Intelligent Reflecting Surface Assisted Anti-Jamming Communications: A Fast Reinforcement Learning Approach

Helin Yang, *Member, IEEE*, Zehui Xiong, *Member, IEEE*, Jun Zhao, *Member, IEEE*, Dusit Niyato, *Fellow, IEEE*, Qingqing Wu, *Member, IEEE*, H. Vincent Poor, *Fellow, IEEE*, and Massimo Tornatore, *Senior Member, IEEE*

Abstract—Malicious jamming launched by smart jammers can attack legitimate transmissions, which has been regarded as one of the critical security challenges in wireless communications. With this focus, this paper considers the use of an 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 a smart jammer, an optimization problem for jointly optimizing power allocation at the base station (BS) and reflecting beamforming at the IRS is formulated while considering quality of service (QoS) requirements of legitimate users. As the jamming model and jamming behavior are dynamic and unknown, a fuzzy win or learn fast-policy hill-climbing (WoLFPHC) learning approach is proposed to jointly optimize the anti-jamming power allocation and reflecting beamforming strategy, where WoLFPHC is capable of quickly achieving the optimal policy without the knowledge of the jamming model, and fuzzy state aggregation can

represent the uncertain environment states as aggregate states. Simulation results demonstrate that the proposed anti-jamming learning-based approach can efficiently improve both 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 open nature of wireless channels [1], [2], wireless transmissions are 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 serious threat in wireless communications. Jamming-related investigations not only provide solutions into wireless security guarantees against jamming, but also offer insights on the vulnerabilities of existing systems. In this regard, lots of technologies have been recently presented to defend wireless security against jamming attacks, including frequency hopping, power control, relay assistance, beamforming, and so on.

Frequency-hopping (FH) is a powerful and widely-adopted techniques where a wireless user is allowed to quickly switch its current operating frequency to other frequency spectrum, thereby avoiding potential jamming attacks [4]-[6]. In [4], a mode-FH approach was presented to jointly utilize mode hopping and conventional FH to decrease bit error rate in the presence of jammers. A stochastic game was developed to study the interaction process between jammer and legitimate user [5], where the jammer and the transmitter can act as two contending players to achieve the optimal attack and defense policies, respectively. In [6], Hanawal *et al.* proposed a joint FH and rate-adaptation scheme to avoid jamming attacks in the presence of a jammer, where the transmitter has the ability to escape the jammer by switching its operating channel, adjusting its rate, or both. However, FH may become ineffective if smart jammers attack multiple channels simultaneously and it also requires extra spectrum resources to evade jammers. Besides FH, power control is another commonly used technique [3], [7]-[12]. As an example, references [7] and [8] investigated a jammed wireless system where the system operator tries to control the transmit power to maximize system rate while guaranteeing quality-of-service (QoS) requirements of legitimate receivers. The authors in

[3], [9]-[11] developed the novel power control strategies for the anti-jamming problem. Specifically, they leveraged game theory to optimize the power control policy of the transmitter against jammers, and simulations were provided to verify the effectiveness of the proposed anti-jamming power control approaches. In addition, a jamming-resistant receiver was designed to improve the robustness of communication system against jamming in [12], and the optimal power control scheme was also developed to improve the achievable rate.

Recently, cooperative communication using trusted relays has been proposed as one promising anti-jamming technique for improving the physical layer security [13]-[17]. To reduce the adverse effects of jamming signals, a joint relay selection and beamforming problem was formulated in relay-aided systems [13], which was solved by applying semi-determined relaxation (SDR) technique. In [14] and [15], the joint cooperative beamforming and jamming designs considering worst-case robust schemes were proposed to maximize the achievable rate under the imperfect channel state information (CSI) of a jammer. In [16] and [17], the advanced anti-jamming schemes were proposed to cancel strong jamming signal, with the goal to maximize the signal-to-interference-plus-noise-ratio (SINR) of legitimate users.

To deal with uncertain and/or unknown jamming attack models, such as jamming policies and jamming power levels, reinforcement learning (RL) algorithms have been applied in some existing studies to optimize the jamming resistance policy in dynamic wireless communication systems [18]-[25]. In [18], a policy hill climbing (PHC)-based Q-Learning approach was developed to improve the vehicular communication performance against jamming without knowing the jamming model. Additionally, a fast PHC-based power control algorithm was also presented to assist base station (BS) to select anti-jamming transmit power strategy over multiple antennas [19]. As traditional Q-learning algorithm has slow convergence to the optimal policy, in [20] and [21], the authors adopted deep reinforcement learning (DRL) algorithms that enable transmitters to quickly obtain an optimal communication policy to guarantee security performance against smart jamming without the need of knowledge of the jammer's characteristics.

However, despite the effectiveness of the above aforementioned anti-jamming schemes [3]-[21], employing a number of active relays incurs an excessive hardware cost and system complexity. Moreover, anti-jamming beamforming and power control in communication systems are generally energy-consuming as more transmit power as well as circuit power need to be consumed to improve the communication performance. To tackle these shortcomings, a new paradigm, called intelligent reflecting surface (IRS) [26]-[39], has been recently proposed as a promising technique to enhance spectrum efficiency and secrecy performance in the fifth-generation (5G) and beyond communication systems. In particular, IRS is a uniform planar array comprising 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, thus enhancing and/or weakening the received signals at different users [26]-[36]. As a result, IRS has been employed

in wireless commutation systems for security performance optimization [37]-[43]. In [37]-[43], the authors investigated the physical layer security enhancement of IRS-assisted communications systems, where both the BSs beamforming and the IRSs phase shifts were jointly optimized to improve secrecy rate in the presence of an eavesdropper. Additionally, two algorithms, called alternative optimization (AO) algorithm and SDP relaxation algorithm were applied to address the joint optimization problem in [37]-[40], and simulation results demonstrated the effectiveness of the presented IRS-assisted system compared with traditional systems. Furthermore, Yang *et al.* in [42] 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 [4]-[43] to enhance the anti-jamming strategy against smart jamming, where the smart jammer attempts to deteriorate the quality of intended transmissions by transmitting jamming signal over the legitimate channels.

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, while guaranteeing the QoS requirements of users against smart jamming. As the communication system environment is complex and the jammer is smart, a fuzzy RL based anti-jamming approach is proposed to effectively jointly optimize the anti-jamming power allocation and reflecting beamforming in uncertain environments. The main contributions are summarized as follows.

- We, for the first time, propose a novel anti-jamming model based on the IRS-assisted system that jointly optimizes the transmit power allocation at the BS and the reflecting beamforming at the IRS to improve secrecy rate against smart jamming.
- Aiming to enhance the anti-jamming performance, an optimization problem for jointly optimizing power allocation and reflecting beamforming is formulated given QoS requirements of legitimate users. Since it is difficult to address the non-convex optimization problem in dynamic environments, the problem is modelled as an RL process.
- As the jamming model and jamming behavior are dynamic and unknown, a fuzzy win or learn fast-policy hill-climbing (WoLFCPHC) learning 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, and fuzzy state aggregation (FSA) has the ability to represent the uncertain environment states as aggregate states.
- Simulation results are provided to validate that the IRS-assisted system significantly improves the anti-jamming communication performance compared to conventional systems without IRS, and also to verify the effectiveness of the proposed learning approach in terms of improving the system rate and service protection level, compared with the existing approaches. For example, the proposed
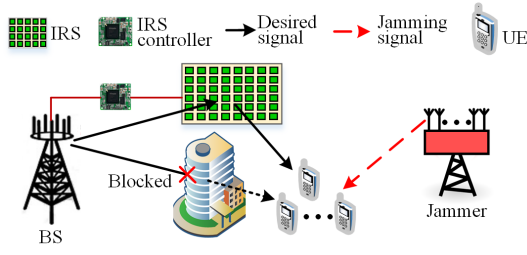
Fig. 1. Illustration of an IRS-assisted communication system against a multi-antenna jammer.

learning approach achieves the system rate and service protection level improvements of 21.29% and 13.36%, respectively, over the existing solution in [39].

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

Notations: In this paper, vectors and matrices are represented by Boldface lowercase and uppercase letters, respectively. $\text{Tr}(\cdot)$, $(\cdot)^*$ and $(\cdot)^H$ respectively stand for the trace, the conjugate and the conjugate transpose operations. $|\cdot|$ and $||\cdot||$ denote the absolute value of a scalar and the Euclidean norm of a vector or matrix, respectively. $\mathbb{C}^{M \times N}$ is the space of $M \times N$ complex-valued matrices. $\mathbb{E}[\cdot]$ denotes the statistical expectation operation.

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 are located in dead zones. In addition, as illustrated in Fig. 1, a malicious (or unfriendly) multi-antenna jammer is located near the UEs who attempts to interfere the transmissions by sending faked or replayed jamming signal for the UEs via N_J antennas, in order to degrade the 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 $\mathbf{G} \in \mathbb{C}^{M \times N}$, $\mathbf{g}_{\text{bu},k}^H \in \mathbb{C}^{1 \times N}$, $\mathbf{g}_{\text{ru},k}^H \in \mathbb{C}^{1 \times M}$, and $\mathbf{h}_{\text{J},k}^H \in \mathbb{C}^{1 \times N_J}$ 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 $\mathbf{\Phi} = \text{diag}(\Phi_1, \Phi_2, ..., \Phi_M) \in \mathbb{C}^{M \times M}$ denote the reflection coefficient matrix associated with effective phase shifts at the IRS, where $\Phi_m = \omega_m e^{j\theta_m}$ comprises both a 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 [26]-[29].

The transmitted signal at the BS can be expressed as

$$\mathbf{x} = \sum_{k=1}^{K} \sqrt{P_k}\mathbf{w}_k s_k \quad (1)$$

where P_k stands for the transmit power allocated for the k-th UE, and we have the power constraint: $\sum_{k=1}^{K} P_k \leq P_{\max}$ 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 study considers the case that the smart jammer attempts to disturb the BS's transmitted signal by emitting jamming signal $\mathbf{z}_k \in \mathbb{C}^{N_J \times 1}$ at the k-th UE. In addition, the transmit power of the faked jamming signal for the k-th UE is denoted as $P_{\text{J},k} = \|\mathbf{z}_k\|^2 = \text{Tr}(\mathbf{z}_k \mathbf{z}_k^H)$. 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

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

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 SINR at the k-th UE can be expressed as

$$SINR_k =$$

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

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 and the minimum received SINR constraints.

Accordingly, the optimization problem can be formulated as

$$\max_{\{P_k\}_{k\in\mathcal{K}},\Phi} \sum_{k\in\mathcal{K}} \log_2\left(1 + SINR_k\right)$$
$$s.t. \ (a): \ \sum_{k=1}^{K} P_k \leq P_{\max},$$
$$(b): \ SINR_k \geq SINR_k^{\min}, \ \forall k \in \mathcal{K}, \quad (4)$$
$$(c): \ |\Phi_m| = 1, \ 0 \leq \theta_m \leq 2\pi, \ \forall m \in \mathcal{M}$$

where $SINR_k^{\min}$ denotes the minimum SINR threshold of the k-th UE. Note that problem (4) is a non-convex optimization problem, where the objective function is non-concave over the reflecting beamforming matrix Φ; furthermore, the transmit power allocation variables $\{P_k\}_{k\in\mathcal{K}}$ and Φ are intricately coupled in the objective function, thus rendering the joint optimization problem difficult to be solved optimally. So far, many optimization algorithms (e.g., AO and SDR) have been proposed to obtain an approximate solution to problem (4), by iteratively updating either $\{P_k\}_{k\in\mathcal{K}}$ or Φ 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 study 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 output SINR in (3) to make an easy presentation [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. The traditional optimization algorithms (e.g., AO and SDR) are capable of addressing one single time slot optimization problem, but they may achieve the suboptimal solution and achieve the greedy-search like performance as the historical system information and the long term benefit are ignored. In addition, the unknown jamming model and channel variation result in dynamic open and uncertain characteristics, which increases difficulty of solving the optimization problem.

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 [44]. Thus, this paper models the optimization problem as an RL, and a fuzzy 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 acts 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., $\{P_{J,k}^{t-1}\}_{k\in\mathcal{K}}$ 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 $\{\mathbf{g}_k^t\}_{k\in\mathcal{K}}$, and it is defined as

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

Actions: The action $a^t \in \mathcal{A}$ is one of the valid selections that the learning agent chooses at the time slot t, and it 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 $\pi(\cdot)$ denote a policy that it maps the current system state to a probability distribution over the available actions which can be taken by the agent, i.e., $\pi(a,s): S \to \mathcal{A}$.

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 threefold: maximizing the UEs' achievable rate while decreasing the power consumption at the BS as much as possible and guaranteeing the SINR constraints against smart jamming.

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

$$r = \underbrace{\sum_{k\in\mathcal{K}} \log_2\left(1 + SINR_k\right)}_{\text{part 1}} - \underbrace{\lambda_1 \sum_{k\in\mathcal{K}} P_k}_{\text{part 2}} - \underbrace{\lambda_2 \sum_{k\in\mathcal{K}} SINR_k^{\text{out}}}_{\text{part 3}} \quad (7)$$

where

$$SINR_k^{\text{out}} = \begin{cases} 0, & \text{if } SINR_k \geq SINR_k^{\min}, \forall k \in \mathcal{K}, \\ 1, & \text{otherwise.} \end{cases} \quad (8)$$

In (7), the part 1 represents the immediate utility (system achievable rate), the part 2 and part 3 are the cost functions which are defined as the transmission cost of the power consumption at the BS and the violation of minimum SINR requirements, respectively, with λ_1 and λ_2 being the corresponding coefficient. The goal of (8) is to impose the SINR protection level, where the cost is zero (i.e., $SINR_k^{\text{out}} = 0$) when the SINR constraint is guaranteed against jamming, otherwise, $SINR_k^{\text{out}} = 1$.

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

$$Q^\pi(s^t, a^t) = E_\pi\left[\sum_{j=0}^{\infty} \gamma^j r^{(t+j+1)} | s^t = s, a^t = a \right]. \quad (9)$$

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

Note that similar to [44], the state-action Q-function $Q^\pi(s^t, a^t)$ satisfies the Bellman equation which is expressed as

$$Q^\pi(s^t, a^t) = E_\pi \left[r^{t+1} + \gamma \sum_{s^{t+1} \in \mathcal{S}} \mathcal{P}(s^{t+1}|s^t, a^t) \right.$$
$$\left. \sum_{a^{t+1} \in \mathcal{A}} \pi(s^{t+1}, a^{t+1}) Q^\pi(s^{t+1}, a^{t+1}) \right] \quad (10)$$

The conventional Q-Learning algorithm is widely utilized to search the optimal policy π^*. From (10), the optimal Q-function (Bellman optimality equation) associated with the optimal policy has the following form

$$Q^*(s^t, a^t)$$
$$= r^{t+1} + \gamma \sum_{s^{t+1} \in \mathcal{S}} \mathcal{P}(s^{t+1}|s^t, a^t) \max_{a^{t+1} \in \mathcal{A}} Q^*(s^{t+1}, a^{t+1}) \quad (11)$$

It is worth noting that the Bellman optimality equation generally does not have any closed-form solution. Thus, the optimal Q-function (11) 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

$$Q(s^t, a^t) \leftarrow (1-\alpha)Q(s^t, a^t)$$
$$+ \alpha \left(r^t + \gamma \max_{a^t \in \mathcal{A}} Q^*(s^{t+1}, a^t) \right) \quad (12)$$

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

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

Most of existing RL algorithms are value-based RL, such as Q-Leaning, 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 [44]. However, it is not easy to achieve the optimal anti-jamming policy without knowing the jamming model and jamming strategy.

In order to obtain the optimal anti-jamming policy against smart jamming, we propose a fast fuzzy 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, and FSA is used to enable Q-Learning to represent the system states with a fixed number of aggregate states and represent continuous state spaces as discrete. 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 procedures of the proposed learning based decision making approach are provided in the following analysis.

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 [44], [45]. Hence, the action selection probability of the learning agent is expressed as

$$\Pr(a = \tilde{a}) = \begin{cases} 1 - \varepsilon, & \tilde{a} = \arg\max_{a \in \mathcal{A}} Q(s, a), \\ \frac{\varepsilon}{|\mathcal{A}|-1}, & \tilde{a} \neq \arg\max_{a \in \mathcal{A}} Q(s, a). \end{cases} \quad (13)$$

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 [46], so this study adopts it to derive the optical power allocation and reflecting beamforming strategy with the unknown jamming model. In the IRS-assisted communication system, the jammer attempts to disturb the communication performance of legitimate UEs by exploiting the action of the learning agent, the WoLF-PHC algorithm can provide uncertainty in the action selection and fools the jammer's attacks in the presence of 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.,

$$\pi(s, a) \leftarrow \pi(s, a)$$
$$+ \begin{cases} \xi, & \text{if } a = \arg\max_{a'} Q(s, a'), \\ -\frac{\xi}{|\mathcal{A}|-1}, & \text{otherwise.} \end{cases} \quad (14)$$

In (14), the WoLF-PHC algorithm has two variable learning rates, i.e., ξ_{win} and ξ_{loss}, with $\xi_{\text{loss}} > \xi_{\text{win}}$ [46]. These two learning rates are employed to improve the learning policy depending on whether the learning agent wins or loses at the current learning step. The expected value is generally used to take this determination, in other words, whether the expected value of the current policy $\pi(s, a)$ is larger than the expected

value of the average policy $\widehat{\pi}(s,a)$. If the current expected value is larger (i.e. the agent is "winning"), then the smaller learning rate ξ_{win} is chosen; otherwise, ξ_{loss} is selected. The corresponding updating rule is given by

$$\xi = \begin{cases} \xi_{\text{win}}, & \text{if } \sum_a \pi(s,a)Q(s,a) > \sum_a \widehat{\pi}(s,a)Q(s,a), \\ \xi_{\text{loss}}, & \text{otherwise.} \end{cases} \tag{15}$$

When the selected action a is executed, the average policy $\widehat{\pi}(s,a)$ of all actions is updated under the given state s as

$$\widehat{\pi}(s,a) \leftarrow \widehat{\pi}(s,a) + \frac{\pi(s,a) - \widehat{\pi}(s,a)}{C(s)} \tag{16}$$

where $C(s)$ denotes the number of the state s from the initial state to the current state.

In practical communication systems, it is difficult to know the accurate jamming model and behaviors of the smart jammer, so the possible estimated state space related to the jamming behaviors increases. Moreover, the uncertain jamming behaviors and mobility of the jammer result in an uncertain dynamic during the learning process. Hence, this paper combines fuzzy state aggregation (FSA) into the WoLF-PHC algorithm to represent the system states with a fixed number of FSA states and thus reduce the number of system states that the WoLF-PHC algorithm must deal with. FSA can transform continuous states as discrete [47], hence enabling the use of the WoLF-PHC algorithm in continuous state spaces. Thus, FSA is combined with the WoLF-PHC algorithm to formulate a fuzzy approximation architecture, where the fuzzy state-action value function under the state-action pair (s, a) is given by

$$FQ(s,a) = \sum_{l=1}^{L} Q_l(s,a)\psi_l(s,a) \tag{17}$$

where L represents the number of fuzzy states, $FQ_l(s,a)$ is the value function of the l-th fuzzy state and $\psi_l(s,a)$ stands for the degree of relationship between the state s and the l-th fuzzy state with the given action a.

The mixed policy $\pi(s,a)$ controls the probabilities which is utilized to choose an action with a given policy. With FSA, the policy decision is determined by both the expected reward value $Q_l(s,a)$ and the policy $\pi_l(s,a)$, i.e.,

$$\pi_{\text{FSA}}(s,a) = \sum_{l=1}^{L} \pi_l(s,a)Q_l(s,a)\psi_l(s,a). \tag{18}$$

And the elements of $\{\pi_l(s,a)\}_{l=1}^{L}$ are initialized by

$$\sum_{a=1}^{|\mathcal{A}|} \sum_{l=1}^{L} \pi_l(s,a)\psi_l(s,a) = 1. \tag{19}$$

The policy $\pi_l(s,a)$ is updated as follow

$$\pi_l(s,a)$$
$$\leftarrow \pi_l(s,a) + \begin{cases} \frac{\xi\psi_l(s,a)}{L}, & \text{if } a = \arg\max_{a'} FQ(s,a') \\ -\frac{\xi\psi_l(s,a)}{L(|\mathcal{A}|-1)}, & \text{otherwise.} \end{cases}, \forall l \tag{20}$$

As WoLF-PHC is combined with FSA, the probability ξ is used to update the entire fuzzy summation

$$\sum_{a=1}^{|\mathcal{A}|} \sum_{l=1}^{L} \pi_l(s,a)\psi_l(s,a) \leftarrow \sum_{a=1}^{|\mathcal{A}|} \sum_{l=1}^{L} \pi_l(s,a)\psi_l(s,a)$$
$$+ \begin{cases} \xi, & \text{if } a = \arg\max_{a'} FQ(s,a') \\ -\frac{\xi}{|\mathcal{A}|-1}, & \text{otherwise.} \end{cases} \tag{21}$$

It is worth pointing out that with a given action a, the probability ξ needs to be carefully set in (20) and (21), to avoid a disproportionate growth in $\pi_l(s,a)$ [47].

Accordingly, the update on the fuzzy state-action value function is given by

$$FQ(s,a) \leftarrow (1-\alpha)FQ(s,a)$$
$$+\alpha \left(r + \gamma \max_{a \in \mathcal{A}} FQ^*(s',a) \right). \tag{22}$$

The fuzzy WoLF-PHC-based joint power allocation and reflecting beamforming approach for the IRS-assisted communication system against smart jamming is summarized in **Algorithm 1**. In the system, 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 the average policy $\widehat{\pi}(s^t,a^t)$, and uses them to select the variable learning rate ξ to improve the learning rate. According to the updated policy $\pi(s^t,a^t)$, FSA is adopted to calculate the fuzzy Q-function $FQ(s^t,a^t)$ and the fuzzy-state policy $\pi_{\text{FSA}}(s,a)$, as well as updating $FQ(s^t,a^t)$ in the next time slot until it converges to the final point. 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.

After adequate training in **Algorithm 1**, the trained model is loaded for the implementation stage. During the implementation stage, the controller uses the trained learning model to output its selected action a by going through fuzzy WoLF-PHC, with the observed state s from the IRS-aided secure communication system. Specifically, it chooses an action a, with the maximum value based on the trained fuzzy WoLF-PHC learning model. Afterward, the environment feeds back an instantaneous reward and a new system state to the a-gent. Finally, the power allocation $\{P_k\}$ at the BS and the phase shift matrix $\mathbf{\Phi}$ (reflecting beamforming) at the IRS are achieved according to the selected action.

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

1: **Input:** Fuzzy WoLF-PHC learning structure and IRS-assisted system with a jammer.
2: **Initialize:** $Q(s,a) = 0$, $\pi(s,a) = 1/|\mathcal{A}|$, $\widehat{\pi}(s,a) = \pi(s,a)$, ξ, $\xi_{\text{loss}} > \xi_{\text{win}}$, $C(s)$, γ, α, and set fuzzy rules.
3: **for** each episode $j = 1, 2, \ldots, N^{\text{epi}}$ **do**
4:    **for** each time step $t = 0, 1, 2, \ldots, T$ **do**
5:      Observe an initial system state s^t;
6:      Select action a^t based on ε-greedy policy by (13):
      $a^t = \arg\max\limits_{a^t \in \mathcal{A}} Q(s^t, a^t)$, with probability 1-$\varepsilon$;
      $a^t = \text{random}\{a_i\}_{a_i \in \mathcal{A}}$, with probability ε;
7:      Execute the exploration action a^t, receive a reward $r(s^t, a^t)$ and the next state s^{t+1};
8:      Update the current policy $\pi(s^t, a^t)$ via (14);
9:      Select the variable learning rate ξ via (15);
10:     Update the average policy $\widehat{\pi}(s^t, a^t)$ via (16);
11:     Set $C(s) = C(s) + 1$;
12:     Compute the fuzzy Q-function $FQ(s^t, a^t)$ via (17);
13:     Update the fuzzy-state policy $\pi_{\text{FSA}}(s, a)$ via (20);
14:     Update $FQ(s^t, a^t)$ by via (22);
15:    **end for**
16:   **end for**
17: **Return:** Fuzzy WoLF-PHC-based learning model;
18: **Output:** Load the learning model to achieve the joint power allocation $\{P_k\}_{k \in \mathcal{K}}$ and reflecting beamforming matrix Φ strategy.

C. Complexity and Convergence Analysis

There have two main notions of complexity for RL: the space complexity and the run-time complexity. Let $\mathcal{S}$ and $\mathcal{A}$ denote the sets of the current state space and action space, respectively. Our proposed WoLFCPHC learning algorithm is an extension framework of the Q-learning algorithm. For RL, [44] proved that the Q-learning algorithm takes $O\left(|\mathcal{S}|^2 \times |\mathcal{A}|\right)$ sample transitions to compute an ε-greedy policy. It also showed that it takes $O\left(|\mathcal{S}| \times |\mathcal{A}|\right)$ running time to compute the iteration process. Hence, at a decision stage, the sample complexity of the action selection and learning update of the proposed WoLFCPHC learning algorithm is $O\left(|\mathcal{S}|^2 \times |\mathcal{A}|\right)$, and the running time complexity the algorithm is $O\left(|\mathcal{S}| \times |\mathcal{A}|\right)$.

According to the aforementioned analysis of both the sample complexity and the run-time complexity, we can observe that both the state and action spaces mainly decide the complexity of the proposed algorithm. In our proposed WoLFCPHC learning algorithm which is based on Q-learning, both the state and action spaces are discretized.

The proposed fuzzy WoLF-PHC-based algorithm is an extension of Q-learning for the decision-making, and it essentially uses a sample average of the fuzzy action-value function, i.e. $FQ(s, a)$, to approximate $FQ^*(s, a)$. Here, we provide a proof that the proposed fuzzy WoLF-PHC-based algorithm can converge to the optimal point of finite Markov decision process (MDP).

Theorem 1: The fuzzy WoLF-PHC-based algorithm given

Fig. 3. Simulation setup.

by

$$
\begin{aligned}
FQ_{t+1}(s_t, a_t) &= (1 - \alpha_t)FQ_t(s_t, a_t) \\
&\quad + \alpha_t \left[r_t + \gamma \max_{a \in \mathcal{A}} FQ_t(s_{t+1}, a_t) \right]
\end{aligned} \quad (23)
$$

can converge to the optimal $FQ^*(s, a)$ values of finite MDP if the following conditions are satisfied [44]:
  i. The state and action spaces are finite;
  ii. The learning rate α_t admits: $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$;
  iii. The instantaneous reward is bounded and the transition probability function $\mathcal{P}(s_{t+1}|s_t, a_t)$ is stationary.
Proof: the details of the proof can be seen in [44].

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 and the IRS are (0, 0) and (75, 100) in meter (m), respectively. There are obstacles which block the direct communication links from the BS to the UEs, so the obstacles cause a large-scale pathloss for the communication links. In addition, a smart jammer is randomly located in the 50 m $\times$ 100 m rectangle area (grey area), since the location behavior of the smart jammer is not easy to be caught and it may move randomly near the UEs.

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

$$
PL = (PL_0 - 10\beta \log_{10}(d/d_0)) \quad (24)
$$

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 β_{bu}, β_{br}, β_{ru}, and β_{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 [26]-[29], we set $PL_0 = 30$ dB and $d_0 = 1$ m. Since there are extensive obstacles as well as scatterers in the channel links from the BS to the UEs, we set the path loss exponent $\beta_{\text{bu}} = 3.75$. As the IRS is carefully located in favorable location to provide a low path loss, then we set $\beta_{\text{br}} = \beta_{\text{ru}} = 2.2$. The smart jammer is located near the UEs to disturb the communication performance of UEs, and its corresponding path loss exponent can be set to $\beta_{\text{ju}} = 2.5$.

We set the background noise at all UEs equal to $\delta^2 = -105$ dBm. The number of antennas at the BS and the jammer are set to $N = N_J = 8$. The maximum transmit power $P_{\max}$ at the BS varies from 15 dBm to 40 dBm, the SINR target value for UEs varies from $SINR^{\min} = 10$ dB to $SINR^{\min} = 25$ dB, 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 0.1. The cost parameters λ_1 and λ_2 in (7) are set to $\lambda_1 = 1$ and $\lambda_2 = 2$ to balance the utility and cost. We set $\xi_{\text{loss}} = 0.04$ and $\xi_{\text{win}} = 0.01$ [46], [47].

In addition, we compare the proposed fuzzy WoLF-PHC-based joint power allocation and reflecting beamforming approach (denoted as fuzzy WoLF-PHC learning) with the following approaches:

- The system rate maximization approach which jointly optimizes the BS's transmit power allocation and the IRS's reflect beamforming against jamming. The approach adopts AO to solve the optimization problem (4), and an iterative algorithm is used to update power allocation and reflect beamforming under QoS constraints, which is similar to the near-optimal solution [39] (denoted as Baseline 1 [39]).
- The popular fast Q-Learning approach [19], which is adopted to optimize the transmit power allocation and reflecting beamforming in IRS-assisted communication systems (denoted as fast Q-Learning [19]).
- The optimal transmit power allocation at the BS without IRS assistance (denoted as optimal PA without IRS).

A. Convergence Comparisons of Different RL Approaches

We first compare the convergence performance of all approaches when $P_{\max} = 30$ dBm, $K = 4$, $M = 60$, and $SINR^{\min} = 10$ dB. We show in Fig. 4 that the reward value with increasing training episodes to investigate the convergence behavior of the proposed fuzzy WoLF-PHC-based approach and compared approaches. Clearly, the proposed learning approach achieves the higher reward performance than those of the fast Q-learning approach and classical Q-learning approach. Especially, the proposed approach has faster convergence speed by adopting the PHC mechanism to improve the learning efficiency and convergence speed. The fast Q-learning approach can achieve the comparable convergence speed by adopting fast learning mechanism, but it achieves poorer performance than that of our proposed approach, leading to the poor reward value. Even though the classical Q-learning approach has the simplest structure, the worst performance fails to optimize the network performance with increasing training episodes. Our proposed approach applies both the PHC and fuzzy state aggregation mechanisms to enhance the convergence speed and learning efficiency,
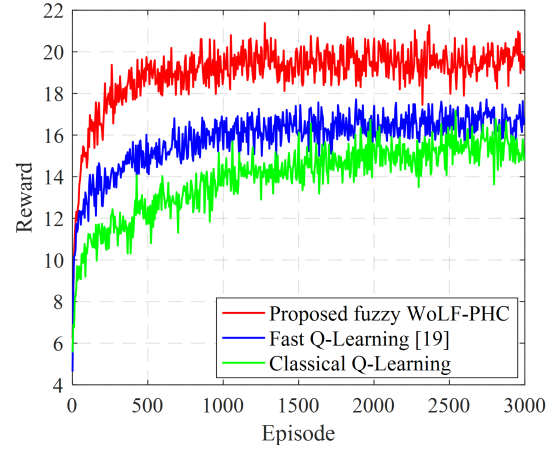

Fig. 4. Convergence behaviors of the various RL approaches.

and the optimized strategy can be learned after a number of training episodes.

From Fig. 4, the reward per episode is improved as training continues, demonstrating the effectiveness of the proposed training approach. When the training episode approximately reaches 1200, the performance gradually converges despite some fluctuations due to mobility-induced channel fading in mobile environments and the adopted ε-greedy policy for the exploration and the exploitation in the training process. In our considered IRS-assisted systems, the environment is complex, and also the action and state space are large, the proposed learning approach requires about 1200 training episodes to appropriately converge.

B. Performance Comparisons versus Maximum Transmit Power

The average system rate and SINR protection level versus the maximum transmit power $P_{\max}$ for various approaches are shown in Fig. 5, when $K = 4$, $M = 60$, and $SINR^{\min} = 10$ dB, which demonstrates that the achieved system rate and SINR protection level improve as $P_{\max}$ increases. We can also observe that both the proposed learning approach and the Baseline 1 approach have good system rate value under different values of $P_{\max}$, and both of them greatly outperform other approaches. However, the SINR protection level of the Baseline 1 approach is obviously lower than that of the proposed learning approach and fast Q-Learning approach, because it is a single time slot optimization solution that ignores the long-term benefit, and thus QoS requirement cannot be effectively guaranteed.

Additionally, the performance improvement achieved by using IRS versus no IRS increases with $P_{\max}$, which indicates the advantage of deploying IRS against smart jamming. In addition, the performance of both the system rate and the SINR protection level of the proposed fuzzy WoLF-PHC-based learning approach is higher than that of the fast Q-Learning approach, which is due to the fact that WoLF-PHC and FSA are adopted to effectively search the optimal joint power allocation and reflecting beamforming strategy against smart jamming in dynamic uncertain environments. From Fig.
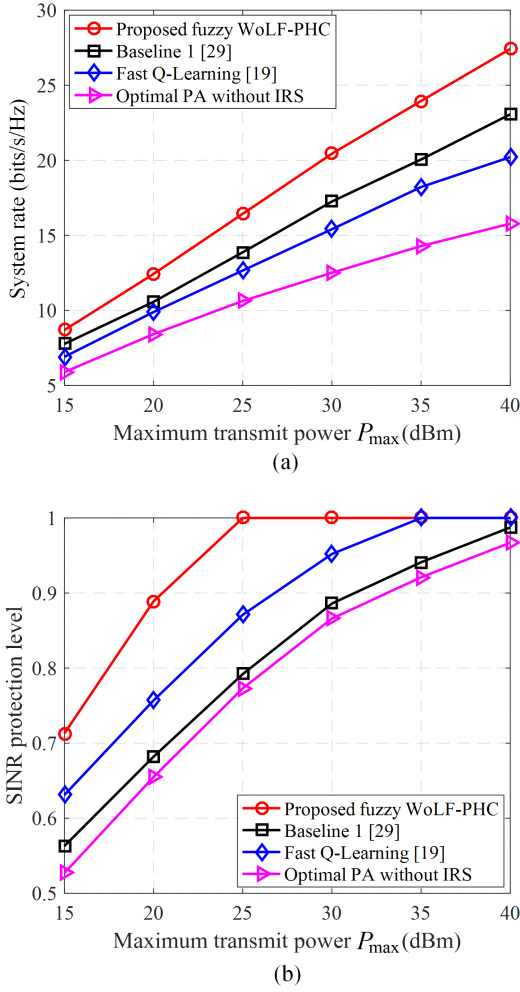
Fig. 5. Performance comparisons versus the maximum transmit power P_{max}.

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

5(b), it is interesting to observe that the optimal PA approach without IRS has the comparable SINR protection level with the fast Q-Learning approach based IRS-assisted system, but the optimal approach is impractical due to the assumption of the prefect CSI and known jamming model in the system.

C. Performance Comparisons versus Number of Reflecting Elements

Fig. 6 compares the performance of the four approaches for different reflecting elements number M when $P_{max} = 30$ dBm, $K = 4$ and $SINR^{min} = 10$ dB. 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. The reason 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 2.36 bits/s/Hz, while this value is improved to 12.21 bits/s/Hz when $M = 100$. In addition, by
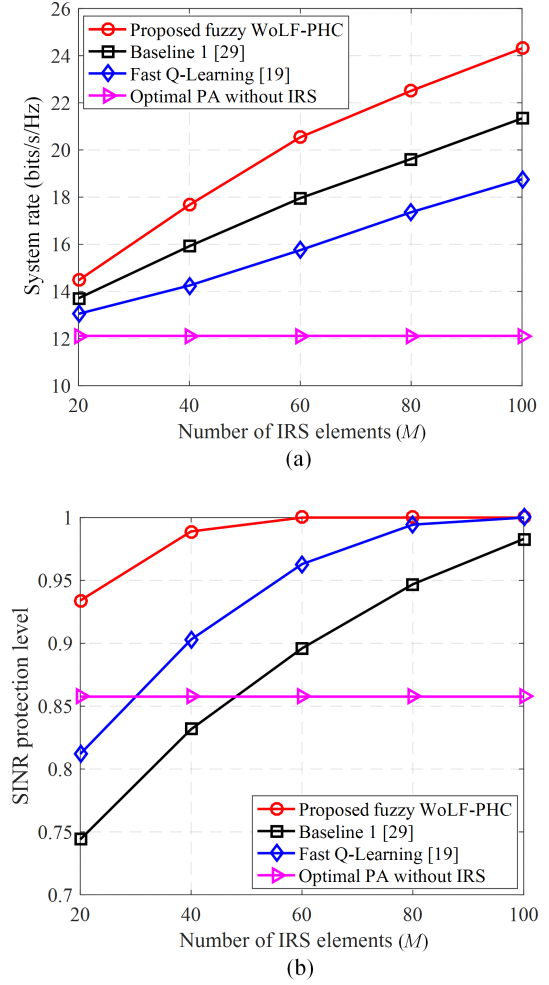
deploying the IRS, the SINR protection level is significantly improved compared with the optimal PA approach without IRS assistance. Such performance improvement results from the fact that higher power can be achieved at the IRS by increasing M, and a higher reflecting beamforming gain is achieved to 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.

In addition, from Fig. 6(a), we can also observe that the achievable rate of the proposed learning approach outperforms both the fast Q-Learning and Baseline 1 approaches, and especially, the performance gap significantly increases with M. At the same time, Fig. 6(b) shows that, as the reflecting elements increases, the proposed learning approach is the first one can achieve 100% SINR protection level compared with other approaches. This is because deploying more reflecting elements, the proposed fuzzy WoLF-PHC-learning based joint power allocation and reflecting beamforming approach becomes more flexible for optimal phase shift (reflecting beamforming) design and hence achieves the higher performance gain. These results also show that employing IRS into wireless communications improves the anti-jamming communication performance against smart jamming.

Fig. 7. Performance comparisons versus UE SINR target.

D. Impact of SINR Target

In Fig. 7, we investigate the effect of the UE SINR target ($SINR^{\min}$) on the system performance for different approaches when $P_{\max} = 30$ dBm, $K = 4$, and $M = 60$. It can be seen from Fig. 7 that, for the low SINR target regime, both the system rate and SINR protection level decline slightly for increasing SINR target value, but the performance drops significantly when $SINR^{\min}$ is larger than a certain threshold. The reason is that that all approaches can still operate an optimized power allocation and reflecting beamforming design to maintain a favorable rate and guarantee the SINR requirement as the SINR target value is small, but the IUI and jamming interference become the performance bottleneck when the UE SINR target is high; especially, the jamming interference from the smart jammer is out of control which mainly limits the performance maintenance when the SINR constraint is stringent.

It is worth noting that the three approaches based on the IRS-assisted system achieve quite higher system rate and SINR protection level performance than that of the optimal PA approach without IRS, and the SINR protection performance gap between them increases significantly for increasing SINR

target value. This further demonstrates that the deployment of IRS with reflecting beamforming design can effectively enhance the desired signal power and mitigate the jamming interference generated from the smart jammer. From Fig. 7(b), among the three approaches in the IRS-assisted system, our proposed fuzzy WoLF-PHC-based leaning approach and the fast Q-Learning approach achieve a significant improvement of the SINR protection level against smart jamming compared with the Baseline 1 approach. The reason is that the two learning approaches have the particular design of the SINR-aware reward function shown in (7) for SINR protection against smart jamming. Furthermore, our proposed learning approach achieves the best performance, which indicates that the joint power allocation and reflect beamforming design is crucial for the anti-jamming communication performance against smart jamming, which needs the use of both the practical reflecting beamforming and more sophisticated optimization.

V. Conclusions

This paper has proposed to improve the anti-jamming performance of wireless communication systems by employing an IRS. With the assistance of IRS, the signal received at legitimate UEs is enhanced while the jamming signal generated from the smart jammer can be mitigated. Specifically, we have formulated an optimization problem by joint optimizing both the transmit power allocation at the BS and the reflecting beamforming at the IRS. As the non-convex optimization problem is not easy to solve and the jamming model is unknown, a fuzzy WoLF-PHC learning approach has been proposed to achieve the optimal anti-jamming strategy, where WoLF-PHC and FSA are capable of quickly achieving the optimal policy without knowing the jamming model in uncertain environments. Simulation results have confirmed that that the IRS-assisted system significantly improves the anti-jamming performance, and also verified the effectiveness of our proposed learning approach in terms of improving system rate as well as service protection level, compared to other existing approaches.

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] Y. -S. Shiu, S. -Y. Chang, H. -C. Wu, S. C. -H. Huang, and H. -H. Chen, "Physical layer security in wireless networks: a tutorial," *IEEE Wireless Commun.*, vol. 18, no. 2, pp. 66C74, Apr. 2011.

[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. DOro, 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] L. Jia, F. Yao, Y. Sun, Y. Xu, S. Feng, and A. Anpalagan, "A hierarchical learning solution for anti-jamming Stackelberg game with discrete power strategies," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 818-821, Dec. 2017.

[11] A. Garnaev, A. Petropulu, W. Trappe, and H. V. Poor, "A power control game with uncertainty on the type of the jammer," in *Proc. IEEE Glob. Conf. Signal Inf. Process. (GlobalSIP)*, Ottawa, ON, Canada, 2019, pp. 1-5.

[12] T. T. Do, E. Bj?rnson, E. G. Larsson, and S. M. Razavizadeh, "Jamming-resistant receivers for the massive MIMO uplink," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 210-223, Jan. 2018.

[13] P. Gu, C. Hua, W. Xu, R. Khatoun, Y. Wu, and A. Serhrouchni, "Control channel anti-jamming in vehicular networks via cooperative relay beamforming," *IEEE Internet of Things J.*, vol. 7, no. 6, pp. 5064-5077, Jun. 2020.

[14] 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.

[15] 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.

[16] Z. Feng *et al.*, "Power control in relay-assisted anti-jamming systems: a Bayesian three-layer stackelberg approach," *IEEE Access*, vol. 7, pp. 14623-14636, 2019.

[17] Q. Liu, M. Li, X. Kong, and N. Zhao, "Disrupting MIMO communications with optimal jamming signal design," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5313-5325, Oct. 2015.

[18] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhuang, "UAV relay in VANETs against smart jamming with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4087-4097, May 2018.

[19] 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.

[20] 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.

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

[22] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838-1851, Mar. 2020.

[23] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. Zhang, "The roadmap to 6G - AI empowered wireless networks, " *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84-90, Aug. 2019.

[24] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge AI: Algorithms and systems," Appear in *IEEE Commun. Surveys Tuts.*, doi: 10.1109/COMST.2020.3007787.

[25] K. Yang, Y. Shi, Y. Zhou, Z. Yang, L. Fu, and W. Chen, "Federated machine learning for intelligent IoT via reconfigurable intelligent surface," *IEEE Network*, vol. 34, no. 5, pp. 16-22, Sep. 2020.

[26] 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.

[27] 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.

[28] 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.

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

[30] M. D. Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. d. Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450-2525, Nov. 2020.

[31] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849-5863, Sept. 2020.

[32] C. Pan *et al.*, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1719-1734, Aug. 2020.

[33] C. Pan *et al.*, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5218-5233, Aug. 2020.

[34] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1809-1822, Aug. 2020.

[35] H. Zhang, B. Di, L. Song, and Z. Han, "Reconfigurable intelligent surfaces assisted communications with limited phase shifts: How many phase shifts are enough?," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4498-4502, Apr. 2020.

[36] M. A. El Mossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Trans. Cognitive Commun. Netw.*, vol. 6, no. 3, pp. 990-1002, Sept. 2020.

[37] 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.

[38] L. Dong and H. Wang, "Secure MIMO transmission via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 787-790, June 2020.

[39] 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.

[40] 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.

[41] H. Gacanin and M. Di Renzo, "Wireless 2.0: Towards an intelligent radio environment empowered by reconfigurable meta-surfaces and artificial intelligence," 2020. [Online]. Available: https://arxiv.org/abs/2002.11040.

[42] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning based intelligent reflecting surface for secure wireless communications," Appear in *IEEE Trans. Wireless Commun.*, doi: 10.1109/TWC.2020.3024860.

[43] S. Hong, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Artificial-noise-aided secure MIMO wireless communications via intelligent reflecting surface," Appear in *IEEE Trans. Commun.*, doi: 10.1109/TCOM-M.2020.3024621.

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

[45] 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.

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

[47] W. Dean and P. Gilbert, "Fuzzy state aggregation and policy hill climbing for stochastic environments," *Int. J. Comp. Intell. App.*, vol. 6, pp. 413-428, Feb. 2006.

[48] H. Yang *et al.*, "Intelligent reflecting surface assisted anti-jamming communications based on reinforcement learning," will be presented in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, Taipei, Taiwan.
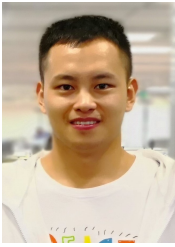
Helin Yang (S'15-M'20) received the B.S. and M.S. degrees in the School of Telecommunications Information Engineering from Chongqing University of Posts and Telecommunications in 2013, and 2016, and the Ph.D. degree from School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2020. His current research interests include wireless communication, visible light communication, Internet of Things and resource management.

Zehui Xiong (S'17-M'20) is currently a researcher with Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University, Singapore. He received the Ph.D. degree at Nanyang Technological University, Singapore. He received the B.Eng degree with the highest honors at Huazhong University of Science and Technology, Wuhan, China. He is the visiting scholar with Princeton University and University of Waterloo. His research interests include network economics, wireless communications, blockchain, and edge intelligence. He has published more than 60 peer-reviewed research papers in leading journals and flagship conferences, and 3 of them are ESI Highly Cited Papers. He has won several Best Paper Awards. He is an Editor for Elsevier Computer Networks (COMNET) and Elsevier Physical Communication (PHYCOM), and an Associate Editor for IET Communications. He is the recipient of the Chinese Government Award for Outstanding Students Abroad in 2019, and NTU SCSE Outstanding PhD Thesis Runner-Up Award in 2020.
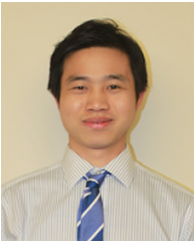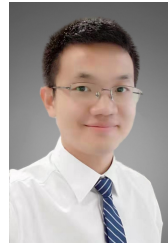
Jun Zhao (S'10-M'15) is currently an Assistant Professor in the School of Computer Science and Engineering at Nanyang Technological University (NTU) in Singapore. He received a PhD degree in Electrical and Computer Engineering from Carnegie Mellon University (CMU) in the USA (advisors: Virgil Gligor, Osman Yagan; collaborator: Adrian Perrig), affiliating with CMU's renowned CyLab Security & Privacy Institute, and a bachelor's degree from Shanghai Jiao Tong University in China. Before joining NTU first as a postdoc with Xiaokui Xiao and then as a faculty member, he was a postdoc at Arizona State University as an Arizona Computing PostDoc Best Practices Fellow (advisors: Junshan Zhang, Vincent Poor). His research interests include communications, networks, security, and AI.

Dusit Niyato (M'09-SM'15-F'17) is currently a professor in the School of Computer Science and Engineering, at Nanyang Technological University, Singapore. He received B.Eng. from King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand in 1999 and Ph.D. in Electrical and Computer Engineering from the University of Manitoba, Canada in 2008. His research interests are in the area of energy harvesting for wireless communication, Internet of Things (IoT) and sensor networks.

Qingqing Wu (S13-M16) received the B.Eng. and the Ph.D. degrees in Electronic Engineering from South China University of Technology and Shanghai Jiao Tong University (SJTU) in 2012 and 2016, respectively. He is currently an assistant professor in the Department of Electrical and Computer Engineering at the University of Macau, China, and also affiliated with the State key laboratory of Internet of Things for Smart City. He was a Research Fellow in the Department of Electrical and Computer Engineering at National University of Singapore. His current research interest includes intelligent reflecting surface (IRS), unmanned aerial vehicle (UAV) communications, and MIMO transceiver design. He has published over 70 IEEE journal and conference papers. He was the recipient of the IEEE WCSP Best Paper Award in 2015, the Outstanding Ph.D. Thesis Funding in SJTU in 2016, the Outstanding Ph.D. Thesis Award of China Institute of Communications in 2017. He was the Exemplary Editor of IEEE Communications Letters in 2019 and the Exemplary Reviewer of several IEEE journals. He serves as an Associate Editor for IEEE Communications Letters, IEEE Open Journal of Communications Society, and IEEE Open Journal of Vehicular Technology. He is the Lead Guest Editor for IEEE Journal on Selected Areas in Communications on "UAV Communications in 5G and Beyond Networks", and the Guest Editor for IEEE Open Journal on Vehicular Technology on "6G Intelligent Communications" and IEEE Open Journal of Communications Society on Reconfigurable Intelligent Surface-Based Communications for 6G Wireless Networks". He is the workshop co-chair for IEEE ICC 2019-2021 workshop on "Integrating UAVs into 5G and Beyond", and the workshop co-chair for IEEE GLOBECOM 2020 workshop on Reconfigurable Intelligent Surfaces for Wireless Communication for Beyond 5G". He serves as the Workshops and Symposia Officer of Reconfigurable Intelligent Surfaces Emerging Technology Initiative and Research Blog Officer of Aerial Communications Emerging Technology Initiative.

H. Vincent Poor (S'72-M'77-SM'82-F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor of Electrical Engineering. During 2006 to 2016, he served as Dean of Princetons School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, machine learning and network science, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the recent book Multiple Access Techniques for 5G Wireless Networks and Beyond. (Springer, 2019).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and a D.Eng. honoris causa from the University of Waterloo awarded in 2019.

Massimo Tornatore (S'03-M'06-SM'13) is currently an Associate Professor with the Department of Electronics, Information, and Bioengineering, Politecnico di Milano. He also holds an appointment as Adjunct Professor at University of California, Davis, USA and as visiting professor at University of Waterloo, Canada. His research interests include performance evaluation, optimization and design of communication networks (with an emphasis on the application of optical networking technologies), cloud computing, and machine learning application for network management. In these areas, he co-authored more than 350 peer-reviewed conference and journal papers (with 18 best paper awards), 2 books and 1 patent. He is a member of the Editorial Board of IEEE Communication Surveys and Tutorials, IEEE Communication Letters, Springer Photonic Network Communications, and Elsevier Optical Switching and Networking. He is active member of the technical program committee of various networking conferences such as INFOCOM, OFC, ICC, and GLOBECOM. He acted as technical program chair of ONDM 2016 and DRCN 2017 and DRCn 2019 conferences. He has participated in several EU RD projects (among others FP7 COMBO, H2020 MEtroHaul, and Cost Action RECODIS) as well as in several projects in USA, Canada and Italy.