Deep Reinforcement Learning Based Intelligent Reflecting Surface for Secure Wireless Communications

Helin Yang[1], Yang Zhao[1], Zehui Xiong[1], Jun Zhao[1], Dusit Niyato[1], Kwok-Yan Lam[1], and Qingqing Wu[2]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, 999078 China
E-mail: {hyang013, s180049, zxiong002, junzhao, dniyato, kwokyan.lam}@ntu.edu.sg

GLOBECOM 2020 - 2020 IEEE Global Communications Conference | 978-1-7281-8298-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/GLOBECOM42002.2020.9322615

***Abstract*—In this paper, we study an intelligent reflecting surface (IRS)-aided wireless secure communication system for physical layer security, where an IRS is deployed to adjust its reflecting elements to secure the communication of multiple legitimate users in the presence of multiple eavesdroppers. Aiming to improve the system secrecy rate, a design problem for jointly optimizing the base station (BS)'s beamforming and the IRS's reflecting beamforming is formulated considering different quality of service (QoS) requirements and time-varying channel conditions. As the system is highly dynamic and complex, a novel deep reinforcement learning (DRL)-based secure beamforming approach is firstly proposed to achieve the optimal beamforming policy against eavesdroppers in dynamic environments. Simulation results demonstrate that the proposed deep learning based secure beamforming approach can significantly improve the system secrecy performance compared with other approaches.**

***Index Terms*—Physical layer security, intelligent reflecting surface, beamforming, secrecy rate, deep reinforcement learning.**

I. INTRODUCTION

PHYSICAL layer security (PLS) has attracted increasing attention, as it exploits the wireless channel characteristics by using signal processing designs and channel coding to support secure communication services without relying on a shared secret key [1]. So far, a variety of approaches have been reported to improve PLS in wireless communication systems. However, employing a large number of active antennas and relays in PLS systems incurs an excessive hardware cost and the system complexity. Moreover, cooperative jamming and transmitting artificial noise require extra transmit power for security guarantees.

To tackle these shortcomings of the existing approaches [2]-[3], a new paradigm, called intelligent reflecting surface (IRS) [4]-[8], has been proposed as a promising technique to achieve high spectrum efficiency and energy efficiency, and enhance secrecy rate in the fifth generation (5G) and beyond wireless communication systems. In particular, IRS is a uniform planar array which is comprised of a number of low-cost passive reflecting elements, where each of elements adaptively adjusts its reflection amplitude and/or phase to control the strength and direction of the electromagnetic wave, hence IRS is capable of enhancing and/or weakening the reflected signals at different users [9]. As a result, the reflected signal by IRS can increase the received signal at legitimate users while suppressing the signal at the eavesdroppers [4]-[8]. Hence, from the PLS perspective, some innovative studies have been recently devoted to performance optimization for IRS-aided secure communications [9].Initial studies on IRS-aided secure communication systems have reported in [9], [10] , where a simple system model with only a single-antenna legitimate user and a single-antenna eavesdropper was considered in these works. The authors in [9] and [10] applied the alternative optimization (AO) algorithm to jointly optimize the transmit beamforming vector at the base station (BS) and the phase elements at the IRS for the maximization of the secrecy rate, but they did not extend their models to multi-user IRS-assisted secure communication systems.

The above mentioned studies [9], [10] mainly applied the traditional optimization techniques e.g., AO, SDP or MM algorithms to jointly optimize the BSs beamforming and the IRSs reflecting beamforming, which are less efficient for large-scale systems. Inspired by the recent advances of artificial intelligence (AI), several works attempted to utilize AI algorithms to optimize IRSs reflecting beamforming [12]-[15]. Deep learning (DL) was exploited to search the optimal IRS reflection matrices that maximize the achievable system rate in an IRS-aided communication system, and the simulation demonstrated that DL significantly outperforms conventional algorithms. Moreover, the authors in [14] and [15] proposed deep reinforcement learning (DRL) based approach to address the non-convex optimization problem, and the phase shifts at the IRS are optimized effectively. However, the works [12]-[15] merely considered to maximize the system achievable rate of a single user without considering the scenario of multiple users, secure communication and imperfect CSI in their models. The authors in [16] and [17] applied reinforcement learning (RL) to achieve smart beamforming at the BS against an eavesdropper in complex environments. To the best of our knowledge, RL or DRL has not been explored yet in prior works to optimize both the BS's transmit beamforming and the IRS's reflect beamforming in dynamic IRS-aided secure communication systems with multiple eavesdroppers.

In this paper, we investigate an IRS-aided secure communication system with the objective to maximize the system secrecy rate of multiple legitimate users in the presence of multiple eavesdroppers, while guaranteeing quality of service (QoS) requirements of legitimate users. The optimization problem is formulated as Markov decision process (MDP), and a novel DRL-based secure beamforming approach is firstly proposed to jointly optimize the beamforming matrix at the BS

978-1-7281-8298-8/20/$31.00 ©2020 IEEE

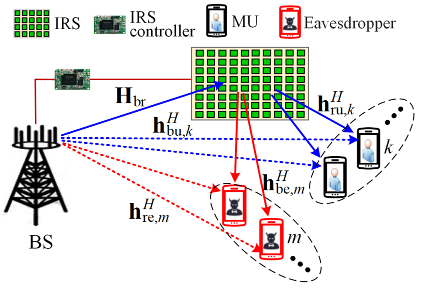

IRS
IRS controller
MU
Eavesdropper
$\mathbf{H}_{\rm br}$
$\mathbf{h}_{{\rm bu},k}^H$
$\mathbf{h}_{{\rm ru},k}^H$
$\mathbf{h}_{{\rm be},m}^H$
$\mathbf{h}_{{\rm re},m}^H$
k
m
BS

Fig. 1. IRS-aided secure communication under multiple eavesdroppers.

and the reflecting beamforming matrix at the IRS in dynamic environments. Simulation results show that the proposed solution can effectively improve the security performance in terms of improving the secrecy rate and the QoS satisfaction probability, compared with other existing approaches.

The rest of this paper is organized as follows. Section II presents the system model and problem formulation. Section III proposes a deep PDS based secure beamforming approach. Section IV provides simulation results and Section V concludes the paper.

II. System Model and Problem Formulation

A. System Model

We consider an IRS-aided secure communication system, as shown in Fig. 1, where the BS is equipped with N antennas to serve K single-antenna legitimate mobile users (MUs) in the presence of M single-antenna eavesdroppers. An IRS with L reflecting elements is deployed in the system to assist secure wireless communications from the BS to the MUs. The IRS is equipped with a controller to coordinate the BS. For the ease of practical implementation, the maximal reflection without power loss at the IRS is considered since the reflecting elements are designed to maximize the reflected desired signal power to the MUs [8]-[12]. In addition, unauthorized eavesdroppers aim to eavesdrop any of the data streams of the MUs. Hence, the use of reflecting beamforming at IRS is also investigated to improve the achievable secrecy rate.

Let $\mathcal{K} = \{1, 2, \ldots, K\}$, $\mathcal{M} = \{1, 2, \ldots, M\}$ and $\mathcal{L} = \{1, 2, \ldots, L\}$ denote the MU set, the eavesdropper set and the IRS reflecting element set, respectively. Let $\mathbf{H}_{\rm br} \in \mathbb{C}^{L\times N}$, $\mathbf{h}_{{\rm bu},k}^H \in \mathbb{C}^{1\times N}$, $\mathbf{h}_{{\rm ru},k}^H \in \mathbb{C}^{1\times L}$, $\mathbf{h}_{{\rm be},m}^H \in \mathbb{C}^{1\times N}$, and $\mathbf{h}_{{\rm re},m}^H \in \mathbb{C}^{1\times L}$ denote the channel coefficients from the BS to the IRS, from the BS to the k-th MU, from the IRS to the k-th MU, from the BS to the m-th eavesdropper, and from the IRS to the m-th eavesdropper, respectively. All the above mentioned channel coefficients in the system are assumed to be small-scale fading with path loss which follows the Rayleigh fading model [6]-[9]. Let $\mathbf{\Psi} = {\rm diag}(\chi_1 e^{j\theta_1}, \chi_2 e^{j\theta_2}, \ldots, \chi_L e^{j\theta_L})$ denote the reflection coefficient matrix associated with effective phase shifts at the IRS, where $\chi_l \in [0, 1]$ and $\theta_l \in [0, 2\pi]$ denote the amplitude reflection factor and the phase shift coefficient on the combined transmitted signal, respectively. As each phase shift is desired to be design to achieve full reflection, we consider that $\chi_l = 1$, $\forall l \in \mathcal{L}$ in the sequel of the paper.

At the BS side, the beamforming vector for the k-th MU is denoted as $\mathbf{v}_k \in \mathbb{C}^{N\times 1}$, which is the continuous linear precoding [6]-[11]. Thus, the transmitted signal for all MUs at the BS is written as $\mathbf{x} = \sum_{k=1}^{K} \mathbf{v}_k s_k$, where s_k is the transmitted symbol for the k-th MU which can be modelled as independent and identically distributed (i.i.d.) random variables with zero mean and unit variance [6]-[10]. The total transmit power at the BS is subject to the maximum power constraint:

$$\mathbb{E}[||\mathbf{x}||^2] = {\rm Tr}(\mathbf{V}\mathbf{V}^H) \le P_{\max} \tag{1}$$

where $\mathbf{V} \triangleq [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K] \in \mathbb{C}^{M\times K}$, and $P_{\max}$ is the maximum transmit power at the BS.

When the BS transmits a secret message to the k-th MU, the MU will receive the signal from the BS and the reflected signal from the IRS. Accordingly, the received signal at MU k can be given by

$$\begin{aligned} y_k = & \underbrace{\left(\mathbf{h}_{{\rm ru},k}^H \mathbf{\Psi} \mathbf{H}_{\rm br} + \mathbf{h}_{{\rm bu},k}^H\right) \mathbf{v}_k s_k}_{\text{desired signal}} \\ & + \underbrace{\sum_{i\in\mathcal{K}, i\neq k} \left(\mathbf{h}_{{\rm ru},k}^H \mathbf{\Psi} \mathbf{H}_{\rm br} + \mathbf{h}_{{\rm bu},k}^H\right) \mathbf{v}_i s_i}_{\text{inter}-\text{user interference}} + n_k \end{aligned} \tag{2}$$

where n_k denotes the additive complex Gaussian noise (AWGN) with the with zero mean and variance δ_k^2 at the k-th MU. In (2), we observe that in addition to the received desired signal, each MU also suffers inter-user interference (IUI) in the system. In addition, the received signal at eavesdropper m is expressed by

$$y_m = \left(\mathbf{h}_{{\rm re},m}^H \mathbf{\Psi} \mathbf{H}_{\rm br} + \mathbf{h}_{{\rm be},m}^H\right) \sum_{k\in\mathcal{K}} \mathbf{v}_k s_k + n_m \tag{3}$$

where n_m is the AWGN of eavesdropper m with variance δ_m^2.

Based on (2), the data rate of the k-th MU in (bits/s/Hz) is given by

$$R_k^{\rm u} = \log_2\left(1 + \frac{\left|(\mathbf{h}_{{\rm ru},k}^H \mathbf{\Psi} \mathbf{H}_{\rm br} + \mathbf{h}_{{\rm bu},k}^H)\mathbf{v}_k\right|^2}{\left|\sum\limits_{i\in\mathcal{K}, i\neq k} (\mathbf{h}_{{\rm ru},k}^H \mathbf{\Psi} \mathbf{H}_{\rm br} + \mathbf{h}_{{\rm bu},k}^H)\mathbf{v}_i\right|^2 + \delta_k^2}\right) \tag{4}$$

If the m-th eavesdropper attempts to eavesdrop the signal of the k-th MU, its wiretapped data rate can be expressed by

$$R_{m,k}^{\rm e} = \log_2\left(1 + \frac{\left|(\mathbf{h}_{{\rm re},m}^H \mathbf{\Psi} \mathbf{H}_{\rm br} + \mathbf{h}_{{\rm be},m}^H)\mathbf{v}_k\right|^2}{\left|\sum\limits_{i\in\mathcal{K}, i\neq k} (\mathbf{h}_{{\rm re},m}^H \mathbf{\Psi} \mathbf{H}_{\rm br} + \mathbf{h}_{{\rm be},m}^H)\mathbf{v}_i\right|^2 + \delta_m^2}\right). \tag{5}$$

Since each eavesdropper can eavesdrop any of the K MUs' signal[9]-[12], the achievable minimum-secrecy rate from the BS to the k-th MU can be expressed by

$$R_k^{\rm sec} = \left[R_k^{\rm u} - \max_{\forall m} R_{m,k}^{\rm e}\right]^+ \tag{6}$$

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

B. Problem Formulation based on MDP

Model-free RL is a dynamic programming tool which can be adopted to solve the decision-making problem by learning the optimal solution in dynamic environments [18]. Hence, we model the secure beamforming optimization problem as an MDP problem. In MDP, the IRS-aided secure communication system is treated as an environment, the central controller at the BS is regarded as a learning agent. The key elements of MDP or RL are defined as follows.

State space: Let $\mathcal{S}$ denote the system state space. The current system state $s \in \mathcal{S}$ includes the channel information of all users, the predicted secrecy rate, the transmission data rate of the last time slot and the QoS satisfaction level, which is defined as

$$s = \left\{ \{\mathbf{h}_k\}_{k \in K}, \{\mathbf{h}_m\}_{m \in \mathcal{M}}, \{R_k^{\mathrm{sec}}\}_{k \in \mathcal{K}}, \{R_k\}_{k \in \mathcal{K}}, \{\mathrm{QoS}_k\}_{k \in \mathcal{K}} \right\} \tag{7}$$

where $\mathbf{h}_k$ and $\mathbf{h}_m$ are the channel coefficients of the k-th MU and m-th eavesdropper, respectively. QoS_k is the feedback QoS satisfaction level of the k-th MU. Other parameters in (7) are already defined in Section II.

Action space: Let $\mathcal{A}$ denote the system action space. According to the observed system state s, the central controller chooses the beamforming vector $\{\mathbf{v}_k\}_{k \in \mathcal{K}}$ at the BS and the IRS reflecting beamforming coefficient (phase shift) $\{\theta_l\}_{l \in \mathcal{L}}$ at the IRS. Hence, the action $a \in \mathcal{A}$ can be defined by

$$a = \left\{ \{\mathbf{v}_k\}_{k \in \mathcal{K}}, \{\theta_l\}_{l \in \mathcal{L}} \right\}. \tag{8}$$

Transition probability: Let $\mathcal{T}(s'|s,a)$ represent the transition probability, which is the probability of transitioning to a new state $s' \in \mathcal{S}$, given the action a executed in the sate s.

Reward function: In RL, the reward acts as a signal to evaluate how good the secure beamforming policy is when the agent executes an action at a current state. The reward function represents the optimization objective, and our objective is to maximize the system secrecy rate of all MUs while guaranteeing the QoS requirements. Thus, the presented QoS-aware reward function is expressed as

$$r = \underbrace{\sum_{k \in \mathcal{K}} R_k^{\mathrm{sec}}}_{\text{part 1}} - \underbrace{\sum_{k \in \mathcal{K}} \mu_1 p_k^{\mathrm{sec}}}_{\text{part 2}} - \underbrace{\sum_{k \in \mathcal{K}} \mu_2 p_k^{\mathrm{u}}}_{\text{part 3}} \tag{9}$$

where

$$p_k^{\mathrm{sec}} = \begin{cases} 1, \text{ if } R_k^{\mathrm{sec}} < R_k^{\mathrm{sec}\,,\min}, \forall k \in \mathcal{K}, \\ 0, \text{ otherwise}, \end{cases} \tag{10}$$

$$p_k^{\mathrm{u}} = \begin{cases} 1, \text{ if } R_k < R_k^{\min}, \forall k \in \mathcal{K}, \\ 0, \text{ otherwise}. \end{cases} \tag{11}$$

where $R_k^{\mathrm{sec}\,,\min}$ and $R_k^{\min}$ are the minimum secrecy rate and data rate thresholds of the k-th MU, respectively.

In (9), the part 1 represents the immediate utility (system secrecy rate), the part 2 and the part 3 are the cost functions which are defined as the unsatisfied secrecy rate requirement and the unsatisfied minimum rate requirement, respectively. The coefficients μ_1 and μ_2 are the positive constants of the part 2 and the part 3 in (9), respectively, and they are used to balance the utility and cost [20]-[22].

The goals of (10) and (11) are to impose the QoS satisfaction levels of both the secrecy rate and the minimum data rate requirements, respectively. If the QoS requirement is satisfied in the current time slot, then $p_k^{\mathrm{sec}} = 0$ or $p_k^{\mathrm{u}} = 0$, indicating that there is no punishment of the reward due to successful QoS guarantees.

The goal of the learning agent is to search for an optimal policy π^* (π is a mapping from states in $\mathcal{S}$ to the probabilities of choosing an action in $\mathcal{A}$: $\pi(s) : \mathcal{S} \to \mathcal{A}$) that maximizes the long-term expected discounted reward, and the cumulative discounted reward function can be defined as

$$U_t = \sum_{\tau=0}^{\infty} \gamma_\tau r_{t+\tau+1} \tag{12}$$

where $\gamma \in (0,1]$ denotes the discount factor. Under a certain policy π, the state-action function of the agent with a state-action pair $(s,\, a)$ is given by

$$Q^\pi(s_t, a_t) = \mathbb{E}_\pi \left[U_t | s_t = s, a_t = a \right]. \tag{13}$$

The conventional Q-Learning algorithm can be adopted to learn the optimal policy. The key objective of Q-Learning is to update Q-table by using the Bellman's equation as follows:

$$Q^\pi(s_t, a_t) = \\ \mathbb{E}_\pi \left[r_t + \gamma \sum_{s_{t+1} \in \mathcal{S}} T(s_{t+1}|s_t, a_t) \sum_{a_{t+1} \in \mathcal{A}} \pi(s_{t+1}, a_{t+1}) Q^\pi(s_{t+1}, a_{t+1}) \right]. \tag{14}$$

The optimal action-value function in (14) is equivalent to the Bellman optimality equation, which is expressed by

$$Q^*(s_t, a_t) = r_t + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}) \tag{15}$$

and the state-value function is achieved as follows:

$$V(s_t) = \max_{a_t \in \mathcal{A}} Q(s_t, a_t). \tag{16}$$

In addition, the Q-value is updated as follows:

$$Q_{t+1}(s_t, a_t) = (1 - \alpha_t) Q_t(s_t, a_t) + \alpha_t \left(r_t + \gamma V_t(s_{t+1}) \right) \tag{17}$$

where $\alpha_t \in (0,1]$ is the learning rate. Q-Learning generally constructs a lookup Q-table $Q(s,a)$, and the agent selects actions based on the greedy policy for each learning step [18]. In the $\varepsilon-$greedy policy, the agent chooses the action with the maximum Q-table value with probability $1 - \varepsilon$, whereas a random action is picked with probability ε to avoid achieving stuck at non-optimal policies [18].

III. Deep PDS Learning Based Secure Beamforming

We propose a deep PDS learning based secure beamforming approach, as shown in Fig. 2, where PDS-learning is utilized to enable the learning agent to learn and adapt faster in dynamic environments. In detail, the agent utilizes the observed state (i.e, CSI, previous secrecy rate, QoS satisfaction level), the feedback reward from environment as well as the historical experience from the replay buffer to train its learning model. After that, the agent employs the trained model to make decision (beamforming matrices $\mathbf{V}$ and $\mathbf{\Psi}$) based on its learned policy. The modified deep PDS-learning can trace the environment dynamic characteristics, and then adjust the transmit beamforming at the BS and the reflecting elements at the IRS accordingly, which can speed up the learning efficiency in dynamic environments.

PDS-learning can be defined as an immediate system state $\tilde{s}_t \in \mathcal{S}$ happens after executing an action a_t at the current

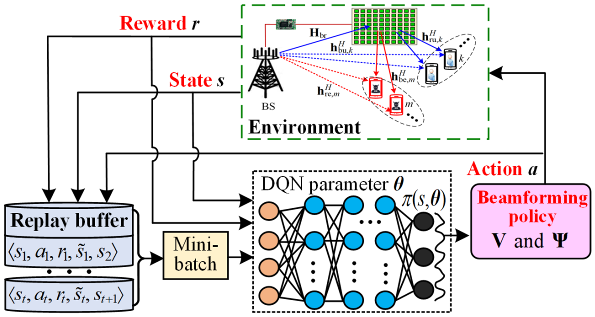

Reward r
State s
$\mathbf{H}_{\rm br}$
$\mathbf{h}_{{\rm bu},k}^H$
$\mathbf{h}_{{\rm ru},k}^H$
$\mathbf{h}_{{\rm be},m}^H$
$\mathbf{h}_{{\rm re},m}^H$
k
m
BS
Environment
Action a
DQN parameter θ
$\pi(s,\theta)$
Beamforming policy
$\mathbf{V}$ and $\mathbf{\Psi}$
Replay buffer
$\langle s_1, a_1, r_1, \tilde{s}_1, s_2\rangle$
$\langle s_t, a_t, r_t, \tilde{s}_t, s_{t+1}\rangle$
Mini-batch

Fig. 2. Deep PDS learning based secure beamforming.

state s_t and before the next time state s_{t+1}. In detail, the PDS-learning agent takes an action a_t at state s_t, and then will receive known reward $r^{\rm k}(s_t, a_t)$ from the environment before transitioning the current state s_t to the PDS state $\tilde{s}_t$ with a known transition probability $\mathcal{T}^{\rm k}(\tilde{s}_t|s_t, a_t)$. After that, the PDS state further transform to the next state s_{t+1} with an unknown transition probability $\mathcal{T}^{\rm u}(s_{t+1}|\tilde{s}_t, a_t)$ and an unknown reward $r^{\rm u}(s_t, a_t)$, where corresponds to the wireless CSI dynamics. In PDS-learning, s_{t+1} is independent of s_t given the PDS state $\tilde{s}_t$, and the reward $r(s_t, a_t)$ is decomposed into the sum of $r^{\rm k}(s_t, a_t)$ and $r^{\rm u}(s_t, a_t)$ at $\tilde{s}_t$ and s_{t+1}, respectively. Mathematically, the state transition probability in PDS-learning from s_t to s_{t+1} admits

$$\mathcal{T}(s_{t+1}|s_t, a_t) = \sum\nolimits_{\tilde{s}_t} \mathcal{T}^{\rm u}(s_{t+1}|\tilde{s}_t, a_t)\mathcal{T}^{\rm k}(\tilde{s}_t|s_t, a_t). \quad (18)$$

Moreover, it can be verified that the reward of the current state-action pair (s_t, a_t) is expressed by

$$r(s_t, a_t) = r^{\rm k}(s_t, a_t) + \sum\nolimits_{\tilde{s}_t} \mathcal{T}^{\rm k}(\tilde{s}_t|s_t, a_t)r^{\rm u}(\tilde{s}_t, a_t). \quad (19)$$

At the time slot t, the PDS action-value function $\tilde{Q}(\tilde{s}_t, a_t)$ of the current PDS state-action pair $(\tilde{s}_t, a_t)$ is defined as

$$\tilde{Q}(\tilde{s}_t, a_t) = r^{\rm u}(\tilde{s}_t, a_t) + \gamma \sum_{s_{t+1}} \mathcal{T}^{\rm u}(s_{t+1}|\tilde{s}_t, a_t)V(s_{t+1}). \quad (20)$$

By employing the extra information (the known transition probability $\mathcal{T}^{\rm k}(\tilde{s}_t|s_t, a_t)$ and known reward $r^{\rm k}(s_t, a_t)$), the Q-function $\hat{Q}(s_t, a_t)$ in PDS-learning can be further expanded under all state-action pairs (s, a), which is expressed by

$$\hat{Q}(s_t, a_t) = r^{\rm k}(s_t, a_t) + \sum\nolimits_{\tilde{s}_t} \mathcal{T}^{\rm k}(\tilde{s}_t|s_t, a_t)\tilde{Q}(\tilde{s}_t, a_t). \quad (21)$$

The state-value function in PDS-learning is defined by

$$\hat{V}_t(s_t) = \sum\nolimits_{s_{t+1}} \mathcal{T}^{\rm k}(s_{t+1}|s_t, a_t)\tilde{V}(s_{t+1}) \quad (22)$$

where $\tilde{V}_t(s_{t+1}) = \max\limits_{a_t \in \mathcal{A}} \tilde{Q}_t(\tilde{s}_{t+1}, a_t)$. At each time slot, the PDS action-value function $\tilde{Q}(\tilde{s}_t, a_t)$ is updated by

$$\tilde{Q}_{t+1}(\tilde{s}_t, a_t) = (1-\alpha_t)\tilde{Q}_t(\tilde{s}_t, a_t) + \alpha_t\left(r^{\rm u}(\tilde{s}_t, a_t) + \gamma\hat{V}_t(s_{t+1})\right). \quad (23)$$

After updating $\tilde{Q}_{t+1}(\tilde{s}_t, a_t)$, the action-value function $\hat{Q}_{t+1}(s_t, a_t)$ can be updated by plugging $\tilde{Q}_{t+1}(\tilde{s}_t, a_t)$ into (23).

After presenting in the above modified PDS-learning, a deep PDS learning algorithm is presented. In the presented learning algorithm, the traditional DQN is adopted to estimatete the action-value Q-function $Q(s, a)$ by using $Q(s, a; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denote the DNN parameter. The objective of DQN is to minimize the following loss function at each time slot

Algorithm 1 Deep PDS Learning Based Secure Beamforming

```
Input: IRS-aided secure communication simulator and QoS requirements of all MUs.
Initialize: DQN with initial Q-function $Q(s, a; \boldsymbol{\theta})$, parameters $\boldsymbol{\theta}$, learning rate $\alpha$ and $\beta$.
Initialize: experience replay buffer $\mathcal{D}$ with size $D$, and mini-batch size $H$.
for each episode = 1, 2, ..., $N^{\rm epi}$ do
    Observe an initial system state $s$;
    for each time step $t$= 0, 1, 2, ..., $T$ do
        Select action based on the $\varepsilon$-greedy policy at current state $s_t$: choose a random action $a_t$ with probability $\varepsilon$;
        Otherwise, $a_t = \arg\max_{a_t \in \mathcal{A}} Q(s_t, a_t; \theta_t)$;
        Execute action $a_t$, receive an immediate reward $r^{\rm k}(s_t, a_t)$ and observe the sate transition from $s_t$ to PDS state $\tilde{s}_t$ and then to the next state $s_{t+1}$;
        Update PDS action-value function $\tilde{Q}(\tilde{s}_t, a_t; \boldsymbol{\theta}_t)$ by (21);
        Update the Q-function $\hat{Q}(s_t, a_t; \boldsymbol{\theta}_t)$ by (23);
        Store experience $e_t = \langle s_t, a_t, r_t, \tilde{s}_t, s_{t+1}\rangle$ in experience replay buffer $\mathcal{D}$;
    end for
end for
Output: Return the deep PDS learning model.
```

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}_t) &= \left[\left\{\hat{V}_t(s_t; \boldsymbol{\theta}_t) - \hat{Q}(s_t, a_t; \boldsymbol{\theta}_t)\right\}^2\right] \\ &= \left[\left\{r(s_t, a_t) + \gamma \max_{a_{t+1} \in \mathcal{A}} \hat{Q}_t(s_{t+1}, a_{t+1}; \boldsymbol{\theta}_t) - \hat{Q}(s_t, a_t; \boldsymbol{\theta}_t)\right\}^2\right]\end{aligned} \quad (24)$$

where $\hat{V}_t(s_t; \boldsymbol{\theta}_t) = r(s_t, a_t) + \gamma \max\limits_{a_{t+1} \in \mathcal{A}} \hat{Q}_t(s_{t+1}, a_{t+1}; \boldsymbol{\theta}_t)$ is the target value. The error between $\hat{V}_t(s_t; \boldsymbol{\theta}_t)$ and the estimated value $\hat{Q}(s_t, a_t; \boldsymbol{\theta}_t)$ is usually called temporal-difference (TD) error, which is expressed by

$$\delta_t = \hat{V}_t(s_t; \boldsymbol{\theta}_t) - \hat{Q}(s_t, a_t; \boldsymbol{\theta}_t). \quad (25)$$

The DNN parameter $\boldsymbol{\theta}$ is achieved by taking the partial differentiation of the objective function (26) with respect to $\boldsymbol{\theta}$, which is given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \beta \nabla \mathcal{L}(\boldsymbol{\theta}_t). \quad (26)$$

where β is the learning rate of $\boldsymbol{\theta}$, and $\nabla(\cdot)$ denotes the first-order partial derivative.

Accordingly, the policy $\hat{\pi}_t(s)$ of the modified deep PDS-learning algorithm is given by

$$\hat{\pi}_t(s) = \arg\max_{a_t \in \mathcal{A}} \hat{Q}(s_t, a_t; \boldsymbol{\theta}_t). \quad (27)$$

Similar to most DRL algorithms, our proposed deep PDS learning based secure beamforming approach consists of two stages, i.e., the training stage and implement stage. The training process of the proposed approach is shown in **Algorithm 1**. A central controller at the BS is responsible for collecting environment information and making decision for secure beamforming.

IV. Simulation Results and Analysis

This section evaluates the performance of the IRS-aided secure communication system. As illustrated in Fig. 3, K

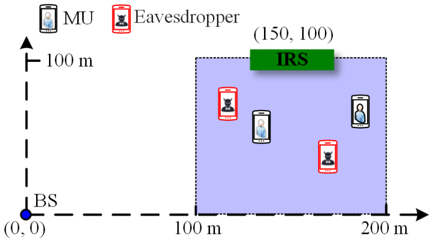


Fig. 3. Simulation setup.

single-antenna MUs and M single-antenna eavesdroppers are randomly located in the $100m \times 100m$ half right-hand side rectangular of Fig. 3 (light blue area) in a two-dimensional plane. The BS and the IRS are located at (0, 0) and (150, 100) in meter (m), respectively. The background noise power of MUs and eavesdroppers is equal to -90 dBm. We set the number of antennas at the BS is $N = 4$, the number of MUs is $K = 2$ and the number of eavesdroppers is $M = 2$. The transmit power $P_{\max}$ at the BS varies between 15 dBm and 40 dBm, the number of IRS elements L varies between 10 and 60, and the outdated CSI coefficient ρ varies from 0.5 to 1 for different simulation settings. The minimum secrecy rate and the minimum transmission data rate are 3 bits/s/Hz and 5 bits/s/Hz, respectively. The path loss model is defined by $PL = (PL_0 - 10\varsigma \log 10(d/d_0))$ dB, where $PL_0 = 30$ dB is the path loss at the reference distance $d_0 = 1$ m [9], $\varsigma = 3$ is the path loss exponent, and d is the distance from the transmitter to the receiver. The learning model consists of three connected hidden layers, containing 500, 250, and 200 neurons [23], [24] , respectively. The learning rate is set to $\alpha = 2 \times 10^{-3}$, the discount factor is set to $\gamma = 0.95$ and the final exploration rate is set to $\varepsilon = 0.1$. The parameters μ_1 and μ_2 in (9) are set to $\mu_1 = \mu_2 = 2$ to balance the utility and cost.

In addition, simulation results are provided to evaluate the performance of the proposed deep PDS learning based secure beamforming approach (denoted as deep PDS- beamforming), and compare it with the following existing approaches:

- The classical DQN based secure beamforming approach (denoted as DQN-based beamforming), where DNN is employed to estimate the Q-value function.
- The secrecy rate maximization approach which optimizes the BS's transmit beamforming and the IRS's reflect beamforming by using an iterative algorithm, which is similar to the suboptimal solution [9] (denoted as Baseline 1 [9]).
- The optimal BS's transmit beamforming approach without IRS assistance (denoted as optimal BS without IRS).

In Fig. 4, we first investigate the convergence performances of DRL-based secure beamforming algorithms in terms of the average secrecy rate per MU, when $P_{\max} = 30$ dBm and $L = 40$. It is observed that the secrecy rate of two DRL-based algorithms first enhances and then converges to a constant level. In addition, it is worth noting that the proposed learning approach has the faster convergence speed and higher secrecy rate than that of the DQN approach by adopting by PDS-learning to enhance the learning efficiency, in order to improve convergence speed and provide the global optimal solution for joint beamforming optimization problem.

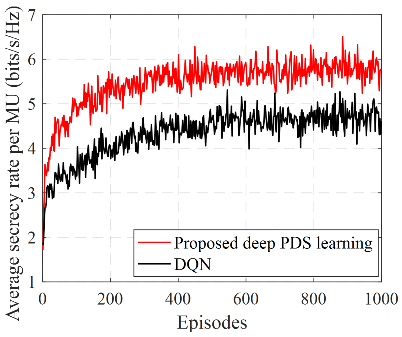
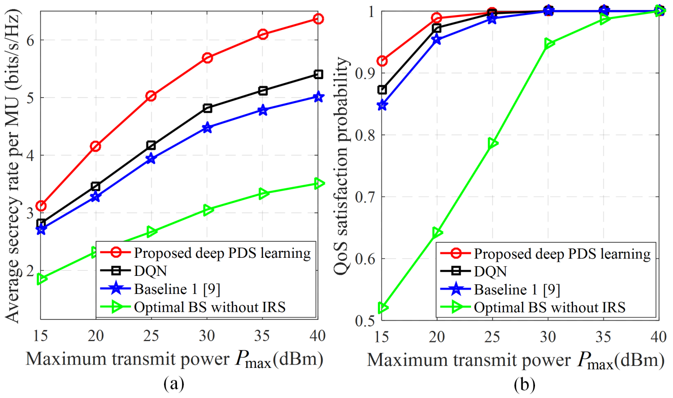


Fig. 4. Convergence comparisons of DRL algorithms.



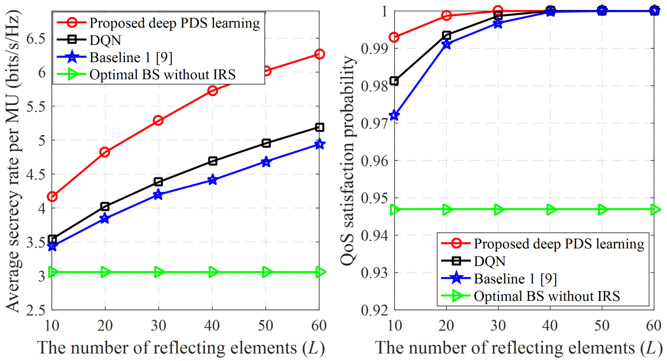
(a)

(b)

Fig. 5. Performance comparisons versus the maximum transmit power.

Fig. 5 shows the average secrecy rate and QoS satisfaction level versus the maximum transmit power $P_{\max}$, when $L = 40$. As expected, both the secrecy rate and QoS satisfaction level of all the approaches enhance monotonically with increasing $P_{\max}$ due to the increase of the received SINR at MUs. In addition, we find that our proposed learning approach outperforms the Baseline 1 approach. In fact, our approach jointly optimizes the beamforming matrixes $\mathbf{V}$ and $\mathbf{\Psi}$, which can simultaneously facilitates more favorable channel propagation benefit for MUs and impair eavesdroppers, while the Baseline1 approach optimizes the beamforming matrixes in an iterative way. Moreover, our proposed approach has higher performance than DQN, due to its efficient learning capacity by utilizing PDS-learning in the dynamic environment. From Fig. 5, the three IRS assisted secure beamforming approaches provide significant higher performance than the traditional system without IRS. This indicates that the IRS can effectively guarantee secure communication and QoS requirements via reflecting beamforming, where reflecting elements at the IRS can be adjusted to maximize the received SINR at MUs and suppress the wiretapped rate at eavesdroppers.

In Fig. 6, the achievable secrecy rate and QoS satisfaction level performance of all approaches are evaluated through changing the IRS elements, i.e., from $L = 10$ to 60, when $P_{\max} = 30$ dBm. For the secure beamforming approaches assisted by the IRS, their performance are obvious increment with the number of the IRS elements. The improvement results from the fact that more IRS elements, more signal paths and signal power can be reflected by the IRS to improve the received SINR at the MUs but to decrease the received SINR at the eavesdroppers. From Fig. 6(a), the secrecy rate



(a)

(b)

Fig. 6. Performance comparisons versus the number of IRS elements.

of the proposed learning approach is higher than those of the Baseline 1 and DQN approaches, especially, their performance gap also obviously increases with L, this is because that with more reflecting elements at the IRS, the proposed deep PDS-PER learning based secure communication approach becomes more flexible for optimal phase shift design and hence achieves higher gains. In addition, from Fig. 6(b) compared with the Baseline 1 and DQN approaches, as the reflecting elements at the IRS increases, we observe that the proposed learning approach is the first one who attains 100% QoS satisfaction level. This superior achievements are based on the particular design of the QoS-aware reward function shown in (9) for secure communication.

V. Conclusion

In this paper, we have investigated the joint BS's beamforming and IRS's reflect beamforming optimization problem. We formulated the secure beamforming optimization problem as an RL problem, and a deep PDS learning based secure beamforming approach has been proposed to jointly optimize both the BS's beamforming and the IRS's reflect beamforming in the dynamic IRS-aided secure communication system. Simulation results have verified that the effectiveness of the proposed learning approach compared with other approaches. We will apply IRS in visible light communication systems in the future [25].

VI. Acknowledgements

This work is support by Nanyang Technological University (NTU) Startup Grant, Alibaba-NTU Singapore Joint Research Institute (JRI), Singapore Ministry of Education Academic Research Fund Tier 1 RG128/18, Tier 1 RG115/19, Tier 1 RT07/19, Tier 1 RT01/19, and Tier 2 MOE2019-T2-1-176, NTU-WASP Joint Project, Singapore National Research Foundation (NRF) under its Strategic Capability Research Centres Funding Initiative: Strategic Centre for Research in Privacy-Preserving Technologies & Systems (SCRIPTS), Energy Research Institute @NTU (ERIAN), Singapore NRF National Satellite of Excellence, Design Science and Technology for Secure Critical Infrastructure NSoE DeST-SCI2019-0012, AI Singapore (AISG) 100 Experiments (100E) programme, and NTU Project for Large Vertical Take-Off & Landing (VTOL) Research Platform..

References

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

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

[3] Z. Mobini, M. Mohammadi, and C. Tellambura, "Wireless-powered full-duplex relay and friendly jamming for secure cooperative communications," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 3, pp. 621-634, Mar. 2019.

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

[5] J. Zhao, "A survey of intelligent reflecting surfaces (IRSs): Towards 6G wireless communication networks," 2019. [Online]. Available: https://arxiv.org/abs/1907.04789.

[6] H. Han, *et al.*, "Intelligent reflecting surface aided power control for physical-layer broadcasting," 2019. [Online]. Available: https://arxiv.org/abs/1912.03468.

[7] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157-4170, Aug. 2019.

[8] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394-5409, Nov. 2019.

[9] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1410-1414, Oct. 2019.

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

[11] D. Xu, X. Yu, Y. Sun, D. W. K. Ng, and R. Schober, "Resource allocation for secure IRS-assisted multiuser MISO systems," 2019. [Online]. Available: http://arxiv.org/abs/1907.03085.

[12] C. Huang, G. C. Alexandropoulos, C. Yuen, and M. Debbah, "Indoor signal focusing with deep learning designed reconfigurable intelligent surfaces," 2019. [Online]. Available: https://arxiv.org/abs/1905.07726.

[13] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," 2019. [Online]. Available: https://arxiv.org/abs/1904.10136.

[14] K. Feng, Q. Wang, X. Li and C. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," Appear in *IEEE Wireless Commun. Lett.*. DOI: 10.1109/LWC.2020.2969167.

[15] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," 2020. [Online]. Available: https://arxiv.org/abs/2002.10072.

[16] C. Li, W. Zhou, K. Yu, L. Fan, and J. Xia, "Enhanced secure transmission against intelligent attacks," *IEEE Access*, vol. 7, pp. 53596-53602, Aug. 2019.

[17] L. Xiao, G. Sheng, S. Liu, H. Dai, M. Peng, and J. Song, "Deep reinforcement learning-enabled secure visible light communication against eavesdropping," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6994-7005, Oct. 2019.

[18] H. Yang, X. Xie, and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157-4169, May 2019.

[19] H. L. Yang A. Alphones, C. Chen, W. D. Zhong, and X. Z. Xie, "Learning-based energy-efficient resource management by heterogeneous RF/VLC for ultra-reliable low-latency industrial IoT networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5565-5576, Aug. 2020.

[20] X. He, R. Jin, and H. Dai, "Deep PDS-learning for privacy-aware offloading in MEC-enabled IoT," *IEEE Internet of Things J.*, vol. 6, no. 3, pp. 4547-4555, Jun. 2019

[21] N. Mastronarde and M. van der Schaar, "Joint physical-layer and system-level power management for delay-sensitive wireless communications," *IEEE Trans. Mobile Comput.*, vol. 12, no. 4, pp. 694-709, Apr. 2013.

[22] T. Schaul, J. Quan, I. Antonoglou, and D. Silver,"Prioritized experience replay," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, US, May. 2016, pp. 1C21.

[23] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," Appear in *IEEE Trans. Commun.*, DOI: 10.1109/TCOMM.2019.2961332.

[24] H. Yang, X. Xie, and M. Kadoch, "Machine learning techniques and a case study for intelligent wireless networks,"*IEEE Network*, vol. 34, no. 3, pp. 208-215, May 2020.

[25] H. Yang *et al.*, "Coordinated resource allocation-based integrated visible light communication and positioning systems for indoor IoT," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4671-4684, Jul. 2020.