Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Joint power control and passive beamforming optimization in RIS-assisted anti-jamming communication*

Yang LIU, Kui XU‡, Xiaochen XIA, Wei XIE, Nan MA, Jianhui XU

School of Communication Engineering, Army Engineering University of PLA, Nanjing 210007, China

†E-mail: 614417393@qq.com; lgdxxukui@sina.com; tjuxxc@sina.com; lgdxxw@outlook.com; manan995@163.com; xujianhui900118@163.com

Received Dec. 27, 2022; Revision accepted Aug. 25, 2023; Crosschecked Nov. 21, 2023

Abstract: Due to the openness of the wireless propagation environment, wireless networks are highly susceptible to malicious jamming, which significantly impacts their legitimate communication performance. This study investigates a reconfigurable intelligent surface (RIS) assisted anti-jamming communication system. Specifically, the objective is to enhance the system's anti-jamming performance by optimizing the transmitting power of the base station and the passive beamforming of the RIS. Taking into account the dynamic and unpredictable nature of a smart jammer, the problem of joint optimization of transmitting power and RIS reflection coefficients is modeled as a Markov decision process (MDP). To tackle the complex and coupled decision problem, we propose a learning framework based on the double deep Q-network (DDQN) to improve the system achievable rate and energy efficiency. Unlike most power-domain jamming mitigation methods that require information on the jamming power, the proposed DDQN algorithm is better able to adapt to dynamic and unknown environments without relying on the prior information about jamming power. Finally, simulation results demonstrate that the proposed algorithm outperforms multi-armed bandit (MAB) and deep Q-network (DQN) schemes in terms of the anti-jamming performance and energy efficiency.

Key words: Reconfigurable intelligent surface (RIS); Power control; Anti-jamming; Reinforcement learning (RL)
https://doi.org/10.1631/FITEE.2200646 **CLC number:** TN929.5

1 Introduction

Wireless networks have gained significance in our daily lives by offering continuous and reliable connectivity for various devices. However, the open nature of the wireless propagation environment makes these networks susceptible to malicious jamming. Malicious jamming attacks, executed by jammers with the intention to disrupt legitimate communications by transmitting jamming signals, have become a pressing concern (Pirayesh and Zeng, 2022). Consequently, optimizing the design of wireless communication networks to mitigate the impact of jamming attacks has emerged as a prominent research area in both academic and industrial domains.

Anti-jamming transmission can be achieved from multiple dimensions, such as the frequency domain, power domain, and beamforming domain. In the frequency domain, anti-jamming can be achieved by dynamically changing the communication frequencies based on known hopping patterns, to evade the jamming attack from hostile jammers (Chang et al., 2017). Machine learning based hopping strategies can also be used to adaptively avoid jamming attacks in dynamic environments (Li et al., 2023). However, frequency hopping anti-jamming has some disadvantages, including spectrum resource waste and high communication overhead. Moreover,

‡ Corresponding author

* Project supported by the Natural Science Foundation of Jiangsu Province, China (Nos. BK 20201334, BK 20200579, and BK 20231485), the National Natural Science Foundation of China (Nos. 62071485, 62271503, and 62001513), and the Basic Research Project of Jiangsu Province, China (No. BK 20192002)

ORCID: Yang LIU, https://orcid.org/0000-0003-1667-6390; Kui XU, https://orcid.org/0000-0001-8533-2255

it becomes ineffective in the presence of malicious jamming across the entire frequency band.

In recent years, significant research progress has been achieved in the field of power-domain anti-jamming methods. In Xu JW et al. (2021), the problem of anti-jamming power control in unmanned aerial vehicle (UAV) communication systems was formulated using two different models of the Stackelberg game and the Nash game. Yu et al. (2017) applied the Stackelberg game to model and analyze the anti-jamming problem, and provided closed-form expressions for the optimal transmission power of users and jammers. In Feng et al. (2019), a three-tier Bayesian Stackelberg anti-jamming game was constructed, with the primary user as the leader, the relay user as the deputy leader, and the jammer as the follower. A multi-user hierarchical iterative algorithm was proposed to achieve the Stackelberg equilibrium using the backward induction method. However, note that the game theory based anti-jamming methods in Feng et al. (2019) and Xu JW et al. (2021) assumed prior knowledge of the jamming model, which might be difficult to attain in practical scenarios.

Reinforcement learning (RL) is a branch of machine learning that relies on actions and feedback to learn and improve. In a specific environment, agents receive feedback in the form of rewards or punishments for their actions, with the goal of maximizing the overall reward in that environment. Through the gradual formation of expectations, the agent can develop an optimal behavior that leads to the maximum rewards (Arulkumaran et al., 2017). RL has widespread application in the field of wireless communication, particularly in optimizing anti-jamming strategies in dynamic wireless communication systems (Luong et al., 2019). In Feng et al. (2020), the discrete power-control problem in anti-jamming relay communication networks was studied, and the power optimization problem was modeled as a multi-armed bandit (MAB) problem. This approach uses an online learning algorithm for discrete power control. In Geng et al. (2022), a power-control optimization strategy was proposed for wireless anti-jamming communication systems based on Q-learning, even when the jamming model and transmission loss were unknown. Xiao ZC et al. (2018) presented a fast power-control algorithm based on strategy climbing for downlink millimeter-wave multi-input multi-output (MIMO) systems. This algorithm, an improved Q-learning approach, considers the signal-to-noise ratio (SNR) and the strength of jamming signals. Ma et al. (2022) proposed a real-time power-control strategy framework based on a deep deterministic policy gradient (DDPG) to achieve higher system rate and energy efficiency. However, note that all of these studies considered jamming power as the state space for agent learning, which might be challenging to obtain in practical scenarios.

The reconfigurable intelligent surface (RIS) is an emerging technology that holds great promise for the sixth generation (6G) mobile communication systems. It offers the potential to enhance the signal coverage, spectral efficiency, and energy efficiency in wireless communication systems (Huang et al., 2019; di Renzo et al., 2020). Unlike existing active massive MIMO and amplify-and-forward technologies, RIS operates by reflecting incident signals without actively processing them. This feature results in lower energy consumption and deployment costs (Basar et al., 2019; Jian et al., 2022). By designing the phase shift and amplitude reflection coefficient of each reflecting element, RIS can achieve passive beamforming capabilities (Wu and Zhang, 2020). In Björnson et al. (2020), the downlink rate of users was maximized through the optimization of the RIS reflective phase-shift coefficient. Additionally, Zhang SW and Zhang (2020) investigated the rate optimization problem in RIS-assisted MIMO systems and proposed an alternating optimization approach to make the objective function converge by optimizing the RIS phase-shift coefficients.

Recent research shows that RIS can be applied to enhance the transmission security of wireless communication networks (Yang et al., 2021a; Xu JD et al., 2023). A novel transmission scheme for secure wireless body area networks (WBANs) using RL and RIS was proposed in Xiao L et al. (2022). The sensor encryption key, transmitting power, and RIS phase shifts were optimized to mitigate the risks of active eavesdropping. In Tang et al. (2021), the deployment of RIS in airspace was explored for achieving anti-jamming communication. The alternating optimization method was used to jointly optimize the beamforming and deployment position of the RIS, leading to an effective enhancement in the transmission rate of legitimate users. On the other hand, Sun et al. (2021) proposed a secure communication

framework for RIS-assisted anti-jamming and anti-eavesdropping scenarios. By jointly optimizing the active beamforming of the base station and RIS-reflected beamforming, the system achievable rate was improved while minimizing the potential information leakage. However, note that the convex optimization methods used in Sun et al. (2021) and Tang et al. (2021) may not guarantee the discovery of the global optimum solution. In contrast, Yang et al. (2020, 2021b) investigated the application of the win-or-learn-fast policy hill-climbing (WoLF-PHC) learning approach to jointly optimize the transmission power control and reflection beamforming in RIS-based anti-jamming systems. Nevertheless, these methods still require estimation of the jamming power. Furthermore, these algorithms may become ineffective when dealing with continuous state spaces.

In this study, the problem of joint optimization of transmitting power control and passive beamforming is modeled as a Markov decision process (MDP). When the prior jamming power is unknown, the issue of anti-jamming communication is transformed into an RL problem by reasonably designing the state, action, and reward. Then we propose a double deep Q-network (DDQN) to implement efficient agent training in dynamic unknown environments. The main contributions of this study are as follows:

1. A RIS-assisted downlink massive MIMO anti-jamming transmission scheme is proposed. The the problem of joint optimization of power control and passive beamforming is modeled as an MDP. The uncertainty of jamming power in dynamic and unknown jamming environments is addressed using RL algorithms.

2. To cope with dynamic and unknown jamming attacks, a DDQN learning algorithm is proposed to handle the uncertainty of jamming power and optimize the joint strategy of power control and passive beamforming. Moreover, phase-shift dithering is introduced to enhance the effectiveness of phase shifting and improve the flexibility and performance of the RIS-assisted anti-jamming scheme.

3. Simulation results demonstrate that the proposed DDQN algorithm optimizes the joint strategy of power control and passive beamforming, enabling effective adaptation to dynamic and unknown jamming attacks, thereby improving the achievable system rate and energy efficiency. Compared with the MAB and deep Q-network (DQN) schemes, the proposed DDQN-based RIS-assisted scheme exhibits superior anti-jamming performance and energy efficiency.

Notations: Scalars are denoted by italic letters; vectors and matrices are denoted by bold-face lower-case and upper-case letters, respectively. $\mathbb{C}^{x\times y}$ denotes the space of $x\times y$ complex-valued matrices. For a complex-valued vector $\boldsymbol{x}$, $\boldsymbol{x}^{\mathrm{H}}$ denotes conjugate transpose, and $\mathrm{diag}(\boldsymbol{x})$ denotes a diagonal matrix with each diagonal element being the corresponding element in $\boldsymbol{x}$. $x\sim\mathcal{CN}(0,\sigma^2)$ means that random variable x is complex circularly symmetric Gaussian with zero mean and variance σ^2, whereas $E\{\cdot\}$ denotes the statistical expectation. $|\cdot|$ denotes modulus.

2 System model and problem formulation

In this section, the system model is described and the RIS-assisted anti-jamming scheme is introduced. The anti-jamming problem is expressed as an optimization problem.

As shown in Fig. 1, in the MIMO wireless communication system, there is a base station equipped with N_{t} antennas and K legitimate users with a single antenna. The direct link between the base station and the users is blocked, while a smart jammer equipped with N_{j} antennas is transmitting a jamming signal to attack users. A RIS with M reflecting elements is deployed on the surface of the building near the base station to assist users in receiving desired signals and resisting jamming. We assume that the users move randomly within a specified range, and the communication system operates in an unknown dynamic environment. In this case, the system achievable rate can be improved and the impact of jamming can be reduced by jointly optimizing the transmitting power at the base station and the passive beamforming at the RIS.

Let $\mathcal{K}=\{1,2,...,K\}$ and $\mathcal{M}=\{1,2,...,M\}$ represent the user equipment (UE) set and the RIS reflecting elements set, respectively. The equivalent channels for the transmitter–user link, transmitter–RIS link, RIS–user link, jammer–user link, and jammer–RIS link are expressed as $\boldsymbol{h}_{\mathrm{tu}}^{\mathrm{H}}\in\mathbb{C}^{1\times N_{\mathrm{t}}}$, $\boldsymbol{G}_{\mathrm{tr}}\in\mathbb{C}^{M\times N_{\mathrm{t}}}$, $\boldsymbol{h}_{\mathrm{ru}}^{\mathrm{H}}\in\mathbb{C}^{1\times M}$, $\boldsymbol{h}_{\mathrm{ju}}^{\mathrm{H}}\in\mathbb{C}^{1\times N_{\mathrm{j}}}$, and $\boldsymbol{G}_{\mathrm{jr}}\in\mathbb{C}^{M\times N_{\mathrm{j}}}$, respectively. $\boldsymbol{\Phi}=\mathrm{diag}(\phi_1,\phi_2,...,\phi_M)$ is defined as the reflecting coefficient matrix of the RIS,

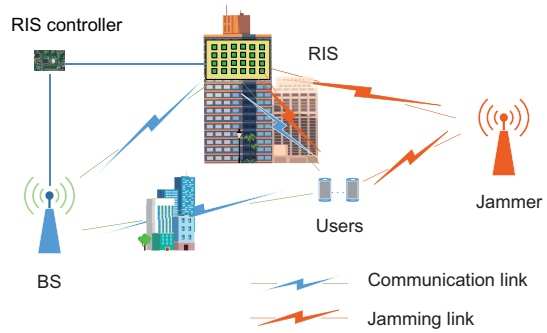

Fig. 1 System model (BS: base station; RIS: reconfigurable intelligent surface)

where $\phi_m = \beta_m e^{j\theta_m} (m = 1, 2, ..., M)$. The reflecting amplitude is described by $\beta_m \in [0, 1]$, while the reflecting phase-shift coefficient of each element is expressed by $\theta_m \in [0, 2\pi]$. We set $\beta_m = 1$ for the maximization of the reflected signal (Wu and Zhang, 2019). The signal received by the user is written as (Wei et al., 2021)

$$\begin{aligned} y_k = & \sqrt{P_{\mathrm{t},k}} \left(\boldsymbol{h}_{\mathrm{tu},k}^{\mathrm{H}} + \boldsymbol{h}_{\mathrm{ru},k}^{\mathrm{H}} \boldsymbol{\Phi} \boldsymbol{G}_{\mathrm{tr}} \right) \boldsymbol{w}_{\mathrm{t},k} x_{\mathrm{t},k} \\ & + \sum_{i=1, i \neq k}^{K} \sqrt{P_{\mathrm{t},i}} \left(\boldsymbol{h}_{\mathrm{tu},k}^{\mathrm{H}} + \boldsymbol{h}_{\mathrm{ru},k}^{\mathrm{H}} \boldsymbol{\Phi} \boldsymbol{G}_{\mathrm{tr}} \right) \boldsymbol{w}_{\mathrm{t},i} x_{\mathrm{t},i} \\ & + \sqrt{P_{\mathrm{j},k}} \left(\boldsymbol{h}_{\mathrm{ju},k}^{\mathrm{H}} + \boldsymbol{h}_{\mathrm{ru},k}^{\mathrm{H}} \boldsymbol{\Phi} \boldsymbol{G}_{\mathrm{jr}} \right) \boldsymbol{w}_{\mathrm{j},k} x_{\mathrm{j},k} + n, \end{aligned} \tag{1}$$

where $\boldsymbol{w}_{\mathrm{t}} \in \mathbb{C}^{N_{\mathrm{t}} \times 1}$ and $\boldsymbol{w}_{\mathrm{j}} \in \mathbb{C}^{N_{\mathrm{j}} \times 1}$ represent the signal beamforming vectors of the base station and jammer, respectively. x_{t} represents the data signal transmitted by the base station and x_{j} is considered as the data signal of the jammer. Specifically, we set $x_i \in \mathbb{C}$, $E\{x_e\} = 0$, and $E\left\{|x_e|^2\right\} = 1 (e = \mathrm{t}, \mathrm{j})$. In addition, P_{t} denotes the transmitting power and P_{j} denotes the jamming power. $n \sim \mathcal{CN}(0, \sigma^2)$ denotes the noise received by the users, which obeys a complex Gaussian distribution.

In this study, we are committed to improving the system achievable rate and energy efficiency by jointly optimizing power control and passive beamforming in anti-jamming communication. Although significant power can bring greater system reachability, the energy problem of power cost must be considered. Therefore, the optimization problem is expressed as

$$\begin{aligned} & \max_{P_{\mathrm{t},k}} \Delta \boldsymbol{\Phi} \sum_{k=1}^{K} \left\{ \log_2 (1 + \mathrm{SINR}_k) - C_{\mathrm{t}} P_{\mathrm{t},k} \right\} \\ & \text{s.t. } C_1 : P_{\mathrm{t}} \leq P_{\mathrm{t}\max}, \\ & \quad C_2 : |\phi_m| = 1, \theta_m \in [0, 2\pi], \quad \forall m \in M. \end{aligned} \tag{2}$$

The signal interference noise ratio (SINR) at the user can be expressed in Eq. (3). The selection of the transmitting power is limited to the set of transmitting power space. The reflecting elements of the RIS provide only phase shift, but cannot amplify the incident signal power. In an unknown environment, the base-station power control and the passive beamforming of RIS can be modeled as an MDP (Ning et al., 2020). The optimal anti-jamming policy can be explored through an RL algorithm, which is developed in the next section.

3 DDQN algorithm in RIS-assisted anti-jamming communication

In this section, we model the optimization problem as an MDP and present the proposed DDQN-based algorithm for joint design of power control and passive beamforming, as shown in Fig. 2. In detail, the designs of the state s, action a, and reward r are introduced.

3.1 Formulation of the optimization problem as an MDP

RL consists of two parts: agent and environment. In this study, the RIS-assisted anti-jamming communication system is considered as an environment, and the central controller of the base station is regarded as an agent. During training, the agent always interacts with the environment with the aim of maximizing its reward in a complex and uncertain environment (Zhang ZD et al., 2020).

The interaction between the agent and the environment can be formulated as an MDP, represented by the tuple (S, A, R, S'). S represents the state space and A denotes the action space. R represents the reward space and S' denotes the next state space. We express the policy as $\psi(S, A) : S \to A$, which is a mapping function from the state space to action space. The agent observes the current state s_n and selects an action a_n. Then, the agent obtains a reward r_n from the environment and moves to the next state s_{n+1} with the transition probability $P(s_{n+1}|s_n, a_n)$. Assume that the jamming power is dynamically unknown. To optimize the problem, s, a, and r are defined as follows.

$$\mathrm{SINR}_k=\frac{P_{\mathrm{t},k}\left|\left(\boldsymbol{h}_{\mathrm{tu},k}^{\mathrm{H}}+\boldsymbol{h}_{\mathrm{ru},k}^{\mathrm{H}}\boldsymbol{\Phi}\boldsymbol{G}_{\mathrm{tr}}\right)\boldsymbol{w}_{\mathrm{t},k}\right|^2}{P_{\mathrm{j},k}\left|\left(\boldsymbol{h}_{\mathrm{ju},k}^{\mathrm{H}}+\boldsymbol{h}_{\mathrm{ru},k}^{\mathrm{H}}\boldsymbol{\Phi}\boldsymbol{G}_{\mathrm{jr}}\right)\boldsymbol{w}_{\mathrm{j},k}\right|^2+\sum\limits_{i=1,i\neq k}^{K}P_{\mathrm{t},i}\left|\left(\boldsymbol{h}_{\mathrm{tu},k}^{\mathrm{H}}+\boldsymbol{h}_{\mathrm{ru},k}^{\mathrm{H}}\boldsymbol{\Phi}\boldsymbol{G}_{\mathrm{tr}}\right)\boldsymbol{w}_{\mathrm{t},i}\right|^2+\sigma^2}. \tag{3}$$

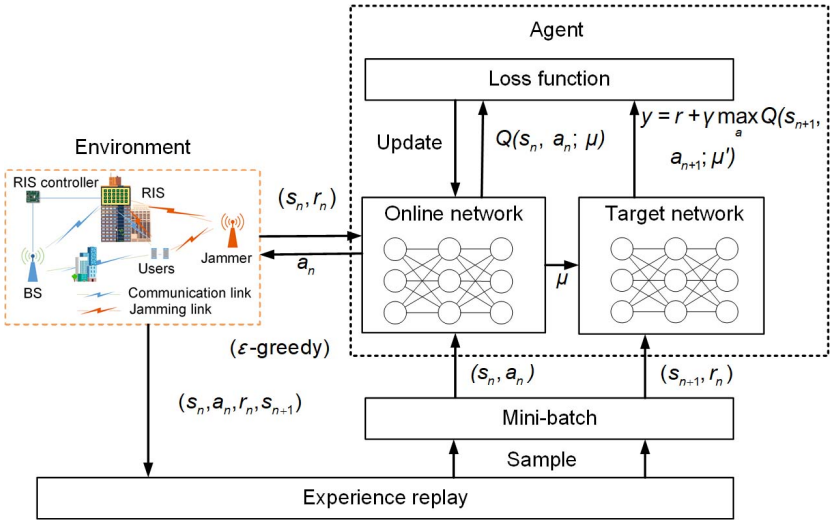

Fig. 2 Double deep Q-network

3.1.1 State

The system state represents the information in the environment observed by the agent at the current moment, including the equivalent channel from the base station to the users $\{\boldsymbol{h}_k\}_{k\in\mathcal{K}}$, the reflection coefficients of RIS $\boldsymbol{\Phi}$, and the $\{\mathrm{SINR}_k\}_{k\in\mathcal{K}}$ of user feedback:

$$S=\left\{\{\boldsymbol{h}_k\}_{k\in\mathcal{K}},\boldsymbol{\Phi},\{\mathrm{SINR}_k\}_{k\in\mathcal{K}}\right\}, \tag{4}$$

where h includes the direct link and reflecting link between the base station and users. SINR_k can be obtained by calculating the statistical quantities of the received symbols in the estimator (Summers and Wilson, 1998; Takizawa et al., 2002).

3.1.2 Action

Action represents the anti-jamming countermeasure performed at the current time, including the transmitting power of the base station $\{P_{\mathrm{t},k}\}_{k\in\mathcal{K}}$ and the incremental reflection coefficients $\Delta\boldsymbol{\Phi}$ of RIS:

$$A=\left\{\{P_{\mathrm{t},k}\}_{k\in\mathcal{K}},\Delta\boldsymbol{\Phi}\right\}, \tag{5}$$

where $\Delta\boldsymbol{\Phi}$ represents the incremental reflection coefficients at the current moment (Wang W and Zhang, 2022a). Therefore, the incremental reflection coefficients at the next moment can be expressed as $\boldsymbol{\Phi}_{n+1}=\boldsymbol{\Phi}_n\odot\Delta\boldsymbol{\Phi}_n$. The subset of column vectors of the discrete Fourier transform (DFT) matrix at point M can be selected as the action set:

$$A_{\Delta\boldsymbol{\Phi}}=\{\tau(-\boldsymbol{s_2}),\tau(-\boldsymbol{s_1}),\tau(\boldsymbol{0}),\tau(\boldsymbol{s_1}),\tau(\boldsymbol{s_2})\}, \tag{6}$$

$$\tau(\boldsymbol{n})=\left[1,\mathrm{e}^{\mathrm{j}\frac{2\pi}{M}n},...,\mathrm{e}^{\mathrm{j}(M-1)\frac{2\pi}{M}n}\right]^{\mathrm{T}}, \tag{7}$$

where $|\boldsymbol{s_1}|<|\boldsymbol{s_2}|\,(|\boldsymbol{s_1}|,|\boldsymbol{s_2}|\in(0,M-1])$. When $\Delta\boldsymbol{\Phi}_n=\tau(0)$, the reflection coefficients do not change. $\tau(\boldsymbol{s_1})$ and $\tau(\boldsymbol{s_2})$ represent the asynchronous lengths of the changes of $\boldsymbol{\Phi}$, while $\tau(-\boldsymbol{s_1})$ and $\tau(-\boldsymbol{s_2})$ represent the corresponding changes of $\tau(\boldsymbol{s_1})$ and $\tau(\boldsymbol{s_2})$ in the opposite direction, respectively. To improve the convergence speed, the incremental reflection coefficient is set to the action subspace, which greatly improves the flexibility of learning.

3.1.3 Phase-shift dither

Using a subset of the DFT matrix as the incremental reflection coefficients will limit the optional states of the phase to M kinds of DFT transformation. To release the optional states of the reflection

coefficient matrix, a random noise of reflection coefficients $\boldsymbol{\Phi}_{\text{noise}}$ can be added to dither the transformed reflection coefficients $\boldsymbol{\Phi}_{n+1}$ by a small magnitude (Wang W and Zhang, 2022a). As shown in Fig. 3, the phase-shift coefficient precision of dither is set as $\frac{2\pi}{qM}$, where M is the number of elements of RIS and q is the phase quantization level. If the system achievable rate after dither is higher than it was before, the dither is performed:

$$\boldsymbol{\Phi}_{n+1} = \boldsymbol{\Phi}_{n+1} \odot \boldsymbol{\Phi}_{\text{noise}}. \tag{8}$$

Otherwise, the dither is performed in the opposite direction:

$$\boldsymbol{\Phi}_{n+1} = \boldsymbol{\Phi}_{n+1} \odot \boldsymbol{\Phi}^*_{\text{noise}}, \tag{9}$$

where $\boldsymbol{\Phi}^*_{\text{noise}}$ denotes the conjugate of $\boldsymbol{\Phi}_{\text{noise}}$.

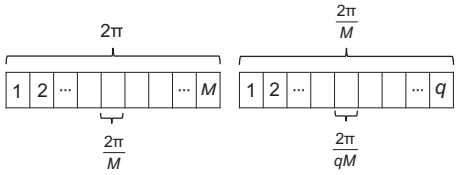

Fig. 3 Phase-shift precision

3.1.4 Reward

The reward function is set as the system achievable rate minus the power consumption:

$$r = \sum_{k=1}^{K} \{\log_2(1+\text{SINR}_k) - C_{\text{t}} P_{\text{t},k}\}, \tag{10}$$

where C_{t} is the transmitting-power cost factor of the base station.

3.2 DDQN-based joint design of power control and passive beamforming

The state-action value function $Q_\psi(s_n, a_n)$ is designed to measure the expected long-term payoff when the action a_n is executed in state s_n. Mathematically, it can be expressed as (Sutton and Barto, 2018)

$$Q_\psi(s_n, a_n) = E_\psi[R_n | s_n = s, a_n = a], \tag{11}$$

where $R_n = \sum_{i=n}^{\infty} \gamma^{n-i} r_i$. $\gamma \in (0,1]$ is a discount factor used to weigh the impact of potential future rewards on the present moment. When the agent selects the optimal strategy $\psi^*(s_n) = \arg\max Q(s_{n+1}, a_{n+1})$, the Q-value function can obtain the maximum value. Thus, the Q-function of the optimal strategy can be written as

$$\begin{aligned} Q^*(s_n, a_n) &= E^*[r_{n+1} | s_n = s, a_n = a] \\ &+\gamma \sum_{s_{n+1} \in S} P(s_{n+1}|s_n, a_n) \max_{a_{n+1} \in A} Q^*(s_{n+1}, a_{n+1}). \end{aligned} \tag{12}$$

According to the recursive derivation of the Bellman equation, the updated formula of Q-function can be given as (Sutton and Barto, 2018)

$$\begin{aligned} Q^*(s_n, a_n) &\leftarrow (1-\alpha) Q^*(s_n, a_n) \\ &+\alpha \left(r_n + \gamma \max_{a_{n+1}} Q_\psi(s_{n+1}, a_{n+1}) \right), \end{aligned} \tag{13}$$

where α is the learning rate.

RL algorithms typically store action values in a table. However, the expansion of the state-action space will lead to dimension disaster (Lyu et al., 2022). To overcome this issue, the function $Q_\mu(s_n, a_n)$ can be used to approximate the value function $Q_\psi(s_n, a_n)$, where $\boldsymbol{\mu}$ denotes the parameter vector of the neural network (Mnih et al., 2015):

$$Q_\psi(s_n, a_n) \approx Q_{\boldsymbol{\mu}}(s_n, a_n). \tag{14}$$

A DQN is proposed to solve such problems by combining value function approximation and neural network technology, using experience replay to train the network. The core of the DQN algorithm is to maintain the Q-function and make decisions based on Q-values.

However, the traditional DQN algorithm will overestimate the Q-values. To make the Q-values close to the objective value, the DDQN algorithm can be used to optimize the agent strategy.

3.2.1 Double estimator method

Two neural networks are used in the DDQN algorithm to decouple action selection and action evaluation into two independent maximum function estimations (van Hasselt et al., 2016). The online network is used for action selection, while the target network is used to generate the target value. This double estimation method can avoid overestimating the action value and lead to the optimal decision result. The target value of the algorithm is defined as (Sutton and Barto, 2018)

$$T_n = r_n + \gamma Q(s_{n+1}, \arg\max_{a^*} Q(s_{n+1}, a^*; \boldsymbol{\mu}); \boldsymbol{\mu}'), \tag{15}$$

where $Q(s_{n+1}, a^*; \boldsymbol{\mu}')$ represents the estimated target value generated by the target network and $\boldsymbol{\mu}'$ represents the parameter vector in the target network. The online network will deliver the parameter vector to the target network periodically.

3.2.2 Experience replay

During learning, the tuple data (s_n, a_n, r_n, s_{n+1}) are stored in the replay buffer as experience samples for training the neural network, and can be used repeatedly. In this way, the amount of interaction with the environment can be significantly reduced. In addition, to obtain optimal training results, the diversity of training data is increased. The loss function adopted during training is written as (Sutton and Barto, 2018)

$$L(\boldsymbol{\mu}) = E\left[\left(T_n - Q(s_n, a_n; \boldsymbol{\mu})\right)^2\right]. \tag{16}$$

The specific DDQN algorithm is summarized in Algorithm 1. In general, the fully connected network can be used to deal with simple tasks. We apply the deep residual network (ResNet) (He et al., 2016) to process the three substates $(\{\boldsymbol{h}_k\}_{k\in\mathcal{K}}, \boldsymbol{\Phi}, \{\mathrm{SINR}_k\}_{k\in\mathcal{K}})$, and then use the three-layer dense network to fuse the information of the three substates after processing. The specific network structure is shown in Fig. 4. The Swish function is used as the activation function (Ramachandran et al., 2017). The gradient descent method is used to minimize the loss function.

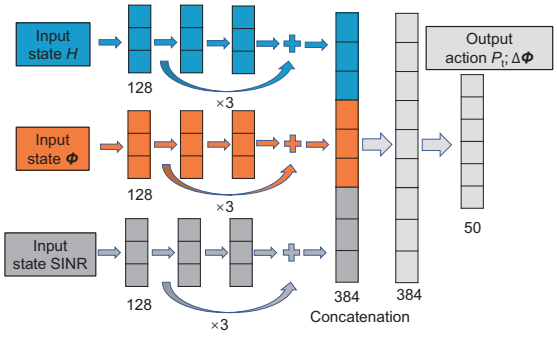

Fig. 4 Structure of the deep neural network

3.3 Computational complexity analysis

The complexity of the DDQN algorithm depends on the structure of the neural network model and the learning process, as referenced in Sharma et al. (2023). The computational complexity of each time step can be determined as follows:

$$\mathcal{O}\left(n_0 n_k + \sum_{k=1}^{K-1} n_k n_{k+1}\right), \tag{17}$$

where n_0 represents the size of the input layer, and n_k represents the number of neurons in the k^{th} layer in the deep neural networks (DNNs). In addition, the complexity as a result of the learning process is reflected in the number of learning episodes N and the number of total time slots T in each episode. Therefore, the computational complexity should be expressed as

$$\mathcal{O}\left(NT\left(n_0 n_k + \sum_{k=1}^{K-1} n_k n_{k+1}\right)\right). \tag{18}$$

Algorithm 1 DDQN-based joint design of power control and passive beamforming

Initialize experience memory and parameters M_{e}, N_{b}, γ , ε;
Initialize online network with a random weight and biases $\boldsymbol{\mu}$;
Initialize the target network as a copy of primary network weights and biases $\boldsymbol{\mu}'$;
for $n = 1, 2, 3, \ldots$ **do**
  Intput state s_n to the DQN and obtain the state-action value $Q_\psi(s_n, a_n)$;
  Select an action a_n based on the ε-greedy policy:
  $$\psi(s,a) = \begin{cases} \arg\max\limits_a Q(s,a), & p_r > \varepsilon, \\ a_{\mathrm{random}}, & p_r \le \varepsilon; \end{cases}$$
  Obtain immediate reward r_n and observe next state s_{n+1};
  Store experience (s_n, a_n, r_n, s_{n+1}) into memory M_{e};
  Randomly sample a mini-batch of N_{b} experience tuples from memory;
  Calculate the target Q-value in the target deep network;
  Set target:
  $T_n = r_n + \gamma Q(s_{n+1}, \arg\max\limits_{a^*} Q(s_{n+1}, a^*; \boldsymbol{\mu}); \boldsymbol{\mu}')$;
  Train online network to minimize the loss function:
  $L(\boldsymbol{\mu}) = E\left[\left(T_n - Q(s_n, a_n; \boldsymbol{\mu})\right)^2\right]$;
  Update the weights of the target network $\boldsymbol{\mu}'$;
end for

4 Simulation results

This section describes the simulated parameter settings and analyzes the simulation results.

Specifically, we compare the anti-jamming performance and energy efficiency of the DDQN, DQN, MAB, random algorithms, and without RIS deployment.

4.1 Simulation scenarios

In the simulation, the base station is located at [0, 10, 10] m; the RIS is located at [5, 12, 5] m; the jammer is located at [−10, −190, 10] m; users are randomly distributed in the region [0, 10) m × [0, 10) m at the height of 1.5 m. The number of base station antennas N_{t} is 2; the number of jammer antennas N_{j} is 2; the number of users K is 2. The maximum power of both the transmitter and jammer is set to 30 dBm, and the power space is $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] \times P_{\max}$. The smart jammer is assumed to use the MAB algorithm to attack legitimate users. The reward function of the smart jammer is set as (Ma et al., 2022)

$$r_{\mathrm{j}} = \sum_{k=1}^{K} \{-\log_2(1+\mathrm{SINR}_k) - C_{\mathrm{j}} P_{\mathrm{j},k}\}. \tag{19}$$

The jamming power cost coefficient C_{j} is the same as the transmitting power cost coefficient. The Rician factor R_{f} in the channel model is set to 10. In the DDQN algorithm, the discount factor is 0.95; the learning rate is 0.001; the greedy coefficient is 0.01. The number of mini-batches is 16.

4.2 Channel model

Due to the limited diffraction ability and predominant line-of-sight (LOS) propagation of millimeter waves (mmWave), the application of RIS in assisting mmWave communication scenarios can effectively compensate for the non-line-of-sight (NLOS) links, particularly highlighting the beamforming capability of the RIS (Wang PL et al., 2020). Specifically, Khawaja et al. (2020) conducted an empirical research to analyze the signal coverage enhancement capabilities of infrared pulse-aided mmWave MIMO in the 28-GHz frequency. The channel models used in this study are based on the Rician fading model (Shen et al., 2021):

$$\boldsymbol{H} = \rho\left(\sqrt{\frac{R}{R+1}}\boldsymbol{H}^{\mathrm{LOS}} + \sqrt{\frac{1}{R+1}}\boldsymbol{H}^{\mathrm{NLOS}}\right), \tag{20}$$

where $\boldsymbol{H}^{\mathrm{LOS}}$ is the LOS link component of the channel and $\boldsymbol{H}^{\mathrm{NLOS}}$ is the NLOS link of the fast-fading channel component. The path loss is $\rho = \frac{\lambda \mathrm{e}^{-\mathrm{j}2\pi d}}{4\pi d}$, where λ is the wavelength and d is the distance between two ends of a link (Wang W and Zhang, 2022b).

The LOS component depends on user positions. Thus, it is slowly time varying. The locations of the users change randomly in each training episode, and the LOS link changes accordingly. Taking the channel for the transmitter–RIS link and the channel for the transmitter–user link as examples, their LOS components are (Wang W and Zhang, 2021)

$$\boldsymbol{G}_{\mathrm{tr}}^{\mathrm{LOS}} = \boldsymbol{a}_M(\vartheta)\boldsymbol{a}_{N_{\mathrm{t}}}{}^{\mathrm{H}}(\upsilon), \tag{21}$$

$$\boldsymbol{h}_{\mathrm{tu},k}^{\mathrm{LOS}} = \boldsymbol{a}_{N_{\mathrm{t}},k}(\varsigma), \tag{22}$$

where $\boldsymbol{a}(\cdot)$ is the steering vector, specifically in the following form:

$$\begin{cases} \boldsymbol{a}_M(\vartheta) = [1, \mathrm{e}^{\mathrm{j}\pi\vartheta}, ..., \mathrm{e}^{\mathrm{j}\pi(M-1)\vartheta}]^{\mathrm{T}}, \\ \boldsymbol{a}_{N_{\mathrm{t}}}(\upsilon) = [1, \mathrm{e}^{\mathrm{j}\pi\upsilon}, ..., \mathrm{e}^{\mathrm{j}\pi(N_{\mathrm{t}}-1)\upsilon}]^{\mathrm{T}}, \\ \boldsymbol{a}_{N_{\mathrm{t}},k}(\varsigma) = [1, \mathrm{e}^{\mathrm{j}\pi\varsigma}, ..., \mathrm{e}^{\mathrm{j}\pi(N_{\mathrm{t}}-1)\varsigma}]^{\mathrm{T}}, \end{cases} \tag{23}$$

where ϑ, υ, and ς are the angular parameters. Assume that the NLOS link changes randomly in each learning time slot, whose components are Gaussian distributed, i.e., $\boldsymbol{H}^{\mathrm{NLOS}} \sim \mathcal{CN}(0,1)$ (Guo et al., 2020). Every 20 learning time steps are taken as a round of training, and the parameters of the online network are passed to the target network. The specific simulation parameters are shown in Table 1.

4.3 Analysis of simulation

In Fig. 5, we compare the system achievable rate using different algorithms. It can be observed that the proposed DDQN algorithm outperforms the DQN, MAB, and random methods. The convergence system achievable rate of the proposed DDQN algorithm reaches 8.5 bit/(s·Hz), while the final convergence value of the DQN algorithm reaches 8.1 bit/(s·Hz) and that of the MAB algorithm reaches 7.5 bit/(s·Hz). In the absence of RIS deployment, the LOS links are obstructed, resulting in communication being almost interrupted, and the system achievable rate cannot be improved even with power optimization. However, in the scenario where the RIS is deployed, the strategy of randomly selecting RIS phases and transmission powers is not effective in improving the system rate during the dynamic

process of jamming. Nevertheless, the system achievable rate can be increased through the reflective links provided by the RIS.

Table 1 Simulation parameters

Notation	Description	Value
R_{f}	Rician factor	10
f	Operating frequency	28 GHz
M_{e}	Length of experience memory	10 000
N_{b}	Number of mini-batches	16
α	Learning rate	0.001
γ	Discount factor	0.95
ε	Greedy coefficient	0.01
σ^2	Power of noise	−100 dBm
B	Communication bandwidth	25 MHz
q	Dither phase quantization level	8
M	Number of elements of RIS	32
$P_{\mathrm{t\,max}}$	Maximum transmitting power	30 dBm
$P_{\mathrm{j\,max}}$	Maximum jamming power	30 dBm
C_{t}	Transmitting power cost coefficient	5×10^{-3}
C_{j}	Jamming power cost coefficient	5×10^{-3}
s_1	Small step of phase shift change	1
s_2	Large step of phase shift change	3

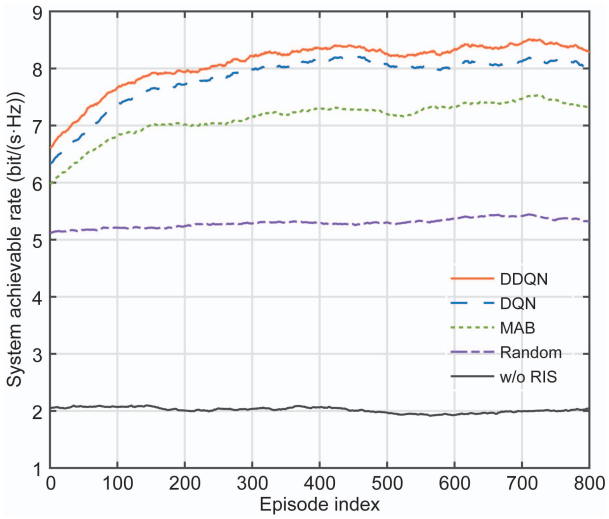

Fig. 5 System achievable rate of the DDQN, DQN, MAB, random algorithms, and without RIS (References to color refer to the online version of this figure)

When the number of users is 2, we assume that communication outage occurs when the system achievable rate is <2 bit/(s·Hz). Fig. 6 compares the outage probability under different algorithms. The results indicate that the highest outage probability occurs without RIS deployment. When the base-station transmission power and RIS reflection phase are randomly adjusted, the outage probability is approximately 0.15. As the number of training rounds increases, the MAB, DQN, and DDQN algorithms are all effective in reducing the outage probability. When learning convergence is achieved, the convergence value of the outage probability for the MAB algorithm is 0.085; for the DQN algorithm it is 0.062; for the DDQN algorithm it is 0.057.

In the reward function setting of RL, a penalty term for power cost is incorporated. This is because the learning objective is to maximize the energy efficiency and explore energy-saving and jamming-resistant dynamic strategies. Assume that the energy efficiency is equal to the instantaneous system achievable rate divided by the instantaneous power. The normalized energy efficiency under different algorithms is compared in Fig. 7. It can be observed that the highest energy efficiency is achieved by the DDQN algorithm, followed by DQN, while MAB exhibits lower energy efficiency. Although the difference in system achievable rates among the three algorithms is not significant, stronger learning capability in exploring energy efficiency is demonstrated by the DDQN algorithm. Relatively low energy efficiency is exhibited by both the random method and the non-RIS approach.

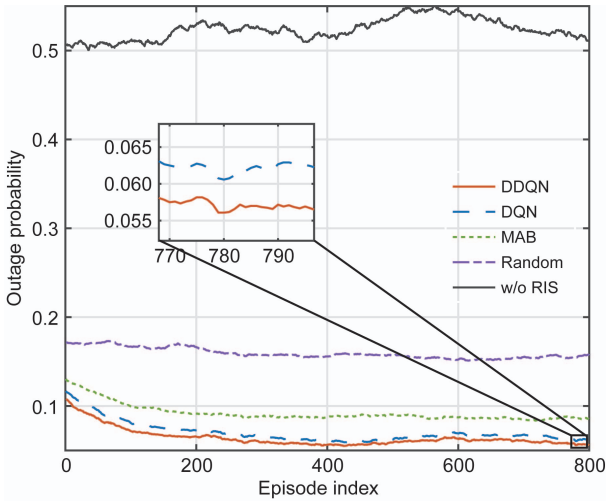

Fig. 6 Outage probability of the DDQN, DQN, MAB, random algorithms, and without RIS (References to color refer to the online version of this figure)

Fig. 8 illustrates the comparison of the power consumed to achieve a rate increase of 1 bit/(s·Hz) in the initial learning stage and cumulative power consumed to achieve a rate increase to 7 bit/(s·Hz) using the DDQN, DQN, and MAB algorithms. In the initial learning stage, the power consumption required to increase the system rate by 1 bit/(s·Hz) using the

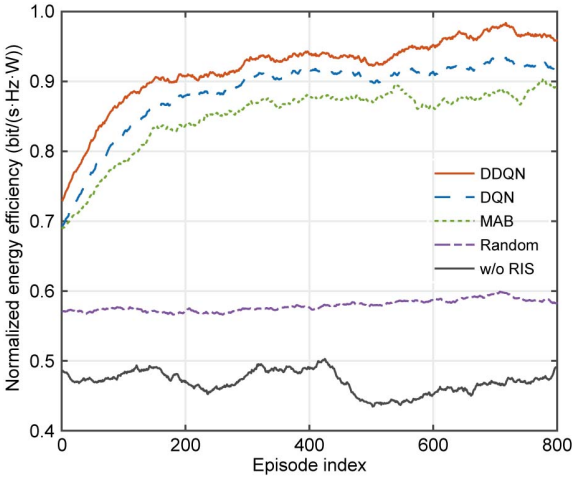

Fig. 7 Normalized energy efficiency of the DDQN, DQN, MAB, random algorithms, and without RIS (References to color refer to the online version of this figure)

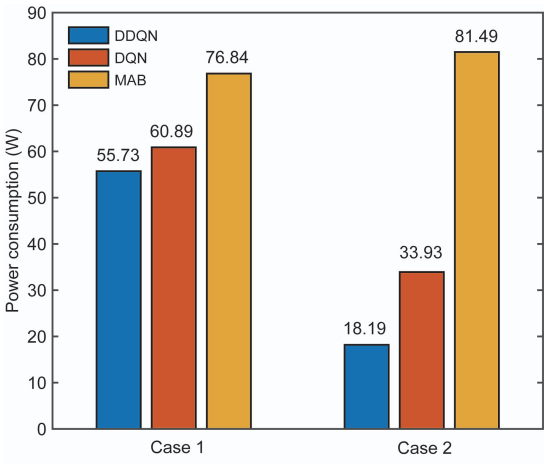

Fig. 8 Power consumed to achieve a rate increase of 1 bit/(s·Hz) (Case 1) and cumulative power consumed to achieve a rate increase to 7 bit/(s·Hz) (Case 2)

DDQN algorithm is approximately 55.73 W, while for the DQN algorithm it is approximately 60.89 W, and for the MAB algorithm it is approximately 76.84 W. When using the DDQN algorithm to increase the system rate to 7 bit/(s·Hz), the cumulative power consumption is approximately 18.19 W, for the DQN algorithm it is 33.93 W, and for the MAB algorithm it is approximately 81.49 W. Obviously, the DDQN algorithm outperforms the DQN and MAB algorithms in terms of energy efficiency.

5 Conclusions

The performance of an anti-jamming RIS-assisted wireless communication system is investigated in this study. By employing RL, the optimal anti-jamming strategy is explored to enhance the anti-jamming communication performance, even without knowledge of the jamming power. Specifically, a DDQN learning algorithm is proposed to jointly optimize power control and RIS passive beamforming, aiming to improve the system's communication performance. Simulations are conducted to compare the performance of the DDQN, DQN, MAB, and random algorithms in terms of the system achievable rate, outage probability, and energy efficiency. The results demonstrate that the proposed DDQN learning algorithm outperforms the other algorithms in terms of performance and energy efficiency.

Contributors

Yang LIU designed the research, processed the data, and drafted the paper. Kui XU, Nan MA, and Jianhui XU helped organize the paper. Kui XU, Xiaochen XIA, and Wei XIE revised and finalized the paper.

Compliance with ethics guidelines

Yang LIU, Kui XU, Xiaochen XIA, Wei XIE, Nan MA, and Jianhui XU declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Arulkumaran K, Deisenroth MP, Brundage M, et al., 2017. Deep reinforcement learning: a brief survey. *IEEE Signal Process Mag*, 34(6):26-38. https://doi.org/10.1109/MSP.2017.2743240

Basar E, di Renzo M, de Rosny J, et al., 2019. Wireless communications through reconfigurable intelligent surfaces. *IEEE Access*, 7:116753-116773. https://doi.org/10.1109/ACCESS.2019.2935192

Björnson E, Özdogan Ö, Larsson EG, 2020. Intelligent reflecting surface versus decode-and-forward: how large surfaces are needed to beat relaying? *IEEE Wirel Commun Lett*, 9(2):244-248. https://doi.org/10.1109/LWC.2019.2950624

Chang GY, Wang SY, Liu YX, 2017. A jamming-resistant channel hopping scheme for cognitive radio networks. *IEEE Trans Wirel Commun*, 16(10):6712-6725. https://doi.org/10.1109/TWC.2017.2728659

di Renzo M, Zappone A, Debbah M, et al., 2020. Smart radio environments empowered by reconfigurable intelligent surfaces: how it works, state of research, and the road ahead. *IEEE J Select Areas Commun*, 38(11):2450-2525. https://doi.org/10.1109/JSAC.2020.3007211

Feng ZB, Ren GC, Chen J, et al., 2019. Power control in relay-assisted anti-jamming systems: a Bayesian three-layer Stackelberg game approach. *IEEE Access*, 7:14623-14636. https://doi.org/10.1109/ACCESS.2019.2893459

Feng ZB, Luo YJ, Chen XQ, et al., 2020. A MAB-based discrete power control approach in anti-jamming relay communication via three-layer Stackelberg game. Proc 6th Int Conf on Computer and Communications, p.267-272. https://doi.org/10.1109/ICCC51575.2020.9344934

Geng SQ, Li PK, Yin XZ, et al., 2022. The study on anti-jamming power control strategy based on Q-learning. Proc 7th Int Conf on Intelligent Computing and Signal Processing, p.182-185. https://doi.org/10.1109/ICSP54964.2022.9778818

Guo HY, Liang YC, Chen J, et al., 2020. Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks. *IEEE Trans Wirel Commun*, 19(5):3064-3076. https://doi.org/10.1109/TWC.2020.2970061

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Huang CW, Zappone A, Alexandropoulos GC, et al., 2019. Reconfigurable intelligent surfaces for energy efficiency in wireless communication. *IEEE Trans Wirel Commun*, 18(8):4157-4170. https://doi.org/10.1109/TWC.2019.2922609

Jian MN, Alexandropoulos GC, Basar E, et al., 2022. Reconfigurable intelligent surfaces for wireless communications: overview of hardware designs, channel models, and estimation techniques. *Intell Converg Netw*, 3(1):1-32. https://doi.org/10.23919/ICN.2022.0005

Khawaja W, Ozdemir O, Yapici Y, et al., 2020. Coverage enhancement for NLOS mmWave links using passive reflectors. *IEEE Open J Commun Soc*, 1:263-281. https://doi.org/10.1109/OJCOMS.2020.2969751

Li XC, Chen JN, Ling X, et al., 2023. Deep reinforcement learning-based anti-jamming algorithm using dual action network. *IEEE Trans Wirel Commun*, 22(7):4625-4637. https://doi.org/10.1109/TWC.2022.3227575

Luong NC, Hoang DT, Gong SM, et al., 2019. Applications of deep reinforcement learning in communications and networking: a survey. *IEEE Commun Surv Tut*, 21(4):3133-3174. https://doi.org/10.1109/COMST.2019.2916583

Lyu L, Shen Y, Zhang SC, 2022. The advance of reinforcement learning and deep reinforcement learning. Proc IEEE Int Conf on Electrical Engineering, Big Data and Algorithms, p.644-648. https://doi.org/10.1109/EEBDA53927.2022.9744760

Ma N, Xu K, Xia XC, et al., 2022. Reinforcement learning-based dynamic anti-jamming power control in UAV networks: an effective jamming signal strength based approach. *IEEE Commun Lett*, 26(10):2355-2359. https://doi.org/10.1109/LCOMM.2022.3193309

Mnih V, Kavukcuoglu K, Silver D, et al., 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529-533. https://doi.org/10.1038/nature14236

Ning WL, Huang XY, Yang K, et al., 2020. Reinforcement learning enabled cooperative spectrum sensing in cognitive radio networks. *J Commun Netw*, 22(1):12-22. https://doi.org/10.1109/JCN.2019.000052

Pirayesh H, Zeng HC, 2022. Jamming attacks and anti-jamming strategies in wireless networks: a comprehensive survey. *IEEE Commun Surv Tut*, 24(2):767-809. https://doi.org/10.1109/COMST.2022.3159185

Ramachandran P, Zoph B, Le QV, 2017. Searching for activation functions. https://arxiv.org/abs/1710.05941

Sharma H, Kumar N, Tekchandani R, 2023. Mitigating jamming attack in 5G heterogeneous networks: a federated deep reinforcement learning approach. *IEEE Trans Veh Technol*, 72(2):2439-2452. https://doi.org/10.1109/TVT.2022.3212966

Shen ZX, Xu K, Xia XC, 2021. 2D fingerprinting-based localization for mmWave cell-free massive MIMO systems. *IEEE Commun Lett*, 25(11):3556-3560. https://doi.org/10.1109/LCOMM.2021.3109645

Summers TA, Wilson SG, 1998. SNR mismatch and online estimation in turbo decoding. *IEEE Trans Commun*, 46(4):421-423. https://doi.org/10.1109/26.664291

Sun YF, An K, Luo JS, et al., 2021. Intelligent reflecting surface enhanced secure transmission against both jamming and eavesdropping attacks. *IEEE Trans Veh Technol*, 70(10):11017-11022. https://doi.org/10.1109/TVT.2021.3104580

Sutton RS, Barto AG, 2018. Reinforcement Learning: an Introduction. MIT Press, Cambridge, USA.

Takizawa K, Sasaki S, Zhou J, et al., 2002. Online SNR estimation for parallel combinatorial SS systems in Nakagami fading channels. Proc Global Telecommunications Conf, p.1239-1243. https://doi.org/10.1109/GLOCOM.2002.1188395

Tang X, Wang DW, Zhang RN, et al., 2021. Jamming mitigation via aerial reconfigurable intelligent surface: passive beamforming and deployment optimization. *IEEE Trans Veh Technol*, 70(6):6232-6237. https://doi.org/10.1109/TVT.2021.3077662

van Hasselt H, Guez A, Silver D, 2016. Deep reinforcement learning with double Q-learning. Proc 30th AAAI Conf on Artificial Intelligence, p.2094-2100.

Wang PL, Fang J, Yuan XJ, et al., 2020. Intelligent reflecting surface-assisted millimeter wave communications: joint active and passive precoding design. *IEEE Trans Veh Technol*, 69(12):14960-14973. https://doi.org/10.1109/TVT.2020.3031657

Wang W, Zhang W, 2021. Joint beam training and positioning for intelligent reflecting surfaces assisted millimeter wave communications. *IEEE Trans Wirel Commun*, 20(10):6282-6297. https://doi.org/10.1109/TWC.2021.3073140

Wang W, Zhang W, 2022a. Intelligent reflecting surface configurations for smart radio using deep reinforcement learning. *IEEE J Select Areas Commun*, 40(8):2335-2346. https://doi.org/10.1109/JSAC.2022.3180787

Wang W, Zhang W, 2022b. Jittering effects analysis and beam training design for UAV millimeter wave communications. *IEEE Trans Wirel Commun*, 21(5):3131-3146. https://doi.org/10.1109/TWC.2021.3118558

Wei L, Huang CW, Alexandropoulos GC, et al., 2021. Channel estimation for RIS-empowered multi-user MISO wireless communications. *IEEE Trans Commun*, 69(6):4144-4157. https://doi.org/10.1109/TCOMM.2021.3063236

Wu QQ, Zhang R, 2019. Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming. *IEEE Trans Wirel Commun*, 18(11):5394-5409. https://doi.org/10.1109/TWC.2019.2936025

Wu QQ, Zhang R, 2020. Towards smart and reconfigurable environment: intelligent reflecting surface aided wireless network. *IEEE Commun Mag*, 58(1):106-112. https://doi.org/10.1109/MCOM.001.1900107

Xiao L, Hong SY, Xu SY, et al., 2022. IRS-aided energy-efficient secure WBAN transmission based on deep reinforcement learning. *IEEE Trans Commun*, 70(6):4162-4174. https://doi.org/10.1109/TCOMM.2022.3169813

Xiao ZC, Gao B, Liu SC, et al., 2018. Learning based power control for mmWave massive MIMO against jamming. Proc IEEE Global Communications Conf, p.1-6. https://doi.org/10.1109/GLOCOM.2018.8647173

Xu JD, Yuen C, Huang CW, et al., 2023. Reconfiguring wireless environments via intelligent surfaces for 6G: reflection, modulation, and security. *Sci China Inf Sci*, 66(3):130304. https://doi.org/10.1007/s11432-022-3626-5

Xu JW, Wang KH, Zhang X, et al., 2021. Anti-jamming strategy based on game theory in single-channel UAV communication network. Proc 6$^{\text{th}}$ Int Conf on Fog and Mobile Edge Computing, p.1-7. https://doi.org/10.1109/FMEC54266.2021.9732602

Yang HL, Xiong ZH, Zhao J, et al., 2020. Intelligent reflecting surface assisted anti-jamming communications based on reinforcement learning. Proc IEEE Global Communications Conf, p.1-6. https://doi.org/10.1109/GLOBECOM42002.2020.9322599

Yang HL, Xiong ZH, Zhao J, et al., 2021a. Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications. *IEEE Trans Wirel Commun*, 20(1):375-388. https://doi.org/10.1109/TWC.2020.3024860

Yang HL, Xiong ZH, Zhao J, et al., 2021b. Intelligent reflecting surface assisted anti-jamming communications: a fast reinforcement learning approach. *IEEE Trans Wirel Commun*, 20(3):1963-1974. https://doi.org/10.1109/TWC.2020.3037767

Yu L, Li YS, Pan C, et al., 2017. Anti-jamming power control game for data packets transmission. Proc 17$^{\text{th}}$ Int Conf on Communication Technology, p.1255-1259. https://doi.org/10.1109/ICCT.2017.8359836

Zhang SW, Zhang R, 2020. Capacity characterization for intelligent reflecting surface aided MIMO communication. *IEEE J Select Areas Commun*, 38(8):1823-1838. https://doi.org/10.1109/JSAC.2020.3000814

Zhang ZD, Zhang DX, Qiu RC, 2020. Deep reinforcement learning for power system applications: an overview. *CSEE J Power Energy Syst*, 6(1):213-225. https://doi.org/10.17775/CSEEJPES.2019.00920