symmetry

MDPI

Article

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

Yuan Liu, Yue Li *, Lin Li and Mengli He

Electronic Engineering School, Heilongjiang University, Harbin 150001, China; 2211816@s.hlju.edu.cn (Y.L.); 2211794@s.hlju.edu.cn (L.L.); ml1006@163.com (M.H.)

* Correspondence: 2017021@hlju.edu.cn

Abstract: To address the need for massive connections in Internet-of-Vehicle communications, local wireless networks utilize non-orthogonal multiple access (NOMA). Scholars have introduced deep reinforcement learning networks for user grouping and power allocation to reduce computational complexity. However, the traditional algorithm based on DQN (Deep Q-Network) still exhibits slow convergence speed and low training stability, while the uniform sampling method in the sample playback process suffers from low sampling efficiency. In order to address these issues, this paper proposes a user grouping and power allocation method for NOMA systems based on Prioritized Dueling DQN-DDPG joint optimization. Firstly, the paper introduces the user grouping network based on Dueling DQN, which considers both the state value and action value in the entire connection layer. The two values compete with each other, are summed up, and re-evaluated. The network significantly improves training stability and increases the convergence speed. Secondly, in this paper, a depth deterministic strategy gradient (DDPG) algorithm with symmetric properties is used. This algorithm works well for continuous action spaces and avoids the power quantization error because of the continuity of power value in the power allocation stage. Finally, the priority sampling based on TD-error (Temporal-difference error) is combined with the Dueling DQN network and DDPG network to ensure random sampling and improve the replay probability of important samples. Simulation results show that the proposed priority-based Dueling DQN-DDPG algorithm significantly improves the convergence speed of sample training. The research results of this paper provide a solid foundation for the following research content, which focuses on NOMA system resource allocation under the mobile user state.

Keywords: non-orthogonal multiple access (NOMA); resource allocation; dueling DQN; prioritized sampling; depth deterministic policy gradient (DDPG)

check for updates

Citation: Liu, Y.; Li, Y.; Li, L.; He, M. NOMA Resource Allocation Method Based on Prioritized Dueling DQN-DDPG Network. *Symmetry* **2023**, *15*, 1170. https://doi.org/10.3390/sym15061170

Academic Editor: Javid Atai

Received: 14 May 2023
Revised: 26 May 2023
Accepted: 28 May 2023
Published: 30 May 2023

1. Introduction

As the 5G network becomes commercially available and the development of 6G technology continues, the demands for communication quality across various industries are increasing. Mobile communication devices are required to provide higher data rates, lower communication delays, and better reliability. The traditional Orthogonal Multiple Access (OMA) technologies cannot fully meet current communication needs, and as a result, Non-Orthogonal Multiple Access (NOMA) technology has become an essential aspect of the development of new generation communication technology. NOMA technology can be primarily 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. SIC is used to demodulate the signals and eliminate interference, allowing the required information to be recovered successfully. SIC works by first decoding the signal with the strongest power, which is usually the signal with the highest quality, and subtracting it from the received superimposed signals. This process is repeated for each user's signal until all signals are successfully demodulated and recovered. 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 a higher transmission rate, lower delay and better transmission reliability [1–3]. By allowing multiple users to share the same resources, NOMA technology can increase the capacity of the communication system while maintaining the same bandwidth. Furthermore, the use of SIC at the receiver allows for efficient decoding of these superimposed signals, improving the overall performance of the system. As a result, NOMA technology has become a promising candidate for the development of future wireless communication systems.

In recent years, many researchers have devoted themselves to the design and implementation of NOMA technology. They have demonstrated the compatibility of power domain NOMA with cooperative communication, relay systems, and MIMO technology. The problems of user grouping, power allocation, and spectrum resource allocation for NOMA have also attracted extensive attention. The system sum rate can be significantly improved by using an efficient scheme to group users and allocate power to them at the transmitter. This can also enhance the accuracy and stability of the system. Islam et al. [4] proposed a random user pairing method in which the base station randomly selected users to form several user sets with the same number of users. They then group two users with a large channel gain difference in each user set. Zhang et al. [5] proposed a user grouping based on channel gain. While these algorithms could improve the system performance, the complexities were too high to apply to practice. In [6], it was 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 the user throughput. A further study [7] examined the user pairing problem of the NOMA system based on fixed power allocation. They discussed the influence of user pairing on the sum rate, studied the power allocation scheme of two users pairing and analyzed its performance. Another study [8] proposed a low-complexity and high-efficiency three-stage alternating optimization algorithm, which comprehensively considered service quality, power budget, and cooperation constraints and optimized the transmit power, power allocation coefficient (PAC), and relay power. In two further studies [9] and [10], 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, which can lead to high computational complexity. In contrast, deep learning is a powerful tool that can be used to solve complex mathematical problems and has shown significant advantages. There have been many studies that combine NOMA technology with deep learning. One study [11] outlines that Deep Neural Networks (DNN) are used for decoding in order to consider user fairness in NOMA. Compared with traditional algorithms, Deep Learning (DL) can effectively reduce computational complexity, achieving fairness and maximizing the system sum rate efficiently. Another study [12] outlines the Attention-Based Neural Network (ANN) approach to allocating channels to users in the 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 computational complexity. One study [13] outlines the training of DNN to simulate the interior point algorithm for power allocation. The introduction of neural networks can improve computational efficiency. The study [14] effectively characterized the nonlinearity between channel diversity and transmission power clustering using a deep neural network-based UC (DNN-UC). The DNN-UC model provides a larger space for hyperparameter optimization to maximize its learning ability. 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. This allows for direct control strategies from high-dimensional raw data, providing faster convergence speed and greater effectiveness for multi-state and action-space systems. In [15], a Deep Q-Network (DQN) is proposed and used as an approximator in various fields. In [16], a DRL-based resource allocation scheme is proposed, 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. Currently, DQN is a commonly used deep reinforcement learning network that is widely applied in resource allocation for NOMA systems. It effectively addresses the high complexity issue in traditional NOMA resource allocation. However, traditional DQN networks are known to have slow training convergence speeds and unstable training processes when training with samples. In practice, the problem with DQN is that under the condition of state st, Q(st, a) cannot fully represent the value of state-action, resulting in slow convergence speed. At times, regardless of what action is taken in a certain state, it will not have a positive impact on the next state. When the state is good, no matter what action is taken, a high value will be obtained, and when the state is poor, the obtained value will also be low. In [17], an improved version of the DQN network called the Dueling DQN is proposed, which decomposes the state value $Q\pi(s_t,\mathrm{a}_{t1})$ into a state value function $V(s_t)$ and an action advantage function $A(s_t,\mathrm{a}_{t1})$ within the neural network. Dueling DQN emphasizes the advantages that can be obtained from each current state and action, and the state value function and advantage function form a competitive network. This effectively enhances the instability of the traditional DQN training process and speeds up the convergence of training. Based on this, the proposed paper applies Dueling DQN in the resource allocation of NOMA systems, which not only addresses the high complexity issue of traditional algorithms in resource allocation but also overcomes the problems of slow convergence speed and unstable training process of traditional DQN networks.

Given that the output of both DQN and Dueling DQN are discrete, when using Dueling DQN to perform power allocation tasks, the continuous user power needs to be quantized, which may result in quantization errors. In order to address this issue, Deep Deterministic Policy Gradient (DDPG) networks can be employed, as they can handle continuous action spaces [18]. In this paper, the power allocation optimization problem in NOMA systems is addressed using the Actor-Critic algorithm. The algorithm dynamically selects the power allocation coefficient and constructs a parameterized policy from the Actor-network part, which is evaluated by the Critic network. The Actor network then adjusts the power allocation policy based on feedback from the Critic network part.

Furthermore, the empirical replay algorithm is employed in the Dueling DQN and DDPG network to reduce the correlation between samples and ensure that the samples exhibit independent and identically distributed characteristics. However, the current sampling method involves uniform sampling, which does not consider the importance of samples. In the sampling process, some valuable samples may not be learned, thus reducing the learning rate. The prioritized sampling method, which is based on TD error, can improve the replay probability of important samples and address this issue [19]. Therefore, this paper proposes a priority sampling-based approach for the Dueling DQN and DDPG network to accelerate the convergence of training.

This paper aims to maximize the system sum rate in the NOMA resource allocation problem and proposes an optimal joint scheme based on a Prioritized Dueling DQN-DDPG network. In this scheme, the Dueling DQN is employed to perform discrete tasks for user grouping, and the DDPG network is used to perform continuous tasks for power allocation among each user. In building on this, this paper proposes a sampling optimization scheme to address the random sampling problem. The scheme is combined with the Dueling DQN-DDPG network, utilizing Time Difference Error (TD-error) to calculate sample priority and improve sampling efficiency and learning rate. Based on the aforementioned approach, this paper solves a series of issues in the resource allocation of NOMA systems, effectively addressing slow sample training convergence speeds and unstable system training, improving the efficiency of sample training, and ultimately enhancing the system sum rate [19].

2. Materials and Methods

2.1. System Model

Figure 1 depicts the transmission model of the NOMA uplink system. In this paper, we investigate the scenario of a multi-user NOMA system in the uplink, where the Base Station (BS) is positioned at the center of the cell, and users are randomly distributed in close proximity to the base station. Our objective is to maximize the system sum rate by addressing issues related to user grouping and power allocation within the cell. Assuming that there are K users per cell, they are randomly distributed throughout various locations within the cell. The base station and users are configured with single antennas. Channel decay follows the Rayleigh distribution, where z_n represents the additive Gaussian white noise with a variance of δ_{n2}. The total bandwidth of system B is evenly distributed among N sub-channels, 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. $\mathrm{c}_{\mathrm{m},n}$ indicates the data signal connected by user n. The power allocated to the user m on the n sub-channel is represented by $b_{m,n}$, $S_{m,n}$ represents the allocation index of the sub-channel, and when user m is assigned to sub-channel n, then $S_{m,n}$ = 1, and $S_{m,n}$ = 0. The signal transmitted on the nth sub-channel is then given by:

$$x_n = \sum_{i=1}^{M} c_{m,n}\sqrt{b_{m,n}}S_{m,n} \tag{1}$$

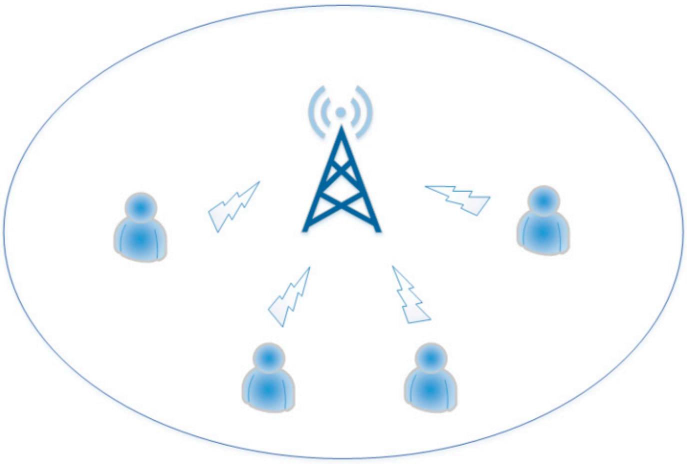


Figure 1. Transmission Model of NOMA Uplink System.

In this model, $g_{m,n}$ represents the channel gain of user m on the sub-channel n. The received signal expression at the base station is then given by:

$$y_n = g_{m,n}c_{m,n}\sqrt{b_{m,n}}S_{m,n} + \sum_{i=1,i\neq m}^{M} g_{i,n}c_{m,n}\sqrt{b_{i,n}}S_{i,n} + z_{m,n} \tag{2}$$

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 remaining signal as interference. Further to this, $z_{m,n}$ represents additive Gaussian white noise, which obeys complex Gaussian distribution, and $z_{m,n} \sim CN(0, \vartheta_n^2)$. Therefore, for users in sub-channel n, the $SINR$ can be expressed as:

$$SINR = \frac{c_{m,n} b_{m,n} |g_{m,n}|^2}{\vartheta_n^2 + \sum_{i=1,|g_{i,n}|^2<|g_{m,n}|^2}^{M} c_{m,n} b_{i,m} |g_{i,n}|^2} \tag{3}$$

According to Shannon's theorem, the rate of the mth user on the sub-channel n can be expressed as:

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

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

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

The system sum rate is:

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

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

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

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

$$\begin{gathered} \text{C1} : 0 \le b_{m,n} \le b_{\max} \\ \text{C2} : R_{m,n} \ge R_{\min} \end{gathered} \tag{8}$$

where b_{max} represents the maximum transmit power of the user, and R_{min} is the minimum data rate requirement for each user. Constraint C1 ensures that the transmit power per user does not exceed b_{max}, while constraint C2 guarantees that the rate per user meets the minimum signal rate requirement. Finding a globally optimal solution for this objective function is a challenging task. Although the global search method can provide the optimal solution by searching all possible user grouping combinations, the computational complexity is too high to be practical. Hence, the predecessors of this research utilized DRL to reduce the complexity of the calculation [20]. In building upon this, the present article proposes a novel approach that combines Prioritized Dueling DQN-DDPG with joint optimization for user grouping and power allocation in NOMA systems. The proposed method aims to enhance the system sum rate, improve learning efficiency, and address the issues of slow convergence speed and unstable training.

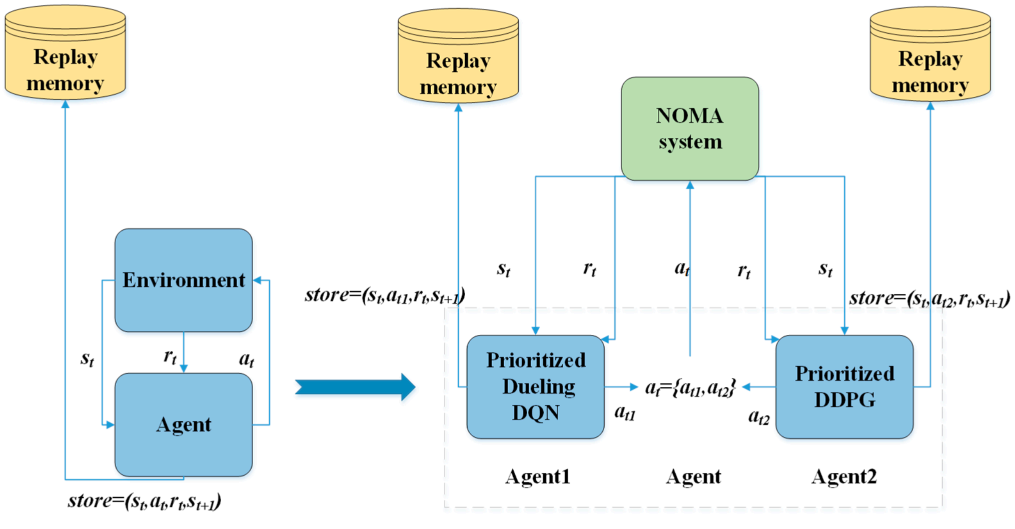
2.2. Resource Allocation Method Based on Prioritized Dueling DQN-DDPG

2.2.1. Resource Allocation Network Architecture

Generic reinforcement learning comprises five components, namely, Agent, Action, State, Reward, and Environment. Agent refers to an entity that produces a corresponding Action based on the input State. The Environment, in turn, receives the Action and returns the State and Reward. The Agent updates the decision function that generates the Action based on the Reward and the current State. This process is repeated until the Agent can produce the optimal Action in any State, i.e., the learning process of the model is completed. The critical aspect of reinforcement learning is ensuring that the State, Action, and Reward correspond one-to-one with the parameters of the NOMA system under study, thereby enabling the reinforcement learning method to achieve the desired outcomes [21].

Based on the structure of reinforcement learning, this paper designs the NOMA system model, as shown in Figure 2. NOMA stands for a 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 $A1$, and the power allocation space is defined as $A2$. 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 [22].

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} \ldots = \sum_{i=0}^{\infty} \gamma^i r_{t+i}, \gamma \in [0,1] \tag{9}$$



(a)Reinforcement learning

(b)Proposed system structure

Figure 2. Resource allocation network based on deep reinforcement learning.

The discount factor, γ, determines the relative importance of immediate rewards and future rewards. The value of γ ranges from 0 to 1. The expected value of the cumulative payoff R_t is defined as the Q value, which is determined by the state s_t. Choice action a_t under certain strategy π. It is expressed as:

$$Q_\pi(s_t, a_t) = E[r_t + \gamma \max Q_\pi(s_{t+1}, a_{t+1})] \tag{10}$$

At every Time Slot (TS), Agent1 and Agent2 obtain the channel gain from the NOMA system, select the user combination and power in the action space based on the current channel gain, and communicate the action results to the NOMA system. The NOMA system generates instantaneous rewards and channel gains for the next TS based on the received action, which are then passed on to Agent1 and Agent2, respectively. Based on the reward, Agent1 and Agent2 update the decision function that selects the optimal action under the current channel gain to complete the interaction. This process is reiterated until the Agent is capable of generating the optimal decision for any channel gain [23]. However, the DQN user grouping scheme proposed by previous researchers has some inherent issues, such as slow convergence speed and unstable training, which adversely impact system performance. Therefore, the present study improves upon the uplink method, proposing a joint optimization scheme for user grouping and power allocation in the NOMA system based on Prioritized Dueling DQN-DDPG, as illustrated in Figure 2.

2.2.2. User Grouping Based on Dueling DQN

This paper applies Prioritized Dueling DQN to accomplish the task of user grouping. DQN, as one of the deep reinforcement learning algorithms, merges the Q-learning algorithm with neural networks, leveraging the neural network's strong representational capability. In reinforcement learning, the input record serves as the state, which is fed into the neural network model (Agent) as input. Subsequently, the neural network model outputs the corresponding value (Q) of each action to determine the action to be executed. However, in numerous deep reinforcement learning tasks, the value functions for different actions in various states are not identical, and in some states, the value functions have no connection with actions. In line with the aforementioned concept, Wang et al. proposed the Dueling network model to substitute the network model in the DQN [17]. The core idea of Dueling DQN is to divide the state value $Q\pi(s_t,\alpha_{t1})$ into the state value function $V(s_t)$ and the action advantage function $A(s_t,\alpha_{t1})$. In this paper, Dueling DQN is implemented in the user grouping stage of the NOMA system. The fundamental concept is that Dueling DQN considers different state values and advantage functions in different states, which can swiftly select the present optimal action in the sample training process.

Dueling DQN-Based User Grouping Network

This section introduces the user grouping framework base on Dueling DQN in the NOMA system. As shown in Figure 3, 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 the 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. In taking s_t as input, this paper adopts the ε-greedy strategy to select α_{t1} as a new user combination, namely:

$$a_{t1} = \arg\max_{a_{t1}\in A1} (s_t, a_t; \theta, \beta, \alpha) \tag{11}$$

where θ is the convolution layer parameter; β and α are the fully connected layer parameters of the two branches.

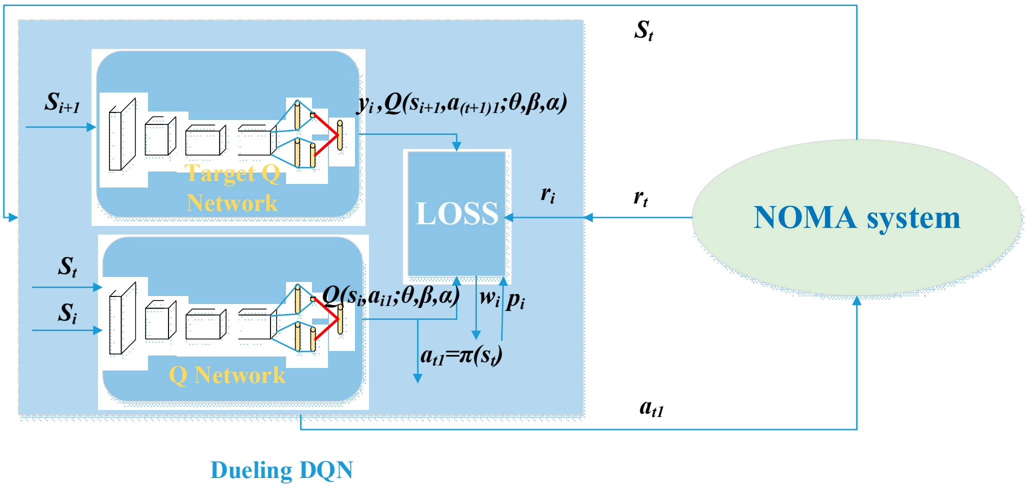


Figure 3. User grouping framework based on Dueling DQN.

This means that the ζ 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 α_{t1} and power α_{t2} (setting the power allocation action to α_{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 $(s_t, a_{t1}, r_t, s_{t+1})$. 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 calculate the sample weight using the sampling probability and train the target Q value in the network generated using the Q-network, namely:

$$y_i = r_i + \gamma \max_{a_{(i+1)1} \in A1} Q_\pi \left(s_{i+1}, a_{(i+1)1}; \theta^-, \beta, \alpha \right) \tag{12}$$

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:

$$LOSS_1 = \frac{1}{N} \sum_{i=1}^{N} w_i (y_i - Q(s_i, a_{i1}; \theta, \beta, \alpha))^2 \tag{13}$$

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 wi is the sampling weight importance of the sample [24,25].

Dueling DQN Network Structure

The architecture of the Dueling DQN model used in the user grouping algorithm is shown in Figure 4a, while the traditional DQN model architecture is given in Figure 4b for comparison. 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 second path outputs the action advantage value(*A(*s_t,α_{t1}*)*), which represents the additional value of selecting an action. Finally, through full connection, it is merged into the action value Q_π*(*s_t,α_{t1}*)*. The state value function is unrelated to the action. In contrast, the action advantage function is related to the action and represents the average reported degree of goodness of the action, which is related to the state and can solve the Reward-bias problem. Based on this competing network structure, the agent can learn a more realistic value *V(*s_t*)* in the environmental state without the influence of action [17].

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

$$V(s_t) \cong V(s_t; \theta, \beta) \tag{14}$$

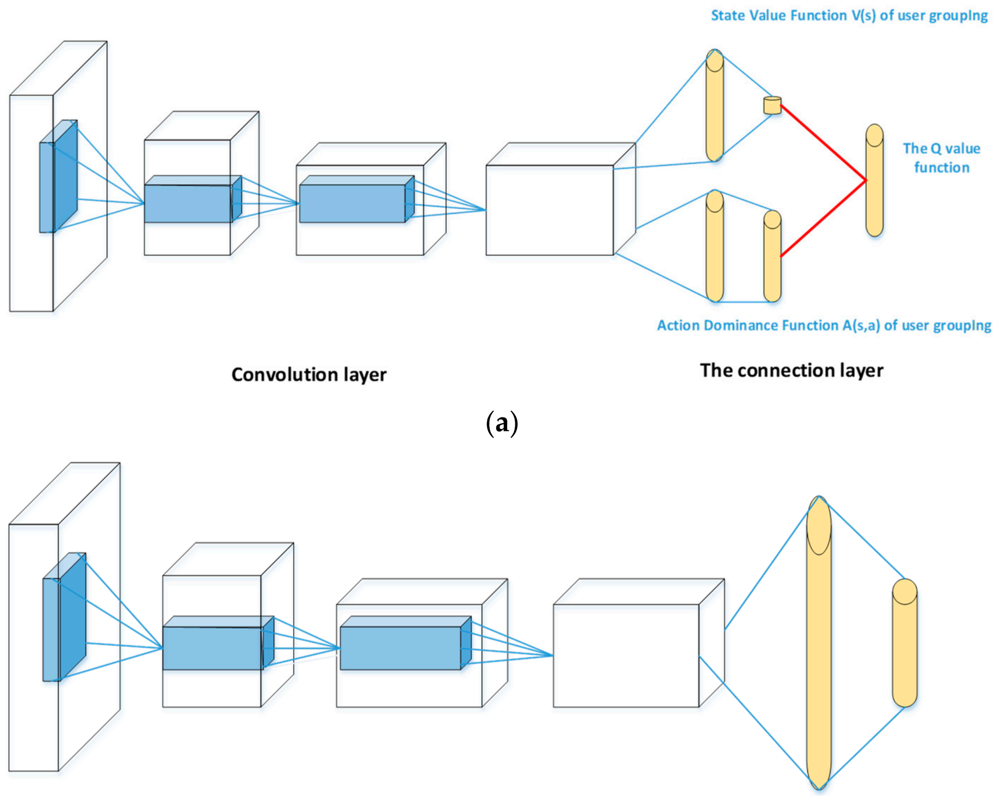
Action advantage function *A(*s_t,α_{t1}*)* can be expressed as:

$$A(s_t, a_{t1}) \cong A(s_t, a_{t1}; \theta, \alpha) \tag{15}$$

where θ is the convolution layer parameter; β and α are the fully connected layer parameters of the two branches.

Dueling DQN only improves the intermediate structure of the neural network, and simply splitting the model is not enough to fundamentally solve the problem. In this article, we need to impose some restrictions on the output of the split two parts. If we do not restrict the output of these two parts, there can be infinite possible combinations of the value function *V(*s_t*)* and the advantage function *A(*s_t,α_{t1}*)* given a constant Q value. However, only a few of these combinations actually make sense and come close to the real number. In order to solve this problem, this paper qualifies the dominant function *A(*s_t,α_{t1}*)*. 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:

$$Q_\pi(s_t, a_{t1}; \theta, \beta, \alpha) = V(s_t; \theta, \beta) + \left(A(s_t, a_{t1}; \theta, \alpha) - \frac{1}{|A|} \sum_{a_{t1}'} A(s_t, a_{t1}'; \theta, \alpha) \right) \tag{16}$$


State Value Function V(s) of user grouping
The Q value function
Action Dominance Function A(s,a) of user grouping
Convolution layer
The connection layer

(**a**)

(**b**)

Figure 4. Comparison of Dueling DQN and DQN network structures. (**a**) Dueling DQN network structure; (**b**) DQN network structure.

The advantage of this expression is that it ensures the stable relative ranking of dominant functions of each action in a given state, reduces the range of Q value, removes the excess degrees of freedom, and improves the stability of the algorithm. Compared to the traditional DQN network structure, Dueling DQN decomposes the Q value into the form of the value function $V(s_t)$ and the advantage function $A(s_t,a_{t1})$, which makes training easier and convergence faster. As the number of actions increases, this advantage becomes even more pronounced. The state value function depends solely on the state and is independent of the behavior, making it easier to train. Multiple behaviors can share the same value $V(s_t)$, in the same state. The difference between different behaviors lies only in the dominance function. The convergence of this part can also be independent of the value function, allowing for the independent learning of relative differences between behaviors. Moreover, the advantage function is introduced to avoid unstable results caused by the large magnitude of Q values and the very small differences between Q values.

The primary advantages of Dueling DQN are as follows:

(1) It can generalize the learning process to all possible actions in the environment without changing the underlying reinforcement learning algorithm.

(2) Since it can learn the most critical state for the agent, it does not need to know the impact of each action on each state, enabling it to quickly identify the best action.

(3) From a network training perspective, less data is required, making the network training more user-friendly and straightforward.

(4) Training the state and advantage functions separately makes it easier to maintain the order between actions. When breaking down the value function, each result part has practical significance, and their combination is uniquely determined, making the network learning more precise and robust.

Therefore, Dueling DQN, as an improved reinforcement learning algorithm, has better performance and higher efficiency and can be employed to address NOMA system problems such as user grouping.

2.2.3. 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, and they are effective in solving complex problems with high dimensions of state space and action space. However, they are 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, iterative optimization within the training set incurs a performance penalty. Therefore, DQN cannot be directly applied to continuous action spaces. DDPG is a model-free, offline learning method based on deterministic policy gradients. The DDPG algorithm has symmetric properties. 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. 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 networks [26]. The results show that the time complexity of the two frameworks is similar, but the performance of the DDPG network is better than that of the multi-DQN network. This is because, in multi-DQN, the quantization of user power results in the loss of some important information. The 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 based on sub-channel assignment in Dueling DQN. DDPG can be easily extended to larger and more complex mobile communication systems. Compared to the discrete method, the continuous resource allocation method proposed in this chapter can achieve a better system sum rate and has stronger processing power for large-scale user access. Figure 5 shows the network structure of DDPG [27–30].

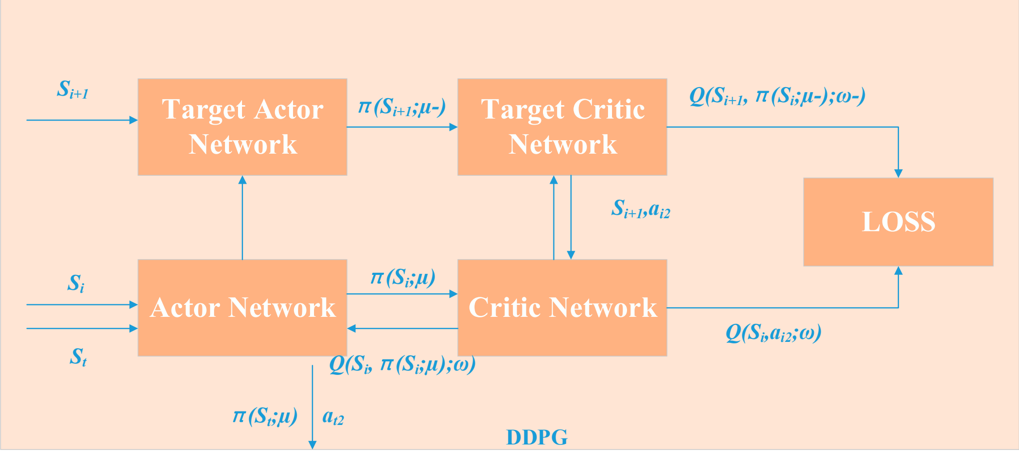


Figure 5. DDPG network structure.

2.2.4. Priority Experience Playback Mechanism

Dueling DQN and DDPG networks introduce an empirical replay mechanism to reduce the correlation and dependence among samples, where all samples are uniformly sampled from the experience pool. However, this approach may ignore some important samples, leading to lower learning efficiency. Therefore, to compensate for the insufficient random sampling of the experience pool, this paper introduces a priority-based empirical replay mechanism for Dueling DQN and DDPG networks. This mechanism solves the problem of low sampling efficiency in the empirical replay and greatly improves the training speed of network models.

The priority experience replay mechanism does not perform random sampling but instead samples according to the importance of each sample in the experience pool. This approach can more effectively find the samples required for training. In priority experience replay, the temporal difference error (TD-error) of each sample is used as the evaluation criterion for sampling. The TD-error formula for the samples in the user grouping is as follows [19].

$$\delta_i = y_i - Q(s_i, a_{i1}; \theta) \tag{17}$$

where δ_i is the TD-error of sample i. The larger the absolute value of the TD-error of a sample, the higher its probability of being sampled. The TD-error of a sample determines the probability of being sampled. The priority sampling probability of samples can be expressed as follows:

$$P(i) = \frac{P_i^k}{\sum j P_j^k} \tag{18}$$

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-errors may never be trained. In order to ensure that samples with lower priority can also be drawn as training samples, it is assumed that ε_0 is a positive value to ensure that the sample priority is always greater than 0. Further to this, k determines the degree of priority: when $k = 0$, it indicates uniform sampling, and when $k = 1$ indicates greedy strategy sampling. Therefore, k does not change the monotonicity of priority and is used to increase or decrease the priority of the TD-error experience.

Since the priority experience replay algorithm frequently replays empirical samples with high TD-errors, it can result in a change in the data distribution of the samples, leading to training bias or overfitting. In order to reduce this bias, the priority experience replay algorithm uses the importance sampling weight method to correct the bias. The importance sampling weight of a sample is defined as follows [31]:

$$w_i = \left(\frac{1}{H}\frac{1}{P(i)}\right)^{\sigma} \tag{19}$$

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

Figure 6 illustrates the priority-based sampling model.

2.2.5. User Grouping and Power Allocation of Prioritized Dueling DQN-DDPG

In this paper, we use Prioritized Dueling DQN to perform user grouping and Prioritized DDPG network for power allocation. The combination of the user group and power distribution is optimized to obtain the optimal user combination and the optimal power distribution mode. In the user grouping module of the resource allocation network, the NOMA system provides the current channel gain, which is s_t, and feeds it into the Dueling DQN. Dueling DQN selects user group action $\mathrm{a}_{\mathrm{t}1}$ from the user group action space according to the current channel gain and inputs it into the NOMA system. The NOMA system generates the next channel gain and reward (i.e., system sum rate) feedback to the Dueling DQN. Dueling DQN updates the action value function that generates this action according to the system sum rate, which is the real Q value. At this time, an interaction is completed. Power distribution action a_{t2} is also obtained by using the DDPG network in the power distribution module [32].

When we jointly optimize user grouping and power allocation, we need to obtain the joint action $a_t = \{a_{t1}, a_{t2}\}$ of user grouping and power allocation in the above process and input it into the NOMA system. According to the action, the system updates the environment, gives feedback to the distribution action, and then feeds back to the base station according to the reward value set. Finally, the base station adjusts the selected distribution action and updates the parameters of the network according to the feedback reward value and the updated state. The following Algorithm 1 shows the Prioritized user grouping and power allocation algorithm based on Prioritized Dueling DQN-DDPG [33]:

Algorithm 1: User grouping and power allocation of Prioritized Dueling DQN-DDPG

```
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 $\theta^{-} = \theta$.
The random weights $\mu$ and $\omega$ are used to initialize the current $\pi(s_i;\mu)$ of the Actor network, and the $Q(s_i;\omega)$ of Critic current network;
Update Actor target network $\pi(s_{i+1};\mu')$ with parameter $\mu' < -\mu$; Update Critic target network $Q(s_{i+1};\omega')$ with parameter $\omega' < -\omega$.
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.
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 $s_{t+1}$.
Save the $(s_t, a_{t1}, r_t, s_{t+1})$ to the memory.
Sample data $(s_t, a_{t1}, r_t, s_{t+1})$ by priority size from the memory.
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. The Target value $y_i = r_i + \gamma \max_{a_{(t+1)1} \in A1} Q\left((s_i, a_{(i+1)1}; \theta^{-})\right)$ of Target Q Network in Dueling DQN network is calculated,
the TD error($\delta_{i1} = y_i - Q(s_i, a_{i1}; \theta)$) of samples is calculated, and the loss function $Loss_1 = \frac{1}{N}\sum_{i=1}^{N} w_i\left(y_i - Q\left(s_i, a_{i1}; \theta\right)\right)^2$ is calculated.
Calculate the Target Q value $y_i = r_i + \gamma Q(s_{i+1}, \pi(s_{i+1}, \mu'); \omega')$ of the Target Critic Network, calculate the sample
$\delta_{i2} = y_i - Q(s_i, a_{i2}; \omega)$, and get the loss function $Loss_2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - Q\left(s_i, a_{i2}^2, \omega\right)\right)^2$.
Through continuous parameter update to train the sample, finally find the appropriate user group and power allocation mode $a_t = \{a_{t1}, a_{t2}\}$.
Through the calculated loss function, update all parameters, recalculate TD error, and then determine all sample priority $p_i$ according to the TD error, update all priority $p_i$.
The weight parameter $\theta$ of Dueling DQN is updated by minimizing the loss function formula $Loss_1$.
Update the weight $\omega$ of Critic current network in DDPG by minimizing loss function formula $Loss_2$.
The resampling strategy gradient formula $J(\mu) = \frac{-1}{N}\sum_{i=1}^{N} Q(s_i, a_i; \mu)$ was used to update the policy parameter $\mu$ of the Actor's current network in DDPG.
Every W interval, update the weight $\theta^{-}$ of the target network with the prediction network weight $\theta$.
Every W time intervals, update parameter $\mu'$ of Actor target network according to $\mu' = \tau\mu + (1-\tau)\mu'$ and parameter $\omega'$ according to Critic target network $\omega' = \tau\omega + (1-\tau)\omega'$.
END
END
```

The priority-based experience replay mechanism mentioned above is used in both the Dueling DQN and DDPG networks. Figure 7 illustrates the structure model for user grouping and power allocation based on Prioritized Dueling DQN-DDPG.

This paper proposes three approaches to optimize user grouping and power allocation in NOMA systems. Firstly, in the user grouping stage, Dueling DQN adopts a competitive network structure. The full connection layer of the Dueling DQN network is divided into two paths. The upper path outputs the state value, which represents the intrinsic value of the static state environment. The lower path outputs the action advantage value, which represents the additional value brought by selecting a certain action. These two values compete with each other and are combined into the action value by a full connection. The state value function is independent of the action, while the action advantage function is related to the action. It reflects the average reward obtained relative to the state s and is used to solve the reward bias problem. Based on this competitive network structure, the agent can learn more accurate values in an environment without action influence. By emphasizing the advantages of each current state and action, Dueling DQN can accelerate convergence speed and improve the stability of the training process. It also enhances the system sum rate.

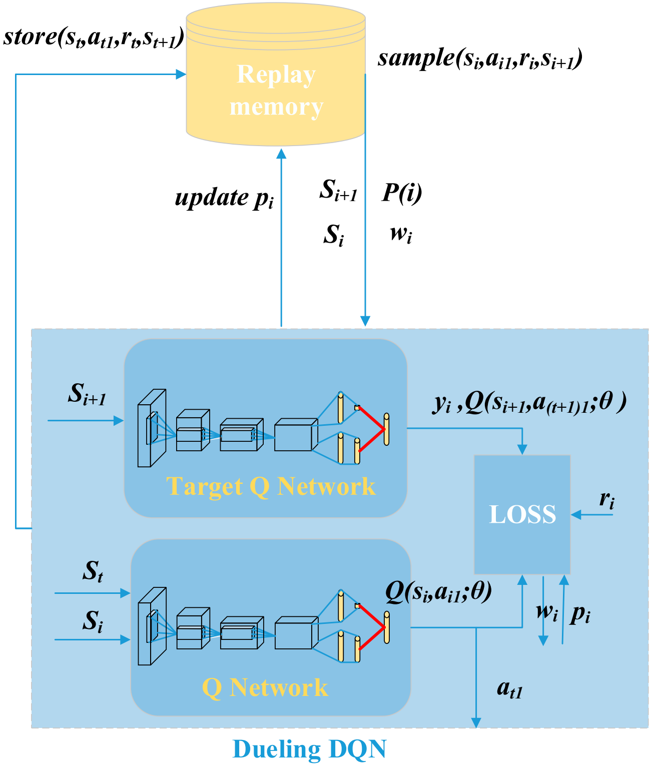


Figure 6. Prioritized Dueling DQN.

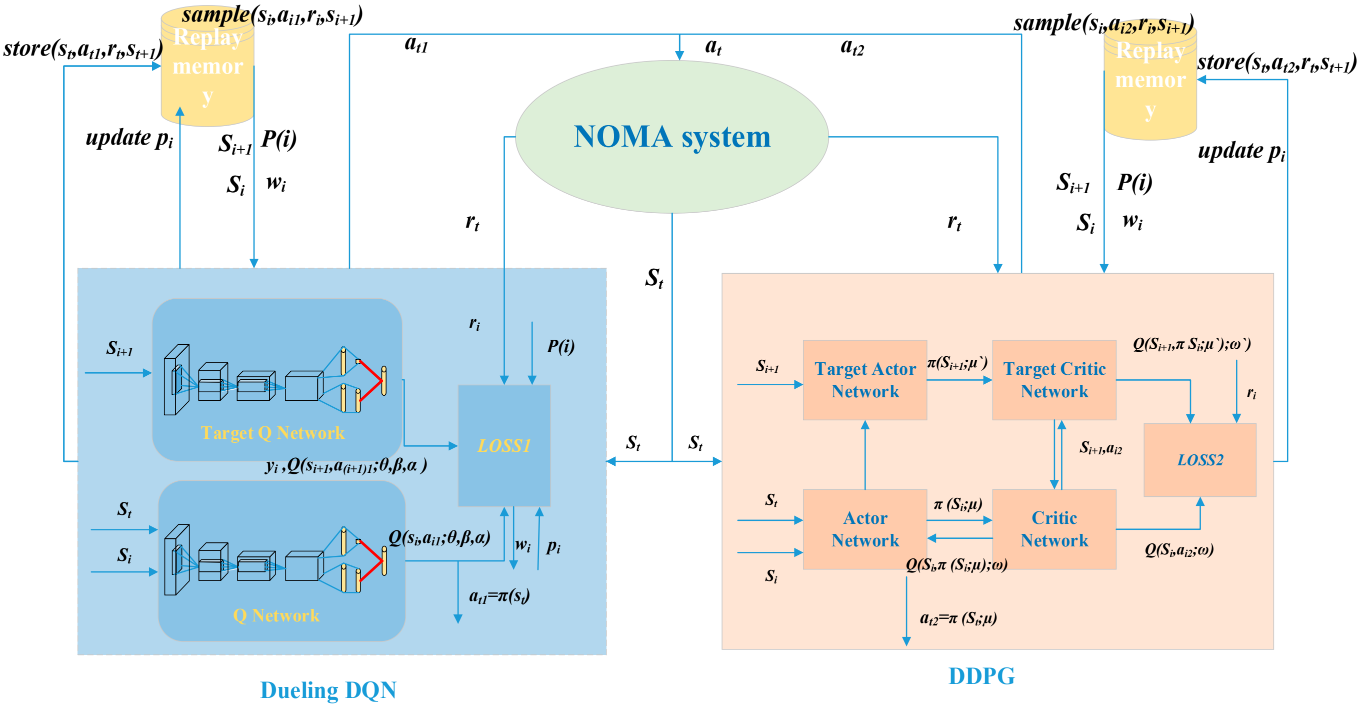


Figure 7. User grouping and power allocation of Prioritized Dueling DQN-DDPG.

Secondly, in the power allocation stage, due to the continuous nature of the user power variable, we use the DDPG network, which can effectively handle continuous actions and solve the problem that DQN cannot handle continuous actions. Compared with DQN, DDPG does not quantify the total power emitted by the base station for power allocation actions. This solves the problem of high dimensionality, and DDPG directly outputs the user's power, which can generate actions based on parameterized policies.

Finally, as Dueling DQN and DDPG introduce experience replay to reduce the correlation and dependence between samples, all samples are uniformly sampled from the experience pool. In this case, some important samples may be ignored, leading to reduced learning efficiency. Therefore, to address the problem of inadequate random sampling in the experience pool, this paper applies a prioritized reinforcement learning method to Dueling DQN and DDPG networks to solve the sampling problem in experience replay. It significantly improves the training speed of the network model. Based on this, this paper proposes a resource allocation method based on the Prioritized Dueling DQN-DDPG algorithm for joint optimization of user grouping and power allocation, which effectively improves the convergence speed of sample training and improve the stability of the training process. It also enhances the system sum rate.

3. Results and Discussion

This paper conducts simulations to evaluate the performance of the proposed Prioritized Dueling DQN-DDPG resource allocation in an uplink NOMA system. The base station is located at the center of the cell, and the users are randomly distributed throughout the cell. The specific parameters are listed in Table 1.

Table 1. Simulation parameter setting.

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

Various learning rates can affect the convergence speed and stability of Dueling DQN training. In this paper, the algorithm's learning rate is first set to 0.001 through parameter selection. Figure 8 illustrates the convergence of the proposed algorithm at different learning rates.

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 the DQN-DDPG method, the user grouping is completed according to DQN and the power allocation is finished according to DDPG. In the 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 stability, and algorithm time complexity of the aforementioned algorithms. As shown in Figure 9, it can be seen that the Prioritized Dueling DQN-DDPG algorithm is superior to the other two algorithms. The following will discuss the advantages and disadvantages of the algorithm proposed in this paper from three aspects.

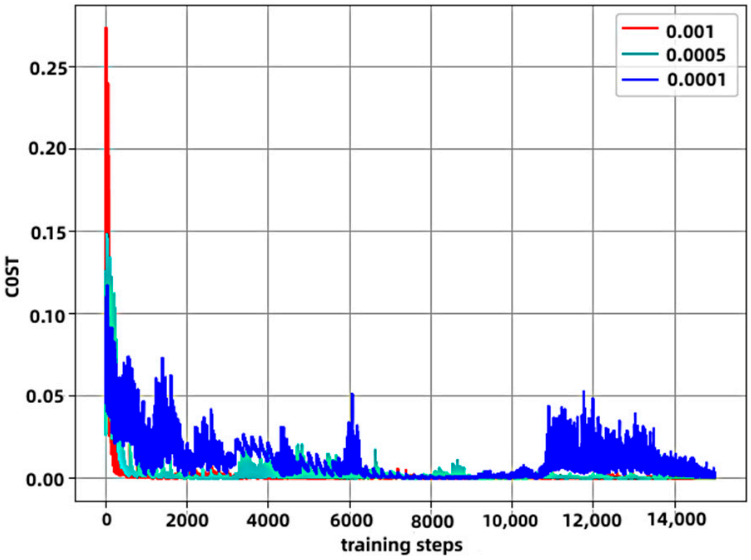


Figure 8. Comparison of learning efficiency parameters.

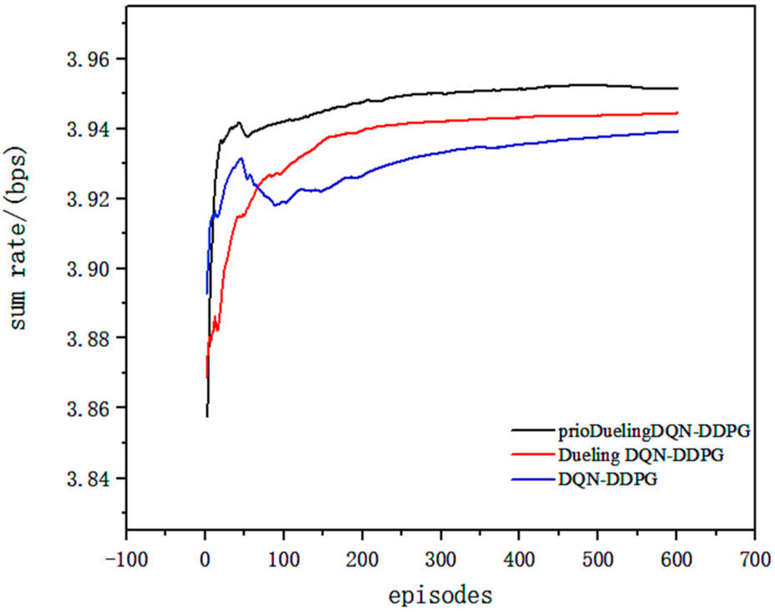


Figure 9. Comparison of different algorithm systems sum rate.

3.1. Convergence of the Proposed Algorithm

For the convergence performance of the proposed algorithm in this paper, Figure 9 shows a 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 150, while DQN-DDPG tends to converge when the number of iterations is nearly 300. We compared the proposed Dueling DQN-DDPG with DQN-DDPG in two aspects. Firstly, the convergence speed of Dueling DQN-DDPG is significantly faster than DQN-DDPG, with an increase of more than double. This is because the main feature of Dueling DQN is to use the model structure to express the value function in a more detailed form, allowing the model to have better performance. DQN only contains one Q network, corresponding to only one Q function, while the Q network in Dueling DQN contains two functions: the state value function and the advantage function. The state value function represents the inherent value of the static environment itself, and the advantage function represents the additional value brought by choosing a certain action in a certain state. Dueling DQN can speed up convergence by paying attention to the advantage that can be obtained for each current state and action, and the training process is more stable. Second, it can be observed that the convergence speed of the Prioritized Dueling DQN-DDPG is significantly faster than that of Dueling DQN-DDPG because the prioritized experience replay stores prioritized learning experience in the experience pool and guides the optimization of model parameters by extracting samples with high TD-error, which improves 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 significant improvement compared with Dueling DQN-DDPG.

Based on the above two discussions, we conclude that the proposed Prioritized Dueling DQN-DDPG method significantly improves the stability and convergence speed of training, with a speed increase of almost double.

3.2. Average Sum Rate Performance of the Proposed Algorithm

Figure 9 shows the experimental results for the system sum rate. All experimental results are averaged every 600 TS to achieve a smoother and clearer comparison. The Prioritized Dueling DQN-DDPG algorithm has obvious advantages over the other two algorithms in the system sum rate. Compared with the DQN-DDPG algorithm, the proposed algorithm improves the system sum rate by 0.5%. There are two reasons for this. Firstly, the network structure of Dueling DQN has more advantages than DQN. DQN only contains one Q network, corresponding to only one Q function, while the Q network in Dueling DQN contains two functions: the state value function and the advantage function. The state value function represents the inherent value of the static environment itself, and the advantage function represents the additional value brought by choosing a certain action in a certain state. By giving importance to the advantage that can be obtained for each current state and action, Dueling DQN can learn more accurate Q values based on 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 the priority for valuable samples that are beneficial to network training, thereby improving the system sum rate.

3.3. 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 Dueling DQN divides the output of the fully connected layer into two parts, decomposing the Q value into the value function and the advantage function, which compete with each other to obtain the optimal solution, and then adds these two parts together. Therefore, some calculation steps are added when training samples, resulting in an increase in computational complexity. Secondly, this paper introduces the priority algorithm based on temporal errors in the Dueling DQN and DDPG networks, which increases the computational complexity. However, during the training process, the convergence speed of the algorithm improved significantly. Table 2 is a time complexity comparison of the two methods. Figure 10 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 s	2724.8576722 s	15.703%
2	2074.8412441 s	2376.3507379 s	14.531%
3	2021.6223315 s	2280.1505739 s	12.788%
4	2042.1567555 s	2304.2756512 s	12.835%
5	2006.6152422 s	2290.4872706 s	14.146%
6	2020.9810086 s	2323.9442205 s	14.990%
7	2031.1689703 s	2305.3912113 s	13.500%
8	2011.4987920 s	2276.6919056 s	13.159%
9	2103.3444909 s	2434.7927646 s	15.758%
10	2043.8314553 s	2370.6670829 s	15.991%

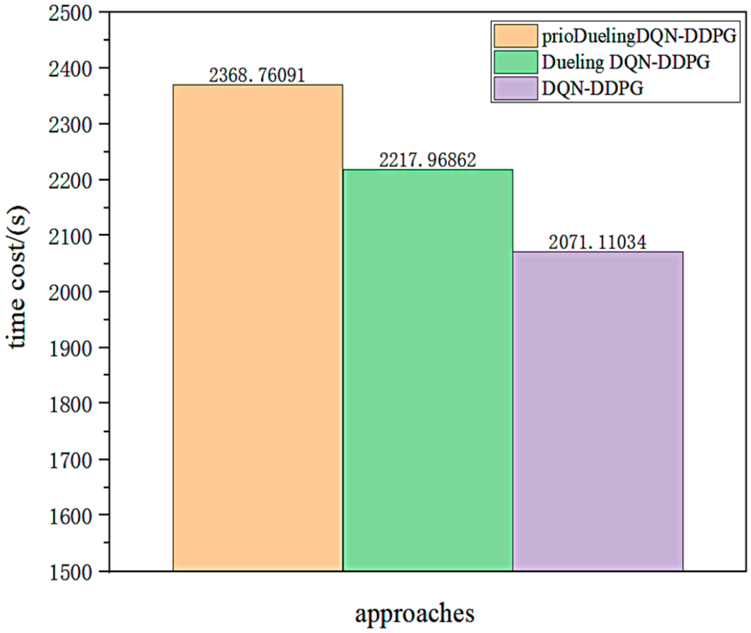


Figure 10. Comparison of time complexity.

The table above shows the time complexity comparison results of ten experiments. It is observed that Prioritized Dueling DQN-DDPG increases the average time complexity by around 15% compared with DQN-DDPG. This paper only compares the running time of the two algorithms:

$$\text{Time complexity increased by percentage} = \frac{\text{Running time of Prioritized Dueling DQN} - \text{DDPG} - \text{DQN} - \text{Running time of DDPG}}{\text{Running time of DQN} - \text{DDPG}} \times 100\%$$

The above method of calculating time complexity is based on the comparison of the results obtained by running both algorithms 600 TS. However, as we have concluded in Section 3.1, our proposed algorithm approaches convergence at around 150 iterations, while DQN-DDPG tends to converge at around 300 TS. This section calculates the running time of both algorithms as they approach convergence. As we can see from the table below, the Prioritized Dueling DQN-DDPG algorithm reaches convergence in only about 36% of the time taken by DQN-DDPG. Therefore, based on the time taken to reach convergence for both algorithms, we can conclude that the time complexity of the Prioritized Dueling DQN-DDPG algorithm has not increased but has actually reduced the training time required to some extent. Table 3 shows a comparison of the convergence time between Prioritized Dueling DQN-DDPG and DQN-DDPG.

Table 3. Comparison of convergence time of the two methods.

	DQN-DDPG (300 TS)	Prioritized Dueling DQN-DDPG (150 TS)	(Prioritized Dueling DQN-DDPG/DQN-DDPG) × 100%
1	597.822680 s	212.271871 s	35.51%
2	617.336156 s	228.685814 s	38.25%
3	607.584831 s	219.122785 s	36.06%
4	582.458203 s	209.875624 s	36.03%
5	584.686530 s	209.240106 s	36.01%
6	586.673505 s	213.803065 s	36.44%
7	561.049539 s	206.073456 s	36.73%
8	568.325278 s	206.477650 s	36.33%
9	551.649149 s	227.379164 s	41.22%
10	558.230064 s	185.991332 s	33.32%
Average value	-	-	36.58%

4. Conclusions

This paper aims to solve 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's sum-rate maximization. A resource allocation method for the 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. In the power allocation part, the Prioritized DDPG network is used to output the power of all users simultaneously. 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 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%. The time required for the Prioritized Dueling DQN-DDPG algorithm to reach convergence is only about 36% of the time required by the DQN-DDPG algorithm.

This paper focuses on the resource allocation of the NOMA system in the non-mobile state. However, in practical applications, resource allocation in the mobile state, such as the Internet of Vehicles, should also be considered. The mobile state can cause system instability, which can lead to additional challenges. The research conducted in this paper establishes a theoretical foundation for more complex practical applications and provides a reliable basis for the implementation of follow-up work. The future direction of this paper is to continue studying the resource allocation of NOMA in the mobile state and consider more complex channel scenarios based on the outcomes of this research [19]. The optimization method proposed in this paper is also applicable to multi-user MIMO. Our future research direction will also consider applying this method to resource allocation in MIMO.

Author Contributions: Y.L. (Yue Li) proposed the framework of the whole algorithm; Y.L. (Yuan Liu) performed the simulations, analysis and interpretation of the results. Y.L. (Yuan Liu), L.L. and M.H. participated in the conception and design of this research and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Basic Scientific Research Project of Heilongjiang Province [grant number 2020-KYYWF-1003].

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: A preprint of this paper [19] was published in Research Square, but the overall structure and content of this paper have been significantly modified. The research content of this paper is original and has been improved upon from the preprint, which has been cited in the references. See literature [19] for details.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. You, X.; Pan, Z.; Gao, X.; Cao, S.; Wu, H. The 5G mobile communication: The development trends and its emerging key techniques. *Sci.-Sin. Inf.* **2014**, *44*, 551–563.
2. Yu, X.H.; Pan, Z.W.; Gao, X.Q.; Cao, S.M.; Wu, H.S. Development trend and some key technologies of 5G mobile communication. *Sci. China Inf. Sci.* **2014**, *44*, 551–563.
3. Goto, J.; Nakamura, O.; Yokomakura, K.; Hamaguchi, Y.; Ibi, S.; Sampei, S. A Frequency Domain Scheduling for Uplink Single Carrier Non-orthogonal Multiple Access with Iterative Interference Cancellation. In Proceedings of the 2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall), Vancouver, BC, Canada, 14–17 September 2014; IEEE: Vancouver, BC, Canada; pp. 1–5.
4. Islam, S.M.R.; Zeng, M.; Dobre, O.A.; Kwak, K.-S. Resource allocation for downlink NOMA systems: Key techniques and open issues. *IEEE Wirel. Commun.* **2018**, *25*, 40–47. [CrossRef]
5. Zhang, H.; Zhang, D.-K.; Meng, W.-X.; Li, C. User pairing algorithm with SIC in non-orthogonal multiple access system. In Proceedings of the International Conference on Communications, Kuala Lumpur, Malaysia, 22–27 May 2016; pp. 1–614.
6. Sun, Y.; Ng, D.W.K.; Ding, Z.; Schober, R. Optimal Joint Power and Subcarrier Allocation for Full-Duplex Multicarrier Non-Orthogonal Multiple Access Systems. *IEEE Trans. Commun.* **2017**, *65*, 1077–1091. [CrossRef]
7. Li, X.; Ma, W.; Luo, L.; Zhao, F. Power allocation of NOMA system in Downlink. *Syst. Eng. Electron.* **2018**, *40*, 1595–1599.
8. Asif, M.; Ihsan, A.; Khan, W.U.; Ranjha, A.; Zhang, S.; Wu, S.X. Energy-Efficient Backscatter-Assisted Coded Cooperative-NOMA for B5G Wireless Communications. *IEEE Trans. Green Commun. Netw.* **2022**, *7*, 70–83. [CrossRef]
9. Shi, J.; Yu, W.; Ni, Q.; Liang, W.; Li, Z.; Xiao, P. Energy Effcient Resource Allocation in Hybrid Non-Orthogonal Multiple Accrss Systems. *IEEE Trans. Commun.* **2019**, *67*, 3496–3511. [CrossRef]
10. Fang, F.; Cheng, J.; Ding, Z. Joint energy effcient subchannel and power optimization for a downlink NOMA heterI ogeneous network. *IEEE Trans. Veh. Technol.* **2019**, *68*, 1351–1364. [CrossRef]
11. Yang, N.; Zhang, H.; Long, K.; Hsieh, H.-Y.; Liu, J. Deep Neural Network for Resource Management in NOMA Networks. *IEEE Trans. Veh. Technol.* **2019**, *69*, 876–886. [CrossRef]
12. He, C.; Hu, Y.; Chen, Y.; Zeng, B. Joint power allocation and channel assignment for NOMA with deep reinforcement learning. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2200–2210. [CrossRef]
13. Shamna, K.F.; Siyad, C.I.; Tamilselven, S.; Manoj, M.K. Deep Learning Aided NOMA for User Fairness in 5G. In Proceedings of the 2020 7th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 23–24 July 2020; IEEE: Chennai, India, 2020; pp. 1–6.
14. Kumaresan, S.P.; Tan, C.K.; Ng, Y.H. Deep Neural Network (DNN) for Efficient User Clustering and Power Allocation in Downlink Non-Orthogonal Multiple Access (NOMA) 5G Networks. *Symmetry* **2021**, *13*, 1507. [CrossRef]
15. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-Level Control Through Deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
16. Ahsan, W.; Yi, W.; Qin, Z.; Liu, Y. Arumugam Nallanathan, Resource allocation in uplink NOMA-IoT networks: A reinforcement-learning approach. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 5083–5098. [CrossRef]
17. Wang, Z.; Schaul, T.; Hessel, M. Dueling network architectures for deep reinforcement learning. *PMLR* **2015**, *48*, 1995–2003.
18. Zhang, S.; Li, L.; Yin, J.; Liang, W.; Li, X.; Chen, W.; Han, Z. A dynamic power allocation scheme in power-domain NOMA using actor-critic reinforcement learning. In Proceedings of the 2018 IEEE/CIC International Conference on Communications in China (ICCC), Beijing, China, 16–18 August 2018; IEEE: Beijing, China, 2018; pp. 719–723.
19. Liu, Y.; Li, Y.; Li, L.; He, M. NOMA Resource Allocation Method Based on Prioritized Dueling DQN-DDPG Network. *Res. Sq.* **2022**, *Preprint*.
20. Schaul, T.; Quan, J.; Antonoglou, I.; Silver, D. Prioritized experience replay. In Proceedings of the International Conference Learning, Representations, San Diego, CA, USA, 7–9 May 2015.
21. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning, in ICLR. *arXiv* **2015**, arXiv:1509.02971.
22. Le, Q.N.; Nguyen, V.-D.; Nguyen, N.-P.; Chatzinotas, S.; Dobre, O.A.; Zhao, R. Learning-assisted user clustering in cell-free massive MIMO-NOMA networks. *IEEE Trans. Veh. Technol.* **2021**, *70*, 12872–12887. [CrossRef]
23. Liu, X.; Zhang, X. NOMA-based resource allocation for cluster-based cognitive industrial internet of things. *IEEE Trans. Ind. Inform.* **2019**, *16*, 5379–5388. [CrossRef]
24. Zhang, Y.; Wang, X.; Xu, Y. Energy-efcient resource allocation in uplink NOMA systems with deep reinforcement learning. In Proceedings of the International Conference on Wireless Communications and Signal Processing (WCSP), Xi'an, China, 23–25 October 2019; pp. 1–6.

25. Salaün, L.; Coupechoux, M.; Chen, C.S.J.I.T.O.S.P. Joint subcarrier and power allocation in NOMA: Optimal and approximate algorithms. *IEEE Trans. Signal Process.* **2020**, *68*, 2215–2230. [CrossRef]
26. Wang, X.; Zhang, Y.; Shen, R.; Xu, Y.; Zheng, F.-C. DRL-based energy-efficient resource allocation frameworks for uplink NOMA systems. *IEEE Internet Things J.* **2020**, *7*, 7279–7294. [CrossRef]
27. Cheng, W.; Zhao, S.; Mei, C.; Zhu, Q. Joint Time and Power Allocation Algorithm in NOMA Relaying Network. *Int. J. Antennas Propag.* **2019**, *2019*, 7842987.
28. Xiao, L.; Li, Y.; Dai, C.; Dai, H.; Poor, H.V. Reinforcement learning-based NOMA power allocation in the presence of smart jamming. *IEEE Trans. Veh. Technol.* **2017**, *67*, 3377–3389. [CrossRef]
29. Feng, L.; Fu, X.; Tang, Z.; Xiao, P. Power Allocation Intelligent Optimization for Mobile NOMA Communication System. *Int. J. Antennas Propag.* **2022**, *2022*, 5838186.
30. Meng, F.; Chen, P.; Wu, L.; Cheng, J. Power allocation in multi-user cellular networks: Deep reinforcement learning approaches. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 6255–6267. [CrossRef]
31. Neto, F.H.C.; Araujo, C.; Mota, M.P.; Macieland, T.; De Almeida, A.L.F. Uplink Power Control Framework Based on Reinforcement Learning for 5G Networks. *IEEE Trans. Veh. Technol.* **2021**, *70*, 5734–5748. [CrossRef]
32. Ge, J.; Liang, Y.-C.; Joung, J.; Sun, S. Deep Reinforcement Learning for Distributed Dynamic MISO Downlink-Beamforming Coordination. *IEEE Trans. Commun.* **2020**, *68*, 6070–6085. [CrossRef]
33. Mismar, F.B.; Evans, B.L.; Alkhateeb, A. Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination. *IEEE Trans. Commun.* **2020**, *68*, 1581–1592. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.