Joint Buffer-Aided Hybrid-Duplex Relay Selection and Power Allocation for Secure Cognitive Networks with Double Deep Q-Network

Chong Huang, *Graduate Student Member, IEEE*, Gaojie Chen, *Senior Member, IEEE*, Yu Gong, and Zhu Han, *Fellow, IEEE*

***Abstract*—This paper applies the reinforcement learning in the joint relay selection and power allocation in the secure cognitive radio (CR) relay network, where the data buffers and full-duplex jamming are applied at the relay nodes. Two cases are considered: maximizing the throughput with the delay and secrecy constraints, and maximizing the secrecy rate with the delay constraint, respectively. In both cases, the optimization relies on the buffer states, the interference to/from the primary user, and the constraints on the delay and/or secrecy. This makes it mathematically intractable to apply the traditional optimization methods. In this paper, the double deep Q-network (DDQN) is used to solve the above two optimization problems. We also apply the a-priori information in the CR network to improve the DDQN learning convergence. Simulation results show that the proposed scheme outperforms the traditional algorithm significantly.**

***Index Terms*—Buffer-aided relay selection, power allocation, secure cognitive radio networks, double deep Q-Network, delay.**

I. Introduction

With the development of 5th generation (5G) communications, both cognitive networks and relay networks have attracted much attention in current research topics [1]. Cognitive relay networks allow users to share the spectrum to improve spectral efficiency, and ensure cooperations among secondary users to enhance communication reliability [2]. Various relay selection algorithms in cognitive relay networks have been proposed. In [3], a Max-Min based relay selection method was intended to improve the outage performance in decode-and-forward (DF) cognitive relay networks. An outage probability-based relay selection was investigated in cognitive relay networks where the relay nodes are randomly distributed [4]. Buffer-aided relay selection as a robust scheme can further enhance the outage performance in cooperative networks [5]. In [6] and [7], the Max-Link relay selection scheme was proposed to select the available link with the highest signal-to-noise (SNR) for DF and amplify-and-forward (AF) buffer-aided relay networks, respectively. To enhance the outage performance in buffer-aided cognitive relay networks, the available link with the highest signal-to-interference ratio (SIR) was selected for the relay selection scheme in [8]. Then, in [9], a state-based relay selection scheme was proposed to group all links with different priorities to further improve the outage performance in buffer-aided cognitive relay networks. Though applying buffer technology in cooperative networks can reduce the outage probability significantly, it often results in higher delays in data transmissions [10]–[12]. Therefore, this paper will consider the delay as one of the constraints to do the buffer-aided relay selection.

On the other hand, the physical layer secrecy has been widely studied in cooperative communications. In [13], a Max-Ratio relay selection scheme was proposed based on selecting the highest signal-to-interference-ratio (SIR) to reduce the secrecy outage probability in buffer-aided relay networks. The Max-Ratio based relay selection scheme was proposed in the buffer-aided cognitive relay network to reduce the secrecy outage probability in [14]. In [15], the secrecy performance was analyzed by applying the Max-Link scheme in the buffer-aided cognitive relay network with the impact of outdated channel state information (CSI). The secrecy outage performance of the Max-Link scheme was studied for the multi-antenna buffer-aided cognitive relay networks in [16]. However, aforementioned studies only consider the security in half-duplex (HD) relay systems.

The full-duplex (FD) is an attractive way to improve the secrecy performance in wireless communications, because of the successful use of the self-interference cancellation scheme [17]–[19]. In [17], the secrecy is improved by directing the FD jamming to the eavesdropper. In [18] and [19], the FD relay network achieves better secrecy performance with the self-interference cancellation. In [20], the optimal secrecy performance was achieved with the convex optimization in the FD relay network by using power allocation. A joint relay selection and power allocation scheme was proposed by using the dominant balance to achieve a high secrecy rate in [21]. In [22], a simple joint relay selection and power allocation scheme

The work of Gaojie Chen was supported by EPSRC grant number EP/R006377/1 ("M3NETs"). The work of Zhu Han was supported by NSF EARS-1839818, CNS1717454, CNS-1731424, and CNS-1702850.

C. Huang and G. Chen are with School of Engineering, University of Leicester, UK, Email: {ch481, gaojie.chen}@leicester.ac.uk.

Y. Gong is with Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, UK, Email: y.gong@lboro.ac.uk.

Z. Han is with the Department of Electrical and Computer Engineering in the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea, 446-701, Email: zhan2@uh.edu.

based on the SNR was proposed to maximize the instantaneous secrecy rate.

The machine learning methods have attracted much attention in the relay network [23]. For example, in [24], a deep Q-Learning based relay selection was proposed to achieve better outage performance. In [25], a relay selection based on deep reinforcement learning was proposed to enhance the robustness. In [26], the deep Q-Learning was applied to select the virtual vehicle relay node. None of these approaches considers buffers and/or full duplex transmission at the relay, which will significantly increase the searching dimension for the learning, making it harder to converge.

In this paper, we consider the joint relay selection and power allocation in the buffer-aided cognitive relay network, where the relay node may work in the standard half-duplex mode or full-duplex jamming mode. Due to the integer nature of the optimization problem, the complexity will go very high with the number of variables. Therefore, to solve this type of problems, deep reinforcement learning can be introduced in wireless communications [27], [28]. For better convergence, we propose to apply the double deep Q-Network (DDQN) [29]. By applying two neural networks to estimate the next action and the target Q-value, respectively, the DDQN algorithm is able to reduce the overestimation effects which is particularly serious in the secure buffer-aided cognitive relay network. This is because applying the data buffers and full-duplex jamming at the relays leads to very high dimension for learning. As a result, the randomly initialized Q-values can be easily overestimated, leading to slow convergence. The main contributions of this paper are listed as follows:

- We investigate the DDQN learning-based joint buffer-aided hybrid-duplex relay selection and transmit power allocation (DDQN-RP) scheme to solve two complicated optimization problems in the secure cognitive networks.
- We propose the a-priori information based DDQN learning, which is used to analyse the impact of variables and tuples in the system and improve the convergence of DDQN.
- The simulation result confirms that the performance of the DDQN-based hybrid-duplex relay selection scheme with priori information (DDQNPI-RP) outperforms DDQN-RP and Max-Ratio scheme with the HD relay.

The rest of the paper is organized as follows: Section II describes the system model; Section III defines the MDP tuples for two relay selection cases; Section IV applies the DDQN-based algorithm in the CR relay network; Section V applies a-priori information in the DDQN approach; Section VI shows simulation result to verify the proposed schemes; Finally, Section VII concludes the paper.

II. System Model and Problem Formulation

A. System Model

The system model of a secure buffer-aided cognitive relay network is shown in Fig. 1, where there are one pair of primary source PS and destination PD, one pair of secondary source SS and destination SD, and a set of K randomize-and-forward[1] relays R_k which is equipped with a data buffer of finite size L, where $k \in \{1, 2..., K\}$. There is an untrusted node[2] E to eavesdrop the transmission between SS to SD. Every relay node R_k and SD can perform at either the HD or the FD mode. When a node performs at the HD mode, it only receives data. When a node is at the FD mode, it receives data and transmits jamming signal at the same time, but cannot transmit data to SD.

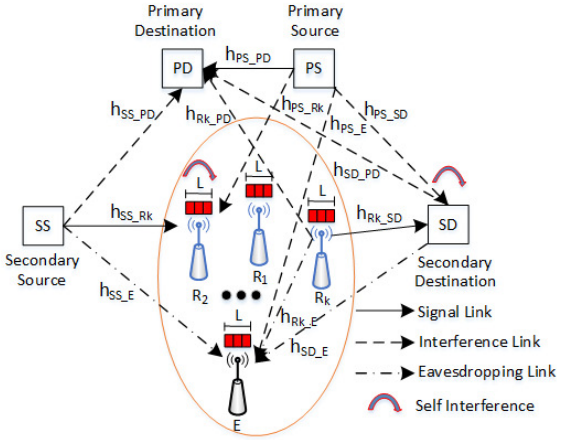


Fig. 1. System model of the secure buffer-aided cognitive relay network.

We assume that all channels are assumed to be quasi-static Rayleigh fading and there is no direct link between SS and SD [31]. The channel coefficients between node m and node n are denoted as $g_{m\_n} = h_{m\_n} d_{m\_n}^{-\alpha/2}$, where h_{m_n} and d_{m_n} are the fading coefficients and the distance between two nodes, respectively, and α is the path loss exponent. $\mathbb{E}[|g_{m\_n}|^2] = \lambda_{m\_n} = d_{m\_n}^{-\alpha}$ is the average channel gain, where $\mathbb{E}[.]$ is the expectation operator, and h_{m_n} is the complex Gaussian random variable with unit variance. At every time slot, only one channel link can be selected for data transmission. The source node SS receives the instantaneous buffer states and the CSI of all channels[3].

This paper considers underlay cognitive relay network that the interference to the primary PD must be smaller than the pre-

[1]Randomize-and-forward strategy employs different codebooks for the two hops, and the legitimate user utilizes the Wyner wiretap code at different transmission time slots [30].

[2]The untrusted node, which is usually a user in the secondary networks, may have potential malicious behaviour such as eavesdropping [22].

[3]In general, either the source node or a central node applies the relay selection, where the required CSI can be obtained through independent feedback links, or based on other methods such as the reciprocal equivalent channel. The details of the channel estimation are out of the scope of this work, which can be found in [32]

defined level I_{th}. On the other hand, the total transmit power[4] from all transmit nodes at any time is also limited to P_b. To be specific, when a $SS \to R_k$ link is selected, the transmit powers for SS and R_k are given by βP_b and $(1-\beta)P_b$, respectively, where $\beta \in (0,1]$ which is the power split factor. It is clear that when $\beta = 1$, R_k operates at the HD mode that it only receives signal from SS and otherwise at the FD mode by transmitting the jamming signal at the same time. Similarly, when a $R_k \to SD$ link is selected, the transmit powers for R_k and SD are given by βP_b and $(1-\beta)P_b$, respectively. With these considerations, the transmit powers at SS, R_k and SD are obtained as

$$\begin{aligned} P_{SS} &= \min\left(\bar{P}_{SS}, \beta P_b\right) \quad \text{and} \quad P^J_{R_k} = \min\left(\bar{P}_{R_k}, (1-\beta)P_b\right), \\ P^T_{R_k} &= \min\left(\bar{P}_{R_k}, \beta P_b\right) \quad \text{and} \quad P_{SD} = \min\left(\bar{P}_{SD}, (1-\beta)P_b\right), \end{aligned} \tag{1}$$

respectively, where $P^J_{R_k}$ and $P^T_{R_k}$ are the data and jamming transmit power at R_k respectively, and

$$\bar{P}_{SS} = \frac{I_{th}}{|g_{SS\_PD}|^2},\ \bar{P}_{R_k} = \frac{I_{th}}{|g_{R_k\_PD}|^2} \text{ and } \bar{P}_{SD} = \frac{I_{th}}{|g_{SD\_PD}|^2} \tag{2}$$

At a time slot, when a $SS \longrightarrow R_k$ link is selected after the use of self-interference cancellation scheme [33], the received signal at R_k is given by

$$\begin{aligned} y_{R_k} =& \sqrt{P_{SS}} g_{SS\_R_k} x_S + \sqrt{P_{PS}} g_{PS\_R_k} x_P \\ &+ \sqrt{\rho P^J_{R_k}} g_{R_k} x_{J1} + n_{R_k}, \end{aligned} \tag{3}$$

where x_S and x_P are the data signals from SS and PS respectively, $\rho \in [0,1]$ is the self-interference cancellation factor, where $\rho = 0$ denotes that there is no self-interference when R_k operates in the HD mode, g_{R_k} is the residual self-interference channel gain due to the FD transmission at R_k, x_{J1} is the jamming signal at R_k, P_{PS} is the transmit power at PS, n_{R_k} is the additive-white-Gaussian-noise (AWGN) noise with variance σ_n^2 at R_k. The intercepted signal at E is given by

$$\begin{aligned} y_{SE} =& \sqrt{P_{SS}} g_{SS\_E} x_S + \sqrt{P_{PS}} g_{PS\_E} x_P \\ &+ \sqrt{P^J_{R_k}} g_{R_k\_E} x_{J1} + n_E, \end{aligned} \tag{4}$$

where n_E is the AWGN noise with variance σ_n^2 at E.

The channel capacities for $SS \longrightarrow R_k$ and $SS \longrightarrow E$ links are given by

$$\begin{aligned} C_{SR_k} &= \frac{1}{2}\log_2\left(1 + \frac{P_{SS}\frac{|h_{SS\_R_k}|^2}{d^\alpha_{SS\_R_k}}}{P_{PS}\frac{|h_{PS\_R_k}|^2}{d^\alpha_{PS\_R_k}} + \rho P^J_{R_k}|h_{r_k}|^2 + \sigma_n^2}\right), \\ C_{SE} &= \frac{1}{2}\log_2\left(1 + \frac{P_{SS}\frac{|h_{SS\_E}|^2}{d^\alpha_{SS\_E}}}{P_{PS}\frac{|h_{PS\_E}|^2}{d^\alpha_{PS\_E}} + P^J_{R_k}\frac{|h_{R_k\_E}|^2}{d^\alpha_{R_k\_E}} + \sigma_n^2}\right), \end{aligned} \tag{5}$$

respectively, where $\rho P^J_{R_k}|h_{r_k}|^2$ is the residual self-interference of node R_k when R_k operates in the FD mode [33]. If the receiver operates in the HD mode, there is no self-interference.

On the other hand, when a $R_k \longrightarrow SD$ link is selected by using self-interference cancellation scheme, the received signal at SD is given by

$$\begin{aligned} y_D =& \sqrt{P^T_{R_k}} g_{R_k\_SD} x_{R_k} + \sqrt{P_{PS}} g_{PS\_SD} x_P \\ &+ \sqrt{\rho P_{SD}} g_{SD} x_{J2} + n_D, \end{aligned} \tag{6}$$

where x_{R_k} is the data signal from R_k, n_D is the AWGN noise with variance σ_n^2 at SD, g_{SD} is the residual self-interference channel gain, x_{J2} is the jamming signal at node SD. And the intercepted signal at E is given by

$$\begin{aligned} y_{R_kE} =& \sqrt{P^T_{R_k}} g_{R_k\_E} x_{R_k} + \sqrt{P_{PS}} g_{PS\_E} x_P \\ &+ \sqrt{P_{SD}} g_{SD\_E} x_{J2} + n_E. \end{aligned} \tag{7}$$

The channel capacities of $R_k \longrightarrow SD$ and $R_k \longrightarrow E$ links are given by

$$\begin{aligned} C_{R_kD} &= \frac{1}{2}\log_2\left(1 + \frac{P^T_{R_k}\frac{|h_{R_k\_SD}|^2}{d^\alpha_{R_k\_SD}}}{P_{PS}\frac{|h_{PS\_SD}|^2}{d^\alpha_{PS\_SD}} + \rho P_{SD}|h_{SD}|^2 + \sigma_n^2}\right), \\ C_{R_kE} &= \frac{1}{2}\log_2\left(1 + \frac{P^T_{R_k}\frac{|h_{R_k\_E}|^2}{d^\alpha_{R_k\_E}}}{P_{PS}\frac{|h_{PS\_E}|^2}{d^\alpha_{PS\_E}} + P_{SD}\frac{|h_{SD\_E}|^2}{d^\alpha_{SD\_E}} + \sigma_n^2}\right). \end{aligned} \tag{8}$$

where $\rho P_{SD}|h_{SD}|^2$ is the residual self-interference of node SD when SD operates in the FD mode, and there is no self-interference when SD operates in the HD mode.

Then the secrecy capacities for $SS \longrightarrow R_k$ and $R_k \longrightarrow SD$ links can be obtained as

$$\begin{aligned} C_{s(SR_k)} &= [C_{SR_k} - C_{SE}]^+, \\ C_{s(R_kD)} &= [C_{R_kD} - C_{R_kE}]^+, \end{aligned} \tag{9}$$

respectively, where $[y]^+ = \max(0, y)$. We assume the target data rate is η, with which the achievable secrecy rates are obtained by letting $C_{SR_k} = C_{R_kD} = \eta$.

B. Problem Formulation

Applying data buffers at the relays improves the data throughput but increases the packet delay. The delay D for a packet is defined as the period between the packet being transmitted from SS and arrived at SD successfully.

We define the binary selection parameter as $V_{k,j}(t)$, where $k = 1, 2, ..., K$ which is the relay index and $j = 0$ or 1 which corresponds to the $S \to R_k$ or $R_k \to D$ link, respectively. Specifically, if $V_{k,0}(t) = 1$ (or 0), the corresponding $S \to R_k$ link is (or is not) selected for data transmission. Similarly, if $V_{k,1}(t) = 1$ (or 0), the corresponding $R_k \to D$ link is (or is not) selected. Particularly, when the buffer of a relay is empty or full, the corresponding $R_k \to D$ or $S \to R_k$ link becomes unavailable for data transmission, respectively. At any time slot, only one link or no link is selected for data transmission.

[4]In this paper, we investigated the total energy consumption of the the secondary system [22], but the proposed algorithm can be generalized to other power allocation schemes.

In this paper, we consider two relay selection cases: to maximize the data throughput subject to the delay- and secrecy-constraints, and to maximize the secure data throughput subject to the delay constraint, respectively.

Case 1 - To improve the transmission efficiency of the secure cognitive relay network, the joint relay selection and power allocation is to maximize the throughput subject to delay- and secrecy- constraints as:

$$\mathcal{O} = \max_{\mathbf{V},\ \beta} \frac{1}{N} \sum_{t=1}^{N} \sum_{k=1}^{K} V_{k,1}(t), \tag{10}$$

$$\text{s.t. } C_{R_k\_SD}(t) \geq V_{k,1}(t) \cdot \eta, \tag{10a}$$

$$C_{SS\_R_k}(t) \geq V_{k,0}(t) \cdot \eta, \tag{10b}$$

$$C_{s(R_k\_SD)}(t) \geq V_{k,1}(t) \cdot \psi, \tag{10c}$$

$$C_{s(SS\_R_k)}(t) \geq V_{k,0}(t) \cdot \psi, \tag{10d}$$

$$V_{k,1}(t) \cdot D(t) \leq \omega, \tag{10e}$$

$$\beta \in (0, 1], \tag{10f}$$

$$V_{k,j}(t) \in \{0, 1\} \tag{10g}$$

$$\sum_{k=1}^{K} \sum_{j=0}^{1} V_{k,j}(t) \in \{0, 1\} \tag{10h}$$

where $\mathbf{V}(t) = [V_{1,0}(t), V_{1,1}(t), ..., V_{K,0}(t), V_{K,1}(t)]$ which is the relay selection vector at time slot t, N is the number of total time slots, η and ψ are the target data rate and target secrecy rate respectively, ω is the target delay, (10a) and (10b) ensure that the selected link satisfies data transmission requirement, (10c) and (10d) ensure that the selected link satisfies data secure transmission requirement, (10e) ensures that only the arrived packets satisfying the delay constraint are included in the throughput, (10f) defines the range of the power allocation ratio, (10g) states that the selection parameter is binary, and (10h) ensures that either only one link or no link is selected at any time slot.

Case 2 - The joint relay selection and power allocation is to maximize the secrecy rate with the delay-constraint as:

$$\mathcal{U} = \max_{\mathbf{V},\ \beta} \frac{1}{N} \sum_{t=1}^{N} \sum_{k=1}^{K} \Bigg(V_{k,1}(t) \cdot \min\Big\{ C_{s(SR_k)}(t - D(t)), C_{s(R_k D)}(t) \Big\} \Bigg), \tag{11}$$

$$\text{s.t. } C_{R_k\_SD}(t) \geq V_{k,1}(t) \cdot \eta, \tag{11a}$$

$$C_{SS\_R_k}(t) \geq V_{k,0}(t) \cdot \eta, \tag{11b}$$

$$V_{k,1}(t) \cdot D(t) \leq \omega, \tag{11c}$$

$$\beta \in (0, 1], \tag{11d}$$

$$V_{k,j}(t) \in \{0, 1\} \tag{11e}$$

$$\sum_{k=1}^{K} \sum_{j=0}^{1} V_{k,j}(t) \in \{0, 1\} \tag{11f}$$

where the constraints are similarly defined as those in (10).

The joint relay selection and power allocation in both Case 1 and 2 are complicated. By applying the buffers at the relays, the problems in Case 1 and 2 can be regarded as a Markov decision process (MDP) which is affected by the fading channels, buffer-states, interactions with the primary networks. The constraint on instantaneous delay further complicates the problem. Moreover, the eavesdropper can intercept the data in both hop 1 and 2, which does not occur consecutively. All these make the optimization in Case 1 and 2 very complicated, if not impossible, to be obtained. The machine learning is thus used to handle the cases.

III. Markov Decision Process

In secure buffer-aided cognitive relay networks, the system can select a link for data transmission at the current state, then and randomly moves into the next system state. Therefore, to reduce the complexity of the joint relay selection and power allocation in secure buffer-aided cognitive relay networks, we model the optimization problems in *Case 1* and *Case 2* as MDPs which include a set of actions, a set of states and rewards. In a Q-Learning based algorithm, there is an agent who can take actions to change the state of the system and learn to maximize the total reward to find a good solution. In the proposed scheme, the agent selects an action by using ϵ-greedy strategy to get a balance between the exploration and exploitation. In the exploration mode of the Q-learning, the agent selects a link randomly so that the algorithm can update the Q-value for all possible actions. On the other hand, in the exploitation mode, the agent chooses an action from its learning experience.

A. Case 1: Maximizing the Throughput

In *Case 1*, the goal is to maximize the throughput with constraints in (10) so that the definition of tuples in this MDP problem should aim to reach the maximum throughput with constraints.

1) Action: In the proposed system, an action is not only to select a link for data transmission but also to decide the transmit power allocation for the transmitter and the receiver. Therefore, we define the action as $a(t) = a_{k,V,\beta}$ where $k \in \{1, 2, ..., K\}$, $V \in \{0, 1\}$ and $\beta \in (0, 1]$. To be specific, at time slot t the agent takes action $a(t) = a_{k,V,\beta}$ which is selecting the $SS \longrightarrow R_k$ link for data transmission when $V = 0$, with the transmit power βP_b for the transmitter and $(1 - \beta)P_b$ for the receiver. Otherwise, the action is selecting $R_k \longrightarrow SD$ link when $V = 1$. Because the action space should be discrete in a MDP problem, we assume a discrete power allocation in the proposed scheme as in [34], and the number of the power levels is ℓ, then we obtain

$$\beta \in \left\{ \frac{1}{\ell}, \frac{2}{\ell}, ..., 1 \right\}. \tag{12}$$

At any time slot, a system with K relays can have $2\ell K$ possible actions, and we consider there is a possible action that no link is selected, thus the total number of actions is $2\ell K + 1$.

2) State: The system state is the combination of the buffer state and channel state. It is clear that a $SS \longrightarrow R_k$ link can not be used for data transmission when the buffer state $l_k(t) = L$. Moreover, a $R_k \longrightarrow SD$ link can not be used for data transmission when $l_k(t) = 0$. The channel state $z_k(t)$ for the corresponding relay R_k denotes the availability of two links for data transmission and secure data transmission. We define $\mu_{m\_n}(t) = 1$ indicates $C_{m\_n}(t) \geq \eta$ and $C_{s(m\_n)}(t) \geq \psi$, otherwise, $\mu_{m\_n}(t) = 0$. Then, we form the channel state $z_k^I(t)$ in this case as

$$z_k^I(t) = \begin{cases} 1, & \mu_{SS\_R_k}(t) = 0, \mu_{R_k\_SD}(t) = 0 \\ 2, & \mu_{SS\_R_k}(t) = 1, \mu_{R_k\_SD}(t) = 0 \\ 3, & \mu_{SS\_R_k}(t) = 0, \mu_{R_k\_SD}(t) = 1 \\ 4, & \mu_{SS\_R_k}(t) = 1, \mu_{R_k\_SD}(t) = 1. \end{cases} \tag{13}$$

Notice that we define the channel state $z_k^I(t)$ with $\beta = 1$, and a learning-based algorithm should learn a function to map this state and the optimum value of β at a given time slot. However, in practice the range of exploration is wide and it may affect the convergence of the algorithm. We discuss the solution to this problem in Section V-B.

Then, we build the system state which includes buffer state $l_k(t)$ and channel state $z_k^I(t)$ as

$$s(t) = \{l_1(t), l_2(t), ..., l_K(t), z_1^I(t), z_2^I(t), ..., z_K^I(t)\},\ k \in \{1, ..., K\}. \tag{14}$$

In a cognitive relay network with K relays, the total number of system states is $(4(L+1))^K$.

3) Reward: The reward is used to help the Q-Learning algorithm maximize the throughput with delay and security constraints in *Case 1*. We consider giving a positive bonus if there is a packet arriving at SD satisfied to delay-and security-constrained. Thus, the algorithm can learn to get more throughput with the constraints. However, we also consider that sometimes the algorithm may make wrong decisions due to its large range of exploration. To solve this problem, we design the negative reward for the proposed scheme to avoid invalid actions, such as selecting unavailable links or allocating small value of β which leads to a failed data transmission. Therefore, the combination of the positive reward and the negative reward can help the learning algorithm reduce the range of exploration and converge faster.

B. Case 2: Maximizing the Secrecy Rate

In *Case 2*, the goal is to maximize the secrecy rate with constraints in (11) so that the definition of tuples in this MDP problem should aim to reach the maximum secrecy rate with constraints.

1) Action: In *Case 2*, the actions are the same as in *Case 1*.

2) State: In *Case 2*, the buffer state $b(t)$ of the system is the same as in *Case 1*. However, the channel state only needs to show the availability of two links for data transmission. Thus,

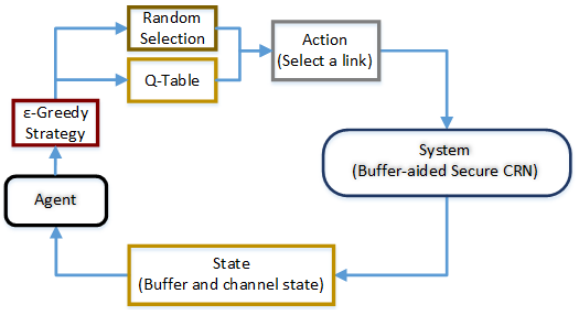


Fig. 2. The structure of the Q-Learning in the buffer-aided cognitive relay networks.

we form the channel state $z_k^{II}(t)$ for *Case 2* as

$$z_k^{II}(t) = \begin{cases} 1, & C_{SS\_R_k}(t) < \eta \ \&\ C_{R_k\_SD}(t) < \eta, \\ 2, & C_{SS\_R_k}(t) \geq \eta \ \&\ C_{R_k\_SD}(t) < \eta, \\ 3, & C_{SS\_R_k}(t) < \eta \ \&\ C_{R_k\_SD}(t) \geq \eta, \\ 4, & C_{SS\_R_k}(t) \geq \eta \ \&\ C_{R_k\_SD}(t) \geq \eta. \end{cases} \tag{15}$$

3) Reward: In *Case 2*, the goal is to maximize the secrecy rate with delay constrained so that the positive reward is designed to encourage the agent selecting high-security links. However, in the exploration mode, the agent may choose a low-security link for data transmission. Therefore, we consider the negative reward for the agent when low-security links are selected.

Therefore, we have defined basic tuples of the environment for a DDQN-based algorithm in both two cases. In order to solve the optimization problem in (10) and (11), we will introduce two DDQN-based algorithms in the following sections.

IV. DDQN-Based Algorithm Without Priori Information

Firstly, we consider the conventional learning-based scheme without prior information. To break the barriers of traditional optimization schemes, the Q-Learning algorithm which is shown in Fig. 2 was proposed. Moreover, to avoid the overestimation problem, the double Q-Learning algorithm was proposed to improve the performance of the Q-Learning. There is an agent in the double Q-learning to make decisions for the system at each time slot. Moreover, the agent can explore the system based on the ϵ-greedy strategy and then store its experience in Q-tables which can help the agent make decisions. When in the exploitation mode, the agent takes the action $a_n(t)$ which is estimated from the Q-table. On the other hand, when in the exploration mode, the agent takes an action randomly. In the double Q-learning, two Q-tables are designed to avoid the overestimation problem, and at each time slot only one Q-table is randomly selected to update. We assume the Q-value $Q^A(s(t), a(t))$ denotes the value for state $s(t)$ and action $a(t)$ at time slot t in Q-table A. Then the function of updating Q-

values in Q-table A at time slot t is given by

$$Q^A(s(t),a(t)) = Q^A(s(t),a(t)) + \delta(r_{(s(t),a(t))} + \tau \cdot Q^B(s(t+1), \text{argmax}_a\{Q^A(s(t+1),a)\}) - Q^A(s(t),a(t))), \tag{16}$$

where $r_{s(t),a(t)}$ is the reward for the state $s(t)$ and action $a(t)$, $\delta \in (0,1)$ denotes the learning rate, $\tau \in (0,1)$ is the discount rate, and $\text{argmax}_a\{Q^A(s(t+1),a)\}$ denotes the next state $at+1$ with the maximum Q-value in Q-table A. To solve the overestimation problem, Q-table A decides the next action with the maximum Q-value, but the evaluation value $Q^B(s(t+1), \text{argmax}_a\{Q^A(s(t+1),a)\})$ is from Q-table B. Then the function of updating Q-values in Q-table B at time slot t is given by

$$Q^B(s(t),a(t)) = Q^B(s(t),a(t)) + \delta(r_{(s(t),a(t))} + \tau \cdot Q^A(s(t+1), \text{argmax}_a\{Q^B(s(t+1),a)\}) - Q^B(s(t),a(t))). \tag{17}$$

However, in a K relays buffer-aided cognitive relay network, the Q-Table is a $(4(L+1))^K \times (2\ell K+1)$ matrix. With such a high-dimensional state-action space, it is difficult to build Q-tables for a double Q-Learning algorithm. Therefore, we introduce the deep neural network in the double Q-Learning as a function approximator to map the actions and states. Furthermore, we introduce Adam [35] as the iterative optimization algorithm for the proposed DDQN-RP scheme.

In the DDQN-RP scheme, the agent can generate a sample for each time slot, and then save the sample to its memory. We define the sample as

$$\{s(t), a(t), r_{(s(t),a(t))}, s(t+1)\}. \tag{18}$$

Then after every M time slots, the agent selects W training samples randomly from the memory and sends them to the deep neural networks for training [36], [37]. Notice that the agent is designed to select W samples rather than select the whole memory to prevent the overfitting problem which is a modelling error. We design two neural networks for DDQN-RP, which are the prediction network and the target network, respectively. The prediction network can estimate the Q-value $Q^P(s(t),a(t))$ for $s(t)$ and $a(t)$ from the sample, while the target network outputs the Q-value $Q^T(s(t+1), \text{argmax}_a Q^P(s(t+1),a))$ which is based on the next action from the prediction network. Therefore, we can calculate the loss between the outputs from the two networks, and then update the prediction network. With considering the reward and the discount factor, the loss function is given by

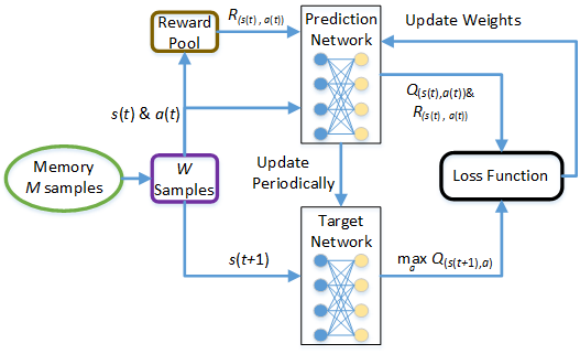


Fig. 3. The structure of training neural networks in DDQN based scheme.

Algorithm 1 DDQN-RP:

```
Initialize the variables.
Repeat:
for v = 1, ..., V do
    for t = 1, ..., M do
        Predict a_n(t) from the prediction network for the exploitation mode.
        Use the ε-greedy strategy to decide the exploration/exploitation mode, and then select a_n(t) or a random action as the current action a(t).
        Get the reward r_(s(t),a(t)) and the next state s(t+1).
        Generate a sample {s(t), a(t), r_(s(t),a(t)), s(t+1)}.
    end for
    for i = 1, ..., W do
        Get Q^P(s(t),a(t))^i from the prediction network based on s(t) and a(t).
        Get Q^T(s(t+1), argmax_a Q^P(s(t+1),a))^i from the target network based on s(t+1).
    end for
    Use the loss function (19) and iterative optimization method Adam to update the prediction network.
end for
Update the target network.
```

$$\varrho = \sum_{t=1}^{W} \Big(r_{(s(t),a(t))} + \tau \cdot Q^T\Big(s(t+1), \text{argmax}_a Q^P(s(t+1),a)\Big) - Q^P\Big(s(t),a(t)\Big)\Big)^2. \tag{19}$$

After updating the prediction network, the agent clears its memory and then continues to generate new samples. After updating the prediction network for V times, the agent updates the target network by copying the parameters from the prediction network. The structure of the training for the two neural networks is shown in Fig. 3. Notice that the DDQN-RP scheme can work in both *Case 1* and *Case 2*. It shows that the

DDQN-based scheme can learn the solution with different goals, due to its low complexity for implementation after building the algorithm. The specific implementation process of DDQN-RP scheme is shown in **Algorithm 1**.

V. DDQN-Based Algorithm with the Priori Information

With introducing the power allocation as well as increasing buffer length and relay number, the range of exploration in secure buffer-aided cognitive relay networks is quite large for the traditional Q-Learning algorithm. To reduce the range of exploration, the priori information can be used to help Q-Learning algorithms converge faster and improve the performance. Therefore, we need to re-define the tuples in MDP to introduce the priori information for the double Q-Learning.

A. Action

Although the total number of the actions is still $2\ell K+1$, not all actions are available at a given time slot due to the constraints of the target data rate, the target secrecy rate and the buffer state in secure buffer-aided cognitive relay networks. Therefore, for *Case 1*, it is clear shown that an action (selecting a link between nodes m and n) cannot work at time slot t unless it satisfies the requirements as follows:

$$\begin{aligned} &C_{m\_n}(t) \geq \eta,\ C_{s(m\_n)}(t) \geq \psi, \\ &l_k(t) < L \text{ for } SS \longrightarrow R_k,\ k \in \{1,2,...,K\}, \\ &l_k(t) > 0 \text{ for } R_k \longrightarrow SD,\ k \in \{1,2,...,K\}. \end{aligned} \tag{20}$$

In an action, the power level β has an impact on the data rate $C_{m\_n}(t)$ and the secrecy rate $C_{s(m\_n)}(t)$ at time slot t. Therefore, we propose the priori information-based learning to delete invalid actions which cannot satisfy (20). Then we embed it in the proposed scheme to improve the performance of the proposed scheme, which will be shown in Section VI.

In *Case 2*, the action set is the same as in *Case 1*, but the priori information is different based on (10) and (11). Therefore, in *Case 2*, an action (selecting a link between nodes m and n) can not work at time slot t unless it satisfies:

$$\begin{aligned} &C_{m\_n}(t) \geq \eta, \\ &l_k(t) < L\ for\ SS \longrightarrow R_k,\ k \in \{1,2,...,K\}, \\ &l_k(t) > 0\ for\ R_k \longrightarrow SD,\ k \in \{1,2,...,K\}. \end{aligned} \tag{21}$$

With considering the effect of β for the data rate, we delete invalid actions which cannot satisfy (21) in the priori information-based learning algorithm and embed it in the proposed scheme.

Furthermore, we consider that the buffer state can also be a priori information when it is not empty and not full. When the value of target delay is not sufficiently large (e.g. close to the buffer length), a trade-off between buffer states that are kept away from empty or full as much as possible can improve the efficiency of overall transmission. Moreover, the average channel gains also have an impact on the delay, a strong link can always have valid actions for transmission, but a weak link only has few valid actions for transmission. Therefore, we introduce the target buffer length of ξ_k for relay R_k as the priori information to help the algorithm converge faster. ξ_k is given by

$$\xi_k = \min\left(\frac{\omega \log_2(1+d_{R_k\_SD}^{-\alpha})}{\sum_{i=1}^{K}\log_2(1+d_{R_i\_SD}^{-\alpha})}, L\right). \tag{22}$$

If the buffer state $l_k(t) \geq \xi_k$ at time slot t, the actions for selecting $SS \longrightarrow R_k$ links are assumed to be all invalid.

B. State

By considering the power allocation in the proposed scheme, the channel state is decided by the channel coefficients and the power allocation. For each action $a(t) = a_{k,V,\beta}$ with different transmit power βP_t at the transmitter, the transmission rate may be different at time slot t. Therefore, we should consider all valid actions which can satisfy (20). The channel state $z_k(t)$ with the prior information for both two cases is given by

$$z_k(t) = \begin{cases} 1, & \text{no link can be selected;} \\ 2, & \text{only exist valid actions for } \mathrm{R_k} \longrightarrow \mathrm{SD;} \\ 3, & \text{only exist valid actions for } \mathrm{SS} \longrightarrow \mathrm{R_k;} \\ 4, & \text{exist valid actions for both two links.} \end{cases} \tag{23}$$

Furthermore, there is a requirement of the target secrecy rate in *Case 1* and the goal of *Case 2* is to maximize the average secrecy rate. Thus, optimizing the secrecy rate is a key point in both two cases. When the successful data transmission is guaranteed, we can easily know that a strong jamming signal has an impact on the untrust node, which will enhance the secrecy rate of the legitimate nodes [38]. Therefore, in *Case 1* the optimal scheme aims to find the smallest β which can guarantee the data transmission requirement $C_{m\_n}(t) \geq \eta$ at time slot t. This power allocation idea can guarantee the data transmission first, and then maximize the secrecy rate to make sure that the transmission is secure. To be specific, in *Case 1* the data transmission has been guaranteed first to maximize the throughput, and then the secrecy rate is maximized to try to meet the requirement of the target secrecy rate. On the other hand, to maximize the average secrecy rate in *Case 2*, the outage can be acceptable when the maximum secrecy rate is sufficiently low. However, the requirement of delay leads to the result that the smallest β which can guarantee the data transmission is still important to get a trade-off between the throughput and the maximum secrecy rate. Thus, we can use this prior information to consider the channel state for each action. When the valid actions are obtained at a given time slot, we can reduce the range of state space by introducing the minimum value of β for each link. To be specific, we can build the system state with the prior information as

$$\begin{aligned} s(t) =&\{l_1(t), l_2(t), ..., l_K(t), z_1(t), z_2(t), ..., z_K(t), \\ &\bar{\beta}_{1,V}(t), \bar{\beta}_{2,V}(t), ..., \bar{\beta}_{K,V}(t)\},\ k \in \{1,...,K\}, \end{aligned} \tag{24}$$

where $\bar{\beta}_{k,V}(t)$ denotes the minimum value of β which can guarantee $C_{m\_n}(t) \geq \eta$ for the corresponding relay R_k at time slot

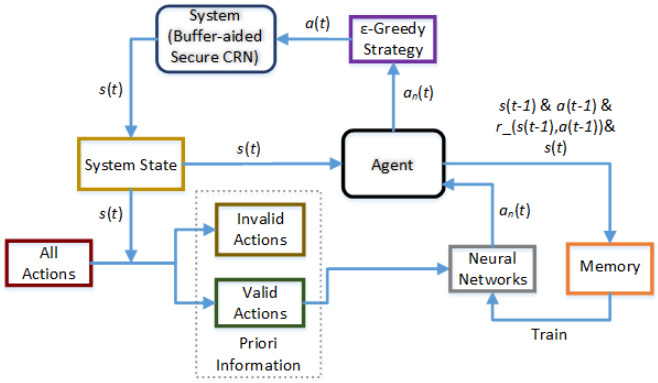


Fig. 4. The structure of DDQN-based scheme with the priori information in buffer-aided cognitive relay networks.

t, $C_{m\_n}(t) = C_{SS\_R_k}(t)$ when $V = 0$, $C_{m\_n}(t) = C_{R_k\_SD}(t)$ when $V = 1$.

C. Reward

The penalty is designed to avoid invalid actions in *Case 1* and *Case 2*. Therefore, we remove the invalid actions in Section V-A for each time slot, and then the double Q-Learning algorithm can reduce the possible selections and converge faster in both two cases. Moreover, in *Case 1* selecting none of the links is the action which we always try to avoid when any other valid action exists. Thus, we can define a strong penalty for the action that no link can be selected when other valid actions exist. On the other hand, no link can be selected may avoid low secrecy rate in *Case 2*, the reward of this action depends on the maximum secrecy rate at a given time slot.

D. Learning Algorithm with the Priori Information

In general Q-Learning algorithms, the reward is used to help the agent reach the target and avoid bad selections. However, in buffer-aided secure cognitive relay networks, the agent can know which actions are invalid from (20), and we can design a method to help the agent avoid selecting invalid actions. Therefore, before using the ε-greedy strategy to decide the action for the current state at a given time slot, we can remove the invalid actions from the action set to reduce the complexity of selection. To be specific, when the neural networks predict the Q-values of all possible actions for the current state $s(t)$, we remove the output Q-values for invalid actions at time slot t for state $s(t)$.

Moreover, the state set with the priori information can help reduce the range of exploration. The structure of DDQNPI-RP scheme is shown in Fig. 4, and the specific implementation process of DDQNPI-RP scheme is shown in **Algorithm 2**. The computational complexity of the proposed algorithm with/without the priori information is $V(M + W)$ as the number of iterations of loops in [39], [40], exploring the priori information does not introduce extra complexity.

Algorithm 2 DDQNPI-RP:

Initialize the variables.
Repeat:
for $v = 1, \cdots, V$ **do**
  for $t = 1, \cdots, M$ **do**
    Based on the current state $s(t)$, remove all invalid actions from the action set at time slot t.
    Predict $a_n(t)$ among the action set from the prediction network for the exploitation mode.
    Use the ε-greedy strategy to decide the exploration/exploitation mode, and then select $a_n(t)$ or a random action as the current action $a(t)$.
    Get the reward $r_{(s(t),a(t))}$ and the next state $s(t+1)$.
    Generate a sample $\{s(t), a(t), r_{(s(t),a(t))}, s(t+1)\}$.
  end for
  for $i = 1, \cdots, W$ **do**
    Get $Q^P(s(t), a(t))^i$ from the prediction network based on $s(t)$ and $a(t)$.
    Get $Q^T(s(t+1), \text{argmax}_a Q^P(s(t+1), a))^i$ from the target network based on $s(t+1)$.
  end for
  Use the loss function (19) and iterative optimization method Adam to update the prediction network.
end for
Update the target network.

VI. Simulation Results

The results of the proposed DDQN-RP scheme and DDQNPI-RP scheme are shown in this section, and the Max-Ratio buffer-aided relay selection algorithm is selected as the benchmark. Unless otherwise stated, we set the parameters for the system as follows: the number of relays $K = 3$, the buffer size $L = 10$, the pre-defined level $I_{th} = 10$, the maximum total transmit power to noise ratio for each time slot $P_b/\sigma_n^2 = 40$ dB, the transmit power at PS is assumed to be unity without losing generality, the self-interference factor $\rho = 0.5$, the number of power levels $\ell = 10$, the path loss exponent $\alpha = 3$, the target rate $\eta = 2$ bps/Hz, the target secrecy rate $\psi = 0.5$ bps/Hz, the target delay $\omega = 12$ time slots. The locations of all nodes are shown in **Table I**.

In this paper, we use the deep learning library Keras/TensorFlow to build the deep neural networks which consists of three fully-connected layers with 64 neurons [41]. The computer with the GPU NVIDIA GeForce GTX-2080 is used to run the simulations. In the neural network, the ϵ decay factor for the ϵ-greedy strategy is 0.999, the minimum ϵ decay value is 0.1, the learning rate is 0.9, the discount factor $\tau = 0.9$, the memory size $M = 500$, the sample size $W = 32$, the iteration number of updating the target network $V = 100$.

Fig. 5 shows the throughput learning curves with and without the a-priori information in *Case 1*. It shows that the throughput of the DDQNPI-RP scheme converges to 0.41 packet/time slot

TABLE I: Positions of Nodes

SS	SD	PS	PD
(0, 0) m	(0, 10) m	(5.5, 2.4) m	(5.4, 2.6) m
R_1	R_2	R_3	E
(0.1, 5.0) m	(-1.1, 4.6) m	(0.7, 5.2) m	(11.6, 5.1) m

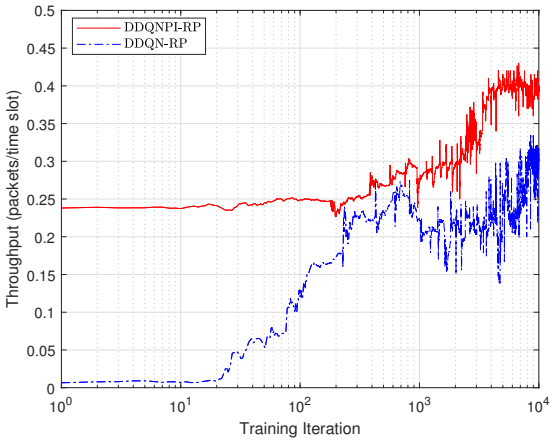


Fig. 5. Throughput with delay and security constrained vs. training iterations in *Case 1*.

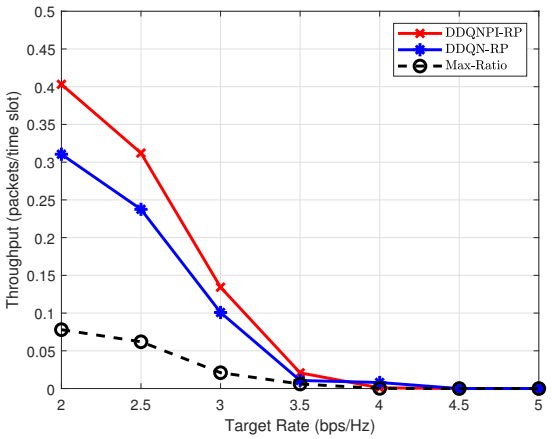


Fig. 6. The comparison of throughput between proposed schemes and Max-Ratio scheme for *Case 1* with different target rate.

after 4,000 training iterations, while that of the DDQN-RP scheme only achieves about 0.31 after 7,000 iterations. This verifies that the DDQNPI-RP uses the priori information to avoid these invalid actions, leading to faster convergence and higher throughput than its DDQN-RP counterpart. Moreover, because the DDQNPI-RP always avoids the invalid actions, even at the beginning of the training that the neural networks are randomly initialized, the DDQNPI-RP has better performance than the DDQN-RP. This is clearly shown in Fig. 5.

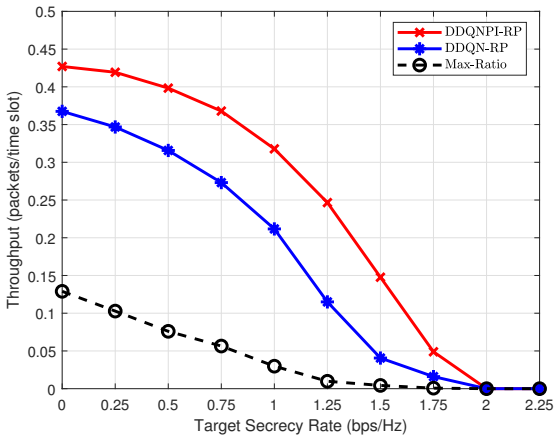


Fig. 7. The comparison of throughput between proposed schemes and Max-Ratio scheme for *Case 1* with different secrecy target rate.

Fig. 6 shows the comparison of throughput between the proposed schemes and Max-Ratio scheme for *Case 1*. It shows that both proposed schemes outperform Max-Ratio significantly, the DDQNPI-RP and DDQN-RP get the throughput of 0.32 and 0.24 with the target rate $\eta = 3$ bps/Hz, respectively, while Max-Ratio only achieves 0.06. This clearly indicates that the Max-Ratio has not considered throughput with delay and security constrained. However, the proposed two schemes can solve this problem because the agent can learn the solution by using the ε-greedy strategy to satisfy the constraints.

Fig. 7 shows the throughput with delay-and security-constrained vs. secrecy target rate for the proposed two schemes and Max-Ratio scheme in *Case 1*. We can observe that the proposed algorithm performs dramatically better than Max-Ratio. The DDQNPI-RP and DDQN-RP achieve 0.41 and 0.31 when the target secrecy rate $\psi = 0.5$ bps/Hz, respectively, while Max-Ratio obtains 0.07. Though Max-Ratio improves Max-Link algorithm by considering the security, many selections can lead to a huge delay by using Max-Ratio because the delay constraint is not considered in Max-Ratio.

Fig. 8 shows the throughput with delay and security constrained vs. target delay for the proposed two schemes and Max-Ratio scheme in *Case 1*, where DDQN-RP(HD) is the DDQN-based scheme for the HD system. It is clearly shown that with a target delay, the DDQNPI-RP scheme achieves around 0.39 when the target delay $\omega = 10$ time slots, which outperforms Max-Ratio dramatically. Both two learning schemes can learn to meet different requirements of target delay, and thus the agent can learn different policies for each case. On the other hand, it is clear that the Max-Ratio is not designed for the data transmissions with delay constraints. Compared with the HD system, we can observe that the hybrid-duplex system can help improve the performance, due to the impact of FD jamming. Notice that all schemes can only achieve throughput when $\omega \geq 2$, because a packet arriving at SD takes at least two

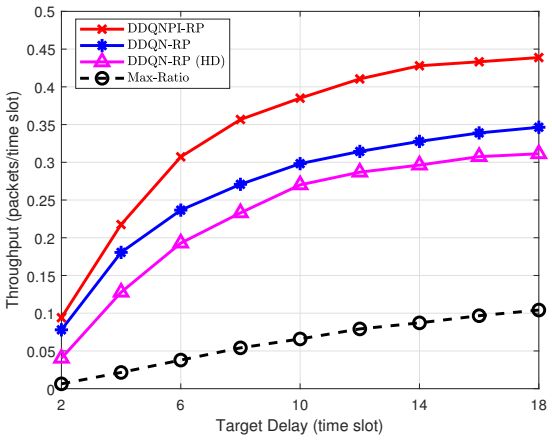


Fig. 8. The comparison of throughput between proposed schemes and Max-Ratio scheme for *Case 1* with different target delay.

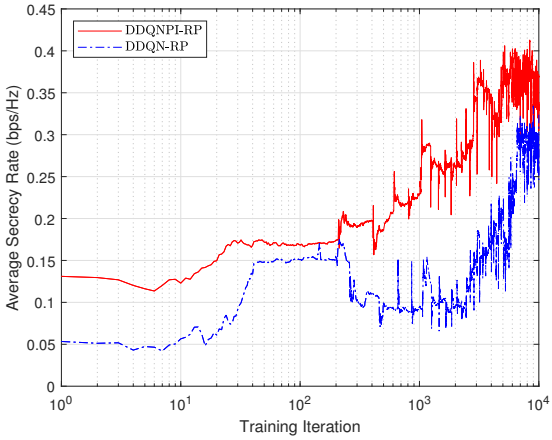


Fig. 9. Average secrecy rate with delay constrained vs. training iterations in *Case 2*.

time slots.

Fig. 9 shows the secrecy rate learning curves of the proposed scheme with and without a priori information in *Case 2*. It shows that the DDQNPI-RP scheme converges to around 0.38 bps/Hz slot after 5,000 training iterations, while DDQN-RP scheme achieves about 0.3 after 7,000 iterations. It is clear that DDQNPI-RP scheme can achieve higher and more stable secrecy rate than the DDQN-RP, which verifies the effectiveness of applying the a-priori information.

Fig. 10 shows the average secrecy rate with the constrained delay vs. target delay for the proposed two schemes and Max-Ratio scheme in *Case 2*. We can observe that the DDQNPI-RP scheme can achieve average secrecy rate of 0.35 when the target delay $\omega = 10$ time slots, while DDQN-RP and Max-Ratio only achieve around 0.29 and 0.08, respectively. The exploration mode leads to many local optimum problems and it makes the performance of DDQN-RP unstable. Therefore, with

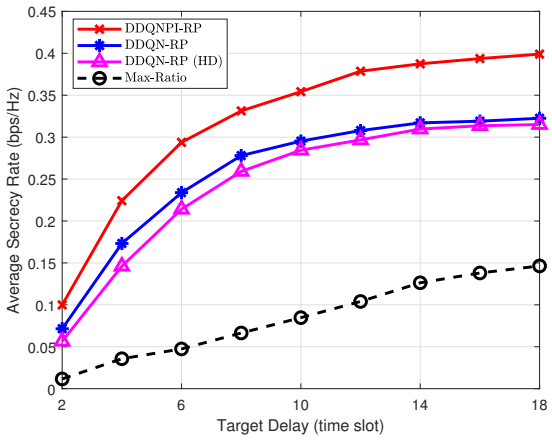


Fig. 10. Average secrecy rate with delay constrained vs. target delay in *Case 2*.

the priori information, DDQN-based scheme can try to avoid local optimums and achieve better performance. Moreover, the hybrid-duplex transmission can switch between the HD and FD modes, and achieve better performance than HD transmission.

VII. Conclusion

In this paper, we applied the DDQN for the joint relay selection and power allocation in the delay and/or secrecy constrained buffer-aided cognitive relay networks. Two cases have been considered, namely maximizing the throughput and the average secrecy rate, respectively. We introduced the a-priori information to improve the convergence. Simulations show that proposed schemes outperform the max-ratio scheme in both two cases. Finally, we note that the proposed scheme can be generalized to more complicated system such as the multi-antenna and/or multi-user cases [42]. While this would further complicate the learning process, the principle of the proposed scheme remains the same. This would be left as an interesting future topic.

References

[1] V. W. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key technologies for 5G wireless systems*. Cambridge, U.K.: Cambridge university press, 2017.

[2] Y. Zou, Y. Yao, and B. Zheng, "Cooperative relay techniques for cognitive radio systems: Spectrum sensing and secondary user transmissions," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 98–103, Apr. 2012.

[3] J. Lee, H. Wang, J. G. Andrews, and D. Hong, "Outage probability of cognitive relay networks with interference constraints," *IEEE Transactions on Wireless Communications*, vol. 10, no. 2, pp. 390–395, Feb. 2011.

[4] J. Bang, J. Lee, S. Kim, and D. Hong, "An efficient relay selection strategy for random cognitive relay networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1555–1566, Mar. 2015.

[5] N. Zlatanov, A. Ikhlef, T. Islam, and R. Schober, "Buffer-aided cooperative communications: opportunities and challenges," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 146–153, Apr. 2014.

[6] I. Krikidis, T. Charalambous, and J. S. Thompson, "Buffer-aided relay selection for cooperative diversity systems without delay constraints," *IEEE Transactions on Wireless Communications*, vol. 11, no. 5, pp. 1957–1967, May. 2012.

[7] Z. Tian, G. Chen, Y. Gong, Z. Chen, and J. A. Chambers, "Buffer-aided max-link relay selection in amplify-and-forward cooperative networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 2, pp. 553–565, May. 2014.

[8] G. Chen, Z. Tian, Y. Gong, and J. Chambers, "Decode-and-forward buffer-aided relay selection in cognitive relay networks," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 9, pp. 4723–4728, Nov. 2014.

[9] R. Zhang, R. Nakai, K. Sezaki, and S. Sugiura, "Generalized buffer-state-based relay selection in cooperative cognitive radio networks," *IEEE Access*, vol. 8, pp. 11644–11657, Jan. 2020.

[10] Z. Tian, Y. Gong, G. Chen, and J. A. Chambers, "Buffer-aided relay selection with reduced packet delay in cooperative networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2567–2575, Mar. 2017.

[11] P. Xu, Z. Ding, I. Krikidis, and X. Dai, "Achieving optimal diversity gain in buffer-aided relay networks with small buffer size," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8788–8794, Oct. 2016.

[12] N. Nomikos, D. Poulimeneas, T. Charalambous, I. Krikidis, D. Vouyioukas, and M. Johansson, "Delay- and diversity-aware buffer-aided relay selection policies in cooperative networks," *IEEE Access*, vol. 6, pp. 73531–73547, Nov. 2018.

[13] G. Chen, Z. Tian, Y. Gong, Z. Chen, and J. A. Chambers, "Max-ratio relay selection in secure buffer-aided cooperative wireless networks," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 4, pp. 719–729, Apr. 2014.

[14] A. Sun, T. Liang, and Y. Zhang, "Performance analysis of secure buffer-aided cognitive radio network," in *IEEE/CIC International Conference on Communications in China (ICCC), Shenzhen, China*, Nov. 2015.

[15] C. Wei, W. Yang, Y. Cai, X. Tang, and T. Yin, "Secrecy outage performance of buffer-aided underlay cognitive relay networks with outdated csi," in *IEEE/CIC International Conference on Communications in China (ICCC), Beijing, China*, Aug. 2018.

[16] C. Wei, W. Yang, Y. Cai, X. Tang, and N. Pu, "Secrecy outage analysis for DF buffer-aided multi-antenna underlay crns," in *10th International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, China*, Oct. 2018.

[17] A. Yadav, O. A. Dobre, and H. V. Poor, "Is self-interference in full-duplex communications a foe or a friend?," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 951–955, Jul. 2018.

[18] G. Chen, Y. Gong, P. Xiao, and J. A. Chambers, "Physical layer network security in the full-duplex relay system," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 574–583, Mar. 2015.

[19] J. Lee, "Full-duplex relay for enhancing physical layer security in multi-hop relaying systems," *IEEE Communications Letters*, vol. 19, no. 4, pp. 525–528, Apr. 2015.

[20] X. Huang, J. He, Q. Li, Q. Zhang, and J. Qin, "Optimal power allocation for multicarrier secure communications in full-duplex decode-and-forward relay networks," *IEEE Communications Letters*, vol. 18, no. 12, pp. 2169–2172, Dec. 2014.

[21] L. Elsaid, L. Jiménez-Rodríguez, N. H. Tran, S. Shetty, and S. Sastry, "Secrecy rates and optimal power allocation for full-duplex decode-and-forward relay wire-tap channels," *IEEE Access*, vol. 5, pp. 10469–10477, Jun. 2017.

[22] A. Kuhestani, A. Mohammadi, and M. Mohammadi, "Joint relay selection and power allocation in large-scale mimo systems with untrusted relays and passive eavesdroppers," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 2, pp. 341–355, Feb. 2018.

[23] Z. Xiong, Y. Zhang, D. Niyato, R. Deng, P. Wang, and L. Wang, "Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 44–52, Jun. 2019.

[24] 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, Oct. 2019.

[25] Y. Zou, Y. Xie, C. Zhang, S. Gong, D. T. Hoang, and D. Niyato, "Optimization-driven hierarchical deep reinforcement learning for hybrid relaying communications," in *IEEE Wireless Communications and Networking Conference (WCNC), Virtual Conference*, May. 2020.

[26] X. Du, H. Van Nguyen, C. Jiang, Y. Li, F. R. Yu, and Z. Han, "Virtual relay selection in lte-v: A deep reinforcement learning approach to heterogeneous data," *IEEE Access*, vol. 8, pp. 102477–102492, May. 2020.

[27] L. Zhu, Y. He, F. R. Yu, B. Ning, T. Tang, and N. Zhao, "Communication-based train control system performance optimization using deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 10705–10717, Dec. 2017.

[28] L. Xiao, Y. Li, G. Han, H. Dai, and H. V. Poor, "A secure mobile crowdsensing game with deep reinforcement learning," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 1, pp. 35–47, Jan. 2018.

[29] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *AAAI conference on artificial intelligence*, Phoenix, AZ, Mar. 2016.

[30] C. Cai, Y. Cai, W. Yang, and W. Yang, "Secure connectivity using randomize-and-forward strategy in cooperative wireless networks," *IEEE Communications Letters*, vol. 17, no. 7, pp. 1340–1343, 2013.

[31] D. S. Michalopoulos and G. K. Karagiannidis, "Performance analysis of single relay selection in rayleigh fading," *IEEE Transactions on Wireless Communications*, vol. 7, no. 10, pp. 3718–3724, Oct. 2008.

[32] Z. Hadzi-Velkov, D. S. Michalopoulos, G. K. Karagiannidis, and R. Schober, "On the effect of outdated channel estimation in variable gain relaying: Error performance and PAPR," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1084–1097, Mar. 2013.

[33] G. Zheng, I. Krikidis, J. Li, A. P. Petropulu, and B. Ottersten, "Improving physical layer secrecy using full-duplex jamming receivers," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 4962–4974, Oct. 2013.

[34] W. Zhong, G. Chen, S. Jin, and K. Wong, "Relay selection and discrete power control for cognitive relay networks via potential game," *IEEE Transactions on Signal Processing*, vol. 62, no. 20, pp. 5411–5424, Oct. 2014.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, Dec. 2014.

[36] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[37] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning, New York, NY*, Jan. 2016.

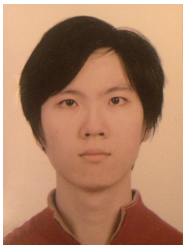
[38] G. Chen, J. P. Coon, and M. Di Renzo, "Secrecy outage analysis for downlink transmissions in the presence of randomly located eavesdroppers," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1195–1206, May. 2017.

[39] F. Hussain, R. Hussain, A. Anpalagan, and A. Benslimane, "A new block-based reinforcement learning approach for distributed resource allocation in clustered IoT networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 2891–2904, Mar. 2020.

[40] I. Budhiraja, N. Kumar, and S. Tyagi, "Deep reinforcement learning based proportional fair scheduling control scheme for underlay D2D communication," *IEEE Internet of Things Journal*, pp. 1–1, (Early Access) 2020.

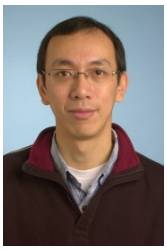
[41] Q. Wang, W. Zhang, Y. Liu, and Y. Liu, "Multi-UAV dynamic wireless networking with deep reinforcement learning," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2243–2246, Dec. 2019.

[42] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5g," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.



Chong Huang (Graduate Student Member, IEEE) received the B.Eng. degree in communication engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2011, and the M.Sc. degree in electrical and electronic engineering from Loughborough University, Loughborough, U.K., in 2015. He is currently pursuing a Ph.D. degree in signal processing with University of Leicester, Leicester, U.K.. His research interests include deep learning, reinforcement learning, physical layer security, cooperative communications, cognitive radio, non-orthogonal multiple access (NOMA), and Internet of Things (IoT).

Gaojie Chen (S'09 – M'12 – SM'18) received the B.Eng. and B.Ec. Degrees in electrical information engineering and international economics and trade from Northwest University, China, in 2006, and the M.Sc. (Hons.) and Ph.D. degrees in electrical and electronic engineering from Loughborough University, Loughborough, U.K., in 2008 and 2012, respectively. From 2008 to 2009, he was a Software Engineer with DT Mobile, Beijing, China. From 2012 to 2013, he was a Research Associate with the School of Electronic, Electrical and Systems Engineering, Loughborough University. He was a Research Fellow with 5GIC, Faculty of Engineering and Physical Sciences, University of Surrey, U.K., from 2014 to 2015. He was also a Research Associate with the Department of Engineering Science, University of Oxford, U.K., from 2015 to 2018. He is currently a Lecturer with the School of Engineering, University of Leicester, U.K. His current research interests include information theory, wireless communications, cooperative communications, cognitive radio, the Internet of Things, secrecy communications, and random geometric networks. He received the Exemplary Reviewer Certificates of the IEEE WIRELESS COMMUNICATIONS LETTERS in 2018 and the IEEE TRANSACTIONS ON COMMUNICATIONS in 2019. He serves as an Associate Editor for the IEEE COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS - MACHINE LEARNING IN COMMUNICATIONS AND NETWORKS and *Electronics Letters* (IET).



Yu Gong has been with School of Electronic, Electrical and Systems Engineering, Loughborough University, UK, since 2012. Dr Gong obtained his BEng and MEng in electronic engineering in 1992 and 1995 respectively, both at the University of Electronics and Science Technology of China. In 2002, he received his PhD in communications from the National University of Singapore. After PhD graduation, he took several research positions in Institute of Inforcomm Research in Singapore and Queen's University of Belfast in the UK respectively. From 2006 and 2012, Dr Gong had been a lecturer in the School of Systems Engineering, University of Reading, UK. His research interests are in the area of signal processing and communications including wireless communications, cooperative networks, non-linear and non-stationary system identification and adaptive filters.

Zhu Han (S'01–M'04-SM'09-F'14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015-2018, AAAS fellow since 2019 and ACM distinguished Member since 2019. Dr. Han is 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."