DRL-Based RIS Phase Shift Design for OFDM Communication Systems

Peng Chen, Xiao Li, *Member, IEEE*, Michail Matthaiou, *Fellow, IEEE* and Shi Jin, *Senior Member, IEEE*

***Abstract*—This letter investigates a downlink orthogonal frequency division multiplexing (OFDM) transmission system aided by a reconfigurable intelligent surface (RIS). To reduce the system overhead and cost, we consider a 1-bit resolution and column-wise controllable RIS, and aim to design the reflection phase shifts of the elements on the RIS to improve the spectral efficiency. By leveraging a deep Q-network (DQN) framework, a deep reinforcement learning (DRL) based optimization algorithm is proposed in order to design the reflection phase shifts. Simulations illustrate that the proposed DRL-based algorithm can achieve significant performance gains in the spectral efficiency, while greatly reducing the calculation delay.**

***Index terms*—Deep reinforcement learning, discrete phase shift, OFDM, reconfigurable intelligent surface.**

I. Introduction

In recent years, it is widely accepted that reconfigurable intelligent surfaces (RISs) represent a disruptive technology to offer high spectral efficiency and coverage cost-effectively [1]. Specifically, a RIS is a metasurface made up of many reflective elements which can reflect the incident signal with dynamically adjusted amplitude and/or phase [2, 3]. Additionally, these metasurfaces show great flexibility and compatibility in practical deployments. They can be installed on the existing infrastructure easily, such as the walls of skyscrapers and other buildings.

Thanks to the above advantages, RIS-aided communication systems have attracted great research attention. For broadband systems, [4] proposed a successive convex approximation (SCA) algorithm to optimize the RIS phase shift. In [5], the SCA algorithm was also adopted to maximize the users' minimum rate. Note that most existing works have considered a continuous phase profile of RISs [6]. However, due to hardware constraints, considering a discrete phase shift for each reflective element is more reasonable in practice. Although discrete phase shifts can be obtained by directly quantizing the continuous phase shifts computed by traditional algorithms, almost all of them are of high computational complexity.

Since artificial intelligence (AI) has been developed rapidly over the past years, the application of deep learning (DL) and deep reinforcement learning (DRL) in wireless communication systems have been widely investigated [7]. In [8–11], a DRL method was utilized to handle the optimization problem of the reflection coefficients at the RIS. Moreover, [8, 9] jointly designed the BS precoding matrix and continuous phase of RIS based on a DRL algorithm. In [10], a DRL-based algorithm and discrete Fourier transform (DFT) codebook were adopted to design the discrete phases of a RIS. A simultaneous transmission and reflection RIS was considered in [11], and a DRL-based algorithm was proposed to design its continuous reflection phase shift and discrete transmission phase shift. Nevertheless, for RIS-aided broadband communication systems, the application of DRL-based algorithms to optimize the low-resolution discrete phase shifts is still open for research exploitation.

In this letter, we elaborate on the discrete phase shift optimization of each reflective element on the RIS in a downlink orthogonal frequency division multiplexing (OFDM) transmission system. In order to reduce the hardware cost, we consider a 1-bit resolution and column-wise controllable RIS. A DRL-based algorithm is proposed for the optimization of the reflection phase shifts by applying a deep Q-network (DQN) [10]. Numerical results verify that the proposed algorithm can attain comparable spectral efficiency to the optimal approach with low time consumption.

II. System model

In this letter, as illustrated in Fig. 1, a RIS-aided downlink OFDM transmission system with K users and a RIS is considered. The RIS, utilized to assist the communication, has $M = M_x \times M_y$ (M_x rows, M_y columns) reflective elements, and is column-wise controllable such that the elements on the same column share the same reflection phase shift. Moreover, each reflective element on the RIS is of 1-bit resolution, i.e., the phase shift could be either 0 or π. Due to severe pass loss, we only take into account the signal reflected once by the RIS.

For typically low-mobility users served by the RIS, we consider that all channels remain constant during each coherence block and are relatively independent between different blocks. The frequency bandwidth is divided into V subcarriers, represented as $\mathcal{V} = \{0, 1, \ldots, V-1\}$, and for user k, the subcarrier set is defined as V_k satisfying $V_1 \cap V_2 \cap \cdots \cap V_K = \emptyset$, $V_1 \cup V_2 \cup \cdots \cup V_K = \mathcal{V}$. We suppose that the channel state information (CSI) is known at the BS and RIS. Let us define the re-

Manuscript received October 6, 2022; revised December 13, 2022; accepted January 31, 2023. The work of X. Li was supported in part by the National Natural Science Foundation of China under Grants 62231009 and 61971126, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20211511, and in part by the Jiangsu Province Frontier Leading Technology Basic Research Project under Grant BK20212002. The work of S. Jin was supported in part by the National Natural Science Foundation of China under Grants 62261160576 and 61921004. The work of M. Matthaiou was supported by a research grant from the Department for the Economy Northern Ireland under the US-Ireland R&D Partnership Programme. The associate editor coordinating the review of this paper and approving it for publication was Dr. Mingzhe Chen. *(Corresponding author: Xiao Li.)*

P. Chen, X. Li, and S. Jin are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China. (e-mail: pengc@seu.edu.cn, li_xiao@seu.edu.cn, jinshi@seu.edu.cn.)

Michail Matthaiou is with the Centre for Wireless Innovation (CWI), Queen's University Belfast, Belfast BT3 9DT, U.K. (e-mail: m.matthaiou@qub.ac.uk.)

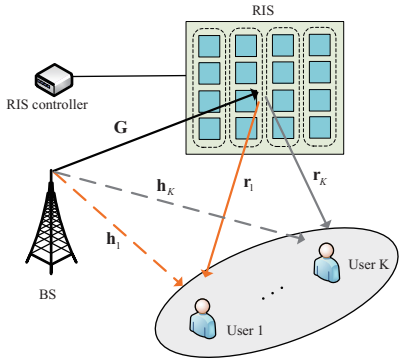

Fig. 1. The considered RIS-aided downlink OFDM transmission system.

flection vector of the RIS as $\boldsymbol{\varphi} = \left[e^{j\theta_1}, \cdots, e^{j\theta_M}\right]^T \in \mathbb{C}^{M\times 1}$, where the m-th reflective element phase shift $\theta_m = 0$ or π, $m = 1, 2, \ldots, M$, while $(\cdot)^T$ represents the transpose. Since the RIS is column-wise controllable, its reflection vector can be reformulated as

$$\boldsymbol{\varphi} = \mathbf{1}_{M_x} \otimes \overline{\boldsymbol{\varphi}}, \tag{1}$$

where $\overline{\boldsymbol{\varphi}} = \left[e^{j\bar{\theta}_1}, \cdots, e^{j\bar{\theta}_{M_y}}\right]^T \in \mathbb{C}^{M_y\times 1}$, with $e^{j\bar{\theta}_{m_y}}$ denoting the reflection coefficient for the m_y-th column, $\mathbf{1}_{M_x}$ is a M_x-dimensional all one vector and $\otimes$ indicates the Kronecker product.

Each user receives the signal transmitted by the BS through a direct channel as well as a reflection channel. For the user k, we denote $\widetilde{\mathbf{h}}_{d,k} = \left[\widetilde{h}_{d,k,0}, \widetilde{h}_{d,k,1}, \cdots, \widetilde{h}_{d,k,L_0-1}, \mathbf{0}_{1\times(V-L_0)}\right]^T \in \mathbb{C}^{V\times 1}$ as the zero-padded time domain direct channel with L_0 delay taps. Moreover, we use $\mathbf{g}_m \in \mathbb{C}^{L_1\times 1}$ to represent the L_1 tap equivalent baseband channel among the BS and the m-th RIS reflective element, and $\mathbf{r}_{k,m} \in \mathbb{C}^{L_2\times 1}$ to represent the L_2 tap equivalent baseband channel among the m-th RIS reflective element and the user k. For the user k, the time domain reflection channel through the m-th RIS reflective element is obtained as $\mathbf{g}_m * e^{j\theta_m} * \mathbf{r}_{k,m} = e^{j\theta_m}\mathbf{g}_m * \mathbf{r}_{k,m} \in \mathbb{C}^{L_3\times 1}$, the number of delayed taps is $L_3 = L_1 + L_2 - 1$, while $*$ indicates the convolution of two vectors. Let us denote the time-domain zero-padded BS-RIS-user k reflection channel as $\mathbf{Z}_k = [\mathbf{z}_{k,1}, \mathbf{z}_{k,2}, \cdots, \mathbf{z}_{k,M}] \in \mathbb{C}^{V\times M}$, where $\mathbf{z}_{k,m} = \left[(\mathbf{g}_m * \mathbf{r}_{k,m})^T, \mathbf{0}_{1\times(V-L_3)}\right]^T \in \mathbb{C}^{V\times 1}$.

At this point, for user k, the total received signal on subcarrier i can be expressed as

$$y_{k,i} = \sqrt{p_i} u_{k,i} s_i + \hat{n}_i, i \in V_k, \tag{2}$$

where p_i represents the transmission power on subcarrier i at the BS satisfying $\sum_{k=1}^{K}\sum_{i\in V_k} p_i \leqslant \hat{P}$, $\hat{P}$ is the maximal BS power, s_i denotes the transmit signal at the BS on subcarrier i satisfying $\mathbb{E}\left[|s_i|^2\right] = 1$, $\mathbb{E}[\cdot]$ is the expectation operation, $\hat{n}_i$ represents the additive Gaussion white noise (AWGN) on subcarrier i with mean 0 and variance σ^2. Moreover, $u_{k,i}$ denotes the frequency domain effective channel response of user k on subcarrier $i \in V_k$, specified as

$$u_{k,i} = \mathbf{f}_i^H \overline{\mathbf{Z}}_k \overline{\boldsymbol{\varphi}} + \mathbf{f}_i^H \widetilde{\mathbf{h}}_{d,k}, \tag{3}$$

where $\mathbf{f}_i^H$ is the i-th row of a DFT matrix $\mathbf{F}_V \in \mathbb{C}^{V\times V}$, $\overline{\mathbf{Z}}_k = \left[\overline{\mathbf{z}}_{k,1}, \overline{\mathbf{z}}_{k,2}, \cdots, \overline{\mathbf{z}}_{k,M_y}\right] \in \mathbb{C}^{V\times M_y}$ represents the zero-padded time domain reflection channel for the column-wise controllable RIS, whose m_y-th column $\overline{\mathbf{z}}_{k,m_y}$ represents the aggregated reflection channel corresponding to the m_y-th column of the reflective elements on the RIS, and is expressed as

$$\overline{\mathbf{z}}_{k,m_y} = \sum_{j=0}^{M_x-1} \mathbf{z}_{k,m_y+jM_y}, \ m_y = 1, 2, \ldots, M_y. \tag{4}$$

Then, we can express the frequency domain reflection channel and direct channel of user k on subcarrier i as

$$\mathbf{h}_{k,i}^r = \mathbf{f}_i^H \overline{\mathbf{Z}}_k, h_{k,i}^d = \mathbf{f}_i^H \widetilde{\mathbf{h}}_{d,k}. \tag{5}$$

Therefore, the spectral efficiency of the user k can be computed as

$$R_k = \frac{1}{N + N_{CP}} \sum_{i\in V_k} \log_2\left(1 + p_i \frac{\left|\mathbf{h}_{k,i}^r \overline{\boldsymbol{\varphi}} + h_{k,i}^d\right|^2}{\kappa\sigma^2}\right), \tag{6}$$

where $\kappa \geqslant 1$ is used to indicate the difference in channel capacity relative to the actual system [12], and N_{CP} is the cyclic prefix (CP) satisfying the constraint $N_{CP} \geqslant \max(L_0, L_3)$.

In this letter, we assume that the BS distributes the subcarrier power evenly, $p_i = \frac{P_t}{V}$, $i = 0, 1, \ldots, V-1$. We try to maximize the broadband transmission system sum spectral efficiency through optimizing the reflection phase shifts of the RIS elements. The problem can be formulated as

$$\begin{aligned} \text{P1}: \quad & \max_{\overline{\boldsymbol{\varphi}}} \sum_{k=1}^{K}\sum_{i\in V_k} \log_2\left(1 + p_i \frac{\left|\mathbf{h}_{k,i}^r \overline{\boldsymbol{\varphi}} + h_{k,i}^d\right|^2}{\kappa\sigma^2}\right), \\ \text{s.t.} \quad & \bar{\theta}_{m_y} = 0 \ \text{ or } \ \pi, m_y = 1, 2, \ldots, M_y. \end{aligned} \tag{7}$$

Note that the above problem is non-convex. Numerical methods, such as the branch and bound algorithm, are usually utilized to solve such discrete optimization problems, while entailing high computational delay. Next, we try to solve it through a DRL-based algorithm.

III. DRL-BASED OPTIMIZATION ALGORITHM

In this section, we propose an efficient method to design the reflection phase shifts of the RIS with low calculation delay. We define the codebook $\mathcal{P}$ as the set containing all the possible reflection vectors of the RIS. Thus, the above problem for the reflection vector $\overline{\boldsymbol{\varphi}}$ that optimizes the sum spectral efficiency can be re-formulated as

$$\begin{aligned} \text{P2}: \quad & \max_{\overline{\boldsymbol{\varphi}}} \sum_{k=1}^{K}\sum_{i\in V_k} \log_2\left(1 + p_i \frac{\left|\mathbf{h}_{k,i}^r \overline{\boldsymbol{\varphi}} + h_{k,i}^d\right|^2}{\kappa\sigma^2}\right), \\ \text{s.t.} \quad & \overline{\boldsymbol{\varphi}} \in \mathcal{P}. \end{aligned} \tag{8}$$

Problem P2 can be solved via an exhaustive search algorithm. However, this would entail a computational complexity of $O(2^{M_y})$. To reduce the calculation delay, we propose an optimization algorithm exploiting DQN to design the reflection vector.

DQN is a DRL algorithm able to solve the non-convex problem with continuous state space and discrete action space, as demonstrated in Fig. 2. The DQN agent contains a state-action Q-network, a target $\hat{Q}$-network, a replay buffer, and an optimizer. The state-action Q-network and the target $\hat{Q}$-network are implemented with a deep neural network (DNN)

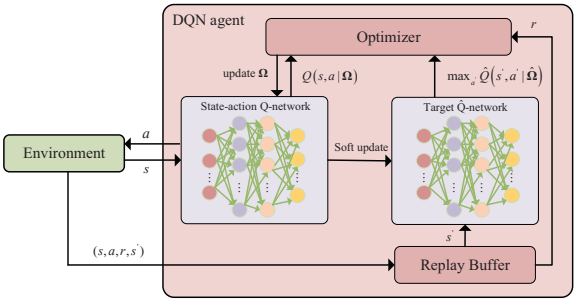

Fig. 2. DQN framework.

with weights $\mathbf{\Omega}$ and $\hat{\mathbf{\Omega}}$, respectively. Specifically, in each learning iteration t, the DQN agent acquires the state s_t by observing the current environment, and then selects an action a_t according to an ϵ-greedy policy, that either selects an action randomly with probability ϵ, or selects an action by the network, i.e.,

$$a_t = \arg\max_{\boldsymbol{a'}} Q(s_t, a'; \mathbf{\Omega}), \tag{9}$$

with probability $(1-\epsilon)$. The reward r_t will be obtained from the environment and the state will update to s_{t+1}, after the action a_t is completed. Then, the agent stores the tuple (s_t, a_t, r_t, s_{t+1}) into the replay buffer as an experience. The optimizer randomly selects $\mathcal{B}$-sized experience tuples from the replay buffer, and update the network with the selected experience tuples so that the state-action Q-network output $Q(s_t, a_t; \mathbf{\Omega})$ approaches the target Q-value:

$$Q(s_t, a_t; \mathbf{\Omega}) \leftarrow r_t + \gamma \max_{\boldsymbol{a'}} \hat{Q}(s_{t+1}, a'; \hat{\mathbf{\Omega}}), \tag{10}$$

where γ is a discount factor and $\hat{Q}(s_{t+1}, a'; \hat{\mathbf{\Omega}})$ represents the target $\hat{Q}$-network output. At certain intervals, the weights $\hat{\mathbf{\Omega}}$ of the target $\hat{Q}$-network are soft-updated according to the state-action Q-network weights. After extensive training, a robust DRL-based algorithm will be obtained that can output the optimal action which has the maximum static-action Q-network output for a given environment state.

In the proposed RIS phase shift optimization algorithm, we regard the CSI of the considered system as the environment. At current learning iteration t, the key elements of the DQN method are defined in details:

1) State: $s_t = \left\{\mathbf{h}_{k,i}^{r(t)}, h_{k,i}^{d(t)}\right\}_{i=0,1,\ldots,V-1}$ is set as the corresponding state, where $\mathbf{h}_{k,i}^{r(t)}$ and $h_{k,i}^{d(t)}$ represents the frequency domain reflection channel and direct channel in learning iteration t. In this way, the DQN method can obtain the channel state characteristics. Moreover, the agent deconstructs the input vector into real and imaginary parts, doubling the dimension of s_t to $2(M_y+1)V$.
2) Action: Set $a_t = \left\{\overline{\theta}_1^{(t)}, \overline{\theta}_2^{(t)}, \cdots, \overline{\theta}_{M_y}^{(t)}\right\} \in \mathcal{P}$ as the action, the elements of which are the reflection phase shifts for each column of RIS.
3) Reward: Note that the design objective of the proposed algorithm is to optimize the total spectral efficiency. Thus, we set $r_t = \sum_{k=1}^{K} R_k^{(t)}$ as the reward value.

Training Procedure: At the initialization time, the state-action Q-network $\mathbf{\Omega}$ and the target $\hat{Q}$-network $\hat{\mathbf{\Omega}}$ are generated so that their weights are uniformly distributed. Moreover, a reflection vector codebook $\mathcal{P}$ and a replay buffer $\mathcal{D}$ with capacity $\mathcal{C}$ are initialized. In each learning iteration t, the CSI is preprocessed as the current state s_t. For exploration besides exploitation, the agent chooses the action based on the ϵ-greedy policy. Then, a_t is reformed into a reflection vector $\overline{\boldsymbol{\varphi}} = \left[e^{j\overline{\theta}_1}, \cdots, e^{j\overline{\theta}_{M_y}}\right]^T$ to compute the current reward r_t and the current environment state is updated to s_{t+1}. Further, the agent obtains the tuple (s_t, a_t, r_t, s_{t+1}) and stores it to the replay buffer $\mathcal{D}$ as one experience. Then, the agent randomly samples $\mathcal{B}$-sized minibatches of the experience tuples, i.e., (s_j, a_j, r_j, s_{j+1}), $j = 1, \ldots, \mathcal{B}$ from $\mathcal{D}$. After that, the target Q-value that the static-action Q-network aims to approximate will be calculated by

$$y_j = \begin{cases} r_j, & j = \mathcal{B}, \\ r_j + \gamma \max\limits_{\boldsymbol{a'}} \hat{Q}(s_{j+1}, a'; \hat{\mathbf{\Omega}}), & j < \mathcal{B}, \end{cases} \tag{11}$$

where $\hat{Q}(s_{j+1}, a'; \hat{\mathbf{\Omega}})$ are the corresponding target $\hat{Q}$-network output. We define the loss function as the following mean squared error, i.e.,

$$L(\mathbf{\Omega}) = \frac{1}{B}\sum\nolimits_{j=1}^{B} \left(y_j - Q(s_j, a_j; \mathbf{\Omega})\right)^2. \tag{12}$$

The DQN agent utilizes a soft-update on the weights $\hat{\mathbf{\Omega}}$ of the target $\hat{Q}$-network. A soft-update can effectively eliminate the instability of state-action network and accelerate the convergence of the algorithm, that is given by

$$\hat{\mathbf{\Omega}} = \tau \mathbf{\Omega} + (1-\tau)\hat{\mathbf{\Omega}}, \tag{13}$$

where $\tau \ll 1$ is the soft update coefficient. Finally, the DQN agent learns to map an input (the CSI) to an output (the reflection phase shift of RIS).

Application Procedure: To further improve the performance, we add a searching procedure within a small range of potential phase shift vectors, based on the output of the trained DQN agent.[1] The agent obtains the current channel information $\left\{\mathbf{h}_{k,i}^{r(t)}, h_{k,i}^{d(t)}\right\}_{i=0,1,\ldots,V-1}$ and proprocesses it to get the state s_t. Then, the most potential L actions, i.e., $\mathcal{A} = \left\{a^{(1)}, a^{(2)}, \cdots, a^{(L)}\right\}$ can be achieved by selecting the L actions with the largest output $Q(s, a; \mathbf{\Omega})$. The optimal action a^* leading to the maximum spectral efficiency will be selected as the reflection vector. By properly choosing the value of L, the proposed approach can effectively improve the spectral efficiency of the network at the cost of a relatively low extra time consumption. Algorithm 1 exemplifies the detailed procedure of the proposed algorithm.

Implementation Consideration: Note that a RIS has usually low or no data processing capabilities, while the BS is usually capable of complex computations and can acquire the overall system information. Thus, we consider to place the DRL agent at the BS, so that the agent can obtain the channel coefficients through a channel estimation method applied at the BS, such as

[1]Unlike the improvements of RL in some existing works, such as [13], which proposed a novel distributed RL approach, and [14, 15], which employed a modulated Hebbian network and a quantum-inspired experience replay, we hereafter utilize a searching procedure to improve RL, which can output the nearly optimal action by comparing the spectral efficiencies of the most potential part of the actions.

Algorithm 1 DRL-based Optimization Algorithm

Input: replay buffer capacity $\mathcal{D}$, γ, τ, codebook of reflection coefficients $\mathcal{P}$, batch size $\mathcal{B}$.

Output: Trained state-action Q-network $Q(s,a;\mathbf{\Omega})$.

Initialization: Initialize the network $Q(s,a;\mathbf{\Omega})$ and $\hat{Q}(s,a;\hat{\mathbf{\Omega}})$ with random weights $\hat{\mathbf{\Omega}}=\mathbf{\Omega}$.

Task I: Network Training

for each episode **do**
  Initialize the outer environment s_0;
  for $t=0,1,2,...,T-1$ **do**
    Obtain the current channel information $\{\mathbf{h}_{k,i}^{r(t)}, h_{k,i}^{d(t)}\}$, $\forall i$ and preprocess it;
    With probability ϵ select an action $a_t \in \mathcal{P}$ at random;
    Otherwise select action $a_t = \arg\max_{\boldsymbol{a'}} Q(s_t, a'; \mathbf{\Omega})$;
    Acquire the reward r_t and next state s_{t+1};
    Put the experience tuple (s_t, a_t, r_t, s_{t+1}) into $\mathcal{D}$;
    if t mod $T_p = 0$ **then**
      Sample $\mathcal{B}$-sized tuples from $\mathcal{D}$ randomly;
      Calculate target Q-value based on (11);
      Update the state-action Q-network and target $\hat{Q}$-network according to (12) and (13);
    end if
    $s_t \leftarrow s_{t+1}$;
  end for
  Decrease ϵ gradually;
end for

Task II: Network Application

Obtain the current channel information $\{\mathbf{h}_{k,i}^{r}, h_{k,i}^{d}\}$, $\forall i$ from the environment and preprocess to s;
Select the most potential L actions, i.e., $\mathcal{A} = \{a^{(1)}, a^{(2)}, \cdots, a^{(L)}\}$ with the largest $Q(s,a;\mathbf{\Omega})$;
Calculate the spectral efficiencies corresponding to each action in the set $\mathcal{A}$;
Select the optimal action a^* as the RIS reflection phase shift by comparing these spectral efficiencies;

[16] and [17]. Then, the agent can calculate the reward based on its output action. The RIS controller obtains the phase shifts from the BS through either a wired or wireless link, and then adjusts the RIS elements correspondingly.

IV. Experiment Results

In this section, as described in Fig. 3, we consider a downlink SISO-OFDM system with eight users, located within a quarter circle near the reconfigurable intelligent surface, the radius of which is set to $d_3 = 10$m. The vertical and horizontal distances among the BS and the RIS are $d_1 = 10$m, $d_2 = 50$m, respectively.

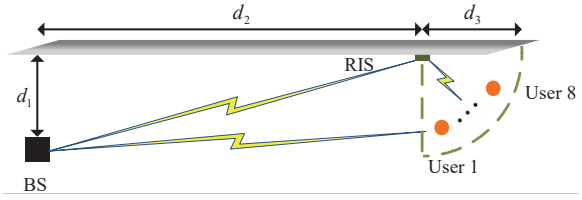

Fig. 3. Simulation setup.

The total number of subcarriers is $V = 64$ and the subcarriers are allocated to the eight users evenly, i.e., $V_k = \{8k, \ldots, 8k+7\}$, $k = 0, 1, \ldots, 7$. The distance between two adjacent reflective elements of the RIS is set to half wavelength. For the maximal delay spread, we set $L_0 = 16$, $L_1 = 13$ and $L_2 = 4$, respectively. Moreover, for the direct channels of BS-users, we consider Rayleigh fading, while the reflection channels, i.e., BS-RIS and RIS-users channels, are considered to be Rician fading. The first tap of the reflection channels represents the line-of-sight (LoS) path, while the other taps are non-line-of-sight (NLoS) paths. The Rician factors of the BS-RIS channel and the RIS-users channel are denoted as $\Im_{\mathrm{BR}}$ and $\Im_{\mathrm{RU}}$, i.e.,

$$\Im_{\mathrm{BR}} = \frac{P_{\mathrm{LoS,BR}}}{P_{\mathrm{NLoS,BR}}}, \Im_{\mathrm{RU}} = \frac{P_{\mathrm{LoS,RU}}}{P_{\mathrm{NLoS,RU}}}, \tag{14}$$

where $P_{\mathrm{LoS,BR}}$, $P_{\mathrm{NLoS,BR}}$, $P_{\mathrm{LoS,RU}}$ and $P_{\mathrm{NLoS,RU}}$ are the powers of the LoS and NLoS path of the corresponding channels. The large-scale fading is modeled as $\mathrm{PL} = \mathrm{PL}_0 - 10\varrho\log_{10}\left(\frac{d}{D_0}\right)$ dB, where $\mathrm{PL}_0 = -30$ dB, $D_0 = 1$m, d is the distance among the transmitter and receiver, and ϱ is the path loss exponent. The corresponding parameters of the OFDM system model are set to $\Im_{\mathrm{BR}} = 2$dB, $\Im_{\mathrm{RU}} = 4$dB, κ = 8.8dB, $N_{\mathrm{CP}} = 16$ and $\sigma^2 = -75$dBm, respectively. The path loss exponents of the BS-RIS channel, RIS-users channel, and BS-users channel are respectively set as $\varrho_{\mathrm{BR}} = 2.2$, $\varrho_{\mathrm{RU}} = 2.4$, and $\varrho_{\mathrm{BU}} = 3.8$. The signal-to-noise ratio (SNR) of the BS-Users direct channel is $\mathrm{SNR}_d = \frac{P_{\mathrm{d}}}{V\sigma^2}$, where $P_{\mathrm{d}} = \frac{P_t}{V}\sum_{k=1}^{K}\sum_{i\in V_k}\left|h_{k,i}^{d}\right|^2$. A single GPU of NVIDIA RTX2080 Ti is applied to train the DRL network. The static-action Q-network and target $\hat{Q}$-network both consist of four fully-connected layers of 1408, 2048, 2048, 1024 nodes, respectively. The capacity $\mathcal{C}$ of the replay buffer is 2000 and the batch size is $\mathcal{B} = 512$, while $T = 1000$ and $T_p = 200$. Moreover, the initial value of the greedy factor ϵ is 0.95 and it decreases by a factor of 1% in every episode until it reaches 0.1. The learning rate and soft update coefficient are set to $\alpha = 0.001$, $\tau = 0.005$, respectively, while $\gamma = 0$.

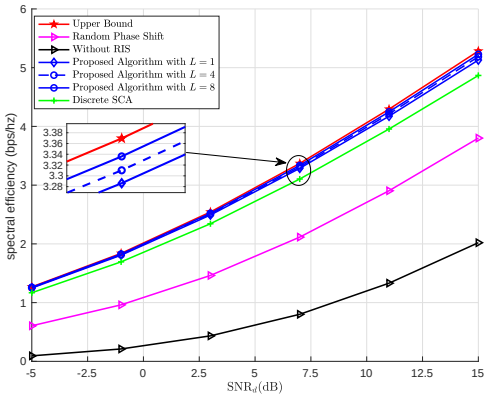

Fig. 4. Spectral efficiency vs. SNR_d.

Figure 4 demonstrates the spectral efficiency achieved by the proposed DRL-based algorithm versus SNR_d under different L. In this figure, the SNR_d varies from -5 to 15dB and the arrangement of reflective elements in the RIS is

$M_x = 10$, $M_y = 10$. For comparison, we also demonstrate the performance of the exhaustive search (as an upper bound), random phase shifts, without RIS and Discrete successive convex approximation (SCA) algorithm. The Discrete SCA algorithm obtains the phase shift by directly quantizing the continuous phase shift obtained by the SCA algorithm [4] to 0 or π, whilst the number of iterations of the SCA algorithm is 8. As shown by the results, regardless of the SNR_d, the appropriate deployment of the RIS (efficient phase shift design) brings significant gains in spectral efficiency. As illustrated, the spectral efficiency of the DRL-based algorithm increases with L. Most importantly, it is obvious that the proposed algorithm can achieve a comparable performance to the upper bound and outperforms all the other algorithms. In

TABLE I
ON-LINE RUNNING TIME AND PERFORMANCE COMPARISON

Algorithm	Time(ms)	Performance
Proposed Algorithm with $L = 1$	0.66	97.34%
Proposed Algorithm with $L = 2$	2.54	97.86%
Proposed Algorithm with $L = 4$	4.71	98.47%
Proposed Algorithm with $L = 8$	8.33	99.27%
Discrete SCA	1494.8	92.13%

Table I, we compare the on-line running time and performance of the proposed DRL-based algorithm and the Discrete SCA algorithm. The performance is represented by the ratio of the achieved spectral efficiency of these algorithms to the upper bound. The other experiment parameters are the same as in Fig. 4. From Table I, it can be seen that the running time of the Discrete SCA algorithm is several hundred, or even thousand, times that of the proposed algorithm. The spectral efficiency achieved by the proposed algorithm outperforms that of the Discrete SCA algorithm. This is because the Discrete SCA algorithm obtains the phase shift by directly quantizing the continuous phase shift obtained by the SCA algorithm to 0 or π, which might not be optimal. Moreover, the achieved spectral efficiency of the proposed algorithm can reach up to 99% of the upper bound when $L = 8$. Thus, the proposed algorithm can attain a spectral efficiency comparable to the optimal approach with low time consumption. Also, the proposed DRL-based algorithm can make a good compromise between time complexity and performance by adjusting L dynamically.

In Fig. 5, the convergence performance of the proposed DRL-based algorithm is illustrated, where $\mathrm{SNR}_d = 15$ dB and the arrangement of reflective elements in the RIS is $M_x = 10$, $M_y = 10$. In this figure, we can notice that the performance of the proposed method approaches the upper bound of the exhaustive search algorithm as the agent learns, while the convergence of DRL requires about 330 episodes.

V. CONCLUSION

In this letter, we investigated the design problem of discrete reflection vector of a RIS in a broadband communication system to optimize the spectral efficiency. We proposed an efficient DRL-based algorithm for designing the 1-bit resolution discrete phase shifts of the column-wise controllable RIS. Experimental results indicated that the proposed method can achieve significant performance gain and is close to the upper bound with low time consumption.

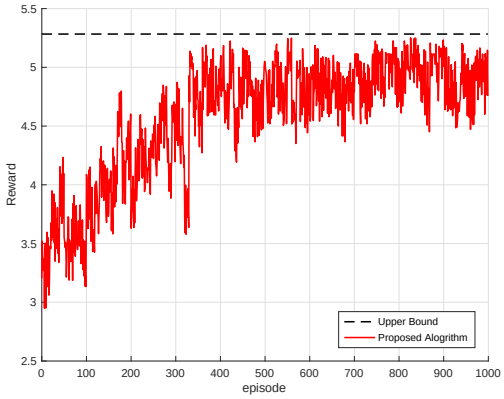

Fig. 5. Convergence performance of the proposed algorithm.

REFERENCES

[1] M. Matthaiou *et al.*, "The road to 6G: Ten physical layer challenges for communications engineers," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 64–69, Jan. 2021.

[2] W. Tang *et al.*, "MIMO transmission through reconfigurable intelligent surface: System design, analysis, and implementation," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2683–2699, Jul. 2020.

[3] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[4] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, "Intelligent reflecting surface meets OFDM: Protocol design and rate maximization," *IEEE Trans. Wireless Commun.*, vol. 68, no. 7, pp. 4522–4535, Jul. 2020.

[5] Y. Yang, S. Zhang, and R. Zhang, "IRS-enhanced OFDMA: Joint resource allocation and passive beamforming optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 760–764, Jun. 2020.

[6] C. Luo, X. Li, S. Jin, and Y. Chen, "Reconfigurable intelligent surface-assisted multi-cell MISO communication systems exploiting statistical CSI," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2313–2317, Oct. 2021.

[7] Q. Wang, K. Feng, X. Li, and S. Jin, "Precodernet: Hybrid beamforming for millimeter wave systems with deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1677–1681, Oct. 2020.

[8] K. Feng *et al.*, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 745–749, May 2020.

[9] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[10] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb, "Deep reinforcement learning for intelligent reflecting surfaces: Towards standalone operation," in *Proc. IEEE SPAWC*, May 2020.

[11] R. Zhong *et al.*, "Hybrid reinforcement learning for STAR-RISs: A coupled phase-shift model based beamformer," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2556–2569, Sept. 2022.

[12] C. You, B. Zheng, and R. Zhang, "Intelligent reflecting surface with discrete phase shifts: Channel estimation and passive beamforming," in *Proc. IEEE ICC*, Jun. 2020.

[13] S. Wang *et al.*, "Distributed reinforcement learning for age of information minimization in real-time IoT systems," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 3, pp. 501–515, Apr. 2022.

[14] P. Ladosz *et al.*, "Deep reinforcement learning with modulated hebbian plus Q-network architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 5, pp. 2045–2056, May 2022.

[15] Q. Wei, H. Ma, C. Chen, and D. Dong, "Deep reinforcement learning with quantum-inspired experience replay," *IEEE Trans. Cybernetics*, vol. 52, no. 9, pp. 9326–9338, Sept. 2022.

[16] Y. Lin, S. Jin, M. Matthaiou, and X. You, "Tensor-based algebraic channel estimation for hybrid IRS-assisted MIMO-OFDM," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3770–3784, June 2021.

[17] S. Jeong *et al.*, "Low-complexity joint CFO and channel estimation for RIS-aided OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 1, pp. 203–207, Jan. 2022.