Journal of Beijing Institute of Technology, 2025, Vol. 34, No. 3

Relay Selection for Cooperative NOMA Systems Based on the DQN Algorithm

Ying Lin✉, Yongwei Xiong, Xingbo Gong, Sifei Zhang, Yinhang Tian
(School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: In this study, a solution based on deep Q network (DQN) is proposed to address the relay selection problem in cooperative non-orthogonal multiple access (NOMA) systems. DQN is particularly effective in addressing problems within dynamic and complex communication environments. By formulating the relay selection problem as a Markov decision process (MDP), the DQN algorithm employs deep neural networks (DNNs) to learn and make decisions through real-time interactions with the communication environment, aiming to minimize the system's outage probability. During the learning process, the DQN algorithm progressively acquires channel state information (CSI) between two nodes, thereby minimizing the system's outage probability until a stable level is reached. Simulation results show that the proposed method effectively reduces the outage probability by 82% compared to the two-way relay selection scheme (Two-Way) when the signal-to-noise ratio (SNR) is 30 dB. This study demonstrates the applicability and advantages of the DQN algorithm in cooperative NOMA systems, providing a novel approach to addressing real-time relay selection challenges in dynamic communication environments.

Keywords: deep Q network (DQN); cooperative non-orthogonal multiple access (NOMA); relay selection; outage probability

1 Introduction

In recent years, smart devices have become prevalent in daily life and have captured a significant share of the market. However, smart devices impose stringent demands on the quality of service when exchanging information. This situation creates additional challenges in non-orthogonal multiple access (NOMA) technology, which has garnered considerable attention due to its advantages in enhancing spectrum utilization and system capacity [1, 2]. Cooperative NOMA systems employ relay nodes to assist communication, which facilitates the expansion of the coverage area for transmitted signals and enhances the success rate of signal transmission [3]. Consequently, this approach improves the transmission reliability of the system and elevates the signal quality for users. Outage probability is one of the key metrics for assessing the communication service quality and reliability of a system. To minimize the outage probability in cooperative NOMA systems, two common approaches are typically employed: optimizing the relay selection strategy and the power allocation scheme.

For multi-relay scenarios, it is often possible to choose the transmission path of the optimal relay, which can achieve greater system reliability and higher transmission efficiency [4, 5]. In [6], the authors proposed a two-way relay selection scheme (Two-Way) based on the coopera-

Manuscript received Nov. 27, 2024; revised Feb. 19, 2025; accepted Apr. 9, 2025. The associate editor coordinating the review of this manuscript was Dr. Yixiao Gu. This work was supported by the National Natural Science Foundation of China (Nos. 61841107 and 62061024), Gansu Natural Science Foundation (Nos. 22JR5RA274 and 23YFGA0062), Gansu Innovation Foundation (No. 2022A-215).
✉ Corresponding author. Email: linying@lut.edu.cn
DOI: 10.15918/j.jbit1004-0579.2024.116

tive NOMA communication system, and the simulation results showed that the outage probability of the proposed strategy was reduced compared with that of the traditional relay selection strategy. Moreover, the optimal diversity gain was obtained, which was more significant than that of the traditional orthogonal multiple access (OMA). In [7] and in [8], the authors employed amplify-and-forward (AF) and decode-and-forward (DF) techniques at the relay to implement collaborative transmission under partial relay selection schemes, respectively. The results showed that the total system rate and user fairness were improved. In [9], the authors proposed a distributed passive relay selection strategy to enhance the outage performance of the system. They simulated and analyzed the effect of changes in the location and number of relay nodes on the outage probability of the system. In [10], the authors proposed a relay selection strategy based on single-source optimal paths to obtain relay users with superior channel conditions. This strategy aimed to eliminate ineffective paths, thereby reducing system time wastage and decreasing outage probability, ultimately achieving full coverage of the NOMA communication network in the designated region. In [11], the authors analyzed the security outage probability of the Two-Way relay selection and optimal relay selection schemes in cooperative NOMA systems. The results showed that the security outage probability of the Two-Way relay selection scheme was equal to the security outage probability of the optimal relay selection scheme. In [12], the authors employed a partial relay selection algorithm to identify the optimal relay for data forwarding. The results indicated that the partial relay approach outperformed the multiple relay approach in the context of OMA techniques. In [13], the authors investigated the impacts of relay selection schemes on cooperative underlay cognitive radio non-orthogonal multiple access networks, where the max-min relay selection scheme (Max-Min) and the Two-Way relay selection scheme are applied in the network. Monte Carlo simulations were presented to validate the accuracy of the numerical results. In [14], the authors addressed the issue of inter-cell interference arising from multi-cell scenario. They proposed a novel relay selection criterion to ensure that each selected relay signal not only experienced favorable channel conditions when transmitted to its corresponding user but also minimized its impact on the performance of other users.

Reinforcement learning (RL) is currently applied in many fields [15, 16]. Several researchers have successfully implemented RL in the domain of cooperative NOMA. RL does not necessitate an exact model of the system. Instead, it acquires information and learns optimal strategies through interactions with the environment. During these interactions, it dynamically adjusts based on the changing environment, uncovering implicit patterns and measures to autonomously optimize overall performance. In [17], the authors proposed a Q-learning based relay selection scheme in a multi-hop network communication scenario. Simulation results demonstrated that the scheme achieved a near-optimal average end-to-end rate while reducing computational complexity and signaling overhead, thereby attaining high performance at a limited cost. In [18], the authors integrated the annealing algorithm with RL to enhance the performance of relay node selection. Simulation results indicated that the proposed relay selection scheme converged more rapidly than both classical RL and random selection algorithms as the number of iterations increased. However, the above study had the limitation of being applicable solely to simple problems with low dimensionality. In [19], the authors proposed a DRL based relay selection scheme for wireless sensor networks, which selected the optimal relay based on outage probability and channel state. The results demonstrated that the proposed method achieved a low system outage probability. In [20], the

authors designed a relay selection scheme based on deep Q network (DQN) to enhance the security of cooperative vehicular networks. This scheme selected the optimal relay from multiple candidates by considering the corresponding outages probabilities, leading to increased secrecy capacity. In [21], the authors proposed a relay selection scheme based on supervised machine learning to reduce the complexity of the system throughput maximization problem. The results demonstrated that the performance of the proposed scheme closely approximated that of the optimal exhaustive search method.

This paper proposes a DQN approach to address the relay selection problem in cooperative NOMA communication systems, aiming to minimize the outage probability. Unlike traditional relay selection methods, the approach proposed in this paper does not rely on predefined channel state information (CSI). Instead, it acquires real-time CSI through interactions with the environment. RL is integrated with neural networks, utilizing the action-value function to represent the immediate reward for executing a given action in the current state. Neural networks are employed to iteratively approximate the action-value function, thereby optimizing the desired objective function. Specifically, the contributions of this paper can be summarized as follows:

1) This paper optimizes the relay selection strategy in the cooperative NOMA system, models the relay selection problem as a Markov decision process, and uses the DQN algorithm to autonomously select the optimal relay node to minimize the system outage probability;

2) This paper constructs a DQN-based reinforcement learning framework, approximates the optimal strategy through a deep neural network, and sets the state space, action space, and reward function based on the outage probability constraint. The system can still efficiently select the optimal relay node in a complex channel environment, changing the limitations of the traditional relay selection method with channel state information;

3) Meanwhile, the traditional relay selection method and the Q-learning relay selection method are introduced to compare with the proposed method, verifying the superiority of the proposed method. The results show that the method proposed in this paper achieves a lower outage probability, showing the effectiveness of the DQN algorithm in the optimization of the collaborative NOMA system.

The remainder of the paper is organized as follows: Section 2 provides a description of the system model. In Section 3, the relay selection method based on the DQN approach is discussed. Section 4 presents the simulation results, while the conclusion is provided in Section 5.

2 System Model

The multi-relay communication system model for downlink NOMA is illustrated in Fig. 1. The system comprises a base station (BS), M decode-and-forward relay nodes (R_m, where m=1, 2, ···, M), and a destination node, which includes User 1 and User 2. Assuming that the source nodes and destination nodes are located at a significant distance from one another, no direct link exists due to obstructing obstacles. Relay nodes are required to facilitate signal transmission. The communication from the BS to the destination node via the selected relay node occupies two time slots. The entire system operates in half-

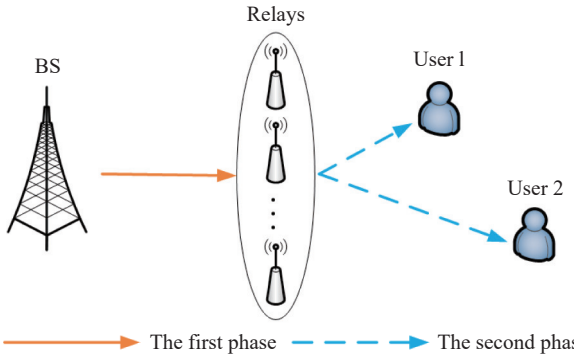

Fig. 1 System model for downlink NOMA with multiple relays

duplex mode. The channels among the nodes are Rayleigh fading channels. The channel gain of each link is generated based on the Rayleigh distribution of independent homogeneous distributions. The channel gain is generated as follows

$$h_i \sim \text{Rayleigh}(\sigma), i = SR_m, R_mD_1, R_mD_2 \quad (1)$$

reflecting the multipath fading and random fluctuations that may occur in real wireless communications.

In the first time slot, the source node transmits the superimposed signals $\sqrt{\alpha_1}x_1 + \sqrt{\alpha_2}x_2$ to the relay node, where $\sqrt{\alpha_1}$ and $\sqrt{\alpha_2}$ represent the power allocation factors for User 1 and User 2, respectively. And $\alpha_1 + \alpha_2 = 1\,(\alpha_1 > \alpha_2)$. Considering the decoding requirements and assuming $\alpha_1 > 0.5$ for the sake of non-generality, x_1 and x_2 are the signals sent by the source node to User 1 and User 2, respectively. Consequently, the received signal at the relay node can be expressed as follows

$$y_{R_m} = h_{SR_m}\sqrt{P_S}\,(\sqrt{\alpha_1}x_1 + \sqrt{\alpha_2}x_2) + n_{R_m} \quad (2)$$

where P_s is the transmit power of the source node, h_{SR_m} is the channel gain between the base station and the relay node R_m, and $n_{R_m} \sim \text{CN}(0, \sigma^2)$ denotes the additive Gaussian white noise.

When employing the DF protocol, the relay node follows the NOMA decoding principle and treats x_2 as noise when decoding x_1. Following this, x_1 is eliminated after decoding, and x_2 is subsequently decoded. At this point, the instantaneous signal-to-noise ratios of the signals received at the relay node are expressed as follows

$$\gamma_{R_m}^{x_1} = \frac{\alpha_1\gamma|h_{SR_m}|^2}{\alpha_2\gamma|h_{SR_m}|^2 + 1} \quad (3)$$

$$\gamma_{R_m}^{x_2} = \alpha_2\gamma|h_{SR_m}|^2 \quad (4)$$

where $\gamma = \dfrac{P_S}{\sigma^2}$ is the system signal-to-noise ratio (SNR).

At the second time slot, the optimal relay R_m forwards the decoded superimposed signal $\sqrt{\beta_1}x_1 + \sqrt{\beta_2}x_2$ to User 1 and User 2. The received signal of User D_i $(i \in \{1, 2\})$ at the destination node is expressed as follows

$$y_{D_i} = \sqrt{P_R}h_{R_mD_i}\left(\sqrt{\beta_1}x_1 + \sqrt{\beta_2}x_2\right) + n_{D_i} \quad (5)$$

where P_R is the transmit power of the relay node, $\sqrt{\beta_1}$ and $\sqrt{\beta_2}$ are the redistribution power allocation factors for User 1 and User 2 respectively, $h_{R_mD_i}$ is the channel gain between the relay node R_m and the destination node, and $n_{D_i} \sim \text{CN}(0, \sigma^2)$ denotes the additive Gaussian white noise.

In NOMA systems, serial interference removal is generally employed at the receiver side to decode the signals. Let $P_S = P_R = P$. User 1 treats x_2 as interference while decoding its own signal, and the instantaneous signal-to-noise ratio for decoding x_1 is expressed as follows

$$\gamma_{R_mD_1}^{x_1} = \frac{\beta_1\gamma|h_{R_mD_1}|^2}{\beta_2\gamma|h_{R_mD_1}|^2 + 1} \quad (6)$$

For User 2 to decode its own signal, it must first decode the x_1 signal, and the received instantaneous signal-to-noise ratio is expressed as follows

$$\gamma_{R_mD_2}^{x_1} = \frac{\beta_1\gamma|h_{R_mD_2}|^2}{\beta_2\gamma|h_{R_mD_2}|^2 + 1} \quad (7)$$

After successfully decoding x_1, the instantaneous signal-to-noise ratio of decoding x_2 by eliminating x_1 and then decoding its own information is expressed as follows

$$\gamma_{R_mD_2}^{x_2} = \beta_2\gamma|h_{R_mD_2}|^2 \quad (8)$$

The relay node and the destination node aim to successfully transmit and decode the signal, ensuring that both time slots facilitate effective transmission. That is, the rate of information transmission in each time slot must exceed a specified interruption threshold; otherwise, the system may experience an interruption. Assuming that the target rates of User 1 and User 2 are ζ_1 and ζ_2, and the corresponding signal-to-noise

ratios are $\gamma_{\text{th1}} = 2^{2\zeta_1} - 1$ and $\gamma_{\text{th2}} = 2^{2\zeta_2} - 1$, the outage probability of the system can be expressed as follows

$$P_{\text{out}} = 1 - \Pr\left(\gamma_{Rm}^{x_1} > \gamma_{\text{th1}}, \gamma_{Rm\,D_1}^{x_1} > \gamma_{\text{th1}}\right) \Pr\left(\gamma_{Rm}^{x_2} > \gamma_{\text{th2}}, \gamma_{Rm\,D_2}^{x_1} > \gamma_{\text{th1}}, \gamma_{Rm\,D_2}^{x_2} > \gamma_{\text{th2}}\right) \tag{9}$$

The minimization of the system outage probability can be expressed as

$$\min P_{\text{out}} \tag{10}$$

When channel conditions are suboptimal or relay node selection is inadequate, transmission interruptions may occur. This hinders the successful completion of signal transmission. A higher outage probability leads to worse communication performance of the system. Therefore, the objective function of this paper is to minimize the outage probability.

3 Relay Selection Based on DQN

A DQN-based relay selection method is introduced to solve the problem in Eq. (10). In real communication environments, the CSI for each time slot is subject to dynamic changes and uncertainty. Traditional relay selection methods typically require precise mathematical modeling of the system. These methods often rely on fixed rules or simplistic strategy optimization. Consequently, they struggle to adapt to dynamic and variable communication environments, such as rapidly changing channel conditions, node heterogeneity, and power allocation for spectrum resources. In this context, traditional methods are inadequate for effectively addressing high-dimensional and dynamically evolving decision-making problems. Therefore, this paper introduces DQN as an innovative approach to address the aforementioned challenges. DQN combines the state feature extraction capabilities of deep learning with the decision optimization abilities of RL. This integration enables the system to autonomously learn efficient and intelligent relay selection strategies through interactions with the dynamically changing communication network environment. As shown in Fig. 2, the basic framework of DQN-based relay selection is presented. It consists of the environment, the Q network, the target network, the DQN loss, and the replay memory. The environment represents the communication system, providing the state and receiving actions decided by the Q network. The target network is regularly updated to improve learning accuracy. The DQN loss function calculates the difference between the predicted and target Q-values, guiding the network's updates. Replay memory stores past experiences, enabling efficient sampling to break data correlation and improve training stability. These components work together to optimize the relay selection process.

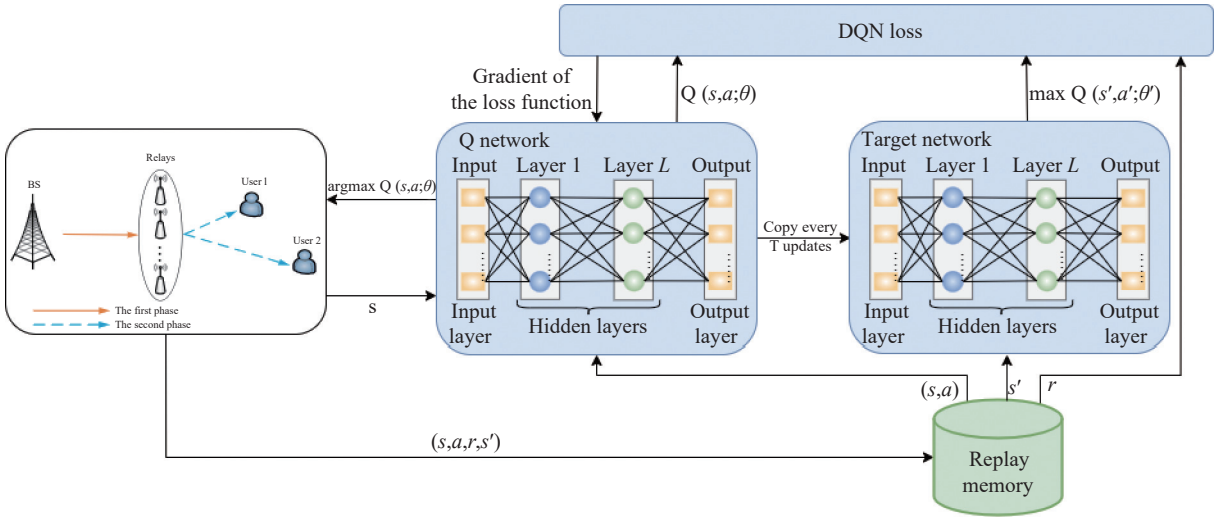

Fig. 2 Basic framework for DQN-based cooperative NOMA relay selection

3.1 Formulating the Relay Selection Problem as a Markov Decision Process (MDP)

In a multi-relay system, the channel states of each relay node are distinct. Consequently, it is necessary to select a relay node that optimizes the overall system performance to participate in the communication process. Minimizing the system's P_{out} is identified as a relay selection strategy employed in this paper. Based on the proposed optimization objective in Eq. (10), the communication process is modeled as an MDP, as specified below.

In this paper, the channel state during the communication process is defined as the state space, as follows

$$S = \left[h_{SR_m}^{(t)}, h_{R_m D_1}^{(t)}, h_{R_m D_2}^{(t)}, \gamma\right] \tag{11}$$

where $h_{SR_m}^{(t)}$, $h_{R_m D_1}^{(t)}$ and $h_{R_m D_2}^{(t)}$ denote the CSI for the S-R_m link, the R_m-D_1 link, and the R_m-D_2 link at time slot t, respectively, reflecting the dynamic variations in channel quality. $h_{SR_m}^{(t)}$ determines the data transmission rate and reliability from the source node to the relay node. $h_{R_m D_1}^{(t)}$ and $h_{R_m D_2}^{(t)}$ jointly influence the relay node's capability to simultaneously serve multiple users in a NOMA scenario, thereby affecting its ability to provide information transmission services to both users.

The selection of relay nodes is defined as the action space, shown as follows

$$A = \{a_t\}, \quad t = 1, 2, \cdots, t_{\max} \tag{12}$$

where $a_t \in A = \{1, 2, \cdots, m, m+1, \cdots, M\}$. If the action $a_t = m$ is selected, it indicates that in the t-th time slot of the communication process, the m-th relay R_m is chosen to assist the base station in transmitting the signal. Each action in the action space corresponds to the selection of a specific relay node.

In an MDP, the reward is used to evaluate the quality of the action taken in the current state. This makes it a crucial component. The reward function is defined as

$$r = 1 - P_{\text{out}} \tag{13}$$

To encourage the agent to select relay nodes that minimize the system outage probability, the reward function is set as shown in Eq. (13). A higher reward value signifies that the correct selection enhances system reliability; conversely, a lower reward indicates that the selected relay node is inadequate, prompting the agent to modify its selection strategy. This approach aims to optimize the long-term performance of the system and reduce the outage probability.

3.2 Proposed DQN-Based Cooperative NOMA Relay Selection Algorithm

In the DQN algorithm, the source node acts as an agent to interact with the communication environment. The agent continuously performs actions to receive feedback from the environment and observe the subsequent state. The Bellman equation is used as an iterative updating method to estimate the action-value function, and these experiences are stored in an experience replay pool for subsequent training. Through multiple iterations, the algorithm converges, thereby enabling the agent to identify the optimal action strategy. As shown in Algorithm 1, DQN is trained for N rounds.

Firstly, a few training parameters are initialized: the discount factor $\mu \in [0, 1]$, the learning rate $\alpha \in (0, 1]$, the greedy factor $\varepsilon \in [0, 1]$, the target DQN weight update frequency T, the batch size for experience tuples L_{b},the replay memory size L_{M}, the weights of the Q-network θ and the weights of the target network $\bar{\theta}$, such that $\bar{\theta} = \theta$.

During the training phase of the algorithm, each time slot t, BS inputs states to the Q network, in order to obtain the Q values of all available actions. It is important to note that the source node does not have prior knowledge of the channel conditions between any two nodes in the communication system. The agent can only perceive the current channel conditions through continuous interactions with the environment. It subsequently employs the deep neural network (DNN) as a nonlinear function approximator to process the observed data. This allows the agent

to estimate the Q-value function of the state space. The dynamic updating of the Q-value function enables the agent to adapt to the changing conditions. This process evaluates the long-term cumulative reward for each relay node in the current state and guides the agent in selecting the relay node with the highest Q-value. A higher Q-value indicates that choosing this relay node will yield better communication quality in the current network conditions, thus facilitating optimal decision-making in the communication system, ensuring the efficiency and stability of relay selection. Based on the recursive Bellman equation, the formula for calculating the Q-value is given by

$$Q_t(s_t, a_t) = E[r_t | s_t = s, a_t = a] \tag{14}$$

where r_t represents the reward generated based on the feedback from the agent's action. According to the iterative Bellman equation, the update formula for the action-value function in each time slot can be expressed as

$$\begin{aligned} Q(s,a) \leftarrow (1-\alpha) Q(s_t, a_t) + \\ \alpha [r_t + \mu \max Q(s_{t+1}, a_{t+1})] \end{aligned} \tag{15}$$

where $\alpha \in (0,1]$ is the learning rate factor. $\mu \in [0,1]$ is the discount factor, indicating the importance of future rewards. $\max Q(s_{t+1}, a_{t+1})$ denotes the maximum Q-value for the next state s_{t+1} after selecting action a_{t+1} at the current state. $r_t + \mu \max Q(s_{t+1}, a_{t+1})$ represents the reward obtained from the current action, which includes both the immediate reward and the potential future reward associated with the action.

The ε-greedy strategy is employed for action selection. This strategy seeks to balance exploration and exploitation, thereby enabling the agent to discover new possibilities and avoid local optima. This strategy facilitates the identification of optimal actions and supports a gradual convergence toward the most effective selection. In the ε-greedy strategy, the agent explores with a probability of ε. This means that it selects actions randomly from the action space. When the probability is $1-\varepsilon$, the agent exploits the ε-greedy strategy by selecting the action that has the highest estimated value from the action space.

The obtained experience tuple $e_t = (s_t, a_t, r_t, s_{t+1})$ is stored in the replay memory $B = \{e_1, e_2, \cdots, e_t\}$, awaiting training as described in Step 7 of Algorithm 1. L_{b} experience tuples are sampled from the replay memory to train the target network and compute the loss function. The loss function measures the discrepancy between the Q-values and the actual returns. By minimizing this discrepancy, the neural network parameters are optimized. This enables accurate assessment of the relative merits of different relay nodes in a dynamically changing communication system. This process tries to make better decisions, thereby improving communication performance and efficiency. Specifically

$$\text{Loss}(\theta) = E_{e_t \sim B}\left\{[y_i - Q(s_t, a_t; \theta)]^2\right\} \tag{16}$$

where

$$y_i = r_t + \mu \max Q\left(s_{t+1}, a_{t+1}; \bar{\theta}\right) \tag{17}$$

where $\bar{\theta}$ denotes the parameters of the target network. This parameter will be replaced by a parameter θ in the Q-network every certain time T, as in step 10 of Algorithm 1.

Algorithm 1 DQN-based cooperative NOMA relay selection algorithm

Initialization:

1: Initialize the training parameters μ,α,ε,T,L_{b} and L_{M}.

2: Initialize the weights of the Q-network θ, the weights of the target network $\bar{\theta}$, such that $\bar{\theta} = \theta$.

Training Phase:

3: **for** $n = 0, 1, \cdots, N$ **do**

4: **for** $t = 0, 1, 2, \cdots, t_{\max}$ **do**

5: Input state s_t into the Q-network and obtain the Q-values for all actions.

6: Select action based on the ε-greedy algorithm, receive reward r_t, and move to next state s_{t+1}.

7: Deposit the experience tuple (s_t, a_t, r_t, s_{t+1}) in the replay memory B.

8: Draw k samples from the replay memory, train the target network, and compute the loss function according to (16).

9: **if** $t\%T = 0$ **then**

10: Update the weights of the target network $\bar{\theta}$.

11: **end if**

12: **end for**

13: **end for**

3.3 Complexity of Proposed Algorithm

The computational complexity of DQN is primarily influenced by the network structure and the experience replay. DQN utilizes a neural network to compute the Q-values for all actions and selects the optimal action. Suppose the neural network in DQN consists of L layers, with an average of d neurons per layer. The complexity of forward propagation is approximately $O(d^2L)$. During training, L_b samples are randomly drawn from the experience replay buffer and used for gradient updates, resulting in a complexity of $O(d^2L_\mathrm{b}L)$. Over the entire training process, assuming a total of N training steps, the overall complexity of the DQN-based relay selection method can be expressed as $O(d^2L_\mathrm{b}LN)$.

As shown in Tab. 1, the complexity of Max-Min and Two-Way traditional methods is $O(M)$. Q-learning has a computational complexity of $O(N|S||A|)$ due to its reliance on Q-table storage. Q-learning is susceptible to dimensional catastrophe in large-scale state spaces. In contrast, the DQN-based relay selection method effectively avoids the need for Q-table storage through deep neural networks and empirical playback, and reduces the outage probability while maintaining high decision quality.

Tab. 1 Complexity of different methods

Methods	Complexity
Max-Min	$O(M)$
Two-Way	$O(M)$
Q-learning	$O(N\|S\|\|A\|)$
DQN	$O(d^2L_\mathrm{b}LN)$

4 Simulation Results

Tab. 2 outlines the configuration of the simulation environment, followed by a detailed analysis of the proposed algorithm through simulation experiments. In the DQN algorithm, the DNN comprises two hidden layers, each containing 64 neurons, with the ReLU function serving as the activation function among these layers. The number of neurons in the input layer corresponds to the dimensionality of the state space. The number of neurons in the output layer corresponds to the dimensions of the action space. The DNN in the DQN algorithm contains a large number of parameters, such as the algorithm's convergence speed. Therefore, optimization of the neural network parameters is performed to determine the optimal network structure.

Tab. 2 Simulation parameters

Parameter	Meaning
μ	The discount factor
α	The learning rate
ε	The greedy factor
T	The target DQN weight update frequency
L_b	The batch size
L_M	The replay memory size
M	Number of relay
α_1, α_2	Transmit power of User 1 and User 2 at the base station
β_1, β_2	Transmit power of User 1, User 2 at relay node
ζ_1, ζ_2	Outage thresholds for User 1, User 2

In Fig. 3, optimizing the learning rate for updating network parameters is critical. The

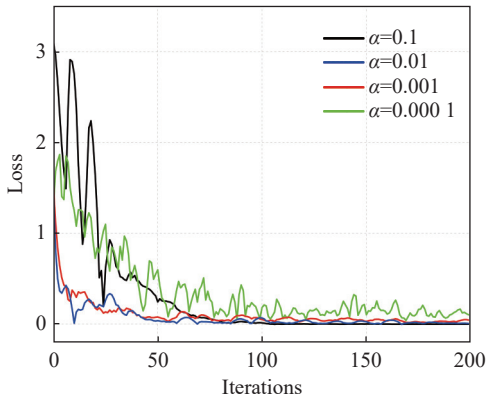

Fig. 3 Loss in different learning rate settings

learning rate affects both the convergence speed of the target network and the stability of the training process in the DQN algorithm. If the learning rate is too large, such as 0.1, the network may bypass certain optimal solutions during training, resulting in prolonged convergence. If the learning rate is too small, such as 0.000 1, the time required for convergence and the discovery of the optimal solution increases significantly. Convergence occurs around the 100th iteration with this rate. Although the network eventually converges, the results indicate that the network may have fallen into a local minimum and fails to find the true optimal solution. When the learning rate is set to 0.01 or 0.001, the network converges around the 50th iteration. However, at a learning rate of 0.01, the network parameters exhibit instability by oscillating in the parameter space, thus affecting the stability of the training process. Considering these observations, a learning rate of 0.001 was selected for the subsequent simulations.

In Fig. 4, the influence of batch size on the performance of the DQN network is analyzed. It has been observed that a smaller batch size reduces the efficiency of data utilization from the experience replay buffer. This leads to slower network convergence and higher memory consumption. Additionally, a small batch size affects the computation of the loss gradient within the network. When a reduced number of data samples are drawn from the experience replay buffer, certain data must inevitably be discarded. If the discarded experiences are critical, this can introduce instability into the training process. On the other hand, a large batch size accelerates network convergence but imposes significant computational demands. Moreover, a large batch size necessitates the availability of substantial data in the experience replay buffer.

The network's hyperparameters are adjusted to their optimal values before proceeding with the subsequent simulations.

In the establishment of the communication model, the number of relays was set to 2. The CSI among nodes follows an exponential distribution with parameters $[10, 1, 10]$. The power allocation factors for both the source and relay nodes are set to $\alpha_1 = \beta_1 = 0.8$, $\alpha_2 = \beta_2 = 0.2$. To evaluate the performance of the proposed method, this study simulates three relay selection schemes based on cooperative NOMA systems: the Max-Min, the Two-Way, and the random relay selection scheme (Random). In Fig. 5, as can be seen, the random relay selection method exhibits a higher outage probability, while the Max-Min and Two-Way relay selection schemes achieve a reduction in the outage probability. The proposed DQN scheme achieves a lower outage probability. At the SNR of 30 dB, the outage probability is lower than that of the other methods. Unlike other methods, the proposed approach makes decisions based on the current state information. This enhances its adaptability to real-

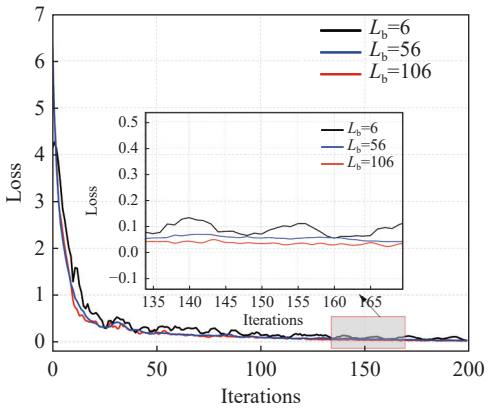

Fig. 4 Loss in different batch size settings

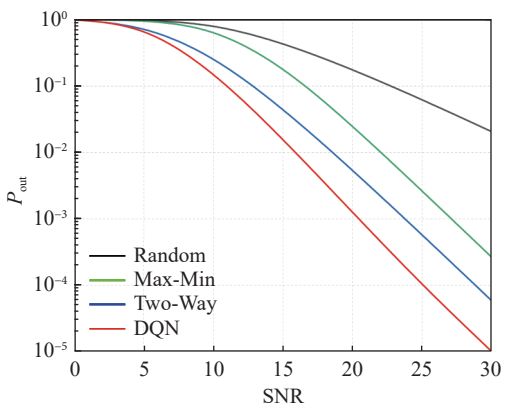

Fig. 5 Outage probability using different methods

world environments. Additionally, it consistently achieves a lower outage probability across different SNR levels, thus ensuring a better communication environment.

To better demonstrate the performance of the proposed method, the efficacy is evaluated with various power allocation factors. In Fig. 6 the power allocation factors of the DQN scheme are compared with those of the Two-Way. In the simulation, the Two-Way and the proposed algorithm are evaluated with power allocation factors denoted as $\alpha_1 = [0.7, 0.8, 0.9]$. The target rates for User 1 and User 2 are set to 0.5 bps/Hz and 1 bps/Hz, respectively. It can be observed that the outage probability is lower when $\alpha_1 = 0.8$ compared to $\alpha_1 = 0.9$. This is due to the excessive reduction of α_2, which decreases the SNR for User 2, thereby increasing the outage probability for User 2. When $\alpha_1 = 0.7$, the outage probability is lower compared to $\alpha_1 = 0.8$. This is due to the fact that User 2, located at the destination node, requires a greater signal-to-noise ratio. Increasing the power allocation factor enhances the signal-to-noise ratio at the receiver, thereby improving the outage probability. The DQN method outperforms the Two-Way with all tested power allocation factors. Notably, when SNR is 25 dB and 30 dB, the proposed method with $\alpha_1 = 0.9$ achieves a lower outage probability than the Two-Way with $\alpha_1 = 0.7$, further demonstrating the feasibility and effectiveness of the proposed approach.

Fig. 7 illustrates the outage probability performance of different relay selection strategies under various outage thresholds. The Random strategy exhibits the highest outage probability, which increases significantly as the outage threshold grows. This indicates that randomly selecting relays is not an effective strategy. This is because it does not take into account the effect of channel conditions, which leads to poor performance. In contrast, the Max-Min strategy achieves a significantly lower outage probability than the Random strategy. But it still increases with the outage threshold. The Max-Min strategy maximizes the minimum channel gain by selecting relays and thus improves performance to some extent. However, this strategy is not globally optimal, so its performance may be limited in complex channel conditions. The Two-Way strategy further reduces the outage probability and outperforms the Max-Min strategy. This suggests that the Two-Way strategy likely considers the channel quality among the relay node, the source, and the destination. Thus, it can significantly improve the system performance in certain scenarios. The DQN strategy demonstrates the lowest outage probability, with a slower growth rate as the outage threshold increases. This indicates that the DQN policy, as an intelligent relay selection method based on DQN, can effectively learn the channel environment and select the optimal relay. Thus, it signif-

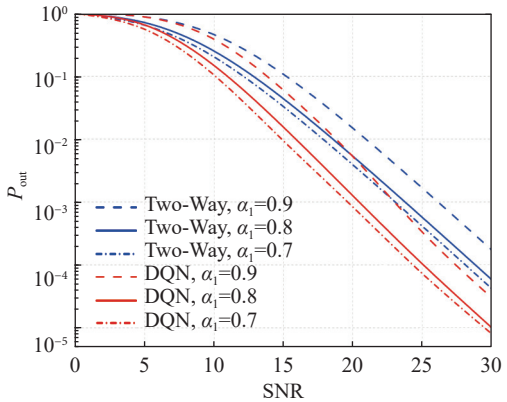

Fig. 6 Outage probability of Two-Way and DQN methods with different power allocation factors

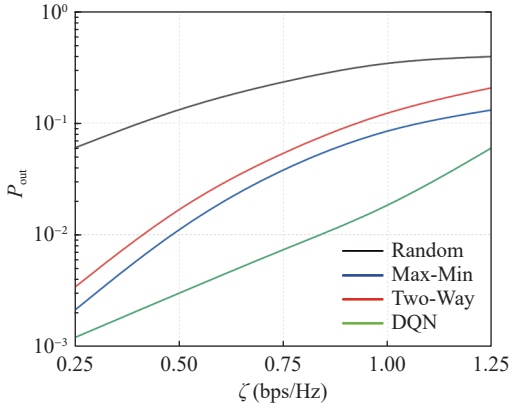

Fig. 7 Outage probabilities of different relay selection strategies under various outage thresholds ($\zeta_1 = \zeta_2 = \zeta$)

icantly reduces the outage probability, especially in the case of higher outage thresholds.

In Fig. 8, to evaluate the performance of the DQN relay selection algorithm compared to the Max-Min relay selection algorithm and the Two-Way relay selection algorithm with different target rates, we established three different user outage threshold cases: 1) $\zeta_1 = 0.5$ bps/Hz, $\zeta_2 = 1$ bps/Hz, 2) $\zeta_1 = 0.75$ bps/Hz, $\zeta_2 = 0.5$ bps/Hz, and 3) $\zeta_1 = 0.5$ bps/Hz, $\zeta_2 = 0.5$ bps/Hz. Firstly, in all three scenarios, the outage probability of the DQN relay selection algorithm is significantly lower than that of the Max-Min and Two-Way relay selection algorithms. Secondly, by comparing scenario 1) and scenario 2), it can be observed that scenario 2) exhibits a lower outage probability after raising the interruption threshold of User 1 and lowering the interruption threshold of User 2. This observation is attributed to the fact that User 1 has superior channel conditions compared to User 2. Additionally, the lower threshold for User 2 facilitates a relatively higher transmission success rate, despite the suboptimal channel conditions. In scenario 3), the target rate is reduced to 0.5 bps/Hz. It can be observed that as the user outage threshold decreases, all three methods achieve a lower outage probability.

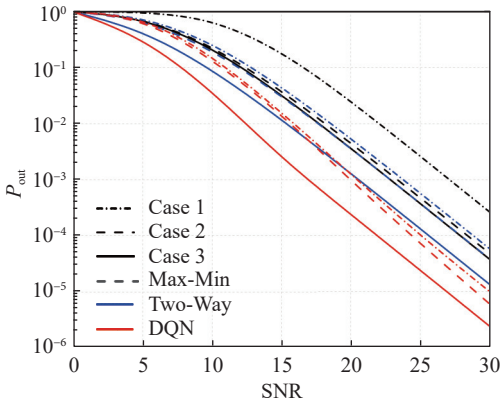

Fig. 8 Outage probability of Max-Min, Two-Way and DQN methods for different outage threshold cases

Examine the variation outage probability as the number of relays increases, with the application of the DQN algorithm with a signal-to-noise ratio of 10 dB. Fig. 9 illustrates that during the initial iterations, the proposed method continuously learns and explores, keeping the communication system's outage probability below 0.2. As the number of iterations increases, the method gradually converges. The system reached convergence at 10 iterations. When the number of relays M=2, the outage probability is below 0.1. When the number of relays M=8, the outage probability is below 0.05. It is evident that as the number of relays increases, the system's outage probability decreases accordingly.

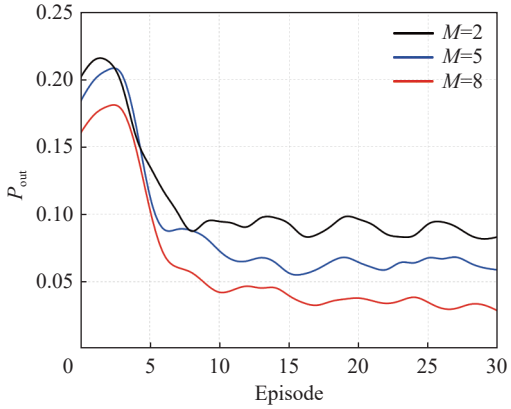

Fig. 9 Outage probability of the DQN method for different numbers of relays

As shown in Fig. 10, the DQN algorithm outperforms the Q-learning algorithm. As SNR increases, the outage probability of both algorithms decreases. The DQN algorithm consistently achieves lower outage probability compared to the Q-learning algorithm across differ-

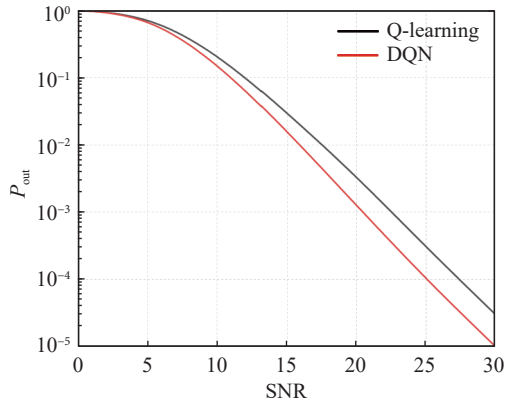

Fig. 10 The outage probability of the DQN and Q-learning methods

ent SNR values. At $\mathrm{SNR} = 30\ \mathrm{dB}$, the DQN algorithm reduces the outage probability by 66%. This improvement is due to its efficient approximation of the Q-value function using deep neural networks. Additionally, the experience replay buffer enables effective experience reuse, enhancing the decision-making capability of the DQN algorithm.

5 Conclusion

A novel relay selection method is introduced in this paper, which utilizes the DQN algorithm to minimize outage probability in cooperative NOMA systems. Unlike traditional methods, which rely on pre-determined CSI, the proposed approach dynamically acquires CSI by interacting with the communication environment. This enables a more adaptable and real-time solution. Through simulation experiments, different relay selection methods were compared. When the SNR is 30 dB, the proposed method reduced the outage probability by 82% compared to the Two-Way method. With a power allocation factor of 0.7, the outage probability decreased to 10^{-5}. Additionally, the outage probability of the proposed method also showed a significant decrease with different outage threshold values. With the deployment of 5th-generation mobile communication technology (5G) and future 6th generation mobile networks (6G) networks, the scenarios for NOMA technology will become more complex and diverse. Therefore, future research could focus on integrating NOMA with advanced deep reinforcement learning to handle dynamic network environments.

References:

[1] Z. Wang, J. Du, Z. Fan, and X. Wan, "Research on resource allocation of multicarrier NOMA systems: A surve, " *Radio Engineering*, vol. 53, no. 9, pp. 2074-2087, 2023.

[2] R. K. Senapati and P. J. Tanna, "Deep learning-based NOMA system for enhancement of 5G networks: A review, " *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 3, pp. 3380-3394, 2024.

[3] T. N. Do, D. B. da Costa, T. Q. Duong, and B. An, "Improving the performance of cell-edge users in NOMA systems using cooperative relaying, " *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 1883-1901, 2018.

[4] C. Gao, B. Yang, D. Zheng, X. Jiang, and T. Taleb, "Cooperative jamming and relay selection for covert communications in wireless relay systems, " *IEEE Transactions on Communications*, vol. 72, no. 2, pp. 1020-1032, 2024.

[5] P. Archana, V. P. Harigovindan, G. G. Faustina, and A. V. Babu, "Incremental relaying with partial relay selection for enhancing the performance of underlay cognitive NOMA networks, " *IEEE Latin America Transactions*, vol. 22, no. 7, pp. 548-556, 2024.

[6] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative NOMA, " *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 416-419, 2016.

[7] S. Lee, D. B. da Costa, Q. T. Vien, T. Q. Duong, and R. T. de Sousa, "Non-orthogonal multiple access schemes with partial relay selection, " *IET Communications*, vol. 11, no. 6, pp. 846-854, 2017.

[8] K. S. G. Kiran and R. Swaminathan, "Performance analysis of DF-relaying-based cooperative noma system with partial relay selection, " in *2022 14th International Conference on Communication Systems & Networks (COMSNETS)*, Bangalore, India, pp. 574-580, 2022.

[9] C. Song, D. Liu, E. Li, R. Zhao, S. Hao, and T. Shen, "Reactive relay selection strategy for NOMA systems, " *Journal of Xidian University*, vol. 48, no. 3, pp. 5662, 2021.

[10] Y. Lu, G. Wang, C. Yang, L. Yang, K. Zhao, and H. Zhu, "NOMA relay selection and cooperative transmission based on single source optimal path, " *Systems Engineering and Electronics*, vol. 44, no. 1, pp. 292-298, 2022.

[11] Z. Wang and Z. Peng, "Secrecy performance analysis of relay selection in cooperative NOMA systems, " *IEEE Access*, vol. 7, pp. 86274-86287, 2019.

[12] T. M. Hoang, B. C. Nguyen, H. M. Nguyen, and P. T. Tran, "Secrecy outage performance of NOMA relay networks using partial relay selection in the presence of multiple colluding eavesdroppers, " *Computer Communications*, vol. 216, pp. 238-250, 2024.

[13] D. Yang, C. Li, B. Ren, H. Li, and K. Guo, "Analysis of relay selection schemes in underlay cogni-

tive radio non-orthogonal multiple access networks, " *International Journal of Distributed Sensor Networks*, vol. 18, no. 4, pp. 1550, 2022.

[14] M. Noori Shirazi and M. Zahabi, "Sum-rate and energy efficiency optimization by novel relay selection in a NOMA-based cooperative network in the presence of interference, " *Wireless Personal Communications*, vol. 134, no. 1, pp. 225-248, 2024.

[15] M. Fayaz, W. Yi, Y. Liu, S. Thayaparan, and A. Nallanathan, "Toward autonomous power control in semi-grant-free NOMA systems: A power pool-based approach, " *IEEE Transactions on Communications*, vol. 72, no. 6, pp. 3273-3289, 2024.

[16] B. Yin, X. Fang, and X. Wang, "Joint optimization of trajectory control, resource allocation, and user association based on DRL for multi-fixed-wing uav networks, " *IEEE Transactions on Wireless Communications*, vol. 23, no. 10, pp. 13330-13343, 2024.

[17] X. Wang and X. Wang, "Reinforcement learning-based multihop relaying: A decentralized Q-learning approach, " *Entropy*, vol. 23, no. 10, pp. 1310, 2021.

[18] Y. Zhang, Y. Su, X. Shen, A. Wang, B. Wang, Y. Liu, and W. Bai, "Reinforcement learning based relay selection for underwater acoustic cooperative networks, " *Remote Sensing*, vol. 14, no. 6, pp. 1-27, 2022.

[19] Y. Su, X. Lu, Y. Zhao, L. Huang, and X. Du, "Cooperative communications with relay selection based on deep reinforcement learning in wireless sensor networks, " *IEEE Sensors Journal*, vol. 19, no. 20, pp. 9561- 9569, 2019.

[20] E. M. Ghourab, L. Bariah, S. Muhai dat, P. C. Sofotasios, M. Al-Qutayri, and E. Damiani, "Secure relay selection with outdated CSI in cooperative wireless vehicular networks: A DQN approach, " *IEEE Access*, vol. 12, pp. 12424-12436, 2024.

[21] M. Alkhawatrah, "The performance of supervised machine learning based relay selection in cooperative NOMA, " *IEEE Access*, vol. 11, pp. 1570-1577, 2023.

Ying Lin received the communications engineering Ph.D. degree in Lanzhou University of Technology in 2022. She is an Associate Professor and Master's Supervisor at School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China. Her research interests include multiple access technology in next generation wireless communication system, signal processing theory in wireless communication system.

Yongwei Xiong received the B.S. degree in communications engineering from Jiangxi University of Science and Technology in 2022. She is currently working towards the master degree in School of Computer and Communication, Lanzhou University of Technology. Her current research is wireless communication theory and technology.

Xingbo Gong received the communications engineering B.S. degree in Lanzhou University of Technology in 2018. He is currently working towards the master degree in Lanzhou University of Technology.

Sifei Zhang received his B.S. degree in communication engineering from Lanzhou University of Technology in 2018 and is currently pursuing a master's degree of Lanzhou University of Technology.

Yinhang Tian received the communications engineering B.S. degree in Tianjin University of Science and Technology in 2023. He is currently working towards the master degree in Lanzhou University of Technology.