NOMA Resource Allocation Method in IoV Based on Prioritized DQN-DDPG Network

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

Research Article

Keywords: Prioritized deep Q network (Prioritized DQN), sum tree, importance sampling, deep deterministic policy gradient (DDPG), non-orthogonal multiple access (NOMA).

Posted Date: November 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1023427/v1

Version of Record: A version of this preprint was published at EURASIP Journal on Advances in Signal Processing on December 1st, 2021. See the published version at https://doi.org/10.1186/s13634-021-00828-1.

NOMA Resource Allocation Method in IoV Based on Prioritized DQN-DDPG Network

Mengli He, Yue Li *, Xiaofei Wang, Zelong Liu
Electronic Engineering School, Heilongjiang University, Harbin, 150001, China

Abstract

To meet the demands of massive connections in the Internet-of-vehicle (IoV) communications, non-orthogonal multiple access (NOMA) is utilized in the local wireless networks. In NOMA technique, power multiplexing and successive interference cancellation techniques are utilized at the transmitter and the receiver respectively to increase system capacity, and user grouping and power allocation are two key issues to ensure the performance enhancement. Various optimization methods have been proposed to provide optimal resource allocation, but they are limited by computational complexity. Recently, the deep reinforcement learning (DRL) network is utilized to solve the resource allocation problem. In a DRL network, an experience replay algorithm is used to reduce the correlation between samples. However, the uniform sampling ignores the importance of sample. Different from conventional methods, this paper proposes a joint prioritized DQN user grouping and DDPG power allocation algorithm to maximize the sum rate of the NOMA system. At the user grouping stage, a prioritized sampling method based on TD-error (temporal-difference error) is proposed to solve the problem of random sampling, where TD-error is used to represent the priority of sample, and the DQN takes samples according to their priorities. In addition, sum tree is used to store the priority to speed up the searching process. At the power allocation stage, to deal with the problem that DQN cannot process continuous tasks and needs to quantify power into discrete form, a DDPG network is utilized to complete power allocation tasks for each user. Simulation results show that the proposed algorithm with prioritized sampling can increase the learning rate and perform a more stable training process. Compared with the previous DQN algorithm, the proposed method improves the sum rate of the system by 2% and reaches 94% and 93% of the exhaustive search algorithm and optimal iterative power optimization algorithm, respectively. While the computational complexity is reduced by 43% and 64% compared with the exhaustive search algorithm and optimal iterative power optimization algorithm, respectively.

Key Words: Prioritized deep Q network (Prioritized DQN), sum tree, importance sampling, deep deterministic policy gradient (DDPG), non-orthogonal multiple access (NOMA).

1 Introduction

Internet of Vehicles (IoV) is applied to support road safety, smart and green transportation and In-vehicle Internet access, which is a promising technique to improve autonomous driving system performance. 5G is the core wireless technology used for IoV networks that provides ubiquitous connectivity and mass data transmission [1]. Among various new technologies in 5G, Non-orthogonal multiple access (NOMA) is utilized to support high capacity data transmissions by multiplexing the same time frequency resources by power division or code division [2-4]. Sparse code multiple access (SCMA) is a popular technology in code domain NOMA. The spreading sequences are sparse sequences, and SCMA can significantly improve system capacity through non-orthogonal resource allocation [5]. The principle of power domain NOMA technology is to use power multiplexing technology to allocate power to different users at the transmitter, and then superimpose multiple users on the same time-frequency resource block by superposition coding (SC) technology, and send them to the receiver by non-orthogonal method. At the receiving end, successive interference cancellation (SIC) is used to eliminate interferences from superimposed users [6].

Meanwhile, the connection ability of NOMA can make it applicable to future wireless communication systems (for example, cooperative communication, multiple-input multiple-output (MIMO), beam forming and Internet of Things (IoT), etc.). Researchers combine NOMA and MIMO to give full play to their advantages, which can further improve the efficiency of the system in terms of capacity and reliability [7-8]. Liu [9] proposed a Ka-band multibeam satellite IIoT, which improved the transmission rate of NOMA by optimizing the power allocation proportion of each node. The results showed that the total transmission rate of NOMA is much larger than that of OMA. He later [10] proposed a cluster-based cognitive industrial IoT (CIIoT), in which data were transmitted through NOMA. The results showed that the NOMA for the cluster-based CIIoT could guarantee transmission performance and improve system throughput.

When the NOMA system was proposed, its resource allocation problem was mainly studied by constructing the joint optimization of user grouping and power allocation, and to find the optimal solution by using typical algorithms such as convex optimization and Lagrange multiplication. Han S and others [5] used a Lagrangian dual decomposition method to solve the non-convex optimization problem of power allocation, and the results showed that the optimized algorithm can significantly improve the system performance. Islam [11] 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, and then put the two users with large channel gain difference in the user set into one group. Benjebbovu [12] proposed an exhaustive user grouping algorithm. Zhang [13] proposed an algorithm for user grouping based on channel gain. These algorithms could improve the system performance, but at the same time, the complexities were too high to apply to practice. Sala [14] proposed a joint subchannel

and power allocation algorithm, and the results showed that this algorithm had low complexity. However, due to the dynamism and uncertainty of the wireless communication system, it is difficult for these joint optimization algorithms of user grouping and power allocation to model the system and derive the optimal scheme. Without an accurate system model, the performance of the NOMA system may be limited.

In recent years, deep learning has been applied to wireless communication. Many scholars use neural networks to approximate optimization problems. Gui [15-16] used the neural network to allocate resources, and proposed a deep learning aided NOMA system. Compared with traditional methods, this method had good performance. Saetan [17] proposed a power allocation scheme with maximum system sum rate. The optimal scheme was found by exhaustive search, and the optimal power allocation scheme was learned by training a deep neural network. The results showed that the scheme could approach the optimal sum rate and reduce the computational complexity. Huang [18] designed an effective deep neural network which implemented user grouping and power allocation through a training algorithm, improved transmission rate and energy efficiency. However, deep learning itself cannot generate the learning goal, but through the use of an optimization algorithm. The deep neural network will be trained according to the learning goal provided by the optimization algorithm, and the advantage of the deep learning network is to improve the calculation speed and reduce the running time. Therefore, the method based on deep learning needs to use the traditional optimization algorithm to generate the best label for training. In a complex system, it is difficult to obtain good training data, and the training is also very time-consuming.

To solve these problems, deep reinforcement learning (DRL) is applied. Deep reinforcement learning is a combination of deep learning and machine learning. It uses the powerful representation ability of the neural network to fit Q table or direct fitting strategy to solve the problem of large state action space or continuous state action space [19]. Ahsan [20] proposed an optimization algorithm based on DRL and SARSA algorithms to maximize the sum rate. The results showed that it could achieve a high accuracy with low complexity. Mnih [21] proposed a Deep Q-Network (DQN) which was used as an approximator in many fields. He [22] *et al*. proposed a resource allocation scheme based on DRL, which expressed the joint channel allocation and user grouping problem as an optimization problem. Compared with other methods, the proposed framework could achieve better system performance.

When using the deep reinforcement learning network to allocate NOMA resources, there are many problems that need to be further solved. Firstly, the experience playback algorithm is used in DQN (the most commonly used deep reinforcement learning network) to reduce the correlation between samples and ensure the independent and identically distributed characteristics between samples, but the current sampling method of the sample pool is uniform sampling which ignores the importance of the sample. In the sampling process, some valuable samples may not be learned, which reduces the learning rate. Prioritized DQN algorithm [23] can solve the

sampling problem in experience replay. It can improve the sampling efficiency and learning rate by using a sum tree and importance sampling. In addition, since the output of DQN can only be discrete, but the user power is continuous, although the power can be quantified, quantization will bring quantization error. Deep deterministic policy gradient (DDPG) network [24] can solve this problem, and use actor-critic structure to improve the stability of learning. Meng [25] performed multi-user power allocation based on the DDPG algorithm. The results showed that the algorithm is superior to the existing models in terms of sum rate, and had better generalization ability and faster processing speed.

Aiming at the above problems in current NOMA resource allocation methods, this paper proposes a joint optimization method of user grouping and power allocation in the NOMA system based on deep reinforcement learning network. Firstly, this paper proposes a joint design of DQN-DDPG network, in which DQN executes discrete tasks to perform user grouping, while DDPG network executes continuous tasks to allocate power to each user. Secondly, this paper proposes one solution to the problems existing in the random sampling methods, where temporal difference error (TD-error) is used to calculate the sample priority, and the valuable samples are sampled according to the priority. Besides, the sum tree is also utilized to speed up the search speed of priority samples.

The paper is organized as follows. Section 2 presents the system model of NOMA and the optimization objective of this paper. Section 3 describes the proposed NOMA system resource allocation algorithm based on deep reinforcement learning, specifically, the method of storing priority and the sampling method in the prioritized experience replay. Section 4 shows the numerical simulation results. Section 5 draws a conclusion.

2 Methods

2.1 System Model

This paper researches on the resource allocation issue of an uplink multi-user NOMA system, where the base station (BS) is located in the center of the cell, and the users are randomly distributed near the base station. The total system bandwidth B is equally divided among S subchannels, and the users in the same subchannel are non-orthogonal. Assume there are U users and S subchannels in the system, and the maximum power transmitted by the base station is $P_{\max}$. The signal transmitted on subchannel s is,

$$x_s(t) = \sum_{u=1}^{U} b_{s,u}(t)\sqrt{p_{s,u}(t)}x_{s,u}(t) \tag{1}$$

where $x_{s,u}(t)$ and $p_{s,u}(t)$ represent the data signal and allocated power of user u on subchannel s, respectively. $b_{s,u}(t)=1$ indicates that subchannel s is allocated to user u, and vice versa. The received signal can be expressed as,

$$y_{s,u}(t) = b_{s,u}(t)h_{s,u}(t)\sum_{u=1}^{U}\sqrt{p_{s,u}(t)}x_{s,u}(t) + \sum_{q=1,q\neq u}^{U} b_{s,q}(t)\sqrt{p_{s,q}(t)}x_{s,q}(t) + z_{s,u}(t) \tag{2}$$

where $h_{s,u}(t) = g_{s,u} PL^{-1}(d_{s,u})$ denotes the channel gain between the base station and user u on subchannel s. Assume that $g_{s,u}$ is Rayleigh fading channel gain [26], $PL^{-1}(d_{s,u})$ is the path loss, and $d_{s,u}$ is the distance between user u and base station on channel s. $z_{s,u}(t)$ represents additive white Gaussian noise which follows the complex Gaussian distribution, i.e. $z_{s,u}(t) \sim CN(0, \sigma_n^2)$.

In the NOMA system, due to the interference introduced by superimposed users, successive interference cancellation (SIC) technique is required to eliminate interference at the receiver. Firstly, the receiver decodes the users with high power levels, then subtracts it from the mixed signal, repeats this process until the desired signal has the maximum power in the superimposed signal, and regards the rest as interference signals. As a result, the signal to interference plus noise ratio (SINR) can be described as,

$$SINR(t) = \frac{b_{s,u}(t) p_{s,u}(t) |h_{s,u}(t)|^2}{\sum_{u=1, |h_{s,q}(t)|^2 < |h_{s,u}(t)|^2}^{U} b_{s,q}(t) p_{s,q}(t) |h_{s,q}(t)|^2 + \sigma_n^2} \tag{3}$$

The data rate of user u on subcarrier s is defined as,

$$R_{s,u}(t) = \frac{B}{S} \log_2(1 + SINR(t)) \tag{4}$$

The user sum rate is,

$$R = \sum_{u=1}^{U} R_{s,u}(t) \tag{5}$$

The optimization objectives and constraints of the joint user grouping and power allocation problem are given as follows,

$$\begin{aligned}
&\text{P1:} \quad \max \ R \\
&\text{C1:} \quad 0 \le \sum_{s=1}^{S} p_{s,u}(t) \le P_{max}, s \in S, u \in U \\
&\text{C2:} \quad b_{s,u}(t) \in \{0,1\}, s \in S, u \in U \\
&\text{C3:} \quad \sum_{s=1}^{S} b_{s,u}(t) \le 1, s \in S, u \in U \\
&\text{C4:} \quad \sum_{u=1}^{U} b_{s,u}(t) \le C, s \in S, u \in U
\end{aligned} \tag{6}$$

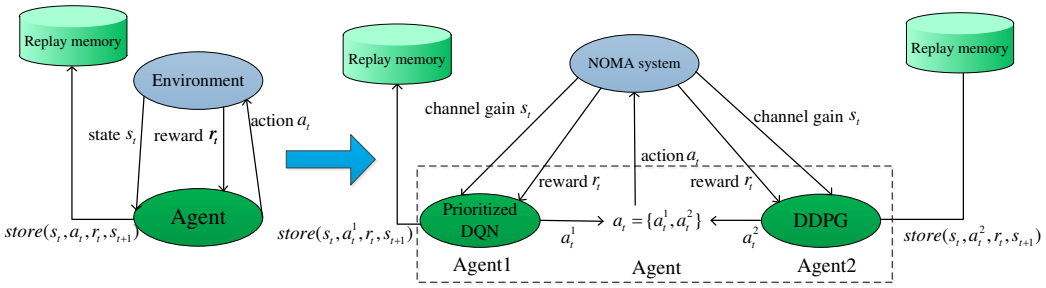
In the above constraints, C1 indicates that the power allocated to each user should be less than the maximum power transmitted by the base station. C3 and C4 indicate that multiple users can be placed on one subchannel. Because this objective function is a Non-Convex optimization problem, it is difficult to find the global optimal solution. Although the global search method can find the optimal solution by searching all the grouping possibilities, the computational complexity is too high to apply in practice. Therefore, a DRL-based method is proposed for user grouping and power allocation in the NOMA system.

2.2 NOMA Resource Allocation Based on DRL Network

In this section, a NOMA resource allocation network based on DRL network is proposed. The description of the system structure is given in the following subsections.

2.2.1 System structure

The system structure is shown in Fig.1. Fig.1 (a) is a general reinforcement learning network structure. The general reinforcement learning is mainly divided into the following five parts: agent, environment, state s_t, action a_t, and immediate reward r_t. The learning process of reinforcement learning can be described as: the agent obtains the state s_t from the environment, and then selects an action a_t from the action space and feeds it back to the environment. At this time, the environment generates a reward r_t, which is generated by choosing this action a_t in the current state s_t, and also generates state s_{t+1} of the next time slot. Then the environment gives them back to the agent. The agent stores learning experience in the experience replay pool to facilitate learning in the next time slot.



(a) Reinforcement learning (b) Proposed system structure

Fig.1 System structure.

According to the structure of reinforcement learning, the system model designed in this paper is shown in Fig.1 (b). Specifically, the NOMA system represents the environment of reinforcement learning. There are two agents, the Prioritized DQN user grouping network represents agent 1, and the DDPG power allocation network represents agent 2. We use channel gain as a characterization of the environment. Accordingly, the state space can be expressed as $S = \{h_{1,1}(\mathrm{t}), h_{2,1}(\mathrm{t}), ..., h_{s,u}(\mathrm{t})\}$, The user group space can be expressed as $A1 = \{b_{1,1}(t), b_{2,1}(t), ..., b_{s,u}(t)\}$, power allocation space are $A2 = \{p_{1,1}(t), p_{2,1}(t), ..., p_{s,u}(t)\}$. Besides, the immediate reward is denoted as $r_t = R$, where R is the system sum rate defined in (5). Our goal is to maximize long-term rewards, which is expressed as,

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ... = \sum_{i=0}^{\infty} \gamma^i r_{t+i}, \quad \gamma \in [0,1] \tag{7}$$

where γ is the fading factor. When γ=0, it means that the agent only pays attention to the reward generated in the current state; when $\gamma \neq 0$, it means that the agent also pays attention to future reward, and future rewards take more weight as γ increases.

The expected value of cumulative return R_t (obtained by (7)) of general

reinforcement learning is defined as the Q value, which is determined by state s_t, and the selection of action a_t under a certain strategy π. It is expressed as,

$$Q_\pi(s_t, a_t) = \mathrm{E}[r_t + \gamma \max Q_\pi(s_{t+1}, a_{t+1}) \mid s_t, a_t] \tag{8}$$

In summary, in each time slot (TS), the agent obtains the channel gain from the NOMA system, selects user combination and power in the action space according to current channel gain, and gives the action (optimal user groups and power) result back to the NOMA system. According to the received action, the NOMA system generates immediate reward and the channel gain of the next time slot, and then passes them to the agent. According to the reward, the agent updates the decision function of selecting this action under the current channel gain, which completes an interaction. Repeat this process until the agent can generate an optimal decision under any channel gain. The specific design of the Prioritized DQN and DDPG network in Fig.1 (b) is illustrated in Fig.2, and the detail description of them will be given in the following subsections.

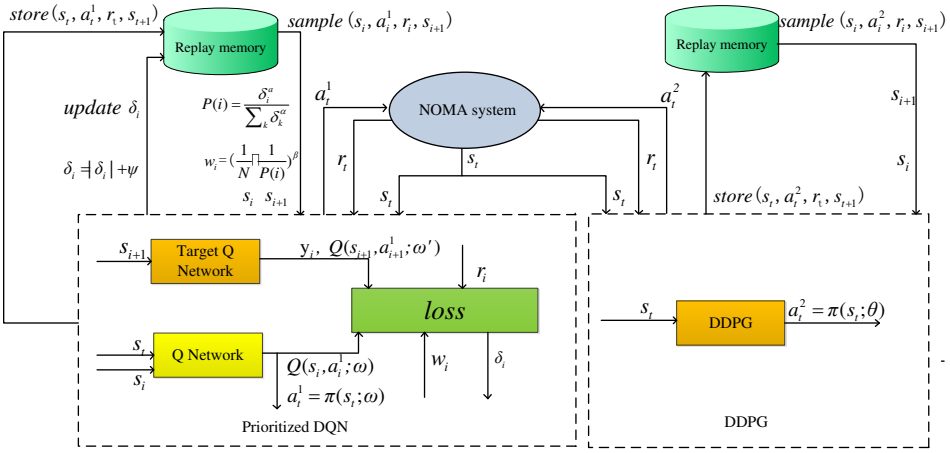


Fig.2 DRL-based NOMA resource allocation system model.

2.2.2 User grouping based on Prioritized DQN

In this article, we use the Prioritized DQN to perform user grouping, which is an improved network of DQN. The DQN includes two networks, the Q network generates the estimated Q value, and the target Q network generates the target Q value. The main idea of the DQN algorithm is to continuously adjust the network weight by optimizing the loss function produced by the estimated Q value and the target Q value. Moreover, experience replay is used in the DQN to reduce the correlation between samples. In the DQN, all the samples are uniformly sampled from the experience replay pool. In this case, some important samples may be neglected, which will reduce the learning efficiency. In order to make up for the shortcomings of the random sampling from experience pool, a reinforcement learning method based on prioritized experience replay is proposed, which mainly solves the sampling problem in experience replay [27]. The main idea is to set priorities for different samples to increase the sampling probability of valuable samples. In this paper, we use Prioritized DQN to perform user grouping. In order to better understand the algorithm, we first introduce the prioritized experience replay knowledge.

(1) Prioritized experience replay

Temporal-difference error (TD-error) indicates the difference between the output action value and the estimated value. TD-errors produced by different samples are different, and their effects on backpropagation are also different. A sample with large TD error indicates that there is a big gap between the current value and the target value, which means that the sample needs to be learned and trained. Therefore, in order to measure the importance of the sample, we use TD-error to represent the priority of sample, which can be expressed as,

$$\delta_i = |y_i - Q(s_i, a_i; \omega)| + \psi \tag{9}$$

where δ_i is the TD-error of sample i, ψ is a very small constant to ensure that samples with a priority of 0 can be selected, y_i is the target value defined in (14).

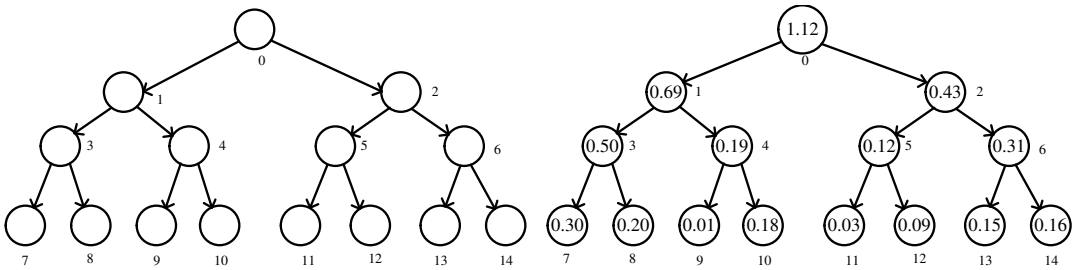
By setting priorities for samples, samples with large TD-errors may be sampled with high probabilities, and they will join the learning process more frequently. In contrast, samples with small TD-errors may not be replayed at all, because the TD-errors of them cannot be updated every time and are always small. In this case, the diversity of samples will be lost and result in over-fitting. It is necessary to ensure that the sample with low priority can be sampled with a certain probability. Therefore, a probability of occurrence is defined as [23],

$$P(i) = \frac{\delta_i^a}{\sum_k \delta_k^\alpha} \tag{10}$$

where α determines the degree of priority. α=0 means uniform sampling.

A. Sum tree

After setting the priority, the samples can be sampled according to the occurrence probability. Prioritized DQN uses a sum tree to solve the problem of sorting samples before sampling. The sum tree is a binary tree, and the structure is shown in Fig.3 (a). The top of the sum tree is the root node, each tree node has only two child nodes, the bottom layer is the leaf node, and the remaining nodes are internal nodes. The number in the figure is the index of the node, which starts from the root node 0. We also use an array T to store the corresponding sample tuples, and the structure is shown in Table 1, where *idx* represents the index of the sample tuple in the array T.



(a) Sum tree before storing priority. (b) Sum tree after storing priority.

Fig.3 Sum tree.

Table1 Array T

T	(s,a,r,s')	(s,a,r,s')	(s,a,r,s')	(s,a,r,s')	.....
idx	0	1	2	3	.....

B. Storing data

Assuming that the layer number is denoted as l, and the number of layers of the binary tree is L, the number of nodes at each layer can be expressed as $2^{l-1}\,(l=1,2,.....,L)$, and the total number of nodes in the binary tree is 2^L-1. It can be found in Fig.3 (a) that the number of the leftmost leaf node can be expressed as $2^{l-1}-1\,(l=1,2,.....,L)$. The index of array T corresponding to the number of the leftmost leaf node is 0, which is denoted as idx = 0. When a priority is stored, the number of the leaf node and idx is increased by 1. The sum tree only stores the priorities of samples at the leaf node, i.e., the nodes at the L^{th} layer, and the priority of a leaf node is matched to a sample tuple in array T. In addition, the priority of an intermediate layer node is the sum of the priorities of its child nodes, and the priority of the root node is the sum of the priorities of all the nodes. The larger the value of leaf nodes is, the higher the priority of samples is. The priority of the sample is stored in the leaf node from left to right. The storage steps are given as follows:

(1) Number the 2^L-1 nodes of the sum tree, and initialize the priorities of all the leaf nodes of the sum tree to 0;

(2) The priority of the current sample is stored in the leftmost leaf node, and the current sample tuple is stored in array T of which the index is idx=0. At the same time, the priorities of the parent nodes of the whole binary tree are updated upward;

(3) Add the priority of the sample at the second leaf node of the sum tree. Then, the number of the leaf node can be expressed as 2^{L-1} (obtained by (2^{L-1}-1)+1). The index of array T corresponding to this leaf node is 1 (obtained by 0+1). Then, add the sample tuple to array T of which the index is 1. Update the priorities of the parent nodes of the whole binary tree upward;

(4) According to the storage method above, the priorities of the samples are added to the leaf nodes one by one. When all the leaf nodes are filled, the subsequent priority will be stored in the first leaf node again.

The difference between the leaf node number of the sum tree and the index of the corresponding T is 2^{L-1}-1. The binary tree after storing the priority is shown in Fig.3 (b). The leaf node numbered 7 has the highest priority, and this indicates that this node has the largest probability of being sampled.

C. Sampling data

Denote the number of samples to be extracted as N, and the priority of the root node as P. Divide P by N, and the quotient M is obtained. Hence the total priority is divided into N intervals, and the j^{th} interval is between [(j-1)*M, j*M]. For example, if the priority of the root node is 1.12 and the number of samples is 8, the priority interval can be expressed as [0,0.14], [0.14,0.28], [0.28,0.42], [0.42,0.56], [0.56,0.70], [0.70,0.84], [0.84,0.98], and [0.98,1.12]. Sample a piece of data uniformly in each

interval, and suppose that 0.60 is extracted in the interval [0.56,0.70]. Start traversing from the root node and compare 0.60 with the left child node 0.69. Since the left child node 0.69 is larger than 0.60, take the path of the left child node and traverse its child nodes. Then compare 0.60 with the left child node 0.50 of the node 0.69. Since 0.60 is larger than 0.50, subtract 0.50 from 0.60 to enter the right child node and traverse its child nodes. Compare 0.10 with the left child node of 0.19. Because 0.10 is larger than 0.01, take the path of right child node. Finally, the priority of the sample is 0.18, and the leaf node number of the sum tree is 10. At the same time, the sample corresponding to this leaf node is extracted from array *T*. After that, a number is uniformly selected from each interval, and then hold this number to sample samples according to the above-mentioned method. Finally 8 samples are sampled.

D. Importance sampling

The distribution of the samples used to train the network should be the same as its original distribution. However, since we tend to replay experience samples with high TD-errors more frequently, the sample distribution will be changed. This change causes a bias in the estimated value, and experience samples with high priority may be used to train the network more frequently. Importance sampling is used to adjust and update the network model by reducing the weight of the sample, so that the introduced error can be corrected [28]. The weight of importance sampling is,

$$w_i = (\frac{1}{N}.\frac{1}{P(i)})^{\beta} \tag{11}$$

where N is the number of samples, $P(i)$ is the probability of the sample which is calculated according to (10), β is used to adjust the degree of deviation. The slight deviation can be ignored at the beginning of learning. The effect of importance sampling to correct deviation is from small to large, so it increases linearly from the initial value, and converges to 1 at the end of training. When $\beta=1$, it indicates that the deviation has been completely eliminated.

Fig.4 shows the relationship between β and the number of iterations (the initial value is 0.4). It can be seen from the figure that at the end of the iteration, β can converge to 1, which means that the non-uniform probability is completely compensated, and the deviation caused by prioritized experience replay can be corrected.

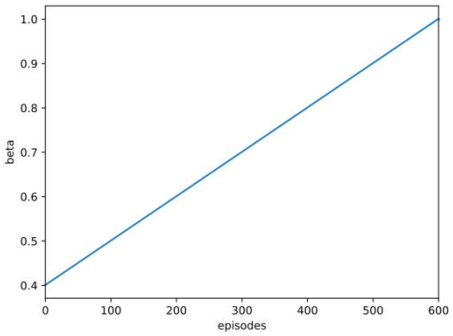


Fig.4 The relationship between β and number of iterations

In order to ensure the stability of learning, we always normalize weights, so (11) can be rewritten as,

$$w_i = \frac{(N \cdot P(i))^{-\beta}}{\max_j w_j} = \frac{(N \cdot P(i))^{-\beta}}{\max_j [(N \cdot P(j))^{-\beta}]}$$
$$= \frac{(P(i))^{-\beta}}{\max_j [(\frac{1}{P(j)})^{\beta}]} = \left(\frac{(P(i))}{\min_j (P(j))} \right)^{-\beta} \tag{12}$$

(2) Prioritized-DQN based user grouping network

In this section, we introduce the user grouping framework based on Prioritized DQN. As shown in Fig.2, the user grouping part contains prioritized experience replay. Prioritized DQN contains two sub-networks, a Q Network is used to generate the estimated Q value of the selected action, and a Target Q Network to generate the target Q value for training the neural network.

In our NOMA system, at the beginning of each TS t, the base station receives channel state information s_t, and inputs it into the estimated Q Network of the Prioritized DQN. With s_t as input, the Q Network outputs all user combinations a_t^1 $(a_t^1 \in A1)$ and estimated Q value $Q(s_t, a_t^1; \omega)$. In this paper, the $\zeta - greedy$ strategy is used to select user combination a_t^1, which randomly selects a user combination from A1 with probability ζ, or a user combination with the highest estimated Q value with probability $(1-\zeta)$. That is,

$$a_t^1 = \arg\max_{a_t^1 \in A1} Q(s_t, a_t^1; \omega) \tag{13}$$

Finally, the user combination a_t^1 and power a_t^2 (produced by the next section) are given back to the NOMA system. According to the selected actions, the NOMA system generates instant rewards r_t and channel state information s_{t+1} of the next time slot. We store the sample tuple $(s_t, a_t^1, r_t, s_{t+1})$ of each TS into the memory block.

In each TS, in order to ensure that all samples can be sampled, Prioritized DQN sets the new samples to the highest priority, and stores the sample tuples and priorities in the experience pool according to the storage steps in section 3.2.1. Select sample tuples according to the sampling method in section 3.2.1. As mentioned above, we use the probability of being sampled to calculate the sample weight (i.e., (12)), and use the target Q network to generate the target Q value for training the network, which is,

$$y_i = r_i + \gamma \max_{a_{i+1}^1 \in A1} Q(s_{i+1}, a_{i+1}^1, \omega') \tag{14}$$

The loss function of Prioritized DQN can be expressed as,

$$loss = \frac{1}{N} \sum_{i=1}^{N} w_i (y_i - Q(s_i, a_i^1; \omega))^2 \tag{15}$$

Update all the weights ω of the Q network in the Prioritized DQN through gradient backpropagation, and update all the parameters of the target Q network by

copying the parameters of their corresponding network in every W TS, i.e. $\omega' = \omega$.

After the parameters of the Q network of the Prioritized DQN are updated, it is necessary to recalculate the TD-error (i.e., (9)) of all the selected samples. Find the corresponding leaf node according to the number of the leaf node obtained by sampling, and set the TD-error to the priority of the sample. Follow the same method of storing data to update the priority of the sum tree leaf node and the priority of all its parent nodes.

2.2.3 Power allocation based on DDPG network

Since the output of DQN is discrete, it cannot be applied to a continuous action space. Fortunately, an Actor-Critic-based DDPG network can handle continuous actions. Wang [29] proposed two frameworks (i.e., DDRA and CDRA) to maximize the energy efficiency of the NOMA system, where DDRA is based on the DDPG network and CDRA is based on multi-DQN. The results showed that the time complexities of the two frameworks were similar, and the performance of the DDPG network was better than that of the multi-DQN network. This is because the user power is quantified in multi-DQN, and some important information is lost, which causes quantization errors and results in poor performance. DDPG network is similar to DQN, using deep neural network and uniform sampling. It is also a deterministic policy gradient network, in which the action is uniquely determined in one state. Moreover, DDPG can handle continuous action tasks without quantifying the transmitted power. Hence this paper uses the DDPG network to perform the user's power allocation task.

3 Results and Discussion

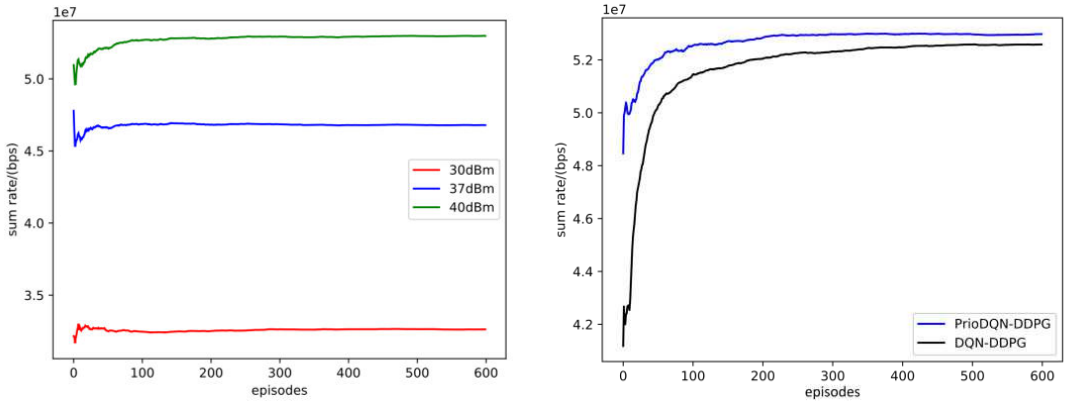
This section shows the simulation results of the above-mentioned DRL based NOMA system user grouping and power allocation algorithms. Assume that there are 4 users to transmit signals on 2 channels, among which 4 users are randomly distributed in a cell with a radius of 500m and the minimum distance between the user and the base station is 30m. The total system bandwidth is 10MHz, and the noise power density is -110dBm/Hz. The maximum transmission power transmitted of the base station is 40dBm, and the minimum power is 3dBm.

In the Prioritized DQN, the number of leaf nodes is 500, the number of samples N is 32, the reward discount factor γ is 0.9, and the greedy selection strategy probability setting ς is 0.9. Set the initial deviation degree β=0.4 and the priority degree α=0.6 in the prioritized experience replay.

3.1 Convergence of the proposed algorithm

Fig.5 shows the convergence performance of the proposed algorithm. Fig.5 (a) shows the convergence of the sum rate of the proposed algorithm under different maximum transmission powers of the base station. Set the maximum transmission power of base station to 30dBm, 37dBm, and 40dBm, respectively. It can be observed that under different base station transmission powers, the sum rate of the system

gradually increases and then tends to converge, which proves that the proposed algorithm has good convergence.



(a) Sum rate of the proposed algorithm under different maximum powers.

(b) Sum rate comparison of PrioDQN-DDPG and DQN-DDPG (The maximum transmission power of the base station $P_{\max} = 40$dBm).

Fig.5 Convergence analysis of the proposed algorithm.

Fig.5 (b) compares the common DQN user grouping algorithm and analyzes the convergence performance of the proposed Prioritized DQN. Power is allocated to users based on the DDPG network. It is clear that the algorithm with prioritized sampling can reduce the training time and make the learning process more stable. It takes about 100 episodes for prioritized experience replay to complete the user grouping task, while around 300 episodes for the uniform experience replay to complete the same task. This is because prioritized experience replay stores the learning experience with priority in the experience pool, and traverses the sum tree to extract samples with high TD-errors to guide the optimization of model parameters, which alleviates the problem of sparse reward and insufficient sampling strategies, and improves learning efficiency. Also, prioritized experience replay not only focuses on samples with high TD-error to help to speed up the training process, but also involves samples with lower TD-error to increase the diversity of training.

3.2 Average sum rate performance of the proposed algorithm

The NOMA system resource allocation algorithm based on Prioritized DQN and DDPG proposed in this paper is denoted as PrioDQN-DDPG. In order to verify the effectiveness of the proposed algorithm, this paper compares several resource allocation algorithms. The algorithms for comparison are ES-EPA, ES-MP, multi-DQN, DQN-DDPG and IPOP. Specifically, ES-EPA uses an exhaustive searching method to select the best user combination, and uses the average power allocation scheme. ES-MP also uses an exhaustive searching method to select the best user grouping, and the maximum power transmission algorithm is utilized to determine the power for each user. The multi-DQN algorithm [29] uses multiple DQN to quantify power, and DQN-DDPG algorithm selects the user grouping based on the DQN and allocates power for each user based on DDPG [30]. IPOP is an iterative power optimization algorithm [31], which finds the optimal solution by constructing a

Lagrangian dual function. Fig.6 shows the experimental results of user sum rates. All the experimental results are averaged every 200 TS to achieve a smoother and clearer comparison.

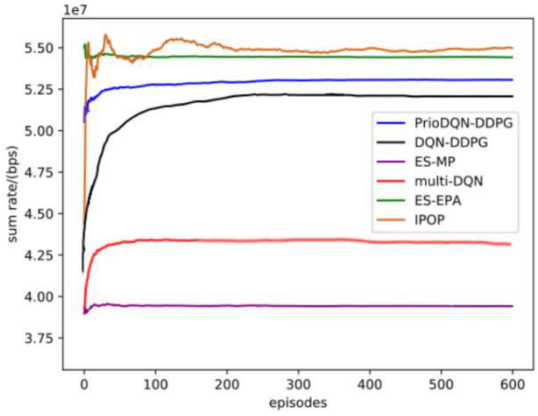


Fig.6 Sum rate comparison of different algorithms.

As can be seen from Fig.6, The sum rate of the ES-MP algorithm is lower than those of the other algorithms. This is because the power allocated to each user is the allowed maximum power, and strong interferences are caused among users. The performances of the multi-DQN, DQN-DDPG and PrioDQN-DDPG algorithms are getting better as the number of episodes increases. The PrioDQN-DDPG algorithm proposed in this paper is better than the other two DRL-based algorithms. Respectively, compared with the DQN-DDPG algorithm, the algorithm proposed in this paper improves the system sum rate by 2%. This is mainly because Prioritized DQN sets the priority for some valuable samples that are beneficial to training the network; moreover, prioritized DQN uses the sum tree to store the priority, so that it is convenient to search experience samples with high priorities, and the valuable experience could be replayed more frequently, which can improve the learning rate and system sum rate. Compared with the multi-DQN algorithm, this article uses the DDPG network to complete the user's power allocation. DDPG can handle continuous action tasks and solves the problem of quantization errors caused by quantization power.

Furthermore, the PrioDQN-DDPG framework proposed in this paper interacts with the NOMA system, finds the optimal resource allocation strategy based on system feedback, and can dynamically find the optimal resource allocation strategy according to the changes of environment, which can reach 93% of the IPOP, and can reach 94% of the ES-EPA algorithm. Although the sum rate of the proposed PrioDQN-DDPG network is lower than IPOP and ES-EPA, it can greatly reduce the computational complexity, which will be discussed in the following part.

3.3 Computational complexity analysis

This section analyzes the computational complexity of the proposed algorithm, and Fig.7 shows the result. The computational time complexity of the algorithm in this paper is 10% higher than that of the traditional DQN-DDPG algorithm. This is

because the prioritized experience replay algorithm is mainly composed of setting sample priority, storing experience samples and extracting samples. It needs extra time to calculate the TD-error and traverse the sum tree. However, the prioritized experience replay algorithm can replay valuable samples frequently, which avoids some unnecessary DRL processes and reduces training time, compared with the optimal ES-EPA and IPOP algorithms, the computational complexity is reduced by 43% and 64%, respectively.

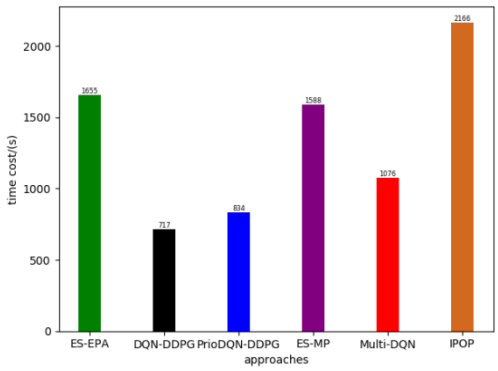


Fig.7 The average time cost comparison of different approaches

4 Conclusion and future work

This paper proposes a joint user grouping and power allocation algorithm to solve the resource allocation problem in the multi-user NOMA system. While ensuring the minimum data rate of all users, we use a DRL-based framework to maximize the sum rate of the NOMA system. In particular, with the current channel state information as input and the sum rate as the optimization goal, we design a Prioritized-DQN-based network to output optimal user grouping strategy, and then use a DDPG network to output the power of all users. The proposed algorithm uses prioritized experience replay to replace previous uniform experience replay, which uses TD-error to evaluate the importance of samples, and uses the binary-tree-based priority queue to store experience. The proposed sampling method allows the samples that are more useful for the learning process to be replayed more frequently. The simulation results show that the proposed algorithm with prioritized sampling can replay valuable samples at a high probability and increase the learning rate. In the power allocation part, there is no need to quantify the transmission power, and the powers of all users are directly output under the current state information. In addition, the joint algorithm proposed in this paper improves the sum rate of the system by 2% compared with the ordinary DQN algorithm, and reaches 94% and 93% of the optimal exhaustive search algorithm and iterative power optimization algorithm, respectively.

In addition, with the promising development prospects of NOMA on complex channels, how to allocate resources for cell-free massive MIMO-NOMA networks is a research focus of this article in the future.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

The picture materials quoted in this article 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.

Funding

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

Authors' Contributions

YL proposed the framework of the whole algorithm; ML performed the simulations, analysis and interpretation of the results. XF and ZL 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
Mengli He, Yue Li, Xiaofei Wang, Zelong Liu

Corresponding author

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

Abbreviations

IoV: Internet-of-vehicle

NOMA: Non-orthogonal multiple access

SIC: Successive interference cancellation

BS: Base station

DRL: Deep reinforcement learning

SC: Superposition coding

TS: Time slot

DQN: Deep Q network

DDPG: Deep deterministic policy gradient network

TD-error: Temporal-difference error

IPOP: Iterative power optimization

References

[1] W. U. Khan, M. A. Javed, T. N. Nguyen et al., Energy-efficient resource allocation for 6G backscatter-enabled NOMA IoV networks. IEEE Trans. Intell. Transp. Syst. (2021)

[2] X. Liu, M. Jia, X. Zhang et al., A novel multichannel Internet of things based on dynamic spectrum sharing in 5G communication. IEEE Internet Things J. 6(4), 5962-5970 (2018)

[3] X. Liu, X. J. I. I. o. T. J. Zhang, Rate and energy efficiency improvements for 5G-based IoT with simultaneous transfer. IEEE Internet Things J. 6(4), 5971-5980 (2018)

[4] X. Liu, X. Zhang, M. Jia et al., 5G-based green broadband communication system design with simultaneous wireless information and power transfer. Phys. Commun. 28, 130-137 (2018)

[5] S. Han, Y. Huang, W. Meng et al., Optimal power allocation for SCMA downlink systems based on maximum capacity. IEEE Trans. Commun. 67(2), 1480-1489 (2018)

[6] K. Yang, N. Yang, N. Ye et al., Non-orthogonal multiple access: Achieving sustainable future radio access. IEEE Commun. Mag. 57(2), 116-121 (2018)

[7] S. R. Islam, N. Avazov, O. A. Dobre et al., Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges. IEEE Communications Surveys and Tutorials, 19(2), 721-742 (2016)

[8] 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)

[9] X. Liu, X. B. Zhai, W. Lu et al., QoS-guarantee resource allocation for multibeam satellite industrial Internet of Things with NOMA. IEEE Trans. Ind. Inform. 17(3), 2052-2061 (2019)

[10] X. Liu, X. J. I. t. o. i. i. Zhang, NOMA-based resource allocation for cluster-based cognitive industrial internet of things. IEEE Trans. Ind. Inform. 16(8), 5379-5388 (2019)

[11] S. R. Islam, M. Zeng, O. A. Dobre et al., Resource allocation for downlink NOMA systems: Key techniques and open issues. IEEE Wirel. Commun. 25(2), 40-47 (2018)

[12] A. Benjebbovu, A. Li, Y. Saito et al. System-level performance of downlink NOMA for future LTE enhancements. IEEE Globecom, 66-70(2013)

[13] H. Zhang, D.-K. Zhang, W.-X. Meng et al. User pairing algorithm with SIC in non-orthogonal multiple access system. Proc.Int.Conf.Commun, 1-6(2016)

[14] 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)

[15] G. Gui, H. Huang, Y. Song et al., Deep learning for an effective nonorthogonal multiple access scheme. IEEE Trans. Veh. Technol. 67(9), 8440-8450 (2018)

[16] M. Liu, T. Song, L. Zhang et al. Resource allocation for NOMA based heterogeneous IoT with imperfect SIC: A deep learning method. Proc. IEEE Annu. Int. Symp. Pers. Indoor Mobile Radio Commun, 1440-1446(2018)

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

[18] H. Huang, Y. Yang, Z. Ding et al., Deep learning-based sum data rate and energy efficiency optimization for MIMO-NOMA systems. IEEE Trans. Wirel. Commun. 19(8), 5373-5388 (2020)

[19] Q. Liu, J. W. Zhai, Z.-Z. Zhang et al., A survey on deep reinforcement learning. Chinese Journal of Computers, 41(1), 1-27 (2018)

[20] 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), 5083-5098 (2021)

[21] V. Mnih, K. Kavukcuoglu, D. Silver et al., Human-level control through deep reinforcement learning. nature, 518(7540), 529-533 (2015)

[22] 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)

[23] T. Schaul, J. Quan, I. Antonoglou et al., Prioritized experience replay. Proc. Int. Conf. Learning, Representations, (2015)

[24] T. P. Lillicrap, J. J. Hunt, A. Pritzel et al., Continuous control with deep reinforcement learning. ICLR, (2015)

[25] 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)

[26] I.-H. Lee, H. J. I. W. C. L. Jung, User selection and power allocation for downlink NOMA systems with quality-based feedback in rayleigh fading channels. IEEE Wirel. Commun. Lett. 9(11), 1924-1927 (2020)

[27] J. Zhai, Q. Liu, Z. Zhang et al. Deep q-learning with prioritized sampling. Proc. Int. Conf. Neural Inf. Process, 13-22(2016)

[28] A. R. Mahmood, H. Van Hasselt, R. S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. proceedings of the NIPS, 3014-3022(2014)

[29] 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)

[30] Y. Zhang, X. Wang, Y. Xu. Energy-efficient resource allocation in uplink NOMA systems with deep reinforcement learning. Proc. International Conference on Wireless Communications and Signal Processing (WCSP), 1-6(2019)

[31] 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)