Multi-Agent Reinforcement Learning Based Buffer-Aided Relay Selection in IRS-Assisted Secure Cooperative Networks

Chong Huang, *Graduate Student Member, IEEE*, Gaojie Chen, *Senior Member, IEEE*, and Kai-Kit Wong, *Fellow, IEEE*

***Abstract*—This paper proposes a multi-agent deep reinforcement learning-based buffer-aided relay selection scheme for an intelligent reflecting surface (IRS)-assisted secure cooperative network in the presence of an eavesdropper. We consider a practical phase model where both phase shift and reflection amplitude are discrete variables to vary the reflection coefficients of the IRS. Furthermore, we introduce the buffer-aided relay to enhance the secrecy performance, but the use of the buffer leads to the cost of delay. Thus, we aim to maximize either the average secrecy rate with a delay constraint or the throughput with both delay and secrecy constraints, by jointly optimizing the buffer-aided relay selection and the IRS reflection coefficients. To obtain the solution of these two optimization problems, we divide each of the problems into two sub-tasks and then develop a distributed multi-agent reinforcement learning scheme for the two cooperative sub-tasks, each relay node represents an agent in the distributed learning. We apply the distributed reinforcement learning scheme to optimize the IRS reflection coefficients, and then utilize an agent on the source to learn the optimal relay selection based on the optimal IRS reflection coefficients in each iteration. Simulation results show that the proposed learning-based scheme uses an iterative approach to learn from the environment for approximating an optimal solution via the exploration of multiple agents, which outperforms the benchmark schemes.**

***Index Terms*—Physical layer security, intelligent reflecting surface, multi-agent reinforcement learning, buffer-aided relay selection, throughput**

I. Introduction

WITH the development of the fifth-generation (5G) wireless communication, physical layer security has been widely studied to provide secure wireless communications in recent years [1]. Unlike cryptographic techniques, physical layer security exploits the dynamics of fading channels for achieving the perfect secrecy performance and dose not require encryption keys [2]–[4]. Security is also particularly relevant for cooperative communication networks, which has been investigated in [5]–[7]. In [5], the secrecy rate performance of full-duplex (FD) decode-and-forward (DF) cooperative networks was studied with a self-interference cancellation technology. The authors in [6] proposed two linear precoding schemes to improve the secrecy rate performance in half-duplex (HD) amplify-and-forward (AF) relaying systems. To maximize the diversity gain, relay selection was also proposed in cooperative networks to reduce the secrecy outage probability in [7]. To further enhance the secrecy performance, a novel max-ratio buffer-aided relay selection was proposed to select the link with the largest signal-to-interference-ratio (SIR) in cooperative networks with buffering technology [8]. Then, the trade-off between the average delay and secrecy rate for the max-ratio scheme was investigated in a buffer-aided cooperative network [9]. In [10], the max-ratio and state-based schemes were amalgamated to reduce the secrecy outage probability and average delay for buffer-aided cooperative networks. Furthermore, in [11], the average secrecy rate in an energy-harvesting based buffer-aided cooperative network was enhanced by an adaptive transmission algorithm considering power constraints, buffer and energy storage. Although using buffer improves outage performance, it increases the instantaneous delay, which is a key issue in Internet of Things (IoT) networks [12]. A buffer-state-based probabilistic relay selection method was proposed to enhance the outage performance with delay constraint in [13]. In [14], the delay constrained throughput was investigated via deep reinforcement learning (DRL). However, the physical layer security has not been considered in delay-constraint buffer-aided cooperative networks. This motivates us to study security communication systems to satisfy instantaneous delay constraints.

On the other hand, intelligent reflecting surface (IRS) is an emerging technique for beyond 5G wireless communications [15]–[17]. IRS is an array which consists of low-cost passive reflecting elements, each of which can be appropriately re-configured to control its reflection coefficient independently to provide controllable phase shift and amplitude for reflecting or refracting the signals to the intended receiver. Therefore, the IRS-assisted secure networks have attracted much attention recently [18]–[20]. In [18], a joint design of the transmit covariance and the IRS phase shifts was proposed to maximize the secrecy rate via an alternating optimization-based algorithm. A joint beamforming vectors and IRS phase shifts optimization scheme based on the combination of the alternating optimization and semidefinite relaxation (SDR) methods were proposed to enhance the secrecy rate in [19]. In [20], a successive convex approximation based algorithm was proposed to improve the secrecy rate by considering the trajectory and power control of an unmanned-aerial-vehicle (UAV), and the phase shifts of the

This work was supported by EPSRC grant number EP/R006377/1 (“M3NETs”).

C. Huang and G. Chen are with School of Engineering, University of Leicester, Leicester LE11 7RH, U.K., Email: {ch481, gaojie.chen}@leicester.ac.uk.

K. Wong is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K., Email: kai-kit.wong@ucl.ac.uk.

IRS. The effect of main parameters on secrecy performance in IRS-assisted wireless networks was investigated in [21]. To improved the average secrecy rate, a alternating optimization method was proposed to jointly optimize the UAV's trajectory, beamforming matrix and transmit power for a IRS-assisted UAV communication network in [22]. However, the above works focused on point-to-point communications. Therefore, to amalgamate the benefits of both IRS and cooperative communications, the authors in [23] and [24] investigated the IRS phase shifts optimization to maximize the achievable rate for hybrid IRS with HD and FD relay networks. Two hybrid relay and IRS-assisted transmission scheme were proposed in [25] to achieve higher error performance and achievable rate, compared with IRS-only and relay-only transmissions. In [26], to further enhance the transmission quality, relay selection was investigated in IRS-assisted cooperative networks. Motivated by this, this paper proposes an IRS-assisted secure buffer-aided relay network. Moreover, we consider the practical phase shift model [27], where both the reflection amplitude and the phase shift vary with the reflection coefficient. Therefore, the design of the IRS reflection coefficient vectors and buffer-aided relay selection will be given to maximize the secrecy rate and throughput with secure and delay constraints in this paper.

The traditional optimization algorithms in most IRS-assisted networks require huge amount of computational resource with low-efficiency. Deep reinforcement learning was therefore introduced to solve the complicated optimization problems [28]. In [29], a joint optimization of transmit beamforming vectors and IRS continuous phase shifts was proposed via DRL to enhance the average sum rate. Furthermore, the DRL algorithm was applied to jointly optimize the beamforming vectors and IRS continuous phase shifts to enhance the secrecy rate for IRS-assisted networks in [30]. However, the phase-dependent amplitude variation has not been considered. Moreover, joint optimization of the buffer-aided relay selection and IRS reflection coefficients is a much more complicated high-dimensional problem for the proposed system. To solve this, we introduce the multi-agent DRL (MA-DRL) algorithm as in [31] to solve two related sub-problems, namely, buffer-aided relay selection and IRS reflection coefficients optimization. Considering the limitations of storage, computation ability, delay and energy for a single device, we further apply distributed DRL in [32] on multiple relay nodes to reduce the training cost and improve the convergence efficiency for IRS reflection coefficients optimization. The main contributions of this paper are summarized as follows:

- We propose a joint buffer-aided relay selection and IRS reflection coefficient optimization scheme in IRS-assisted secure cooperative networks. Two optimization problems are considered: maximizing the average secrecy rate with the delay constraint and maximizing the throughput with the delay and secrecy constraints.
- We introduce the MA-DRL method to solve the complicated optimization problem. In terms of optimizing the IRS reflection coefficients, an asynchronous distributed framework is proposed to apply multi-agent on relay nodes to train the global model without sharing the local data. On the other hand, the agent at the source for solving the buffer-aided relay selection problem learns its strategy by using the IRS reflection coefficients optimization solution from the distributed framework.
- Simulation results show that the proposed MA-DRL algorithm can apply multi-agents to explore the environment to learn approximate solutions. Based on the rewards, the agents can build a global model to solve the two related sub-problems to generate the joint optimization strategy and achieve higher performance than the benchmark.

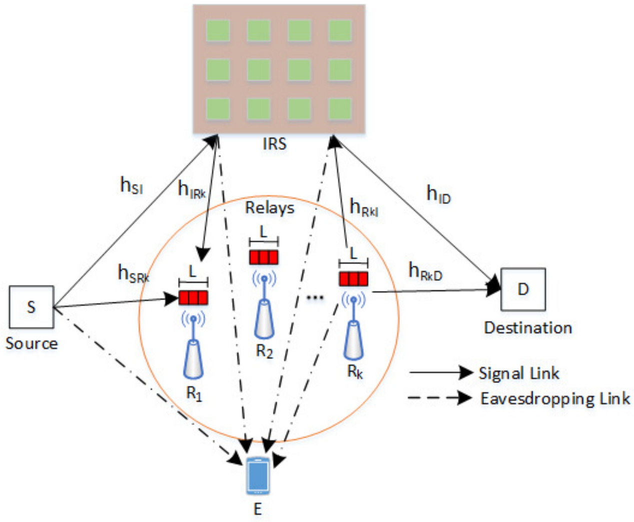

Fig. 1. System model of the secure hybrid buffer-aided relay and IRS network.

The rest of this paper is organized as follows: The system model and problem formulation are introduced in Section II. In Section III, MA-DRL scheme is proposed. Section IV provides the simulation results and verifies the performance of the proposed scheme. Finally, Section V concludes this paper.

II. System Model and Problem Formulation

A. System Model

In this paper, we focus on a two-hop IRS-assisted secure relay network which is consisted of one source S, one IRS I, one destination D, which is equipped with N reflecting elements, and K half-duplex (HD) randomize-and-forward[1] relays R_k $(k \in \{1, 2, ..., K\})$, each of which is equipped with a data buffer of finite size L. The system model is shown in Fig. 1. Based on the buffer-aided relaying technique, a $S \rightarrow R_k$ transmission is considered available when the buffer of R_k is not full, and a $R_k \rightarrow D$ transmission is considered available when the buffer of R_k is not empty. Furthermore, there is

[1] Randomize-and-forward strategy applies different codebooks for two hops, each hop can ensure a secure transmission independently. Thus, the end-to-end transmission is secure when both two hops support secure transmissions [33], [34].

an untrusted node[2] E which can eavesdrop the signal sent from the S and R. All nodes are assumed to be equipped with a single antenna. The IRS can control the reflection coefficients to change the phase shift and reflection amplitude for each IRS reflecting element independently. We assume that S is the controller of the network[3] to determine the relay selection and IRS reflection coefficients, and can receive the required information from all relay nodes, while each relay node R_k can receive the channel state information (CSI) of channels for the transmissions to and from itself[4], and receives the related CSI of the eavesdropper[5] from S. Furthermore, we assume no direct link between nodes S and D due to severe blocking or deep fading. The channels of $S \rightarrow R_k$ and $R_k \rightarrow D$ links are assumed to experience Non-Line-of-Sight (NLoS) Rayleigh fading, and the channels from and to I are assumed to experience Rician fading [45]. Therefore, the channel coefficient h_{ij} between nodes i and j is given by

$$h_{ij} = \begin{cases} \bar{h}_{ij}, & \text{Rayleigh} \\ \sqrt{\frac{\kappa_{ij}}{\kappa_{ij}+1}}\hat{H}_{ij} + \sqrt{\frac{1}{\kappa_{ij}+1}}\hat{h}_{ij}, & \text{Rician,} \end{cases} \tag{1}$$

where κ_{ij} is the Rician factor for h_{ij}. We assume $\bar{h}_{ij} = \bar{g}_{ij} d_{ij}^{-\bar{\alpha}/2}$ in NLoS Rayleigh channels, where $ij \in \{SR_k, SE, R_kD, R_kE\}$, $\bar{g}_{ij}$ is modeled by a complex Gaussian fading with zero-mean and unit-variance, d_{ij} is the distance between nodes i and j, $\bar{\alpha}$ is the path loss exponent of NLoS Rayleigh fading links. On the other hand, we assume $\hat{H}_{ij} = \hat{\mathrm{g}}_{ij} d_{ij}^{-\hat{\alpha}_L/2}$ and $\hat{h}_{ij} = \hat{g}_{ij} d_{ij}^{-\hat{\alpha}_N/2}$ are the Line-of-Sight (LoS) component and the NLoS component in Rician fading channels, respectively, where $ij \in \{SI, IR_k, IE, R_kI, ID\}$, $\hat{\alpha}_L$ and $\hat{\alpha}_N$ are the path loss exponents for the LoS and NLoS Rician fading channel, respectively, $\hat{g}_{ij}$ is modelled by a complex Gaussian fading with zero-mean and unit-variance, and $\hat{\mathrm{g}}_{ij}$ is given by

$$\hat{\mathrm{g}}_{ij} = \sqrt{\beta}[1, e^{-j\pi \sin \partial_{ij}}, ..., e^{-j\pi(M-1)\sin \partial_{ij}}]^T, \tag{2}$$

where β denotes the signal attenuation for the reference distance $d_0 = 1$ m [46], $\partial_{ij} \in [0, 2\pi]$ denotes the angle of arrival (AoA) or the angle of departure (AoD) of the corresponding signal.

At a given time slot, when a $S \rightarrow R_k$ link is selected, S transmits the signal x_S to both R_k and I. Then, I reflects the signal to R_k. The received signal at R_k is thus given by

$$y_{SR_k} = \sqrt{P}\left(h_{SR_k} + \boldsymbol{h}_{IR_k}^H \boldsymbol{\Theta} \boldsymbol{h}_{SI}\right) x_S + n_{R_k}, \tag{3}$$

where P is the transmit power at all nodes, n_{R_k} denotes the additive-white-Gaussian-noise (AWGN) with variance σ_n^2 at R_k, $\boldsymbol{\Theta} = \mathrm{diag}(\eta_1 e^{j\theta_1}, \eta_2 e^{j\theta_2}, ..., \eta_N e^{j\theta_N})$ denotes the diagonal reflection matrix for I, with $\theta_n \in [0, 2\pi)$ and $\eta_m \in [0, 1]$ denoting the phase shift and reflection amplitude of the nth IRS reflecting element, respectively. Without loss of generality, we denote $\boldsymbol{v} = [v_1, v_2, ..., v_N]$ as the reflection coefficient vector for I, $\eta_n = |v_n|$ and $\theta_n = \arg(v_n)$ for the nth IRS reflecting element. We adopt the practical IRS phase shift model to achieve the discrete reflection amplitudes and phase shifts based on the reflection coefficients from [27, Fig. 3(b)], where the effective resistance R is 2 Ω. Moreover, in order to support the practical implementation [47], we assume discrete phase shifts for IRS reflecting elements. The range of discrete phase shifts is

$$\chi \triangleq \left\{0, \frac{2\pi}{\lambda}, ..., \frac{(\lambda-1)2\pi}{\lambda}\right\}, \tag{4}$$

where λ is the number of quantization levels.

Then, based on (3), the channel capacity for a $S \rightarrow R_k$ link is given by

$$C_{SR_k} = \log_2\left(1 + \frac{P\left|h_{SR_k} + \boldsymbol{h}_{IR_k}^H \boldsymbol{\Theta} \boldsymbol{h}_{SI}\right|^2}{\sigma_n^2}\right). \tag{5}$$

Notice that E can also receive signals from S during the $S \rightarrow R_k$ transmission. Therefore, the received signal at E can be expressed as

$$y_{SE} = \sqrt{P}\left(h_{SE} + \boldsymbol{h}_{IE}^H \boldsymbol{\Theta} \boldsymbol{h}_{SI}\right) x_S + n_E, \tag{6}$$

where n_E denotes the AWGN with variance σ_n^2 at E. The channel capacity for the $S \rightarrow E$ link is given by

$$C_{SE} = \log_2\left(1 + \frac{P\left|h_{SE} + \boldsymbol{h}_{IE}^H \boldsymbol{\Theta} \boldsymbol{h}_{SI}\right|^2}{\sigma_n^2}\right). \tag{7}$$

Similarly, when a $R_k \rightarrow D$ link is selected, the received signals at D and E can be expressed as

$$\begin{aligned} y_{R_kD} &= \sqrt{P}\left(h_{R_kD} + \boldsymbol{h}_{ID}^H \boldsymbol{\Theta} \boldsymbol{h}_{R_kI}\right) x_{R_k} + n_D, \\ y_{R_kE} &= \sqrt{P}\left(h_{R_kE} + \boldsymbol{h}_{IE}^H \boldsymbol{\Theta} \boldsymbol{h}_{R_kI}\right) x_{R_k} + n_E, \end{aligned} \tag{8}$$

respectively, where x_{R_k} denotes the signal from relay node R_k. Therefore, the channel capacities for the $R_k \rightarrow D$ and $R_k \rightarrow E$ links can be expressed as

$$\begin{aligned} C_{R_kD} &= \log_2\left(1 + \frac{P\left|h_{R_kD} + \boldsymbol{h}_{ID}^H \boldsymbol{\Theta} \boldsymbol{h}_{R_kI}\right|^2}{\sigma_n^2}\right), \\ C_{R_kE} &= \log_2\left(1 + \frac{P\left|h_{R_kE} + \boldsymbol{h}_{IE}^H \boldsymbol{\Theta} \boldsymbol{h}_{R_kI}\right|^2}{\sigma_n^2}\right), \end{aligned} \tag{9}$$

respectively. Then, the secrecy rates for the $S \longrightarrow R_k$ and

[2] In this paper, we only consider a single untrust relay as an eavesdropper in the proposed network to give a clear exploration and results. Considering that the effect of multiple eavesdroppers, including colluding and non-colluding cases, has been studied for other scenarios in our previous works, we will consider multiple eavesdroppers scenario more complex challenge to solve in our future work.

[3] The source S can be a Macro/Micro base station or an access point in cellular networks. Generally, S has sufficient resource to receive related channel information by using a backhaul link and control the transmissions in the wireless network. In many optimization works, a centralized node (e.g. a base station or an access point) is required to receive global information and make decisions [35]–[39].

[4] Obtaining the CSI is a key issue in IRS-assisted communications, and there are some existing works about the channel estimation for IRS-assisted networks [40], [41]. However, the channel estimation details are out of the scope of this work.

[5] The untrusted node may have malicious behaviour, e.g. eavesdropping; however, it belongs to the proposed networks so that the channel estimation is available between other nodes and itself [42]–[44].

$R_k \longrightarrow D$ links can be expressed as

$$\begin{aligned} C_{s(SR_k)} &= [C_{SR_k} - C_{SE}]^+, \\ C_{s(R_kD)} &= [C_{R_kD} - C_{R_kE}]^+, \end{aligned} \tag{10}$$

respectively, where $[x]^+ = \max(0, x)$.

B. Problem Formulation

We assume the link between nodes i and j is available for single-packet transmission when its capacity satisfies $C_{(ij)} \geq \omega$, where ω denotes the target rate. Moreover, a secure transmission between i and j happens when $C_{s(ij)} \geq \varsigma$, where ς denotes the target secrecy rate. Because the use of buffer leads to delay of data transmissions, we consider the delay constraint in this paper. The delay Δ of a given packet is defined as the time of transmitting this packet from S to D. To be specific, if a packet is transmitted successfully from S to R_k at time slot t, and then arriving at D at time slot $t+1$ successfully, it takes two time slots to arrive at D and its delay is $\Delta = 2$. Furthermore, the use of the buffer-aided relay selection, therefore, at a given time slot, selecting a relay R_k to transmit one packet from its buffer to the destination, or receive one packet from the source and save it in its buffer. Notice that only one link can be selected for transmission at a given time slot. Simultaneously, the IRS reflection coefficient vector $\boldsymbol{v}$ is used to alter the reflected signal to boost the signal quality collaboratively. We use $F = 0$ to specify that $S \rightarrow R_k$ transmission is selected and $F = 1$ to denote that $R_k \rightarrow D$ transmission is chosen. We aim to design an algorithm to obtain solutions to two different optimization problems in IRS-assisted secure cooperative networks. Thus, we have the following two cases:

Case 1 - To find out the optimal k, F and $\boldsymbol{v}$ to maximize the achievable average secrecy rate[6] with delay constraint for the end-to-end transmission, the optimization problem can be formulated as

$$\begin{aligned} O^{(1)} = \max_{k(t),F(t),\boldsymbol{v}(t)} \frac{1}{T}\sum_{t=1}^{T} \bigg(& F(t)\mu(\Delta(t) \leq \Delta_T) \\ & \times \mu\big(C_{R_kD}(t) \geq \omega\big)\mu\big(C_{SR_k}(t-\Delta(t)) \geq \omega\big) \\ & \times \min\Big\{C_{s(SR_k)}(t-\Delta(t)), C_{s(R_kD)}(t)\Big\}\bigg), \end{aligned} \tag{11}$$

$$\begin{aligned} \text{s.t. } & k(t) \in \{1, 2, ..., K\}, && \text{(11a)} \\ & F(t) \in \{0, 1\}, && \text{(11b)} \\ & \boldsymbol{v}(t) = [v_1(t), v_2(t), ..., v_N(t)], && \text{(11c)} \\ & \theta_n(t) = \arg(v_n(t)) \in \chi, \forall m, && \text{(11d)} \\ & \eta_n(t) = |v_n(t)|, && \text{(11e)} \\ & l_k(t) > 0, \text{ when } F(t) = 1, && \text{(11f)} \\ & l_k(t) < L, \text{ when } F(t) = 0, && \text{(11g)} \end{aligned}$$

where T denotes the number of total time slots, $\mu(\cdot) = 1$ if the input event is true and $\mu(.) = 0$ otherwise, and $\Delta(t)$ denotes the delay of the corresponding packet which is transmitted at time slot t, and Δ_T denotes the target delay, and l_k denotes the buffer state of relay R_k. (11f) and (11g) ensure that the buffer should not be empty for $R_k \rightarrow D$ transmissions and not full for $S \rightarrow R_k$ transmissions constraints.

Case 2 - To find out the optimal k, F and $\boldsymbol{v}$ to achieve the maximum throughput with delay and secrecy constraint, the optimization problem can be formulated as

$$\begin{aligned} O^{(2)} = \max_{k(t),F(t),\boldsymbol{v}(t)} \frac{1}{T}\sum_{t=1}^{T} \bigg(& F(t)\mu(\Delta(t) \leq \Delta_T) \\ & \times \mu\big(C_{ij}(t) \geq \omega\big)\mu\big(C_{s(ij)}(t) \geq \varsigma\big)\bigg), \end{aligned} \tag{12}$$

$$\begin{aligned} \text{s.t. } & k(t) \in \{1, 2, ..., K\}, && \text{(12a)} \\ & F(t) \in \{0, 1\}, && \text{(12b)} \\ & \boldsymbol{v}(t) = [v_1(t), v_2(t), ..., v_N(t)], && \text{(12c)} \\ & \theta_n(t) = \arg(v_n(t)) \in \chi, \forall m, && \text{(12d)} \\ & \eta_n(t) = |v_n(t)|, && \text{(12e)} \\ & l_k(t) > 0, \text{ when } F(t) = 1, && \text{(12f)} \\ & l_k(t) < L, \text{ when } F(t) = 0, && \text{(12g)} \end{aligned}$$

where the constraints are similar to that in (11). Due to the discrete integer decision variables in the function and discrete set in each constraint, the functions and constraints are non-convex in (11) and (12). Thus, they are complicated non-convex optimization problems and impossible to be solved by an exhaustive search method in large-scale networks. Besides, many traditional algorithms require high computational complexity to solve the IRS phase shift optimization problem with fixed reflection amplitude [28], and traditional methods for IRS-assisted wireless communications only consider the continuous ideal IRS model. Furthermore, considering the buffering technology provides more choices for relay selection at each time slot, maximizing the throughput with secrecy and delay constraints at D in T time slots is a challenge for traditional optimization algorithms in buffer-aided cooperative networks. Therefore, we introduce the MA-DRL algorithm to obtain the feasible k, F and $\boldsymbol{v}$ to solve these two optimization problems. The DRL algorithm can be used to optimize IRS coefficients with discrete practice phase shift model, and buffer-aided relay selection with secrecy and delay constraints. Moreover, multi-agent framework is applied to reduce the training cost for each single node and communication cost for the proposed network, compared with the centralized DRL training.

III. MA-DRL-Based Algorithm

Considering both the optimization problems in (11) and (12) which contain variables k and F for buffer-aided relay selection, and $\boldsymbol{v}$ for adjusting IRS reflection coefficients, we split each of the optimization problems into two sub-tasks: 1) IRS reflection coefficient optimization, and 2) the buffer-aided relay selection, to reduce the space of exploration for the proposed algorithm. We propose a multi-agent framework for

[6] We assume that the fixed data transmission rate is ω in the proposed buffer-aided system. The achievable secrecy rates are obtained by letting $C_{SR_k} = C_{R_kD} = \omega$ when $C_{SR_k} \geq \omega$ and $C_{R_kD} \geq \omega$.

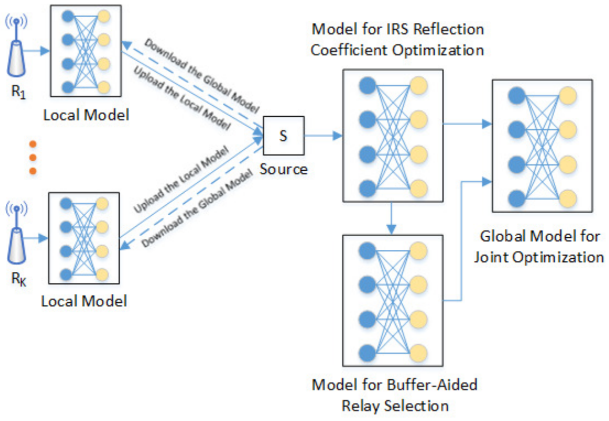
R1
Local Model
RK
Local Model
Download the Global Model
Upload the Local Model
Upload the Local Model
Download the Global Model
S
Source
Model for IRS Reflection Coefficient Optimization
Global Model for Joint Optimization
Model for Buffer-Aided Relay Selection

Fig. 2. The framework of multi-agent DRL in the proposed network.

DRL in secure IRS-assisted buffer-aided cooperative networks, as shown in Fig. 2. The framework of the proposed scheme consists of a controller S and K relay nodes participating in the training process. Considering the large state-action space for IRS reflection coefficient optimization in large-scale IRS systems, we apply the distributed DRL algorithm for IRS reflection coefficient optimization to reduce the training cost of each device for data processing and improve the convergence efficiency. Each relay node represents an agent which learns the solution for IRS reflection coefficient optimization to build its local model (local solution for the IRS reflection coefficient optimization problem), then an agent on S updates the global model (global solution for the IRS reflection coefficient optimization problem) based on the local models from all relay nodes. In each iteration, the proposed approach first learns to address the task of optimizing the IRS reflection coefficient via the asynchronous distributed DRL algorithm, and subsequently train the model for buffer-aided relay selection based on the optimized solution from the former task, finally aggregates the results from the two sub-tasks to generate the joint optimization model (joint solution for the optimization problem in (11) or (12)). Thus, rather than directly exploring the large space of the whole environment, MA-DRL can obtain the optimized K, F and $\boldsymbol{v}$ efficiently for (11) and (12) with low computational complexity.

A. Asynchronous Distributed DRL Algorithm for IRS Reflection Coefficient Optimization

Considering the fact that the complexity of the exhaustive search algorithm is KL^M for searching the optimal IRS reflection coefficients for each iteration, the space of exploration for DRL is huge in large scale networks. Therefore, distributed DRL is introduced to solve this problem. The asynchronous method is considered to improve the convergence efficiency among all nodes participating in the training process. We will first introduce the framework of the distributed DRL and the elements of DRL, then give the details of DRL algorithm.

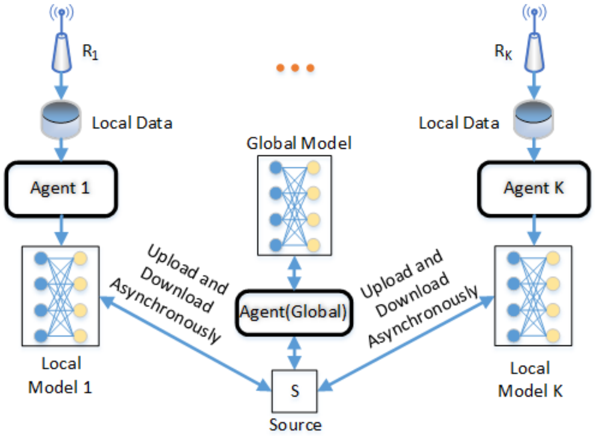

Fig. 3. The framework of the asynchronous distributed DRL network for IRS reflection coefficient optimization.

1. The Distributed Learning Framework and Basic Elements of DRL

As shown in Fig. 3, the process of the asynchronous distributed learning framework has the following steps.

- ***Step 1***: Each of the relay nodes can train and update its own model in parallel. For example, the local agent on relay node R_k trains its local model based on its local data, which contains its buffer state and CSI information. To be specific, relay node R_k explores its environment to train the local deep neural network (DNN) model via DRL. After W_L local iterations of training, R_k updates its local model and obtains the accumulated gradients.
- ***Step 2***: After updating the local model, the agent on R_k uploads its accumulated gradients to the controller S for updating the global model of IRS reflection coefficient optimization. Considering the differentiation between the computational resources of all relay nodes, we apply the asynchronous updating method to improve the efficiency of convergence for the global model. Thus, each local model uploads its accumulated gradients to update the global model on S asynchronously.
- ***Step 3***: After updating the global model on S in Step 2, the corresponding relay node R_k downloads the global model from S to update its local model.

To be specific, there is an agent on each relay node to learn from its own environments with the local data via DRL, then train its local model to find out the local solution of IRS reflection coefficient optimization. In terms of training local models, the DRL algorithm generally consists of states, actions and rewards, which are defined for *Case 1* and *Case 2* as follows.

1) ***Case 1***: To achieve the maximum average secrecy rate with delay constraint.
State: In the IRS-assisted transmissions, we define the state $s(t) = \{h_{ij}(t), h_{Ij}^H(t)h_{iI}(t), h_{iE}(t), h_{IE}^H(t)h_{iI}(t)\}$ at time slot t, to describe the dynamic IRS-assisted network according to (5), (7) and (9). In DRL, the

environment can transit from one state to another possible future state by taking an action, which will be defined as follows.

Action: For a secure IRS-assisted buffer-aided relay network, an action for the task of IRS reflection coefficient optimization is to control $\boldsymbol{v}$ in (11) to reflect or refract the signal to the receiver. Therefore, we define the action $a(t) = [v_1(t), v_2(t), ..., v_N(t)]$ at time slot t to provide controllable phase shifts and reflection amplitudes. Based on the action and state, the DRL algorithm can learn to make decisions for the proposed network with the predefined rewards, which will be introduced to help the DRL algorithm converge.

Reward: The objective of this sub-task is to adjust the IRS reflection coefficients to improve the secure transmission quality with a delay constraint. To be specific, the design of IRS reflection coefficients can improve the secure transmission between nodes i and j with the delay constraint. In other words, the requirements of $C_{(ij)} \geq \omega$ and $\Delta \leq \Delta_T$ are satisfied in *Case 1*. The reward in DRL is designed to help the agent learn the solution of the task, and we give the agent a positive reward when the channel rate of the transmission satisfies the requirements based on the current state-action pair. To be specific, in *Case 1*, when the transmission requirements are satisfied, a packet is transmitted to D within the target delay and the agent can receive a positive reward. Moreover, the agent is designed to receive a negative reward when the corresponding transmission cannot support the requirements. Therefore, the proposed DRL based algorithm can map the relationship between the states and actions using the rewards.

2) ***Case 2***: To achieve the maximum throughput with delay and secrecy constraints.

State: In *Case 2*, the states are designed as the same as in *Case 1*.

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

Reward: In *Case 2*, the goal is to achieve the maximum throughput with delay and secrecy constraints. Therefore, the secure transmission in *Case 2* requires $C_{(ij)} \geq \omega$, $C_{s(ij)} \geq \varsigma$, and $\Delta \leq \Delta_T$. On the other hand, in *Case 2*, when the requirements are satisfied, to encourage the agent giving higher priority to selecting high-security links, the positive reward is designed based on the secrecy rate of the transmitted packet. Moreover, the agent receives a negative reward when the requirements are not satisfied. This design can help to find out the optimal state-action pair for reinforcement learning.

Moreover, unlike traditional reinforcement learning, to combine the advantage of value-based or policy-based algorithms for efficient convergence, we will introduce a distributed asynchronous advantage actor-critic (A3C) algorithm [48] without sharing the global training environment, as the DRL-based solution for the task of IRS reflection coefficient optimization.

2. The Distributed Asynchronous Advantage Actor-Critic Algorithm

To enhance the convergence performance and robustness of training in DRL, we apply the A3C algorithm to train the local models for each agent on relay nodes. In the A3C algorithm, there is an actor network and a critic network to evaluate the relation and advantage of each state-action pair, respectively. For example, at time slot t, the input for both the actor and critic networks is the current state $s(t)$. At the same time, the outputs are the probabilities for each corresponding action-state pair and the evaluation value Q for estimating the advantage of the state, respectively. Therefore, the estimation value V for $s(t)$ is given by

$$\begin{aligned} V(s(t)) = r_{s(t),a(t)} + \rho r_{s(t+1),a(t+1)} + ... \\ + \rho^{T-1} r_{s(t+T-1),a(t+T-1)} + \rho^T Q\big(s(t+T);\theta_c\big), \end{aligned} \tag{13}$$

where ρ is the discount factor of the critic network, $r_{s(t),a(t)}$ denotes the reward for the $\big(s(t), a(t)\big)$ pair, θ_c is the DNN weights set of the critic network, the actions are estimated from the actor network. Then, we can obtain the advantage of the $\big(s(t), a(t)\big)$ pair as

$$A\big(s(t)\big) = V\big(s(t)\big) - Q\big(s(t+N);\theta_c\big), \tag{14}$$

which is used to help the agent learn the advantage (or the disadvantage) of the corresponding state-action pairs from the actor network.

We introduce the DNN to form the actor and critic networks. The actor network θ_a is designed to estimate actions for optimizing a policy π to achieve the maximum throughput with delay and secrecy constraints. In the A3C algorithm, the estimated action for a given state is determined with the maximum probability value from the results of the actor network, where the loss function of the actor network can be expressed as

$$\varphi_A(t) = \log\pi\big(a(t), s(t);\theta_a\big)\big(V\big(s(t)\big) - Q\big(s(t);\theta_c\big)\big). \tag{15}$$

On the other hand, the critic network is used to calculate the loss for the actor network, to evaluate the advantage or the disadvantage of the current policy π in the actor network. The critic network can be trained by the corresponding loss function, which is given by

$$\varphi_C(t) = \Big(V\big(s(t)\big) - Q\big(s(t);\theta_c\big)\Big)^2. \tag{16}$$

To calculate the gradients based on the loss functions in (15) and (16), we apply the RMSProp method [49] in DNN. In A3C, each agent in the proposed distributed DRL algorithm performs the advantage actor-critic (A2C) algorithm asynchronously in its thread, then upload its accumulated gradients to update the global model. To be specific, each relay node has its own agent to perform A2C asynchronously to train its local model and obtain the accumulated gradients as in Step 1. In each local iteration, the local agent generates a training sample as $\{s(t), a(t), r_{s(t),a(t)}\}$ to calculate the gradients. After W_L

local iterations, the relay node will upload its accumulated gradients to update the global model on S as in Step 2, and then download the global model to update its local model as in Step 3. Besides, the output of the global model on S can provide a solution for adjusting the reflection coefficients of IRS elements to meet the requirements of the channel rate and secrecy rate. This result can help the agent on S to solve the other sub-task, which is for buffer-aided relay selection.

B. DRL Algorithm for Buffer-Aided Relay Selection

Now we introduce DRL to select the best buffer-aided relay with delay and secrecy constraints. By considering the long term benefit in buffer-aided relay networks, we apply a DRL agent on the source S to solve the buffer-aided relay selection. The states, actions and rewards in this sub-task are defined as follows.

1) ***Case 1***: To achieve the maximum average secrecy rate with a delay constraint.

State: We introduce A3C algorithm to solve the IRS reflection coefficients problem. The channel capacities for $S \rightarrow R_k$, $S \rightarrow E$ and $R_k \rightarrow D/E$ can be obtained from (5), (7), (9), respectively, and the secrecy rate can be obtained based on (10). We assume that a valid transmission satisfies $C_{(ij)} \geq \omega$ in *Case 1*. Thus, the channel state of relay R_k can be described as $c_k \in \{1, 2, 3, 4\}$, where c_k denotes the availability of $S \rightarrow R_k$ and $R_k \rightarrow D$ transmissions as

- $c_k = 1$: none of the two transmission is valid,
- $c_k = 2$: $S \rightarrow R_k$ is valid,
- $c_k = 3$: $R_k \rightarrow D$ is valid,
- $c_k = 4$: both transmissions are valid.

Therefore, it is easy for each relay to find out its channel states for secure transmission. Furthermore, the buffer state $l_k(t)$ of relay R_k denotes the number of packets in relay node R_k's buffer at time slot t. Thus, we form the state of DRL for buffer-aided relay selection as $s_r(t) = \{c_1(t), c_2(t), ..., c_K(t), l_1(t), l_2(t), ..., l_K(t)\}$, which includes the buffer states and the channel states of the proposed network.

Action: At a given time slot, the buffer-aided relay network needs to select a link for transmission. Considering that there are $2K$ links for transmission in the proposed network, we design the action as $a_r(t) = \{k, F\}$, where $k \in \{1, 2, ..., K\}$, $F = 0$ means to select $S \rightarrow R_k$ transmission and $F = 1$ denotes the fact that $R_k \rightarrow D$ transmission is selected. Therefore, the action is used to determine the variables k and F of the optimization problem in (12).

Reward: This sub-task aims to select the best link in the buffer-aided cooperative network to achieve the maximum average secrecy rate with a delay constraint. To be specific, the optimization of relay selection can help avoid invalid transmissions and address the delay issue. Therefore, we design the positive reward for the proposed scheme similar to that in Section III-A.1. Based on this reward, the DRL algorithm can learn to achieve the maximum delay constrained average secrecy rate for *Case 1*. Furthermore, we assign a negative reward for selecting an invalid transmission, unless all possible transmissions are invalid at a given time slot. This design can encourage

Algorithm 1 MA-DRL:

```
Initialize the variables.
Initialize the actor network $\theta_a^r$ and critic network $\theta_c^r$ for buffer-aided relay selection.
***Train the sub-task for IRS reflection coefficient optimization:***
Initialize the global actor network $\theta_a$ and critic network $\theta_c$.
Initialize the local actor network $\theta_a^{'}$ and critic network $\theta_c^{'}$ for each relay node.
**repeat** for each relay node thread:
    Synchronize the local networks $\theta_a^{'} = \theta_a$ and $\theta_c^{'} = \theta_c$.
    **for** $t = 1, 2, ..., W_L$ **do**
        Use the policy $\pi\big(a(t), s(t); \theta^{'}\big)$ to select the action $a(t)$.
        Obtain the reward $r_{s(t),a(t)}$.
        Save the sample $\{s(t), a(t), r_{s(t),a(t)}\}$.
    **end for**
    $V(s(t)) = \begin{cases} 0, \text{ for final convergence} \\ Q\big(s(t); \theta_c^{'}\big), \text{ otherwise} \end{cases}$
    **for** $t = W_L - 1, W_L - 2, ..., 1$ **do**
        $V(s(t)) = r_{s(t),a(t)} + \rho V(s(t+1))$.
        Get the gradients $\nu(t)$ and $\nu_c(t)$ for $\theta^{'}$ and $\theta_c^{'}$ based on (15) and (16) via RMSProp.
        Accumulate gradients $\nu$ for $\theta^{'}$: $\nu = \nu + \nu(t)$.
        Accumulate gradients $\nu_c$ for $\theta_c^{'}$: $\nu_c = \nu_c + \nu_c(t)$.
    **end for**
    Asynchronous update $\theta_a$ with $\nu$ and $\theta_c$ with $\nu_c$.
    ***Train the sub-task for buffer-aided relay selection:***
    **for** $t = 1, 2, ..., W_L$ **do**
        Use the policy $\pi_r\big(a_r(t), s_r(t); \theta_a^r\big)$ to select the action $a(t)$.
        Obtain the reward $r_{rs_r(t),a_r(t)}$.
        Save the sample $\{s_r(t), a_r(t), r_{rs_r(t),a_r(t)}\}$.
    **end for**
    $V(s_r(t)) = \begin{cases} 0, \text{ for final convergence} \\ Q\big(s_r(t); \theta_c^r\big), \text{ otherwise} \end{cases}$
    **for** $t = W_L - 1, W_L - 2, ..., 1$ **do**
        $V(s_r(t)) = r_r s_r(t), a_r(t) + \rho V(s_r(t+1))$.
        Obtain the gradients $\nu_r(t)$ and $\nu_{r_c}(t)$ for $\theta_a^r$ and $\theta_c^r$ based on (15) and (16) via RMSProp.
        Accumulate gradients $\nu_r$ for $\theta^{'}$: $\nu = \nu + \nu(t)$.
        Accumulate gradients $\nu_{r_c}$ for $\theta_c^{'}$: $\nu_c = \nu_c + \nu_c(t)$.
    **end for**
    Update $\theta_a^r$ with $\nu_r$ and $\theta_c^r$ with $\nu_{r_c}$.
    $\nu_r = 0$ and $\nu_{r_c} = 0$.
    Update the global model for joint optimization.
**until** final convergence.
```

the agent to avoid taking invalid actions and improve the convergence efficiency.

2) ***Case 2***: To achieve the maximum throughput with delay and secrecy constraints.

State: In *Case 2*, the state are designed as the same as in *Case 1*, except that a valid transmission between nodes i and j is assumed to satisfy $C_{(ij)} \geq \omega$ and $C_{s(ij)} \geq \varsigma$.

Action: In *Case 2*, we design the actions as the same as in *Case 1*.

Reward: In *Case 2*, we aim to achieve the maximum throughput with delay and secrecy constraints. Therefore, the reward is designed to encourage the agent selecting links with high secrecy rate. To be specific, we give the positive reward to the agent based on the secrecy rate of the transmitted packet, while the negative reward is used when the requirements of the transmission are not satisfied.

Similar to the task of optimizing IRS reflection coefficients, we apply the A2C algorithm on S to address the problem of buffer-aided relay selection. Then, after training the models for two sub-tasks, we can form the states $s(t)$ and $s_r(t)$ as the input, the variables k, F, and $\boldsymbol{v}$ as the outputs, to train a global model for joint design of buffer-aided relay selection and IRS reflection coefficients. Furthermore, the space of exploration for the DRL algorithm is significantly reduced from $2K \times \lambda^N$ to $2K + \lambda^N$ by splitting each of the optimization problems into the two sub-tasks. Thus, the convergence efficiency of the DRL can be enhanced by the proposed algorithm. The pseudo-code of MA-DRL is summarized in Algorithm 1. Notice that the proposed algorithm is designed for both cases, but the states, actions and rewards are different in the two cases, as mentioned before.

IV. Simulation Results

In this section, we analyze the performance of the proposed scheme via simulations. For comparison, we consider selecting the max-ratio buffer-aided relay selection scheme [8] with random IRS reflection coefficients as the benchmark. Unless otherwise stated, the parameters for the proposed network and algorithm are shown in Table I. Furthermore, the locations of S, I, D, R_1, R_2, R_3, R_4, R_5 and E are (0, 0) m, (2, 24) m, (0, 40) m, (0, 20) m, (-0.88, 18.32) m, (-4.1, 21.92) m, (0.88, 19.52) m, (5.28, 18.8) m and (-28.4, 10.2) m, respectively[7]. Notice that the positions of relay nodes are randomly generated because the DRL algorithm can learn from different environments to achieve corresponding satisfactory solutions. The AOD or AOA ∂_{ij} between nodes i and j is randomly distributed within $[0, 2\pi)$ [23], [45], [50]. We build both the actor network and critic network with 256, 128 and 128 neurons for DRL algorithms in two tasks. We run the simulations on GPU Geforce GTX-2080 with the deep learning library TensorFlow.

TABLE I: Simulation Parameters

Parameter	Value
Number of relays, K	5
Buffer size, L	10
Number of IRS elements, N	32
Quantization bit of IRS, $B = \log_2(\lambda)$	4 bits ($\lambda = 16$)
Transmit power to noise ratio, P/σ_n^2	40 dB
Path loss exponent, $\hat{\alpha}$, $\bar{\alpha}_L$ (LoS)	2
Path loss exponent, $\bar{\alpha}_N$ (NLoS)	2.5
Rician factor, κ	10 dB
Target rate, ω	2 bps/Hz
Target secrecy rate, ς	0.5 bps/Hz
Target delay, Δ_T	12 time slots
Discount factor in MA-DRL, ρ	0.9
Number of local iterations, W_L	500

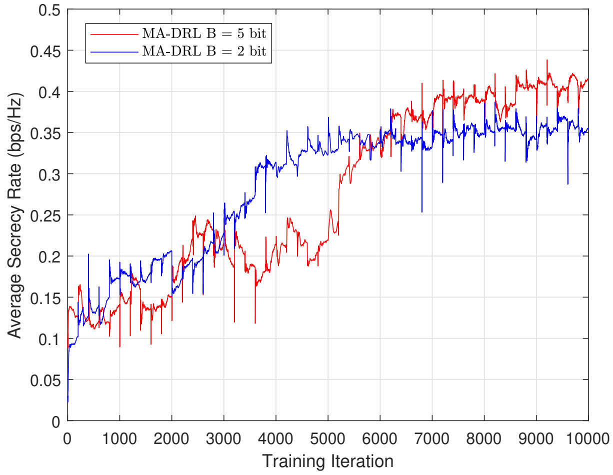

Fig. 4. Average secrecy rate versus training iterations in *Case 1*.

Fig. 4 shows the average secrecy rate with delay constraint versus training iterations for the proposed scheme in *Case 1*. According to the results, it is clear that the average secrecy rate with delay constraint increases with the number of training iterations. The proposed MA-DRL algorithm achieves approximately 0.4 bps/Hz when $B = 5$ bits after 7,000 iterations, while the case with $B = 2$ bits obtains a solution of 0.35 bps/Hz after 5,000 iterations. Thus, the average secrecy rate increases with the quantization bits, resulting from the increase in the quantization levels as B increases. Moreover, more quantization bits lead to slower convergence due to the larger exploration space for reinforcement learning. Due to exploration mode of DRL in training, the curve of the convergence is not very stable. In terms of the implementation for IRS, the quantization bits vary in different scenarios. This result shows that the proposed reinforcement learning-based algorithm can learn from the environment to find the optimal solution in dynamic networks.

The comparison of the average secrecy rate with delay constraint between the proposed scheme and the benchmark

[7] Notice that the proposed DRL algorithm can learn from different environment, i.e., the different locations of the raleys, to achieve corresponding satisfactory solutions. Therefore, to clearly show the related simulation results, we use the one of the snapshots of the relay locations as an example.

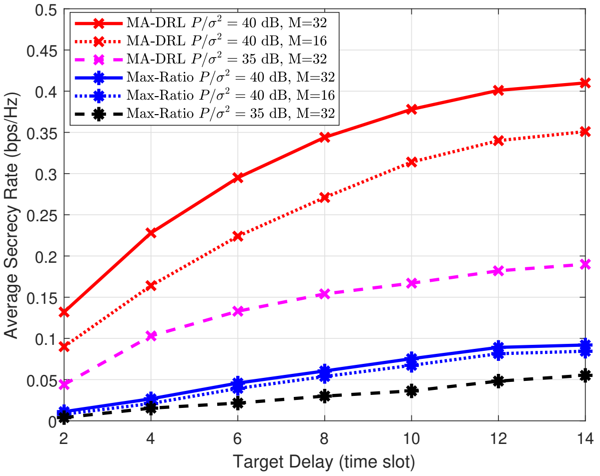

Fig. 5. Average secrecy rate versus target delay in *Case 1*.

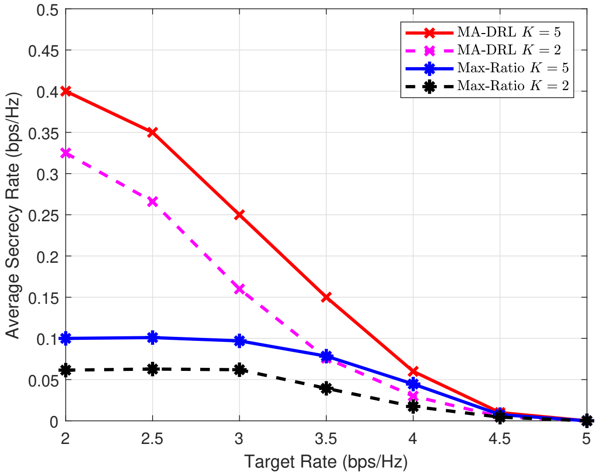

Fig. 6. Average secrecy rate versus target rate in *Case 1*.

in *Case 1* is provided in Fig. 5 . First, we can see that the average secrecy rate increases when the target delay becomes larger. This is because a larger value of target delay leads to decreasing the average delay of packets. Secondly, the results illustrate that the proposed learning-based algorithm can analyze the network to optimize its solution. For example, the proposed MA-DRL algorithm achieves approximately 0.37 bps/Hz when the target delay is $\varsigma = 10$ time slots and the transmit power to noise ratio is $P/\sigma_n^2 = 40$ dB, while the max-ratio scheme only achieves 0.07 bps/Hz. In addition, compared with the case of $P/\sigma_n^2 = 35$ dB, high signal-to-noise ratio (SNR) has more performance gain when the target delay is small. The reason is that if the target delay is large, each packet can be stored in the corresponding buffer for a long period, the relay selection algorithm has more possible choices at a given time slot to reduce the impact of low SNR. Moreover, the proposed MA-DRL achieves more performance gain with a larger value of IRS elements N, while the performance of the benchmark max-ratio with different N is similar. The reason is that the proposed algorithm optimizes the IRS reflection coefficients to improve the signal quality, while max-ratio selects random IRS reflection coefficients.

As we can see in Fig. 6, the average secrecy rate in each result decreases as the target rate increases due to the increasing outage probability. Furthermore, the proposed algorithm performs much better than the benchmark, and the reason is that the MA-DRL algorithm can learn to obtain the solution, which includes the buffer-aided relay selection and IRS reflection coefficients optimization under all constraints, while max-ratio only considers the buffer-aided relay selection in secure transmissions. Results show that the proposed MA-DRL algorithm achieves approximately 0.25 bps/Hz when the target rate $\omega = 3$ bps/Hz and the number of relay $K = 5$, while the max-ratio scheme only achieves 0.1 bps/Hz. Moreover, MA-DRL also performs much better than max-ratio when $K = 2$; this result shows the learning-based algorithm can adjust its variables for different environments to find out the solution.

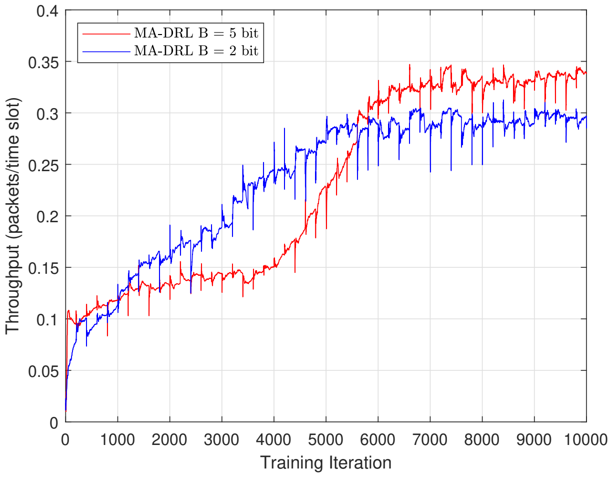

Fig. 7. Throughput versus training iterations in *Case 2*.

The results in Fig. 7 show the investigation of throughput with delay and secrecy constraints under increasing training iterations for the proposed scheme in *Case 2*. As we can see, the MA-DRL converges to a good solution after thousands of training iterations, and we also compare the convergence performance of MA-DRL under different quantization bits. The proposed MA-DRL algorithm achieves approximately 0.34 packets/time slot when $B = 5$ bits after 6,500 iterations, while the case with $B = 2$ bits obtain a solution of 0.3 packets/time slot after 5,000 iterations. This result shows that due to the quantization error and the space of exploration, a larger value of quantization bits leads to better throughput in IRS-assisted networks, but slower convergence.

Fig. 8 indicates the impact of target delay on the throughput with the secrecy constraint in *Case 2*. From this figure, we can see that the throughput increases with the target delay because

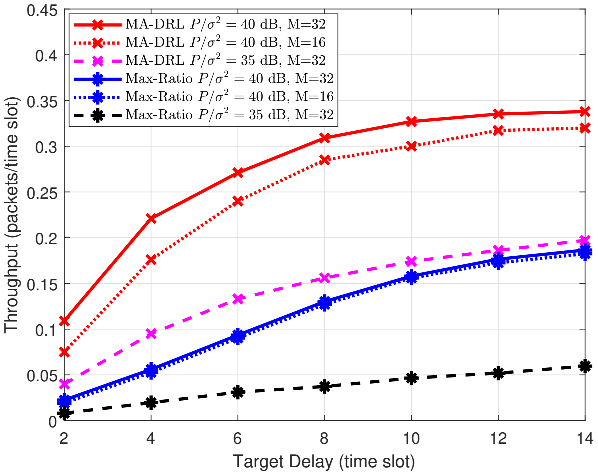

Fig. 8. Throughput versus target delay.

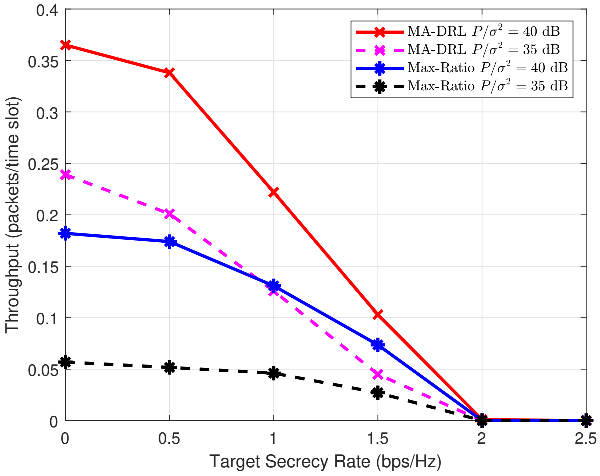

Fig. 10. Throughput versus target secrecy rate.

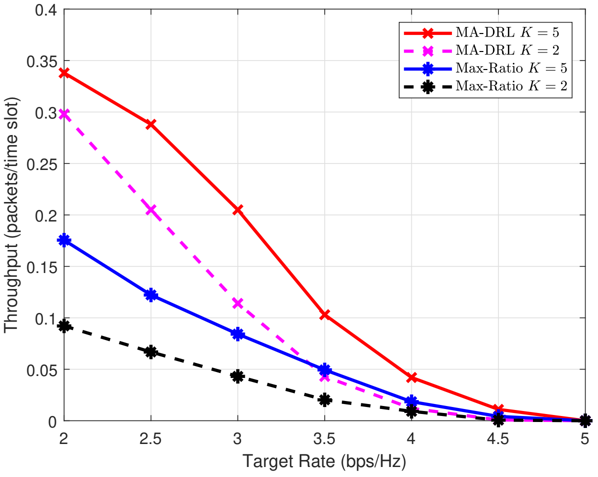

Fig. 9. Throughput versus target rate.

the packets can stay in the corresponding buffer for more time slots, leading to the decreasing of the outage probability. Furthermore, the proposed MA-DRL obtains a better solution from learning experience, compared with the benchmark. The proposed MA-DRL algorithm achieves approximately 0.33 packets/time slot when the target delay $\varsigma = 10$ time slots and the transmit power to noise ratio $P/\sigma_n^2 = 40$ dB, while the max-ratio scheme only achieves 0.18 packets/time slot. In addition, the comparison between $P/\sigma_n^2 = 40$ dB and $P/\sigma_n^2 = 35$ dB shows that higher SNR leads to better performance, due to the decreasing of average delay of packets. Moreover, with a larger value of IRS element N, the proposed algorithm obtains better throughput with constraints, while it is difficult for the benchmark to get gain from the increasing number of N. This result shows the benefit of IRS reflection coefficient optimization in IRS-assisted cooperative networks.

In Fig. 9, we analyze the performance of the proposed scheme and the benchmark under different target rate. It can be seen that in all results, the throughput decreases as the target rate increases. This is because a larger value of target rate leads to more outages in the proposed network. As expected, MA-DRL can optimize the solution from learning experience and achieve better result than the benchmark. As we can see, MA-DRL algorithm achieves approximately 0.21 packets/time slot when the target rate $\omega = 3$ bps/Hz and the number of relay $K = 5$, while the max-ratio scheme only achieves 0.08 packets/time slot. Moreover, when the number of relay K varies, the proposed MA-DRL algorithm always improves the performance compared with max-ratio. This comparison shows the MA-DRL algorithm exhibits robust performance in dynamic networks.

The results in Fig. 10 show that all performance decrease as the target secrecy rate increases due to the increasing secrecy outage probability. The proposed MA-DRL algorithm learns from the environments with different target secrecy rates, and obtains a solution of approximately 0.34 packets/time slot when the target secrecy rate $\varsigma = 0.5$ bps/Hz and the transmit power to noise ratio $P/\sigma_n^2 = 40$ dB, while the max-ratio scheme only achieves 0.17 packets/time slot. As the target rate, high SNR helps achieve high throughput with delay and secrecy constraints when the target secrecy rate is high. Moreover, compared with the max-ratio scheme, the proposed MA-DRL algorithm can learn to reduce the impact of low SNR efficiently and gain better performance. The MA-DRL algorithm not only optimizes the buffer-aided relay selection rule to improve the throughput but also adjusts the IRS reflection coefficients to reduce the secrecy outage probability for the selected transmission.

V. Conclusion

In this paper, the multi-agent DRL-based joint design of relay selection and IRS reflection coefficients was investigated in IRS-assisted secure buffer-aided cooperative networks. For practical implementation, discrete IRS phase shifts and reflection amplitudes were considered. Two optimization problems were considered, namely, maximizing the average secrecy rate with the delay constraint, and maximizing the throughput with the delay and secrecy constraints. To obtain the solution, we split each of the optimization problems into two sub-tasks to reduce the space of exploration in DRL, and combine the solutions from sub-tasks as a joint optimization scheme. Considering the limitation of computation ability for wireless devices, we applied the distributed framework to address the sub-task of IRS reflection coefficient optimization by sharing the accumulated gradients instead of the sharing training data. The simulation results showed the benefits of jointly optimizing buffer-aided relay selection and IRS reflection coefficients, and provided a possible way for solving optimization problems in future wireless networks. Finally, we note that the proposed scheme can also be applied with the 3D mmWave channel model. This is a worthy research direction and would be left as our future work.

References

[1] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. Di Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 20–27, Apr. 2015.

[2] Y. Shiu, S. Y. Chang, H. Wu, S. C. . Huang, and H. Chen, "Physical layer security in wireless networks: a tutorial," *IEEE Wireless Communications*, vol. 18, no. 2, pp. 66–74, Apr. 2011.

[3] 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.

[4] L. Wang, K. Wong, S. Jin, G. Zheng, and R. W. Heath, "A new look at physical layer security, caching, and wireless energy harvesting for heterogeneous ultra-dense networks," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 49–55, Jun. 2018.

[5] 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.

[6] Y. Huang, J. Wang, C. Zhong, T. Q. Duong, and G. K. Karagiannidis, "Secure transmission in cooperative relaying networks with multiple antennas," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6843–6856, Oct. 2016.

[7] L. Yang, J. Chen, H. Jiang, S. A. Vorobyov, and H. Zhang, "Optimal relay selection for secure cooperative communications with an adaptive eavesdropper," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 26–42, Jan. 2017.

[8] 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.

[9] X. Liao, Y. Zhang, Z. Wu, Y. Shen, X. Jiang, and H. Inamura, "On security-delay trade-off in two-hop wireless networks with buffer-aided relay selection," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1893–1906, Mar. 2018.

[10] R. Nakai and S. Sugiura, "Physical layer security in buffer-state-based max-ratio relay selection exploiting broadcasting with cooperative beamforming and jamming," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 431–444, Feb. 2019.

[11] Y. Nie, X. Lan, Y. Liu, Q. Chen, G. Chen, L. Fan, and D. Tang, "Achievable rate region of energy-harvesting based secure two-way buffer-aided relay networks," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1610–1625, Nov. 2020.

[12] H. Yang and M. Alouini, "Data-oriented transmission in future wireless systems: Toward trustworthy support of advanced Internet of Things," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 78–83, Sep. 2019.

[13] P. Xu, G. Chen, Z. Yang, and H. Lei, "Buffer-state-based probabilistic relay selection for cooperative networks with delay constraints," *IEEE Wireless Communications Letters*, vol. 9, no. 11, pp. 1855–1859, Nov. 2020.

[14] C. Huang, G. Chen, and Y. Gong, "Delay constrained buffer-aided relay selection in the Internet of Things with decision-assisted reinforcement learning," *IEEE Internet of Things Journal*, Early Access, 2021.

[15] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[16] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface aided wireless communications: A tutorial," *IEEE Transactions on Communications*, Early Access, 2021.

[17] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[18] H. Shen, W. Xu, S. Gong, Z. He, and C. Zhao, "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications," *IEEE Communications Letters*, vol. 23, no. 9, pp. 1488–1492, Sep. 2019.

[19] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Communications Letters*, vol. 8, no. 5, pp. 1410–1414, Oct. 2019.

[20] S. Fang, G. Chen, and Y. Li, "Joint optimization for secure intelligent reflecting surface assisted UAV networks," *IEEE Wireless Communications Letters*, vol. 10, no. 2, pp. 276–280, Feb. 2021.

[21] L. Yang, J. Yang, W. Xie, M. O. Hasna, T. Tsiftsis, and M. D. Renzo, "Secrecy performance analysis of RIS-aided wireless communication systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12296–12300, Oct. 2020.

[22] S. Li, B. Duo, M. Di Renzo, M. Tao, and X. Yuan, "Robust secure UAV communications with the aid of reconfigurable intelligent surfaces," *IEEE Transactions on Wireless Communications*, pp. 1–1, Early Access, 2021.

[23] Z. Abdullah, G. Chen, S. Lambotharan, and J. A. Chambers, "A hybrid relay and intelligent reflecting surface network and its Ergodic performance analysis," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1653–1657, 2020.

[24] Z. Abdullah, G. Chen, S. Lambotharan, and J. A. Chambers, "Optimization of intelligent reflecting surface assisted full-duplex relay networks," *IEEE Wireless Communications Letters*, vol. 10, no. 2, pp. 363–367, Feb, 2021.

[25] I. Yildirim, F. Kilinc, E. Basar, and G. C. Alexandropoulos, "Hybrid RIS-empowered reflection and decode-and-forward relaying for coverage extension," *IEEE Communications Letters*, vol. 25, no. 5, pp. 1692–1696, May. 2021.

[26] C. Huang, G. Chen, Y. Gong, M. Wen, and J. A. Chambers, "Deep reinforcement learning based relay selection in intelligent reflecting surface assisted cooperative networks," *IEEE Wireless Communications Letters*, Early Access, 2021.

[27] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5849–5863, Sep. 2020.

[28] K. Feng, Q. Wang, X. Li, and C. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 745–749, May. 2020.

[29] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[30] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning based intelligent reflecting surface for secure wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[31] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *International Conference on Machine Learning (ICML)*, Sydney, NSW, Australia, Aug. 2017, pp. 2681–2690.

[32] X. Zhang, X. Zhu, W. Bao, L. T. Yang, J. Wang, H. Yan, and H. Chen, "Distributed learning on mobile devices: a new approach to data mining in the Internet of Things," *IEEE Internet of Things Journal*, Early Access, 2020.

[33] J. Mo, M. Tao, and Y. Liu, “Relay placement for physical layer security: A secure connection perspective,” *IEEE Communications Letters*, vol. 16, no. 6, pp. 878–881, Jun. 2012.

[34] 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, Jul. 2013.

[35] M. Mohammadkhani Razlighi and N. Zlatanov, “Buffer-aided relaying for the two-hop full-duplex relay channel with self-interference,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 477–491, Jan. 2018.

[36] M. Alkhawatrah, Y. Gong, G. Chen, S. Lambotharan, and J. A. Chambers, “Buffer-aided relay selection for cooperative NOMA in the Internet of Things,” *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5722–5731, Jun. 2019.

[37] P. Xu, Z. Yang, Z. Ding, I. Krikidis, and Q. Chen, “A novel probabilistic buffer-aided relay selection scheme in cooperative networks,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4548–4552, Apr. 2020.

[38] J. Kochi, R. Nakai, and S. Sugiura, “Hybrid NOMA/OMA broadcasting-and-buffer-state-based relay selection,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1618–1631, Feb. 2021.

[39] W. Mei and R. Zhang, “Performance analysis and user association optimization for wireless network aided by multiple intelligent reflecting surfaces,” *IEEE Transactions on Communications*, pp. 1–1, Early Access, 2021.

[40] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, “Channel estimation for RIS-empowered multi-user MISO wireless communications,” *IEEE Transactions on Communications*, Early Access, 2021.

[41] B. Zheng, C. You, and R. Zhang, “Efficient channel estimation for double-IRS aided multi-user MIMO system,” *IEEE Transactions on Communications*, Early Access, 2021.

[42] T. Mekkawy, R. Yao, N. Qi, and Y. Lu, “Secure relay selection for two way amplify-and-forward untrusted relaying networks,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11979–11987, Dec. 2018.

[43] A. Arafa, W. Shin, M. Vaezi, and H. V. Poor, “Secure relaying in non-orthogonal multiple access: Trusted and untrusted scenarios,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 210–222, Apr. 2019.

[44] J. Xia, L. Fan, W. Xu, X. Lei, X. Chen, G. K. Karagiannidis, and A. Nallanathan, “Secure cache-aided multi-relay networks in the presence of multiple eavesdroppers,” *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7672–7685, Nov. 2019.

[45] Y. Han, W. Tang, S. Jin, C. Wen, and X. Ma, “Large intelligent surface-assisted wireless communication exploiting statistical CSI,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8238–8242, Aug. 2019.

[46] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[47] P. Xu, G. Chen, Z. Yang, and M. Di Renzo, “Reconfigurable intelligent surfaces assisted communications with discrete phase shifts: How many quantization levels are required to achieve full diversity?,” *IEEE Wireless Communications Letters*, vol. 10, no. 2, pp. 358–362, Early Access, 2020.

[48] 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 (ICML)*, New York, NY, USA, Jan. 2016, pp. 1928-1937.

[49] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural networks*, vol. 61, pp. 85–117, 2015.

[50] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, and L. Hanzo, “Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1719–1734, Aug. 2020.