Intelligent Reflecting Surface Assisted Anti-Jamming Communications Based on Reinforcement Learning

Helin Yang, Zehui Xiong, Jun Zhao, Dusit Niyato, Qingqing Wu, Massimo Tornatore, Stefano Secci

Intelligent Reflecting Surface Assisted Anti-Jamming Communications Based on Reinforcement Learning

Helin Yang[1], Zehui Xiong[1], Jun Zhao[1], Dusit Niyato[1], Qingqing Wu[2], Massimo Tornatore[3], and Stefano Secci[4]
[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
[3]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
[4] Cnam, Cedric Lab, Paris, France
E-mail: {hyang013, zxiong002, junzhao, and dniyato}@ntu.edu.sg

***Abstract*—Malicious jamming launched by smart jammer, which attacks legitimate transmissions has been regarded as one of the critical security challenges in wireless communications. Thus, this paper exploits intelligent reflecting surface (IRS) to enhance anti-jamming communication performance and mitigate jamming interference by adjusting the surface reflecting elements at the IRS. Aiming to enhance the communication performance against smart jammer, an optimization problem for jointly optimizing power allocation at the base station (BS) and reflecting beamforming at the IRS is formulated. As the jamming model and jamming behavior are dynamic and unknown, a win or learn fast policy hill-climbing (WoLFCPHC) learning approach is proposed to jointly optimize the anti-jamming power allocation and reflecting beamforming strategy without the knowledge of the jamming model. Simulation results demonstrate that the proposed anti-jamming based-learning approach can efficiently improve both the the IRS-assisted system rate and transmission protection level compared with existing solutions.**

***Index Terms*—Anti-jamming, intelligent reflecting surface, power allocation, beamforming, reinforcement learning.**

I. Introduction

DUE to the inherent broadcast and openness nature of wireless channels [1], [2], wireless transmissions can be easily vulnerable to jamming attacks. In particular, malicious jammers can intentionally send jamming signals over the legitimate channels to degrade communication performance [1]-[3], which has been considered as one of the serious threats in wireless communications. In this regard, lots of jamming-related studies have been recently presented to defend jamming attacks, including frequency hopping, power control, relay assistance, beamforming, and so on.

Frequency-hopping (FH) is one of the powerful techniques which has been widely adopted to allow a wireless user to quickly switch its current operating frequency to other frequency spectrum, thereby avoiding potential jamming attacks [4]-[6]. In [4] and [5], a mode-FH approach was presented to jointly utilize conventional FH to further improve the communication performance in the presence of jammers. In [6], Hanawal *et al*. proposed a joint FH and rate adaptation scheme to avoid jamming attacks in the presence of a jammer. Besides FH, power control is another commonly used technique, e.g., [3], [7]-[9]. As an example, [7] and [8] investigated a jammed wireless system where the system operator tries to control the transmit power to maximize system rate. The authors in [3] and [9] studied the anti-jamming problem with power control strategies, by leveraging the game theory to optimize the power control policy of the transmitter against jammers. Moreover, cooperative communication using trusted relays has been proposed as one promising anti-jamming technique for improving the physical layer security [10]-[12], and robust joint cooperative beamforming and jamming designs were proposed to maximize the achievable rate under the imperfect channel state information (CSI) of a jammer.

To deal with uncertain and/or unknown jamming attack models, such as jamming policies and jamming power levels, some existing studies utilized reinforcement learning (RL) algorithms have been applied in some existing studies to optimize the jamming resistance policy in dynamic wireless communication systems [13]-[15]. In [13], a policy hill climbing (PHC)-based Q-Learning approach was studied to improve the communication performance against jamming without knowing the jamming model. In [14] and [15], the authors adopted deep reinforcement learning (DRL) algorithms that enable transmitters to quickly obtain an optimal policy to guarantee security performance against jamming.

However, despite the effectiveness of the above mentioned anti-jamming schemes [3]-[15], employing a number of active relays incurs an excessive hardware cost, and anti-jamming beamforming and power control in communication systems is generally energy-consuming. To tackle these shortcomings, a new paradigm, called intelligent reflecting surface (IRS) [16], [17], has been recently proposed as a promising technique to enhance the secrecy performance. In particular, IRS comprises of a large number of low-cost passive reflecting elements, where each of the elements adaptively adjusts its reflection amplitude and/or phase to control the strength and direction of the reflected electromagnetic wave [16], [17]. As a result, IRS has been employed in wireless commutation systems to devote to security performance optimization [18]-[22]. In [18]-[21], the authors investigated the physical layer security enhancement of IRS-assisted communications systems, where both the BS¡s beamforming and the IRS¡s phase shifts were jointly optimized to improve secrecy rate in the presence of an eavesdropper. Furthermore, Yang *et al*. in [22] applied DRL to learn the secure beamforming policy in multi-user IRS-aided secure systems, in order to maximize the system secrecy rate in the presence of multiple eavesdroppers. To the best of our knowledge, IRS has not been explored yet in the existing works [3]-[22] to enhance the anti-jamming strategy against smart jamming.

In this paper, we propose an IRS-assisted anti-jamming solution for securing wireless communications. In particular, we aim to maximize the system rate of multiple legitimate users in the presence of a smart multi-antenna jammer. As the jamming

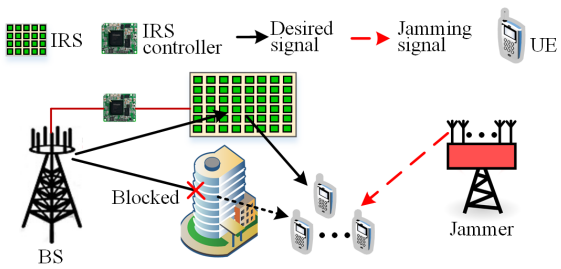

Fig. 1. Illustration of an IRS-assisted communication system against a multi-antenna jammer.

model and jamming behavior are dynamic and unknown, a win or learn fast policy hill-climbing (WoLFCPHC) anti-jamming approach is proposed to achieve the optimal anti-jamming policy, where WoLFCPHC is capable of quickly achieving the optimal policy without knowing the jamming model. Simulation results verify the effectiveness of the proposed learning approach in terms of improving the system rate, compared with the existing approaches.

The remainder of this paper is organized as follows. Section II provides the system model and problem formulation. Section III proposes the WoLFCPHC-based learning approach. Simulation results are provided in Section IV, and the paper is concluded in Section V.

II. System Model and Problem Formulation

A. System Model

As shown in Fig. 1, this paper considers an IRS-assisted communication system, which consists of one BS with N antennas and K single-antenna legitimate user equipments (UEs) located at the cell-edge. The IRS comprised of M reflecting elements is deployed to provide additional communication links so as to improve the performance for the UEs over a given frequency band. The direct communication links of cell-edge UEs may suffer high signal attenuation and these links are severely blocked by obstacles when these UEs located in dead zones. In addition, as illustrated in Fig. 1, a malicious multi-antenna jammer is located near the legitimate UEs who attempts to interfere the legitimate transmissions by sending faked or replayed jamming signal for the UEs via N_{J} antennas, in order to degrade the legitimate communication performance. In this case, deploying the IRS can effectively enhance the desired signal power and mitigate the jamming interference generated from the jammer by designing the reflecting beamforming at the IRS.

Let $\mathcal{K} = \{1, 2, ..., K\}$ and $\mathcal{M} = \{1, 2, ..., M\}$ represent the UE set and the IRS reflecting element set, respectively. Let $\mathbf{G} \in \mathbb{C}^{M\times N}$, $\mathbf{g}_{\mathrm{bu},k}^H \in \mathbb{C}^{1\times N}$, $\mathbf{g}_{\mathrm{ru},k}^H \in \mathbb{C}^{1\times M}$, and $\mathbf{h}_{\mathrm{J},k}^H \in \mathbb{C}^{1\times N_{\mathrm{J}}}$ denote the channel coefficients between the BS and the IRS, between the BS and the k-th UE, between the IRS and the k-th UE, and between the jammer and the k-th UE, respectively. The quasi-static flat-fading model is assumed for all the above channels. Let $\mathbf{\Phi} = \mathrm{diag}(\Phi_1, \Phi_2, ..., \Phi_M) \in \mathbb{C}^{M\times M}$ denotes the reflection coefficient matrix associated with effective phase shifts at the IRS, where $\Phi_m = \omega_m e^{j\theta_m}$ comprises both an reflection amplitude $\omega_m \in [0, 1]$ and a phase shift coefficient $\theta_m \in [0, 2\pi]$ on the combined received signal. Since each phase shift is favorable to be designed to achieve maximum signal reflection, we consider that $\omega_m = 1$, $\forall m \in \mathcal{M}$ in this paper [16]-[22].

The transmitted signal at the BS can be expressed as

$$\mathbf{x} = \sum\nolimits_{k=1}^{K} \sqrt{P_k}\mathbf{w}_k s_k \tag{1}$$

where P_k stands for the transmit power allocated for the k-th UE and we have the power constraint: $\sum_{k=1}^{K} P_k \le P_{\max}$ with $P_{\max}$ being the maximum transmit power of the BS, s_k is the transmitted symbol for the k-th UE, $s_k \in \mathbb{C}$, $\mathbb{E}\{s_k\} = 0$ and $\mathbb{E}\{|s_k|^2\} = 1$ which denotes the unit power information symbol, and $\mathbf{w}_k \in \mathbb{C}^{N\times 1}$ is the beamforming vector for the k-th UE with $\|\mathbf{w}_k\|^2 = 1$, respectively.

This paper considers the case that the smart jammer attempts to disturb the BS's transmitted signal by emitting jamming signal $\mathbf{z}_k \in \mathbb{C}^{N_{\mathrm{J}}\times 1}$ to attack the k-th UE. In addition, the transmit power of the faked jamming signal for the k-th UE is denoted as $P_{\mathrm{J},k} = \|\mathbf{z}_k\|^2 = \mathrm{Tr}(\mathbf{z}_k\mathbf{z}_k^H)$. In this case, for UE k, the received signal consists of the signal coming from its associated BS, the reflected signal from the IRS and the jamming signal from the jammer, which is written by

$$y_k = \underbrace{\left(\mathbf{g}_{\mathrm{ru},k}^H\mathbf{\Phi}G + \mathbf{g}_{\mathrm{bu},k}^H\right)\sqrt{P_k}\mathbf{w}_k s_k}_{\text{desired signal}} + \underbrace{\sum_{i\in\mathcal{K}, i\neq k}\left(\mathbf{g}_{\mathrm{ru},k}^H\mathbf{\Phi}G + \mathbf{g}_{\mathrm{bu},k}^H\right)\sqrt{P_i}\mathbf{w}_i s_i}_{\text{inter-userinterference}} + \underbrace{\sqrt{P_{\mathrm{J},k}}\mathbf{h}_{\mathrm{J},k}^H\mathbf{z}_k}_{\text{jamming signal}} + n_k \tag{2}$$

where n_k denotes the additive complex Gaussian noise with the zero mean and variance δ_k^2 at the k-th UE. In (2), in addition to the received desired signal, each UE also suffers inter-user interference (IUI) and the jamming interference signal in the system. According to (2), the received signal-to-interference-plus-noise-ratio (SINR) at the k-th UE can be expressed as

$$SINR_k = \frac{P_k\left|\left(\mathbf{g}_{\mathrm{ru},k}^H\mathbf{\Phi}\mathbf{G} + \mathbf{g}_{\mathrm{bu},k}^H\right)\mathbf{w}_k\right|^2}{\sum\limits_{i\in\mathcal{K}, i\neq k} P_i\left|\left(\mathbf{g}_{\mathrm{ru},k}^H\mathbf{\Phi}\mathbf{G} + \mathbf{g}_{\mathrm{bu},k}^H\right)\mathbf{w}_i\right|^2 + P_{\mathrm{J},k}\left|\mathbf{h}_{\mathrm{J},k}^H\mathbf{z}_k\right|^2 + \delta_k^2}. \tag{3}$$

B. Problem Formulation

In this paper, we aim to jointly optimize the transmit power allocation $\{P_k\}_{k\in\mathcal{K}}$ at the BS and the reflecting beamforming matrix $\mathbf{\Phi}$ at the IRS to maximize the system achievable rate of all UEs against smart jamming, subject to the transmit power constraint. Accordingly, the optimization problem can be formulated as

$$\begin{aligned} &\max_{\{P_k\}_{k\in\mathcal{K}},\mathbf{\Phi}} \sum_{k\in\mathcal{K}} \log_2\left(1 + SINR_k\right) \\ &s.t.\ (\mathrm{a}):\ \textstyle\sum_{k=1}^{K} P_k \le P_{\max}, \\ &\quad (\mathrm{b}):\ |\Phi_m| = 1,\ 0 \le \theta_m \le 2\pi,\ \forall m \in \mathcal{M} \end{aligned} \tag{4}$$

Note that problem (4) is a non-convex optimization problem, where the objective function is non-concave over the reflecting beamforming matrix $\mathbf{\Phi}$; furthermore, the transmit power allocation variables $\{P_k\}_{k\in\mathcal{K}}$ and $\mathbf{\Phi}$ are intricately coupled in the objective function, thus rendering the joint optimization problem difficult to be solved optimally. So far, many optimization algorithms [16]-[21] have been proposed to obtain an approximate solution to problem (4), by iteratively updating either $\{P_k\}_{k\in\mathcal{K}}$ or $\mathbf{\Phi}$ with the other fixed at each

iteration. Hence, this paper proposes an effective solution to address such kind of the optimization problem, which will be provided in the next section. In addition, it is worth noting that this paper mainly pays attention to jointly optimize the power allocation and the reflecting beamforming, so the transmit beamforming vector $\mathbf{w}_k$ is set by maximizing the received signal power at the IRS as the directin link from the BS to the UEs suffer high signal attenuation by obstacles [16], [17].

III. Joint Power Allocation and Reflecting Beamforming Based on RL

The problem formulated in (4) is difficult to be solved as mentioned at the end of the last section. Model-free RL is one of the dynamic programming tools which has the ability to address the decision-making problem by achieving an optimal policy in dynamic uncertain environments [33]. Thus, this paper models the optimization problem as an RL, and a WoLF-PHC-based joint power allocation and reflecting beamforming approach is proposed to learn the optimal anti-jamming strategy.

A. Optimization Problem Transformation Based on RL

In RL, the IRS-assisted communication system is acted as an environment and the central controller at the BS is regarded as a learning agent. In addition to the environment and the agent, an RL also includes a set of possible system states $\mathcal{S}$, a set of available actions $\mathcal{A}$, and a reward function r, where the learning agent continually learns by interacting with the environment. The main elements of RL are introduced as follows:

States: The system state $s^t \in \mathcal{S}$ is the discretization of the observed information from the environment at the current time slot t. The system state s^t includes the previous jamming power, i.e., $\{P_{\mathrm{J},k}^{t-1}\}_{k\in\mathcal{K}}$ according the channel quality, the previous UEs' SINR values $\{SINR_k^{t-1}\}_{k\in\mathcal{K}}$, as well as the current estimated channel coefficients $\{\mathbf{g}_k^t\}_{k\in\mathcal{K}}$, which is defined as

$$s^t = \left\{\{P_{\mathrm{J},k}^{t-1}\}_{k\in\mathcal{K}}, \{\mathbf{g}_k^t\}_{k\in\mathcal{K}}, \{SINR_k^{t-1}\}_{k\in\mathcal{K}}\right\}. \quad (5)$$

Actions: The action $a^t \in \mathcal{A}$ is one of the valid selections that the learning agent chooses at the time slot t, which includes the transmit power $\{P_k\}_{k\in\mathcal{K}}$ and the reflecting beamforming coefficient (phase shift) $\{\theta_m\}_{m\in\mathcal{M}}$. Hence, the action a^t is given by

$$a^t = \left\{\{P_k^t\}_{k\in\mathcal{K}}, \{\theta_m^t\}_{m\in\mathcal{M}}\right\}. \quad (6)$$

Transition probability: $\mathcal{P}(\cdot)$ is a transition model which represents the probability of taking an action a at a current state s and then ending up in the next state s', i.e., $\mathcal{P}(s'|s,a)$.

Policy: Let $\pi(\cdot)$ denotes a policy and it maps the current system state to a probability distribution over the available actions which is taken by the agent, i.e., $\pi(a,s): S \rightarrow \mathcal{A}$.

Reward function: The reward function design plays an important role in the policy learning in RL, where the reward signal correlates with the desired goal of the system performance. In the optimization problem considered in Section II.B, our objectives are twofold: maximizing the UEs' achievable rate while decreasing the power consumption at the BS as much as possible.

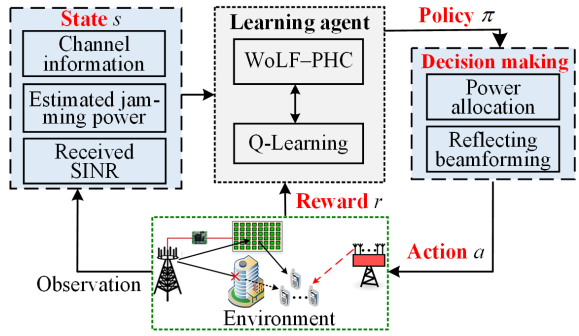

Fig. 2. WoLF-PHC-based anti-jamming policy for IRS-assisted systems.

Based on the above analysis, the reward function is set as

$$r = \underbrace{\sum_{k\in\mathcal{K}} \log_2\left(1 + SINR_k\right)}_{\text{part 1}} - \underbrace{\lambda_1 \sum_{k\in\mathcal{K}} P_k}_{\text{part 2}} \quad (7)$$

In (7), the part 1 represents the immediate utility (system achievable rate), the part 2 is the cost functions which is defined as the transmission cost of the power consumption at the BS, with λ_1 being the corresponding coefficient.

B. WoLF-PHC-Based Joint Power Allocation and Reflecting Beamforming

Most of existing RL algorithms are value-based RL, such as Q-Learning, Deep Q-Network (DQN) and double DQN. These RL algorithms can estimate the Q-function with low variance as well as adequate exploration of action space, which can be ensured by using the greedy scheme. In addition, policy gradient based RL algorithm has the ability to tackle the continuous action space optimization problems, but it may converge to suboptimal solutions [22].

In order to obtain the optimal anti-jamming policy against smart jamming, we propose a fast WoLF-PHC-based joint power allocation and reflecting beamforming for IRS-assisted communication systems, as shown in Fig. 2, where WoLF-PHC is utilized to enable the learning agent to learn and adapt faster in dynamic uncertain environments. In the IRS-assisted system, the learning agent observes a system state and receives an instantaneous reward by interacting with the environment. Then, such information is leveraged to train the learning model to choose the anti-jamming policy with the maximum Q-function value. After that, according to the selected policy, the action is chosen to make decision in terms of power allocation and reflecting beamforming.

The objective of the learning agent is to obtain an optimal policy that optimizes the long-term cumulative discounted reward instead of its immediate reward, which can be expressed as $R_t = \sum_{j=0}^{\infty} \gamma^j r^{(t+j+1)}$, where $\gamma \in (0,1]$ denotes the discount factor. Adopting $Q^\pi(s^t,a^t)$ as the state-action value function, which represents the value of executing an action a in a state s under a policy π, it can be expressed as

$$Q^\pi(s^t,a^t) = E_\pi\left[\sum_{j=0}^{\infty} \gamma^j r^{(t+j+1)} | s^t = s, a^t = a\right]. \quad (8)$$

Similar to [23], the state-action Q-function $Q^{\pi}(s^t, a^t)$ satisfies the Bellman equation which is expressed as

$$Q^{\pi}(s^t, a^t) = E_{\pi}\Bigg[r^{t+1} + \gamma \sum_{s^{t+1}\in\mathcal{S}} P(s^{t+1}|s^t, a^t) \sum_{a^{t+1}\in\mathcal{A}} \pi(s^{t+1}, a^{t+1}) Q^{\pi}(s^{t+1}, a^{t+1})\Bigg] \tag{9}$$

The conventional Q-Learning algorithm is widely utilzied to search the optimal policy π^*. From (9), the optimal Q-function (Bellman optimality equation) associated with the optimal policy has the following form

$$Q^*(s^t, a^t) = r^{t+1} + \gamma \sum_{s^{t+1}\in\mathcal{S}} \mathcal{P}(s^{t+1}|s^t, a^t) \max_{a^{t+1}\in\mathcal{A}} Q^*(s^{t+1}, a^{t+1}). \tag{10}$$

It is worth noting that the Bellman optimality equation generally does not have any closed-form solution. Thus, the optimal Q-function (10) can be solved recursively to achieve the optimal $Q^*(s^t, a^t)$ by using an iterative method. Accordingly, the updating on the state-action value function $Q(s^t, a^t)$ is expressed as

$$\begin{aligned} Q(s^t, a^t) \leftarrow (1-\alpha)Q(s^t, a^t) \\ +\alpha\left(r^t + \gamma \max_{a^t\in\mathcal{A}} Q^*(s^{t+1}, a^t)\right) \end{aligned} \tag{11}$$

where $\alpha \in (0, 1]$ stands for the learning rate for the update of Q-function.

The $\varepsilon-$greedy policy is capable of balancing the tradeoff between an exploitation and an exploration in an RL, in order to avoid converging to local optimal power allocation and reflecting beamforming strategy. In the $\varepsilon-$greedy policy, the agent selects 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 [23]. Hence, the action selection probability of the learning agent is expressed as

$$\Pr(a = \tilde{a}) = \begin{cases} 1-\varepsilon, & \tilde{a} = \arg\max_{a\in\mathcal{A}} Q(s, a), \\ \frac{\varepsilon}{|\mathcal{A}|-1}, & \tilde{a} \neq \arg\max_{a\in\mathcal{A}} Q(s, a). \end{cases} \tag{12}$$

As the WoLF-PHC algorithm is capable of not only keeping the Q-function but also quickly learning the decision-making policy under uncertain characteristics [24], so this paper adopts it to derive the optical power allocation and reflecting beamforming strategy with the unknown jamming model.

In WoLF-PHC, the mixed policy $\pi(s, a)$ is updated by increasing the probability that it selects the most valuable action with the highest Q-function value by a learning rate $\xi \in (0, 1]$, and reducing other probabilities by $-\xi/(|\mathcal{A}|-1)$, i.e.,

$$\pi(s, a) \leftarrow \pi(s, a) + \begin{cases} \xi, \text{ if } a = \arg\max_{a'} Q(s, a'), \\ -\frac{\xi}{|\mathcal{A}|-1}, \text{ otherwise.} \end{cases} \tag{13}$$

The WoLF-PHC-based joint power allocation and reflecting beamforming approach for the IRS-assisted communication system against smart jamming is summarized in **Algorithm 1**. At each episode training step, the learning agent observes its system state s^t (i.e., the estimated jamming power, SINR values, and channel coefficients) by interacting with the environment. At each learning time slot t, the joint action a^t (i.e., power allocation and reflecting beamforming) is selected by using the probability distribution $\pi(s^t, a^t)$. The ε-greedy policy method is employed to balance the exploration and the exploitation, for example, the action with the maximum Q-function value is chosen with probability $1-\varepsilon$ according to the known knowledge, while a random action is chosen with probability ε based on the unknown knowledge. After executing the selected action a^t, the environment will feedback a reward $r(s^t, a^t)$ and a new system state s^{t+1} to the learning agent. Then, the WoLF-PHC algorithm updates both the current policy $\pi(s^t, a^t)$ and updates the variable learning rate ξ to improve the learning rate. Finally, the learning model is trained successfully, and it can be loaded to search the joint power allocation $\{P_k\}_{k\in\mathcal{K}}$ and reflecting beamforming matrix $\mathbf{\Phi}$ strategies according to the selected action.

Algorithm 1 WoLF-PHC-Based Joint Power Allocation and Reflecting Beamforming

Input: WoLF-PHC learning structure and IRS-assisted system with a jammer.
Initialize: $Q(s, a) = 0$, $\pi(s, a) = 1/|\mathcal{A}|$, ξ, γ, and α.
for each episode $j = 1, 2, \ldots, N^{\text{epi}}$ **do**
  for each time step $t = 0, 1, 2, \ldots, T$ **do**
    Observe an initial system state s^t;
    Select an action a^t based on the ε-greedy policy:
      $a^t = \arg\max_{a^t\in\mathcal{A}} Q(s^t, a^t)$, with probability 1-$\varepsilon$;
      $a^t = \text{random}\{a_i\}_{a_i\in\mathcal{A}}$, with probability ε;
    Execute the exploration action a^t, receive a reward $r(s^t, a^t)$ and the next state s^{t+1};
    Update $Q(s^t, a^t)$ by via (11);
    Update the current policy $\pi(s^t, a^t)$;
  end for
end for
Return: WoLF-PHC-based learning model;
Output: Load the learning model to achieve the joint power allocation and reflecting beamforming matrix strategy.

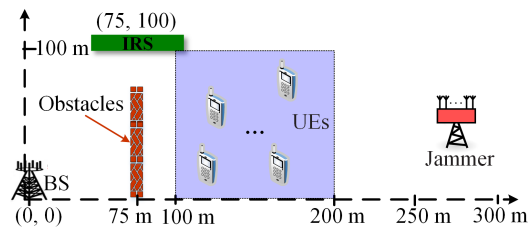

Fig. 3. Simulation setup.

IV. Simulation Results and Analysis

This section evaluates the performance of the IRS-assisted communication system against smart jamming shown in Fig. 3, where a number of single-antenna UEs are randomly located in the 100 m $\times$ 100 m right-hand side rectangular area (light blue area). The locations of the BS, the IRS, and the jammer are (0, 0), (75, 100), and (0, 0) in meter (m), respectively. There exists obstacles which block the direct communication links from the BS to the UEs, so the obstacles cause the large-scale pathloss for the communication links.

As for the communication channel coefficients, the path loss in dB is expressed as

$$PL = (PL_0 - 10\beta \log_{10}(d/d_0)) \tag{14}$$

where PL_0 denotes the path loss at the reference distance d_0, β is the path loss exponent, and d is the distance from the transmitter to the receiver, respectively. Here, we use $\beta_{\rm bu}$, $\beta_{\rm br}$, $\beta_{\rm ru}$, and $\beta_{\rm ju}$ to denote the path loss exponents of the channel links between the BS and the UEs, between the BS and the IRS, between the IRS and the UEs, and between the jammer and the UEs, respectively. According to [18]-[22], we set $PL_0 = 30$ dB, $d_0 = 1$ m, $\beta_{\rm bu} = 3.75$, $\beta_{\rm br} = \beta_{\rm ru} = 2.2$ and $\beta_{\rm ju} = 2.5$. We set that the background noise at all UEs is equal to $\delta^2 = -100$ dBm. The number of antennas at the BS and the jammer are set to $N = N_{\rm J} = 8$. The maximum transmit power $P_{\max}$ at the BS varies from 15 dBm to 40 dBm, and the number of IRS elements M varies from 20 to 100 for different simulation settings. In addition, the jamming power of the smart jammer ranges from 15 dBm to 40 dBm according to its jamming behavior, and the BS cannot know the current jamming power levels, but it can estimate the previous jamming power levels according to the historical channel quality. The learning rate is set to $\alpha = 0.5 \times 10^{-2}$, the discount factor is set to $\gamma = 0.9$ and the final exploration rate is set to $\varepsilon = 0.1$. The cost parameter λ_1 in (7) is set to $\lambda_1 = 1$. We set $\xi = 0.04$ [23], [24]. In addition, we compare the proposed WoLF-PHC-based joint power allocation and reflecting beamforming approach with the following approaches:

- The popular fast Q-Learning approach [13], which is adopted to optimize the transmit power allocation and reflecting beamforming in IRS-assisted communication systems (denoted as fast Q-Learning [13]).
- The greedy approach which jointly optimizes the BS's transmit power allocation and the IRS's reflect beamforming (denoted as Greedy).
- The optimal transmit power allocation at the BS without IRS assistance (denoted as Optimal PA without IRS).

We first compare the convergence performance of all approaches when $P_{\max} = 30$ dBm, $K = 4$, and $M = 60$. It is observed that the system rate of all approaches (except the optimal PA approach) increases with the number of iterations, and the proposed WoLF-PHC learning approach accelerates the convergence rate and enhances the system rate compared with both the fast Q-Learning approach and the greedy approach. Because the proposed leaning approach adopts WoLF-PHC to increase the learning rate and enhance the learning efficiency against smart jamming, yielding a faster learning rate under the dynamic environment. Among all approaches, the fast Q-Learning requires the largest number of convergence iterations to optimize the Q-function estimator, where the slow convergence may fail to protect anti-jamming performance against smart jamming in real-time systems. Moreover, the optimal PA approach without IRS has the fastest convergence speed, but it obtains the worst performance among all approaches, because it does not employ an IRS for system performance improvement and jamming resistance.

The average system rate versus the maximum transmit power $P_{\max}$ for various approaches are shown in Fig. 5 when $K = 4$, and $M = 60$, which demonstrates that the

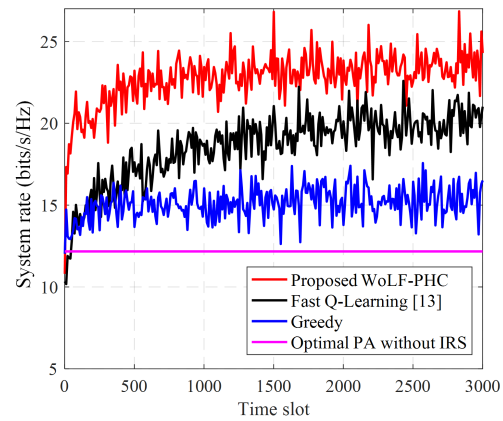

Fig. 4. Convergence behaviors of the various approaches.

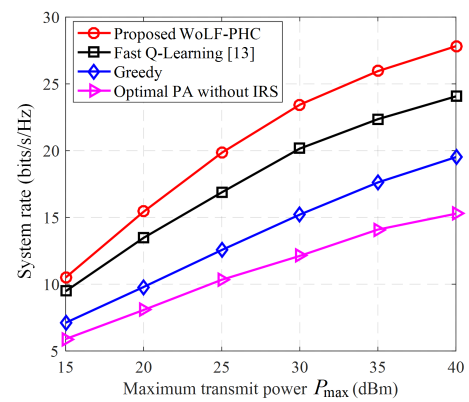

Fig. 5. Performance comparisons versus the maximum transmit power $P_{\max}$.

achieved system rate improve as $P_{\max}$ increases. We can also observe that both the proposed learning approach and the fast Q-Learning approach have good system rate value under different values of $P_{\max}$, and both of them greatly outperform other approaches. Additionally, the performance improvement achieved by using IRS versus without IRS increases with $P_{\max}$, which indicates the advantage of deploying the IRS against smart jamming. In addition, the performance of both the system rate of the proposed WoLF-PHC-based learning approach is higher than that of the fast Q-Learning approach, which is due to the fact that WoLF-PHC is adopted to effectively search the optimal joint power allocation and reflecting beamforming strategy against smart jamming in dynamic uncertain environments.

Fig. 6 compares the performance of the four approaches with the different reflecting elements number M when $P_{\max} = 30$ dBm and $K = 4$. It can be seen that except the optimal PA approach without IRS, the performance of all IRS-based approaches increases with M, and greatly outperforms the optimal PA approach without IRS. This is that the IRS has the ability to support higher degrees of freedom for performance optimization, resulting in the great performance gains obtained by employing the IRS against smart jamming over the traditional system without IRS. Specifically, when $M = 20$, the system achievable rate gain of the proposed learning approach over the optimal PA approach without IRS is only about 4.21 bits/s/Hz, while this value is improved to 15.47 bits/s/Hz when

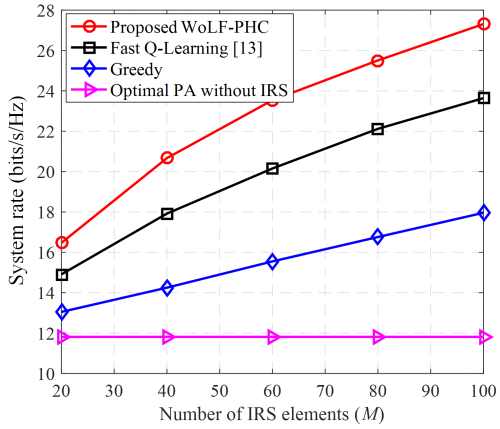

Fig. 6. Performance comparisons versus the number of IRS elements.

$M = 100$. Such performance improvement results from the facts that the more power can be achieved at the IRS by increasing M, and the higher reflecting beamforming gain is achieved to design the IRS phase shifts to improve the received desired signal as well as mitigate the jamming interference from the smart jammer by increasing M.

V. Conclusions

This paper proposed to improve the anti-jamming performance of wireless communication systems by employing an IRS. Specifically, we formulated an optimization problem by joint optimizing both the transmit power allocation at the BS and the reflecting beamforming at the IRS. A WoLF-PHC learning approach was proposed to achieve the optimal anti-jamming strategy, where WoLF-PHC is capable of quickly achieving the optimal policy without knowing the jamming model. Simulation results confirmed that the IRS-assisted system significantly improves the anti-jamming performance compared with other approaches. We will pay attention to apply IRS in visible light communication systems in the future [25], [26].

VI. Acknowledgements

This work is support by Nanyang Technological University (NTU) Startup Grant, Alibaba-NTU Singapore Joint Research Institute , 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 under its Strategic Capability Research Centres Funding Initiative: Strategic Centre for Research in Privacy-Preserving Technologies & Systems, Energy Research Institute @NTU , Singapore NRF National Satellite of Excellence, Design Science and Technology for Secure Critical Infrastructure NSoE DeST-SCI2019-0012, AI Singapore 100 Experiments (100E) programme, and NTU Project for Large Vertical Take-Off & Landing Research Platform..

References

[1] A. Mukherjee, S. A. A. Fakoorian, J. Huang, and A. L. Swindelhurst, "Principles of physical layer security in multiuser wireless networks: a survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1550-1573, third quarter, 2014.

[2] 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.

[3] L. Xiao, J. Liu, Q. Li, N. B. Mandayam, and H. V. Poor, "User-centric view of jamming games in cognitive radio networks," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 12, pp. 2578-2590, Dec. 2015.

[4] L. Liang, W. Cheng, W. Zhang, and H. Zhang, "Mode hopping for anti-jamming in radio vortex wireless communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7018-7032, Aug. 2018.

[5] Y. Gao, Y. Xiao, M. Wu, M. Xiao, and J. Shao, "Game theory-based anti-jamming strategies for frequency hopping wireless communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5314-5326, Aug. 2018.

[6] M. K. Hanawal, M. J. Abdel-Rahman, and M. Krunz, "Joint adaptation of frequency hopping and transmission rate for anti-jamming wireless systems," *IEEE Trans. Mobile Comput.*, vol. 15, no. 9, pp. 2247-2259, Sept. 2016.

[7] S. D¡Oro, E. Ekici and S. Palazzo, "Optimal power allocation and scheduling under jamming attacks," *IEEE/ACM Trans. Net.*, vol. 25, no. 3, pp. 1310-1323, Jun. 2017.

[8] Z. Dou, G. Si, Y. Lin, and M. Wang, "An adaptive resource allocation model with anti-jamming in IoT network," *IEEE Access*, vol. 7, pp. 93250-93258, May 2019.

[9] L. Xiao, T. Chen, J. Liu, and H. Dai, "Anti-jamming transmission stackelberg game with observation errors," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 949-952, Jun. 2015.

[10] P. Gu, C. Hua, W. Xu, R. Khatoun, Y. Wu, and A. Serhrouchni, "Control channel anti-jamming in vehicular networks via cooperative relay beamforming," Appear in *IEEE Internet of Things J.*. Doi: 10.1109/JIOT.2020.2973753.

[11] C. Wang and H. M. Wang, "Robust joint beamforming and jamming for secure AF networks: low-complexity design," *IEEE Trans. Veh. Technol.*, vol. 64, no. 5, pp. 2192C2198, May 2015.

[12] M. A. Maleki Sadr, M. Ahmadian Attari, and R. Amiri, "Robust relay beamforming against jamming attack," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 312-315, Feb. 2018.

[13] Z. Xiao, B. Gao, S. Liu, and L. Xiao, "Learning based power control for mmWave massive MIMO against jamming," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, 2018, pp. 1-6.

[14] N. Van Huynh, D. N. Nguyen, D. T. Hoang, and E. Dutkiewicz, "Jam me if you can: defeating jammer with deep dueling neural network architecture and ambient backscattering augmented communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2603-2620, Nov. 2019.

[15] S. Liu *et al.*, "Pattern-aware intelligent anti-jamming communication: a sequential deep reinforcement learning approach" *IEEE Access*, vol. 7, pp. 169204-169216, 2019.

[16] 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.

[17] 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.

[18] Z. Chu, W. Hao, P. Xiao, and J. Shi, "Intelligent reflecting surface aided multi-antenna secure transmission," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 108-112, Jan. 2020.

[19] L. Dong and H. Wang, "Secure MIMO transmission via intelligent reflecting surface," Appear in *IEEE Wireless Commun. Lett.*. Doi: 10.1109/LWC.2020.2969664.

[20] 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.

[21] 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, Sept. 2019.

[22] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning based intelligent reflecting surface for secure wireless communications," 2020. [Online]. Available: https://arxiv.org/abs/2002.12271.

[23] M. Wiering and M. Otterlo, Reinforcement learning: Stateof-the-art, Springer Publishing Company, Incorporated, 2014.

[24] M. Bowling and M. Veloso, "Rational and convergent learning in stochastic games,¡ in *Proc. Int Joint Conf. Artificial Intell.*, Seattle, WA, 2001, pp. 1021C1026.

[25] 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.

[26] 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.