# NOMA Resource Allocation Method Based on Prioritized Dueling DQN -DDPG Network

**Yuan Liu**
Heilongjiang University
**Yue Li** (2017021@hlju.edu.cn)
Heilongjiang University https://orcid.org/0000-0002-8880-9773
**Lin Li**
Heilongjiang University
**Mengli He**
Heilongjiang University

**Research Article**

**Keywords:** Non-orthogonal Multiple Access (NOMA), Resource Allocation, Dueling DQN, Prioritized Sampling, Depth Deterministic Policy Gradient (DDPG)

**Posted Date:** December 13th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-2341741/v1

# NOMA Resource Allocation Method Based on Prioritized Dueling DQN -DDPG Network

Yuan Liu, Yue Li *,Lin Li, Mengli He
Electronic Engineering School, Heilongjiang University, Harbin, 150001, China

## Abstract

In the mobile communication system, non-orthogonal multiple access (NOMA) technology is introduced to improve spectrum efficiency. Because the combination of users and the transmission power of each user are very important to the performance of NOMA system, the resource allocation technology of NOMA system has been widely studied. In recent years, scholars have introduced deep reinforcement learning network for user grouping and power allocation, which can effectively reduce the computational complexity and improve the system sum rate. However, the traditional algorithm based on DQN network still has the problems of slow convergence speed and low training stability, and the uniform sampling method in the sample playback process has the problem of low sampling efficiency. To address these problems, this paper proposes a priority-based user grouping and power allocation method of NOMA system optimized by Dueling DQN-DDPG, which can effectively improve the convergence speed and training stability. Firstly, in the user grouping stage, a user grouping network based on Dueling DQN is proposed. This network considers both the state value and the action value in the whole connection layer. The two values compete with each other, and then they are summed up and re-evaluated. The proposed network can effectively improve the stability of the traditional DQN network training process and speed up the training convergence. Secondly, considering the continuity of power value, DDPG network, which is suitable for dealing with continuous action space, is adopted in the power allocation stage, which can avoid the power quantization error. Finally, the priority sampling based on TD-error is combined with Dueling DQN network and DDPG network respectively, which can ensure random sampling and improve the replay probability of important samples. Simulation results show that the priority based Dueling DQN -DDPG algorithm proposed in this paper can greatly improve the convergence speed of sample training. At the same time, this scheme has the advantage of priority sampling, which can improve the learning speed and make the training process more stable. Compared with the traditional DQN algorithm, the convergence speed of the proposed algorithm is nearly doubled, and the training process is more stable, but the computational complexity is only increased by about 15%.

## Keywords

Non-orthogonal Multiple Access (NOMA), Resource Allocation, Dueling DQN, Prioritized Sampling, Depth Deterministic Policy Gradient (DDPG).

# 1.Introduction

With the commercialization of 5G network and the continuous development of 6G technology, the requirements of communication quality in various industries are increasing. Mobile communication devices need to provide higher data rate, lower communication delay and better reliability. The traditional Orthogonal Multiple Access (OMA) technology cannot meet the current communication needs, and the Non-Orthogonal Multiple Access (NOMA) technology has become an important part of the new generation communication technology development. NOMA is mainly classified into two types: power domain multiplexing and code domain multiplexing. The main principle of power domain multiplexing is to allocate power to different users at the transmitter according to the real-time Channel State Information (CSI) of users. Then the user information is superimposed on the same time-frequency resource block by Superposition Coding (SC) technology. At the receiving end, the Successive Interference Cancellation technology is used to detect multi-users in a certain order from the received superimposed signals, correctly demodulate signals to eliminate the interference, and finally recover the required information. At the transmitting end of the base station, different signal powers will be allocated to different users, so as to obtain the maximum performance gain of the system and achieve the purpose of distinguishing users. NOMA technology based on power reuse can effectively improve spectrum utilization, and provide higher transmission rate, lower delay and better transmission reliability[1]~[3].

In recent years, many researchers have devoted themselves to the design and implementation of NOMA technology, and proved the compatibility of power domain NOMA with cooperative communication, relay and MIMO. The problems of user grouping, power allocation and spectrum resource allocation for NOMA have also attracted extensive attention. The system sum rate can be greatly improved, and the accuracy and stability of the system can be improved by using an efficient scheme to group and assign power to the users at the transmitter. Reference [4] pointed out that for a given set of scheduled users, the classical iterative water injection power allocation algorithm can achieve the maximum weighted sum of user throughput. Reference [5] studied the user pairing problem of NOMA system based on fixed power allocation, discussed the influence of user pairing on the sum rate, studied the power allocation scheme of two users pairing and analyzed its performance. In reference [6] and reference [7], the authors considered sub-channel allocation and power allocation jointly, but this joint resource allocation problem is usually NP-hard, and it is difficult to obtain an optimal solution with conventional optimization methods.

Conventional methods rely on system modeling, and the computational complexity is high. In contrast, deep learning is a powerful tool to solve complex mathematical problems, which shows great advantages. There have been many studies on the combination of NOMA technology and deep learning. In reference [8], considering the user fairness of NOMA, Deep Neural Networks (DNN) are used for decoding. Compared with traditional algorithms, DL can effectively reduce the computational complexity, so as to efficiently achieve fairness and finally maximize the system sum rate. Reference [9] uses the Attention-Based Neural Network (ANN) to allocate channels to users in NOMA system. Compared with the traditional random allocation and exhaustive search calculation methods, the introduction of neural networks can effectively improve the total throughput of the system and reduce the computational complexity. Reference [10] trains DNN to simulate the interior point algorithm for power allocation, the introduction of neural networks can improve computational efficiency. Through the combination of deep learning and reinforcement learning, Deep Reinforcement Learning (DRL) can make full use of the perceptual advantages of deep learning and the decision-making advantages of reinforcement learning, and directly control strategies from high-dimensional raw data to provide faster convergence speed, which is more effective for multi-state and action-space systems. Reference [11] proposes a Deep Q-Network (DQN), which is used as an approximator in many fields. Reference [12] proposes a DRL based resource allocation scheme, which formulates the joint channel allocation and user grouping problem as an optimization problem. Compared with other methods, the proposed framework can achieve better system performance. DQN is currently a more commonly used deep reinforcement learning network, which is widely used in the resource allocation of NOMA system, and effectively solves the problem of high complexity of resource allocation of traditional NOMA system. However, when using the traditional DQN network to train the samples, the training convergence speed is slow and the training process is unstable. Reference [13] provides an improved network based on DQN, the Dueling DQN, whose core idea is to decompose the state value $Q_{\pi}(s_t,a_{t1})$ into a state value function $V(s_t)$ and an action advantage function $A(s_t,a_{tl})$ within the neural network. The state value and advantage functions form a competitive network, which can effectively improve the instability of the traditional DQN training process and speed up the training convergence. Based on this, this paper proposes Dueling DQN in the resource allocation of NOMA system, which not only solves the problem of high complexity of traditional algorithms in resource allocation, but also solves the problems of low convergence speed and unstable training process of traditional DQN.

Since the output of DQN and Dueling DQN can only be discrete, if we use Dueling DQN to complete power allocation task, the continuous user power needs to be quantized, and quantization will bring quantization error. Deep Deterministic Policy Gradient (DDPG) networks can solve this problem[14]. This paper uses the Actor-Critic algorithm to solve the power allocation optimization problem in NOMA systems. The Actor-Critic algorithm is used to dynamically select the power allocation coefficient, and a parameterized policy is constructed from the Actor network part, which is evaluated by the Critic network. Finally, the Actor network adjusts the power allocation policy according to the feedback of the Critic network part.

In addition, empirical replay algorithm is used in Dueling DQN and DDPG network to reduce the correlation between samples and ensure the independent and identically distributed characteristics between samples. However, the current sampling method is uniform sampling, which ignores the importance of samples. In the sampling process, some valuable samples may not be learned, thus reducing the learning rate. The prioritized sampling method based on TD-error can improve the replay probability of important samples[15]. Therefore, this paper proposes priority sampling based Dueling DQN and DDPG network to speed up the convergence of training.

Aiming at maximizing the system sum rate in NOMA resource allocation problem, this paper proposes a joint optimal scheme based on Prioritized Dueling DQN-DDPG network, where Dueling DQN performs discrete tasks to complete user grouping, and DDPG network performs continuous tasks to allocate power to each user. On this basis, this paper proposes a prioritized sampling method based on TD-error to improve sampling efficiency and learning rate.

## 2. System model

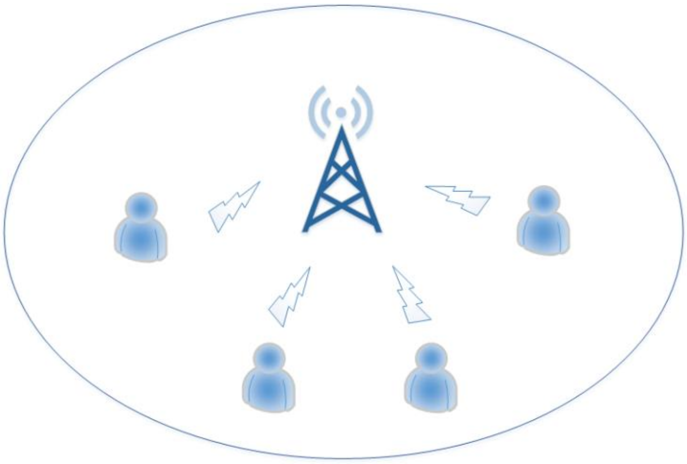


Figure 2-1 Transmission Model of NOMA Uplink System

Figure 2-1 shows the transmission model of NOMA uplink system. In this paper, we study the uplink multi-user NOMA system scenario where the Base Station (BS) is located in the center of the cell and the users are randomly distributed near the base station. We need to solve the problems of user grouping and power allocation in the cell by maximizing the system sum rate. Assuming that the number of users per cell is $K$, the users are randomly distributed in various locations in the cell, and the base station and the users are single-antenna configured. Channel decay follows the Rayleigh distribution, with $\boldsymbol{z_n}$ representing the additive Gaussian white noise with a variance of $\boldsymbol{\delta_n^2}$.The total bandwidth of the system $B$ is evenly distributed among $N$ sub-channels, and users in the same sub-channel are non-orthogonal, and the bandwidth of each sub-channel is $B_s=B/N$. Since multiple users in a NOMA system can reuse the same resource block, the maximum number of users on each sub-channel is set to $M$. The power allocated to the user $m$ on the $n$ sub-channel is $P_{m,n}$ , $S_{m,n}$ is the allocation index of the sub-channel, and when user $m$ is assigned to sub-channel $n$, then $S_{m,n}$ =1, otherwise $S_{m,n}$ =0. Then the signal sent on the $n$th sub-channel is:

$$x_n = \sum_{i=1}^{M} b_{m,n} \sqrt{P_{m,n}} S_{m,n} \quad (1)$$

$g_{m,n}$ is the channel gain of user *m* on the sub-channel *n*. Then at the base station , the expression of the received signal is:

$$y_n = g_{m,n} b_{m,n} \sqrt{P_{m,n}} S_{m,n} + \sum_{i=1, i \neq m}^{M} g_{i,n} b_{i,n} \sqrt{p_{i,n}} S_{i,n} + z_n \quad (2)$$

In NOMA systems, due to interference introduced by the superimposed user, SIC technology is usually used at the receiving end, and the base station will receive multiple different superimposed signals and demodulate them in a certain order. The receiver first demodulates the high-power signal, subtracts it from the mixed signal, and treats the rest as interference. Thus, for users in sub-channel *n*, the *SINR* can be expressed as:

$$SINR = \frac{b_{m,n} P_{m,n} \left| g_{m,n} \right|^2}{\delta_n^2 + \sum_{i=1, \left| g_{i,n} \right|^2 < \left| g_{m,n} \right|^2}^{M} b_{m,n} P_{i,m} \left| g_{i,n} \right|^2} \quad (3)$$

According to Shannon's theorem, the rate of the *m*th user on the sub-channel *n* is:

$$R_{m,n} = B_s \log(1 + SINR) \quad (4)$$

The sum rate of the corresponding sub-channel *n* is:

$$R_n = \sum_{i=1}^{M} R_{m,n} \quad (5)$$

The system sum rate is:

$$R = \sum_{j=1}^{N} R_n = \sum_{i=1}^{M} \sum_{j=1}^{N} R_{m,n} \quad (6)$$

In this paper, the problem is to maximize the system sum rate under the constraints of each user meeting the minimum transmission rate requirements. Optimization problems can be modeled as:

$$\max \sum_{i=1}^{M} \sum_{j=1}^{N} R_{m,n} \quad (7)$$

The constraints of the joint user grouping and power allocation are as follows:

$$\mathrm{C1}: 0 \leq P_{m,n} \leq P_{\max} \quad (8)$$

$$\mathrm{C2}: R_{m,n} \geq R_{\min} \quad (9)$$

where $P_{\max}$ is the maximum transmit power of the user and $R_{\min}$ is the minimum data rate of the user. Constraint C1 ensures that the transmit power per user does not exceed $P_{\max}$. Constraint C2 guarantees that the rate per user is not less than the minimum signal rate. It is difficult to find a globally optimal solution for this objective function. Although the global search method can provide the optimal solution by searching all grouping possibilities, the computational complexity is too high to be applied in practice. Therefore, the predecessors utilized DRL to reduce the complexity of the calculation[16]~[17].On this basis, this article proposes a method based on the joint optimization of Prioritized Dueling DQN-DDPG for user grouping and power allocation in NOMA system.The proposed method can increase the system sum rate, improve learning efficiency, and solve the problems of slow convergence speed and unstable training.

# 3. Resource allocation method based on Prioritized Dueling DQN-DDPG

## 3.1 Resource allocation network architecture

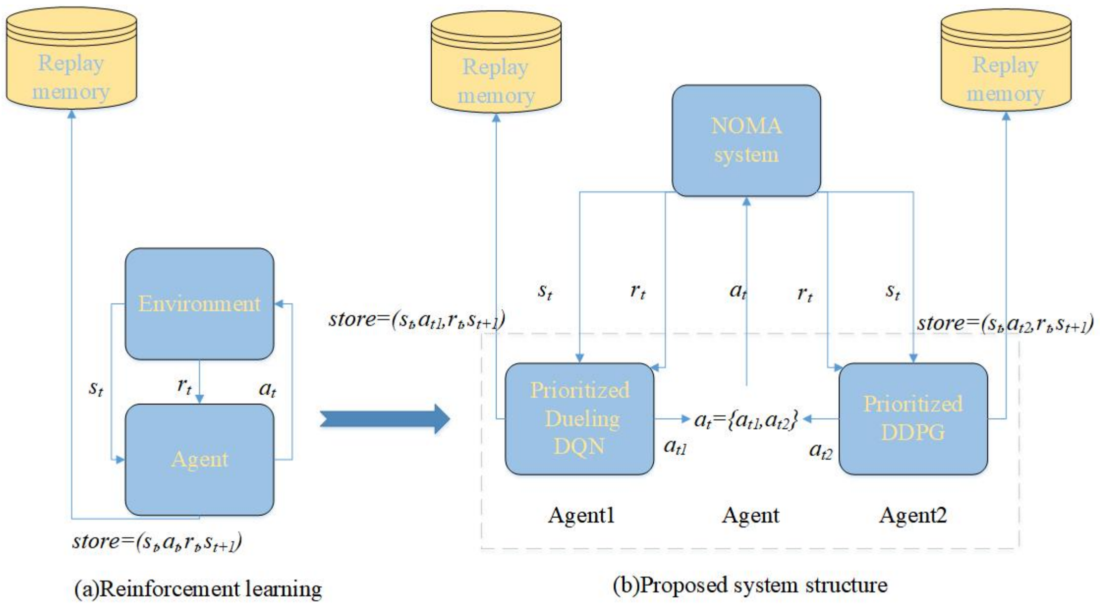



Figure 3-1 Resource allocation network based on deep reinforcement learning

General reinforcement learning is mainly composed of five parts: Agent, Action, State, Reward, and Environment. Agent represents an agent that makes a corresponding Action based on the input State, and the Environment receives the Action and returns the State and Reward. The agent updates the decision function that produces the action based on the reward. This process is repeated until the Agent can make the optimal Action in any State, that is, the model learning process is completed. The key point of reinforcement learning is that the state, action and return should be one-to-one corresponding to the NOMA system parameters studied, so that the reinforcement learning method can achieve the desired effect[18].

According to the structure of reinforcement learning, this paper designs the NOMA system model, as shown in Figure 3-1. NOMA stands for reinforcement learning environment with two agents: One is the Prioritized Dueling DQN, which is responsible for user grouping; the other is the Prioritized DDPG network, which performs power allocation. In this paper, the state space is defined as $S=\{g_{m,1},g_{m,2},\ldots,g_{m,n}\}$, the user grouping space is defined as $A1=\{b_{1,1},b_{2,1},\ldots,b_{m,n}\}$, and the power allocation space is defined as $A2=\{p_{1,1},p_{2,1},\ldots,p_{m,n}\}$. Instant rewards are denoted by $r_t=R$, where $R$ is the optimization target system sum rate, and $R_t$ is used to represent the sum of the rewards and rewards obtained[19].

$$R_t=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}\ldots=\sum_{i=0}^{\infty}\gamma^i r_{t+i},\gamma\in[0,1] \tag{10}$$

$\gamma$ is the discount factor, indicating the importance of future rewards and immediate rewards. The value of $\gamma$ ranges from 0 to 1. The expected value of the cumulative pay off $R_t$ is defined as the Q value, which is determined by the state $s_t$.The choice of actions under certain strategies. It is expressed as:

$$Q_\pi(s_t, a_t) = E[r_t + \gamma \max Q_\pi(s_{t+1}, a_{t+1})] \tag{11}$$

At each Time Slot (TS), Agent1 and Agent2 obtain the channel gain from the NOMA system, select the user combination and power in the action space according to the current channel gain, and return the action result to the NOMA system. Based on the received action, the NOMA system generates instant rewards and channel gains for the next TS, which are then passed to Agent1 and Agent2, respectively. Based on the reward, Agent1 and Agent2 update the decision function that selects the action under the current channel gain to complete the interaction. This process is repeated until the Agent can generate the best decision at any channel gain[20]. For the DQN user grouping scheme proposed by predecessors, there are problems such as slow convergence speed and unstable training, which lead to system performance loss. In order to solve this problem, uplink is improved in this paper. The NOMA system of user grouping and power allocation joint optimization based on Prioritized Dueling DQN-DDPG is shown in Figure 3-1.

### 3.2 User grouping based on Dueling DQN

This paper uses Prioritized Dueling DQN to complete the user grouping task. DQN is one of the deep reinforcement learning algorithms. It combines neural network with Q learning algorithm, uses the powerful representation ability of neural network, takes input record as the state in reinforcement learning, and serves as the input of neural network model (Agent).Then the neural network model outputs the corresponding value (Q) of each action to get the action to be executed. However, in many deep reinforcement learning tasks, the value functions corresponding to actions in different states are not the same, or in some states, the value functions are unrelated to actions. According to the above ideas, Wang et al.[13] proposed the Dueling network model to replace the network model in the DQN. The core idea of Dueling DQN is to decompose the state value $Q_\pi(s_t,a_{t1})$ into the state value function $V(s_t)$ and the action advantage function $A(s_t,a_{t1})$. In this paper, Dueling DQN is applied to the user grouping stage of NOMA system. The main idea is that Dueling DQN considers different state values and advantage functions in different states, which can quickly select the current optimal action in the sample training process.

#### 3.2.1 Dueling DQN based user grouping network

This section introduces the user grouping framework base on Dueling DQN, in NOMA system. As shown in Figure 3-2, Dueling DQN contains two sub-networks, Q-network and target Q-network. Q-network is used to generate the estimated Q value of the selected action, and target Q-network is used to generate the target Q value of the training neural network. In the NOMA system, the current environment is first initialized to obtain the initial state $s_t$, which is fed into the estimated Q-network of the Dueling DQN. Taking $s_t$, as input, this paper adopts the $\varepsilon$ -greedy strategy to select $a_{t1}$ as new user combination, namely:

$$a_{t1} = \arg\max_{a_{t1} \in A1} \left(s_t, a_t; \theta, \beta, \alpha\right) \tag{12}$$

This means that the $\zeta$ probability is to randomly select the action from the action space A1 as the user combination, or the user combination with the highest estimated Q value with a probability of (1-$\varepsilon$). Finally, all user combinations $a_{t1}$ and power $a_{t2}$ (setting the power allocation action to $a_{t2}$) are returned to the NOMA system. Based on the chosen action, the NOMA system generates the immediate reward and the status information $s_{t+1}$ at the next moment, which is then stored in memory, $\left(s_t, a_{t1}, r_t, s_{t+1}\right)$.To ensure that all samples in the sample pool can be sampled, we set the new sample as the highest priority and store this sample tuple in the experience pool. We use the sampling probability to calculate the sample weight, and train the target Q value in the network to be generated using the Q-network, namely:

$$y_i = r_i + \gamma \max_{a_{(i+1)1} \in A1} Q_\pi\left(s_{i+1}, a_{(i+1)1}; \theta^-, \beta, \alpha\right) \tag{13}$$

The purpose of the training process is to make the prediction error between the estimated Q value and the real Q value infinitely close to 0. Therefore, in this paper, the prediction error is defined as a loss function, namely:

$$LOSS = \frac{1}{N}\sum_{i=1}^{N} w_i \left(y_i - Q\left(s_i, a_{i1}; \theta, \beta, \alpha\right)\right)^2 \tag{14}$$

Finally, the loss function is used to update and estimate the weights of the Q-network. Then, after a certain number of iterations, the weight parameters of the target Q-network are updated with the weight parameters of the estimated network. Where $w_i$ is the importance sampling weight of the sample[21]~[23].

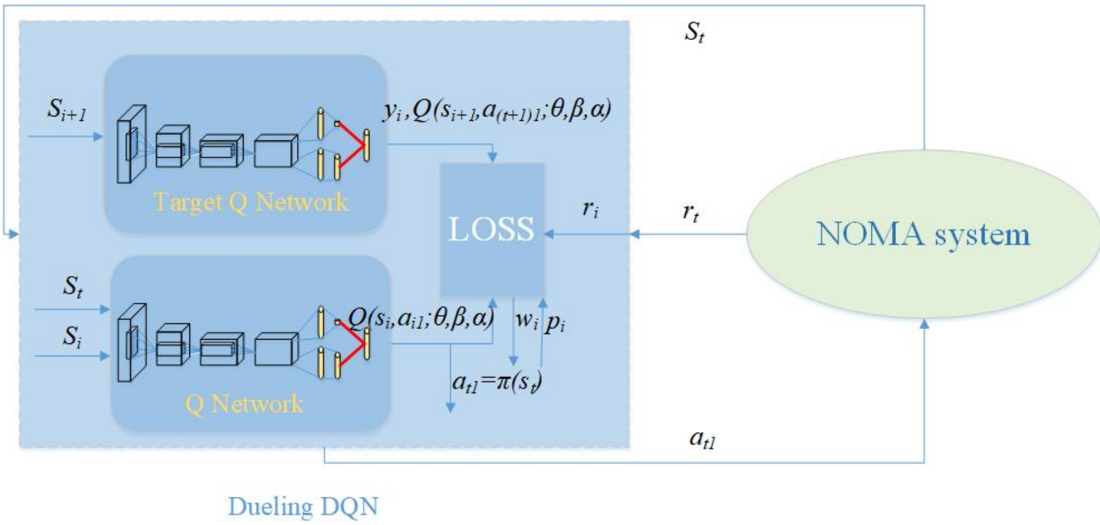



Figure 3-2 User grouping framework based on Dueling DQN

Algorithm 1：User grouping algorithm based on Dueling DQN

(1) Initialize the memory $D$, store the maximum value of the experience sample to $N$, and the weight update interval $W$. Initialize the prediction Q-network and weight $\theta$ of all Dueling DQN units, the target Q-network and weight $\theta^{-}=\theta$, and initialize parameters $\beta$ and $\alpha$.

(2) Initialize state $s_1$, action $a_{t1}$ and ambient noise $z_n$。

Repeat The time step in the empirical trajectory, from $t$=1 to T。

(3) The Dueling DQN network chooses action $a_{t1} \in A1$ according to the $\varepsilon$ - greedy strategy, and otherwise chooses $a_{t1} = \arg\max_{a_{t1}\in A1}(s_t, a_{t1}; \theta, \beta, \alpha)$, and get the return reward $r_t$ and the next state $s_{t+1}$.

(4) Save the $(s_t, a_{t1}, r_t, s_{t+1})$ to the memory.

(5) Sample data $(s_t, a_{t1}, r_t, s_{t+1})$ by priority size from the memory.

(6) The target value of each state is calculated, and the value of Q is updated by the reward $r_t$ after the action is performed by the target network Q.

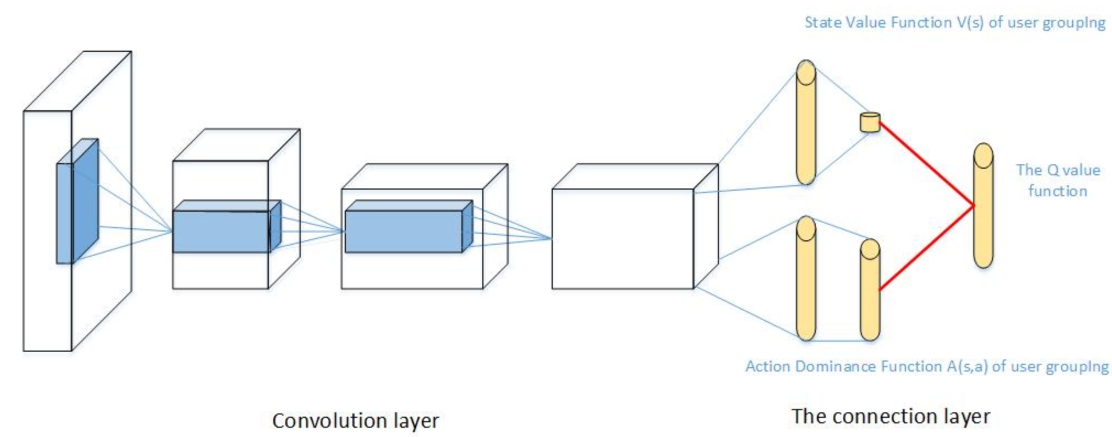
(7) The weight parameter $\theta$ of Dueling DQN is updated by minimizing the loss function formula.

(8) Every $W$ interval, update the weight $\theta^{-}$ of the target network with the prediction network weight $\theta$.

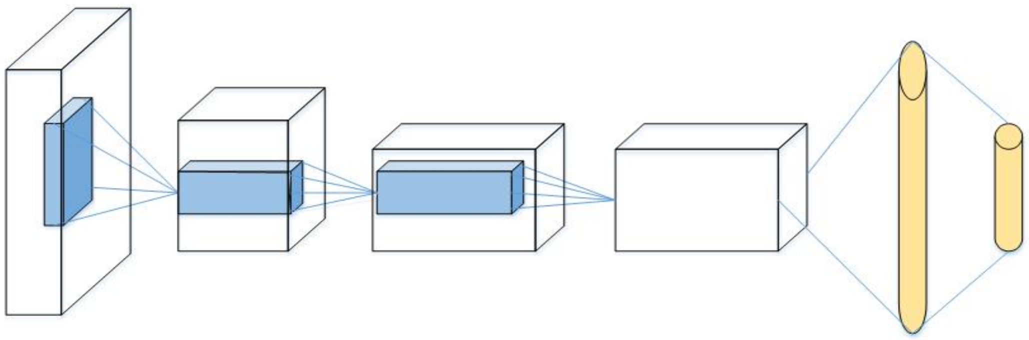
(9) END

### 3.2.2 Dueling DQN network structure

The architecture of the Dueling DQN model in the user grouping algorithm is shown in Figure 3-3 (a). For comparison, the traditional DQN model architecture is given in Figure 3-3(b). Compared with DQN, Dueling DQN first divides the fully connected layer into two branches. The first path is the output state value($V(s_t)$), which represents the value of the static state environment itself. The next path outputs the action advantage value($A(s_t,a_{t1})$), which represents the additional value of selecting an action. Finally, through full connection, it is merged into the action value $Q_\pi(s_t,a_{t1})$. The state value function is unrelated to the action. In contrast, the action advantage function is related to the action, and it is the average reported degree of goodness of the action, which is related to the state, and can solve the Reward-bias problem. Based on the above competing network structure, the agent can finally learn a more realistic value $V(s_t)$ in the environmental state without the influence of action[13].




(a) Dueling DQN network structure



(b)DQN network structure

Figure 3-3 Comparison of Dueling DQN and DQN network structures

In this paper, the state value function $V(s_t)$ of Dueling DQN in user grouping is expressed as:

$$V\left(s_t\right) \cong V\left(s_t;\theta,\beta\right) \tag{15}$$

Action advantage function $A(s_t,a_{t1})$ can be expressed as:

$$A(s_t,a_{t1}) \cong A(s_t,a_{t1};\theta,\alpha) \tag{16}$$

Where $\theta$ is the convolution layer parameter; $\beta$ and $\alpha$ are the fully connected layer parameters of the two branches.

In practice, action dominance is generally set as a separate action dominance function minus the average of all action advantage functions in a certain state. Therefore, the final action Q value of the user grouping in this paper is expressed as:

$$Q_\pi(s_t,a_{t1};\theta,\beta,\alpha) = V\left(s_t;\theta,\beta\right) + \left(A(s_t,a_{t1};\theta,\alpha) - \frac{1}{|A|}\sum_{a_{t1}'} A(s_t,a_{t1}';\theta,\alpha)\right) \tag{17}$$

The advantage of this expression is that it can ensure that the relative ranking of the dominant functions of each action in this state is stable, and can reduce the range of Q value, remove the excess degrees of freedom, and then improve the stability of the algorithm. Compared with the traditional DQN network structure, Dueling DQN decomposes the Q value into the form of value function $V(s_t)$ and advantage function $A(s_t,a_{t1})$, which makes training easier and converges faster. As the number of actions increases, the advantage becomes more obvious. The state value function depends only on the state and is independent of the behavior, so it's easier to train; In the same state, multiple behaviors can share the same value $V(s_t)$. The difference between different behaviors is only in the dominance function. The convergence of this part can also be independent of the value function, so that the relative differences between behaviors can be learned independently. Moreover, the advantage function is introduced to avoid the problem of unstable results caused by the large magnitude of Q values and the very small difference between Q values.

### 3.3 Power allocation based on DDPG network

Deep reinforcement learning methods such as DQN and Dueling DQN use deep neural networks to approximate Q-valued functions, which can effectively solve complex problems with high dimensions of state space and action space. But it is only suitable for dealing with discrete action spaces. This is because DQN needs to find the action with the largest Q value, and if the action is an infinite number of consecutive values, then iterative optimization needs to be performed within TS in a performance penalty. Therefore, DQN cannot be directly applied to the continuous action space. DDPG is a model-free, off-line learning method based on deterministic policy gradients. It follows the Actor-Critic architecture and can effectively deal with problems with continuous action Spaces by using a deep neural network approximation strategy. Wang et al.[24] proposed two frameworks (i.e., DDRA and CDRA) to maximize the energy efficiency of NOMA systems, where DDRA is based on DDPG networks and CDRA is based on multiple DQN[13]. The results show that the time complexity of the two frameworks is similar, but the performance of DDPG network is better than that of DQN network. This is because in multi-DQN, the user power is quantized, resulting in the loss of some important information. DDPG network is similar to DQN, using deep neural networks and uniform sampling. It is also a deterministic policy gradient network where behavior is uniquely determined in one state. Moreover, DDPG can handle sequential action tasks without quantifying the transmission power. Therefore, in this section, the power allocation network based on DDPG is designed on the basis of sub-channel assignment in Dueling DQN. DDPG can be easily extended to larger and more complex mobile communication systems. Compared with the discrete method, the continuous resource allocation method proposed in this chapter can achieve better system sum rate, and has stronger processing power for large-scale user access[25]~[28]. Figure 3-4 shows the network structure of DDPG.

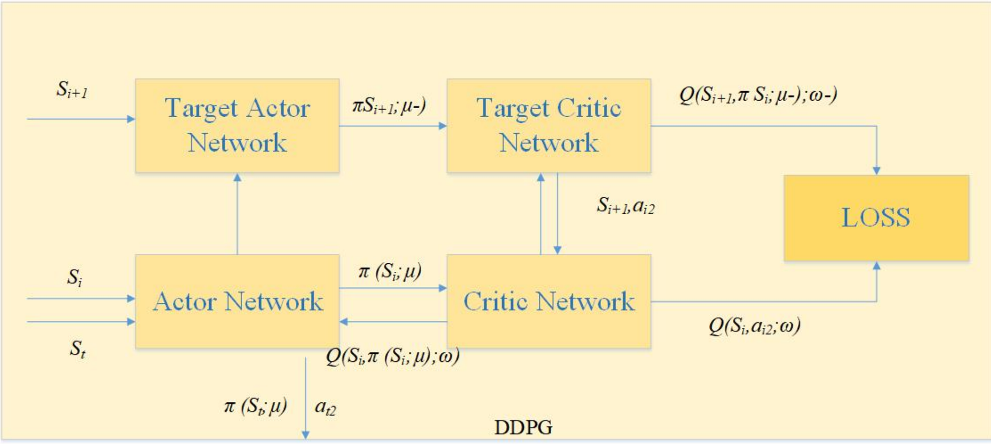



Figure 3-4 DDPG network structure

### 3.4 Priority experience playback mechanism

The priority experience playback mechanism is not randomly sampling, but sampling according to the importance of each sample in the experience pool, which can find the samples required for training more effectively[15]. In priority experience playback, the Temporal-difference error (TD-error) of each sample is used as the evaluation criterion for sampling, and the TD-error formula for the samples in the user grouping is as follows:

$$\delta_i = y_i - Q\left(s_i, a_{i1}; \theta, \beta, \alpha\right) \tag{18}$$

Where $\delta_i$ is the TD-error of sample $i$. If the absolute value of TD-error of a sample is larger, its probability of being sampled is higher. The TD-error of a sample determines the probability of being sampled. The priority sampling probability of samples can be expressed as:

$$P\left(i\right) = \frac{P_i^k}{\sum jP_j^k} \tag{19}$$

where $P_i$ represents the priority of the sample, it is calculated according to the TD-error of the sample, $P_i=|\delta_i|+\varepsilon_0$. $P_i>0$, $\varepsilon_0>0$。By setting the priority of the samples, samples with high probability will be added to the learning process frequently and samples with small TD-error may never be trained. In order to ensure that samples with lower priority can also be drawn as training samples, it is assumed that $\varepsilon_0$ is a positive value to ensure that the sample priority is always greater than 0. $k$ determines the degree of priority, when $k$=0, it indicates uniform sampling, and when $k$=1 indicates greedy strategy sampling. $k$ does not change the monotonicity of priority and is used to increase or decrease the priority of TD-error experience.

As the priority experience replay algorithm frequently replayed empirical samples with high TD-error, it resulted in a change in the data distribution of the samples, and the training will be biased or over fitting, in order to reduce the bias, priority experience replay uses the importance sampling weight method to correct the bias. The importance sampling weight of the sample is defined as:

$$w_i = \left( \frac{1}{N} \frac{1}{P(i)} \right)^{\sigma} \tag{20}$$

where $N$ is the number of samples, $P(i)$ is the sample probability, $\sigma$ is used to adjust the degree of deviation, and $\sigma$ =1 indicates that the deviation is completely eliminated.

Figure 3-5 shows the priority-based sampling model.

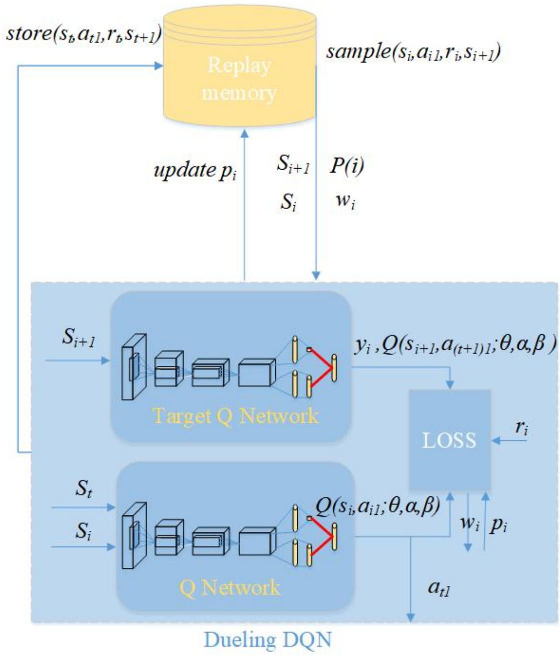



Figure 3-5 Prioritized Dueling DQN on Large clusters

The above mentioned priority based experience replay mechanism is used both in the Dueling DQN and DDPG network. The following figure shows the user grouping and power allocation structure model based on Prioritized Dueling DQN-DDPG.

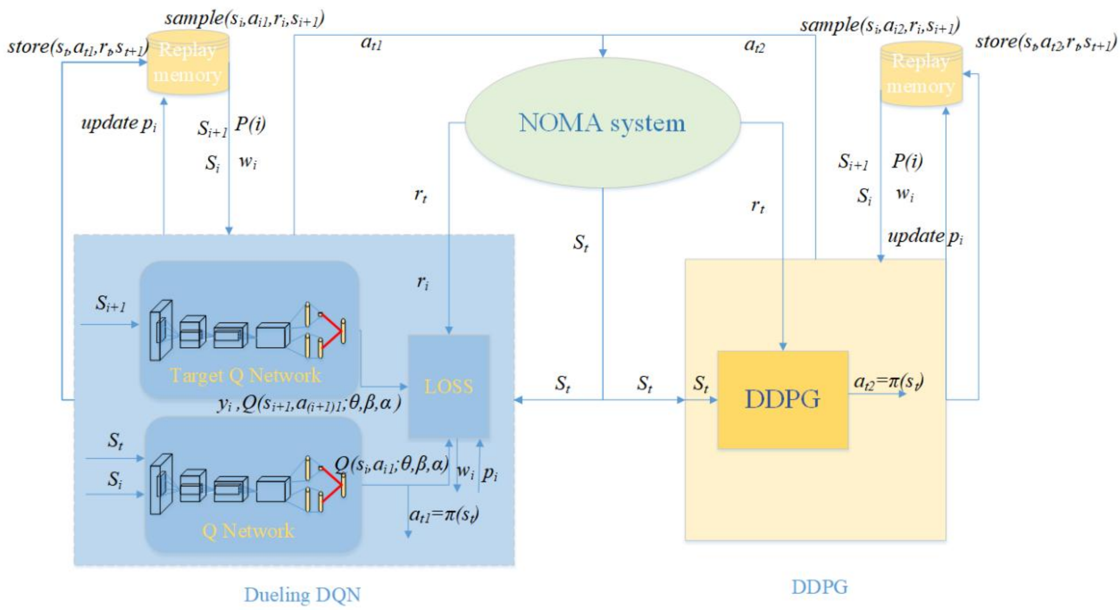



Figure 3-6 User grouping and power allocation of Prioritized Dueling DQN-DDPG

## 4. Results and discussion

In this paper, the performance of the proposed Prioritized Dueling DQN -DDPG resource allocation is simulated in the uplink NOMA system. The base station is located in the center of the cell, and the users are randomly distributed in the cell. Specific parameters are shown in Table 1.

Table 1 Simulation parameter setting

| Parameter | Numerical |
|---|---|
| The number of users | 4 |
| Radius of neighborhood | 500m |
| Path loss factor | 3 |
| Number of samples | 64 |
| Noise power density | -110dBm/Hz |
| The minimum power | 3dBm |
| Total system bandwidth | 10MHz |
| Discount factor $\gamma$ | 0.9 |
| Greedy choice strategy probability $\varsigma$ | 0.9 |
| Algorithm learning rate | 0.001 |

Different learning rates will affect the convergence speed and stability of Dueling DQN training, and this paper first determines that the learning rate of the algorithm is 0.001 through parameter selection. Figure 4-1 shows the convergence of the proposed algorithm at different learning rates.

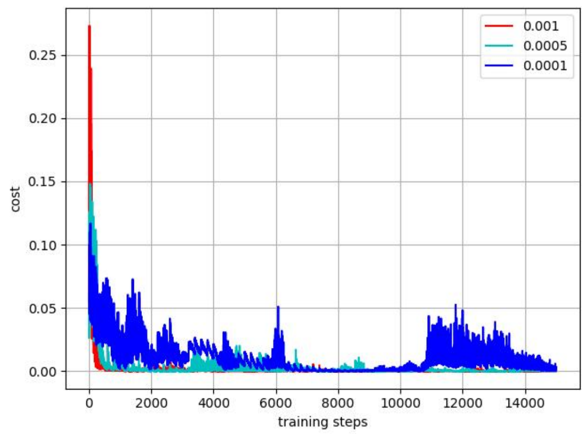



Figure 4-1 Comparison of learning efficiency parameters

In this paper, the NOMA system resource allocation algorithm of Prioritized Dueling DQN and DDPG proposed is denoted as Prioritized Dueling DQN-DDPG. In order to verify the effectiveness of the proposed algorithm, this paper makes a comparison between DQN-DDPG, Dueling DQN-DDPG and Prioritized Dueling DQN-DDPG. In DQN-DDPG method, the user grouping is completed according to DQN and the power allocation is finished according to DDPG. In Dueling DQN-DDPG method, Dueling DQN performs user grouping and DDPG performs power allocation. Prioritized Dueling DQN-DDPG is put forward in this paper, where Prioritized Dueling DQN makes user grouping and Prioritized DDPG makes power allocation. This paper compares the system sum rate performance, training convergence speed, and training stability of the above algorithms. It can be observed that the Prioritized Dueling DQN-DDPG is superior to several other algorithms respectively.

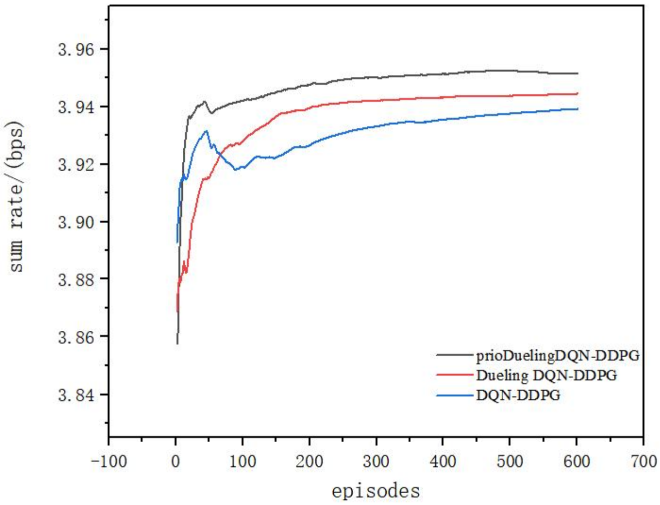



Figure 4-2 Comparison of different algorithm systems sum rate

### 4.1 Convergence of the proposed algorithm

For the convergence performance of the proposed algorithm in this paper, figure 4-2 shows the comparison between the convergence performance of the proposed Prioritized Dueling DQN-DDPG , Dueling DQN-DDPG and DQN-DDPG methods. As the system sum rate gradually increases, the algorithm proposed in this paper is close to convergence when the number of iterations is 50, while the DQN-DDPG tends to converge when the number of iterations is nearly 200. By comparison, the convergence speed of Dueling DQN-DDPG proposed is significantly faster than that of DQN-DDPG, and the convergence speed is more than doubled. It can effectively reduce the training time and make the training process more stable. It can be observed that the convergence speed of the Prioritized Dueling DQN-DDPG is significantly faster than that of the Dueling-DDPG, because the prioritized experience replay stores the prioritized learning experience in the experience pool, and guides the optimization of model parameters by extracting samples with high TD-error, which improves the learning efficiency. In addition, prioritized experience replay not only focuses on samples with high TD-error to help speed up the training process, but also involves samples with low TD-error to increase the diversity of training. Therefore, it is concluded that the convergence speed of the Prioritized Dueling DQN-DDPG has a very significant improvement compared with the Dueling DQN-DDPG.

### 4.2 Average sum rate performance of the proposed algorithm

Figure 4-2 shows the experimental results for system sum rate. All the experimental results are averaged every 600 TS to achieve a smoother and clearer comparison. Prioritized Dueling DQN-DDPG algorithm has obvious advantages over the other two algorithms. Compared with DQN-DDPG algorithm, the proposed algorithm improves the system sum rate by 0.5%.There are two reasons. First, the network structure of Dueling DQN has more advantages than that of DQN, and it can learn more real Q value according to the value function and advantage function. Therefore, the system sum rate of Dueling DQN-DDPG is improved compared with DQN-DDPG. At the same time, Prioritized Dueling DQN-DDPG sets priority for some valuable samples that are beneficial to network training, so as to improve the system sum rate.

### 4.3 Computational complexity analysis

This section analyzes the computational complexity of the proposed algorithm. Based on the computer program runtime (computer configuration: 64-bit operating system, x64-based processor), the time complexity of the Prioritized Dueling DQN-DDPG increases by about 15% compared to the DQN-DDPG. This is because that Dueling DQN divides the output into two parts at the fully connected layer, decomposes the Q value into a value function and a dominance function, and then adds the two parts, so some calculation steps are added when training samples, and the Prioritized algorithm based on timing error is introduced, resulting in an increase in computational complexity. But during the training process, the convergence speed of the algorithm has improved significantly.Table 2 is time complexity comparison of the two methods.Figure 4-3 shows the time complexity of the three methods.

Table 2 Time complexity comparison of the two methods

| Number | DQN-DDPG | Prioritized Dueling DQN-DDPG | Time complexity increased by percentage |
|---|---|---|---|
| 1 | 2355.0431316 Seconds | 2724.8576722 Seconds | 15.703% |
| 2 | 2074.8412441 Seconds | 2376.3507379 Seconds | 14.531% |
| 3 | 2021.6223315 Seconds | 2280.1505739 Seconds | 12.788% |
| 4 | 2042.1567555 Seconds | 2304.2756512 Seconds | 12.835% |
| 5 | 2006.6152422 Seconds | 2290.4872706 Seconds | 14.146% |
| 6 | 2020.9810086 Seconds | 2323.9442205 Seconds | 14.990% |
| 7 | 2031.1689703Seconds | 2305.3912113 Seconds | 13.500% |
| 8 | 2011.4987920 Seconds | 2276.6919056 Seconds | 13.159% |
| 9 | 2103.3444909 Seconds | 2434.7927646 Seconds | 15.758% |
| 10 | 2043.8314553 Seconds | 2370.6670829 Seconds | 15.991% |

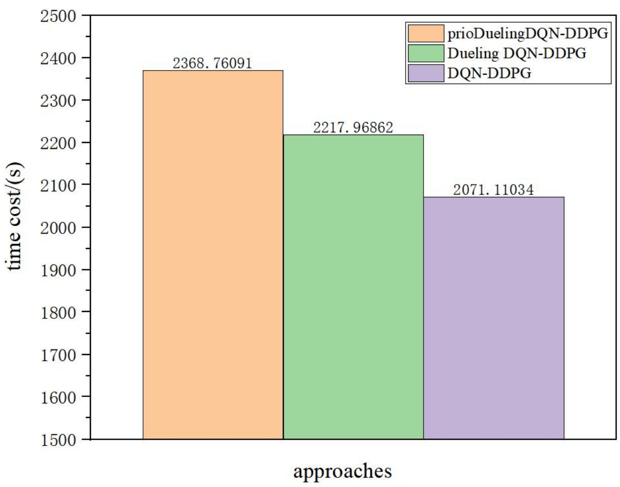



Figure 4-3 Comparison of time complexity

## 5. Conclusion

This paper aims at solving the problems of slow convergence speed and unstable training of DQN under the constraint of ensuring the minimum transmission rate of each user and ensuring the system sum rate maximization. A resource allocation method for NOMA system with Prioritized Dueling DQN-DDPG joint optimization is proposed. Prioritized Dueling DQN is designed with the current channel state information as input and the sum rate as the optimization objective, so that it can output the optimal user grouping policy. The algorithm uses priority experience replay instead of previous randomly distributed experience replay, and uses TD-error to evaluate the importance of samples. Thus, the optimal strategy can be selected more quickly. Simulation results show that when Dueling DQN is used for user grouping, the training convergence speed is significantly accelerated and the training process is relatively stable. The proposed combined priority sampling algorithm can replay valuable samples with high probability, improve the learning rate and make the training more stable. In the power allocation part, the Prioritized DDPG network is used to output the power of all users simultaneously. In addition, compared with the common DQN-DDPG, the convergence speed of the proposed joint algorithm is nearly doubled, and the complexity is only increased by 15%

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

The picture materials quoted in this paper have no copyright requirements, and the source has been indicated.

### Availability of data and materials

Please contact author for data requests.

### Competing interests

The authors declare that they have no competing interests.

### Authors' Contributions

YL* proposed the framework of the whole algorithm; YL performed the simulations, analysis and interpretation of the results. YL, LL and ML have participated in the conception and design of this research, and revised the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

Not applicable.

**Authors' information**

**Affiliations**

Electronic Engineering School, Heilongjiang University, Harbin, 150001, China
`Yuan Liu, Yue Li, Lin Li,Mengli He

**Corresponding author**

Correspondence to Yue Li, Email:2017021@hlju.edu.cn.

## Abbreviations

OMA:Orthogonal multiple access

NOMA: Non-orthogonal multiple access

CSI:Channel State Information

SIC: Successive interference cancellation

BS: Base station

DNN: Deep Neural Networks

ANN:Attention-Based Neural network

RL:reinforcement learning

DRL: Deep reinforcement learning

SC: Superposition coding

TS: Time slot

DQN: Deep Q network

Dueling DQN:Dueling Deep Q network

DDPG: Deep deterministic policy gradient network

TD-error: Temporal-difference error

# References


[1] The 5G mobile communication: the development trends and its emerging key techniques［J］.SCIENTIA-SINICA Information，2014，44（5）：(551-563.).

[2] YU X H, PAN Z W, GAO X Q, et al. Development trend and some key technologies of 5G mobile communication [J]. Science China Information Science, 2014,44 (5) : 551-563.(YOU XH, PAN Z W, GAO X Q, et al).

[3] Goto J, Nakamura O, Yokomakura K, et al. A Frequency Domain Scheduling for Uplink Single Carrier Non-orthogonal Multiple Access with Iterative Interference Cancellation[C].2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall). IEEE, 2014: 1-5.

[4] Sun Y, Ng W K, Ding Z, et al. Optimal Joint Power and Subcarrier Allocation for Full-Duplex Multicarrier Non-Orthogonal Multiple Access Systems[J]. IEEE Transactions on Communications, 2017, 65(3):1077-1091.

[5] LI Xiaoyu, MA Wenping, LUO Lianfei, ZHAO Feifei. Power allocation of NOMA system in Downlink [J]. Systems Engineering and Electronics, 2018,40(07):1595-1599.

[6] J.Shi,W.Yu,Q.Ni,W.Liang,Z.Li and P.Xiao.Energy Effcient Resource Allocation in Hybrid Non-Orthogonal Multiple Accrss Systems[J].IEEE Transactions on Communications,2019,67（5）：3496-3511.

[7] F. Fang, J. Cheng, and Z. Ding. Joint energy effcient subchannel and power optimization for a downlink NOMA heterI ogeneous network [J]. IEEE Trans. Veh. Technol. ,2019,68 (2) : 1351-1364.

[8] Yang N, Zhang H, Long K, et al. Deep Neural Network for Resource Management in NOMA Networks[J]. IEEE Transactions on Vehicular Technology, 2019, 69(1): 876-886.

[9] He C, Hu Y, Chen Y, et al. Joint power allocation and channel assignment for NOMA with deep reinforcement learning [J ]. IEEE Journal on Selected Areas in Communications , 2019, 37(10): 2200-2210.

[10] Shamna K F, Siyad C I, Tamilselven S, et al. Deep Learning Aided NOMA for User Fairness in 5G[C]. 2020 7th International Conference on Smart Structures and Systems (ICSSS). IEEE, 2020: 1-6.

[11] V. Mnih et al. Human-Level Control Through Deep Reinforcement Learning[J]. Nature, 2015,518(7540): 529-533.

[12] W. Ahsan, W. Yi, Z. Qin et al., Resource allocation in uplink NOMA-IoT networks: a reinforcement-learning approach. IEEE Trans. Wirel. Commun. 20(8), 5083–50

[13] Z. Wang, T. Schaul, M. Hessel. Dueling network architectures for deep reinforcement learning[J]. 2015:1995-2003.

[14] Zhang S, Li L, Yin J, et al. A dynamic power allocation scheme in power -domain NOMA using actor-critic reinforcement learning[C]. 2018 IEEE/CIC International Conference on Communications in China (ICCC). IEEE, 2018:719-723.

[15] T. Schaul, J. Quan, I. Antonoglou et al., Prioritized experience replay, in Proceedings of International Conference Learning, Representations (2015)

[16] C. He, Y. Hu, Y. Chen et al., Joint power allocation and channel assignment for NOMA with deep reinforcement learning. IEEE J. Sel. Areas Commun. 37(10), 2200−2210 (2019)

[17] T.P. Lillicrap, J.J. Hunt, A. Pritzel et al., Continuous control with deep

reinforcement learning, in ICLR (2015)

[18] Q. Le, V.-D. Nguyen, O.A. Dobre et al., Learning-assisted user clustering in cell-free massive MIMO-NOMA networks. IEEE Trans. Veh. Technol. (2021).

[19] X. Liu, X. Zhang, NOMA-based resource allocation for cluster-based cognitive industrial internet of things. IEEE Trans. Ind. Inform. 16(8), 5379–5388 (2019).

[20] Y. Zhang, X. Wang, Y. Xu. Energy-efcient resource allocation in uplink NOMA systems with deep reinforcement learning, in Proceedings of International Conference on Wireless Communications and Signal Processing (WCSP) (2019), p. 1–6.

[21] L. Salaün, M. Coupechoux, C.S.J.I.T.O.S.P. Chen, Joint subcarrier and power allocation in NOMA: optimal and approximate algorithms. IEEE Trans. Signal Process. 68, 2215–2230 (2020).

[22] He, Y. Hu, Y. Chen et al., Joint power allocation and channel assignment for NOMA with deep reinforcement learning[C]. IEEE J. Sel. Areas Commun. 37(10), 2200-2210 (2019)..

[23] X. Wang, Y. Zhang, R. Shen et al., DRL-based energy-efficient resource allocation frameworks for uplink NOMA systems. IEEE Internet Things J. 7(8), 7279–7294 (2020).

[24] X. Wang, R. Chen, Y. Xu et al., Low-complexity power allocation in NOMA systems with imperfect SIC for maximizing weighted sum-rate. IEEE Access 7, 94238–94253 (2019).

[25] Xiao L, Li Y, Dai C, et al. Reinforcement learning-based NOMA power allocation in the presence of smart jamming[J]. IEEE Transactions on Vehicular Technology, 2017, 67(4): 3377 -3389.

[26] W. Saetan, S. Thipchaksurat. Power allocation for sum rate maximization in 5G NOMA system with imperfect SIC: a deep learning approach, in Proceedings of the 4th International, Conference on Information Technology (2019), p. 195–198.

[27] F. Meng, P. Chen, L. Wu et al., Power allocation in multi-user cellular networks: deep reinforcement learning approaches. IEEE Trans. Wirel. Commun. 19(10), 6255–6267 (2020).

[28] F.H. Costa Neto，D. Costa Araujo，M． Pontes Mota，T. Macieland A． L． F． De Almeida,Uplink Power Control Framework Based on Reinforcement Learning for 5G Networks［C］． in IEEE Transactions on Vehicular Technology，doi: 10． 1109 /TVT(2021).