*electronics*

MDPI

*Article*

# RIS-Assisted Robust Beamforming for UAV Anti-Jamming and Eavesdropping Communications: A Deep Reinforcement Learning Approach

**Chao Zou [1,2], Cheng Li [2,*], Yong Li [2,*] and Xiaojuan Yan [3]**

[1] College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20211249675@nuist.edu.cn
[2] The Sixty-Third Research Institute, National University of Defense Technology, Nanjing 210007, China
[3] School of Information Science and Engineering, Southeast University, Nanjing 214135, China; yxj9609@163.com
* Correspondence: licheng@nudt.edu.cn (C.L.); liy771121@163.com (Y.L.)

**Abstract:** The reconfigurable intelligent surface (RIS) has been widely recognized as a rising paradigm for physical layer security due to its potential to substantially adjust the electromagnetic propagation environment. In this regard, this paper adopted the RIS deployed on an unmanned aerial vehicle (UAV) to enhance information transmission while defending against both jamming and eavesdropping attacks. Furthermore, an innovative deep reinforcement learning (DRL) approach is proposed with the purpose of optimizing the power allocation of the base station (BS) and the discrete phase shifts of the RIS. Specifically, considering the imperfect illegitimate node's channel state information (CSI), we first reformulated the non-convex and non-conventional original problem into a Markov decision process (MDP) framework. Subsequently, a noisy dueling double-deep Q-network with prioritized experience replay (Noisy-D3QN-PER) algorithm was developed with the objective of maximizing the achievable sum rate while ensuring the fulfillment of the security requirements. Finally, the numerical simulations showed that our proposed algorithm outperformed the baselines on the system rate and at transmission protection level.

**Keywords:** reconfigurable intelligent surface; unmanned aerial vehicle; anti-jamming; robust beamforming design; deep reinforcement learning



**Citation:** Zou, C.; Li, C.; Li, Y.; Yan, X. RIS-Assisted Robust Beamforming for UAV Anti-Jamming and Eavesdropping Communications: A Deep Reinforcement Learning Approach. *Electronics* **2023**, *12*, 4490. https://doi.org/10.3390/electronics12214490

Academic Editor: Francisco Falcone

Received: 14 September 2023
Revised: 18 October 2023
Accepted: 30 October 2023
Published: 1 November 2023



## 1. Introduction

Recently, the advancement of next-generation wireless communications has led to exponential growth in data transmission and connected nodes [1]. However, owing to the open nature of wireless channels, wireless communications are progressively susceptible to active jamming and passive eavesdropping [2,3]. With this as the focus, the academic community has studied various techniques to combat jamming and eavesdropping attacks, e.g., power control [4], frequency hopping [5], artificial-noise-aided beamforming [6], and cooperative relaying scheme [7]. However, power control cannot handle the jamming attacks with high power, and frequency hopping consumes additional spectrum resources. On the other hand, releasing artificial noise consumes extra power, and employing relays may incur additional hardware cost [4–7].

The above-mentioned shortcomings have motivated a new paradigm called the reconfigurable intelligent surface (RIS) [8]. This technology has recently been regarded as a promising solution for enhancing the power/spectral efficiency of wireless communication systems [8–11]. Specifically, the RIS consists of massive passive elements, which can dynamically adjust the reflection coefficient on the elements according to the needs of different communication scenarios to increase the received signal power or significantly reduce the impact of interference in the network [9–11]. Therefore, the RIS has garnered

extensive research attention in the domain of secure communications [12–25]. However, in the face of increasingly complex electromagnetic environments, there is an urgent need for highly efficient and reliable beamforming algorithms for RIS-aided secure communications.

### *1.1. Related Works*

In recent years, several fundamental technical challenges of RIS-assisted secure communication systems have been addressed [12–14]. In [12], the joint beamforming scheme was proposed to protect secure transmission from eavesdropping attacks, where several optimization algorithms were applied, including alternating optimization (AO) and semidefinite relaxation (SDR). To maximize the secrecy rate of the RIS-assisted Gaussian multiple-input multiple-output (MIMO) channel, the authors in [14] used the AO algorithm to jointly optimize the transmit covariance at the transmitter and the phase shift coefficient at the RIS and further proposed the minimization–maximization (MM) algorithm to optimize the local optimal phase shift. However, these works assumed that the base station (BS) can acquire the ideal channel state information (CSI) of all nodes, which is impractical due to the uncooperative relationship between the BS and the illegitimate nodes. To tackle this matter, a robust algorithm has been developed to jointly optimize active beamforming and passive reflecting beamforming to secure the wireless transmission system against jammer attacks, where the CSI of illegitimate nodes at the BS is not completely known [15–17]. In addition, the authors in [18] iteratively solved an energy-efficient secure transmission problem with the probabilistic outage constraint by low-complexity first-order algorithms in the presence of imperfect information about the eavesdropper's channel state.

Considering that the actual communication environment may become increasingly complex, such as in densely populated areas with clusters of buildings, the links between the RIS and various nodes may encounter obstacles. Unmanned aerial vehicles (UAVs) have been widely used in complex communication networks due to their low cost and flexible maneuverability [19–23]. In addition, when we mount the RIS on a UAV, the channel attenuation of the ground-to-air channel is much lower than that of the ground channel, which can significantly reduce the energy loss of passive reflection. The authors in [21] utilized UAVs carrying reflective surfaces to facilitate power delivery to intelligent devices, while simultaneously transmitting information. Liu et al. used an AO framework to study a multi-controllable system for RIS-aided UAV communication [22]. In [23], the authors studied the secrecy problem of RIS-based integrated satellite UAV relay networks with multiple eavesdroppers.

The obvious challenge is that traditional convex optimization algorithms may be less efficient for large-scale communication systems. Besides, the practical RIS's coefficient adjustment is discrete, which makes the traditional algorithms no longer applicable. Benefiting from the rapid development of artificial intelligence (AI), reinforcement learning (RL) has attracted much interest in beamforming design in RIS-assisted wireless communication systems [24–30], which can effectively deal with the large-scale discrete RIS's coefficients. The authors in [24] proposed a passive phase shift design to maximize the downlink received signal-to-noise ratio based on deep reinforcement learning (DRL). In [25], DRL and extremum seeking control were incorporated for the purposes of model-free control of the RIS. In response to increased network demand and interference challenges from nearby UAV cells, a direct collaborative-communication-enabled multi-agent decentralized double-deep Q-network (CMAD–DDQN) approach facilitates direct collaboration among UAVs, optimizing their 3D flight trajectories to maximize energy efficiency while outperforming existing methods by up to 85% [26]. However, these works did not explore the issue of the security of AI in RIS-enhanced communication systems. In [27,28], the authors proposed secure DRL-based beamforming methods for protecting RIS-assisted wireless communications from active jamming or passive eavesdropping. Furthermore, in order to maximize the energy efficiency of multi-UAV-assisted wireless coverage, the authors in [29] proposed a cooperative multi-agent decentralized double-deep Q-network (MAD-DDQN) approach, but the algorithm could not be directly applied to optimize the

reflecting beamforming for the RIS. To the best of our knowledge, no exiting work has considered the design of DRL in RIS-assisted secure transmission strategies in the presence of both jammers and eavesdroppers and imperfect CSI conditions.

### *1.2. Contributions*

In this paper, we aimed to delve into the anti-jamming and anti-eavesdropping problems in an RIS-assisted UAV transmission system and introduce an innovative robust DRL-based approach to design discrete RIS coefficients in the presence of imperfect CSI from illegitimate nodes. In conclusion, our principal contributions are itemized as follows:

- Considering the illegitimate nodes' imperfect CSI, the joint optimization problem of power allocation at the BS and reflecting beamforming at the RIS is formulated to maximize the achievable system rate, while ensuring fulfillment of the security requirements.
- To cope with the non-convex and non-conventional optimization problem, we first used a robust method to process the imperfect CSI, and subsequently, the optimization problem was reformulated into a Markov decision process (MDP) framework. Then, a noisy dueling double-deep Q-network with prioritized experience replay (Noisy-D3QN-PER) algorithm with safety performance awareness is proposed, where the D3QN is the improvement of the DQN, the NoisyNet can be encouraged to avoid falling into local optima, and the PER accelerates the convergence.
- The numerical results indicated that the Noisy-D3QN-PER algorithm outperformed conventional approaches in improving the safety performance protection level and achievable sum rate. For example, the proposed algorithm improved the system rate and transmission protection level by 27.43% and 11.11%, respectively, compared to the conventional DQN of the benchmark scheme.

## 2. System Model and Problem Formulation

### *2.1. System Model*

Figure 1 depicts the secure transmission scenario under consideration, wherein a BS aided by a fixed aerial RIS-UAV seeks to establish dependable links with $K$ single-antenna users in the presence of a smart jammer and a single-antenna eavesdropper. Here, we assumed that the BS and the jammer are equipped with $N$, $NJ$ antennas, respectively, and the RIS deployed on the UAV has $L$ reflecting units. For the ease of exposition, we further denote the channel matrix between the BS and the RIS-UAV, the smart jammer and the RIS-UAV, the BS and the $k$-th user, the RIS-UAV and the $k$-th user, the smart jammer and the $k$-th user, the BS and the eavesdropper, and the RIS-UAV and the eavesdropper by $\mathbf{G}_{\mathrm{BR}} \in \mathbb{C}^{L\times N}$, $\mathbf{G}_{\mathrm{JR}} \in \mathbb{C}^{L\times NJ}$, $\mathbf{h}_{\mathrm{BU},k}^{H} \in \mathbb{C}^{1\times N}$, $\mathbf{h}_{\mathrm{RU},k}^{H} \in \mathbb{C}^{1\times L}$, $\mathbf{h}_{\mathrm{JU},k}^{H} \in \mathbb{C}^{1\times NJ}$, $\mathbf{h}_{\mathrm{BE}}^{H} \in \mathbb{C}^{1\times N}$, and $\mathbf{h}_{\mathrm{RE}}^{H} \in \mathbb{C}^{1\times L}$, respectively. Due to the cooperation between the legitimate nodes, we assumed that the CSI of the involved legitimate channel $\left\{\mathbf{G}_{\mathrm{BR}}, \mathbf{h}_{\mathrm{BU},k}, \mathbf{h}_{\mathrm{RU},k}\right\}$ is accurately available at the BS. However, in light of the expectation that illegitimate nodes will not collaborate with the BS to perform channel estimation, we took the practical assumption into account that the CSI of illegitimate channels, namely $\left\{\mathbf{G}_{\mathrm{JR}}, \mathbf{h}_{\mathrm{JU},k}, \mathbf{h}_{\mathrm{BE}}, \mathbf{h}_{\mathrm{RE}}\right\}$, cannot be perfectly obtained. To elaborate on this, considering a more-practical and more-general situation, rather than using a statistical or bounded uncertainty model [15], we further characterized the illegitimate CSI as a given angle-based range, i.e.,

$$\begin{aligned}\Delta_{J,G} =&\{\mathbf{h}_{\mathrm{JR}} | \theta_{G}^{J,R} \in \left[\theta_{G,L}^{J,R}, \theta_{G,U}^{J,R}\right], \varphi_{G}^{J,R} \in \left[\varphi_{G,L}^{J,R}, \varphi_{G,U}^{J,R}\right],\\ &\left|g_{G}^{J}\right| \in \left[\left|g_{G,L}^{J}\right|, \left|g_{G,U}^{J}\right|\right]\},\end{aligned} \tag{1}$$

$$\begin{aligned}\Delta_{J,h} =&\{\mathbf{h}_{\mathrm{JU},k} | \theta_{k}^{J} \in \left[\theta_{k,L}^{J}, \theta_{k,U}^{J}\right], \varphi_{k}^{J} \in \left[\varphi_{k,L}^{J}, \varphi_{k,U}^{J}\right],\\ &\left|g_{k}^{J}\right| \in \left[\left|g_{k,L}^{J}\right|, \left|g_{k,U}^{J}\right|\right], \forall k \in K\},\end{aligned} \tag{2}$$

$$\Delta_E = \{\mathbf{h}_i | \theta_i^E \in \left[\theta_{i,L}^E, \theta_{i,U}^E\right], \varphi_i^E \in \left[\varphi_{i,L}^E, \varphi_{i,U}^E\right], \left|g_i^E\right| \in \left[\left|g_{i,L}^E\right|, \left|g_{i,U}^E\right|\right], i \in (\mathrm{BE}, \mathrm{RE})\}, \tag{3}$$

where $\Delta_J = \{\Delta_{J,h}, \Delta_{J,G}\}$, $\theta_L$ represents the minimum vertical angle of AoD (AoA), while $\theta_U$ represents the maximum vertical angle of AoD (AoA). Similarly, $\varphi_L$ represents the minimum horizontal angle of AoD (AoA), while $\varphi_U$ represents the maximum horizontal angle of AoD (AoA). Finally, $g_L$ and $g_U$ represent the lower and upper limits of the channel gain amplitude, respectively.

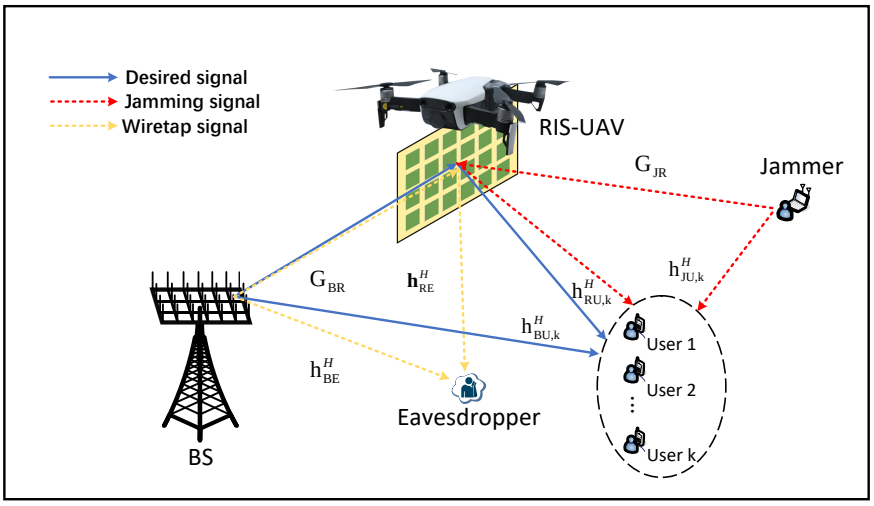


**Figure 1.** System model.

Let $s_k$ be defined as the information symbol transmitted to the $k$-th user, satisfying $\mathrm{E}[s_k]$ and $\mathrm{E}\left[|s_k|^2\right] = 1$. Before transmission, $s_k$ should be multiplied by the beamforming vector $\mathbf{w}_k \in \mathbb{C}^{N\times 1}$ satisfying $\|\mathbf{w}_k\|^2 = 1$. Consequently, the total transmitted signal at the BS can be written as $\mathbf{x} = \sum_{k=1}^{K} \sqrt{P_k}\mathbf{w}_k s_k$, where $P_k$ denotes the allocated transmit power assigned to the $k$-th user. Meanwhile, the smart jammer endeavors to disrupt the legitimate communication by transmitting jamming signal $\mathbf{w}_\mathrm{J} s_\mathrm{J} \in \mathbb{C}^{NJ\times 1}$. As such, the RIS receives the superimposed signals and imposes the phase shift coefficient $\mathbf{\Phi} = diag\left(\beta_1 e^{j\phi_1}, ..., \beta_l e^{j\phi_l}, ..., \beta_L e^{j\phi_L}\right)$ on them, where $\phi_l \in [0, 2\pi]$ and $\beta_l \in [0, 1]$ represent the phase shift and the amplitude of the $l$-th RIS reflective element, respectively. Hence, the received signal at the $k$-th user and the eavesdropper can be, respectively, expressed by

$$y_{\mathrm{U},k} = \bar{\mathbf{h}}_{\mathrm{BU},k}\sqrt{P_k}\mathbf{w}_k s_k + \sum_{i\neq k}\bar{\mathbf{h}}_{\mathrm{BU},k}\sqrt{P_i}\mathbf{w}_i s_i + \bar{\mathbf{h}}_{\mathrm{JU},k}\mathbf{w}_\mathrm{J} s_\mathrm{J} + n_{\mathrm{U},k}, \tag{4}$$

$$y_\mathrm{E} = \bar{\mathbf{h}}_\mathrm{BE}\sqrt{P_k}\mathbf{w}_k s_k + \sum_{i\neq k}\bar{\mathbf{h}}_\mathrm{BE}\sqrt{P_i}\mathbf{w}_i s_i + n_\mathrm{E}, \tag{5}$$

where $\bar{\mathbf{h}}_{\mathrm{BU},k} = \mathbf{h}_{\mathrm{RU},k}^H \mathbf{\Phi}\mathbf{G}_\mathrm{BR} + \mathbf{h}_{\mathrm{BU},k}^H$, $\bar{\mathbf{h}}_{\mathrm{JU},k} = \mathbf{h}_{\mathrm{RU},k}^H \mathbf{\Phi}\mathbf{G}_\mathrm{BR} + \mathbf{h}_{\mathrm{JU},k}^H$, $\bar{\mathbf{h}}_\mathrm{BE} = \mathbf{h}_{\mathrm{RE},k}^H \mathbf{\Phi}\mathbf{G}_\mathrm{BR} + \mathbf{h}_\mathrm{BE}^H$. The symbol $n_{\mathrm{U},k} \sim \mathcal{CN}\left(0, \sigma_{\mathrm{U},k}^2\right)$ represents the additive white Gaussian noise (AWGN) at the $k$-th user, and $n_\mathrm{E} \sim \mathcal{CN}\left(0, \sigma_\mathrm{E}^2\right)$ is the AWGN at the eavesdropper. Hence, the achievable system rate of the $k$-th user and the wiretap rate of the eavesdropper can be, respectively, expressed as

$$R_{\mathrm{U},k} = \log_2\left(1 + \frac{P_k\left|\widetilde{\mathbf{h}}_{\mathrm{BU},k}\mathbf{w}_k\right|^2}{\sum\limits_{i\neq k} P_i\left|\widetilde{\mathbf{h}}_{\mathrm{BU},k}\mathbf{w}_i\right|^2 + \left|\widetilde{\mathbf{h}}_{\mathrm{JU},k}\mathbf{w}_{\mathrm{J}}\right|^2 + \sigma_{U,k}^2}\right), \tag{6}$$

$$R_{\mathrm{E},k} = \log_2\left(1 + \frac{P_k\left|\widetilde{\mathbf{h}}_{\mathrm{BE}}\mathbf{w}_k\right|^2}{\sum\limits_{i\neq k} P_i\left|\widetilde{\mathbf{h}}_{\mathrm{BE}}\mathbf{w}_i\right|^2 + \sigma_E^2}\right). \tag{7}$$

The secrecy rate of the $k$-th user can be written as

$$R_{\mathrm{sec},k} = [R_{\mathrm{U},k} - R_{\mathrm{E},k}]^+, \tag{8}$$

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

### *2.2. Problem Formulation*

Our objective is to maximize the achievable sum rate through jointly optimizing the transmit power allocation $\{P_k\}_{k\in\mathrm{K}}$ and the reflecting beamforming matrix $\mathbf{\Phi}$ under the imperfect illegitimate node's CSI, while meeting the worst-case secrecy/achievable rate constraints. As such, the optimization problem can be formulated as

$$\begin{aligned}
\mathcal{F}: \quad & \max_{\{P_k\}_{k\in K},\mathbf{\Phi}} \min_{\Delta J} \sum_{k\in K} R_{\mathrm{U},k}, \\
& s.t.\ \mathrm{C1}: \min_{\Delta E} R_{\mathrm{sec},k} \geq R_{\mathrm{sec},k}^{\min},\ \forall k \in K, \\
& \quad \mathrm{C2}: \min_{\Delta J} R_{\mathrm{U},k} \geq R_k^{\min},\ \forall k \in K, \\
& \quad \mathrm{C3}: \textstyle\sum_{k=1}^{K} P_k \leq P_{\max}, \\
& \quad \mathrm{C4}: \left|\beta_l e^{j\phi_l}\right| = 1,\ 0 \leq \theta_l \leq 2\pi,\ \forall l \in L,
\end{aligned} \tag{9}$$

where $R_{\mathrm{sec},k}^{\min}$ and $R_k^{\min}$ represent the minimum secrecy rate and the target rate of the $k$-th user. The power allocation is restricted to C3 due to the limited energy supply at the BS, and $P_{\max}$ is the BS's maximum transmit power. Note that, due to the non-convexity of both the objective function and the constraints, (9) is a non-convex and non-trivial problem. Many existing traditional optimization methods, such as the SDR algorithm and the AO algorithm, obtain the solution in each time slot, where the correlation of consecutive instants is ignored, and phase adjustment is usually discrete in form on practical RIS elements, which leads traditional methods to no longer be applicable. In addition, in the scenario we are considering, the jammer is intelligent and can change the unknown jamming strategy in real-time. In order to be able to optimize in real-time and from the perspective of long-term interests, instead of directly solving this problem mathematically, we propose a robust DRL-based approach that can constantly interact with the environment that contains eavesdroppers and smart jammers to learn the optimal solution.

## 3. DRL-Based Algorithm Design

### *3.1. Robust Channel Processing*

As stated in Section 2, the imperfect CSI results in infinite non-convexity in both the objective function and constraints. With this as the focus, according to the works [28–30], the equivalent worst-case CSI of the illegitimate channel that can be obtained through utilizing the discretization method is given, respectively, by

$$\tilde{\mathbf{G}}_{\mathrm{JR}}=\sum_{i_1=1}^{NJ_1}\sum_{i_2=1}^{NJ_2}\sum_{j_1}^{N_1}\sum_{j_2}^{N_2}(1/(Nj+N))\mathbf{G}_{\mathrm{JR}}^{(i_1,i_2)}, \tag{10}$$

$$\tilde{\mathbf{h}}_{\mathrm{M}}=\sum_{i_1=1}^{M_{N_1}}\sum_{i_2=1}^{M_{N_2}}(1/M_N)\mathbf{h}_{\mathrm{M}}^{(i_1,i_2)}, \tag{11}$$

where $\mathrm{M}\in(\mathrm{BE},\mathrm{RE},\{\mathrm{JU},k\})$, $M_N\in(N,L,NJ)$, and $\mathbf{G}_{\mathrm{JR}}^{(i_1,i_2)}$, $\mathbf{h}_{\mathrm{JU},k}^{(i_1,i_2)}$, $\mathbf{h}_{\mathrm{BE}}^{(i_1,i_2)}$, $\mathbf{h}_{\mathrm{RE}}^{(i_1,i_2)}$ are the discrete CSI by uniformly discretizing all the angles in the set of $\Delta_J$ and $\Delta_U$, respectively, i.e.,

$$\theta^{(i_1)}=\theta_L+(i_1-1)(\theta_U-\theta_L)/(Q_1-1), i_1=1,...,Q_1, \tag{12}$$

$$\varphi^{(i_2)}=\varphi_L+(i_2-1)(\varphi_U-\varphi_L)/(Q_2-1), i_2=1,...,Q_2, \tag{13}$$

where $Q_1$ and $Q_2$ are the sample numbers of $\theta$ and $\varphi$. Here, the detail is omitted for brevity, which can be referenced in [31,32].

### *3.2. Overview of DRL*

DRL amalgamates the feature acquisition prowess inherent to deep learning (DL) with the decision-making capabilities intrinsic to RL. It comprises two fundamental constituents: the agent and the environment. The agent continuously improves its strategy by receiving feedback through interactions with the environment to achieve maximum return. This learning process is described as an MDP [33]. The MDP framework can be defined by a tuple $\{S,A,P,R\}$. Herein, $S$ represents the state space denoting the set of observations characterizing the environment. $A$ denotes the set of potential choices. $P$ is the state transition probability denoting the distribution of the next state $s_{t+1}$ given the action $a_t$ taken in the current state $s_t$. Lastly, $R$ is the immediate reward, which provides the quality evaluation $r_t(s_t,a_t)$ of the state–action pair $(s_t,a_t)$. At each time step $t$, the agent obtains the state $s_t\in S$ from the environment and executes an action $a_t\in A$ according to the policy function $\pi(a_t|s_t)=\Pr(A_t=a_t|S_t=s_t)$. Subsequently, the environment will transit to a new state $s_{t+1}$ with probability $\mathrm{P}(s_{t+1}|s_t,a_t)=\Pr(S_{t+1}=s_{t+1}|S_t=s_t,A_t=a_t)$; in the meantime, the agent will receive the immediate reward $r_t\in R$. The agent aims at learning strategies maximizing the long-term reward, i.e., the cumulative discounted future reward $U_t=\sum_{\tau=0}^{\infty}\gamma^\tau R_{t+\tau+1}$, where $\gamma\in[0,1]$ is the discount factor. Therefore, the tuples $(s_1,a_1,r_1,s_2,...,s_{t-1},a_{t-1},r_{t-1},s_t)$ constitute the trajectory in an episode used for the iterative updating of the agent.

To accommodate the proposed algorithm in our problem, we first reformulated Problem (9) into an MDP framework. The corresponding elements of the MDP problem are specified as follows:

**State** $\mathcal{S}$: The state $s_t$ fed back from the RIS-UAV-assisted communication system is given as

$$\left\{\{\mathbf{h}_k^t\}_{k\in K},\{\mathbf{h}_e^t\},\left\{R_{\mathrm{U},k}^{t-1}\right\}_{k\in K}\right\}, \tag{14}$$

where $\mathbf{h}_k$ and $\mathbf{h}_e$ denote the composite channel coefficients of the $k$-th user and eavesdropper, respectively.

**Action** $\mathcal{A}$: Based on the current state $s_t$, the agent needs to make a coordinated decision on the phase shift at the RIS and the power allocation at the BS. Hence, the action $a_t$ at each time step $t$ is given as

$$a_t=\left\{\{\Delta\phi_l\}_{l\in L},\{\Delta P_k\}_{k\in K}\right\}, \tag{15}$$

where $\Delta\phi_l\in\left\{-\frac{\pi}{4},0,\frac{\pi}{4}\right\}$ is the variable for the phase shift of the $l$-th reflection element and $\Delta P_k\in\{-\tilde{p},0,\tilde{p}\}$ is the variable for assigning the $k$-th user's transmit power.

**Reward** $\mathcal{R}$: Our goal was not only to maximize the achievable rate, but also to ensure the system safety performance requirements. Therefore, we designed a composite reward function expressed as

$$r_t = \underbrace{\sum_{k\in K} R_{\mathrm{U},k}}_{\text{basic}} - \underbrace{\sum_{k\in K} \rho_1 p_{\mathrm{U},k} - \sum_{k\in K} \rho_2 p_{\mathrm{E},k}}_{\text{penalty}}, \tag{16}$$

where

$$p_{\mathrm{E,k}} = \begin{cases} 1, if R_{\mathrm{sec},k} < R_{\mathrm{sec},k}^{\min}, \forall k \in K, \\ 0, otherwise. \end{cases}, \tag{17}$$

$$p_{\mathrm{U,k}} = \begin{cases} 1, if R_{\mathrm{U,k}} < R_{\mathrm{U,k}}^{\min}, \forall k \in K, \\ 0, otherwise. \end{cases}. \tag{18}$$

In (16), the base reward is the sum of the rates of all users, and when the constraints in (17) or (18) are not satisfied, we add a penalty term to encourage the agent's behavioral strategy to be closer to our needs. The coefficients $\rho_1$ and $\rho_2$ are the positive constants.

With DRL, a well-known function measuring the expected return for the agent to execute action $a_t$ in the state $s_t$ under the policy $\pi$ is the action value function $Q$:

$$Q_\pi(s_t, a_t; \boldsymbol{w}) = \mathrm{E}[U_t | S_t = s_t, A_t = a_t], \tag{19}$$

where $\boldsymbol{w}$ represents the parameters of the deep neural networks (DNNs). In the learning process, the agent intends to find optimal policy $\pi^*$. Thus, the optimal Q function is expressed as

$$Q_*(s, a) = \max_{\pi} Q_\pi(s, a), \forall s \in S, a \in A. \tag{20}$$

In order to obtain the above equation, the optimal Q function can be constantly approximate by updating the parameter $\boldsymbol{w}$ using the temporal difference (TD) algorithm:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \alpha \nabla_{\boldsymbol{w}} L(\boldsymbol{w}), \tag{21}$$

where $\alpha \in (0, 1)$ is the learning rate for the update on $\boldsymbol{w}$ and $\nabla_{\boldsymbol{w}} L(\boldsymbol{w})$ is the gradient of the loss function $L(\boldsymbol{w})$ with respect to $\boldsymbol{w}$, which is given by

$$L(\boldsymbol{w}) = \left[ r_t + \gamma \max_{a\in A} Q(s_{t+1}, a; \boldsymbol{w}) - Q(s_t, a_t; \boldsymbol{w}) \right]^2, \tag{22}$$

where $r_t + \gamma \max_{a\in A} Q(s_{t+1}, a; \boldsymbol{w})$ refers to the TD target value.

### *3.3. Joint Power Allocation and Reflecting Beamforming Using Noisy-D3QN-PER*

Prevailing reinforcement learning techniques, such as Q-learning, the policy gradient, and the deep Q-network (DQN), have demonstrated notable accomplishments in diverse control tasks. However, regarding the safety beamforming policy requirements discussed in Section 2, the policy gradient algorithm is inadequate for addressing Problem (9), as it involves continuous action space optimization and may converge to suboptimal solutions [34]. Furthermore, although the DQN performs well in in environments characterized by high-dimensional continuous state spaces and discrete action spaces, it remains plagued by several inherent limitations, which adversely affect algorithmic efficacy [35]. Therefore, the Noisy-D3QN-PER algorithm was developed to deal with the challenges in this paper, as shown in Figure 2, which can overcome the constraints associated with the aforementioned methods and significantly enhance the attainable performance.

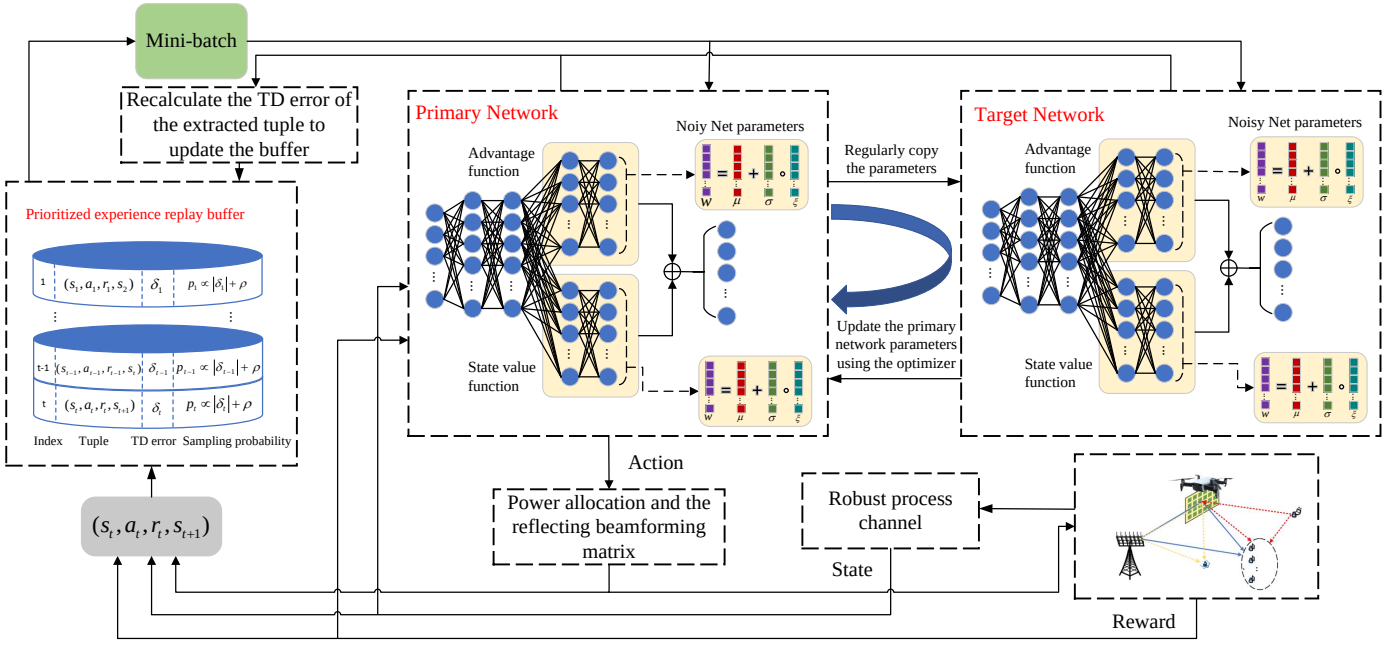


**Figure 2.** The process of the Noisy-D3QN-PER algorithm.

It is noteworthy that a significant disadvantage inherent to the DQN algorithm is over-estimation of the $Q$ function value. The overestimation issue is primarily attributable to two principal factors. First, the process of maximization causes the target value to overestimate the value of the true value. Second, bootstrapping engenders the propagation of bias. In order to address this issue, the double-DQN was adopted in the algorithm [36]. We applied another neural network, i.e., the target network $Q_\pi(s_t, a_t; \boldsymbol{w}^-)$, whose neural network architecture is identical to that of the primary network, but the parameter $\boldsymbol{w}^-$ is different from $\boldsymbol{w}$. Specifically, the primary network was used to choose an action that maximizes the output of the Q function $a^* = \arg\max_{a \in A} Q(s_{t+1}, a; \boldsymbol{w})$, and then, the target network calculates the TD target value $r_t + \gamma Q(s_{t+1}, a^*; \boldsymbol{w}^-)$ with the selected action. Thus, the primary network parameter is updated with the following loss function:

$$L(\boldsymbol{w}) = \left[r_t + \gamma Q\left(s_{t+1}, a^*; \boldsymbol{w}^-\right) - Q(s_t, a_t; \boldsymbol{w})\right]^2. \tag{23}$$

Subsequently, the parameter of the target network is updated with $\boldsymbol{w}$ and $\boldsymbol{w}^-$ every regular interval.

In order to further enhance the algorithm's performance, we incorporated the dueling layer [37], resulting in the formation of the dueling double-DQN (D3QN). The core concept underlying the dueling layer is the decomposition of the optimal action value $Q_*$ into the optimal state value $V_*$ and the optimal advantage $D_*$. As such, the expression of the optimal advantage function is formulated as follows:

$$D_*(s, a) \triangleq Q_*(s, a) - V_*(s). \tag{24}$$

The advantage of modeling the state value function and the advantage function separately is that, in some specific situations, agents only pay attention to the value of the state and do not care about the differences caused by different actions. More specifically, in the optimization problem we are considering, the state values differ greatly, while the action in the same state differs little. The agent pays attention to the difference in the advantage value of different actions, which makes the algorithm converge more stably. As shown in Figure 3, the dueling layer comprises two distinct neural networks. The neural network denoted by $D\left(s, a; \boldsymbol{w}^D\right)$ is an approximation of the optimal advantage function $D_*(s, a)$, and the other neural network is $V(s; \boldsymbol{w}^V)$, which is an approximation of the optimal state value

function $V_*(s)$. The corresponding optimal action value function can be approximated as the following neural network:

$$Q(s,a;\boldsymbol{w}) \triangleq V(s;\boldsymbol{w}^V) + D(s,a;\boldsymbol{w}^D) - \underset{a\in A}{\text{mean}} D(s,a;\boldsymbol{w}^D) \tag{25}$$

where $\underset{a\in A}{\text{mean}} D(s,a;\boldsymbol{w}^D)$ ensures the stability of the parameters in the training process and $\boldsymbol{w} \triangleq (\boldsymbol{w}^V;\boldsymbol{w}^D)$, since, at each iteration, the function $V(s;\boldsymbol{w}^V)$ is updated, which also affects the action value of the other actions.

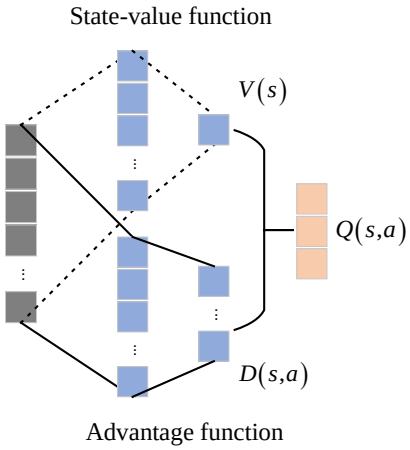


**Figure 3.** Dueling layer.

In addition, there is a dilemma of exploration and exploitation in RL that greatly affects the performance of the algorithm. By gathering more information, or sufficient information, the agent can achieve the optimal long-term strategies on a macro-level at the expense of some short-term benefits. In an effort to attain a good tradeoff between exploration and exploitation, several basic strategies have been proposed, such as Boltzmann exploration and the $\varepsilon$-greedy policy. However, these methods only utilize action dithering, which results in a low exploration rate, especially in complex and unstable environments. Therefore, we propose a NoisyNet technique to improve the exploration efficiency, i.e., adding parameterized noise to the DNN layer [38]. Specifically, as shown in Figure 4, the weight parameter $\boldsymbol{w}$ of the DNN is replaced with

$$\boldsymbol{w} = \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\xi}, \tag{26}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are learnable parameters and denote the mean and standard deviation, respectively, and $\boldsymbol{\xi} \sim \mathcal{N}(0,1)$ is the noise. Here, the term $\circ$ denotes the multiplication of the corresponding elements, i.e.,

$$\boldsymbol{w}_{ij} = \boldsymbol{\mu}_{ij} + \boldsymbol{\sigma}_{ij}\boldsymbol{\xi}_{ij}. \tag{27}$$

Hence, the Q function is written as

$$\tilde{Q}(s,a,\xi;\mu,\sigma) \triangleq Q(s,a;\mu+\sigma\circ\xi). \tag{28}$$

The loss function can be further rewritten as

$$L(\boldsymbol{\mu},\boldsymbol{\sigma}) = (r_t + \gamma\tilde{Q}(s_{t+1},a^*,\boldsymbol{\xi}';\boldsymbol{\mu}^-,\boldsymbol{\sigma}^-) - \tilde{Q}(s_t,a_t,\xi;\boldsymbol{\mu},\boldsymbol{\sigma}))^2, \tag{29}$$

where $a^* = \underset{a\in A}{\arg\max}\, \tilde{Q}(s_{t+1},a,\boldsymbol{\xi};\boldsymbol{\mu},\boldsymbol{\sigma})$, and the noise value $\boldsymbol{\xi}$ is different from $\boldsymbol{\xi}'$. In the training process, noise is added to the training parameters to force the algorithm to minimize the error in the case of parameters with noise, which means that it is forced to tolerate the disturbance of the parameters. It does not matter if the parameters are not strictly equal to

the mean; as long as the parameters are in the neighborhood of the mean, the prediction made by the agent can be reasonable. Therefore, the NoisyNet is not only beneficial to enhance exploration, but also to enhance the robustness of the algorithm.

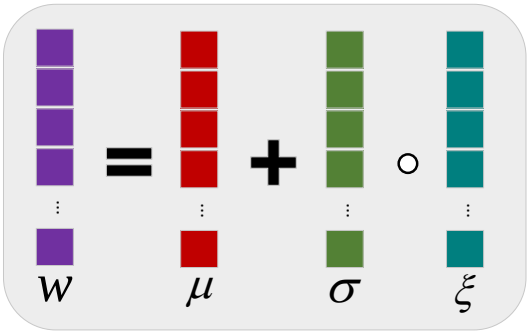


**Figure 4.** NoisyNet.

Experience replay is often utilized in the classical DQN to store and uniformly sample experience transitions, which help in reusing experiences and breaking the correlation of experience transition sequences. However, due to the uncertainty of the jamming strategy of the jammer, the importance of different transitions is different, and uniformly sampling may be ineffective. Hence, we adopted prioritized experience replay (PER) to make the algorithm learn more efficiently and converge faster [39]. PER non-uniformly samples each transition, where the priority of the transition is proportional to its TD error value. Therefore, the sampling probability of transition $j$ is given by

$$P(j) = |\delta_i|^{\alpha_j} / \sum_n |\delta_n|^{\alpha_j}, \tag{30}$$

where $\alpha_j$ adjusts the importance of the priority. In addition, the loss function needs to be multiplied by importance sampling weights to counteract the bias caused by varying the sampling probabilities. Thus, the parameters of the proposed algorithm are update by a mini-batch transition:

$$\boldsymbol{\sigma}_{t+1} = \boldsymbol{\sigma}_t - \alpha_\sigma \nabla_\sigma \frac{1}{m} \sum_{j=1}^{m} ((N \cdot P(j))^{-\omega} L_j(\boldsymbol{\mu},\boldsymbol{\sigma})), \tag{31}$$

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \alpha_\mu \nabla_\mu \frac{1}{m} \sum_{j=1}^{m} ((N \cdot P(j))^{-\omega} L_j(\boldsymbol{\mu},\boldsymbol{\sigma})), \tag{32}$$

where $\alpha_\sigma$ and $\alpha_\mu$ are the learning rate, $m$ is the mini-batch size, $N$ represents the number of samples in the buffer, and $\omega \in (0,1)$ is a hyperparameter that determines the extent to which PER affects the convergence result.

The detailed training process of the Noisy-D3QN-PER algorithm is shown in Algorithm 1. At the beginning of the training, we sample new channel realizations and randomly choose the phase shifts and power allocation to compute the first state $s_0$. Since the NoisyNet is inherently random, exploration can be encouraged. Based on the current state $s_t$, the $\varepsilon$-greedy policy is implemented to select action $a_t$ and, subsequently, receive feedback reward $r_t$ and the next state $s_t + 1$. The transition sequence $(s_t, a_t, r_t, s_{t+1})$ is saved in the experience replay buffer $\mathcal{D}$. After storing enough experiences transitions, the training of the primary networks starts, and mini-batch transitions are selected according to the PER principle and put into the neural networks to obtain the loss function according to Equation (29). Then, the parameters of the primary networks are updated by the Adam optimizer according to Equations (31) and (32), and the target network copies the parameters of the primary networks in every $T_{NET}$ time interval. In addition, each time the experience transitions are sampled, the selected transitions need to update the priority with the new TD error.

**Algorithm 1** Noisy-D3QN-PER algorithm

**Require:** environment simulator, experience replay buffer $\mathcal{D}$, learning rate $\alpha_\sigma$ and $\alpha_\mu$, mini-batch size $m$.

**Initialize**: experience replay buffer $\mathcal{D}$ with size $D$, mini-batch size $m$, primary network parameters $(\mu, \sigma)$, target network parameters $(\mu^- = \mu, \sigma^- = \sigma)$.
**for** each episode = 1, 2, ... , $N^{\text{epi}}$ **do**
  Perceive an initial system state $s$.
  **for** each step = 1, 2, ... , $T$ **do**
    Select action $a_t$ using $\varepsilon$-greedy policy, i.e., select the action that yields the largest action value with a probability of $1 - \varepsilon$, or randomly select from all the possible actions with the probability of $\varepsilon$.
    Receive an instantaneous reward $r_t$, and obtain the next state $s_{t+1}$.
    Store the experience transitions $(s_t, a_t, r_t, s_{t+1})$.
    **if** $|D| \geq m$ **then**
      Sample mini-batch transitions based on PER using (30), and then, update the priority of the selected transition based on its TD error.
      Calculate the loss function for the mini-batch according to (29).
      Perform gradient descent, and update the parameters of the primary networks using (31) and (32).
      **if** $t$ mod $T_{NET} = 0$ **then**
        target network copies the parameters of the primary networks.
      **end if**
    **end if**
  **end for**
**end for**

**Ensure:** joint power allocation and RIS phase shift design strategy.

## 4. Simulation Results

This section presents an evaluation of the Noisy-D3QN-PER algorithm. We varied the maximum transmission power $P$ max between 10 dBm and 30 dBm. The number of antennas on both the BS and the jammer were $N = NJ = 64$, and the number of users was $K = 2$. The fixed deployment height of the RIS-UAV was 100 m. The minimum secrecy rate and target data rate were $R_{\text{sec},k}^{\min} = 0.5$ bits/s/Hz and $R_k^{\min} = 1$ bits/s/Hz, respectively. The background noise at each user and eavesdropper was set to $\sigma_{\text{U},k}^2 = \sigma_{\text{E}}^2 = -90$ dBm. All involved neural networks were considered to be fully connected. The learning rates $\alpha_\sigma$ and $\alpha_\mu$ were set as $\alpha = 0.001$. The initial exploration rate $\varepsilon$ was 1, then was linearly annealed to 0.1. The parameters $\rho_1$ and $\rho_2$ in (12) were set to $\rho_1 = \rho_2 = 2$. The replay buffer size was $D = 100{,}000$, and the mini-batch size was $m = 32$. In addition, the jammer chooses power was from 10 dBm to 30 dBm based on its own jamming strategy, which the BS could not access. Besides, we chose three conventional approaches as benchmarks, namely the classical DQN, the DDQN, and the optimal transmit power allocation without the RIS approach. All of the displayed illustrations are the average results of over 100 independently executed implementations.

Figure 5 shows the average gain graph of the Noisy-D3QN-PER algorithm and the benchmark algorithm. It can be observed that, in the initial phase of training, the algorithms obtained approximately the same reward gain. However, after 100 episodes of training, the Noisy-D3QN-PER algorithm significantly achieved higher gains and faster convergence compared to the benchmark algorithm. This was due to the fact that the preferred empirical playback and competition layers included in the proposed algorithm were better able to adapt to the dynamic and complex interference environment. Specifically, the dueling layer helps to analyze the state bias due to unknown jammer power and unknown location information, and the NoisyNet encourages the exploration of more reflecting beamforming strategies for higher long-term benefits. Moreover, it can be observed that both the DDQN and the proposed algorithm outperformed the classical DQN, which suggests that the use of the DDQN can effectively mitigate the overestimation problem.

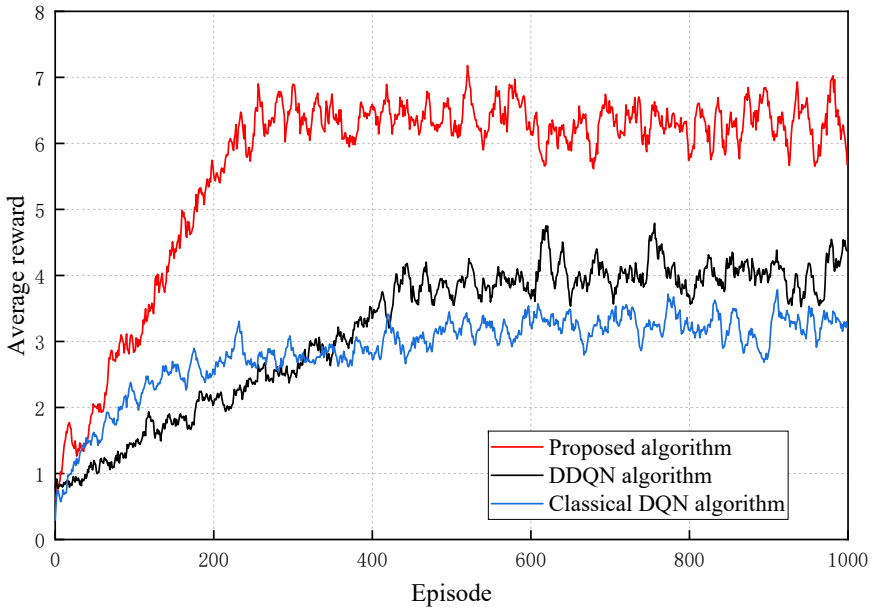


**Figure 5.** Average reward of the Noisy-D3QN-PER algorithm and other comparison approaches.

Figure 6 shows the achievable sum rate with varying maximum transmit power $P_{\max}$. Here, we set $L = 64$. As expected, the proposed algorithm outperformed other approaches. This was because the dueling layers modeling the advantage function and the state value function separately can better focus on states that are less correlated with the current strategy–action relationship and better predict the jammer's strategy when the transmit power changes. Besides, the NoisyNet can prevent the proposed algorithm from becoming stuck at the undesired suboptimal solutions. It can be also observed that the three RIS-UAV-assisted approaches can obtain a much higher achievable rate than that without the RIS, which indicates that deploying the RIS-UAV can efficiently enhance the secure performance. To elaborate on this, the system can enhance the desired signals at the users and eliminate the jamming signal by adjusting the reflecting beamforming at the RIS.

To further highlight the security performance enhancement of the proposed algorithm, the security requirement satisfaction probability (the probability of the satisfaction of the rate constraints [27,28]) of different approaches is shown in Figure 7. It is evident from the figure that the security performance of the optimal PA without the RIS approach cannot be guaranteed when the $P_{\max}$ is low, and the security performance protection improved until $P_{\max}$ was raised to a certain value. However, the other approaches with the RIS-UAV can obtain satisfactory performance at different $P_{\max}$, which further confirmed the superiority of deploying the RIS-UAV in wireless communication systems. Furthermore, it is noteworthy that the proposed algorithm achieved the best result as compared to other conventional approaches. This can be explained by the fact that the comparison approaches usually fell into the suboptimal solution, which only increased the achievable sum rate, but ignored the security performance requirement. However, due to the adopted NoisyNet and the security-aware reward function, the proposed Noisy-D3QN-PER algorithm can explore strategies and make a desirable balance between the security performance and the achievable rate.

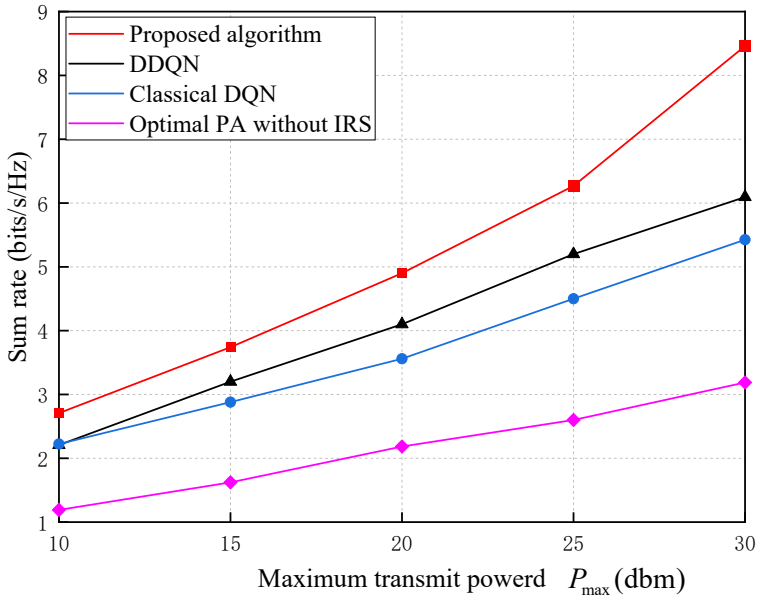


**Figure 6.** Achievable sum rate with varying maximum transmit power $P_{\max}$.

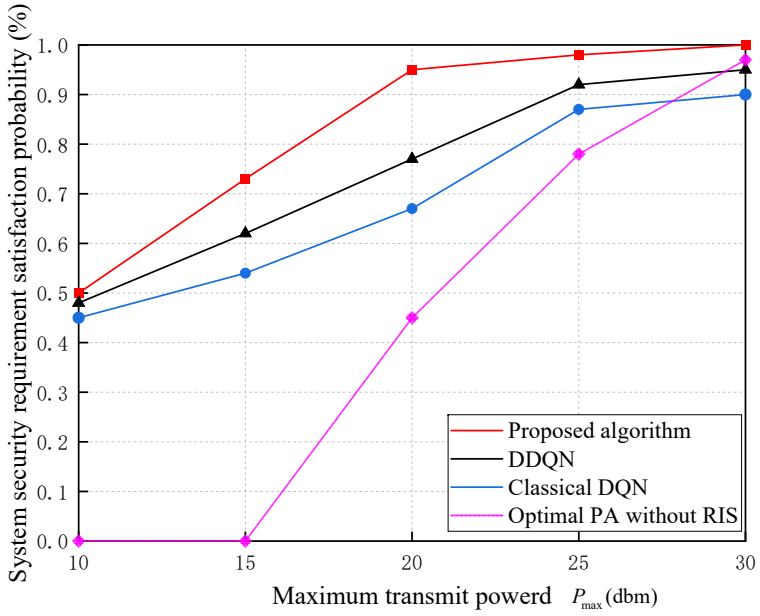


**Figure 7.** System security requirement satisfaction probability versus the maximum transmit power $P_{\max}$.

## 5. Conclusions

This paper delved into the optimization of joint power allocation and reflecting beamforming regarding secure communication via RIS-UAV assistance with imperfect CSI. Specifically, the original optimization problem was formulated into an MDP framework and solved by a Noisy-D3QN-PER algorithm, in which the agent can estimate the unknown jamming strategy through constantly interacting with the environment to quickly adapt to the dynamic environment and, finally, obtain the optimal policy that maximizes the achievable rate and meets the requirements of system security performance, which provides technical support for the realization of the intelligence of the RIS-assisted robust beamforming system. The numerical results confirmed the predominance of the proposed Noisy-D3QN-PER algorithm over other existing conventional approaches in improving the achievable sum rate and system security performance. Although the method proposed in this paper can effectively resist the jamming attack with the uncertainty of the CSI, it is

still necessary to know the variation range of interference. The next step needs to focus on the following two aspects of research: one is to study the anti-jamming method without any interference information; the other is to explore the AI interpretability, to improve the trustworthiness and effectiveness of the AI method.

**Author Contributions:** Conceptualization, C.Z., C.L., Y.L. and X.Y.; methodology, C.L. and Y.L.; validation, C.Z., C.L. and Y.L.; formal analysis, C.Z., C.L., Y.L. and X.Y.; investigation, C.Z., C.L., Y.L. and X.Y.; resources, C.Z., C.L., Y.L. and X.Y.; data curation, C.Z., C.L., Y.L. and X.Y.; writing—original draft preparation, C.Z.; writing—review and editing, C.L. and Y.L.; supervision, C.L. and Y.L.; project administration, C.L.; funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Due to institutional data privacy requirements, our data is unavailable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Cao, H.; Du, J.; Zhao, H.; Luo, D.X.; Kumar, N.; Yang, L.; Yu, F.R. Resource-Ability Assisted Service Function Chain Embedding and Scheduling for 6G Networks With Virtualization. *IEEE Trans. Veh. Technol.* **2021**, *70*, 3846–3859. [CrossRef]
2. Mukherjee, A.; Fakoorian, S.A.A.; Huang, J.; Swindlehurst, A.L. Principles of Physical Layer Security in Multiuser Wireless Networks: A Survey. *IEEE Commun. Surv. Tutorials* **2014**, *16*, 1550–1573. [CrossRef]
3. Zou, Y.; Zhu, J.; Wang, X.; Hanzo, L. A Survey on Wireless Security: Technical Challenges, Recent Advances, and Future Trends. *Proc. IEEE* **2016**, *104*, 1727–1765. [CrossRef]
4. Feng, S.; Haykin, S. Cognitive Risk Control for Anti-Jamming V2V Communications in Autonomous Vehicle Networks. *IEEE Trans. Veh. Technol.* **2019**, *68*, 9920–9934. [CrossRef]
5. Liang, L.; Cheng, W.; Zhang, W.; Zhang, H. Mode Hopping for Anti-Jamming in Radio Vortex Wireless Communications. *IEEE Trans. Veh. Technol.* **2018**, *67*, 7018–7032. [CrossRef]
6. Yan, S.; Yang, N.; Land, I.; Malaney, R.; Yuan, J. Three Artificial-Noise-Aided Secure Transmission Schemes in Wiretap Channels. *IEEE Trans. Veh. Technol.* **2018**, *67*, 3669–3673. [CrossRef]
7. Mayouche, A.; Spano, D.; Tsinos, C.G.; Chatzinotas, S.; Ottersten, B. Learning-Assisted Eavesdropping and Symbol-Level Precoding Countermeasures for Downlink MU-MISO Systems. *IEEE Open J. Commun. Soc.* **2020**, *1*, 535–549. [CrossRef]
8. Liaskos, C.; Nie, S.; Tsioliaridou, A.; Pitsillides, A.; Ioannidis, S.; Akyildiz, I. A New Wireless Communication Paradigm through Software-Controlled Metasurfaces. *IEEE Commun. Mag.* **2018**, *56*, 162–169. [CrossRef]
9. Huang, C.; Zappone, A.; Alexandropoulos, G.C.; Debbah, M.; Yuen, C. Reconfigurable Intelligent Surfaces for Energy Efficiency in Wireless Communication. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 4157–4170. [CrossRef]
10. Hu, S.; Rusek, F.; Edfors, O. Beyond Massive MIMO: The Potential of Data Transmission With Large Intelligent Surfaces. *IEEE Trans. Signal Process.* **2018**, *66*, 2746–2758. [CrossRef]
11. Wu, Q.; Zhang, R. Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 5394–5409. [CrossRef]
12. Cui, M.; Zhang, G.; Zhang, R. Secure Wireless Communication via Intelligent Reflecting Surface. *IEEE Wirel. Commun. Lett.* **2019**, *8*, 1410–1414. [CrossRef]
13. Shen, H.; Xu, W.; Gong, S.; He, Z.; Zhao, C. Secrecy Rate Maximization for Intelligent Reflecting Surface Assisted Multi-Antenna Communications. *IEEE Commun. Lett.* **2019**, *23*, 1488–1492. [CrossRef]
14. Dong, L.; Wang, H.-M. Secure MIMO Transmission via Intelligent Reflecting Surface. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 787–790. [CrossRef]
15. Sun, Y.; An, K.; Luo, J.; Zhu, Y.; Zheng, G.; Chatzinotas, S. Intelligent Reflecting Surface Enhanced Secure Transmission Against Both Jamming and Eavesdropping Attacks. *IEEE Trans. Veh. Technol.* **2021**, *70*, 11017–11022. [CrossRef]
16. Sun, Y.; An K.; Zhu, Y.; Zheng, G.; Wong, K.K.; Chatzinotas, S.; Yin, H.; Liu, P. RIS-Assisted Robust Hybrid Beamforming Against Simultaneous Jamming and Eavesdropping Attacks. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 9212–9231. [CrossRef]
17. Sun, Y.; An, K.; Luo, J.; Zhu, Y.; Zheng, G.; Chatzinotas, S. Outage Constrained Robust Beamforming Optimization for Multiuser IRS-Assisted Anti-Jamming Communications With Incomplete Information. *IEEE Internet Things J.* **2022**, *9*, 13298–13314. [CrossRef]
18. Li, Z.; Wang, S.; Wen, M.; Wu, Y.C. Secure Multicast Energy-Efficiency Maximization With Massive RISs and Uncertain CSI: First-Order Algorithms and Convergence Analysis. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 6818–6833. [CrossRef]
19. Guo, K.; An, K. On the Performance of RIS-Assisted Integrated Satellite-UAV-Terrestrial Networks With Hardware Impairments and Interference. *J. Abbr.* **2022**, *11*, 131–135. [CrossRef]
20. Wu, W.; Zhou, F.; Wang, B.; Wu, Q.; Dong, C.; Hu, R.Q. Unmanned Aerial Vehicle Swarm-Enabled Edge Computing: Potentials, Promising Technologies, and Challenges. *IEEE Wirel. Commun.* **2022**, *29*, 78–85. [CrossRef]

21. Mei, C.; Fang, Y.; Qiu, L. Dual Based Optimization Method for IRS-Aided UAV-Enabled SWIPT System. In Proceedings of the 2022 IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA, 10–13 April 2022; pp. 890–895.
22. Liu, Z.; Zhao, S.; Wu, Q.; Yang, Y.; Guan, X. Joint Trajectory Design and Resource Allocation for IRS-Assisted UAV Communications With Wireless Energy Harvesting. *IEEE Commun. Lett.* **2022**, *26*, 404–408. [CrossRef]
23. Zhou, F.; Li, X.; Alazab, M.; Jhaveri, R.H.; Guo, K. Secrecy Performance for RIS-Based Integrated Satellite Vehicle Networks With a UAV Relay and MRC Eavesdropping. *IEEE Trans. Intell. Veh.* **2023**, *8*, 1676–168. [CrossRef]
24. Feng, K.; Wang, Q.; Li, X.; Wen, C.K. Deep Reinforcement Learning Based Intelligent Reflecting Surface Optimization for MISO Communication Systems. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 745–749. [CrossRef]
25. Wang, W.; Zhang, W. Intelligent Reflecting Surface Configurations for Smart Radio Using Deep Reinforcement Learning. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 2335–2346. [CrossRef]
26. Omoniwa, B.; Galkin, B.; Dusparic, I. Communication-enabled deep reinforcement learning to optimise energy-efficiency in UAV-assisted networks. *Veh. Commun.* **2023**, *43*, 100640. [CrossRef]
27. Yang, H.; Xiong, Z.; Zhao, J.; Niyato, D.; Xiao, L.; Wu, Q. Deep Reinforcement Learning-Based Intelligent Reflecting Surface for Secure Wireless Communications. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 375–388. [CrossRef]
28. Yang, H.; Xiong, Z.; Zhao, J.; Niyato, D.; Wu, Q.; Poor, H.V.; Tornatore, M. Intelligent Reflecting Surface Assisted Anti-Jamming Communications: A Fast Reinforcement Learning Approach. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 1963–1974. [CrossRef]
29. Omoniwa, B.; Galkin, B.; Dusparic, I. Optimizing Energy Efficiency in UAV-Assisted Networks Using Deep Reinforcement Learning. *IEEE Wirel. Commun. Lett.* **2022**, *11*, 1590–1594. [CrossRef]
30. Thanh, P.D.; Giang, H.T.H.; Hong, I.-P. Anti-Jamming RIS Communications Using DQN-Based Algorithm. *IEEE Access* **2022**, *10*, 28422–28433. [CrossRef]
31. Sun, Y.; An, K.; Zhu, Y.; Zheng, G.; Wong, K.K.; Chatzinotas, S.; Ng, D.W.K.; Guan, D. Energy-Efficient Hybrid Beamforming for Multilayer RIS-Assisted Secure Integrated Terrestrial-Aerial Networks. *IEEE Trans. Commun.* **2022**, *70*, 4189–4210. [CrossRef]
32. Sun, Y.; Zhu, Y.; An, K.; Zheng, G.; Chatzinotas, S.; Wong, K.K.; Liu, P. Robust Design for RIS-Assisted Anti-Jamming Communications With Imperfect Angular Information: A Game-Theoretic Perspective. *IEEE Trans. Veh. Technol.* **2022**, *71*, 7967–7972. [CrossRef]
33. Picard, R.W.; Papert, S.; Bender, W.; Blumberg, B.; Breazeal, C.; Cavallo, D.; Machover, T.; Resnick, M.; Roy, D.; Strohecker, C. Affective Learning—A Manifesto. *BT Technol. J.* **2004**, *22*, 253–269. [CrossRef]
34. Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; Riedmiller, M. Deterministic policy gradient algorithms. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 387–395.
35. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv* **2013**, arXiv:1312.5602.
36. Van Hasselt, H.; Guez, A.; Silver, D. Deep Reinforcement Learning with Double Q-learning. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2094–2100.
37. Wang, Z.; Schaul, T.; Hessel, M.; Hasselt, H.; Lanctot, M.; Freitas, N. Dueling network architectures for deep reinforcement learning. *Proc. Mach. Learn. Res.* **2016**, *48*, 1995–2003.
38. Fortunato, M.; Azar, M.G.; Piot, B.; Menick, J.; Osband, I.; Graves, A.; Mnih, V.; Munos, R.; Hassabis, D.; Pietquin, O.; et al. Noisy Networks for Exploration. *arXiv* **2017**, arXiv:1706.10295.
39. Schaul, T.; Quan, J.; Antonoglou, I.; Silver, D. Prioritized Experience Replay. *arXiv* **2015**, arXiv:1511.05952.