Deep Reinforcement Learning Based Dynamic Resource Allocation in 5G Ultra-Dense Networks

Zhiyong Liu*, Xin Chen*, Ying Chen*, Zhuo Li*
* *School of Computer Science, Beijing Information Science & Technology University*
Beijing, China
Email: 2018020368@mail.bistu.edu.cn, {chenxin, chenying, lizhuo}@bistu.edu.cn

***Abstract*—The rapid development of Internet of things (IoT) technology has promoted the densification of network infrastructure. Ultra-dense networks (UDN) will become a key technology in 5G networks. We investigated the dynamic resource allocation problem over 5G UDN. We considered the energy efficiency (EE) and spectral efficiency (SE) of the network. Therefore, the resource allocation problem at different moments was expressed as a joint optimization problem. Considering the dynamic nature of the environment, the EE and SE were dynamically weighted. In order to guarantee the long-term performance of UDN system, the joint optimization problem was described as a markov decision process (MDP). In view of the fact that the densification of the network makes the space explosion of MDP and makes it difficult to solve by traditional methods, the dueling deep Q network (Dueling DQN) method was proposed. Simulation results showed that compared with traditional Q-learning and DQN, this algorithm has obvious performance improvement.**

***Index Terms*—Ultra-dense Network (UDN); Resource Allocation; Energy Efficiency (EE); Spectral Efficiency (SE); Markov Decision Process (MDP); Dueling Deep Q Network (Dueling DQN).**

I. Introduction

In recent years, with the rapid development of mobile Internet and Internet of things (IoT) technology, the demand for mobile communication business is also growing rapidly. Emerging businesses such as 4K ultra-high definition video, virtual reality, augmented reality and unmanned driving have higher requirements for 5G network performance [1]–[3]. The communication needs of intelligent devices and mass IoT environment promote the densification of network infrastructure [4]–[6]. Ultra-dense network (UDN) emerged at the right moment and will become a key technology in 5G network [7]. In the UDN architecture, the radius of cells will be further reduced, resulting in increased interference between cells [8]. Moreover, due to the non-uniform characteristics of user devices in space and time, it is more difficult to manage resources of UDN [9]. Therefore, how to allocate resources adaptively in UDN is worth further discussion.

Resource allocation policies in UDN affect network performance and user experience. A large amount of works systematically studied resource allocation in heterogeneous network scenarios (HetNet) and UDN scenarios from different perspectives. In [10], the maximization of sum rate of small cell users in NOMA system is studied. The optimization of energy efficiency in a dense small cell network is studied in [11]. The balance between energy efficiency and spectral efficiency was studied in [12] and [13]. However, these studies only considered network performance and user requirements in a fixed state. As the network environment changes, the performance requirements of the system may also change. Therefore, network resource allocation should consider the interaction with the environment.

Model-free reinforcement learning (RL) framework can be adopted to solve the stochastic optimization problem in wireless networks. In the unknown environment, RL will get the optimal policy through the interaction with the environment [14], [15]. Q-learning is one of the most popular RL algorithms. However, due to the explosion of action state space in practical problems, Q-learning convergence is slow and it is difficult to find the optimal action to solve the problem. Deep Q network (DQN) is a new deep RL (DRL) algorithm, which can combine the process of RL with a kind of neural network called deep neural network to approximate the action - state value function. Therefore, the limitations of Q-learning are solved in DQN. In [16], adopts DQN in optimizing the ON/OFF strategy of small base stations to enhance EE while meeting QoS requirements. A decentralized resource allocation mechanism based on DQN is proposed in [17] to allocate the communication resources of vehicle-to-vehicle (V2V), so that the can meet strict delay limit and minimize interference. In [18], DQN was used to solve the problem of resource allocation in wireless MEC, and the total cost was significantly reduced.

In this paper, a dynamic model for simultaneous optimization of EE and SE in UDN system is established. Then, in order to overcome the instability of common natural DQN, an dueling deep Q network (Dueling DQN) algorithm is utilized to solve it. Simulation results show that this algorithm has better performance than traditional Q-learning and DQN. The major contributions of this paper are summarized as follows:

- MOOP that is jointly optimized by EE and SE is converted to SOOP by dynamically weighting EE and SE. The dynamic weight coefficient changes with the number of users. In this way, the optimized system can meet the requirements of different network states.
- When resources are allocated in each time slot , they are allocated only to users who have just arrived. If a resource block is already occupied in a small area at this time, its allocation will not be considered. Until the user occupying the block leaves. So the future state depends

978-1-7281-3488-8/19/$31.00 ©2019 IEEE
DOI 10.1109/SmartIoT.2019.00034

on the current state. The transition of state conforms to markov property. The solving process of the model can be transformed into MDP. The impact of future states needs to be considered when considering current decisions.

- Considering that traditional RL Q-learning will face the problem of state and action space explosion when solving the MDP problem, and the instability of ordinary DQN. We use Dueling DQN to solve MDP. The dimensional explosion of state and action space is resolved and the stability of system performance is guaranteed.

The remainder of this paper is as follows. In Section II, we provide the system model and optimization problems. We propose a dynamic resource allocation algorithm based on Dueling DQN in Section III. In Section IV, we discuss the simulation results. Section V concludes this paper and discusses the future work.

II. System Model

The considered downlink 5G ultra-dense network scenario is shown in Fig. 1, where N small cells are distributed in one macro cell. A macro gNB (MgNB) is deployed in the macro cell. Each small cell deploys a small gNB (SgNB). The set of SgNB as $\mathfrak{N} = \{1, 2, ..., N\}$. It is assumed that there are M RBs are available. RBs set can be represented by $\mathfrak{M} = \{1, 2, ..., M\}$ and the bandwidth of each RB is B_m. The MgNB as an agent to collect information. The agents determine which RBs are available. Each SgNB chooses one RB from the available RBs, and assigns a small UE (SUE).

Similar with [19], we model the user arrival and departure process in every network as two independent stochastic processes. In each time slot, UEs arrive to the each small cell according to a Poisson process with the parameter λ_t. Thus, the probability that there are x new SUE arriving to the small cell during the period τ is

$$P(x) = \frac{(\lambda_t \tau)^x}{x!} e^{-\lambda_t \tau}. \tag{1}$$

Similarly, SUE depart from the small cell following a Poisson process too, with the parameter μ_t. Therefore, the probability that there are y SUE departing from the small cell during the period τ is

$$P(y) = \frac{(\mu_t \tau)^y}{y!} e^{-\mu_t \tau}. \tag{2}$$

In our work, we assume that the SUE association to the small cell are completed prior to the resource allocation. Then, we can define the set of user devices associated with SgNB n in time slot t as

$$U_n(t) = \{1, 2, ...s_n(t)..., S_n(t)\}. \tag{3}$$

In this paper, we allow that the same resource block will be reused by multiple SgNBs at the same time. So each SUE of the small cell will be interfered by SgNBs of other small cell. Therefore, in the small cell n, when the RB m is allocated to the $s_n(t)$, its signal to interference-plus-noise ratio (SINR) is given by

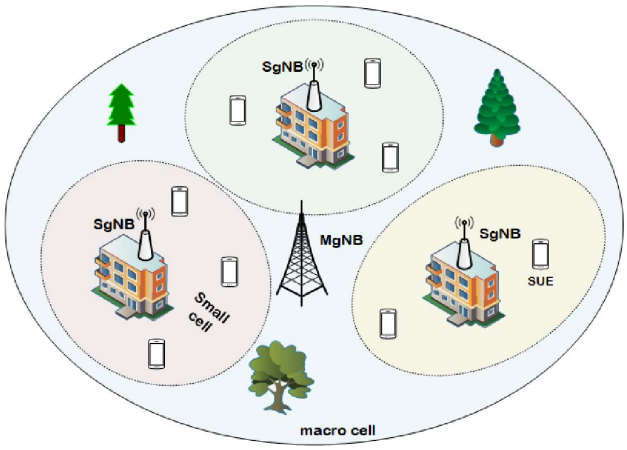

Fig. 1. 5G Ultra-Dense Network Scenario

$$\gamma_{s_n}(t) = \frac{p_n g_n^m}{\sum\limits_{\substack{i=1 \\ i \neq n}}^{N} x_n^m(t) p_n g_i^m(t) + \sigma^2}, \tag{4}$$

where g_n^m is the channel gains from the SgNB n to the SUE s_n when RB m is reused, p_n is the power assigned by SgNB n to each SUE, σ is the variance of additive white gaussian noise (AWGN). $x_n^m(t)$ is a binary variable that represents the RB allocation indicator. That is

$$x_n^m(t) = \begin{cases} 1, & \text{the RB } m \text{ is allocated to the small cell } n, \\ 0, & \text{otherwish.} \end{cases} \tag{5}$$

According to the SINR expression of SUE s_n, the downlink sum throughput in the n-th small cell can be calculated by shannon's formula as

$$R_n(t) = \sum_{s_n=1}^{S_n} R_{s_n}(t) = \sum_{s_n=1}^{S_n} B_m \log_2[1 + \gamma_{s_n}(t)], \tag{6}$$

and The total throughput of the whole system is

$$R(t) = \sum_{n=1}^{N} R_n(t) = \sum_{n=1}^{N} \sum_{s_n=1}^{S_n} B_m \log_2[1 + \gamma_{s_n}(t)]. \tag{7}$$

In this paper, energy efficiency (EE) of the 5G ultra-dense network system is defined as the ratio between the sum throughput and the power consumption of the whole system. In the time slot t the EE can be expressed as

$$\eta_{EE}^{(t)} = \frac{R(t)}{p_M + \sum\limits_{n=1}^{N} \sum\limits_{s=1}^{S} s_n(t) p_n + p_C}, \tag{8}$$

where p_M and p_C are respectively the power of MgNB and consumption generated by all circuit transmissions.

Spectrum efficiency (SE) of the 5G ultra-dense network system is defined as the ratio between the sum throughput and the bandwidth of the whole system, which in the time plot t is given by

$$\eta_{SE}^{(t)} = \frac{R(t)}{\sum_{m=1}^{M} B_m}. \tag{9}$$

Maximizing system SE is equivalent to maximizing the resources available to the system. The maximization of available resources will increase the power consumption of the system, which may lead to the reduction of system EE. Conversely, maximization of EE may result in reduction of SE. Therefore, it is impossible to meet the system performance requirements only by considering the maximization of SE or EE. It is necessary to consider the tradeoff between SE and EE. Solving the tradeoff between SE and EE is a multi-objective optimization problem.

However, during peak and off-peak periods, users have different requirements for SE and EE in the system. Therefore, we use the method of dynamic weighted sum to transform the MOOP of the tradeoff between SE and EE into a SOOP. Dynamic weights are given according to user requirements for SE and EE. The more users there are, the greater the proportion of SE will be. The smaller the number of users, the more important EE becomes. Therefore, the weight between SE and EE in the time plot t can be expressed as

$$\xi^{(t)} = \frac{\sum_{n=1}^{N} U_n(t)}{M \times N}. \tag{10}$$

Then the tradeoff between SE and EE can be defined as

$$\max_{x} \sum_{t=1}^{T} \eta(t) = \sum_{t=1}^{T} [(1-\xi^{(t)})\eta_{EE}^{(t)} + \xi^{(t)}\eta_{SE}^{(t)}] \tag{11}$$

subject to:

$$R_{s_n}(t) \geq R_0, \forall t \in T, \forall n \in \mathfrak{N}, \forall s_n \in U_n, \tag{11a}$$

$$x_n^m(t) \in \{0,1\}, \forall t \in T, \forall n \in \mathfrak{N}, \forall m \in \mathfrak{M}, \tag{11b}$$

$$\sum_{m=1}^{M} x_n^m(t) = S_n(t), \forall t \in T, \forall n \in \mathfrak{N}, \forall m \in \mathfrak{M}, \tag{11c}$$

$$\sum_{s=1}^{S} s_n(t) \leq S_{max}, \forall t \in T, \forall n \in \mathfrak{N}, \forall s_n \in U_n, \tag{11d}$$

where constraint (11a) guarantees the QoS of SUE by requiring higher than expected throughput thresholds; constraints (11b) means that at time plot t, Small cell can only choose whether or not to reuse RB m. '1' means yes and '0' means no; constraints (11c) means that each RB can be assigned to a maximum of one UE in each small cell. S_{max} is the maximum number of UE a small cell can serve. Constraints (11d) means that means that each small cell can hold a finite amount of UE.

III. Algorithm Description

In this paper, in addition to EE, SE is also considered comprehensively, so the resource allocation problem becomes a np-hard problem, and it is difficult to obtain the optimal solution. When making decisions for each slot MgNB, it only allocates resources to the SUE that enters the slot. There will be no redistribution of the allocated SUE. In order to weigh the EE and SE of the system, the resource allocation problem under the 5G UDN scenario can be expressed as a MDP. This paper adopts the method of deep reinforcement learning to solve this problem.

A. The Basic Bodel of Reinforcement Learning

Reinforcement learning sees learning as a heuristic evaluation process. The Agent selects an action for the environment. The state changes after the environment accepts the action. Meanwhile, a reward feedback is generated to the Agent. The Agent selects the next action according to the reward and the current state of the environment. The selection principle is to increase the reward probability, so as to obtain the optimal strategy. Therefore, environment, state and reward are three key factors in reinforcement learning. For the system considered in this paper, we define state space, action space and reward in time slot t based on the framework of reinforcement learning, namely

- ***State Space***: The decisions are made by the agent MgNB. The agent should know the state of each small cell to determine the action. Therefore, the state of agent in time slot t is

$$s_t = \{s_1(t), R_1(t), ..., s_n(t), R_n(t), x(t)\}. \tag{12}$$

  This means that the agent will know the number and throughput of all small cells and the allocation of all RBs in the system.
- ***Action space***: The agent decides which RBs are reused by the small cell. So the action is

$$a_t = \{x_1(t), x_2(t), ..., x_n(t)\}. \tag{13}$$

  The action space increases exponentially with the increase of SgNB. The explosion of action space will be an important and difficult problem to deal with. Each action affects a state, which means that the number of state spaces is also very large.
- ***Reward function***: When agent MgNB takes action a by observing state s_t, it will get immediate reward r_t. Our goal is to maximize $\sum_{t=1}^{T} \eta(t)$, so the reward function is $\eta(t)$ at time t, namely,

$$r_t = \eta(t) = (1-\xi^{(t)})\eta_{EE}^{(t)} + \xi^{(t)}\eta_{SE}^{(t)}. \tag{14}$$

Because the number of RBs is very large and the number of cells increases, the action space becomes very large and the corresponding state space becomes very large. In order to limit the size of the action space, we propose a pre-screening step before learning. For some time t, if the throughput

$R_n(t)$ calculated for the corresponding action does not satisfy constraint (11a), the action will not be executed. In this way, we can reduce the possible value of $x_n(t)$ to limit the action space of agent MgNB.

B. Deep Q Network Strategy

At each instant t, agent MgNB determines the action $a_t = \pi(s_t)$ through the policy π according to the current state s_t. That is, MgNB is rewarded by allocating the available resource blocks to SgNB. In reinforcement learning, the expected return is defined by the state-action value function $Q^{\pi}(s_t, a_t)$, which can be expressed as

$$Q^{\pi}(s_t, a_t) = E_{\pi}\left[\sum_{t=1}^{T} \gamma^t \eta(t|s = s_t, a = a_t)\right], \tag{15}$$

where γ is the discount factor. The goal is to make future returns less relevant to the present. So the sum of $\sum_{t=1}^{T} \gamma^t \eta(t|s = s_t, a = a_t)$ converges. The choice of actions cannot be completely determined due to the user's arrival and departure obeys the poisson distribution. Therefore, $Q^{\pi}(s_t, a_t)$ is essentially the mathematical expectation of the long-term returns that are generated by a strategy.

The goal of MDP is to find an optimal strategy, that is, the strategy that gets the most rewards. This means that for all actions selected by the optimal strategy $\pi^*(s_t) = \arg\max Q(s_{t+1}, a_{t+1})$, $a_t = \pi^*(s_t)$ will maximize the state-action value function $Q^{\pi}(s_t, a_t)$ of (15), and satisfy $Q^{\pi^*}(s_t, a_t) > Q^{\pi}(s_t, a_t)$ for all $Q^{\pi}(s_t, a_t)$. Q-learning is a famous method to solve MDP. In Q-learning, according to behrman equation, $Q(s_t, a_t)$ can be expressed as

$$Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha\left[\eta(t) + \gamma \max Q(s_{t+1}, a_{t+1})\right], \tag{16}$$

where α is learning rate, which affects the learning speed of $Q(s_t, a_t)$.

When the space is very large, it is very difficult to find the optimal strategy by looking up the Q-value table in Q-learning. In recent years, in order to solve the problem of large space, deep neural network (DNN) was introduced in the Q-learning framework. DQN is the most famous method. In DQN, DNN can be used to fit the optimal strategy and optimal value function, i.e

$$Q^*(s_t, a_t; \theta) \approx Q(s_t, a_t), \tag{17}$$

where θ is the parameter of the neural network.

In order to ensure the stability of $Q^*(s_t, a_t; \theta)$, each step needs to train the evaluated neural network to minimize the loss function $L(\theta)$, so as to approach the real $Q(s_t, a_t)$. $L(\theta)$ can be expressed as

$$L(\theta) = E\left[(\eta(t) + \gamma \max Q(s_{t+1}, a_{t+1}; \theta^-) - Q^*(s_t, a_t; \theta))^2\right], \tag{18}$$

where θ^- is the parameter of the target network, and θ is the parameter of the behavior network.

Algorithm 1 Dynamic resource allocation algorithm based on Dueling DQN

Input: reinforcement learning:(state,action), QoS, set of available RBs $\mathfrak{M}$

Output: Optimal action sequence $\{x_n^m(t)\}$, weighted sum $\sum_{t=1}^{T} \eta(t)$ of EE and SE at all time instants

```
Initialize the replay buffer D with the capacity of N
Initializes the state-action values network Q*(s_t, a_t; θ) with weights θ
Initialize target network Q*(s'_t, a'_t; θ⁻) with weights θ⁻
for episode = 1 : K do
    Initialize the 5G UDN system environment, and MgNB receives the initial state s_1
    for t = 1 : T do
        MgNB choose an action a_t at the state s_t using ε-greedy policy from a_t = max Q*(s_t, a_t, θ)
        MgNB executes action a_t to allocate the selected RBs to SUE and calculates the immediate reward r_t based on (15)
        MgNB receives the system state at the next moment s_{t+1}
        MgNB stores experience {s_t, a_t, r_t, s_{t+1}} in replay buffer D
        if the capacity of D has reached N then
            MgNB randomly selects a batch of samples {s_j, a_j, r_j, s_{j+1}} from D
            MgNB calculate two streams of state-action values network, including V(s_t, θ, μ) and A(s_t, a_t, θ, ω), and combine them as Q*(s_t, a_t; θ, μ, ω) based on (19)
            if s_{j+1} is s_T then
                y_j = r_j
            else
                y_j = r_j + γ max Q*(s_{j+1}, a'_{j+1}; θ⁻, μ⁻, ω⁻)
            end if
            MgNB minimizes loss function L(θ, μ, ω) = E[y_j − Q*(s_t, a_t; θ))²] through gradient descent based on (18)
            MgNB completes target network parameter update θ⁻=θ every C step
        end if
    end for
end for
```

C. Dueling Deep Q Network Strategy

In many states, the value function size of RBs assigned to different users is different. However, in some states, different allocation policies may result in identical value functions. According to [20], Dueling DQN is an improved algorithm based on DQN, which uses the model structure to express the value function in a more detailed form. So that the model can have a better performance. In particular, the state - action value function is decomposed into a state - based value function and

a advantage function, namely

$$Q^*(s_t, a_t; \theta, \mu, \omega) = V(s_t; \theta, \mu) + A(s_t, a_t; \theta, \omega), \quad (19)$$

where μ, ω and θ represent the parameters of the state value streams, the action advantage streams and the remaining parts of the model respectively. However, in practice, the advantage streams of actions is generally set as the value of the advantage function of individual actions minus the average value of all the advantage functions of actions under a certain state. Therefore, the advantage function is actually expressed as

$$A(s_t, a_t; \theta, \omega) = A(s_t, a_t; \theta, \omega) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s_t, a'_t; \theta, \omega). \quad (20)$$

The operation of (20) can not only ensure that the dominant function of each action in this state remains unchanged, but also reduce the range of Q-value and remove the excess degrees of freedom, and improve the stability. Then, the dynamic resource allocation process in 5G ultra-dense system is represented by algorithm 1.

IV. Performance Evaluation

In this section, we evaluate the Dueling DQN for solving the dynamic resource allocation problem. First, we determine the value of discount factor. Secondly, the network performance of resource allocation using Q-learning, DQN and Dueling DQN is compared. Finally, we compare and analyze the convergence of DQN and Dueling DQN algorithms. Simulation results show that compared with classical Q-learning algorithm and ordinary DQN algorithm, Dueling DQN method is improved. When the number of SgNB increases, the performance improvement is especially obvious. The convergence of the algorithm is also verified.

A. Simulation Environment

We consider an MgNB, an ultra-dense network of SGNB-S.In the simulation, each user can only occupy one RB. Each RB consists of 12 continuous subcarriers. In the initial state, users in each small region follow the poisson distribution $\Gamma(\lambda_t)$ where $\lambda_t = |\sin(t)|$ and arrive randomly. In the following time slots, users will randomly leave according to the poisson distribution $\Gamma(\mu_t)$ where $\mu_t = |\cos(t)|$, and MgNB will then allocate the available RBs to the new users, and the already allocated users will not be redistributed. In Dueling DQN, in order to avoid the convergence of optimization target to local optimum, an adaptive ϵ-greedy strategy is used. The simulation parameters used are shown in table I [12].

B. Parameters Calculation

Discount factor γ affects algorithm performance. When γ is too small, agent MgNB is short-sighted and values immediate interests. When the γ = 0, the algorithm is similar to the greedy algorithm. The convergence rate is fast, but it is easy to premature convergence. When the γ is too large, the algorithm is not easy to converge. In order to balance the performance and convergence speed of the algorithm, the optimal γ needs to be calculated.

TABLE I
Parameter Settings

Simulation Parameter	Value
System total bandwidth	10 MHz
Total number of RBs	50 RBs
Number of SeNB (N)	6, 8, 10, 12, 14, 16
Small cell inter site distance	50m
power of MgNB p_M	46dBm
power of SgNB p_n	30dBm
Circuit Power of SeNB P_C	6.8W
R_0 of SUE	1.0Mbps
Path loss of SgNB	140.7+37.6log(d)
Shadowing fading variance of SgNB	10dB
Effective thermal noise power σ	-174 dBm/Hz
maximum number of UE a small cell S_{max}	30
Time slot τ	15s
Number of time slots T	1000
Shadowing fading variance of SgNB	10dB
Effective thermal noise power σ	-174 dBm/Hz
maximum number of UE a small cell S_{max}	30
Size of replay memory D	10000
Learning rate α	0.1

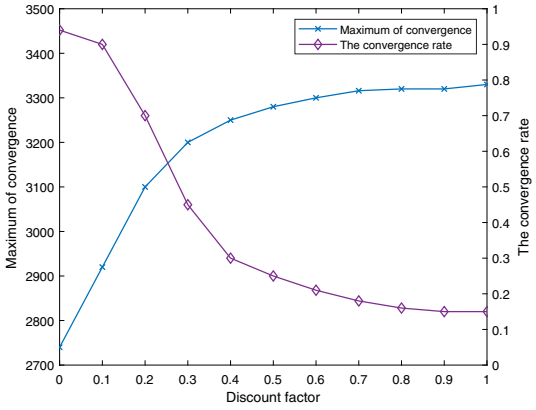

Fig. 2. The influence of discount factor on convergence speed and quality.

During the training, the comparison between the convergence rate and the maximum convergence value is shown in Fig. 2. After several iterations, when γ=0, the fastest convergence rate is obtained, and the convergence rate decreases rapidly with the increase of γ. After 0.5, the convergence rate began to decline in a gentle trend. The maximum convergence is reached at γ=1 and decreases with the decrease of γ. When γ <0.5, the maximum convergence starts to flatten out. By comprehensive consideration, γ is set as 0.3 to ensure the convergence speed and quality at the same time.

C. System Performance Calculation

In order to compare our scheme with other methods, namely Q-learning and DQN, when the number of SgNB changes, we first use these three algorithms to simulate the long-term EE and SE of the system. The total weighted energy-spectral efficiency of the system over a long period varies with the number of SgNB, as shown in Fig. 3. It can be seen that with the increase of the number of SgNB, the total weighted energy-spectral efficiency calculated by Q-learning, DQN and

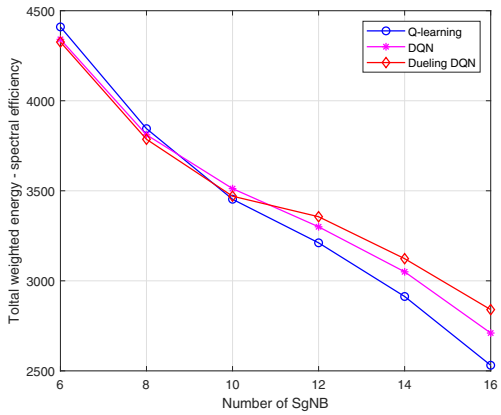

Fig. 3. Long-term weighted energy-spectral rate with numbers of SeNB.

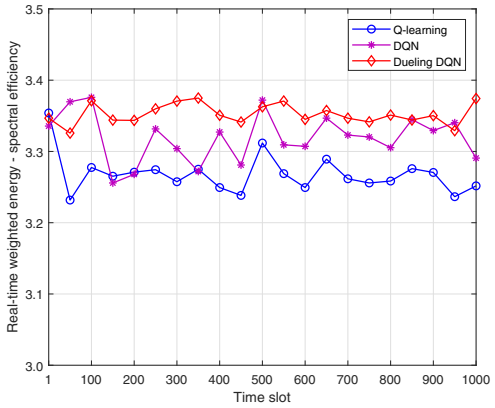

Fig. 4. Real-time weighted energy-spectral efficiency fluctuation.

Dueling DQN shows a declining trend.

When there are 6 SgNB, the total weighted energy-spectral efficiency calculated by Q-learning is the largest, but it is not much different from the other two algorithms. However, with the increase of the number of small SgNB, the total weighted energy-spectral efficiency calculated by Q-learning decreases the fastest. When the number of SgNB is 10, the calculated value of Dueling DQN reaches the maximum. Moreover, with the increasing number of SgNB, Dueling DQN has more obvious advantages. Therefore, Dueling DQN has more obvious performance improvement in the case of dense base station distribution.

In addition, we have also analyzed in detail the fluctuation curves of the energy-spectrum efficiency of each time slot of the 12 SgNB, as shown in Fig. 4. It can be seen that the real-time weighted energy-spectral efficiency is fluctuating. This is because the number of people in each time slot cell changes, and greedy strategy will be considered in the choice of strategy, not necessarily the optimal decision. Therefore, the real - time weighted energy - spectral efficiency fluctuation is caused.

Obviously, DQN has the largest fluctuation range. Sometimes exceeding the real-time weighted energy-spectral efficiency of Dueling DQN calculations. Sometimes even lower than Q-learning. The one with the smallest fluctuation amplitude is Dueling DQN. Therefore, Dueling DQN not only can obtain higher total weighted energy-spectral efficiency, but also has an advantage in stability.

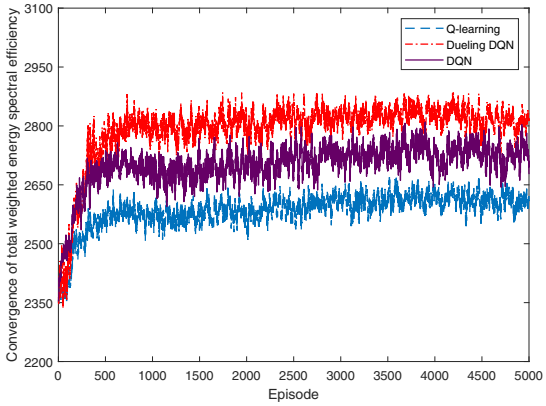

Fig. 5. Algorithm convergence.

D. Comparison and Analysis of algorithm convergence

The convergence of the algorithm will also affect the performance of the system. Therefore, we also compare and analyze the convergence of the algorithm. Considering the density of base stations of the system, we evaluate the convergence of the algorithm when the number of base stations is 16. Fig. 5 shows the convergence of Q-learning, DQN and Dueling DQN when the number of base stations is 16. It can be seen that all three algorithms can converge. Both Dueling DQN and Q-learning converge to the optimal when the number of iterations is 750. However, it can be clearly seen that the optimal value obtained by convergence of Dueling DQN is far greater than Q-learning. For DQN and Dueling, we can see that the convergence rate of DQN is similar to that of Dueling DQN. DQN converges at 500 iterations. However, at this time, the weighted energy spectrum efficiency calculated by Dueling DQN has been better than that of DQN, and has not reached the optimal level yet. Therefore, when Dueling DQN converges, the optimized total weighted energy spectrum efficiency of the system is superior to that of DQN. Finally, the convergence performance of Q-learning, DQN and Dueling DQN is integrated, and Dueling DQN is obviously better than the other two algorithms.

V. Conclusions

In this paper, we investigated the dynamic resource allocation of 5G UDN. We considered both the EE and SE of the network. Therefore, the resource allocation problem at different moments was expressed as a joint optimization problem. In order to ensure the long-term performance of UDN system, the joint optimization problem was described as an MDP. The Dueling DQN method was proposed. Simulation results showed that the Dueling DQN algorithm has the best performance. Compared with classical Q-learning

algorithm and ordinary DQN algorithm, Dueling DQN can greatly improve network performance. In the future research, we will consider power allocation and multi-resource joint configuration in 5G UDN.

ACKNOWLEDGMENT

This work is partly supported by the National Natural Science Foundation of China (Nos. 61872044, 61502040), Beijing Municipal Program for Excellent Teacher Promotion (no. PXM2017_014224.000028), The Supplementary and Supportive Project for Teachers at Beijing Information Science and Technology University (No. 5111823401), The Key Research and Cultivation Projects at Beijing Information Science and Technology University (No.5211910958), Beijing Municipal Program for Top Talent Cultivation (CIT&TCD201804055), Open Program of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (ICDDXN001), Qinxin Talent Program of Beijing Information Science and Technology University.

REFERENCES

[1] Y. Li, Y. Zhang, K. Luo, T. Jiang, Z. Li and W.Peng, "Ultra-Dense HetNets Meet Big Data: Green Frameworks, Techniques, and Approaches," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 56-63, Jun. 2018.

[2] A. Gupta, and R.K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," *IEEE Access*, vol. 3, pp. 1206-1232, Jul. 2015.

[3] P. Gandotra, R.K. Jha and S. Jain, "Green Communication in Next Generation Cellular Networks: A Survey," *IEEE Access*, vol. 5, pp. 11727-11758, Jun. 2017.

[4] T. Qiu, J. Liu, W. Si and D.O. Wu, "Robustness optimization scheme with multi-population co-evolution for scale-free wireless sensor networks," *IEEE/ACM Transactions on Networking*, pp. 1-15, Apr. 2019.

[5] T. Qiu, R. Qiao and D.O. Wu, "EABS: An Event-Aware Backpressure Scheduling Scheme for Emergency Internet of Things," *IEEE Transactions on Mobile Computing*, vol. 17, no. 1, pp. 72-84, Jan. 2018.

[6] J. Chen, K. Hu, Q. Wang, et al. 'Narrowband internet of things: Implementations and applications," *IEEE Internet of Things Journal*, vol.4, no. 6: 2309-2314, Dec. 2017.

[7] M. Kamel, W. Hamouda and A. Youssef, "Ultra-Dense Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2522-2545, Fourthquarter 2016.

[8] Y. Teng, M. Liu, F.R. Yu, V.C.M. Leung, M. Song, and Y. Zhang, "Resource Allocation for Ultra-Dense Networks: A Survey, Some Research Issues and Challenges," *IEEE Communications Surveys & Tutorials*, pp. 1-1, Aug. 2018.

[9] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5G: challenges, methodologies, and directions," *IEEE Wireless Communications*, vol. 23, no. 2, pp. 78-85, Apr. 2016.

[10] J. Zhao, Y. Liu, K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Resource Allocation for Non-Orthogonal Multiple Access in Heterogeneous Networks," in *Proc. IEEE International Conference on Communications(ICC)*, Paris, May. 2017, pp. 1-6.

[11] S. Wu, Z. Zeng, and H. Xia, "Load-Aware Energy Efficiency Optimization in Dense Small Cell Networks," *IEEE Communications Letters*, vol. 21, no. 2, pp. 366-369, Feb. 2017.

[12] C. C. Coskun, and E. Ayanoglu, "Energy-spectral efficiency tradeoff for heterogeneous networks with QoS constraints," in *Proc. IEEE International Conference on Communications(ICC)*, Paris, May. 2017, pp. 1-7.

[13] S. Xu, R. Li and Q. Yang, "Improved Genetic Algorithm Based Intelligent Resource Allocation in 5G Ultra Dense Networks," *IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, Jun. 2018, pp. 1-6.

[14] Y. Wei, F. R. Yu, M. Song and Z. Han, "User Scheduling and Resource Allocation in HetNets With Hybrid Energy Supply: An Actor-Critic Reinforcement Learning Approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 680-692, Jan. 2018.

[15] A. Asheralieva, "Bayesian reinforcement learning-based coalition formation for distributed resource sharing by device-to-device users in heterogeneous cellular networks", *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5016-5032, Aug. 2017.

[16] H. Li, H. Gao, T. Lv and Y. Lu, "Deep Q-Learning Based Dynamic Resource Allocation for Self-Powered Ultra-Dense Networks," *IEEE International Conference on Communications Workshops (ICC Workshops)*, Kansas City, MO, May. 2018, pp. 1-6.

[17] H. Ye, G. Y. Li and B. F. Juang, "Deep Reinforcement Learning Based Resource Allocation for V2V Communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3163-3173, Apr. 2019.

[18] J. Li, H. Gao, T. Lv and Y. Lu, "Deep reinforcement learning based computation offloading and resource allocation for MEC," *IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, Apr. 2018, pp. 1-6.

[19] X. Chen, Z. Li, K. Wang, and L. Xing, "MDP-Based Network Selection with Reward Optimization in HetNets," *Chinese Journal of Electronics*, Vol.27, No.1, pp. 183-190, Jan. 2018.

[20] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling network architectures for deep reinforcement learning," *arXiv preprint arXiv:1511.06581*, 2015.