Energy Efficiency Optimization in Downlink NOMA-Enabled Fog Radio Access Network Based on Deep Reinforcement Learning

Feng Wu, Li Zhang, Yejun He*
Guangdong Engineering Research Center of Base Station Antennas and Propagation
Shenzhen Key Laboratory of Antennas and Propagation
College of Electronics and Information Engineering, Shenzhen University, 518060, China
Email:1928740956@qq.com, wzhangli@szu.edu.cn, heyejun@126.com*

2021 Computing, Communications and IoT Applications (ComComAp) | 978-1-6654-2798-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/COMCOMAP53641.2021.9652884

***Abstract*—In this era of Internet of Everything (IoE) and rapid development of mobile wireless communication technology, time delay and spectrum efficiency are difficult to achieve in the development process. Fog computing (FC) is a kind of edge computing, which is more suitable for the Internet of Things (IoT) with high dense connectivity. Non-orthogonal multiple access (NOMA), as a promising multiple access technology combined with FC, is considered here. In this paper, we study how to adopt deep reinforcement learning (DRL) algorithms to optimize energy efficiency (EE) in the multi-server and multi-user scenario of downlink fog radio access network (F-RAN) based on NOMA. In order to solve this nonconvex problem, it is divided into two subproblems: subchannel allocation and power allocation. The former adopts deep Q network (DQN) and the latter adopts deep deterministic policy gradient (DDPG) to obtain the best allocation strategy. DRL has high applicability in dealing with high-dimensional data, so it can be well applied in dynamic communication environment. The simulation results show that the combination scheme of "DQN-DDPG" has achieved remarkable advantages in terms of the faster convergence speed and the better final result of the system EE than other schemes.**

***Index Terms*—Internet of Things (IoT), fog computing (FC), non-orthogonal multiple access (NOMA), deep reinforcement learning (DRL), Energy Efficiency (EE)**

I. Introduction

With the development of the fifth generation (5G) mobile communication technology, the era of Internet of Everything (IoE) has emerged as an outstanding tool for spectrum efficiency. The large-scale application of smart devices has generated huge traffic. However, under the traditional cloud computing architecture, the long-distance connection and frequent data transmission between users and the cloud cause traffic congestion under the condition of limited spectrum resources, which can not satisfy the users demand for ultra-low delay and bring bad experience to users. In this case, it is particularly necessary to deploy the server closer to the network edge of users. Fog computing (FC) is more suitable for the Internet of Things (IoT) in edge computing. Users under the fog radio access network (F-RAN) architecture no longer need to send their requests to the remote cloud, but only need to choose the fog access point (F-AP) closer to themselves. This not only alleviates the congestion of cloud servers, but also solves the problem of time delay.

A unified multi-layer cost model in a multi-layer FC network with one fog control node (FCN), multiple fog access nodes (FANs), and user equipments (UEs) is studied in [1], including service delay, linear inverse demand dynamic payment scheme, and the resulting cost minimization user scheduling problem. The author focused on studying user scheduling and neglected the allocation of files. Y. Xiao *et al.* proposed a new distributed optimization framework for collaborative fog computing based on dual decomposition [2]. A generalized Nash Equilibrium problem, which is used to solve the problem that the solution space is too large due to complex task decomposition, is called parallel offloading problem of separable tasks [3].

The key idea of non-orthogonal multiple access (NOMA) technology is to conduct multiple access through power domain, which means that different users can obtain different power levels and transmit data in the same frequency, time, and code domain [4]. X. Wen *et al.* [5] focuses on solving the problem of computing offloading under the integrated fog-cloud network architecture, adopts a quasi-static scenario and puts forward a Stackelberg game model with macro remote radio head (MRRH) as the leader and small remote radio heads (SRRHs) as the followers. The main research is the optimization of energy efficiency (EE) in the NOMA-based fog hierarchy network in the downlink. R. Rai *et al.* proposed a fog-cloud structure that enables NOMA in F-RAN system to solve different aspects of high-density and low-density areas [6].

As a branch of machine learning (ML), reinforcement learning (RL) has high applicability in solving optimization problems. Because of the variability of communication environment, deep reinforcement learning (DRL) combined with deep learning (DL) is widely used. In [7], DRL is applied to the grant-free NOMA system to reduce the interference by obtaining the competition status of each user in the system and allocating the subchannel and power, thereby improving the system throughput in unknown communication environment. The resource allocation problem of multi-user in uplink NO-

MA system is studied in [8], and three resource optimization frameworks based on discrete DRL, continuous DRL, and joint DRL are proposed, respectively. The results show that all of the frameworks can obviously improve the EE of the system. Most systems based on NOMA use DRL algorithm to solve the situation of single server and multiple users, such as [7]-[9].

In this paper, we study the environment of multiple F-APs and multiple users in NOMA-based downlink F-RAN. Users' downloading requirements can be satisfied and the system EE can be optimized through reasonable resource allocation. On the issue of resource allocation, deep Q network (DQN) is used to allocate subchannels and deep deterministic policy gradient (DDPG) is used to allocate power. The main contributions of this paper are as follows.

1) Under the system model of F-RAN, we solve the scarcity of spectrum resources based on NOMA to satisfy the requirements of users in a better way. Moreover, the resource allocation problem in the situation of multi-server and multi-user are considered.

2) Since the problem of obtaining the optimal EE of the system is nonconvex, we adopt DRL to obtain the resource allocation strategy in the multi-server and multi-user system environment to avoid the traditional tedious formula derivation for solving the nonconvex problem. Afterwards, DQN and DDPG are considered to obtain the subchannel allocation and power allocation strategies, respectively.

3) Simulation results show that the "DQN+DDPG" scheme can achieve better results no matter whether the channel transformation conditions are severe or not than other schemes. And it has great applicability in our system environment.

The residual of this paper is as follows. The system model of downlink F-RAN based on NOMA is proposed in section II. Section III is the description of the problem and the corresponding resource allocation scheme of joint DQN and DDPG algorithm. Section IV is a more intuitive analysis of simulation results. The last section is a summary of all the contents.

II. System Model

The communication environment considered in this paper is the downlink F-RAN based on NOMA, as shown in Fig. 1. $\mathcal{N} = \{1, 2, ..., N\}$ is used to represent the set of N F-APs, and the total set of M users is represented by $\mathcal{M} = \{1, 2, ..., M\}$. For the sake of simplicity, it is assumed that both the F-AP and the UE only have a single antenna [5], [6]. And the number of selectable subchannels is Q, which is represented by set $\mathcal{Q} = \{1, 2, ..., Q\}$, B represents the total bandwidth of the system. Because it is divided into Q orthogonal subchannels, the bandwidth of each subchannel should be B/Q. Each F-AP has a certain service range with a radius of r. Users who exceed the service range will not realize their own services, so they must reselect F-APs within the service range. And the served users can occupy all subchannels, it is also assumed that the users served by each F-AP are fixed. However, each user can only occupy one

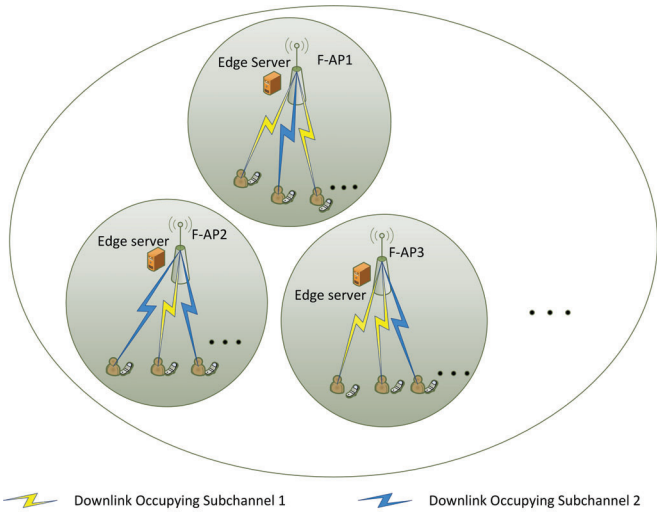

Fig. 1. The system model of NOMA-based downlink F-RAN.

subchannel in each time slot. Therefore, the set of users served under F-AP n is $\mathcal{K}_n = \{1, 2, ..., K_n\}$. The power allocated for the signal required by user n under F-AP m is Pn, m, and cannot exceed the maximum power p_{max}. We assume that the value of p_{max} is the same for each user.

Unlike previous multiple access technologies, NOMA uses non-orthogonal power domain to distinguish users. The non-orthogonal means that the data between users can be transmitted in the same time slot and the same frequency. And NOMA only rely on different power to distinguish users. The development of superposition coding at the transmitting terminal and successive interference cancellation (SIC) at the receiving terminal has greatly improved the practicability of NOMA. In our environment, at each time slot t, the signals of all users occupying the same channel under each F-AP are superimposed and transmitted to the users. And the signal after superimposing all users occupying subchannel q under F-AP n is given by

$$SS_n^q(t) = \sum_{m=1}^{K_n} b_{n,m}^q(t)\sqrt{\lambda_{n,m}(t)p_f}x_{n,m}(t) \qquad (1)$$

where $\lambda_{n,m}(t)$ represents the proportion of power allocated for the signal required by user m under F-AP n. It is assumed that the total distributable power of each F-AP is consistent, which is represented by p_f. And $x_{n,m}(t)$ shows the symbols transmitted between the F-AP n and the served user m. In addition, $b_{n,m}^q(t)$ represents the subchannel allocation coefficient, where $b_{n,m}^q(t) = 1$ means that the user m serving under the F-AP n occupies the subchannel q, and $b_{n,m}^q(t) = 0$ otherwise.

Then, it is assumed that the receivers can obtain instantaneous channel gain at the beginning of each time slot, and the channel gain in the same time slot remains unchanged [8], [10]. The signal received by the user m occupying subchannel

q from the served F-AP n is

$$RS^q_{n,m}(t) = h^q_{n,m}(t)\sum_{r=1}^{K_n} b^q_{n,r}(t)\sqrt{\lambda_{n,r}(t)p_f}x_{n,r}(t) + n^q_{n,m}(t) \tag{2}$$

where $n^q_{n,m}(t)$ represents additive white Gaussian noise with variance σ^2.

SIC technology can be used to decode data packets coming to the receivers [11]. Generally speaking, in downlink communication system, the signal with worst channel quality is decoded first, if it is not the signal desired by the user, subtract it and start decoding the worst channel quality of the remaining signal until the user obtains the desired signal. Therefore, the signal-to-interference-noise ratio (SINR) of the user m served under the F-AP n can be expressed as

$$SINR_{n,m}(t) = \frac{\lambda_{n,m}(t)p_f\left|h^q_{n,m}(t)\right|^2}{I^\theta_{n,m}(t) + I^\beta_{n,m}(t) + \sigma^2} \tag{3}$$

where $I^\theta_{n,m}(t)$ represents the interference of other users occupying subchannel q under the same F-AP n, and $I^\beta_{n,m}(t)$ means the interference of users occupying subchannel q under different F-APs. In addition,

$$I^\theta_{n,m}(t) = \sum_{u=1,|h^q_{n,u}(t)|^2>|h^q_{n,m}(t)|^2}^{K_n} b^q_{n,u}(t)\lambda_{n,u}(t)p_f\left|h^q_{n,m}(t)\right|^2 \tag{4}$$

and

$$I^\beta_{n,m}(t) = \sum_{i\neq n, i\in\mathcal{N}}\sum_{j=1}^{K_i}\lambda_{i,j}(t)p_f b^q_{i,j}(t)\left|h^q_{i,n,m}(t)\right|^2 \tag{5}$$

where $h^q_{n,m}(t) = \left|\tilde{h}^q_{n,m}(t)\right|^2 d_{n,m}(t)^{-\alpha}$ and $h^q_{i,n,m}(t) = \left|\tilde{h}^q_{i,n,m}(t)\right|^2 d_{i,n,m}(t)^{-\alpha}$ represents the subchannel gain between F-AP n and user m on subchannel q, and the subchannel gain between F-AP i and user m served by F-AP n on subchannel q in time slot t, respectively. In addition, $\tilde{h}^q_{n,m}(t), \tilde{h}^q_{i,n,m}(t) \sim \mathcal{CN}(0,1)$, $d_{n,m}(t)$ and $d_{i,n,m}(t)$ represent the distance between the F-AP n and the served user m and the distance between the F-AP i and the user m under F-AP n, respectively. α is the path loss factor.

Based on (3), (4) and (5), the data transmission rate of user m under F-AP n is

$$\begin{aligned}\eta_{n,m}(t) &= B/Q\log_2\left(1 + SINR_{n,m}(t)\right)\\ &= B/Q\log_2\left(1 + \frac{\lambda_{n,m}(t)p_f\left|h^q_{n,m}(t)\right|^2}{I^\theta_{n,m}(t) + I^\beta_{n,m}(t) + \sigma^2}\right).\end{aligned} \tag{6}$$

In this paper, the weighted sum of EE of all F-APs is used to judge the system performance, and the EE of each F-AP is given by

$$EE_n(t) = \frac{\sum_{m=1}^{k_n}\eta_{n,m}(t)}{\sum_{m=1}^{k_n}\lambda_{n,m}(t)p_f + P_c} \tag{7}$$

where P_c is the constant power consumption of the whole communication link. And the weighted sum of EE of all F-APs is expressed as

$$EE_{sum}(t) = \sum_{n=1}^{N} w_n EE_n(t) \tag{8}$$

where w_n represents the weight of the EE of F-AP n.

III. Problem description And Proposed Approaches

In this section, we formulate the optimization index of system performance, that is, maximizing the weighted sum of system EE of all F-APs and put forward the corresponding solutions. Since this is a nonconvex and non-deterministic polynomial (NP)-hard problem, therefore, it is decomposed into two subproblems: subchannel allocation and power allocation. DQN and DDPG in DRL algorithm are used to obtain subchannel and power allocation strategies, respectively.

A. Problem Formulation

The original optimization problem is formulated as

$$\begin{aligned}\text{P1:}\ &\max_{\{b^q_{n,m}(t),\lambda_{n,m}(t)\}} EE_{sum}(t)\\ \text{s.t.}\ \ &\text{C1}: \sum_{q=1}^{Q}\sum_{n=1}^{N}\sum_{m=1}^{K_n} b^q_{n,m}(t)\lambda_{n,m}(t) \le 1\\ &\quad \forall q\in\mathcal{Q}, n\in\mathcal{N}, m\in\mathcal{K}_n\\ &\text{C2}: \lambda_{n,m}(t)\ge 0 \quad \forall n\in\mathcal{N}, m\in\mathcal{K}_n\\ &\text{C3}: \lambda_{n,m}(t)p_f \le p_{max} \quad \forall n\in\mathcal{N}, m\in\mathcal{K}_n\\ &\text{C4}: b^q_{n,m}(t)\in\{0,1\} \quad \forall n\in\mathcal{N}, m\in\mathcal{K}_n, q\in\mathcal{Q}\end{aligned} \tag{9}$$

where $b^q_{n,m}(t)$ and $\lambda_{n,m}(t)$ represent subchannel and power allocation strategies in each time slot t, respectively. Constraints C1 and C2 represent the upper bound of the ratio of a given power allocation and the nonnegativity of the ratio, respectively. And C3 represents the power allocated to the served users under each F-AP can not exceed p_{max}. The restriction C4 represents whether the user has selected the subchannel q.

B. DRL Framework in the Downlink NOMA System

Here, we introduce the meanings of three elements of DRL in our proposed downlink NOMA system, which are as shown follows.

$States$: The channel gain of different subchannels occupied by each user for all F-APs in time slot t is taken as the observed state of the agent. Therefore, the states can be represented by

$$S_t = \left\{h^q_{n,m}(t), h^q_{i,n,m}(t)|\forall n,i\in\mathcal{N}, i\neq n, m\in\mathcal{K}_n, q\in\mathcal{Q}\right\}. \tag{10}$$

The change of state is reflected in the mobility of users and the time-varying characteristics of channels. In order to better represent the time-varying characteristics of the channel, we adopt the first-order Gaussian-Markov channel model to obtain the small-scale fading coefficients of the next time slot [9].

$$\tilde{h}^q_{n,m}(t+1) = \epsilon\tilde{h}^q_{n,m}(t) + \sqrt{1-\epsilon^2}u^q_{n,m}(t+1) \tag{11}$$

and

$$\tilde{h}^q_{i,n,m}(t+1) = \epsilon \tilde{h}^q_{i,n,m}(t) + \sqrt{1-\epsilon^2} u^q_{i,n,m}(t+1) \quad (12)$$

where ϵ means the time-varying degree of the channel and $u^q_{n,m}(t+1), u^q_{i,n,m}(t+1) \sim \mathcal{CN}(0,1)$.

$Actions$: The agents select two actions from the action space according to the observed state in the environment, one is the subchannel allocation strategy $A_1(t)$, and the other is the power allocation strategy $A_2(t)$. The actions could be given by

$$A_t = \left\{A^1_t, A^2_t\right\}. \quad (13)$$

In addition,

$$A^1_t = \left\{b^q_{n,m}(t) | \forall n \in \mathcal{N}, m \in \mathcal{K}_n, q \in \mathcal{Q}\right\} \quad (14)$$

and

$$A^2_t = \left\{\lambda_{n,m}(t) | \forall n \in \mathcal{N}, m \in \mathcal{K}_n\right\}. \quad (15)$$

$Reward$: Under the framework of DRL, the ultimate learning goal is to get the maximum total reward. We use the weighted sum of system EE in NOMA-based F-RAN system of each F-AP in time slot t as the immediate reward R_t according to the current time slot states and chosen actions. And it can be expressed by

$$R_t = EE_{sum}(t) = \sum_{n=1}^{N} w_n EE_n(t). \quad (16)$$

C. Problem Approach by Joint DQN And DDPG

We adopt DQN algorithm to solve the subproblem of subchannel allocation. The biggest difference between DQN and Q-learning is as follows. Generally speaking, DQN makes the Q table selected by actions in the original Q-learning into a neural network, and outputs the Q values of the total actions through the neural network according to the current state. The action through an exploration method is selected by the largest Q value or the randomly.

DQN has great limitations for continuous action problems such as the power allocation subproblem. As we all know, these problems need to be solved by DDPG algorithm. Similar to DQN, if the update target is constantly changing during the update, it will be difficult to update, which requires the help of fixed network technology. In addition, the approximate method of calculating and updating the actor network after sampling a mini-batch is as follows.

$$\nabla_\phi J(\zeta) \approx \frac{1}{N} \sum_i \left[\nabla_\phi \zeta(S;\phi)|_{S=S_v} \nabla_A Q(S,A;\psi) \Big|_{S=S_v, A=\zeta(S_v;\phi)} \right]. \quad (17)$$

Therefore, the algorithm needs four networks, namely actor, critic, target-actor, and target-critic. Among them, the function of critic is to estimate the Q value, and the role of actor is to output an action to critic to get the maximum Q value. The pseudo-code expression is shown in Algorithm 1.

Algorithm 1: Subchannel Allocation and Power Allocation by Joint DQN and DDPG

Initialize replay memory as RPM;
Initialize the Q network $Q(S,A;\vartheta)$ and the target Q network $Q^{'}(S,A;\vartheta^{'})$ with the same weights ϑ and $\vartheta^{'}$;
Initialize the actor network $\zeta(S;\phi)$ and the critic network $Q(S,A;\psi)$ with the weights ϕ and ψ;
Initialize the target-actor network $\zeta^{'}(S;\phi^{'})$ and the target-critic network $Q^{'}(S,A;\psi^{'})$ with weights $\phi^{'} = \phi$ and $\psi^{'} = \psi$;
Initialize the maximum episodes ep_{max}, the maximum time slots t_{max} for each episode, the random noise disturbance n^q_t, the update interval U of the target Q network and the total time slot t_{total} set zero;
for $episode = 1, 2, 3, ..., ep_{max}$ **do**
  Randomly initialize the state S_1;
  for *time slot* $t = 1, 2, 3, ..., t_{max}$ **do**
    $t_{total} + 1$;
    The sub-channel allocation DQN agent randomly select the action A^2_t with probability ϵ, or select the action with the maximum $Q(S_t, A^2_t; \vartheta)$ The power allocation DDPG agent select the action A^2_t according to $\zeta(S;\phi) + n^q_t$, and is limited between 0 and 1. Put the stateS_t and the obtained action $A_t = \left\{A^1_t, A^2_t\right\}$ into the environment to interact, get rewards $R(t)$, and the state of the next time slot S_{t+1};
    Store $(S_t, A^1_t, R_t, S_{t+1})$ in the RPM;
    Sample a random mini-batch $(S_v, A^1_v, R_v, S_{v+1})$ from RPM;
    Set y^1_v=

$$\begin{cases} R_v, & t = max \\ R_v + \max\limits_{A^1_{v+1}} Q\left(S_{v+1}, A^1_{v+1}; \vartheta^{'}\right), & otherwise \end{cases}$$

    The Q network minimize the loss function to update the weights according to $\left(y_v - Q\left(S_v, A^1_v; \vartheta\right)\right)^2$;
    The actor network update its weights by (16);
    Set $y^2_v = R_v + \gamma Q\left(S_{v+1}, \zeta\left(S_{v+1}; \phi^{'}\right); \psi^{'}\right)$;
    The critic network update its weights by $\frac{1}{N}\sum_i \left(y^2_v - Q\left(S_v, A^2_v; \psi\right)\right)^2$;
    if t_{total} *is divisible by* U **then**
      The weights of Q network is assigned to the target Q network $\vartheta^{'} = \vartheta$;
    end
    Update the weights of target-actor network and target-critic network by $\phi^{'} = \tau\phi + (1-\tau)\phi^{'}$ and $\psi^{'} = \tau\psi + (1-\tau)\psi^{'}$
  end
end

IV. Simulation Results

In this section, the analysis of simulation results of EE of multiple F-APs and multiple users in NOMA-based downlink F-RAN. "DQN-MP", "RS-DDPG", "RS-MP", and "SA" are used to compare the advantages of using DQN algorithm to obtain subchannel allocation strategy and DDPG algorithm to obtain power allocation strategy, i.e., "DQN-DDPG". "RS" represents the subchannel allocation strategy by random selection, and "MP" represents a scheme in which each F-AP allocates the maximum power of the served users. "SA" is a heuristic algorithm called simulated annealing algorithm. Each network adopts four fully connected layers, that is, one input layer, one output layer, and two hidden layers. In order to facilitate network training, the total bandwidth of the system is normalized [8]. Unless otherwise specified, the default parameter settings are shown in Table 1 below.

TABLE I
Simulation Parameters

Parameters	Value
The radius of each F-AP r	50m
The total power of each F-AP p_f	4W
The learning rate of DQN	10^{-4}
The number of neurons in hidden layers of DQN	$680, 540$
The learning rate of Actor network	10^{-4}
The learning rate of Critic network	10^{-4}
The number of neurons in hidden layers of DDPG	$400, 300$
The path loss factor α	4
The time-varying degree of the channel ϵ	0.9
The batch size of the network	256
The discount factor γ	0.95

Fig. 2 shows the results change in the iterative process of the four schemes for achieving the optimal weighted sum of EE. We can find that the scheme of "DQN-DDPG" has obvious advantages on the whole, and it first converges to 2.9 at about 700 episodes, then starts to rise at about 2,000 episodes. This is because the stability tended to before this has fallen into a local optimal situation, and the rise since then is due to the network finding of a better solution. Finally, it converges to the maximum value of 3.1 at about 2,400 episodes.

However, the "RS-DDPG" scheme can quickly converge to at 2.3 about 50 episodes, and it can be found that although the convergence of the iterative curve is stable, it can still be seen that there are drastic fluctuations, which is due to the contingency of "RS", a scheme of randomly selecting subchannel allocation, which will cause for the fluctuations in the results. Comparatively speaking, the results of "DQN-MP" scheme fluctuate smoothly after the stability of 1,200 episodes, and the total rising amplitude is lower than the above two. The fourth scheme is basically stable in the whole iterative process. This is because the scheme does not have a learning process.

As shown in Fig. 3, different combinations of γ and learning rate have different results under the scheme of "DQN-DDPG", and the learning rate of 0.0001 is better than that of 0.0005. What's more, different γ bring different convergence results at the same learning rate. For example, although the final convergence results of two curves with the same learning rate

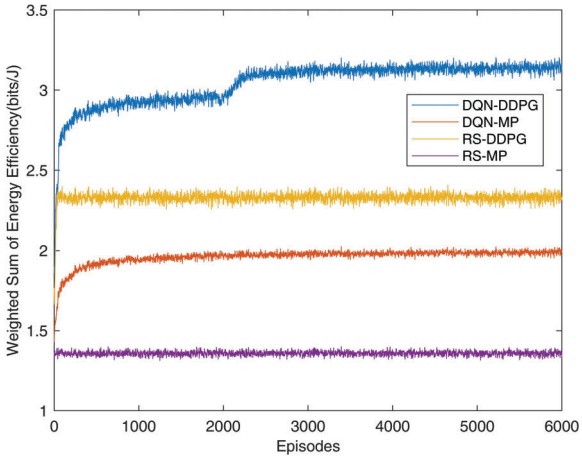

Fig. 2. Weighted sum of EE comparison of the four different approaches.

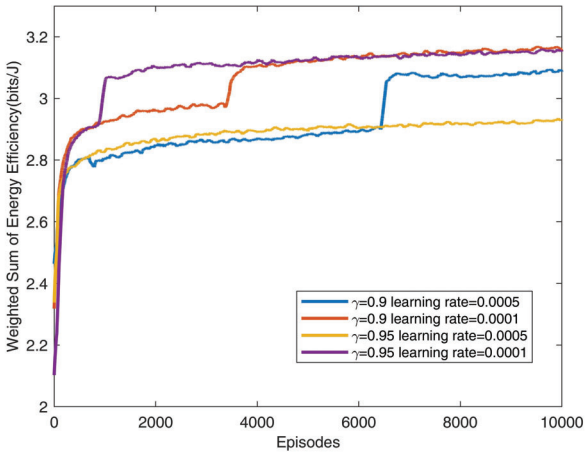

Fig. 3. Influence of different discount factors and learning rates on weighted sum of EE.

of 0.0001 are roughly the same, the convergence speed of the curve with γ = 0.95 is obviously faster than that of the curve with value of 0.9. The former converges to the maximum value around 1800 episodes, while the latter needs to reach the maximum value around 3700 episodes. However, when the learning rate is 0.0005, it is not satisfactory. It may be because the learning rate is too high, the position of finding the optimal value is neglected, and the optimal convergence result cannot be achieved.

Fig. 4 is the comparison between the change of the total power given to each F-AP and the results in these five cases. We can find that the results of "DQN-DDPG" strategy are generally better than other strategies, and with the increase of power, the overall results are in a fluctuating state, and the weighted sum of EE can reach the maximum at 2W or 3W. This shows that blindly increasing the total transmission power of F-AP can not necessarily achieve good results. Only

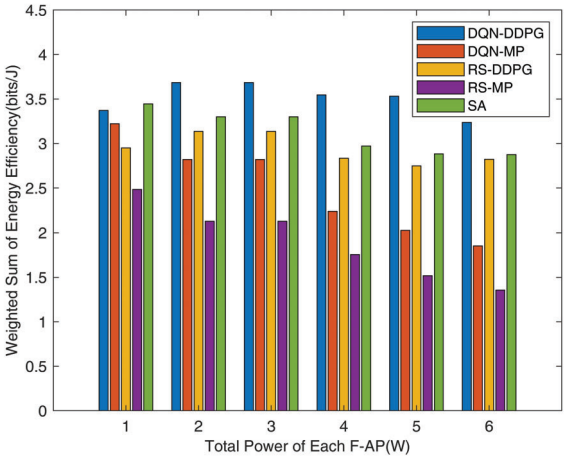

Fig. 4. Weighted sum of EE versus the power of each F-AP.

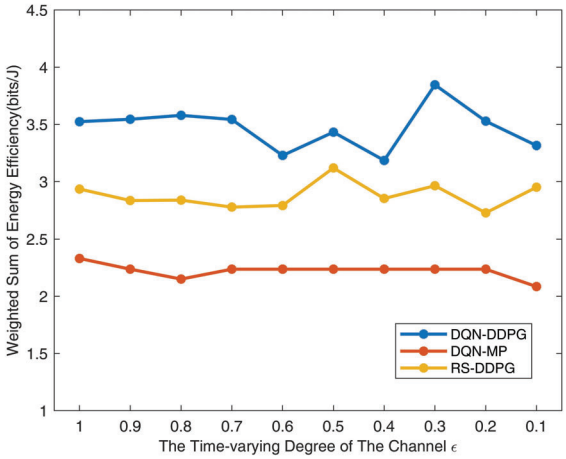

Fig. 5. Weighted sum of EE versus the time-varying degree of the channel.

when p_f = 1W, which is a relatively smaller power, the "SA" algorithm can be equivalent to "DQN-DDPG". This also shows the limitations of "SA". By observing the comparison results of "DQN-MP" and "RS-DDPG", it can be found that the increase of power also leads to the greater impact on EE.

The results change of the same state for testing under the three schemes that need to undergo network training in different channel time-varying degrees is shown in Fig. 5. On the whole, the impact of different channel time-varying degrees ϵ under "DQN-DDPG" and "RS-DDPG" schemes on the results is more severe than that of "DQN-MP". However, the fluctuation is more gentle when ϵ is larger. And the relatively smaller ϵ, that is, the time-varying characteristics of the channel are larger, will have a greater impact on the network learning and will cause the fluctuation of the results. Moreover, the results of "DQN-DDPG" scheme are better than the other two schemes, and can make better decisions in different time-varying characteristics of channels.

V. Conclusion

In this article, we have studied how to optimize the EE in the system environment of downlink F-RAN based on NOMA with multiple F-APs and multiple users. The major problem of optimizing weighted sum of EE is divided into two subproblems: subchannel allocation and power allocation. The former adopted DQN and the latter adopted DDPG algorithms to obtain the best allocation strategy. Simulation results show that under different schemes, "DQN-DDPG" has great advantages, including faster convergence and better optimization results than other schemes. In addition, the limitations of traditional "SA" algorithm in solving such problems are also discussed.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62071306, and in part by Shenzhen Science and Technology Program under Grants JCYJ20200109113601723 and JSG-G20210420091805014.

References

[1] Z. Liu, Y. Yang, Y. Chen, K. Li, Z. Li and X. Luo, "A Multi-tier Cost Model for Effective User Scheduling in Fog Computing Networks," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 1-6.

[2] Y. Xiao and M. Krunz, "Distributed Optimization for Energy-Efficient Fog Computing in the Tactile Internet," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2390-2400, Nov. 2018.

[3] Z. Liu, Y. Yang, K. Wang, Z. Shao and J. Zhang, "POST: Parallel Offloading of Splittable Tasks in Heterogeneous Fog Networks," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3170-3183, April 2020.

[4] Y. Liu, F. R. Yu, X. Li, H. Ji and V. C. M. Leung, "Distributed Resource Allocation and Computation Offloading in Fog and Cloud Networks With Non-Orthogonal Multiple Access," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 12137-12151, Dec. 2018.

[5] X. Wen, H. Zhang, H. Zhang and F. Fang, "Interference Pricing Resource Allocation and User-Subchannel Matching for NOMA Hierarchy Fog Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 467-479, June 2019.

[6] R. Rai, H. Zhu and J. Wang, "Performance Analysis of NOMA Enabled Fog Radio Access Networks," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 382-397, Jan. 2021.

[7] J. Zhang, X. Tao, H. Wu, N. Zhang and X. Zhang, "Deep Reinforcement Learning for Throughput Improvement of the Uplink Grant-Free NOMA System," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6369-6379, July 2020.

[8] X. Wang, Y. Zhang, R. Shen, Y. Xu and F. -C. Zheng, "DRL-Based Energy-Efficient Resource Allocation Frameworks for Uplink NOMA Systems," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7279-7294, Aug. 2020.

[9] S. Wang, T. Lv, W. Ni, N. C. Beaulieu and Y. J. Guo, "Joint Resource Management for MC-NOMA: A Deep Reinforcement Learning Approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 9, pp. 5672-5688, Sept. 2021.

[10] M. Chu, H. Li, X. Liao and S. Cui, "Reinforcement Learning-Based Multiaccess Control and Battery Prediction With Energy Harvesting in IoT Systems," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2009-2020, April 2019.

[11] S. Sen, N. Santhapuri, R. R. Choudhury, and S. Nelakuditi, "Successive interference cancellation: A back-of-the-envelope perspective," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, 2010, pp. 1-6.