A Deep Reinforcement Learning Approach to Two-Timescale Transmission for RIS-aided Multiuser MISO systems

Huaqian Zhang, Xiao Li, *Member, IEEE,* Ning Gao, *Member, IEEE,* Xinping Yi, *Member, IEEE,* and Shi Jin, *Senior Member, IEEE*

***Abstract*—Reconfigurable intelligent surface (RIS) has drawn great attention recently as a promising technology for future wireless networks. In this letter, considering the two-timescale transmission protocol, we investigate the joint design of the transmit beamforming at the base station (BS) with instantaneous channel state information (CSI) and the RIS phase shifts with statistical CSI. Due to the large number of RIS elements, this design issue usually suffers from high computational complexity. To resolve the non-convexity issue with low complexity, we propose a novel deep reinforcement learning (DRL) framework, which contains two agents applying proximal policy optimization (PPO) based algorithm. Experiment results demonstrate that the proposed algorithm has comparable spectral efficiency performance to the state-of-the-art methods with substantially reduced computational delay.**

***Index Terms*—Deep reinforcement learning, reconfigurable intelligent surface, two-timescale optimization, beamforming.**

I. Introduction

As the research of sixth-generation (6G) mobile communication emerges, RIS [1]–[3] has attracted great attention thanks to its green, cost-effective and plug-and-play characteristics. By dynamically adjusting its reflection elements, RIS has made a significant impact on energy efficiency [4], system throughput [5], [6], to name just a few.

However, the requirement of instantaneous CSI (I-CSI) in many previous works inevitably causes a large amount of CSI acquisition overhead. An alternative apporach is to use statistical CSI (S-CSI), which changes more slowly and is easier to obtain. In [7] and [8], only S-CSI was utilized to design BS beamforming and RIS phase shifts. However, for a complex and time-varying environment, or the line-of-sight (LOS) component is relatively weak, using S-CSI alone will cause system performance degradation. Thus, [9]–[11] further adopted the two-timescale (TTS) transmission protocol, which design the RIS phase shifts using S-CSI, while design the beamforming at the BS using I-CSI. Although effective, the computational complexity of the applied iterative optimization algorithms is relatively high. In recent years, with the development of artificial intelligence, attempts of DRL in RIS-aided transmission were carried out, such as [12] and [13] with I-CSI, leading to low computational complexity solutions.

In this letter, to improve the spectral efficiency with low complexity, we study the joint design of the BS beamforming and RIS phase shifts using DRL under the TTS transmission scheme. To deal with the non-convex issue in the joint optimization, we propose a two-agents PPO based algorithm, where one agent is responsible for determining RIS phase shifts and the other is responsible for determining BS beamforming. The simulation results show that the proposed algorithm has comparable performance to the near-optimal numerical algorithm with substantially reduced computation latency.

II. System Model

Consider a multi-user multiple-input single-output (MISO) downlink communication system in Fig. 1. It consists of a BS equipped with a uniform linear array (ULA) of M antennas and K single-antenna users. To better serve the users in a specific area, a RIS is employed, which is constructed as a uniform planar array (UPA) with N ($N = N_x \times N_y$) reflection elements. To obtain the phase shifts configuration, the RIS is linked to an intelligent controller which communicates with the BS via a separate link. The signals reflected twice or more by the RIS are omitted due to significant path loss.

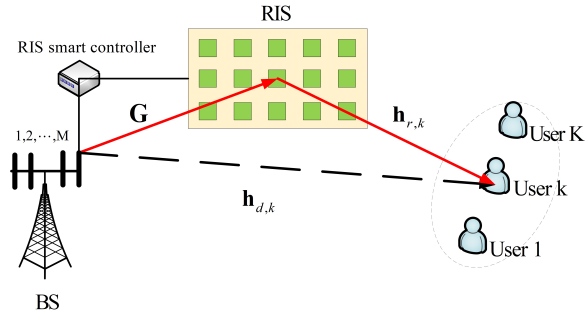

Fig. 1. RIS-assisted multi-user MISO communication system.

Manuscript received March 23, 2023; revised May 7, 2023; accepted May 15, 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 associate editor coordinating the review of this paper and approving it for publication was Dr. Somayeh Kafaie. (*Corresponding author: Xiao Li.*)

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

N. Gao is with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: ninggao@seu.edu.cn).

X. Yi is with the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3BX, U.K. (e-mail: Xinping.Yi@liverpool.ac.uk).

Assuming that all channels are quasi-static frequency flat-fading, we denote the channels from the BS to the RIS, from the RIS to the k-th user, and from the BS to the k-th user as $\mathbf{G} \in \mathbb{C}^{N\times M}$, $\mathbf{h}_{r,k} \in \mathbb{C}^{N\times 1}$, and $\mathbf{h}_{d,k} \in \mathbb{C}^{M\times 1}$, respectively. As both LOS and non-line-of-sight (NLOS) components can be present, the channels are modeled as follows

$$\mathbf{G} = \sqrt{L_G}\left(\sqrt{\frac{\beta_1}{\beta_1+1}}\bar{\mathbf{G}} + \sqrt{\frac{1}{\beta_1+1}}\tilde{\mathbf{G}}\right), \tag{1}$$

$$\mathbf{h}_{r,k} = \sqrt{L_{r,k}}\left(\sqrt{\frac{\beta_{2,k}}{\beta_{2,k}+1}}\bar{\mathbf{h}}_{r,k} + \sqrt{\frac{1}{\beta_{2,k}+1}}\tilde{\mathbf{h}}_{r,k}\right), \tag{2}$$

$$\mathbf{h}_{d,k} = \sqrt{L_{d,k}}\left(\sqrt{\frac{\beta_{3,k}}{\beta_{3,k}+1}}\bar{\mathbf{h}}_{d,k} + \sqrt{\frac{1}{\beta_{3,k}+1}}\tilde{\mathbf{h}}_{d,k}\right), \tag{3}$$

where β_1, $\beta_{2,k}$ and $\beta_{3,k}$ are the Rician factors of the corresponding channels, L_G, $L_{r,k}$ and $L_{d,k}$ are the corresponding path-loss coefficients. Moreover, $\tilde{\mathbf{G}}$, $\tilde{\mathbf{h}}_{r,k}$, and $\tilde{\mathbf{h}}_{d,k}$ are the NLOS components, each element of which is modeled as an independent and identically distributed complex Gaussian random variable with zero mean and unit variance, while $\bar{\mathbf{G}}$, $\bar{\mathbf{h}}_{r,k}$, and $\bar{\mathbf{h}}_{d,k}$ are the LOS components modeled as follows

$$\bar{\mathbf{G}} = \mathbf{a}_N\left(\theta_{AoA}, \phi_{AoA}\right)\mathbf{a}_M^H(\theta_{AoD,1}), \tag{4}$$

$$\bar{\mathbf{h}}_{r,k} = \mathbf{a}_N\left(\theta_{AoD}, \phi_{AoD}\right), \tag{5}$$

$$\bar{\mathbf{h}}_{d,k} = \mathbf{a}_M\left(\theta_{AoD,2}\right), \tag{6}$$

where θ_{AoA} and ϕ_{AoA} are the azimuth and elevation angles of arrival (AoAs) from the BS to the RIS, $\theta_{AoD,1}$ is the azimuth angle of departure (AoD) from the BS to the RIS, θ_{AoD} and ϕ_{AoD} are the azimuth and elevation AoDs from the RIS to the *k*-th user, and $\theta_{AoD,2}$ is the azimuth AoD from the BS to the *k*-th user. $\mathbf{a}_M(\theta) = [1, \cdots, e^{j2\pi\frac{d_1}{\lambda}\sin(\theta)}, \cdots, e^{j2\pi(M-1)\frac{d_1}{\lambda}\sin(\theta)}]^H$ and $\mathbf{a}_N(\theta,\phi) = [1, \cdots, e^{j2\pi\frac{d_2}{\lambda}(n_x\cos(\theta)\sin(\phi)+n_y\sin(\theta)\sin(\phi))}, \cdots, e^{j2\pi\frac{d_2}{\lambda}((N_x-1)\cos(\theta)\sin(\phi)+(N_y-1)\sin(\theta)\sin(\phi))}]^H$ are the array responses of ULA and UPA, where d_1 and d_2 are the antenna spacings, and λ is the carrier wavelength.

The signal received at the *k*-th user can be expressed as

$$y_k = (\mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G} + \mathbf{h}_{d,k}^H)\sum_{i=1}^K \mathbf{w}_i x_i + n_k, \tag{7}$$

where $\mathbf{\Theta} = \mathrm{diag}(\phi_1, \phi_2, \cdots, \phi_N)$ represents the RIS diagonal phase shifts matrix with $\phi_n = e^{j\theta_n}$, $\theta_n \in [0, 2\pi)$, and n_k is the complex Gaussian noise with variance σ_k^2, $\mathbf{w}_i \in \mathbb{C}^{M\times 1}$ is the transmit beamforming vector for the *i*-th user with the power constraint $\sum_{i=1}^K \|\mathbf{w}_i\|^2 \le P$, P is the maximum transmit power at the BS, and x_i represents the transmitted signal for the *i*-th user satisfying $\mathbb{E}[x_i^2] = 1$. Let us define $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_K]$.

Thus, for the k-th user, we can obtain its signal-to-interference-plus-noise ratio (SINR)

$$SINR_k = \frac{|(\mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G} + \mathbf{h}_{d,k}^H)\mathbf{w}_k|^2}{\sum_{i=1,i\neq k}^K |(\mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G} + \mathbf{h}_{d,k}^H)\mathbf{w}_i|^2 + \sigma_k^2}, \tag{8}$$

and the corresponding spectral efficiency is

$$r_k = \log_2(1 + SINR_k). \tag{9}$$

We consider the TTS transmission scheme [11], where the BS beamforming and the RIS phase shifts are designed exploiting I-CSI and S-CSI, respectively. The design objective is to maximize the average sum spectral efficiency, subject to the transmit power constraint and the unit module constraint of $\mathbf{\Theta}$. Thus, we have the following optimization problem

$$\begin{aligned} (\mathrm{P1}): \max_{\mathbf{\Theta}} \quad & \mathbb{E}\left\{\max_{\mathbf{w}_k} \sum\nolimits_{k=1}^K r_k\right\}, \\ \text{s.t.} \quad & \sum\nolimits_{k=1}^K \|\mathbf{w}_k\|^2 \le P, \\ & |\phi_n| = 1. \end{aligned} \tag{10}$$

The objective function of problem (P1) includes an internal rate-maximization problem, which optimizes the short-term transmint beamforming at BS to maximize the spectral efficiency of each time slot given the RIS phase shifts. Additionally, there is an outer rate-maximization problem that optimizes the long-term RIS phase shifts to maximize the average spectral efficiency within the channel statistics coherence time. However, the coupling of the BS beamforming and the RIS phase shifts in the objective function makes it difficult to solve. Although numerical methods such as the stochastic successive convex approximation (SSCA) algorithm [11] have been proposed, the computational complexity is extremely high.

III. DRL Based Optimization Algorithm

In this section, to solve the optimization problem (P1) with low complexity, a novel two-agents PPO based algorithm, referred to as PPOΘ-PPOW, is proposed. Firstly, we will provide a brief introduction of the PPO algorithm. Then, the details of the proposed PPO based algorithm will be described.

A. PPO Description

PPO algorithm [14] is a policy based DRL algorithm proposed by OpenAI, which can deal with the problems for both continuous space and discrete space. The PPO agent contains a critic network, an actor network, a replay buffer, and an optimizer. The critic network is used to output the value function that evaluates the current state with parameter vectors $\mathbf{\Omega}_c$, the actor network is used to obtain action with parameter vectors $\mathbf{\Omega}_a$, the replay buffer U is used to store experience, and the optimizer is used to optimize network parameters. During timestep t in each episode, the agent observes the environment state s_t and selects action a_t according to the policy $\pi_{\mathbf{\Omega}_a}(a_t|s_t)$. When the action is done, the environment goes to the next state s_{t+1}, and the reward r_t is provided. The value function $V_{\mathbf{\Omega}_c}^{\pi}(s) = \mathbb{E}_\pi[G_t|s_t = s]$ is the expectation of accumulated reward of state s under policy π, where $G_t = \sum_{\tau=0}^{\infty}\gamma^\tau r_{t+\tau}$ is the cumulative reward value with the discount factor $\gamma \in (0, 1]$. Then, the agent puts the experience $\{s_t, a_t, \pi_{\mathbf{\Omega}_a}(a_t|s_t), r_t, s_{t+1}\}$ into the replay buffer. The optimizer selects experience from the replay buffer with minibatch B to update the network so as to maximize the objective function

$$L(\mathbf{\Omega}_a) = \mathbb{E}_t\left[\min\left(r_t(\mathbf{\Omega}_a)A_t, \mathrm{clip}(r_t(\mathbf{\Omega}_a), 1-\varepsilon, 1+\varepsilon)A_t\right)\right], \tag{11}$$

where $r_t(\mathbf{\Omega}_a) = \frac{\pi_{\mathbf{\Omega}_a}(a_t|s_t)}{\pi_{\mathbf{\Omega}_{a,old}}(a_t|s_t)}$ denotes the probability ratio, $\mathbf{\Omega}_{a,old}$ is the policy parameters before the update, $\text{clip}\,(a,b,c)$ is the function that clips the probability ratio a within the range of $[b,c]$, ε is a hyperparameter controlling the clip range, A_t is the advantage function calculated by the general advantage estimation (GAE) method to make the trade-off between variance and bias, i.e.,

$$A_t = \sum\nolimits_{l=0}^{\infty} (\gamma\xi)^l \delta_{t+1}^V, \tag{12}$$

$$\delta_t^V = r_t + \gamma V_{\mathbf{\Omega}_c}^{\pi}(s_{t+1}) - V_{\mathbf{\Omega}_c}^{\pi}(s_t), \tag{13}$$

where ξ is the GAE parameter.

To further improve the exploration ability of the algorithm, a policy entropy with coefficient c can be added to (11), thus, leading to the final objective function

$$\begin{aligned} L(\mathbf{\Omega}_a) =& \mathbb{E}_t[\min\left(r_t(\mathbf{\Omega}_a)A_t, \text{clip}(r_t(\mathbf{\Omega}_a), 1-\varepsilon, 1+\varepsilon)A_t\right) \\ &+ cH(\pi_{\mathbf{\Omega}_a}(\cdot|s_t))], \end{aligned} \tag{14}$$

where $H(\pi_{\mathbf{\Omega}_a}(\cdot|s_t))$ is the policy entropy.

B. Proposed PPO Based Algorithm

To tackle the optimization problem (P1), we propose a two-agent PPO based algorithm to jointly optimize the beamforming at the BS and RIS phase shifts, as shown in Fig. 2.

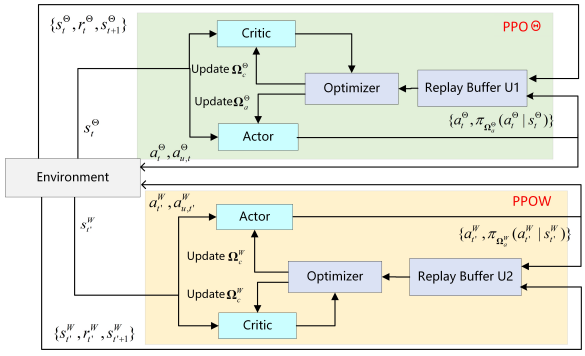

Fig. 2. The proposed PPOΘ-PPOW algorithm framework.

The proposed algorithm contains two agents named as PPOΘ and PPOW, which are constructed based on PPO framework. PPOΘ uses S-CSI to obtain RIS phase shifts, while PPOW uses I-CSI within each channel statistics coherence time to obtain beamforming at the BS. Both PPOΘ and PPOW contain a critic network with parameter vectors $\mathbf{\Omega}_c^\Theta$, $\mathbf{\Omega}_c^W$ and an actor network with parameter vectors $\mathbf{\Omega}_a^\Theta$, $\mathbf{\Omega}_a^W$. It is worth noting that the action a_t is obtained by sampling the Gaussian distribution, which is formed by the mean and variance in the actor network in this letter. Therefore, we also take the mean in the actor network as action $a_{u,t}$ for the training of another agent. Both agents also contain an experience replay buffer of their own named U1 and U2. The state, action, and reward of the two agents are designed as follows, where the timestep t corresponds to the channel statistics coherence time, and t' corresponds to the time slot during each channel statistics coherence time.

1) State: The state of PPOΘ at timestep t is set as $s_t^\Theta = \{\sqrt{L_{r,1}(t)}\bar{\mathbf{h}}_{r,1}(t), \sqrt{L_{d,1}(t)}\bar{\mathbf{h}}_{d,1}(t), \cdots, \sqrt{L_{r,K}(t)}\mathbf{h}_{r,K}(t), \sqrt{L_{d,K}(t)}\bar{\mathbf{h}}_{d,K}(t)\}$. PPOW observes the equivalent channel vector $\mathbf{h}_k = \mathbf{h}_{r,k}^H\mathbf{\Theta}\mathbf{G} + \mathbf{h}_{d,k}^H$, so its state at the time slot t' is set as $s_{t'}^W = \{\mathbf{h}_1(t'), \cdots, \mathbf{h}_K(t')\}$. The state with respect to the real and imaginary parts of CSI is separately input into neural network.
2) Action: Set $a_t^\Theta = \{\theta_1(t), \cdots, \theta_N(t)\}$ as the action of PPOΘ at timestep t, the element of which is the reflection phase shifts of the RIS. Set $a_{t'}^W = \{\text{Re}(\mathbf{W}(t')), \text{Im}(\mathbf{W}(t'))\}$ as the action of PPOW at time slot t', which is normalized and reformulated into the beamforming matrix as $\mathbf{W}(t') = \frac{\sqrt{P}\mathbf{W}(t')}{\|\mathbf{W}(t')\|_F}$. It is worth noting that in this algorithm, PPOΘ and PPOW need to output action a_t^Θ, $a_{t'}^W$, as well as action $a_{u,t}^\Theta$, $a_{u,t'}^W$. $a_{u,t}^\Theta$ is generated by PPOΘ for the training of PPOW, while $a_{u,t'}^W$ is generated by PPOW for the training of PPOΘ.
3) Reward: PPOΘ takes into account the average performance of all time slots under the same S-CSI, so the reward at timestep t is set as $r_t^\Theta = \mathbb{E}\left[\sum_{k=1}^K r_k(t')\right]$, which represents the expectation with respect to the I-SCI during the channel statistical coherence time. The reward of PPOW at time slot t' is set as $r_{t'}^W = \sum_{k=1}^K r_k(t')$,

Algorithm 1 The PPOΘ-PPOW Based Algorithm

Initialize $\mathbf{\Omega}_a^\Theta$, $\mathbf{\Omega}_c^\Theta$, $\mathbf{\Omega}_{a,old}^\Theta \leftarrow \mathbf{\Omega}_a^\Theta$, $\mathbf{\Omega}_{c,old}^\Theta \leftarrow \mathbf{\Omega}_c^\Theta$.
Initialize $\mathbf{\Omega}_a^W$, $\mathbf{\Omega}_c^W$, $\mathbf{\Omega}_{a,old}^W \leftarrow \mathbf{\Omega}_a^W$, $\mathbf{\Omega}_{c,old}^W \leftarrow \mathbf{\Omega}_c^W$.

for $episode = 1, 2, \cdots, L$ **do**
  Initialize the environment and replay buffer U1 and U2;
  for $t = 1, 2, \cdots, T$ **do**
    PPOΘ observes state s_t^Θ and selects action a_t^Θ;
    for $n = 1, 2, \cdots, T_H$ **do**
      $t' = tT_H + n$
      PPOW observes state $s_{t'}^W$ and selects action $a_{u,t'}^W$;
    end for
    Get the reward r_t^Θ and the next state s_{t+1}^Θ;
    Store transition $\{s_t^\Theta, a_t^\Theta, \pi_{\mathbf{\Omega}_a^\Theta}(a_t^\Theta|s_t^\Theta), r_t^\Theta, s_{t+1}^\Theta\}$ into replay buffer U1;
  end for
  for $t = T+1, T+2, \cdots, 2T$ **do**
    PPOΘ observes state s_t^Θ and selects action $a_{u,t}^\Theta$;
    for $n = 1, 2, \cdots, T_H$ **do**
      $t' = tT_H + n$;
      PPOW observes state $s_{t'}^W$ and selects action $a_{t'}^W$;
      Get the reward $r_{t'}^W$ and the next state $s_{t'+1}^W$;
      Store transition $\{s_{t'}^W, a_{t'}^W, \pi_{\mathbf{\Omega}_a^W}(a_{t'}^W|s_{t'}^W), r_{t'}^W, s_{t'+1}^W\}$ into replay buffer U2;
    end for
  end for
  Compute advantages $\{A_t^\Theta\}_{t=1}^T$, $\{A_{t'}^W\}_{t'=1}^{T_HT}$ using (12);
  Compute $\{y_t^\Theta\}_{t=1}^T$, $\{y_{t'}^W\}_{t'=1}^{T_HT}$ using (17);
  for $j = 1, 2, .., J$ **do**
    Update $\mathbf{\Omega}_a^\Theta$, $\mathbf{\Omega}_c^\Theta$, $\mathbf{\Omega}_a^W$, $\mathbf{\Omega}_c^W$ with minibatch size B^Θ, B^W respectively using (15) and (18);
  end for
  update $\mathbf{\Omega}_{a,old}^\Theta \leftarrow \mathbf{\Omega}_a^\Theta$, $\mathbf{\Omega}_{c,old}^\Theta \leftarrow \mathbf{\Omega}_c^\Theta$, $\mathbf{\Omega}_{a,old}^W \leftarrow \mathbf{\Omega}_a^W$, $\mathbf{\Omega}_{c,old}^W \leftarrow \mathbf{\Omega}_c^W$;
end for

which is the short-term sum rate.

C. Training and Online Working Process

Noting that usually the BS can obtain global channel state information through the channel estimation method and has strong information processing capability, we consider deploying the two agents at the BS. In this case, the agents can obtain channel state information from the BS, take it as the input, and output the corresponding action. Then, the action of PPOΘ is transmitted through a separate link to the RIS controller to adjust the RIS phase shifts. Based on the channel state information and output actions, each agent can calculate the its reward at the BS.

In the l-th episode, $2T$ sets of S-CSI are used for training, and each S-CSI contains T_H time slots. The training process is described as follows.

The first T S-CSI is for generating PPOΘ experience: At timestep t, PPOΘ observes the current state s_t^Θ, inputs it to critic network and actor network: the former outputs value function $V^\pi_{\mathbf{\Omega}^\Theta_{c,old}}(s_t^\Theta)$, and the latter outputs action a_t^Θ. Then the RIS phase shifts of timestep t are configured using action a_t^Θ. At time slot t', PPOW observes the current state $s_{t'}^W$ and the actor network outputs the action $a_{u,t'}^W$. Then the beamforming of BS is configured using the action $a_{u,t'}^W$. After T_H time slots, the environment issues a reward r_t^Θ and is updated to the next state s_{t+1}^Θ, then $\{s_t^\Theta, a_t^\Theta, \pi_{\mathbf{\Omega}_a^\Theta}(a_t^\Theta|s_t^\Theta), r_t^\Theta, s_{t+1}^\Theta\}$ is stored in the replay buffer U1.

The last T S-CSI is for generating PPOW experience: At timestep t, PPOΘ observes the current state s_t^Θ and the actor network outputs the action $a_{u,t}^\Theta$. Then the RIS phase shifts of timestep t are configured using action $a_{u,t}^\Theta$. At time slot t', PPOW observes the current state $s_{t'}^W$, critic network and actor network output value function $V^\pi_{\mathbf{\Omega}^W_{c,old}}(s_{t'}^W)$ and action $a_{t'}^W$ respectively. Then the transmit beamforming of BS is configured using the action $a_{t'}^W$. After the action $a_{t'}^W$ is executed, the environment issues a reward $r_{t'}^W$ and is updated to the next state $s_{t'+1}^W$, then $\{s_{t'}^W, a_{t'}^W, \pi_{\mathbf{\Omega}_a^W}(a_{t'}^W|s_{t'}^W), r_{t'}^W, s_{t'+1}^W\}$ is stored in the replay buffer U2.

Each episode carries out J times of parameter updates. In each time j, the optimizer randomly samples minibatches of size B from the replay buffer. The purpose of updating actor network is to maximize $L(\mathbf{\Omega}_a)$, with the policy gradient

$$\begin{aligned}\Delta\mathbf{\Omega}_a =& \frac{1}{B}\sum\nolimits_{i=1}^{B}\nabla_{\mathbf{\Omega}_a}[\min(r_i(\mathbf{\Omega}_a)A_i,\\ &\mathrm{clip}(r_i(\mathbf{\Omega}_a),1-\varepsilon,1+\varepsilon)A_i)+cH(\pi(\cdot|s_i))].\end{aligned} \tag{15}$$

The loss function of the critic network is given by

$$J(\mathbf{\Omega}_c)=\mathbb{E}_t\left[(y_t-V_{\mathbf{\Omega}c}(s_t))^2\right], \tag{16}$$

where

$$y_t = A_t + V_{\mathbf{\Omega}_{c,old}}(s_t). \tag{17}$$

The update gradient of critic network is denoted as

$$\Delta\mathbf{\Omega}_c=\frac{1}{B}\sum\nolimits_{i=1}^{B}\nabla_{\mathbf{\Omega}_c}[(y_i-V_{\mathbf{\Omega}_c}(s_i))^2]. \tag{18}$$

The details of the proposed algorithm are given in Algorithm 1. In the online working stage, during the channel statistics coherence time, PPOΘ observes the current S-CSI state and gets actions $a_{u,t}^\Theta$ to configure the RIS phase shifts. At each time slot, PPOW observes the current I-CSI state and gets actions $a_{u,t'}^W$ to configure the beamforming of the BS, which takes a low runtime. As the RIS phase shift is discrete in practice, it can be quantized to the closest discrete value to meet the requirements of practical configuration.

IV. Experiment results

In this section, we evaluate the performance of our proposed algorithm through the simulation results. The BS and the RIS are positioned at (0m, 0m, 30m) and (100m, 20m, 10m), respectively. Users are uniformly distributed within a circular region centered at (150m, 0m, 1.5m) with a radius of 8m. The large-scale fading is modeled as $L = C_0(\frac{d}{D_0})^{-\alpha}$, where $C_0 = -30$dB is the path loss at the reference distance $D_0 = 1$ m, α represents the path loss exponent, and d represents the link distance. The path loss exponents for the BS-RIS, BS-users, and RIS-users links are set to $\alpha_{Bu} = 3.8$, $\alpha_{BR} = 2.4$, $\alpha_{Ru} = 2.2$ [15]. Unless otherwise specified, other system parameters are set as follows: $\sigma_k^2 = -90$dBm, $P = 10$dBm, $T_H = 10$, $M = 8$, $K = 2$, $\beta_{1,k} = \beta_{2,k} = 5$dB, $\beta_{3,k} = 2$dB.

In the proposed PPOΘ-PPOW framework, all neural networks are four-layered DNN, which consists of two hidden layers of 256 and 128 neurons. We use the Adam optimizer to update the parameters. In terms of the hyperparameter settings of PPO, the batch sizes are set to $B^\Theta = 64$, $B^W = 128$, and the learning rates of the actor network and critic network are $u_a = 0.0001$, $u_c = 0.0003$, respectively. We set the discount reward factor $\gamma = 0.9$, the GAE parameter $\xi = 0.95$, the clip fraction $\varepsilon = 0.2$, the number of parameter updates $J = 10$, and the policy entropy coefficient $c = 0.01$. Moreover, 2200 sets of S-CSI are simulated, where 2000 are used for training and 200 for validation.

For comparison, we consider the WMMSE-RCG algorithm [5] and SSCA algorithm [11] as baselines, which are near-optimal numerical optimization algorithms with only I-CSI and under TTS transmission schemes respectively. At the same time, we also show the performance of a so-called PPOΘ-WMMSE algorithm, which replaces the RIS phase shift design part of the SSCA algorithm in [11] by the PPOΘ we proposed.

Figure 3 illustrates the convergence performance of the proposed PPOΘ-PPOW algorithm, where the number of reflecting elements on RIS is $N = 64$. We can notice that with the increase of episodes, the rewards of both agents rise steadily and reach convergence at about 3000 episodes.

Figure 4 demonstrates the spectral efficiency performance of various algorithms versus the BS maximum transmit power. In this figure, $N = 64$. It shows that the deployment of RIS and proper phase shift can significantly improve the system performance. Notably, the difference in performance between the SSCA algorithm and the WMMSE-RCG algorithm is small, further confirming the feasibility of the TTS transmission scheme. Most importantly, it is evident that the spectral efficiencies of the proposed PPOΘ-WMMSE and PPOΘ-PPOW algorithms are close to that of the SSCA algorithm, which can reach 95.85% and 91.82% of the SSCA respectively when $P = 10$dBm, and the proposed PPOΘ-WMMSE is

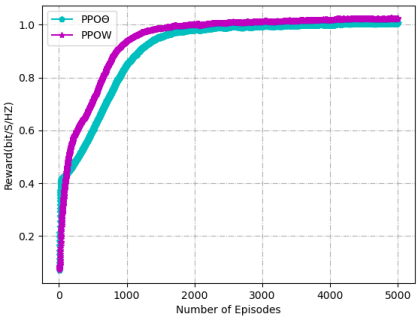

Fig. 3. Convergence performance of the proposed algorithm.

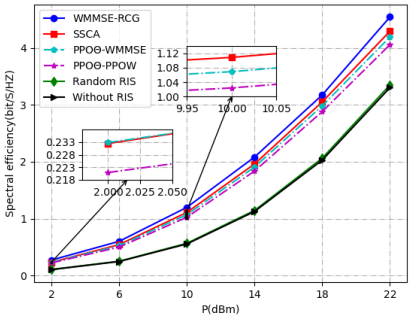

Fig. 4. Spectral efficiency vs. total power budget P.

even better than the SSCA algorithm when $P = 2$dBm. In addition, to illustrate the generalization ability of the proposed algorithm, we consider expanding the test area of users by keeping the center of the test area unchanged and extending the radius to 12m in the online working stage. In this case, the spectral efficiencies of the PPOΘ-WMMSE and PPOΘ-PPOW algorithm can still reach 93.54% and 89.82% of SSCA algorithm respectively when $P = 10$dBm.

In Fig. 5, we compare the spectral efficiency under different number of RIS elements with $N_x = 8$. As the number of reflecting elements in the RIS increases, the performance of all algorithms will become better. Meanwhile, the proposed algorithm is stable under different number of RIS elements.

Table I compares the online running time of various algorithms under different number of RIS reflecting elements. It is evident that the time consumption of the SSCA algorithm increases significantly with N. In contrast, the PPOΘ-PPOW algorithm can achieve comparable performance with an extremely short time, while the PPOΘ-WMMSE algorithm has a longer online processing time but slightly better performance.

V. Conclusion

In this letter, we investigated beamforming optimization of RIS-assisted multi-user MISO system with the TTS transmission scheme. A two-agents PPO based algorithm, referred to as PPOΘ-PPOW algorithm, was proposed to jointly optimize short-term transmit beamforming at the BS and long-term RIS phase shifts. Experimental results indicated that the proposed algorithm can approximate the performance of the SSCA algorithm with extremely low time overhead.

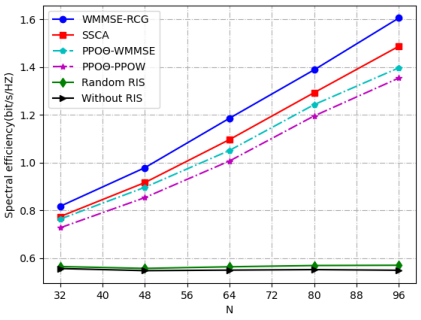

Fig. 5. Spectral efficiency vs. the number of elements N on RIS.

TABLE I
RUNNING TIME COMPARISON

N	Running Time(ms)		
	SSCA	PPOΘ-WMMSE	PPOΘ-PPOW
48	2340.98	23.28	0.31
64	3218.56	26.47	0.42
80	3942.88	31.12	0.49

References

[1] T. J. Cui, M. Q. Qi, X. Wan, J. Zhao, and Q. Cheng, "Coding metamaterials, digital metamaterials and programmable metamaterials," *Light: Science & Applications*, vol. 3, no. 10, p. e218, Oct. 2014.

[2] J. Sang, Y. Yuan, W. Tang, Y. Li, X. Li, S. Jin, Q. Cheng, and T. J. Cui, "Coverage Enhancement by Deploying RIS in 5G Commercial Mobile Networks: Field Trials," *IEEE Wireless Commun.*, 2023, early access.

[3] C. Huang *et al.*, "Holographic MIMO Surfaces for 6G Wireless Networks: Opportunities, Challenges, and Trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118-125, Oct. 2020.

[4] C. Huang *et al.*, "Reconfigurable Intelligent Surfaces for Energy Efficiency in Wireless Communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157-4170, Aug. 2019.

[5] H. Guo *et al.*, "Weighted Sum-Rate Maximization for Reconfigurable Intelligent Surface Aided Wireless Networks," *Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3064-3076, May. 2020.

[6] K. Feng, X. Li, Y. Han and Y. Chen, "Joint beamforming optimization for reconfigurable intelligent surface-enabled MISO-OFDM systems," *China Commun.*, vol. 18, no. 3, pp. 63-79, Mar. 2021.

[7] X. Gan, C. Zhong, C. Huang, and Z. Zhang, "RIS-Assisted Multi-User MISO Communications Exploiting Statistical CSI," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6781-6792, Oct. 2021.

[8] L. Jiang, X. Li, M. Matthaiou, and S. Jin, "Joint User Scheduling and Phase Shift Design for RIS Assisted Multi-cell MISO Systems," *IEEE Wireless Commun. Lett.*, vol. 12, no. 3, pp. 431-435, Mar. 2023.

[9] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large Intelligent Surface-Assisted Wireless Communication Exploiting Statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8238-8242, Aug. 2019.

[10] H. Guo, Y. C. Liang, and S. Xiao, "Intelligent Reflecting Surface Configuration With Historical Channel Observations," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1821–1824, Jun. 2020.

[11] M.-M. Zhao *et al.*, "Intelligent Reflecting Surface Enhanced Wireless Networks: Two-Timescale Beamforming Optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 2-17, Jan. 2021.

[12] K. Feng *et al.*, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, pp. 745–749, May. 2020.

[13] 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, pp. 1839–1850, Aug. 2020.

[14] J. Schulman *et al.*, "Proximal policy optimization algorithms," Jul. 2017. [Online]. Available: arXiv:1707.06347.

[15] W. Huang *et al.*, "Reconfigurable Intelligent Surface-Enhanced Broadband OFDM Communication Based on Deep Reinforcement Learning," in *Proc. IEEE VTC-Fall*, pp. 1-6, 2021.