IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 July 2024, accepted 24 July 2024, date of publication 29 July 2024, date of current version 7 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435145

RESEARCH ARTICLE

Power Allocation for Secure NOMA Network Based on Q-Learning

WEIDONG GUO AND XINYAN WANG
School of Cyber Science and Engineering, Qufu Normal University, Qufu, Shandong 273165, China
Corresponding author: Weidong Guo (gud2001@qfnu.edu.cn)

This work was supported by the Horizontal Project under Grant kj2023hx.

ABSTRACT This paper investigates the secrecy performance of a multi-relay non-orthogonal multiple access (NOMA) network. Considering the presence of eavesdroppers, an optimal relay selection scheme and a jamming signal transmission scheme for both amplify-and-forward (AF) and decode-and-forward (DF) strategies are proposed. However, the resource allocation problem aimed at maximizing the effective secure throughput (EST) of the system is non-convex. It is difficult to directly solve this optimization problem using conventional methods. As such, a Q-learning approach to solve the resource allocation problem in this system is applied, and an innovative reward function that can maximize the communication quality of edge users while ensuring secure communication of nearby users is designed. According to the analysis of the simulation results, the convergence of the proposed scheme is verified. Under the same conditions, DF relays achieve a higher signal-to-noise ratio (SINR) at the user terminal, and the EST is closely related to the transmission power of the source node and the relays. The numerical results also show that compared to conventional power allocation methods, the proposed method achieves a larger average EST and provides better confidentiality performance.

INDEX TERMS NOMA, relay selection, power allocation, physical layer security, Q-learning, effective secrecy throughput.

I. INTRODUCTION

With the rapid development of fifth generation (5G) mobile communication technology, the scale of wireless communication networks has greatly increased. As the number of communication devices continues to grow, there is a growing demand from users for high throughput and high access density. Non-orthogonal multiple access (NOMA) technology allows multiple users to share the same orthogonal resources, effectively improving user access and spectrum efficiency. This has become a key technology for 5G mobile communication systems [1], [2]. NOMA employs superposition coding (SC) and successive interference cancellation (SIC). The transmitter overlays information for multiple users via the SC, while the receiver utilizes SIC to decode the signals for each user, thereby meeting the requirements of massive access [3]. Although NOMA can enhance the spectral efficiency, the distance of the signal transmission is limited. Integrating NOMA with cooperative communication offers an excellent solution for overcoming the coverage limitation of NOMA [4]. The introduction of relay cooperation technology not only enhances the coverage capability of the communication network, but also effectively addresses issues such as vulnerability to eavesdropping during the communication process. Currently, a wealth of literature combines these two techniques and conducts research [5], [6], [7]. However, with a notable surge of users that access in NOMA cooperative network, it has also introduced some security issues that need to be addressed.

Considering the inherent broadcast nature of wireless communication, it is challenging to achieve the secure transmission of confidential information in the presence of eavesdroppers. Therefore, physical layer security (PLS) techniques in wireless communication systems are rapidly developing and receiving widespread attention [8], [9]. Eavesdroppers typically conduct both active and passive

The associate editor coordinating the review of this manuscript and approving it for publication was Walid Al-Hussaibi.

types of attack, and the two types of attack have different objectives. Active attack is aimed at interfering with the reception of communications, while passive attack is aimed at eavesdropping on information [10]. Most papers focus on passive attack, which will be explained extensively in subsequent references. If an eavesdropper takes active measures, it may expose channel information. Most studies focus on improving the secrecy performance of systems by optimizing power allocation or relay selection strategies. Nodes can operate more efficiently with limited resources by optimizing power allocation, thereby maximizing the duration of their normal operation. The authors in [11] considered strong users in the NOMA network as potential eavesdroppers and investigated a scheme to achieve optimal power distribution. Therefore, the security rate of the system is maximized under the user data rate and total power constraints. In addition, relay selection is also an effective method to enhance communication security. Based on NOMA networks containing multiple eavesdroppers, [12] proposed a secure relay selection scheme and measured the system performance with security outage probability (SOP). But most of the previous studies relied on mathematical methods to solve the problem, resulting in high computational complexity. To overcome this problem, improving the computational efficiency has become an important goal.

Over the past few years, a considerable number of researchers have employed machine learning (ML) in wireless communication systems to enhance the efficiency of the system in dynamic scenarios [13], [14], [15], [16]. Reference [17] proposed a physical layer authentication method to improve the authentication efficiency by training neural networks. Also, using machine learning for resource allocation can improve the performance of the communication system. Reference [18] adopted the NOMA scheme in the mobile edge computing (MEC) system. To minimize the delay of the system, the offloading strategy is optimized and reinforcement learning (RL) is used to solve the power allocation problem. However, the communication model of [17] only considered the case of a single relay, while [18] did not consider relay collaboration. Among these ML techniques, Q-learning in RL has gained widespread application owing to its model-free and distributed nature [19]. It is suitable for massive machine type communication (mMTC) devices [20], [21].

It is worth noting that previous studies mainly addressed PLS problems by optimizing power allocation or relay selection individually, and few have investigated the joint optimization of relay selection and power allocation using Q-learning. Therefore, for cooperative NOMA systems with passive eavesdroppers, this study formulates optimized strategies for relay selection and power allocation. To further reduce computational complexity, a Q-learning algorithm is also introduced to improve the performance of the system. Both amplify-and-forward (AF) and decode-and-forward (DF) relays are considered to ensure the integrity of the analysis. Since Q-learning stores the Q-values of all state-action pairs through a Q-table, it is difficult for the Q-table to cover the entire state space when dealing with multi-dimensional states, so the convergence speed is often improved by the deep Q-network (DQN) algorithm [22]. This paper compares two RL algorithms and shows that the Q-learning algorithm is more efficient when the state space is small. The contributions of this study are summarized as follows.

- Based on the preceding context, the integration of NOMA with cooperative communication yields significant improvements in frequency efficiency and coverage capacity. Moreover, the security of these networks is important. In this study, a relay cooperative network model with eavesdroppers, multiple relays and two users is proposed. One relay is selected to forward the message whereas another relay sends a jamming signal to the eavesdropper, thus confusing the eavesdropper and reducing the decoding ability of the eavesdropper, further improving the communication and confidentiality performance of the system. Furthermore, all the relays use fixed gains, eliminating the need for real-time gain adjustments and effectively reducing system complexity, which is more in line with practical communication scenarios.
- To reduce the computational complexity and improve the system efficiency, this study employs Q-learning to address the power allocation issue. In addition, a novel reward function that enhances the sum rate while satisfying power conditions is designed. This reward function can maximize the communication quality of edge users while ensuring secure communication of nearby users which is one of innovations of this study.
- This study conducts a comprehensive comparative analysis of the use of AF relays and DF relays in terms of effective secure throughput (EST) and runtime. The effects of different parameters on EST of the two protocols are examined. These parameters include the number of relays, transmit power of the source and relay nodes, and the interference power. Furthermore, the proposed method is compared with the deep Q-learning based power allocation method and common used methods, demonstrating its significant advantages in terms of confidentiality.

The remainder of this paper is organized as follows. Section II introduces the research background and related work. In Section III, a relay-assisted NOMA system model is proposed and the relay selection and power allocation problems are described. Section IV presents a Q-learning based power allocation scheme. In Section V, a comprehensive comparison of the system performance obtained using the DF and AF protocols is presented. Finally, Section VI concludes the paper.

II. RELATED WORK

For improving the security of the communication process, many studies have investigated optimization problems such as power allocation and relay selection. Reference [23] investigated the secrecy performance of reconfigurable intelligent surface (IRS)-assisted millimeter wave system by optimizing parameters such as power and phase shifts of the RIS elements. For single-input single-output (SISO) NOMA system in the presence of passive eavesdroppers, [24] investigated maximizing the rate of confidentiality of the system while meeting the quality of QoS requirements of the users. However, the models proposed in the above papers do not consider the introduction of relays for collaboration, which is an effective technique to improve the secrecy performance of the system. For an untrusted NOMA system with DF relay cooperation, [25] proposed an optimized power allocation scheme to maximize the secrecy rate of the nearby user while satisfying the quality of service (QoS) requirement of the distant user. However, that study only investigated the use of DF relay. Therefore, [26] considered an AF relay cooperation NOMA system with untrusted users, and obtained a closed-form expression for optimal power distribution between the source and relay nodes.

To further enhance the confidentiality of the system, [27] introduced a friendly jammer. Attacks during communication are reduced by sending jamming signals to the eavesdropper, but relay collaboration techniques are not considered. The optimal power allocation for the AF relay cooperative NOMA network was obtained by [28] using a cooperative jamming (CJ) scheme. CJ achieves a higher secrecy rate by jamming the eavesdropping link. However, it should be noted that this study considered only the case of single relay cooperative communication. When multiple relays are present in the network, in addition to optimizing the power allocation, relay selection strategies can be used to enhance the confidentiality of the system. Reference [29] proposed and compared three relay selection schemes, showing that the proposed optimal relay selection scheme achieved a higher SOP. For relay cooperative networks containing passive eavesdroppers, [30] proposed a new relay selection scheme that combined Wyner coding with linear network coding to improve the security of the communication.

To improve the computational efficiency, many studies have introduced RL algorithms to improve the communication performance in various scenarios. For satellite relay networks, [31] proposed a Q-learning based NOMA random access protocol to optimize access slots and achieve an optimal throughput. Reference [32] focused on MTC and investigated a Q-learning and NOMA based approach to dynamically allocate random access slots to MTC devices. Reference [33] investigated the selection of underwater relays through DQN in a collaborative Internet of Underwater Things (IoUT) system, which effectively improves the transmission performance of the system. In a non-orthogonal amplify-and-forward (NAF) cooperative NOMA network, [34] presented a Q-learning based power allocation algorithm to maximize the system throughput. They further introduced a neighboring strategy searching based power allocation algorithm, striking a balance between throughput and computational complexity.

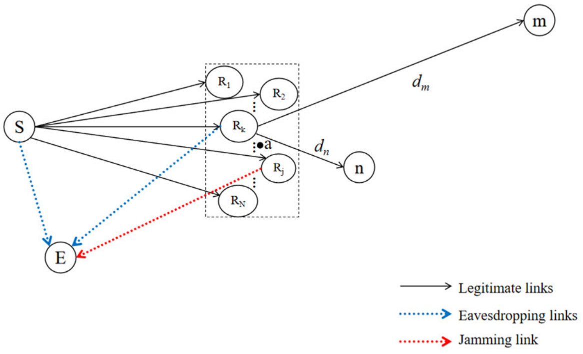


FIGURE 1. Relay cooperative system model based on NOMA networks.

In recent years, RL algorithms have also been widely applied to address the PLS problems. For multi-user networks, [35] considered a model of passive eavesdropping in which the receiving nodes act as both jammers and receivers to disrupt eavesdropping. They proposed a Q-learning based jamming user selection scheme that significantly enhanced the confidentiality performance of the network. Furthermore, [36] investigated the performance of secure communication between unmanned aerial vehicles (UAVs) and ground users in the presence of passive eavesdroppers. The study utilized the Q-learning algorithm to maximize the average secrecy rate (ASR). However, these studies did not consider the effects of the relays on system confidentiality. In [37], a deep learning algorithm (DL) was used to optimize the power management to improve the secrecy rate of AF relay collaborative network. This study improves EST performance by optimizing power allocation and relay selection which achieves a trade-off between throughput and security. However, only one relay is considered in [37] and the relay uses the AF forwarding protocol, while our proposed method considers two forwarding protocols and two reinforcement learning methods with multi-relay simultaneously. Therefore, our approach is more comprehensive. In [38], the use of Q-learning for relay selection in a DF relay cooperation network was studied. Compared with other common relay selection methods, Q-learning was shown to achieve a higher EST. In collaborative communication, [39] used the DQN algorithm to optimize the relay selection scheme. However, they did not consider the optimization of power allocation.

III. SYSTEM MODEL

As shown in Figure 1, a relay cooperative network based on downlink NOMA is considered. The network comprises a source node (S), N relays $R_i(i \in \{1, 2, \cdots, N\})$, two legitimate users (m, n), and an eavesdropper (E). All the relays are randomly distributed within a certain range between the source node and destination nodes and all operate in the half-duplex mode (HD). The distribution of these relays has a certain geometric range both horizontally and vertically, forming a rectangular area. Point a in the figure is the centre of this rectangular area. The distance between a user and the relays refers to the distance from the user to the center of this rectangular region. Denote the distances of the two users to the relay set as d_m, d_n respectively, where $d_m \gg d_n$. The user m is located far from the relay set, referred to as the weak user, whereas the user n is closer to the relay set, known as the strong user, and the two users are not on the same line. The distance of the two users from the set of relay nodes is much larger than the vertical or horizontal range of the relay set. Assume that strong user n carries out communication in small packets of a few bits, while weak user m carries out communication in large packets of cache class, such as movies or short videos. User m is at the edge of the coverage and the propagation experiences occlusion such as buildings or woods. Since all the relays are close to user n, no matter how the forwarding relay is selected it has little effect on user n, and that user gets a better user experience. The relay selection method proposed in this paper allows the edge user to get the best user experience and does its best to achieve the consistency of the system user experience which will be presented in the following.

The channels between all nodes are assumed to be independent Rayleigh fading channels, and each node is equipped with a single antenna. Additionally, the noise of all channels is modelled as additive white Gaussian noise (AWGN) with zero mean and variance N_0. Assuming that there is no direct connection between S and users, E can eavesdrop on the information transmitted by S and the relays. To reduce the decoding capability of E, one relay transmits interference signals to the eavesdropper.

This study adopts the well-known Wyner eavesdropping code to ensure reliable transmission, encoding x_m and x_n as $(\tau_{m,t}, \tau_{m,s})$ and $(\tau_{n,t}, \tau_{n,s})$. $\tau_{m,t}$ and $\tau_{n,t}$ denote the codeword rates of m and n, respectively, $\tau_{m,s}$ and $\tau_{n,s}$ denote the secrecy rates of m and n, respectively. When the channel capacity of the user channel is greater than the codeword rate, the user can successfully decode confidential information, which ensures the reliability of the information. Similarly, when the channel capacity of the eavesdropping channel is less than the secrecy rate, E cannot correctly decode confidential information, thereby ensuring information security. In the following, the communication processes of the DF relay system and the AF relay system will be described separately.

A. DF PROTOCOL

Each communication time slot consisted of two phases. In the first phase, S adopts SC technique to broadcast the signal

$$x_s = \sqrt{a_m P_S} x_m + \sqrt{a_n P_S} x_n, \tag{1}$$

with power P_S, where x_m and x_n represent the signals sent to m and n. And $E\left[x_m^2\right] = 1$, $E\left[x_n^2\right] = 1$, where $E[\bullet]$ is the expectation of the random variable. a_m, a_n represent the power distribution factors of m and n, respectively, satisfying $a_m + a_n = 1$ and $a_m > a_n$. The received signals of $R_i (1 \leq i \leq N)$ and E are denoted as

$$y_{SR_\mathrm{i}} = h_{SR_i} x_S + n_{SR_i}, \tag{2}$$

$$y_{SE} = h_{SE} x_S + n_{SE}, \tag{3}$$

where h_{SR_i} is the channel coefficient from S to R_i, h_{SE} is the channel coefficient from S to E. n_{SR_i} and n_{SE} are AWGNs.

Every relay attempts to decode the signal sent by S. According to the decoding principle of NOMA technology, all relays employ the SIC technique, which first decodes the signal x_m, regards the signal x_n as interference, and eliminates the interference of the signal x_m after successful decoding. Thus, the signal-to-noise ratios (SINRs) of the decoded signals x_m and x_n at relay R_i are expressed as

$$\gamma_{SR_i \to m} = \frac{a_m |h_{SR_i}|^2 P_S}{a_n |h_{SR_i}|^2 P_S + N_0}, \tag{4}$$

$$\gamma_{SR_i \to n} = \frac{a_n |h_{SR_i}|^2 P_S}{N_0}, \tag{5}$$

where $|h_{SR_i}|^2$ represents the channel gain from S to R_i.

The relay selection strategy is designed to achieve two purposes simultaneously. One is to ensure that the codeword rate of user n is achieved and the other is to maximize the communication rate of user m. Specifically, the relay selection strategy is divided into the two stages which can be represented as follows. In the first stage, the relays that can successfully decode both x_m and x_n are placed in decoding set U which can be expressed as

$$\begin{aligned} \mathrm{U} = \{i : 1 \leq i \leq N, \frac{1}{2}\log(1 + \gamma_{SR_i \to m}) \geq \tau_{m,t}, \\ \frac{1}{2}\log(1 + \gamma_{SR_i \to n}) \geq \tau_{n,t}\}. \end{aligned} \tag{6}$$

Then in the second stage the eligible relays are selected from the set U to form a new relay set Ω:

$$\begin{aligned} \Omega = \{\frac{1}{2}\log(1 + \gamma_{R_i n \to m}) \geq \tau_{m,t}, \\ \frac{1}{2}\log(1 + \gamma_{R_i n \to n}) \geq \tau_{n,t}, i \in \mathrm{U}\}, \end{aligned} \tag{7}$$

where $\gamma_{R_i n \to m}$, $\gamma_{R_i n \to n}$ denote the SINRs when user n decodes user m's signal and when user n decodes its own signal, respectively.

Then, in set Ω, the relay that enables user m to have the largest signal-to-noise ratio (SINR) when decoding its own signal is selected, which can be expressed as

$$k = \arg\max_{\substack{k \in \mathrm{U} \\ k \in \Omega}} \left\{\gamma_{R_k m \to m, DF}\right\}, \tag{8}$$

where $\gamma_{R_k m \to m, DF}$ denotes the SINR when user n decodes its own signal. If U is an empty set, it means that no relay is able to successfully decode the received signals, and thus the optimal relay cannot be selected and communication will not proceed in this time slot. On the contrary, if U is not an empty set, the next step is to select the optimal relay $R_k (1 \leq \mathrm{k} \leq N)$ as the forwarding relay from U, based on the relay selection rules set above. The proposed relay selection strategy ensures that the strong user n can communicate correctly, while maximizing the SINR for decoding by the weak user m. In addition, this method only requires comparing the relays in the decoding set U, instead of comparing all the relays in the system which effectively reduces the computational complexity. After relay selection is completed, the selected relay R_k re-encodes the received information and forwards it to m and n with power P_R.

Considering the worst eavesdropping scenario, the eavesdropper possesses excellent demodulation capabilities and can successfully employ SIC in each demodulation process. In other words, during the demodulation of x_n, E is always able to eliminate the interference of x_m. Assuming that the jamming relay only transmits interference signals in the second stage, the SINRs for eavesdropping on m and n using the SIC technique in the first stage are represented as

$$\gamma_{SE\to m} = \frac{a_1|h_{SE}|^2 P_S}{a_2|h_{SE}|^2 P_S + N_0}, \tag{9}$$

$$\gamma_{SE\to n} = \frac{a_2|h_{SE}|^2 P_S}{N_0}, \tag{10}$$

where $|h_{SE}|^2$ represents the channel gain from S to E.

Due to the concealment of the eavesdropper, the position and channel information of the eavesdropper are unknown, it is impossible to accurately select an interference relay based on conditions such as the SINR. Therefore, another relay $R_j (1 \leq j \leq N, j \neq k)$ is randomly selected as the jamming relay and the jamming signal is sent with power P_J to E in the second stage, reducing the SINR for decoding at E and thereby lowering the risk of information being eavesdropped by E. We assume that m and n possess knowledge of the transmitted interference signals and have the capability to eliminate their impacts. The signals received at m and n in the second stage are

$$y_{m,DF} = h_{R_k m}(\sqrt{a_m P_{R_k}} x_m + \sqrt{a_n P_{R_k}} x_n) + n_{R_k m}, \tag{11}$$

$$y_{n,DF} = h_{R_k n}(\sqrt{a_m P_{R_k}} x_m + \sqrt{a_n P_{R_k}} x_n) + n_{R_k n}, \tag{12}$$

where $h_{R_k m}$ and $h_{R_k n}$ are the channel gains from R_k to m and n, respectively. $n_{R_k m}$ and $n_{R_k n}$ are AWGNs.

Similar to the first stage, when the weak user m performs decoding, it considers the signal of n as an interference. The strong user n decodes the signal of m before decoding its own signal. The SINRs when m and n demodulate the signal of m and n demodulates its own signal are expressed as

$$\gamma_{R_k m\to m,DF} = \frac{a_1|h_{R_k m}|^2 P_{R_k}}{a_2|h_{R_k m}|^2 P_{R_k} + N_0}, \tag{13}$$

$$\gamma_{R_k n\to m,DF} = \frac{a_1|h_{R_k n}|^2 P_{R_k}}{a_2|h_{R_k n}|^2 P_{R_k} + N_0}, \tag{14}$$

$$\gamma_{R_k n\to n,DF} = \frac{a_2|h_{R_k n}|^2 P_{R_k}}{N_0}, \tag{15}$$

where $|h_{R_k m}|^2$ and $|h_{R_k n}|^2$ represent the channel gains from R_k to m and n respectively. The observation at E at this stage can be expressed as

$$y_{RE} = h_{R_k E}(\sqrt{a_m P_R} x_m + \sqrt{a_n P_R} x_n) + h_{R_j E}\sqrt{P_J} x_0 + n_{RE}, \tag{16}$$

where $h_{R_k E}$ and $h_{R_j E}$ are the channel gains from R_k and R_j to E, respectively. x_0 represents the interference signal transmitted by R_j, and $E\left[x_0^2\right] = 1$. n_{RE} is the AWGN.

Based on the above analysis, when E is eavesdropping, it first decodes the signal of m and then decodes the signal of n. Therefore, the SINRs during the decoding stage of E in the second phase can be written as

$$\gamma_{RE\to m,DF} = \frac{a_1|h_{R_k E}|^2 P_{R_k}}{a_2|h_{R_k E}|^2 P_{R_k} + P_{R_j}|h_{R_j E}|^2 + N_0}, \tag{17}$$

$$\gamma_{RE\to n,DF} = \frac{a_2|h_{R_k E}|^2 P_{R_k}}{P_{R_j}|h_{R_j E}|^2 + N_0}, \tag{18}$$

where $|h_{R_k E}|^2$ and $|h_{R_j E}|^2$ represent the channel gains from R_k to E and R_j to E, respectively.

In the existing literature, the security performance of a system is often measured using the secrecy capacity, defined as $C_S = [C_B - C_E]^+$, where $[x]^+ = \max\{0, x\}$, C_B represents the communication capacity of the main channel, and C_E represents the channel capacity of the eavesdropper channel. Because it is difficult for the network to obtain the channel state information (CSI) of the eavesdropper, it may not obtain C_E. Therefore, instead of using the common secrecy capacity to measure system performance, EST is used as a measure of the system performance in this study to achieve a trade-off between the communication performance and secrecy performance. EST represents the average rate at which confidential information is transmitted from the transmitter to the receiver without being intercepted. The EST of m and n can be expressed as

$$\begin{aligned} EST\_m_{DF} = {} & \tau_{m,s} \Pr(\gamma_{SR_i\to m} > 2^{2\tau_{m,t}} - 1, \\ & \gamma_{R_k m\to m,DF} > 2^{2\tau_{m,t}} - 1, \gamma_{SE\to m} < 2^{2(\tau_{m,t}-\tau_{m,s})} \\ & - 1, \gamma_{RE\to m,DF} < 2^{2(\tau_{m,t}-\tau_{m,s})} - 1), \end{aligned} \tag{19}$$

$$\begin{aligned} EST\_n_{DF} = {} & \tau_{n,s} \Pr(\gamma_{SR_i\to m} > 2^{2\tau_{m,t}} - 1, \\ & \gamma_{SR_i\to n} > 2^{2\tau_{n,t}} - 1, \gamma_{R_k n\to m,DF} > 2^{2\tau_{m,t}} - 1, \\ & \gamma_{R_k n\to n,DF} > 2^{2\tau_{n,t}} - 1, \gamma_{SE\to n} < 2^{2(\tau_{n,t}-\tau_{n,s})} - 1, \\ & \gamma_{RE\to n,DF} < 2^{2(\tau_{n,t}-\tau_{n,s})} - 1). \end{aligned} \tag{20}$$

where Pr(•) represents a probability operation.

B. AF PROTOCOL

For the AF protocol, the communication process is similar to that described above. In the first stage, similar to the DF protocol, S broadcasts the signal to all relays and E attempts to decode the message. The signals received by R_k and E are given by (2) and (3) respectively.

In the second stage, traverse all relays and similarly place the relays that can ensure that user n correctly decode its own information in the set Ω. Then R_k in Ω that enables user m to decode its own signal with the largest SINR is selected, which can be expressed as

$$k = \underset{\substack{k\in\{1,2,\cdots K\}\\ k\in\Omega}}{\arg\max} \left\{\gamma_{R_k m\to m,AF}\right\}. \tag{21}$$

The signals received at m and n can be written as

$$y_{m,AF} = G h_{R_k m} y_{SR_k} + n_{R_k m}, \tag{22}$$

$$y_{n,AF} = G h_{R_k n} y_{SR_k} + n_{R_k n}, \tag{23}$$

where $G = \sqrt{\frac{P_{R_k}}{P_S E(|h_{SR_i}|^2)+N_0}}$ is the AF relay amplification factor. Owing to the unknown CSI, the gain cannot be adjusted in real time, and a variable-gain relay cannot be used. Therefore, the system uses fixed-gain relaying, which effectively reduces the complexity of the system. The SINR when m decodes the signal x_m can be expressed as

$$\gamma_{R_k m\to m,AF} = \frac{a_1 P_S |h_{SR_k}|^2 |h_{R_k m}|^2 G^2}{a_2 P_S |h_{SR_k}|^2 |h_{R_k m}|^2 G^2 + N_0 |h_{R_k m}|^2 G^2 + N_0}. \tag{24}$$

Subsequently, n first decodes the signal x_m, and then decodes its own signal. The two SINRs are as follows,

$$\gamma_{R_k n\to m,AF} = \frac{a_1 P_S |h_{SR_k}|^2 |h_{R_k n}|^2 G^2}{a_2 P_S |h_{SR_k}|^2 |h_{R_k n}|^2 G^2 + N_0 |h_{R_k n}|^2 G^2 + N_0}, \tag{25}$$

$$\gamma_{R_k n\to n,AF} = \frac{a_2 P_S |h_{SR_k}|^2 |h_{R_k n}|^2 G^2}{N_0 |h_{R_k n}|^2 G^2 + N_0}. \tag{26}$$

The SINRs decoded by the eavesdropper at this stage are expressed as

$$\gamma_{RE\to m,AF} = \frac{a_1 P_S |h_{SR_k}|^2 |h_{R_k E}|^2 G^2}{a_2 P_S |h_{SR_k}|^2 |h_{R_k E}|^2 G^2 + N_0 |h_{R_k E}|^2 G^2 + P_J |h_{R_j E}|^2 + N_0}, \tag{27}$$

$$\gamma_{RE\to n,AF} = \frac{a_2 P_S |h_{SR_k}|^2 |h_{R_k E}|^2 G^2}{N_0 |h_{R_k E}|^2 G^2 + P_J |h_{R_j E}|^2 + N_0}. \tag{28}$$

When using the AF protocol, the EST for m and n can be expressed as

$$\begin{aligned} EST\_m_{AF} = \tau_{m,s} \Pr(&\gamma_{R_k m\to m,AF} > 2^{2\tau_{m,t}} - 1, \\ &\gamma_{SE\to m} < 2^{2(\tau_{m,t}-\tau_{m,s})} - 1, \\ &\gamma_{RE\to m,AF} < 2^{2(\tau_{m,t}-\tau_{m,s})} - 1), \end{aligned} \tag{29}$$

$$\begin{aligned} EST\_n_{AF} = \tau_{n,s} \Pr(&\gamma_{R_k n\to m,AF} > 2^{2\tau_{m,t}} - 1, \\ &\gamma_{R_k n\to n,AF} > 2^{2\tau_{n,t}} - 1, \gamma_{SE\to n} < 2^{2(\tau_{n,t}-\tau_{n,s})} - 1, \\ &\gamma_{RE\to n,AF} < 2^{2(\tau_{n,t}-\tau_{n,s})} - 1). \end{aligned} \tag{30}$$

C. PROBLEM FORMULATION

To enhance the security of the relay cooperation network further, the optimal power allocation problem for S and R_k can be formulated as follows:

$$\begin{aligned} &\max\ EST\_SUM \\ &s.t.\ P_{\max} > P_S > P_{\min} \\ &\qquad P_{\max} > P_R > P_{\min}, \end{aligned} \tag{31}$$

where EST_SUM represents the sum of the EST values of m and n. $P_{\max}$ and $P_{\min}$ denote the maximum and minimum values of power, respectively. In the formulated problem, the parameters that need to be optimized are the power of the source node and the relay node, which is explained in (31). The purpose of optimizing the power is to extend the survival time of the source node and the relay node. In practical communication scenarios, the transmitting node may be powered by solar or other energy harvesting methods and the stored energy is very limited, so the transmission power of these nodes needs to be optimized. This formula aims to allocate the optimal transmit power to both nodes within a given power range to maximize the sum of the EST of users, which optimizes the average rate of message transmission in the system under secrecy conditions. Considering (19), (20), (29) and (30), it can be concluded that the formulated optimization problem is non-convex. Subsequently, we will propose a Q-learning based approach to solve this problem.

IV. PROPOSED POWER ALLOCATION ALGORITHMS

In this study, the optimal power allocation scheme of the source node and the forwarding relay obtained through Q-learning is investigated. The performance of the system is measured using EST_SUM, considering both the communication rate and security performance of the system.

As reinforcement learning in machine learning is more suitable for solving dynamic decision-making problems, it is often used to study the resource allocation problem in communication systems, thus effectively improving computational efficiency. The principle is to update the action selection strategy based on feedback from the environment by continuously interacting with it.

The Q-learning algorithm, as a classic model-free algorithm in reinforcement learning, has been widely applied because of its ability to be implemented in distributed environments. The working principle of this algorithm is to continuously try and make mistakes to obtain the optimal strategy for action selection. Q-learning is a type of Markov Decision Process (MDP) in which the learning model is typically composed of a combination of $\{S, A, R, P\}$. During the learning process, the agent needs to create and continuously update a Q-table that records the expected rewards for different states and actions. In the current state s_i within the state set S, the agent selects the optimal action a_j from the set of available actions A and executes it, aiming to maximize the reward $R(s_i, a_j)$. The purpose is to adjust the action selection strategy according to the reward. If the reward for an action is large, the agent will be more inclined to select this action; conversely, if the reward is small or negative, the action will be avoided. P represents the probability distribution of transitioning from the current state s_i to the next state $s_i{}'$, denoted as $p(s_i{}'|s_i, a_j)$.

During the learning process, Q-learning continuously updates the Q-table based on the rewards received. After multiple iterations, the Q-table converges and the agent can determine the optimal strategy based on the magnitude of the Q-values. The update rule for the Q-values is defined as follows

$$Q(s_i, a_j) \leftarrow (1-\alpha)Q(s_i, a_j) + \alpha[R(s_i, a_j) + \gamma \max_{a'} Q(s', a')], \tag{32}$$

where $Q(s_i, a_j)$ represents the expected reward obtained by taking action a_j in the current state s_i. Based on this value, the agent can determine how to choose actions in the current state to maximize the rewards. $\alpha \in (0, 1)$ is the learning rate, which indicates the learning speed of an agent. When α is small, according to (32), the first term of the expression occupies a larger proportion, which means that the agent focuses more on the previous learning outcomes. $\gamma \in (0, 1)$ is a discount factor that reflects the importance of future rewards. A higher value of γ indicates greater importance of future rewards.

In the system model proposed in this paper, the network serves as the agent for the Q-learning algorithm, and the state-action pairs (S, A) are represented by the pair of source node transmit power and relay node transmit power (P_S, P_R). To ensure a continuous update of the Q-table, it is necessary to ensure that all state-action pairs have a probability of being selected. This means that action selection needs to consider the trade-off between exploration and exploitation. The ε-greedy strategy [40] is employed to select actions and update the Q-table to prevent the algorithm from becoming stuck in local optima. The principle is to set a parameter ε, then choose the power levels of source and selected relay nodes with the maximum Q-value with probability ε, and randomly select the power levels of the two nodes with probability 1 - ε, where $\varepsilon \in [0, 1]$. This ensures that even if the agent becomes trapped in a local optimum, there is an opportunity to break out of such a situation and eventually achieve the global optimum.

Because the goal of Q-learning is to learn an action selection strategy that maximizes the cumulative reward, this paper defines the reward of the DF(AF) relay system as the difference between the current and previous iterations of $\gamma_{R_k m\to \mathrm{m,DF}}(\gamma_{R_k m\to \mathrm{m,AF}})$, while ensuring secure message transmission and successful decoding. If the SINR for weak user m increases, the instantaneous reward is positive; otherwise, the instantaneous reward is negative. The reward function defined in this manner achieves two objectives simultaneously: ensuring that the weak user m achieves optimal communication performance while ensuring successful decoding for the strong user n. Specifically, the reward functions for the two scenarios are defined as follows:

$$R_{DF} = \begin{cases} -1000, & if \ \min\{\gamma_{SR_k\to m}, \gamma_{R_k n\to m,DF}\} \le \tau_{m,t} \\ & or \ \min\{\gamma_{SR_k\to n}, \gamma_{R_k n\to n,DF}\} \le \tau_{n,t} \\ 0, & if \ \max\{\gamma_{SE\to m}, \gamma_{RE\to m,DF}\} \ge \tau_{m,t} - \tau_{m,s} \\ & or \ \max\{\gamma_{SE\to n}, \gamma_{RE\to n,DF}\} \ge \tau_{n,t} - \tau_{n,s} \\ \gamma_{R_k m\to m,DF,t+1} - \gamma_{R_k m\to m,DF,t}, & else, \end{cases} \tag{33}$$

$$R_{AF} = \begin{cases} -1000, & if \ \gamma_{R_k n\to m,AF} \le \tau_{m,t} \ or \ \gamma_{R_k n\to n,AF} \le \tau_{n,t} \\ 0, & if \ \max\{\gamma_{SE\to m}, \gamma_{RE\to m,AF}\} \ge \tau_{m,t} - \tau_{m,s} \\ & or \ \max\{\gamma_{SE\to n}, \gamma_{RE\to n,AF}\} \ge \tau_{n,t} - \tau_{n,s} \\ \gamma_{R_k m\to m,AF,t+1} - \gamma_{R_k m\to m,AF,t}, & else, \end{cases} \tag{34}$$

where R_{DF} and R_{AF} represent the reward functions for Q-learning when the system adopts DF and AF protocols, respectively.

Throughout the entire communication process in the system, if there is a communication interruption, the reward is -1000, and the SINR for decoding by m is set to 0. If the message is intercepted by E in a successful transmission scenario, the reward is 0, and the SINR for decoding by m is set to 0. Under the condition of successful and secure transmission, the reward is the difference in the SINR for decoding by m between two consecutive iterations. The purpose of this setup is to optimize the communication performance of the edge user as much as possible while maintaining a secure transmission. After each iteration, the Q-table is updated based on (32), and actions are selected based on the updated Q-table for the next iteration, until the loop ends.

Algorithm 1 Power Allocation Algorithm for Cooperative Relay Networks Based on Q-Learning

```
Input: $Q(S, A) = 0$, $\alpha \in (0, 1)$, $\gamma \in (0, 1)$, $\varepsilon \in [0, 1]$
Output: Optimum transmit power pair $(P_S{}^*, P_R{}^*)$
for $i = 1$ to max_episodes do
    for $j = 1$ to iteration times do
        Initialize a random number $\mu \in [0, 1]$
        if $\mu > \varepsilon$ then
            Select the $(P_S, P_R)$ with the highest Q-value
        else
            Select the $(P_S, P_R)$ randomly
        end if
        Select the optimal forwarding relay
        Obtain immediate reward via (29) and (30)
        Update the Q-value for $(P_S, P_R)$ pair via (28)
    end for
end for
```

Algorithm 1 describes a power allocation algorithm using Q-learning. First, the Q-table is initialized to 0. Subsequently, the transmit power of the source node and relay is selected based on the ε-greedy strategy. When the Q-table has more than one maximum value at the time of selecting, one of the maximum values is randomly selected. Subsequently, the two nodes transmit signals based on the selected power levels, and the reward is calculated to update the Q-table. After multiple iterations, the Q-table gradually stabilizes, resulting in an optimal power allocation strategy.

V. SIMULATION AND NUMERICAL RESULTS

In this section, based on the proposed power allocation scheme, the performance of DF relay and AF relay

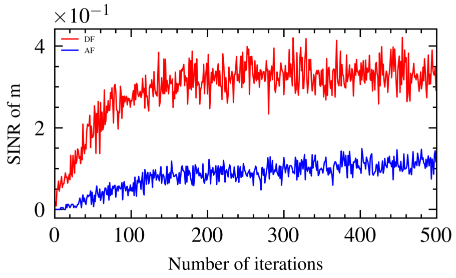


FIGURE 2. SINR versus number of iterations for *m* decoding.

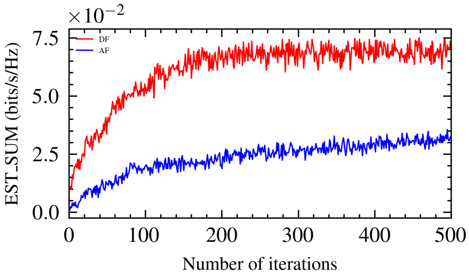
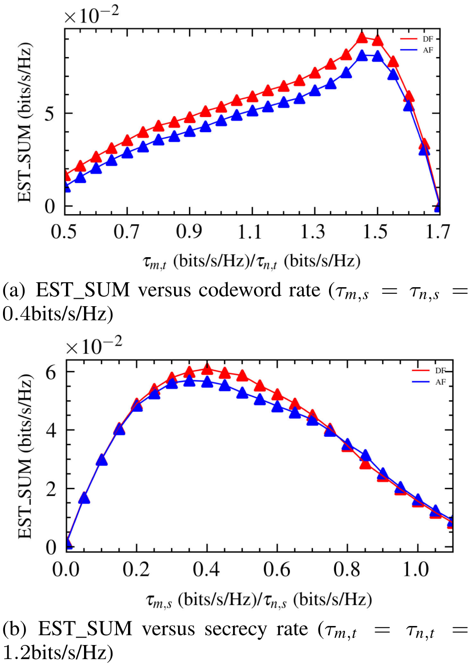


FIGURE 3. EST_SUM of the system versus number of iterations.



(a) EST_SUM versus codeword rate ($\tau_{m,s} = \tau_{n,s} = 0.4$bits/s/Hz)

(b) EST_SUM versus secrecy rate ($\tau_{m,t} = \tau_{n,t} = 1.2$bits/s/Hz)

FIGURE 4. EST_SUM of the system versus codeword rate and secrecy rate.

is compared. The effects of various parameters such as the number of relays N, transmit power of the source node P_S, forwarding power of the relay node P_R, and interference power P_J of the relay R_j on the communication performance are also investigated. This study considers a scenario that includes a source node, an eavesdropper, 8 relays, and 2 users. All relays are randomly distributed in a rectangular area with a horizontal distance of 30 m and a vertical distance of 20 m. One relay is responsible for forwarding the message, whereas another relay interferes with the eavesdropper. The mean of the Rayleigh fading channel is 0, and the variance is 1. The other parameters are set as: $d_m = 800$ m, $d_n = 400$ m. $\tau_{m,t} = \tau_{n,t} = 1.2$ bits/s/Hz, $\tau_{m,s} = \tau_{n,s} = 0.4$ bits/s/Hz, $P_{\max} = 60$ dB, $P_{\min} = -40$ dB, $P_J = 15$ dB, $N_0 = -30$ dB, $a_1 = 0.9$, $a_2 = 0.1$, discount factor $\gamma = 0.01$ and greedy factor $\varepsilon = 0.4$. Because the impact of changing the learning rate on the results is not significant, the learning rate is set to $\alpha = 0.01$.

Figure 2 shows the SINR for decoding its own signal by m versus the number of iterations using the DF and the AF protocols. The figure shows the average results of 1000 simulation runs, each with 500 iterations. It is shown that the decoding SINR fluctuates for both protocols owing to the introduction of the ε-greedy strategy to avoid local optima [41]. However, the overall trend is gradually increasing and both converge to the maximum value after about 200 iterations. In addition, during the iteration process, the decoding SINR obtained using the DF protocol for m is higher than that obtained using the AF protocol. This is because in the AF relay system, the amplification gain of the relay node is fixed and cannot be adaptively adjusted based on the signal strength. Amplifying the signal at the relay node also amplifies the noise in the signal, which may cause signal distortion. On the other hand, the DF relay system improves signal reliability through decoding and re-encoding. Therefore, the DF protocol has a significant advantage in terms of improving the communication quality of edge users.

Figure 3 shows the EST_SUM of the system versus the number of iterations for the same scenario. This represents the average results obtained from 1000 runs. We can observe that the EST_SUM of both protocols gradually increases and converges with an increase in the number of iterations, reaching their maximum values at approximately 200 iterations. In addition, during the Q-learning process, the EST_SUM obtained by the DF protocol is always greater than that obtained by the AF protocol. Therefore, the DF protocol can achieve better security performance through Q-learning.

The effects of codeword rate and secrecy rate on the system performance are shown in Figure 4. The figure shows the EST that the system can achieve after the Q-learning process with different parameter settings. In Figure 4(a), the secrecy rate is set to $\tau_{m,s} = \tau_{n,s} = 0.4$ bits/s/Hz for both users, and the codeword rate is varied in the range of 0.5 bits/s/Hz to 1.7 bits/s/Hz. In Figure 4(b), the codeword rate is set to $\tau_{m,t} = \tau_{n,t} = 1.2$ bits/s/Hz for both users, and the secrecy rate is varied in the range of 0bits/s/Hz to 1.1bits/s/Hz. It can be seen that the EST of the system increases and then decreases as both the codeword rate and the secrecy rate increase. And the EST reaches the maximum value at about $\tau_{m,t} = \tau_{n,t} = 1.5$ bits/s/Hz, $\tau_{m,s} = \tau_{n,s} = 0.4$ bits/s/Hz. Therefore, setting the values of the parameters codeword rate and secrecy rate reasonably can effectively enhance the security performance of the system.

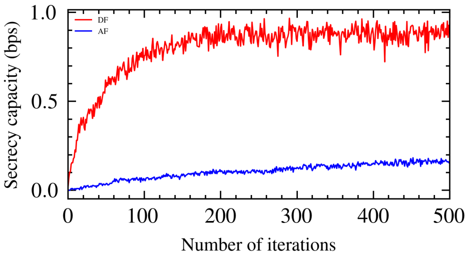


FIGURE 5. Secrecy capacity of the system versus number of iterations.

As Q-learning proceeds, the change in the secrecy capacity is shown in Figure 5. The sum of the secrecy capacities of the two users can be expressed as $C_{SUM} = C_m + C_n$, where the secrecy capacities of m and n are $C_m = [C_{m\_t} - C_{m\_e}]^+$, $C_n = [C_{n\_t} - C_{n\_e}]^+$ respectively. C_{m_t} and C_{n_t} refer to the communication channel capacity of the two users, C_{m_e} and C_{n_e} refer to the capacity of the eavesdropping channel. From the figure, it can be observed that as the number of iterations increases, the secrecy capacities obtained by both protocols show an initial increase, followed by convergence at approximately 200 iterations. Compared with the results obtained in [38], the communication system proposed in this paper can achieve better secrecy capacity. Moreover, a larger secrecy capacity can be obtained with the help of the DF protocol compared to the AF protocol, which indicates that the DF relaying system has higher security.

Figure 6 shows the EST_SUM of the system for different numbers of relays versus the transmit power of the source node. The number of relays is considered to be 4, 12 and 20. In scenarios with different numbers of relays, it is clear that using DF relays can lead to a larger EST_SUM and achieve better performance. EST_SUM of the system first increases and then decreases with increasing P_S. This indicates that increasing P_S is not always beneficial for improving the EST_SUM of the system. When P_S is small, transmission interruption is prone to occur, and when P_S is large, E decodes with a large SINR. This is prone to message leakage, which leads to the disruption of confidentiality. Therefore, there exists the optimal value of P_S that maximizes the EST of the system. Furthermore, P_S within the range of -20 dB to 0 dB can achieve a non-zero EST_SUM, and within this range, increasing the number of relays improves the performance for both approaches.

The relationship between EST_SUM of the system and the transmit power of the forwarding relay P_R with different numbers of relays is shown in Figure 7. The case where the number of relays is 4, 12 and 20 is also considered. Similarly, the EST of the system increases and then decreases with an increase in P_R, indicating that the transmission power of both nodes affects the security performance. It is clear that changing the number of relays has no significant impact on the two protocols when P_S is fixed and 4 relays are sufficient. When P_R>−20 dB, the EST of the AF relay system gradually decreases and is significantly smaller than that of the DF relay system. And at P_R>10dB, EST_SUM of the AF relay system is zero, whereas the DF relay system can obtain a non-zero EST_SUM. Therefore, compared with the AF protocol, the DF protocol can achieve better performance when P_R is large.

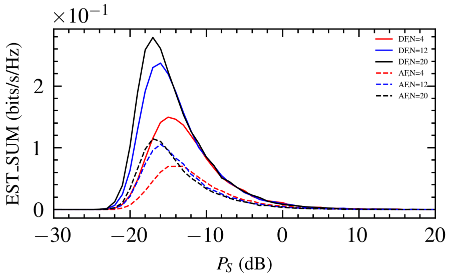


FIGURE 6. EST_SUM of the system for different number of relays versus P_S. (P_R = −8dB).

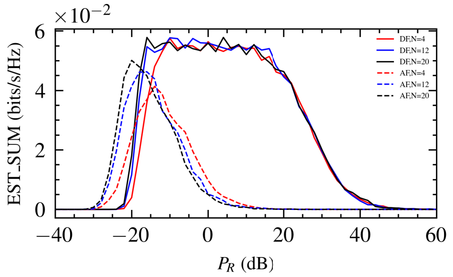


FIGURE 7. EST_SUM of the system for different number of relays versus P_R. (P_S = −8dB).

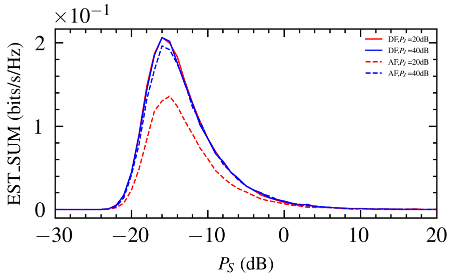


FIGURE 8. EST_SUM of the system for different interference powers versus P_S. (P_R = −8dB).

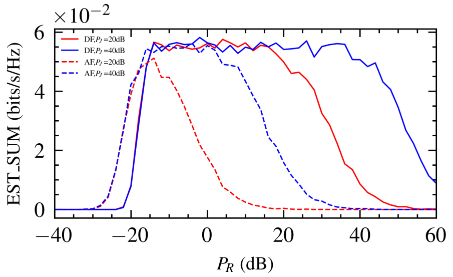


FIGURE 9. EST_SUM of the system for different interference powers versus P_R. (P_S = −8dB).

Figure 8 and Figure 9 show the EST_SUM of the system versus the transmit power of the source node and relay node

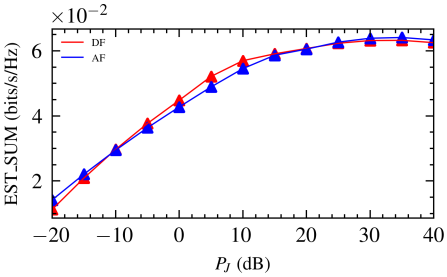


FIGURE 10. EST_SUM of the system versus P_J.

for different interference powers. In Figure 8, EST_SUM of the system increases and then decreases with increasing P_S. The EST_SUM curves for the two protocols overlap at $P_S = 10$dB, at which point the change in interference power has no effect on the EST of the system. This is because the transmit power is too high and the decoding SNR of E is also large, and increasing the interference power at this point does not significantly help reduce the decoding capability of E. The EST_SUM of both systems peaks at approximately $P_S = -17$ dB. Moreover, the influence of the interference power on the performance of the AF relay system is more obvious. In Figure 9, When P_R is small, the two systems have similar performance, which is due to the fact that the lower forwarding power is prone to communication interruptions, and at this time, changing the decoding SINR of E will not significantly improve the communication performance. When P_R is large, for both systems, increasing the interference power yields a larger EST_SUM and effectively improves the security of communication.

In addition, the effect of interference power P_J on EST is shown in Figure 10. The EST of the system first increases with the increase of P_J. However, Continuously increasing P_J does not improve the EST significantly and 20 dB is enough. According to (17), (18), (27) and (28), when the interference power P_J to E is increased, the decoded SINR during eavesdropping decreases, thus improving the security of the system by interfering with eavesdropping. However, when P_J increases to a certain value, the decoded SINR of E decreases to a critical value and is unable to meet the requirements for successful eavesdropping, thus failing to satisfy the eavesdropping condition. Therefore, when P_J increases to a certain value, continuing to increase P_J will not significantly improve the security performance of the system. As a result, reasonable adjustment of the settings of parameters such as codeword rate, secrecy rate, and interference power can effectively enhance the communication performance of the system.

Figure 11 shows the average EST_SUM for 1000 simulation runs with different numbers of relays using the DF protocol. The values of P_S and P_R range from −20dB to 10dB, assuming $P_T = -5$dB as the total power constraint, that is, $P_S + P_R \leq P_T$. Five scenarios for the number of relays are considered: 4, 12, 20, 28 and 36. And the proposed Q-learning based power allocation method is compared with two common power allocation schemes: the equal power allocation (EPA) and the random power allocation (RPA) algorithms. For the EPA algorithm, both nodes have the same transmit power, that is, $P_S = P_R = P_T/2$. For the RPA algorithm, the transmission power of both nodes is randomly chosen from a range of −20dB to 10dB. It can be observed that, compared to the other two approaches, the power allocation algorithm using Q-learning achieves a larger average EST_SUM for different numbers of relays, and EST increases as the number of relays increases. The system performance obtained by the EPA algorithm does not change significantly with an increase in the number of relays. Moreover, as the number of relays increases, the EST_SUM obtained by the RPA algorithm becomes correspondingly larger, but the increase is slower than that of the Q-learning algorithm. Therefore, the Q-learning algorithm can improve the communication performance better, especially when the

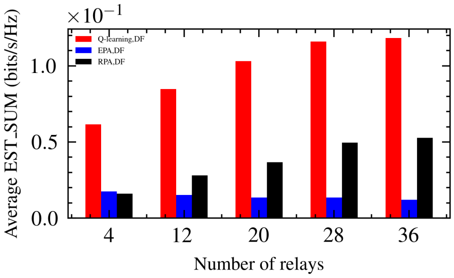


FIGURE 11. Average EST of the system versus the number of relays in different methods when using DF protocols.

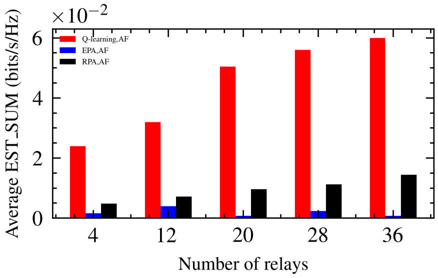


FIGURE 12. Average EST of the system versus the number of relays in different methods when using AF protocols.

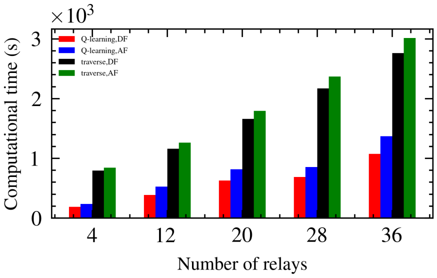


FIGURE 13. Computation time of the algorithm versus number of relays for both protocols.

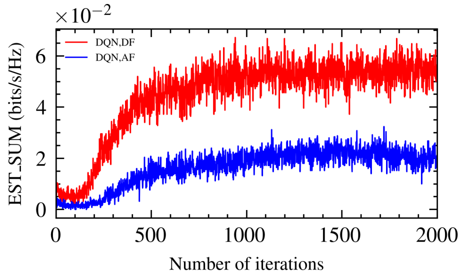
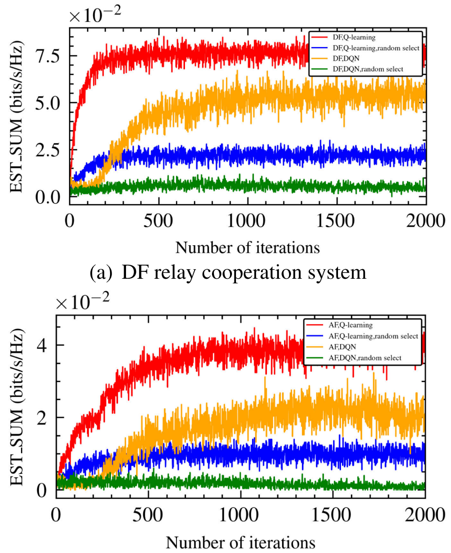


FIGURE 14. EST_SUM of the system obtained using the DQN method versus number of iterations.



(a) DF relay cooperation system

(b) AF relay cooperation system

FIGURE 15. Comparison of DQN and Q-learning methods with and without relay selection.

system contains a large number of relays. This is because of the high probability of selecting the best relay when the number of relays is high, and the Q-learning method is more likely to obtain the best power pair and achieve a better safety performance.

Figure 12 shows an image of the AF protocol under the same conditions. Similarly, for different numbers of relays, the EPA scheme obtains the worst system performance, whereas the proposed Q-learning based algorithm shows a significant advantage compared to the other two schemes. Similar to Figure 12, the EPA algorithm obtains a similar average EST_SUM for different numbers of relays, whereas the RPA method obtains an increasing EST with an increasing number of relays. Overall, Q-learning approach offers significant advantages in terms of improved confidentiality.

Figure 13 shows the time performance of the two protocols for different numbers of relays. The performance of finding the optimal power allocation pair using Q-learning and traversal methods is compared. From the graph, it can be observed that the running time of both methods increases linearly with the number of relays under both protocols. This is due to the fact that the higher the number of relays, the longer it takes to traverse the relays. For both methods, the running time of the DF protocol is slightly shorter than that of the AF protocol. Therefore, a better time performance can be obtained using the DF protocol. In addition, it is clear that the time required to use the traversal method is much greater than that of the Q-learning algorithm, particularly when the number of relays is high. This shows that the use of the Q-learning algorithm significantly reduces computational complexity and improves communication efficiency.

In this paper, the DQN algorithm from [22] is also considered to solve the power allocation problem. A simple neural network model is constructed in this paper, which consists of two fully connected (FC) layers. The size of the experience buffer is 30, the batch size is 20, and the weights of the target network are updated every 20 time slots. Other than that, the other learning parameters are kept consistent with the Q-learning method. The convergence of the DQN algorithm is shown in Figure 14, which illustrates the results of running 1000 times and calculating the average. Each run consisted of 2000 iterations. Compared to Figure 3, the DQN algorithm converges slower and the final optimal EST obtained is slightly smaller than that obtained by the Q-learning algorithm. This is due to the smaller state action space, where the same state may be sampled multiple times, resulting in lower utilization of the samples and slowing down the convergence of the algorithm. At the same time, the complexity of the neural network approximation also affects the performance of the system.

In addition, this paper also considers the case with no optimal relay selection under both RL methods. Figure 15 compares the system performance obtained by selecting the optimal relay and randomly selecting a relay when using the two RL methods. It is clear that the EST obtained by selecting the optimal relay is significantly higher than the EST obtained by randomly selecting a relay in both Q-learning and DQN methods. Therefore optimizing the relay selection strategy is also an effective way to improve security.

TABLE 1. Computation overhead of the two methods.

Scheme	Memory (MB)	Computation time (s)	Convergence time (time slot)
Q-learning	110.48	6140.7	200
DQN	259.97	20050.3	1000

The computational efficiency of the two algorithms is shown in Table 1. Due to the introduction of a deep neural network and an experience replay buffer, the DQN algorithm requires larger parameter storage. Additionally, during each iteration, the DQN algorithm needs to perform forward and backward propagation of the neural network, while the Q-learning algorithm only requires updating a Q-table. Therefore, when the number of states is small, the Q-learning algorithm has lower computational overhead and converges faster and for a relatively small state space like this paper, it is simpler and more efficient to use the Q-learning algorithm.

VI. CONCLUSION

In this study, we investigated a Q-learning based power allocation scheme for relay collaboration networks. The scheme utilizes relay cooperation techniques to reduce the decoding SINR of the eavesdropper and selects a forwarding relay based on certain rules. It then optimizes the transmit power of the source node and the forwarding relay using Q-learning. The goal is to maximize the decoding SINR for a weak user while ensuring successful decoding for a strong user. We comprehensively compared the security performance of the system using the DF and AF relays. Our experimental results indicate that the number of relays, transmit power of the two nodes, and interference power all affect the performance of the system. Furthermore, compared with other commonly used power allocation methods, our proposed algorithm achieves a larger EST_SUM for the system and improves the security performance of system effectively.

REFERENCES

[1] W. Peng, W. Gao, and J. Liu, "A novel perspective on multiple access in 5G network: Framework and solutions," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 154–160, Jun. 2019.

[2] Y. Yin, M. Liu, G. Gui, H. Gacanin, H. Sari, and F. Adachi, "QoS-oriented dynamic power allocation in NOMA-based wireless caching networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 1, pp. 82–86, Jan. 2021.

[3] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1299–1306, Apr. 2018.

[4] 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 Trans. Commun.*, vol. 66, no. 5, pp. 1883–1901, May 2018.

[5] P. Huu, M. A. Arfaoui, S. Sharafeddine, C. M. Assi, and A. Ghrayeb, "A low-complexity framework for joint user pairing and power control for cooperative NOMA in 5G and beyond cellular networks," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6737–6749, Nov. 2020.

[6] H. Pan, J. Liang, and S. C. Liew, "Practical NOMA-based coordinated direct and relay transmission," *IEEE Wireless Commun. Lett.*, vol. 10, no. 1, pp. 170–174, Jan. 2021.

[7] A. Tregancini, E. E. B. Olivo, D. P. M. Osorio, C. H. M. de Lima, and H. Alves, "Performance analysis of full-duplex relay-aided NOMA systems using partial relay selection," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 622–635, Jan. 2020.

[8] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[9] J. Lin, Q. Li, J. Yang, H. Shao, and W.-Q. Wang, "Physical-layer security for proximal legitimate user and eavesdropper: A frequency diverse array beamforming approach," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 671–684, Mar. 2018.

[10] L. Xiao, C. Xie, M. Min, and W. Zhuang, "User-centric view of unmanned aerial vehicle transmission against smart attacks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3420–3430, Apr. 2018.

[11] C. Zhang, F. Jia, Z. Zhang, J. Ge, and F. Gong, "Physical layer security designs for 5G NOMA systems with a stronger near-end internal eavesdropper," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13005–13017, Nov. 2020.

[12] Y. Feng, S. Yan, C. Liu, Z. Yang, and N. Yang, "Two-stage relay selection for enhancing physical layer security in non-orthogonal multiple access," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 6, pp. 1670–1683, Jun. 2019.

[13] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.

[14] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, "Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6369–6379, Jul. 2020.

[15] Y. Xu, J. Yu, and R. M. Buehrer, "The application of deep reinforcement learning to distributed spectrum access in dynamic heterogeneous environments with partial observations," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4494–4506, Jul. 2020.

[16] H. Song, L. Liu, J. Ashdown, and Y. Yi, "A deep reinforcement learning framework for spectrum management in dynamic spectrum access," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11208–11218, Jul. 2021.

[17] Z. Ezzati Khatab, A. Mohammadi, V. Pourahmadi, and A. Kuhestani, "A machine learning multi-hop physical layer authentication with hardware impairments," *Wireless Netw.*, vol. 30, no. 3, pp. 1453–1464, Dec. 2023.

[18] K. Wang, H. Li, Z. Ding, and P. Xiao, "Reinforcement learning based latency minimization in secure NOMA-MEC systems with hybrid SIC," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 408–422, Jan. 2023.

[19] S. Zhang, L. Li, J. Yin, W. Liang, X. Li, W. Chen, and Z. Han, "A dynamic power allocation scheme in power-domain NOMA using actor-critic reinforcement learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2018, pp. 719–723.

[20] J. Moon and Y. Lim, "Access control of MTC devices using reinforcement learning approach," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2017, pp. 641–643.

[21] C. J. Watkins and P. Dayan, "Technical note: Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992.

[22] Y. Zhang, Z. Cheng, D. Guo, S. Yuan, T. Ma, and Z. Zhang, "Downlink resource allocation for NOMA-based hybrid spectrum access in cognitive network," *China Commun.*, vol. 20, no. 9, pp. 171–184, Sep. 2023.

[23] M. Ragheb, A. Kuhestani, M. Kazemi, H. Ahmadi, and L. Hanzo, "RIS-aided secure millimeter-wave communication under RF-chain impairments," *IEEE Trans. Veh. Technol.*, vol. 73, no. 1, pp. 952–963, Jan. 2024.

[24] Y. Zhang, H.-M. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 930–933, May 2016.

[25] I. Amin, D. Mishra, R. Saini, and S. Aïssa, "QoS-aware secrecy rate maximization in untrusted NOMA with trusted relay," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 31–34, Jan. 2022.

[26] I. Amin, D. Mishra, R. Saini, and S. Aïssa, "Secrecy rate maximization in relay-assisted NOMA with untrusted users," in *Proc. IEEE 33rd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2022, pp. 1134–1138.

[27] M. Forouzesh, F. S. Khodadad, P. Azmi, A. Kuhestani, and H. Ahmadi, "Simultaneous secure and covert transmissions against two attacks under practical assumptions," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10160–10171, Jun. 2023.

[28] N. Nayak Vankudoth and K. Kumar Gurrala, "Power allocation assisted new control jamming scheme for NOMA enabled AF relay network," in *Proc. IEEE 17th India Council Int. Conf. (INDICON)*, Dec. 2020, pp. 1–6.

[29] A. Salem, L. Musavian, E. A. Jorswieck, and S. Aïssa, "Secrecy outage probability of energy-harvesting cooperative NOMA transmissions with relay selection," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 4, pp. 1130–1148, Dec. 2020.

[30] H. He and P. Ren, "Secure communications in cooperative D2D networks by jointing Wyner's code and network coding," *IEEE Access*, vol. 7, pp. 34533–34540, 2019.

[31] D. A. Tubiana, J. Farhat, G. Brante, and R. D. Souza, "Q-learning NOMA random access for IoT-satellite terrestrial relay networks," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1619–1623, Aug. 2022.

[32] M. V. da Silva, R. D. Souza, H. Alves, and T. Abrão, "A NOMA-based Q-Learning random access method for machine type communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1720–1724, Oct. 2020.

[33] Y. Su, M. Liwang, Z. Gao, L. Huang, X. Du, and M. Guizani, "Optimal cooperative relaying and power control for IoUT networks with reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 791–801, Jan. 2021.

[34] P. Gong, H. Li, X. Gao, D. O. Wu, and X. Xiao, "Relay power allocation for NAF cooperation assisted NOMA network," *IEEE Wireless Commun. Lett.*, vol. 9, no. 8, pp. 1169–1172, Aug. 2020.

[35] X. Qiu, G. Li, X. Sun, and Z. Du, "Exploiting user selection algorithm for securing wireless communication networks," in *Proc. IEEE 19th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Dec. 2020, pp. 1552–1555.

[36] J. Zhang, "A Q-learning based method F or secure UAV communication against malicious eavesdropping," in *Proc. 14th Int. Conf. Comput. Autom. Eng. (ICCAE)*, Brisbane, QLD, Australia, Mar. 2022, pp. 168–172.

[37] V. Shahiri, M. Forouzesh, H. Behroozi, A. Kuhestani, and K.-K. Wong, "Deep learning aided secure transmission in wirelessly powered untrusted relaying in the face of hardware impairments," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 2196–2210, 2024.

[38] J. Lu, D. He, and Z. Wang, "Learning-assisted secure relay selection with outdated CSI for finite-state Markov channel," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Apr. 2021, pp. 1–5.

[39] E. M. Ghourab, L. Bariah, S. Muhaidat, 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, 2023.

[40] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

[41] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan, and D. Matolak, "A machine learning approach for power allocation in HetNets considering QoS," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

WEIDONG GUO received the B.E. degree in electronic and information engineering from Hohai University, Nanjing, China, in 2005, and the Ph.D. degree in communication and information system from Shandong University, Jinan, China, in 2012. From 2017 to 2022, he was a Postdoctoral Fellow with the School of Information Science and Engineering, Shandong University. Since July 2012, he has been with the School of Cyber Science and Engineering, Qufu Normal University. His current research interest includes the application of machine learning techniques to the physical layer of 5G and beyond.

XINYAN WANG received the B.S. degree in information and computational science from Qufu Normal University, Shandong, China, in 2022, where she is currently pursuing the M.S. degree in computer technology. Her research interests include non-orthogonal multiple access, wireless communications, and reinforcement learning.