a) means that both higher data rate of each user and higher sum rate of the system result in higher reward. In the same term, the sum rate is squared. Term (b) and term (c) guarantee the capacity requirement for each user and power budget for the system, respectively. The data rate which does not meet the requirement will decrease the reward. Moreover, when the transmission power at the current time step exceeds the power budget, the reward at the current time step will be decreased.

C. ALGORITHM DESCRIPTION

In Q-learning, if the number of iterations is not limited, the agent will be able to visit all the state-action pairs. Hence, the Q-table will be updated until the Q-value finally converges to the optimal value with probability 1. However, the Q-learning algorithm always selects the action which derives the maximum Q-value on each iteration, which will cause the agent to be trapped in a limited search area and the algorithm to converge slower. In addition, the number of iterations is always limited in practice, the final Q-value may not be optimal.

In order to solve the above problem, ϵ-greedy algorithm is used in [34] to accelerate the learning process. To be specific, the agent takes action randomly with probability ϵ (known as exploration) and takes the action corresponding to the maximum Q-value (known as exploitation) with probability 1-ϵ. As such, the probability of selection of the action a_m at state s_m is given by

$$\pi_m(s_m, a_m) = \begin{cases} 1-\epsilon, & \text{if } Q_m \text{ of } a_m \text{ is the highest,} \\ \epsilon, & \text{otherwise.} \end{cases} \tag{10}$$

The search area can be controlled by adjusting the value of ϵ, so as to realize the trade-off between exploration and exploitation. It is shown in [34] that the ϵ-greedy algorithm can accelerate the learning process and has a faster convergence rate. Therefore, we adopt ϵ-greedy algorithm in the rest of this paper. The pseudo code of the proposed cooperative Q-learning based power allocation method in downlink NOMA system is shown in **Algorithm 1**.

We first set the main parameters, such as learning rate α, discount factor γ, and ϵ for ϵ-greedy algorithm. The capacity constraint for users and power budget of the system are also preset. In the initialization phase, we define the state space and action space, create Q-tables for multiple antennas and initialize them to 0. The detailed description of the **Algorithm 1** is summarized as follows:

At the beginning of the iteration, the base station observes the current environment state and selects the power level to be allocated to each user at each antenna from the action space according to (10). Next, after the power allocation, the users calculate their capacity and feed them back to the base station. Then, the base station receives the reward for all Q-tables from the environment according to (9). The message that the BS receives can be shared between multiple antennas. All the Q-tables are cooperatively updated according to (8) and transit to the next state. The above process is repeated until the error is below the threshold ζ or the current episodes reach the preset number.

Algorithm 1 Cooperative Q-Learning Based Power Allocation Algorithm of the Downlink NOMA System

Input:
learning rate α;
discount factor γ;
ϵ-greedy parameter ϵ;
capacity threshold C_{th};
power budget P_{tot}.
Output:
the optimal beamforming matrix $\mathbf{W}$.
Initialization:
create Q-tables for each antenna;
set the total iterations I_{tot} and error threshold ζ;
define state space and action space.
while *error* $\geq \zeta$ *and episodes* $\leq I_{tot}$ **do**
  initial state s;
  for *all steps of episode* **do**
    for *all Q-tables* **do**
      choose a from the action space based on ϵ-greedy algorithm;
    end
    perform action a and calculate the sum rate C_{sum} of the system;
    for *all Q-tables* **do**
      measure the reward R and new state s';
      update $Q(s, a)$ according to the updating rule;
      let $s = s'$;
    end
  end
end

IV. SIMULATION RESULTS

In this section we give the simulation set up in detail and evaluate the performance of the proposed Q-learning based power allocation method (QPA). We use the NOMA system with random power allocation algorithm (NOMA random), orthogonal frequency division multiple acces (OFDMA) and the SRMax method (SRPA) proposed in [33] as benchmarks. Different from NOMA system in which all users share bandwidth B, in OFDMA, each user occupies a bandwidth B/N, where N is the number of users.

To have a comprehensive comparison between these four algorithms, we use three metrics: the sum rate (SR), the energy efficiency (EE) and the spectral efficiency (SE). The EE is defined as the ratio of the sum rate and the total power consumption of the system, including circuit power consumption and transmission power. The SE is defined as the ratio of the sum rate to the occupied bandwidth.

In simulations, the positions of the users are distributed based on the uniform distribution and in a circle with the center at the base station. The maximum distance from the user to the base station is 300m, and the minimum distance is 50m. As in [7], [35], we set the total power budget of the BS and the circuit power consumption as 41dBm and 30dBm, respectively. The capacity requirement for users is 1b/s/Hz [34], and the noise power spectral density is set to -120dBm. The bandwidth is 5MHz and the number of power level is 30. Unless otherwise stated, the BS is equipped with two antennas. As for Q-learning, the learning rate α is 0.5 and

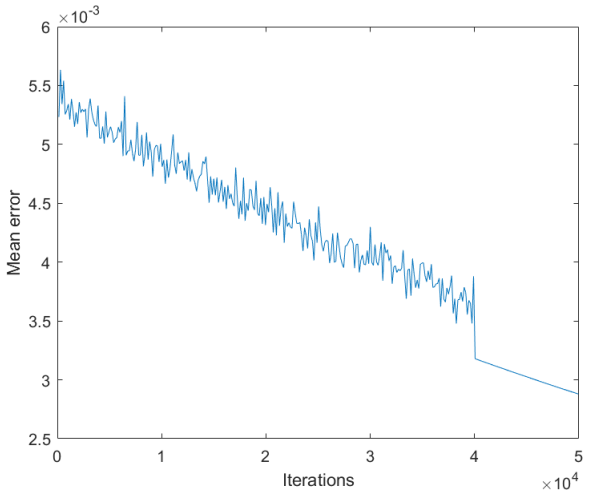

FIGURE 3. Mean error versus the number of iterations.

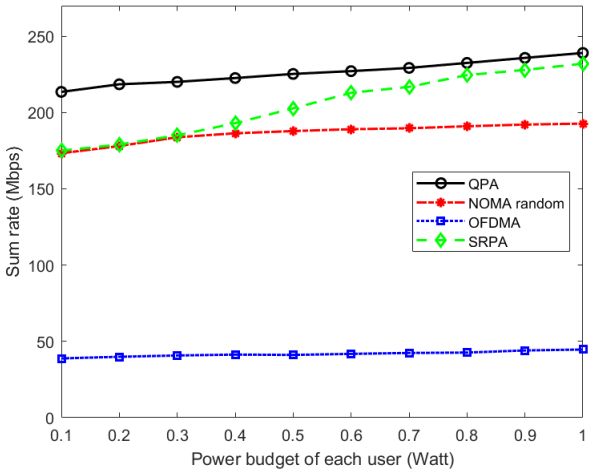

FIGURE 4. Sum rate versus power budget of each user.

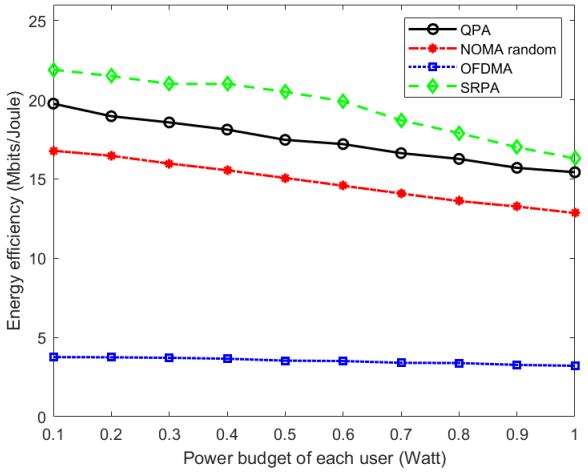

FIGURE 5. Energy efficiency versus power budget of each user.

the discount factor γ is 0.9. We adopt the ϵ-greedy algorithm to accelerate the learning and set the ϵ as 0.1.

A. CONVERGENCE ANALYSIS

We analyze the convergence of the proposed Q-learning based power allocation method. In our simulations, we set the maximum number of iterations as 50,000 and run the code 200 times. Fig. 3 demonstrates the average error versus the number of iterations when the learning rate is 0.1. As it is shown in Fig. 3, we find that the proposed algorithm requires about 40,000 iterations. The initial mean error is large. As the number of iterations increases, the mean error keeps fluctuating and gradually decreases until it eventually becomes zeros. The algorithm always converges at about 40,000 iterations.

B. SR AND EE AGAINST POWER BUDGET

We study the SR and EE performance of the four power allocation methods against different power budgets of each user. The number of users is set as 5 in this scenario. The simulation results are presented as follows:

Fig. 4 depicts the SR of the proposed Q-learning based power allocation algorithm, the random power allocation, the OFDMA system, and the SRPA method for different power budgets of each user. From Fig. 4 we can see that NOMA schemes (QPA, SRPA, and NOMA random) always outperform OFDMA. Specifically, the SR of the proposed QPA algorithm is the best, followed by the SRPA scheme. The SRPA and the NOMA random algorithms have nearly the same SR performance when the power budget is small. As the power budget increases, the SR of the QPA scheme increases slowly, while the SR of the SRPA algorithm increases rapidly until it approaches the SR of the QPA method.

As it is shown in Fig. 5, the EE performance of the three NOMA algorithms are significantly better than that of the OFDMA method. Overall, the EE of the four algorithms decreases slowly as the power budget increases. The EE performance of the SRPA algorithm is better than that of the QPA method, but the gap between the two schemes gradually decreases as the power increases.

C. EE AND SE AGAINST NUMBER OF USERS

In this section, we evaluate the EE and SE performance of the four power allocation schemes under different number of users. The power budget of each user is set as 1 Watt in this scenario. The simulation results are presented as follows:

Fig. 6 displays the EE performance versus the number of users. It can be seen from Fig. 6 that although there are small fluctuations, the EE performance of the QPA algorithm is still better than that of the NOMA random and OFDMA methods. As the number of users increases, the energy efficiency of the QPA method gradually decreases while the energy efficiency of the SRPA scheme increases slowly. When the number of users is 2, the EE of the QPA scheme is much higher than that of the SRPA method, and when the number of users is larger than 12, the EE of the SRPA outperforms QPA.

Fig. 7 shows the SE versus the number of users. From Fig. 7 we can see that the SE of QPA scheme is significantly

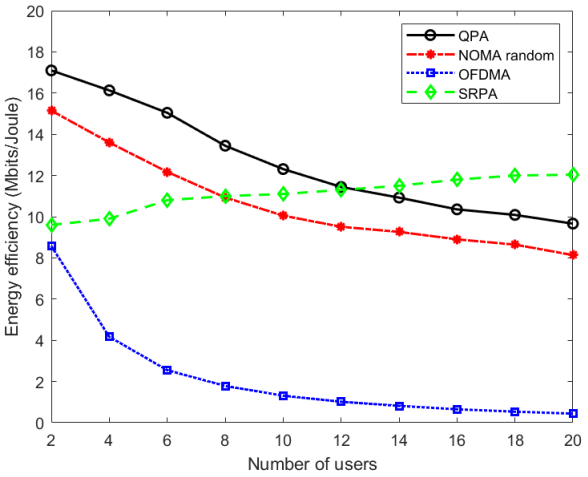

FIGURE 6. Energy efficiency versus number of users.

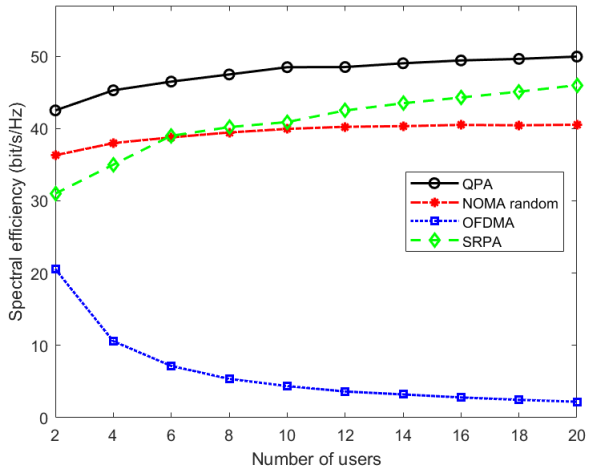

FIGURE 7. Spectral efficiency versus number of users.

higher than that of other algorithms. As the number of users increases, the SE of the three NOMA schemes gradually increases while the SE of the OFDMA decreases. At the beginning, the SE of the SRPA method is lower than that of the NOMA random algorithm. With increasing the number of users, the SE of the SRPA increases rapidly, and exceeds that of NOMA random method when the number of users is 8.

The simulation results show that, compared to OMA, NOMA can significantly improve the spectral efficiency and the proposed method has better performance in terms of SR and EE. We can infer that as the number of users and power budget increase, the proposed Q-learning based power allocation method will yield excellent performance.

Finally, Fig. 8 illustrates the impact of the number of antennas on the sum rate of the proposed algorithm. On the whole, as the value of M increases, the sum rate increases, which indicates that more antennas at the base station can bring better performance. We observe that when the number of users is small, there is little difference in SR performance corresponding to different M. As the number of

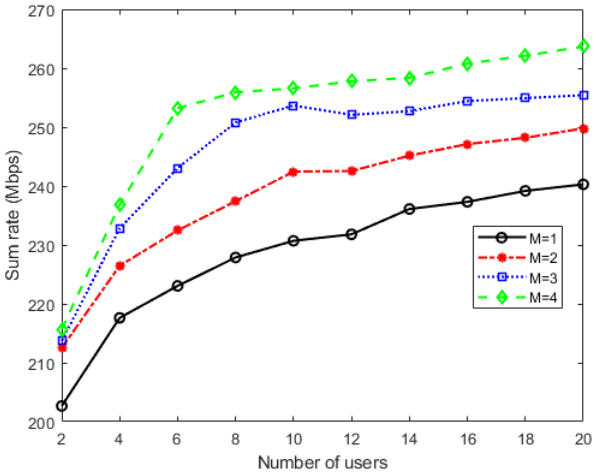

FIGURE 8. Sum rate versus number of users.

users increases, the impact of the number of antennas on SR performance gradually increases. Moreover, as the number of users increases, the sum rate increases rapidly at first, and tends to flatten when the number of users is 12.

D. COMPUTATIONAL COMPLEXITY ANALYSIS

We know that the complexity of the reinforcement learning algorithm mainly depends on the state space size and action space size. According to [36], we can estimate that the computational complexity of the Q-learning algorithm with the ϵ-greedy is $\mathcal{O}(SAH)$ per iteration, where S is the number of states, A is the number of actions, and H is the number of steps per episode. According to the state space and action space defined before, the amount of work per iteration is $\mathcal{O}(MN^2K)$, which mainly increases with the square of the number of users, while the computational complexity of the scheme proposed in [33] is proportional to the third power of the number of users.

Combined with the performance analysis results shown earlier, we find that our proposed method outperforms the traditional optimization technique in terms of sum rate and spectral efficiency. In addition, when the number of users is large, our proposed method has advantages in terms of computational complexity. It is worth noting that unlike deep learning algorithm such as convolutional neural network, which is a process of learning from a training data set and then applying that to a new data set, QPA is able to adjust to the changes in the network based on the feedback from the environment.

E. LIMITATION OF THE PROPOSED SCHEME

As shown in Part B and C of Section IV, in terms of SR, EE and SE, the overall performance of the proposed Q-learning based power allocation method is better than that of the NOMA system with random power allocation algorithm, as well as the performance of the OFDMA system. In addition, the SR and SE performance of the QPA are better than that of EEPA. However, the proposed approach does

not consider the fairness of the system. There may be some extreme cases where the data rates of some users are too high while the data rates of other users are too low. Therefore, the proposed approach will fail in scenarios requiring fairness among users.

V. CONCLUSION

In this paper, we apply a reinforcement learning algorithm, Q-learning, to deal with the power allocation problem in downlink MISO NOMA communication system. By interacting with the environment, the agent, that is, the base station, intelligently allocates different power levels to users by multiple antennas, and iterates until the optimal performance is obtained. The simulation results show that the Q-learning based algorithm achieves better performance than NOMA random and OFDMA algorithms in terms of sum rate, energy efficiency and spectral efficiency under different power budgets and different number of users. In addition, the proposed QPA method is better than SRPA scheme in terms of sum rate, spectral efficiency and computational complexity. Future work will aim to solve the user clustering and power allocation problems and improve the fairness of the network as much as possible while obtaining better performance. In addition, probabilistic methods could be applied to provide uncertainty estimations for the resource allocation problems.

REFERENCES

[1] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018, doi: 10.1109/COMST.2018.2835558.

[2] L. Zhang, M. Xiao, G. Wu, M. Alam, Y.-C. Liang, and S. Li, "A survey of advanced techniques for spectrum sharing in 5G networks," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 44–51, Oct. 2017, doi: 10.1109/MWC.2017.1700069.

[3] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016, doi: 10.1109/COMST.2016.2532458.

[4] M. Elbayoumi, M. Kamel, W. Hamouda, and A. Youssef, "NOMA-assisted machine-type communications in UDN: State-of-the-art and challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1276–1304, 2nd Quart., 2020, doi: 10.1109/COMST.2020.2977845.

[5] S. M. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 40–47, Apr. 2018, doi: 10.1109/MWC.2018.1700099.

[6] H. Zhang, F. Fang, J. Cheng, K. Long, W. Wang, and V. C. M. Leung, "Energy-efficient resource allocation in NOMA heterogeneous networks," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 48–53, Apr. 2018, doi: 10.1109/MWC.2018.1700074.

[7] M. T. P. Le, G. C. Ferrante, G. Caso, L. D. Nardis, and M. D. Benedetto, "On information-theoretic limits of code-domain NOMA for 5G," *IET Commun.*, vol. 12, no. 15, pp. 1864–1871, Sep. 2018, doi: 10.1049/iet-com.2018.5241.

[8] S. Tang, Z. Ma, M. Xiao, and L. Hao, "Hybrid transceiver design for beamspace MIMO-NOMA in code-domain for mmWave communication using lens antenna array," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2118–2127, Sep. 2020, doi: 10.1109/JSAC.2020.3000885.

[9] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017, doi: 10.1109/JSAC.2017.2725618.

[10] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017, doi: 10.1109/COMST.2016.2621116.

[11] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016, doi: 10.1109/ACCESS.2016.2604821.

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

[13] F. A. Rabee, K. Davaslioglu, and R. Gitlin, "The optimum received power levels of uplink non-orthogonal multiple access (NOMA) signals," in *Proc. IEEE 18th Wireless Microw. Technol. Conf. (WAMICON)*, Cocoa Beach, FL, USA, Apr. 2017, pp. 1–4.

[14] P. Wang, J. Xiao, and L. Ping, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Veh. Technol. Mag.*, vol. 1, no. 3, pp. 4–11, Sep. 2006, doi: 10.1109/MVT.2006.307294.

[15] M. B. Shahab, M. F. Kader, and S. Y. Shin, "A virtual user pairing scheme to optimally utilize the spectrum of unpaired users in non-orthogonal multiple access," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1766–1770, Dec. 2016, doi: 10.1109/LSP.2016.2619371.

[16] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016, doi: 10.1109/TWC.2016.2606100.

[17] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, Dec. 2016, doi: 10.1109/TWC.2016.2616310.

[18] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster," *IEEE Access*, vol. 6, pp. 5170–5181, Feb. 2018, doi: 10.1109/ACCESS.2017.2779855.

[19] M. W. Baidas, Z. Bahbahani, and E. Alsusa, "User association and channel assignment in downlink multi-cell NOMA networks: A matching-theoretic approach," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–21, Sep. 2019, doi: 10.1186/s13638-019-1528-8.

[20] M. A. Sedaghat and R. R. Muller, "On user pairing in uplink NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3474–3486, May 2018, doi: 10.1109/TWC.2018.2815005.

[21] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. M. Leung, "Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2874–2885, Dec. 2017, doi: 10.1109/JSAC.2017.2777672.

[22] H. Zhang, N. Yang, K. Long, M. Pan, G. K. Karagiannidis, and V. C. M. Leung, "Secure communications in NOMA system: Subcarrier assignment and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1441–1452, Jul. 2018, doi: 10.1109/JSAC.2018.2825559.

[23] J. Zhou, X. Liu, and C. Huang, "Machine learning for power allocation of a D2D network," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[24] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018, doi: 10.1109/TVT.2018.2848294.

[25] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019, doi: 10.1109/TWC.2018.2879433.

[26] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in Millimeter-Wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7425–7440, Nov. 2018, doi: 10.1109/TWC.2018.2867180.

[27] R. Amiri, M. A. Almasi, J. G. Andrews, and H. Mehrpouyan, "Reinforcement learning for self organization and power control of two-tier heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3933–3947, Aug. 2019, doi: 10.1109/TWC.2019.2919611.

[28] D. Li, H. Zhang, K. Long, W. Huangfu, J. Dong, and A. Nallanathan, "User association and power allocation based on Q-learning in ultra dense heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–5.

[29] L. Lei, L. You, Q. He, T. X. Vu, S. Chatzinotas, D. Yuan, and B. Ottersten, "Learning-assisted optimization for energy-efficient scheduling in deadline-aware NOMA systems," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 3, pp. 615–627, Sep. 2019, doi: 10.1109/TGCN.2019.2902838.

[30] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019, doi: 10.1109/JSAC.2019.2933762.

[31] K. N. Doan, M. Vaezi, W. Shin, H. V. Poor, H. Shin, and T. Q. S. Quek, "Power allocation in cache-aided NOMA systems: Optimization and deep reinforcement learning approaches," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 630–644, Jan. 2020, doi: 10.1109/TCOMM.2019.2947418.

[32] Y. Sun, Y. Wang, J. Jiao, S. Wu, and Q. Zhang, "Deep learning-based long-term power allocation scheme for NOMA downlink system in S-IoT," *IEEE Access*, vol. 7, pp. 86288–86296, 2019, doi: 10.1109/ACCESS.2019.2926426.

[33] P. Liu, Y. Li, W. Cheng, W. Zhang, and X. Gao, "Energy-efficient power allocation for millimeter wave beamspace MIMO-NOMA systems," *IEEE Access*, vol. 7, pp. 114582–114592, 2019, doi: 10.1109/access.2019.2935495.

[34] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan, and D. Matolak, "A machine learning approach for power allocation in HetNets considering QoS," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–7.

[35] D. Ni, L. Hao, Q. T. Tran, and X. Qian, "Transmit power minimization for downlink multi-cell multi-carrier NOMA networks," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2459–2462, Dec. 2018, doi: 10.1109/LCOMM.2018.2872991.

[36] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. Jordan, "Is Q-learning provably efficient?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31. Red Hook, NY, USA: Curran Associates, 2018, pp. 4863–4873.

QI ZHAI received the B.S. degree in communication engineering from Harbin Engineering University, Harbin, China, in 2016, and the M.S. degree in communication and information system from Northwestern Polytechnical University, Xi'an, China, where she is currently pursuing the Ph.D. degree in information and communication engineering. Her current research interests include resource allocation of wireless communication networks, non-orthogonal multiple access communication, and physical layer security.

MIODRAG BOLIĆ received the B.S. and M.S. degrees in electrical engineering from the University of Belgrade, Belgrade, Yugoslavia, in 1996 and 2001, respectively, and the Ph.D. degree in electrical engineering from Stony Brook University, NY, USA, in 2004. Since 2004, he has been an Assistant Professor with the School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada, and was promoted to an Associate Professor, in 2009. Since 2018, he has been a Professor with the University of Ottawa. His research interests include computer architecture, signal processing, wireless communications, and biomedical engineering.

YONG LI received the B.S. degree in avionics engineering, the M.S. and Ph.D. degrees in Circuits and Systems form Northwestern Polytechnical University, Xi'an, China, in 1983, 1988, and 2005, respectively. In 1993, he joined the School of Electronic Information, Northwestern Polytechnical University and was promoted to a professor, in 2002. His research interests include digital signal processing and radar signal processing.

WEI CHENG received the B.S., M.S., and Ph.D. degrees in communication and information system from Northwestern Polytechnical University, Xi'an, China, in 2003, 2006, and 2011, respectively. From April 2011 to April 2013, he worked in the postdoctoral research station with the School of Electronic Information, Northwestern Polytechnical University. Since 2013, he has been a Lecturer with the School of Electronic Information, Northwestern Polytechnical University and was promoted to an Associate Professor, in 2015. His research interests include wireless sensor networks and Ad Hoc networks, and radar signal processing.

CHENXI LIU received the B.S. degree in electronic information engineering and the M.S. degrees in communication and information system from Northwestern Polytechnical University, Xi'an, China, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include physical layer security, machine learning, and wireless communications.