sensors

MDPI

Article

RIS-Assisted Multi-Antenna AmBC Signal Detection Using Deep Reinforcement Learning

Feng Jing [1,2,3], Hailin Zhang [1,3], Mei Gao [1,2,3,*], Bin Xue [2,3,4] and Kunrui Cao [2]

[1] School of Telecommunication Engineering, Xidian University, Xi'an 710126, China
[2] School of Information and Communication, National University of Defense Technology, Xi'an 430035, China
[3] Shaanxi Key Laboratory of Intelligence Coordination Networks, Xi'an 710048, China
[4] Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
* Correspondence: gao_mei@nudt.edu.cn

Abstract: Signal detection is one of the most critical and challenging issues in ambient backscatter communication (AmBC) systems. In this paper, a multi-antenna AmBC signal detection method is proposed based on reconfigurable intelligent surface (RIS) and deep reinforcement learning. Firstly, an efficient multi-antenna AmBC system is developed based on RIS, which can achieve information transmission and energy collection simultaneously. Secondly, a smart twin delayed deep deterministic (TD3) AmBC signal detection method is presented, based on deep reinforcement learning. Extensive quantitative and qualitative experiments are performed, which show that the proposed method is more compelling than the outstanding comparison methods.

Keywords: reconfigurable intelligent surface; ambient backscatter communication; signal detection; multi-antenna; deep reinforcement learning

check for updates

Citation: Jing, F.; Zhang, H.; Gao, M.; Xue, B.; Cao, K. RIS-Assisted Multi-Antenna AmBC Signal Detection Using Deep Reinforcement Learning. *Sensors* **2022**, *22*, 6137. https://doi.org/10.3390/s22166137

Academic Editor: Francisco J. González-Castaño

Received: 5 July 2022
Accepted: 10 August 2022
Published: 16 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, with the development of modern digital society, new services and applications such as smart cities, smart houses, digital Internet of Everything, and wired/wireless distributed sensing systems have been constantly emerging, putting forward significantly higher requirements for communication capability [1].

Surprisingly, ambient backscatter communication (AmBC) has been one of the most promising paradigms for the modern communication technologies [2]. Particularly, the reciprocal communication among energy-free equipment is achieved with backscattering radio frequency (RF) signals [3,4]. Different from the traditional backscatter systems [5], the requirements for a part of dedicated infrastructures are weakened with the appearance and growth of AmBC. Furthermore, the exploration of the scarce and expensive wireless spectrum is improved by the available AmBC signals. Particularly, AmBC signal detection is a fundamental and difficult problem to be solved.

So far, there has been some research focused on AmBC signal detection that can be roughly divided into three categories: the physical methods [6], the statistical methods [7], and the learning methods [8]. However, there are several common thorny problems on most of the approaches above: (1) the obtained parameters are incomplete; (2) the direct link interference is serious; (3) the hide intrinsic representations and characteristics are difficult to extract. These issues above make it difficult for the most existing AmBC signal detection methods to be applied well in practice. Therefore, efficient and practical AmBC signal detection methods are urgently needed in both the theoretical and industrial fields.

It is worth noting that deep learning, especially deep reinforcement learning (DRL), has been widely used in wireless communications [9–11]. Learning rules are ingeniously developed, and intrinsic features are learned autonomously from the environment by

DRL, which is difficult to obtain by traditional methods in wireless communication systems. It has been proven that DRL may be an effective way to perform the design, operation, and optimization of the reconfigurable intelligent surface (RIS)-based wireless communication systems.

Therefore, in this paper, the RIS-assisted multi-antenna AmBC signal detection is investigated, and an efficient signal detection method is proposed based on DRL and RIS. Specifically, the contributions of this work can be summarized as follows:

(1) An RIS-assisted multi-antenna AmBC signal model is developed, which can achieve information transmission and energy collection cooperatively.
(2) A twin delayed deep deterministic (TD3) AmBC signal detection method is developed based on deep reinforcement learning and reconfigurable intelligent surface.
(3) Extensive quantitative and qualitative experiments are performed, showing that the presented method is more compelling than the state-of-the-art comparison approaches.

This paper is organized as follows. The related studies are given in Section 2. The AmBC system model and the problem formulation are shown in Section 3. The proposed solutions are described in Section 4. The experimental results, comparisons, and analysis are presented in Section 5. The conclusions are offered in Section 6.

2. Related Works

Hu et al. [12] proposed a signal detection method for a backscatter communication system in a supervised learning manner, where the label signal of AmBC is detected by transforming the detection problem into a classification problem. Particularly, support vector machine and random forest are used to decode label symbols. Simultaneously, efficient features are extracted to minimize the bit error rate of the AmBC system. Experimental results show that with different extended gains, the proposed detector has a lower bit error rate and higher data rate than the traditional minimum mean squared error (MMSE) detectors at low signal-to-noise ratio.

Wang et al. [13] proposed an AmBC signal detection system in the binary phase shift keying (BPSK) modulation based on machine learning. Particularly, BPSK-based backscattering signal encoded by Hadamard code can be decoded with this method. Firstly, the direct path signal is eliminated, the residual signal is associated with the rough environment signal estimation to extract the tag signal's learnable features. Then, the k-nearest neighbor classification algorithm is used to recover the tag signal. The recovered signal is decoded by Hadamard to obtain the original information bits. Finally, the effectiveness of the method is verified by experimental simulation.

In reference [14], the AmBC signal detection problem is transformed into a clustering problem, and two kinds of known labels are transmitted from the tags as prior knowledge to assist cluster initialization and signal detection. By directly using the received signals, two clustering-based detection methods are developed, i.e., one is the clustering of labeled signals (CLS), and the other one is the clustering of labeled and unlabeled signals (CLUS). Both two methods are developed based on the proposed modulation constrained Gaussian mixture model (GMM). Compared with the optimal detection method with perfect correlation channel state information, the proposed method has only a small gap.

An AmBC signal detection method is constructed in reference [15] in an unsupervised learning manner. The characteristics of the received signals are directly utilized, which are clustered with an unsupervised learning means. In addition, the tag bits are transmitted for cluster bit mapping to assist signal detection without the estimation of channel coefficients and noise power. For the spread spectrum gain $N > 1$ and $N = 1$, two detection methods are proposed, in which the features follow different mixed distributions. The detection threshold is derived with the learned parameters to optimize the detection performance. Finally, a large number of simulation results are given to verify the performance of the proposed scheme.

In reference [16], the tag signal detection problem for AmBC systems were studied by adopting the deep transfer learning (DTL) technology. Firstly, a universal DTL-based

tag signal detection framework is designed, which uses a deep neural network (DNN) to implicitly extract the features of communication channels and directly recover the tag symbols. Based on the established pre-trained DNN and a few pilots, a DTL-based likelihood ratio test (DTL-LRT) was obtained through transfer learning. Moreover, exploiting the advantages of the convolutional neural networks' powerful capability in exploring features of data in a matrix form, a covariance matrix-aware neural network (CMNet) for the sample covariance matrix is designed, and a CMNet-based detection algorithm is proposed. Finally, the simulation shows that the proposed CMNet method can achieve a close-to-optimal performance without explicitly obtaining the channel state information.

There are some differences between our proposed method and [12–16]; the main differences are as follows.

First, supervised learning is used in references [12–14,16], manual intervention is required during learning, and the operation is complex and time-consuming. While in the proposed method, the unsupervised learning manner is applied, which does not need manual intervention, and the parameter learning is performed according to the data and model structure; thus, it is easy to operate. Moreover, the proposed method does not require a large number of manual annotation data in advance, so complex environments can be actively explored and better solutions can be obtained.

Second, for [12], it is difficult to deal with missing data and the multi-classification problem, which can be handled well through deep reinforcement learning in our proposed method. Moreover, the relationships between data are easy to be ignored in [12]. While for the proposed method, the explicit data, implicit data, and the relationship between data can be explored. For [13], an appropriate proximity measure K and data preprocessing are required; otherwise, wrong predictions can be achieved by k-nearest neighbor. While this problem does not exist in our proposed method, which is simple, convenient, and flexible. For [14], before signal detection using the CLS and CLUS clustering methods, the signal needs to be preprocessed with the modulation-constrained GGM, which is inflexible and has difficulty dealing with the situations different from agglomerative hierarchical clustering. While in the proposed method, pre-modulation is not needed, and both the same and different agglomerative hierarchical distributions can be processed. For [15], unsupervised learning is used, which also does not need to label the data, but RIS is not used to improve the channel. The difference of signal characteristics when sending 0 and 1 tags is relatively small, which influences the improvement of detection effect. For [16], an AmBC signal detection method is proposed based on a deep transfer learning framework and convolutional neural network, so the requirement of channel state information computation is eliminated. The problem of insufficient annotation data is solved. However, a large amount of pre-training, fine-tuning, and maul intervention are required, which is relatively complex and not flexible enough. While in our proposed method, pre-training, fine-tuning, and maul intervention are not needed.

Moreover, compared with references [12–16], RIS and multi-antenna are introduced into AMBC signal detection in our proposed method; furthermore, the signal detection problem is converted into the optimization problem of the RIS phase shift matrix and receiving antenna combination coefficient, and then deep reinforcement learning without data labels is presented to effectively solve the optimization problem, which improves the signal detection performance of AmBC system.

3. System Model and Problem Formulation

3.1. Scenario Definition

As shown in Figure 1, a downlink AmBC system assisted by RIS is described, which consists of an ambient radio-frequency (RF) source with single antenna, a reader equipped with $K(K \geq 1)$ antennas, a tag equipped with single antenna, and a RIS with $N(N \geq 1)$ reflectors [17]. The RF signal to the backscatter tag is offered by the ambient RF source, while its own information is transmitted. The backscatter tag's information is modulated by on-off modulation to the ambient RF signal. A large number of reflective elements

and passive beamforming are used to enhance the tag-to-reader wireless transmission link, while counteracting the direct link from ambient RF sources to the reader at the RIS. Multiple antennas are utilized at the reader to linearly combine the received signals to enhance the signal's reception ability.

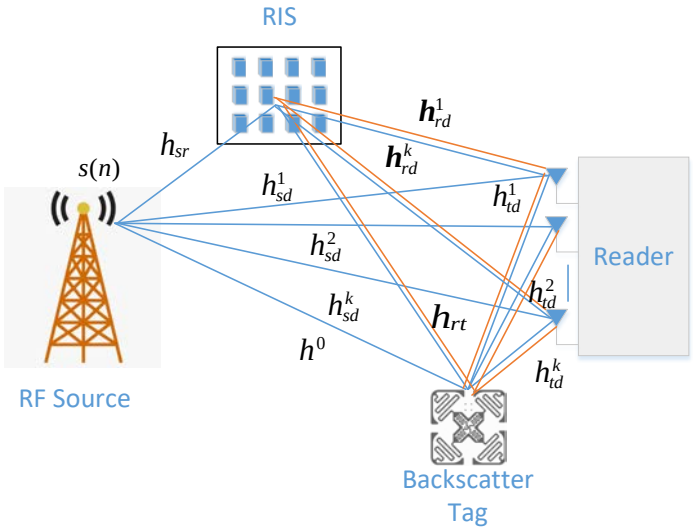


Figure 1. RIS-assisted AmBC system.

In this section, the flat-fading channel is assumed. The channel from the ambient RF source to the tag, from the RF source to the k-th antenna of reader, from the RF source to the n-th reflector of the RIS, from the backscatter tag to the k-th antenna of the reader, from the tag to the n-th reflector of RIS, and from the n-th reflector of RIS to the k-th antenna of the reader, is denoted by $h_0 \in \mathbb{C}^{1\times1}$, $\boldsymbol{h}_{sd} = \left[h_{sd}^1 \quad h_{sd}^2 \quad \cdots \quad h_{sd}^k\right]^T \in \mathbb{C}^{K\times1}$, $\boldsymbol{h}_{sr} = \left[h_{sr}^1 \quad h_{sr}^2 \quad \cdots \quad h_{sr}^n\right]^T \in \mathbb{C}^{N\times1}$, $\boldsymbol{h}_{td} = \left[h_{td}^1 \quad h_{td}^2 \quad \cdots \quad h_{td}^k\right]^T \in \mathbb{C}^{K\times1}$, $\boldsymbol{h}_{rt} = \left[h_{rt}^1 \quad h_{rt}^2 \quad \cdots \quad h_{rt}^k\right]^T \in \mathbb{C}^{N\times1}$, $\boldsymbol{h}_{rd}^k = \left[h_{rd}^1 \quad h_{rd}^2 \quad \cdots \quad h_{rd}^n\right]^T \in \mathbb{C}^{N\times1}$ and $\boldsymbol{H}_{rd} = \left[\boldsymbol{h}_{rd}^1 \quad \boldsymbol{h}_{rd}^2 \quad \cdots \quad \boldsymbol{h}_{rd}^k\right] \in \mathbb{C}^{N\times K}$, respectively. Assume that all the channels follow the independent and identically distributed circularly symmetric complex Gaussian (CSCG) random variables with a mean value of zero and a variance of $\mathrm{E}\left[|h_k|^2\right] = \sigma_h^2$. Assume also that the channel state information (CSI) at the reader is not known, which is consistent with the actual situation. In practice, due to the complexity of the device, the channel often cannot be observed by the reader.

3.2. Problem Description

According to the above scenario, the equivalent complex baseband signal sent by the ambient RF source is denoted as $s(n)$. The signals reach the tag by the reflected RF source-RIS-tag path and the direct RF source-tag path, which is denoted by the symbol $x(n)$, and can be obtained by the following formula:

$$x(n) = \left(h_0 + \boldsymbol{h}_{sr}^H \boldsymbol{\Phi} \boldsymbol{h}_{rt}\right) s(n) \tag{1}$$

where $\boldsymbol{\Phi} = \mathrm{diag}\left(e^{j\theta_1}, e^{j\theta_2}, \cdots, e^{j\theta_N}\right) \in \mathbb{C}^{N\times N}$ denotes the phase shift matrix of the RIS, and $\theta_n \in [0, 2\pi)$ for $n = 1, 2, \cdots, N$ is the phase shift induced by the n-th reflect element. The tag can be switched between reflection and non-reflection by adjusting the impedance, so that two kinds of information "0" and "1" can be carried, which is also called on-off modulation. The tag is a passive device that cannot achieve the complete reflection due to the material limitations, and hence there will be a certain attenuation during reflection. Therefore, after being modulated by backscatter, the transmitted signal of the tag can be expressed as

$$t(n) = \eta d(l) x(n),\ d(l) = 0 \text{ or } 1 \tag{2}$$

where η represents a plural number indicating the attenuation of the tag, $d(l)$ denotes a 1-bit effective information that the tag needs to send to the reader. The tag information's symbol rate is often lower than that of the ambient RF source, due to the limitations in tag equipment and energy consumption. Therefore, assume that L RF sources symbols are spanned in 1 backscattered tag symbol. Thus, L satisfies the following relationship as:

$$L = \frac{T_{tag}}{T_{source}} \tag{3}$$

where T_{source} and T_{tag} are the duration of the RF source and tag symbol, respectively.

3.3. Multi-Antenna Receiver Design

In general, the AmBC receiver owes larger physical size and more power and thus can support multiple antennas receiving together and more complex receiving algorithms than the AmBC tag [18]. In addition, due to the insurmountable BER floor problem in a single antenna, while the spatial diversity can be used with multiple antennas to effectively improve the system's transmission performance, the multiple antennas at the backscatter receiver are exploited for joint reception.

Signals from the RF source and the tag are received at the reader. There are two components in the signal from the RF source, i.e., one is the direct RF source-receiver link, and the other one is the reflected RF source-RIS-receiver link. Furthermore, there are also two components in the signal from the tag; i.e., one is the direct tag-receiver link, and the other one is the reflected tag-RIS-receiver link. Thus, for the k-th receiver antenna, the received signal $r_k(n)$ can be expressed as:

$$r^k(n) = \boldsymbol{H}_{sd}^k s(n) + \boldsymbol{H}_{td}^k t(n) + u^k(n), \tag{4}$$

where $\boldsymbol{H}_{sd}^k = h^k + \boldsymbol{h}_{sr}^H \boldsymbol{\Phi} \boldsymbol{h}_{rd}^k$, $\boldsymbol{H}_{td}^k = h_{td}^k + \boldsymbol{h}_{tr}^H \boldsymbol{\Phi} \boldsymbol{h}_{rd}^k$, and $u^k(n) \sim CN(0, \sigma^2)$ denotes the equivalent noise at the receiver. The received signal at all K receiver antennas is expressed as $\boldsymbol{r}(n) = [r^1(n)\, r^2(n) \cdots r^K(n)]^T$. Separating the two cases of $d = 0$ and $d = 1$, the received signal $\boldsymbol{r}(n)$ can be written as:

$$\boldsymbol{r}(n) = \begin{cases} \boldsymbol{H}_{sd} s(n) + \boldsymbol{u}(n) & , d = 0 \\ \boldsymbol{H}_{sd} s(n) + \boldsymbol{H}_{td} t(n) + \boldsymbol{u}(n) & , d = 1 \end{cases}. \tag{5}$$

Taking Equations (1) and (2) into Equation (7), the following equation can be achieved:

$$\boldsymbol{r}(n) = \begin{cases} \boldsymbol{H}_1 s(n) + \boldsymbol{u}(n) & , d = 0 \\ (\boldsymbol{H}_1 + \boldsymbol{H}_2) s(n) + \boldsymbol{u}(n) & , d = 1 \end{cases} \tag{6}$$

where

$$\boldsymbol{H}_1 = \left(\boldsymbol{h} + \boldsymbol{h}_{sr}^H \boldsymbol{\Phi} \boldsymbol{h}_{rd} \right) \tag{7}$$

$$\boldsymbol{H}_2 = \eta \left(\boldsymbol{h}_{td} + \boldsymbol{h}_{tr}^H \boldsymbol{\Phi} \boldsymbol{h}_{rd}^k \right) \left(h_0 + \boldsymbol{h}_{sr}^H \boldsymbol{\Phi} \boldsymbol{h}_{rt} \right) \tag{8}$$

Finally, the received signals on the multiple antennas are combined by a linear combiner at the receiver. The combine coefficient is represented by the vector $\boldsymbol{f} \in \mathbb{C}^{K \times 1}$, generally, $\|\boldsymbol{f}\|^2 = 1$ without loss of generality. Therefore, the received signal after the combiner can be expressed as $y(n) = \boldsymbol{f}^H \boldsymbol{r}(n)$.

3.4. Energy Detection

After the combiner, energy detection is then used to recover the signal transmitted from the backscatter tag. According to the proposed RIS AmBC system model, both $s(n)$ and $u(n)$ follow the circularly symmetric complex Gaussian (CSCG) distribution, and the

received signal $r(n)$ and the combined signal $y(n)$ also follow CSCG distribution, which can be express as:

$$y(n) \sim \begin{cases} \mathcal{CN}(0, W_0), d = 0 \\ \mathcal{CN}(0, W_1), d = 1 \end{cases} \tag{9}$$

where

$$W_0 = \boldsymbol{f}^H \boldsymbol{H}_1 \boldsymbol{H}_1^H \boldsymbol{f} + \sigma^2 \boldsymbol{I} \tag{10}$$

$$W_1 = \boldsymbol{f}^H (\boldsymbol{H}_1 + \boldsymbol{H}_2)(\boldsymbol{H}_1 + \boldsymbol{H}_2)^H \boldsymbol{f} + \sigma^2 \boldsymbol{I} \tag{11}$$

According to [17], the energy detector is the optimal detector to estimate the information of d:

$$Z = \frac{\rho}{L} \sum_{n=1}^{L} |y(n)|^2 \underset{d=0}{\overset{d=1}{\gtrless}} \frac{\rho W_0 W_1}{W_1 - W_0} \ln \frac{W_1}{W_0} \tag{12}$$

where $\rho = \text{sgn}(W_1 - W_0)$. As shown in Equation (12), when $d = 0$, the mean of Z is $\mu_0 = W_0$, and the variance of Z is $\sigma_0^2 = W_0^2/N$. When $d = 1$, the mean of Z is $\mu_1 = W_1$, and the variance of Z is $\sigma_1^2 = W_1^2/N$. In the case of very large N, applying the central limit law, the BER for backscatter tag signal detection can be obtained when $\rho > 0$ as follows:

$$P_e = \begin{cases} \frac{1}{2} Q\left(\frac{T_h - \mu_0}{\sigma_0}\right) + \frac{1}{2} Q\left(\frac{\mu_1 - T_h}{\sigma_1}\right), \text{ when } \rho > 0 \\ \frac{1}{2} Q\left(\frac{T_h - \mu_1}{\sigma_1}\right) + \frac{1}{2} Q\left(\frac{\mu_0 - T_h}{\sigma_0}\right), \text{ when } \rho < 0 \end{cases} \tag{13}$$

where $T_h = \frac{\mu_0 \mu_1}{\mu_1 - \mu_0} \ln \frac{\mu_1}{\mu_0}$.

3.5. Problem Formulation

Assume that γ denotes the maximum energy difference of the received symbols, $\gamma = \max\left(\frac{\mu_1}{\mu_0}, \frac{\mu_0}{\mu_1}\right)$, and γ is brought into Equation (13). Thus the two cases of Equation (13) can be combined as:

$$P_e = \frac{1}{2} Q\left(\sqrt{L}\left(\frac{\gamma}{\gamma - 1} \ln \gamma - 1\right)\right) + \frac{1}{2} Q\left(\sqrt{L}\left(1 - \frac{\ln \gamma}{\gamma - 1}\right)\right) \tag{14}$$

As shown in Equation (14), the BER of the AmBC system decreases monotonically with γ increasing. This is consistent with the real situation. As γ increases, the gap between μ_0 and μ_1 becomes greater. Thus, the parts of the received signals' conditional probability distributions under $d = 0$ and $d = 1$ are greater. So, the area of overlap between the two probability distribution curves is smaller, which means that the BER is smaller.

In this section, our goal is to maximize the maximum energy ratio γ by jointly optimizing the active beamforming vector at the reader and the passive phase shift of the RIS, subjected to the RIS phase constraint, as well as the energy limitation of the beamforming vector. Mathematically, this problem can be formulated as follows:

$$\mathbf{P1}: \max_{f \ \theta} \gamma \tag{15a}$$

$$s.t. |\boldsymbol{\Phi}_i| = 1, \quad \forall i = 1, 2, \cdots, N \tag{15b}$$

$$\|\boldsymbol{f}\|^2 = 1 \tag{15c}$$

Since the objective function and constraints are non-convex, **P1** is a non-convex optimization problem. Generally, the alternating optimization (AO) method is used to solve the fixed weighting coefficient problem, but there is relatively high complexity. Furthermore, for the problem **P1**, it is difficult for AO to handle the weighting coefficient's dynamic change. In this article, a DRL framework is proposed to solve the problem **P1** effectively.

4. Proposed Solution

In this section, for the problem above, the proposed DRL-based optimization problem solution is detailed. First, the progress of DRL is summarized, and then the proposed TD3 algorithm is given.

4.1. Overview of Deep Reinforcement Learning

Deep reinforcement learning is the combination between reinforcement learning and deep neural network [19]. A conventional DRL system consists of agents, environment, and the mutual transfer of information between them constituting a Markov process. The current state s_t of the environment is obtained through the agent, then the action a_t is taken based on the policy Π and it is transmitted to the environment; a_t is evaluated by the environment to send reward r_t, and the new state is updated to s_{t+1}. With the goal of maximizing long-term returns, deep neural networks and the accumulated empirical data are combined in the agent to continuously optimize the policies. So far, three value-based algorithms, including DQN, DDPG, and TD3, have emerged successively.

(1) DQN: Deep Q network (DQN) aims to solve the optimization problem of the discrete action space [20]. A Q function is constructed with a deep neural network to evaluate the action-state pair, which is expressed as $Q(s, a; \boldsymbol{\theta})$ and is updated by the maximum value of the Q function as Equation (16). However, Equation (16) cannot be used to solve the continuous action space problem well.

$$Q(S, A) \leftarrow Q(S, A) + \alpha \Big[R + \gamma \max_{a} Q\big(S', a\big) - Q(S, A) \Big] \tag{16}$$

(2) DDPG: To solve the untractable problem that DQN cannot deal with the continuous action space, deep deterministic policy gradient (DDPG) is proposed [21]. There are two important components in DDPG: an Actor network and a Critic network. In the Actor network, a deep neural network is used to generate a deterministic policy $\mu\big(s; \boldsymbol{\theta}_\mu\big)$ and then select the optimal action from the continuous action space, which can obtain the maximum value of Q. In the Critic network, $Q\big(s, a; \boldsymbol{\theta}_q\big)$ is used to evaluate the action's effect as the same as DQN. To solve the problem that the updated target continues to change when the network is updated, which makes the updating difficult, fixed network technology is used to generate the copies of the Actor network and the Critic network as the target network, and to generate the target value. The target network is softly updated by the following formula:

$$\theta_{i'} = \tau\theta_i + (1 - \tau)\theta_{i'}, \; i = \mu \text{ or } q \tag{17}$$

Since a certain policy is generated by the Actor network, fully exploring the environment, the noise $N_a \sim \mathcal{N}(a_t, 1)$ is added to μ to overlap a random sample of normal distribution with a_t as the mean. Here, the new policy $\widetilde{\mu}$ can be expressed as

$$\widetilde{\mu}(s_t) = \mu\big(s_t; \theta_\mu\big) + \mathcal{N} \tag{18}$$

(3) TD3: The Twin Delayed Deep Deterministic (TD3) policy gradient algorithm is an optimized version of DDPG [22]. DDPG is mainly improved by TD3 through three aspects: double Q network, delayed actor update, and the target policy smoothing regularization. **Target policy smoothing regularization:** When the target policy network outputs actions, noise following standard normal distribution is added, so that the estimation of Q is more accurate and the network is more robust. The output action is denoted as $\widetilde{a}$, which is expressed by

$$\widetilde{a} = \Pi_{\varphi'}(s') + \varsigma \tag{19}$$

where $\varsigma \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c)$ represents the normal noise with its upper and lower limits $\pm c$. **Double Q network:** Two sets of networks are designed in TD3 to estimate

Q, and the smaller one is regarded as the update target. The update target Q network is denoted as y_t and expressed as follows:

$$y_t \leftarrow r_t + \gamma \min_{i=1,2} Q_{\theta_i}(s'_{t+1}, \widetilde{a}) \tag{20}$$

where s'_{t+1} denotes the next state in minibatch, $\widetilde{a}_t$ is the action after smoothing regularization, γ is the discount factor, r_t is the reward what we designate at the current state. **Delayed actor update:** Delay here means the time delay of the actor network updating, that is, the actor is updated after the critic is updated several times. This is because the actor network's goal is to achieve the maximum Q value, and the critic network generating the Q value is constantly updated. Therefore, it is difficult for the actor networks to find the optimal Q value, or even get stuck in the sub-optimal value. By introducing a delay, the actor network is updated when the Q value of the critic network is relatively determined. The actor network is updated by the following equation:

$$\theta_i \leftarrow -\min_{i=1,2} \sum (y_t - Q_{\theta_i}(s'_t, a'))^2 \tag{21}$$

where (s_n, a_n) is the state-action pair from replay buffer and $n = 1, 2, \cdots D$, D is the size of the replay buffer.

The target critic network and the actor network are soft-updated by the following equations, respectively:

$$\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i \tag{22}$$

$$\varphi'_i \leftarrow \tau\varphi_i + (1-\tau)\varphi'_i \tag{23}$$

4.2. TD3-Based Framework

Analyzing the above algorithm, DQN is not suitable to solve the problem **P1**, because DQN can only handle the discrete action spaces. It is worthy that both the active beamforming and the passive RIS phase shift are continuously variable. While DDPG is easily influenced by the overestimation of the Q value, the optimal performance cannot be achieved, and DDPG is very sensitive to the external conditions. In this paper, an optimization algorithm based on TD3 is proposed to solve the non-convex optimization in problem **P1**, and the block diagram of TD3 is shown in Figure 2. We used six fully connected deep neural networks (DNNs) to build the algorithm, containing one Actor network, one Actor target network, two Critic networks, and two Critic target networks. Each DNN network contains one input layer, two hidden layers and one output layer.

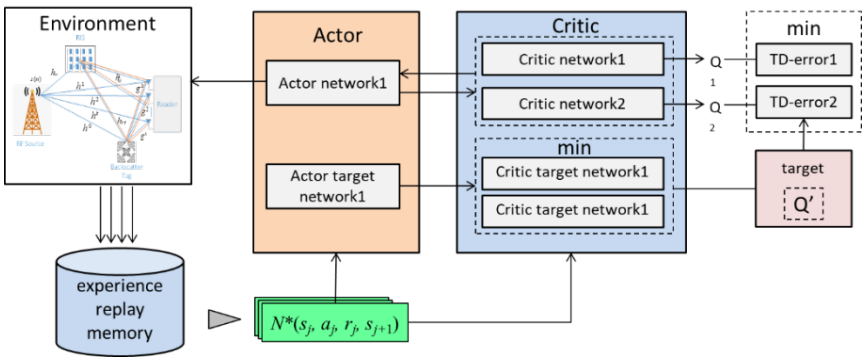


Figure 2. The TD3-based framework.

(1) Construction of TD3: In the paper, the AmBC system and the RIS controller are regarded as the environment and the agent, respectively. In order to perform the TD3 algorithm, the state, action, and reward are defined as follows: **State space:** The state space represents the current state of the system, which consists of the optimization

parameters of the previous time. Since the optimization parameter $\boldsymbol{f}$ denotes complex number value, and the neural network cannot handle the complex number, $\boldsymbol{f}$ is decomposed into two parts: the real part $\boldsymbol{f}'$ and the imaginary part $\boldsymbol{f}^*$, and they are propagated to the neural network at the same time.

$$S_t = \left[{f'}_1^{(t-1)}, \ldots, {f'}_K^{(t-1)}, f_1^{*(t-1)}, \ldots, f_K^{*(t-1)} \right] \tag{24}$$

Action space: There are two key components in the action space: the combining vector $\boldsymbol{f}$ and the phase shift $\boldsymbol{\phi}$. Similarly, each complex variable is decomposed into the real and the imaginary part.

$$A_t = \left[{\phi'}_1^t, \ldots, {\phi'}_N^t, \phi_1^{*t}, \ldots, \phi_N^{*t}, {f'}_1^t, \ldots, {f'}_K^t, f_1^{*t}, \ldots, f_K^{*t} \right] \tag{25}$$

Reward: In this article, the optimization goal is to achieve the maximum energy ratio γ. The reward at step t is a linear transformation of maximum energy ratio γ, i.e.,

$$r^t = 100(\gamma - 1) \tag{26}$$

(2) Algorithm Description: The detail of the TD3 based framework is shown as follows (Algorithm 1):

Algorithm 1 The TD3-based detector for AmBC

Input: The batch size, the soft update coefficient, the replay capacity, the policy
network learning rate, the Q network learning rate, the policy target update interval,
the explore steps, the action noise for exploration and evaluation
Output: The optimal phase shift matrix $\boldsymbol{\phi}_{opt}$, the combining vector $\boldsymbol{f}_{opt}$, the optimal
energy ratio γ, and the BER for backscatter tag detection
Initialization: Randomly initialize the six networks
for episode $j = 1, \ldots, M$ **do**
    Obtain the current CSI
    Randomly chose combining vector to obtain s_0
    for step $t = 1, \ldots, T$ **do**
    if $t >$ exp *lore steps*, choose action based on the policy $s_t = \mu(s_t; \theta_\mu) + \mathcal{N}$
else generate action randomly
    Obtain the reward r_t and observe new state s_{t+1} by the environment, store
the tuple $\{s_t, a_t, r_t, s_{t+1}\}$ into the replay memory
    if the length of replay buffer > batchsize, sample a minibatch from the
: experience replay memory
    Learn and update the Critics network by minimizing the loss
    if reach the policy target update interval, update the actor policy network
    Update the target network parameters of Actor and Critics network
    end for
end for
    Obtain the optimal phase shift matrix $\boldsymbol{\phi}_{opt}$ and the combining vector $\boldsymbol{f}_{opt}$ by final action
A_t. Obtain the AmBC BER according to Equation (14) by the final r^t

It is worth noting that in practical systems, due to the limitations of the physical properties of the device, continuous phase shifts are approximated by multi-bit quantization of discrete phase shifts at the present [23–25]. For this purpose, the method proposed in this paper is still valid and only requires quantization of the optimal phase shift output to the nearest discrete phase shift. Of course, we can also replace the TD3 agent that solves continuous non-convex optimization with a DQN agent for discrete states and actions within the framework proposed in this paper.

5. Numerical Results and Analysis

In this section, extensive quantitative and qualitative experiments, analysis, and discussion are performed to demonstrate the effectiveness of the proposed TD3 signal detection method for RIS-based multi-antenna AmBC.

5.1. Experimental Implementations

The related critical simulation parameters are set as follows: (1) all the RIS channels undergo Rician fading with Rician factor 3; (2) all other channels experience Rayleigh fading; (3) the path loss exponent is 2.5 for all the channels; (4) the ambient signal frequency is 2.4 GHz with transmit power 20 dBm; (5) the initial phase shift of the RIS and the initial received combined vector are set randomly and then are optimized gradually with TD3; (6) the critic and actor networks are fully connected deep neural networks (DNNs), which have one input layer, one output layer and two hidden layers; (7) the size of the actor and critic nets are ($2K$, $2N + 2K$, $4N + 4K$, $2N + 2K$) and ($2N + 4K$, $4N + 8K$, $2N + 4K$, 1), respectively. The DNN parameters are shown in Table 1.

Table 1. Simulation parameters.

Parameter	*Description*	*Value*
μ_c	learing rate for q_net	0.00003
μ_a	learing rate for policy_net	0.0003
Int_{update}	delayed steps for updating the policy network and target networks	3.0
$n_{explore}$	range of action noise for exploration	1.0
n_{eval}	range of action noise for evaluation of action value	0.5
D	buffer size for experience replay	500,000
I	the number of episodes	5000
T	the number of steps in each episode	20,000
W	the number of experiences in the mini-batch	64
γ	discounted rate for future reward	0.99
k	Rician factor	3
d_h	the distance from the BS to the IR	202 m
d_g	the distance from the RIS to the IR	30 m
d_H	the distance from the BS to the RIS	200 m

The simulated performance is achieved by performing the average of 5000 independent channels. All the training is executed by Windows 10 and CPU with 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80 GHz.

5.2. The Effect of the Number of Reflectors

In this section, the proposed TD3-based RIS-assisted multi-antenna AmBC signal detection method (denoted as RIS-TD3) is simulated to evaluate the effect of the receiving antennas number and the RIS reflective elements on the maximum energy ratio. The comparison benchmark algorithms include: (1) the RIS-assisted DDPG method, denoted as RIS-DDPG; (2) the RIS-assisted conventional method to optimize the phase shifts at the RIS, where the successive convex appropriation (SCA) method is used [26] denoted as SCA; (3) the semi-define relaxation (SDR) method that is compared [27] denoted as SDR; (4) the random RIS phase shift and receiver combiner coefficients, denoted as RIS-RND. The ambient RF source's transmit power and the ambient noise variance are set as 20 and -95 dBm, respectively. The simulation results are illustrated in Figure 3.

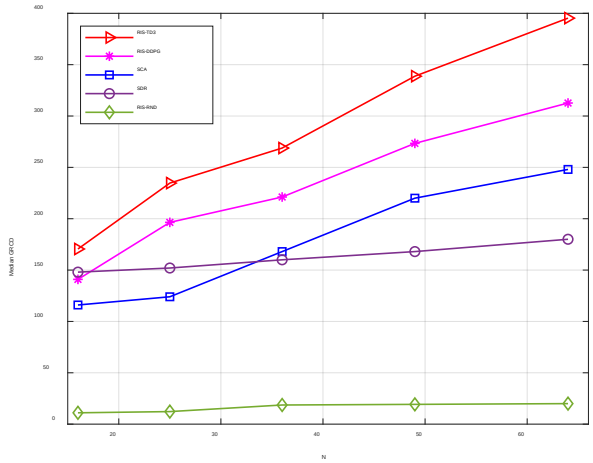


Figure 3. Median GRCD versus the number of antenna N at the reader with $K = 4$, Hidden_dim = 128.

Figure 3 shows the median GRCD versus the number of antennas at the reader with different algorithms. It is observed from Figure 3 that the proposed RIS-TD3 algorithm significantly outperforms the DDPG and the conventional methods. Particularly, with the increased antenna number, the three algorithms' system performance can be enhanced, while the benefit of the proposed TD3 algorithm is the most important. This is because both the phase shift of RIS and the beamforming at the receiver are optimized by TD3, hence achieving the overall optimization, which obtains better median GRCD.

5.3. The Effect of Communication Distance on System Performance

This subsection presents a simulation of the convergence performance of the RIS-TD3 algorithm. The simulation conditions are as follows: the number of RIS reflection units is set to 49, the hidden layer dimensions of the Actor and Critic networks are set to 256, the number of exploration episode is 50, the agent-begins-to-learn episode is 100, and the total training episode is set to 200. We observe the cumulative reward of the episode obtained by the algorithm in each training episode, denoted as all_episode_reward, and apply a sliding average to it. Moreover, it is compared with the random RIS phase shift method, denoted as all_episode_reward_rnd. The simulation results are shown in Figure 4.

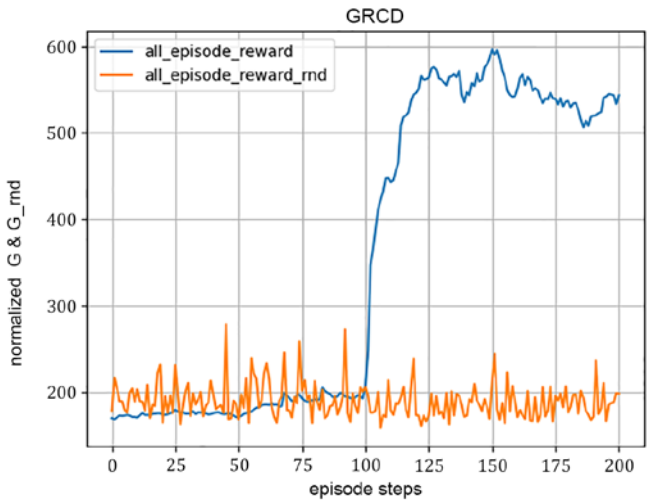


Figure 4. Convergence performance of the proposed algorithms with $N = 49$, Hidden_dim = 256.

From Figure 4, it can be seen that the algorithm is in the random exploration and data accumulation phase before it starts learning (the first 100 rounds), and the accumulated episode reward obtained at this time is small and comparable to the average of the accu-

mulated episode reward of the random RIS phase shift, which is within 200; when the algorithm starts to learn, the episode gain is able to increase rapidly, rising to a higher level of about 560 within about 25 episodes, and after that until the end of learning phase, the accumulated episode reward fluctuates in a smaller range, with the total gain above 500, which is much higher than that of the random RIS phase shifts. This demonstrates that the algorithm is able to be converged quickly and the rewards after convergence are greater, resulting in better convergence performance.

5.4. Computation Complexity Analysis

This subsection provides a time-consuming simulation for the complexity of the proposed algorithm. The benchmark algorithm used the RIS-DDPG algorithm, the SCA algorithm, and the SDR algorithm in the same way as in Section 5.2. The simulation condition is $N = 36$, the number of iterations for the SCA algorithm is 100, the iteration termination condition for the SDR algorithm is an error of less than 0.1, and the other simulation conditions are the same as in Section 5.2. Since both RIS-TD3 and RIS-DDPG can be learned offline and run online, we take the simulation time spent per episode in the testing phase after training as the algorithm consumption time. We recorded the time consumed by each algorithm for 10 simulations and averaged them to obtain the algorithm time-consuming simulation results shown in Table 2.

Table 2. Time consumed of various algorithms at N = 36 (unit: second).

Algorithms	RIS-TD3	RIS-DDPG	SCA	SDR
1	0.4188	0.3231	0.6420	10.4801
2	0.3837	0.3349	0.5787	7.0571
3	0.3684	0.3371	0.4785	7.5617
4	0.3647	0.3241	0.5332	9.7516
5	0.3868	0.4239	0.4490	5.9844
6	0.382	0.3531	0.4675	7.6004
7	0.369	0.3291	0.4493	6.6297
8	0.3786	0.3142	0.4527	6.0317
9	0.3526	0.3331	0.4411	7.2425
10	0.361	0.3577	0.5159	7.2832
Average	0.3766	0.3430	0.5008	7.5622

As can be seen from Table 2, the simulation time consumed by the RIS-TD3 algorithm proposed in the paper for one episode (200 steps) is comparable to that of the RIS-DDPG algorithm, with an increase of only 0.03 s. This is due to the fact that the proposed algorithm uses six DNNs while the RIS-DDPG algorithm only uses four DNNs. However, combining the results of 5.2, the RIS-TD3 algorithm obtains a median GRCD gain of 50 at $N = 36$ compared with that of RIS-DDPG, so the algorithm achieves a significant performance gain with a small time cost. Furthermore, the time consummation of the traditional SCA algorithm is 1.3 times longer than that of the proposed algorithm for only 100 iterations, but the average GRCD obtained is lower than that of the proposed RIS-TD3 algorithm. Finally, the traditional SDR algorithm needs the longest simulation time and has the lowest median GRCD performance. This indicates that the complexity of the proposed algorithm is significantly lower than that of the conventional algorithm.

5.5. Convergence Analysis

In this section, the hyperparameters' effect on the RIS-TD3 intelligent signal detection algorithm's performance is investigated. In the RIS-TD3, the main hyperparameters are as follows: the hidden layer size, exploration step size, discount rate, and learning rate. The

effect of these hyperparameters on the RIS-TD3 algorithm's performance is simulated. The other parameters and their corresponding values are as follows: the number of reflection units at RIS is 36, the length of batch learning data is 512, the discount rate is 0.99, the learning rate is 15×10^{-5}, the hidden layer is 128, and the exploration step is 5000. The exploration noise gradually decreases when the empirical data are greater than two times the exploration step, with a decrease factor of 0.9999 until it decreases to 0.1.

(1) The hidden layer size:

As shown in Figure 5, as the hidden layer gradually increases from 32 to 64, the maximum cumulative reward value of rounds of the RIS-TD3 algorithm keeps increasing, and the average energy ratio of normalized rounds increases to over 1000. However, when the hidden layer is further increased to 128 or 196, RIS-TD3's performance gradually decreases. This is because as the hidden layer increases, the DNN network is larger, and its learning and training complexity increase significantly, making it difficult to perform the training in the limited number of rounds. This suggests that the choice of hidden layers is a compromise between the system performance and algorithm complexity.

(2) The exploration step:

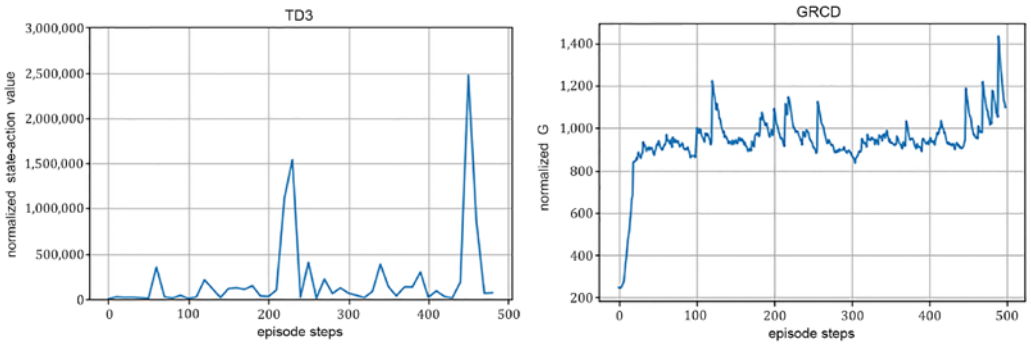
As illustrated in Figure 6, when the exploration step length is 4000, RIS-TD3 owes its first peak at around 80 rounds. As the iterations increase, RIS-TD3 owes its maximum peak at around 270 rounds, with a maximum peak of 131,703.35; when the exploration step length is 5000, RIS-TD3 has its maximum peak at around 90 rounds, with a maximum peak of 1,270,947.24. The optimal value can be achieved by RIS-TD3 quickly; when the exploration step length increases to 10,000, RIS-TD3 shows the first peak at 200 rounds and the maximum peak at 260 rounds, with a maximum peak of 463,016.65. According to the simulation results, it shows that as the exploration step length increases and more rounds are used, the later the DNN network starts learning, and the later the peak appears; however, with more exploration data as the initial sample for network learning, the faster the learning, and the higher the maximum peak can be achieved. When the exploration step is too large, the proposed RIS-TD3 algorithm cannot find the maximum value in the limited number of the iteration rounds, and the efficiency of the algorithm decreases.

(3) The discount rate:

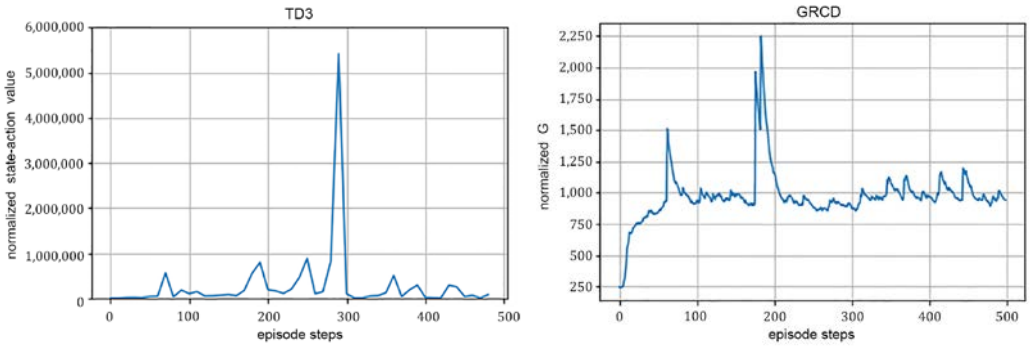
Figure 7 illustrates that when the discount rate is 0.95, RIS-TD3 owes the maximum value of 265,010.03 at round 140; when the discount rate increases to 0.99, the algorithm has the maximum value at round 90, and the maximum value reaches 1,270,947.24. This shows that the acquisition of the optimal value can be influenced significantly by the discount rate. The larger the discount rate is, the larger the reward of the algorithm can be achieved.

(4) The learning rate:

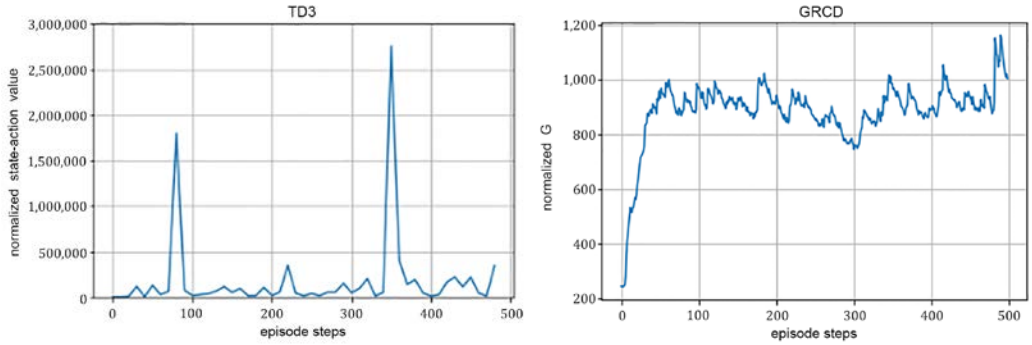
As shown in Figure 8, when the learning rate is 3×10^{-5}, RIS-TD3 peaks at round 140 with a peak value of 3174.04; when the learning rate is 15×10^{-5}, the algorithm peaks at round 90 with a peak pick higher than that of the former. This indicates that the learning rate has significant influence on the optimal value, and when the learning rate is small, the proposed algorithm obtains a smaller optimal value.



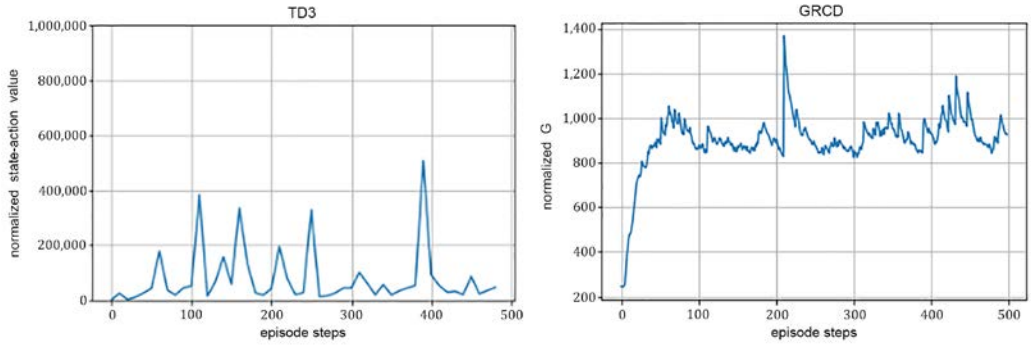
(**a**) HIDDEN_DIM = 32



(**b**) HIDDEN_DIM = 64

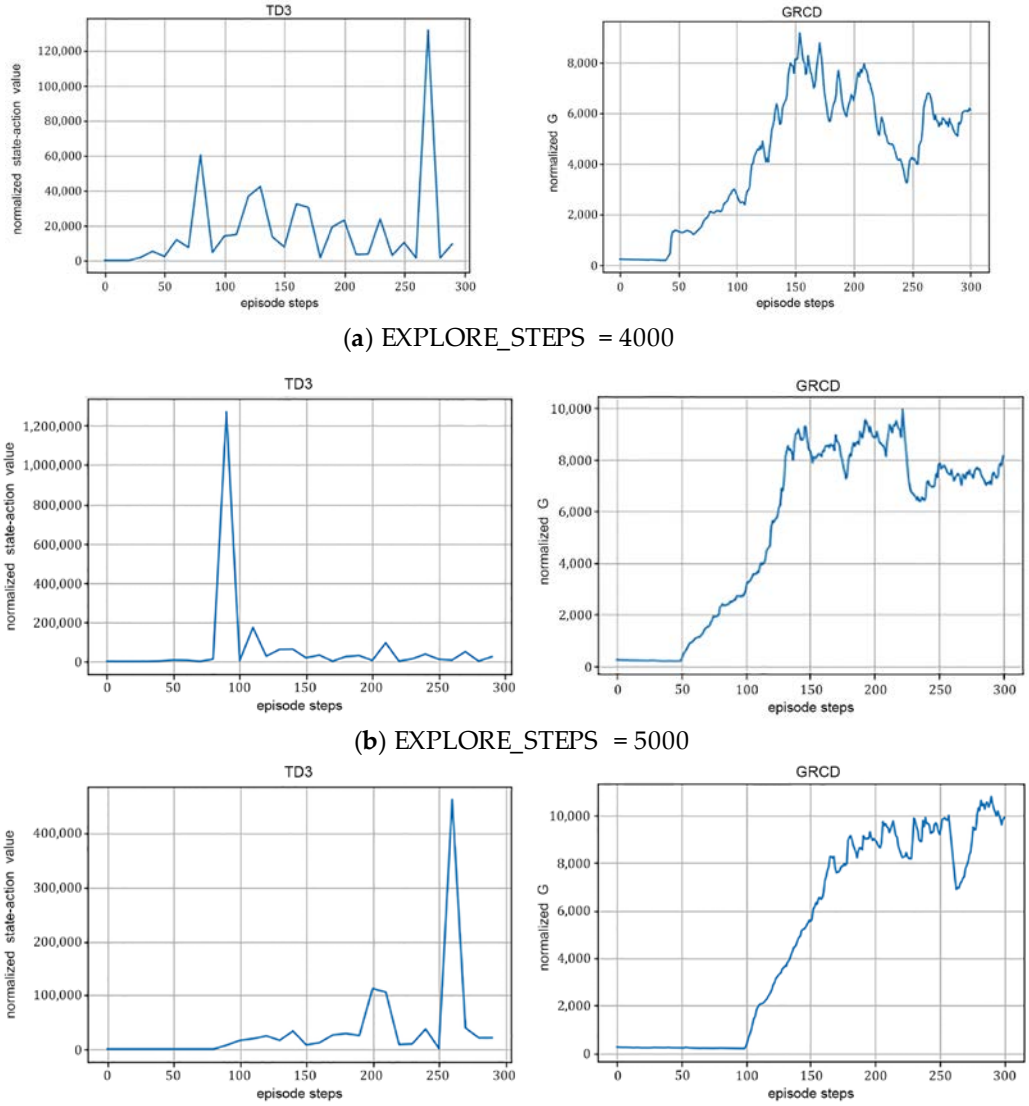


(**c**) HIDDEN_DIM = 128



(**d**) HIDDEN_DIM = 196

Figure 5. Effect of different hidden layer sizes on the algorithm performance with $N = 36$, $\gamma = 0.99$.

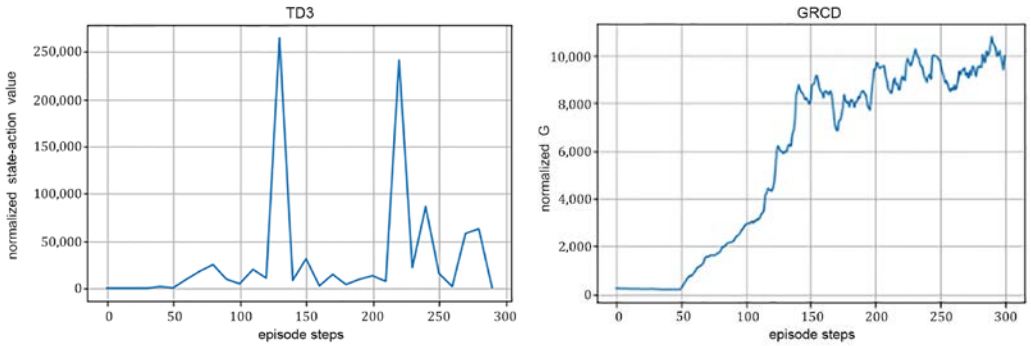


(**a**) EXPLORE_STEPS = 4000

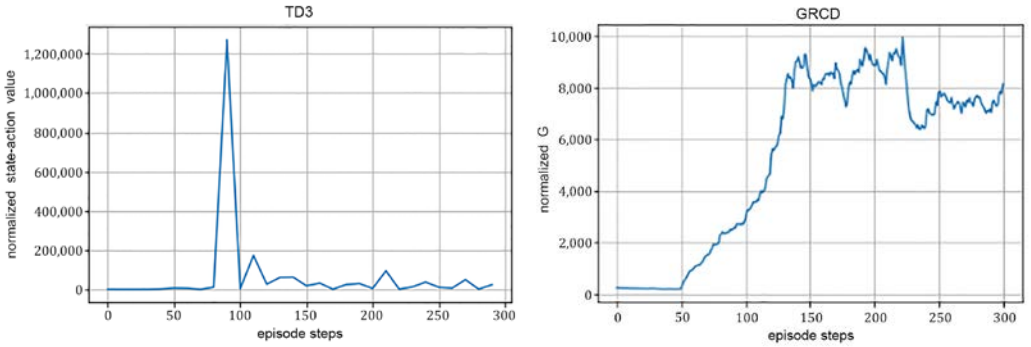
(**b**) EXPLORE_STEPS = 5000

(**c**) EXPLORE_STEPS = 10,000

Figure 6. Effect of different exploration step lengths on the performance of the algorithm with N = 36, γ = 0.99 and Hidden_dim = 128.

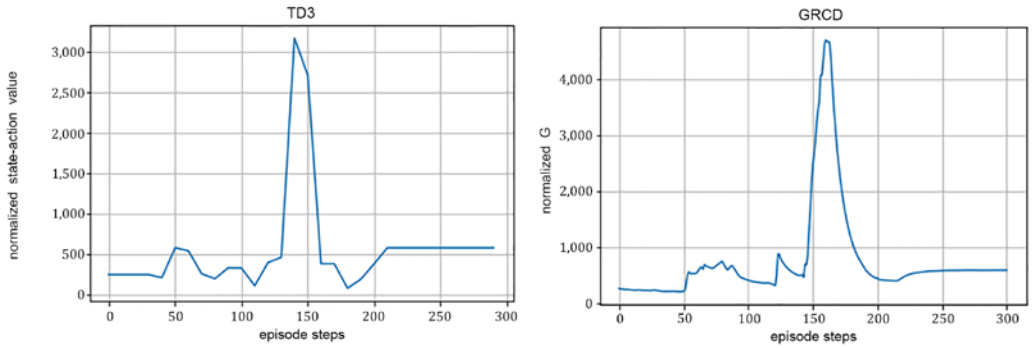


(**a**) GAMMA = 0.95

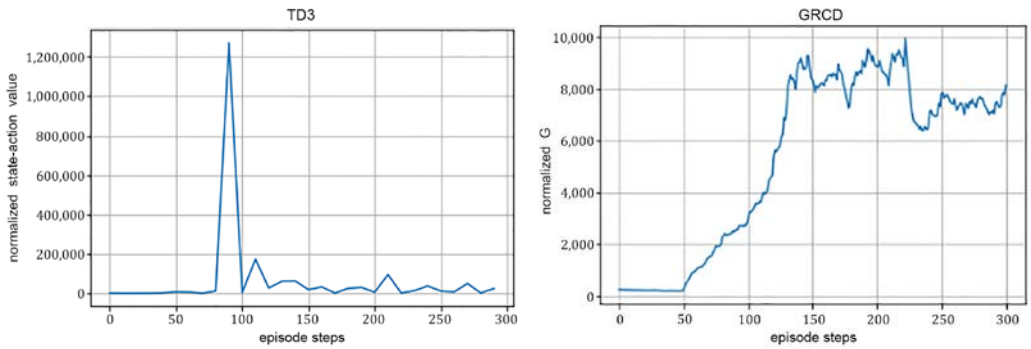


(**b**) GAMMA = 0.99

Figure 7. Effect of different discount rates on algorithm performance with N = 36, Explore_steps = 5000 and Hidden_dim = 128.



(**a**) Q_LR = 3×10^{-5}



(**b**) Q_LR = 15×10^{-5}

Figure 8. Effect of different learning rates on algorithm performance; N = 36, Explore_steps = 5000, $\gamma = 0.99$, and Hidden_dim = 128.

6. Conclusions

In this paper, an RIS-assisted multi-antenna AmBC system model is investigated, and an efficient RIS-assisted AmBC signal detection method based on deep reinforcement learning, named RIS-TD3, is developed. Particularly, an RIS-assisted multi-antenna AmBC signal model is presented, which enables the information transmission and energy collection simultaneously. Furthermore, a Twin Delayed Deep Deterministic (TD3) signal detection method is built for the AmBC system. Compared with several state-of-the-art comparison methods, the presented RIS-TD3 method achieves better performance. Because of the powerful ability of deep reinforcement learning, more essential recessive features are well explored.

There are some important advantages about RIS-TD3: (1) RIS-TD3 has good stability by delay updating and smooth regularization of the target policy network, so as to obtain the optimal solution of the nonconvex optimization problem, lower bit error rate is obtained. (2) The effect of a hyperparameter on the performance of the algorithm is explored, and the hyperparameter is chosen appropriately to further improve detection performance. (3) RIS-TD3 does not need to estimate channel information in advance. (4) RIS-TD3 has lower bit error rate and more stable convergence performance than the reference method under the condition of lack of channel information. (5) The problems of incomplete parameter acquisition and serious direct link interference in traditional AmBC signal detection methods are avoided.

To further improve the signal detection performance, we will consider introducing hierarchical deep multi-task joint learning to learn more discriminative features efficiently. For model speed acceleration, a series of ingenious model compression technologies will be tried to combine into the deep reinforcement learning framework. Moreover, the effects of delicate adaptive self-training and practical unsupervised training schemes will be explored.

Author Contributions: Formal analysis, M.G.; Project administration, H.Z.; Software, B.X.; Supervision, K.C.; Writing—original draft, F.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key R&D Program of China under Grant 2018YFE0100500, in part by the National Natural Science Foundation of China under Grant 61871387, 62106277, and 61871471, Scientific Research Plan of National University of Defense Technology (ZK21-39, and ZK21-44), Information and Communication Business Cost Project of JSCMC (ZLKT-XT-2147), and the Natural Science Basic Research Program of Shaanxi under Grant 2019JM-019.

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

References

1. Latva-aho, M.; Leppänen, K.; Clazzer, F.; Munari, A. Key drivers and research challenges for 6G ubiquitous wireless intelligence. In *6G Flag-Ship*; University of Oulu: Oulu, Finland, 2019.
2. Ye, Y.; Shi, L.; Chu, X.; Lu, G. On the Outage Performance of Ambient Backscatter Communications. *IEEE Internet Things J.* **2020**, *7*, 7265–7278. [CrossRef]
3. Jing, F.; Zhang, H.; Gao, M.; Xue, B. Bayesian-MLE signal detection for multi-antenna ambient ackscatter communication. *IET Commun.* **2022**, *16*, 672–684. [CrossRef]
4. Jia, X.; Zhou, X. IRS-Assisted Ambient Backscatter Communications Utilizing Deep Reinforcement Learning. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 2374–2378. [CrossRef]
5. Al-Nahari, A.; Jantti, R.; Duan, R.; Mishra, D.; Yigitler, H. Multi-Bounce Effect in Multi-Tag Monostatic Backscatter Communications. *IEEE Wirel. Commun. Lett.* **2021**, *11*, 43–47. [CrossRef]
6. Tao, Y.; Li, B.; Zhao, C.; Liang, Y.C. Hardware-efficient signal detection for ambient backscattering communications. *IEEE Commun. Lett.* **2019**, *23*, 2196–2199. [CrossRef]
7. Van Huynh, N.; Nguyen, D.N.; Hoang, D.T.; Vu, T.X.; Dutkiewicz, E.; Chatzinotas, S. Defeating super-reactive jammers with deception strategy: Modeling, signal detection, and performance analysis. *IEEE Trans. Wirel. Commun.* **2022**, 1. [CrossRef]
8. Toro, U.S.; ElHalawany, B.M.; Wong, A.B.; Wang, L.; Wu, K. Machine-learning-assisted signal detection in ambient backscatter com-munication networks. *IEEE Netw.* **2021**, *35*, 120–125. [CrossRef]

9. Xue, B.; He, Y.; Jing, F.; Ren, Y.; Jiao, L.; Huang, Y. Robot target recognition using deep federated learning. *Int. J. Intell. Syst.* **2021**, *36*, 7754–7769. [CrossRef]
10. Nguyen, T.T.; Nguyen, N.D.; Nahavandi, S. Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications. *IEEE Trans. Cybern.* **2020**, *50*, 3826–3839.
11. Xue, B.; Yi, W.; Jing, F.; Wu, S. Complex ISAR target recognition using deep adaptive learning. *Eng. Appl. Artif. Intell.* **2020**, *97*, 104025. [CrossRef]
12. Hu, Y.; Wang, P.; Lin, Z.; Ding, M.; Liang, Y.C. Machine learning based signal detection for ambient backscatter communications. In Proceedings of the 2019 IEEE International Conference on Communications (ICC), Shanghai, China, 21–23 May 2019.
13. Wang, X.; Duan, R.; Yigitler, H.; Menta, E.; Jantti, R. Machine Learning-Assisted Detection for BPSK-Modulated Ambient Backscatter Communication Systems. In Proceedings of the 2019 IEEE Global Communications Conference, Big Island, HI, USA, 9–13 December 2019; pp. 1–6.
14. Zhang, Q.; Guo, H.; Liang, Y.-C.; Yuan, X. Clustering-Inspired Signal Detection for Ambient Backscatter Communication Systems. In Proceedings of the 2018 IEEE Global Communications Conference, Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–7.
15. Zhang, Q.; Liang, Y.-C. Signal Detection for Ambient Backscatter Communications Using Unsupervised Learning. In Proceedings of the 2017 IEEE Globecom Workshops, Singapore, 4–8 December 2017; pp. 1–6.
16. Liu, C.; Wei, Z.; Ng, D.W.K.; Yuan, J.; Liang, Y.-C. Deep Transfer Learning for Signal Detection in Ambient Backscatter Communications. *IEEE Trans. Wirel. Commun.* **2020**, *20*, 1624–1638.
17. Chen, C.; Wang, G.; Guan, H.; Liang, Y.-C.; Tellambura, C. Transceiver Design and Signal Detection in Backscatter Communication Systems with Multiple-Antenna Tags. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 3273–3288.
18. Guo, H.; Zhang, Q.; Xiao, S.; Liang, Y.C. Exploiting multiple antennas for cognitive ambient backscatter communication. *IEEE Internet Things J.* **2019**, *6*, 765–775.
19. Wei, Q.; Ma, H.; Chen, C.; Dong, D. Deep Reinforcement Learning with Quantum-Inspired Experience Replay. *IEEE Trans. Cybern.* **2021**, 1. [CrossRef]
20. Zhu, C.; Chiang, Y.-H.; Xiao, Y.; Ji, Y. FlexSensing: A QoI and Latency-Aware Task Allocation Scheme for Vehicle-Based Visual Crowdsourcing via Deep Q-Network. *IEEE Internet Things J.* **2020**, *8*, 7625–7637.
21. Zhang, Z.; Chen, J.; Chen, Z.; Li, W. Asynchronous Episodic Deep Deterministic Policy Gradient: Toward Continuous Control in Computationally Complex Environments. *IEEE Trans. Cybern.* **2019**, *51*, 604–613. [CrossRef]
22. Cao, K.; Ding, H.; Wang, B.; Lv, L.; Tian, J.; Wei, Q.; Gong, F. Enhancing Physical Layer Security for IoT with Non-Orthogonal Multiple Access Assisted Semi-Grant-Free Transmission. *IEEE Internet Things J.* **2022**, 1. [CrossRef]
23. Woo, J.H.; Wu, L.; Park, J.B.; Roh, J.H. Real-time optimal power flow using twin delayed deep deterministic policy gradient algorithm. *IEEE Access* **2020**, *8*, 213611–213618. [CrossRef]
24. Di, B.; Zhang, H.; Song, L.; Li, Y.; Han, Z.; Poor, H.V. Hybrid Beamforming for Reconfigurable Intelligent Surface based Multi-User Communications: Achievable Rates With Limited Discrete Phase Shifts. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 1809–1822. [CrossRef]
25. Sanchez, J.; Bengtsson, E.; Rusek, F.; Flordelis, J.; Zhao, K.; Tufvesson, F. Optimal, Low-Complexity Beamforming for Discrete Phase Reconfigurable Intelligent Surfaces. In Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM), Madrid, Spain, 7–11 December 2021; pp. 1–6.
26. Guo, H.; Liang, Y.; Chen, J.; Larsson, E.G. Weighted Sum-Rate Maximization for Reconfigurable Intelligent Surface Aided Wireless Networks. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 3064–3076. [CrossRef]
27. Taha, A.; Alrabeiah, M.; Alkhateeb, A. Enabling Large Intelligent Surfaces With Compressive Sensing and Deep Learning. *IEEE Access* **2021**, *9*, 44304–44321. [CrossRef]