Statistical CSI-based Beamforming for RIS-Aided Multiuser MISO Systems via Deep Reinforcement Learning

Mahdi Eskandari, Huiling Zhu, Arman Shojaeifard, and Jiangzhou Wang. *Fellow, IEEE*

***Abstract*—This letter presents a novel joint beamforming algorithm for reconfigurable intelligent surfaces (RIS) in multiuser multiple-input single-output (MISO) wireless communications. At first, by utilizing statistical channel state information (CSI) instead of instantaneous CSI, we significantly reduce channel estimation overhead. Then, the optimization of beamforming weights is accomplished using the proximal policy optimization (PPO) algorithm, a well-established actor-critic-based reinforcement learning (RL) approach. The impact of system parameters on user sum rate is also analyzed through simulations. The results show the PPO algorithm outperforms the existing methods by combining beamforming techniques with statistical CSI.**

I. Introduction

Reconfigurable intelligent surfaces (RIS) consist of reflecting components, each of which can be electronically adjusted to manipulate the phase shift of incoming signals. By strategically setting the phase shift of individual RIS elements, deploying RIS between transmitting and receiving nodes can improve transmission reliability [1].

However, previous studies have shown that the inquiry of obtaining instantaneous channel state information (I-CSI) results in a significant burden on using RISs [2]–[5]. Statistics channel state information (S-CSI) can be used as an alternative, which evolves more slowly and is more easily available. In [6] and [7], only S-CSI was utilized to design BS beamforming and RIS phase shifts. Despite their effectiveness, the iterative optimization algorithms employed exhibit a relatively high level of computational complexity. In recent years, the field has seen the application of deep reinforcement learning (RL) in RIS-aided transmission to develop solutions with reduced computational complexity, e.g. in [8] and [9]. However, the research is still based on I-CSI.

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 deep RL. Specifically, the downlink data rate of the users is derived in closed-form which is a function of S-CSI variables including the Rician factor, large-scale path loss, and angles of arrival and departure. The derived data rate is then maximized with respect to active beamforming at the BS and passive beamforming at the RIS. Simulation results show the effectiveness of the proposed method. The main contributions of this paper are summarised as follows:

- For multi-user MISO (MU-MISO) systems an approximated closed-form expression of the ergodic sum rate is obtained. This leads to an optimization of a joint active-passive beamforming design.
- The joint optimization problem is intractable due to the nonconvex constraint and the intricate relationship between the BS transmit beamformer and the RIS passive beamformer. To tackle this issue, a proximal policy optimization (PPO) based algorithm is proposed to solve the optimization problem.

Mahdi Eskandari, Huiling Zhu and Jiangzhou Wang are with the School of Engineering, University of Kent, UK. (email: {me377, H.Zhu, j.z.wang}@kent.ac.uk). Arman Shojaeifard, is with the InterDigital, London EC2A 3QR, U.K. (email: arman.shojaeifard@interdigital.com.)

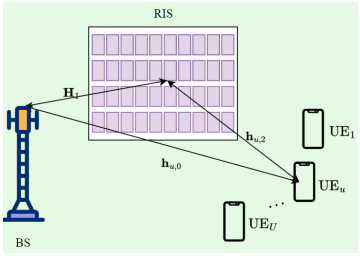

Fig. 1: RIS-aided multi-user MISO system

II. System Model

Consider an MU-MISO communication system with a BS of N antennas communicating with U single-antenna UEs as shown in Fig. 1. Transmission between the BS and UEs is assisted by a RIS equipped with M reflecting elements. Denoting ξ_m as the reflection coefficient of the m-th element of the RIS, the reflection matrix of the RIS panel can be expressed as $\mathbf{\Xi} = \mathrm{diag}(\xi_1, \ldots, \xi_M)$ where $\xi_m = e^{j\phi_m}$ with ϕ_m denoting the phase shift of the m-th element of the RIS. Denote $\mathbf{h}_{u,0} \in \mathbb{C}^{N\times 1}$, $\mathbf{H}_1 \in \mathbb{C}^{M\times N}$ and $\mathbf{h}_{u,2} \in \mathbb{C}^{M\times 1}$ as the direct channel from the BS to the u-th UE, from the BS to RIS, and from the RIS to u-th UE, respectively. Hence, the equivalent effective channel from the transmitter to the u-th UE could be obtained as $\mathbf{h}_u^T \triangleq \mathbf{h}_{u,0}^T + \mathbf{h}_{u,2}^T \mathbf{\Xi} \mathbf{H}_1$. Considering the line-of-sight (LOS) link exists between the BS and the RIS and between the RIS and UEs, Rician distribution is used to model these channels. The channel matrix $\mathbf{h}_{u,2}$ between the RIS and u-th UE is represented as $\mathbf{h}_{u,2} = \sqrt{\frac{\delta_{u,2}\kappa_{u,2}}{1+\kappa_{u,2}}}\bar{\mathbf{h}}_{u,2} + \sqrt{\frac{\delta_{u,2}}{1+\kappa_{u,2}}}\tilde{\mathbf{h}}_{u,2}$ [7] where $\sqrt{\delta_{u,2}}$ denotes the distance dependent path-loss factor, and $\kappa_{u,2}$ denotes the Rician factor between

the RIS and the u-th UE. On the other side, the channel of the BS-RIS link, and the channel of the link from the BS to the u-th UE can be expressed as $\mathbf{H}_1 = \sqrt{\frac{\delta_1\kappa_1}{1+\kappa_1}}\bar{\mathbf{H}}_1 + \sqrt{\frac{\delta_1}{1+\kappa_1}}\tilde{\mathbf{H}}_1$ and $\mathbf{h}_{u,0} = \sqrt{\frac{\delta_{u,0}\kappa_{u,0}}{1+\kappa_{u,0}}}\bar{\mathbf{h}}_{u,0} + \sqrt{\frac{\delta_{u,0}}{1+\kappa_{u,0}}}\tilde{\mathbf{h}}_{u,0}$ [7] with δ_1 and κ_1 being the distance-dependent path-loss and Rician factor of the BS-RIS link. $\delta_{u,0}$ and $\kappa_{u,0}$ denote the distance-dependent path-loss and Rician factor of the link between the BS and the u-th UE. Furthermore, $\bar{\mathbf{h}}_{u,2} = \mathbf{a}_{\text{RIS}}(\phi_u^{(\text{RIS})}, \psi_u^{(\text{RIS})})$, where $\phi_u^{(\text{RIS})}$ ($\psi_u^{(\text{RIS})}$) is the azimuth (elevation) angle of departure (AoD) from the RIS to the u-th UE. By assuming a uniform planar arrays (UPA) RIS with M_{H} and M_{V} reflecting elements at horizontal and vertical directions, respectively (i.e., $M = M_{\text{H}}M_{\text{V}}$), the array response vector at the RIS is given by

$$\mathbf{a}_{\text{RIS}}(\phi_u^{(\text{RIS})},\psi_u^{(\text{RIS})}) = \frac{1}{\sqrt{M_{\text{H}}M_{\text{V}}}}[1, \ldots, e^{\jmath\frac{2\pi}{\lambda}D(h\sin\phi_u^{(\text{i})}\sin\psi_u^{(\text{i})})+v\cos\psi_u^{(\text{i})})}, \ldots, e^{\jmath\frac{2\pi}{\lambda}D(X_{\text{H}}\sin\phi_u^{(\text{i})}\sin\psi_u^{(\text{i})})+(X_{\text{V}}-1)\cos\psi_u^{(\text{i})})}]^T, \tag{1}$$

where λ is the wavelength and D is the distance between antenna elements and also (h, v) is the index of a reflective element in horizontal and vertical directions, respectively. Moreover, $\bar{\mathbf{H}}_1 = \mathbf{a}_{\text{RIS}}(\phi^{(\text{RIS})}, \psi^{(\text{RIS})})\mathbf{a}_{\text{BS}}(\phi^{(\text{BS})}, \psi^{(\text{BS})})^H$ where $\phi^{(\text{RIS})}$ and $\psi^{(\text{RIS})}$ are the azimuth and elevation angle of arrival (AoA) to the RIS, and $\phi^{(\text{BS})}$ and $\psi^{(\text{BS})}$ denote the azimuth and elevation AoD from the BS to the RIS. $\bar{\mathbf{h}}_{u,0} = \mathbf{a}_{\text{BS}}(\phi_u^{(\text{BS})}, \psi_u^{(\text{BS})})$, where $\phi_u^{(\text{BS})}$ and $\psi_u^{(\text{BS})}$ being the azimuth and elevation AoD from the BS in the direction of the u-th UE. With the appropriate changes in notation, the array response vectors of the BS arrays can be written similarly to those for the RIS. By considering the fact that there is no spatial correlation among the antennas, the non-line-of-sight (NLoS) components of the channels, $\tilde{\mathbf{h}}_{u,2}$, $\tilde{\mathbf{h}}_{u,0}$ and $\tilde{\mathbf{H}}_1$ are independently and identically distributed (i.i.d.) complex Gaussian random variables with zero mean and unit variance. The deterministic components of the channel, i.e., $\bar{\mathbf{H}} = \{\bar{\mathbf{h}}_{u,2}, \bar{\mathbf{H}}_1, \bar{\mathbf{h}}_{u,0}\}$ and the Rician factors and large-scale path-losses of all the channels are considered as the known statistical variables. A method for estimating these variables can be found in [10]. Note that these S-CSI values do not inherently contain information related to the phase shifts of the RIS. Also, since, it is assumed that the users are static, the variations in S-CSI and RIS phase shifts are decoupled and operate independently of each other. Based on the above descriptions, the received signal at the u-th UE $y_u \in \mathbb{C}$ is given by $y_u = \sum_{u=1}^{U}\sqrt{P_u}\mathbf{h}_u^T\mathbf{f}_u x_u + n_u$ where P_u and x_u are the transmit power and signal for the u-th UE, respectively. Furthermore, $\mathbf{f}_u$ is the beamforming vector for the u-th UE, and $n_u \in \mathbb{C}$ is a circularly symmetric complex additive Gaussian noise with $\mathbb{E}[n_u n_u^H] = \sigma^2$.

III. Problem Formulation

This paper aims to maximize the sum data rate of the RIS-aided wireless system. Since the collection of I-CSI is time-consuming in RIS-aided environments, the joint active and passive beamforming is performed based on S-CSI, i.e. angular information and Rician factors of the channels, which are easier to measure and feedback [11]. Given the transmit power P at the BS and equal power allocation, the ergodic sum-rate maximization problem can be formulated as

$$\mathcal{P}_1 : \max_{\boldsymbol{\Xi},\mathbf{f}_u} \sum_{u=1}^{U} \underbrace{\mathbb{E}\left[\log_2\left(1 + \frac{\frac{P}{U}\mathbf{f}_u^H\mathbf{h}_u^*\mathbf{h}_u^T\mathbf{f}_u}{\sigma_2 + \frac{P}{U}\sum_{i\neq u}\mathbf{f}_i^H\mathbf{h}_u^*\mathbf{h}_u^T\mathbf{f}_i}\right)\right]}_{R_u} \tag{2}$$

$$\text{s.t} \quad \phi_m \in [0, 2\pi), \quad m = 1, \ldots, M, \tag{2a}$$

$$\|\mathbf{f}_u\|_2^2 = 1, \quad u = 1, \ldots, U, \tag{2b}$$

where the expectation is over a long period of channel realizations. The maximization problem $\mathcal{P}_1$ is mathematically intractable due to the existence of the expectation and unit norm constraints. In order to find the optimal solution, firstly, we will derive the expectation and find the upper bound of it. By using Jensen's inequality, R_u can be written as

$$R_u \leq \log_2\left(\sigma^2 + \frac{P}{U}\sum_{i=1}^{U}\mathbf{f}_u^H\mathbb{E}\left[\mathbf{h}_u^*\mathbf{h}_u^T\right]\mathbf{f}_u\right) - \log_2\left(\sigma^2 + \frac{P}{U}\sum_{i\neq u}\mathbf{f}_i^H\mathbb{E}\left[\mathbf{h}_u^*\mathbf{h}_u^T\right]\mathbf{f}_i\right). \tag{3}$$

Based on (3), the solution of the problem relies on finding $\mathbb{E}[\mathbf{h}_u^*\mathbf{h}_u^T]$.

Theorem 1. *$\mathbb{E}[\mathbf{h}_u^*\mathbf{h}_u^T]$ could be approximated as follows*

$$\begin{aligned}\mathbb{E}[\mathbf{h}_u^*\mathbf{h}_u^T] \triangleq \mathbf{C}_u &= \frac{\delta_{u,0}\kappa_{u,0}}{1+\kappa_{u,0}}\bar{\mathbf{h}}_{u,0}^*\bar{\mathbf{h}}_{u,0}^T \\ &+ \left(\frac{\delta_{u,0}}{1+\kappa_{u,0}} + \frac{M\delta_1\delta_{u,2}}{1+\kappa_1}\right)\mathbf{I}_N \\ &+ \varrho_1\left(\bar{\mathbf{h}}_{u,0}^*\bar{\mathbf{h}}_{u,2}^T\boldsymbol{\Xi}\bar{\mathbf{H}}_1 + \bar{\mathbf{H}}_1^H\boldsymbol{\Xi}^*\bar{\mathbf{h}}_{u,2}^*\bar{\mathbf{h}}_{u,0}^T\right) \\ &+ \varrho_2\mathbf{a}_{BS}(\phi^{(BS)}, \psi^{(BS)})^H\mathbf{a}_{BS}(\phi^{(BS)}, \psi^{(BS)}) \\ &+ \varrho_3\bar{\mathbf{H}}_1^H\boldsymbol{\Xi}^*\bar{\mathbf{h}}_{u,2}^*\bar{\mathbf{h}}_{u,2}^T\boldsymbol{\Xi}\bar{\mathbf{H}}_1,\end{aligned} \tag{4}$$

with $\varrho_1 = \sqrt{\frac{\delta_1\delta_{u,0}\delta_{u,2}\kappa_1\kappa_{u,0}\kappa_{u,2}}{(1+\kappa_1)(1+\kappa_{u,0})(1+\kappa_{u,2})}}$, $\varrho_2 = \frac{M\delta_{u,2}\delta_1\kappa_1}{(1+\kappa_1)(1+\kappa_{u,2})}$ *and* $\varrho_3 = \frac{\delta_1\delta_{u,2}\kappa_1\kappa_{u,2}}{(1+\kappa_1)(1+\kappa_{u,2})}$.

Proof. See Appendix A. ∎

According to Theorem 1, $\mathbf{C}_u$ only depends on the channel's statistics. Also, the sum rate is dependent upon the active beamforming vector at the BS and the passive beamforming matrix at the RIS, therefore both active and passive beamformers should be designed jointly. Using Theorem 1, the problem $\mathcal{P}_1$ can be written as

$$\mathcal{P}_2 : \max_{\boldsymbol{\Xi},\mathbf{f}_u} \sum_{u=1}^{U}\log_2\left(1 + \frac{\frac{P}{U}\mathbf{f}_u^H\mathbf{C}_u\mathbf{f}_u}{\sigma_2 + \sum_{i\neq u}\frac{P}{U}\mathbf{f}_i^H\mathbf{C}_u\mathbf{f}_i}\right), \tag{5}$$

$$\text{subject to} \quad (2a), (2b). \tag{5a}$$

It can be seen in $\mathcal{P}_2$, the new optimization problem is a function of distance-dependent path-loss, Rician factors, and other statistical components of the channel, and the effect of small-scale fading has been averaged out. In the next section, a PPO algorithm is proposed to solve problem $\mathcal{P}_2$.

IV. Proximal Policy Optimization Approach

In this section, considering the complexity of the optimization problem $\mathcal{P}_2$, the PPO algorithm is adopted to solve the optimization problem. PPO has been shown to perform better than the other algorithms [12] which is a model-free, on-policy, actor-critic and policy gradient method. Consider an infinite-horizon discounted Markov decision process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the finite set of states, $\mathcal{A}$ is the finite set of actions, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition probability distribution, $r : \mathcal{S} \to \mathbb{R}$ is the reward function and finally, and $\gamma \in [0, 1]$ is the discount factor. Furthermore, $\hat{A}_t$ is an estimator of the advantage function at timestep t and is given by $A_t = Q(s_t, a_t) - V(s_t)$, where $Q(.,.)$ and $V(.)$ are the action-value and the value functions, respectively, and are defined as $Q(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \ldots}\left[\sum_{\ell=0}^{\infty} \gamma^{\ell} r(s_{t+1})\right]$ and $V(s_t) = \mathbb{E}_{a_t, s_{t+1}, \ldots}\left[\sum_{\ell=0}^{\infty} \gamma^{\ell} r(s_{t+1})\right]$ where $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$. The estimate of the advantage function in the interval $t \in [0, T]$ is given by $\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}$, with $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ and λ being a hyperparameter which denotes the factor for the trade-off of bias and variance for generalized advantage estimator (GAE) [12]. Then, let $\rho_t(\boldsymbol{\theta})$ denote the probability ratio $\rho_t(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t|s_t)}$. The PPO maximizes the following objective function

$$\mathcal{L}^{\text{CLIP}}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_t[\min(\rho_t(\boldsymbol{\theta})\hat{A}_t, \text{clip}\left(\rho_t(\boldsymbol{\theta}), 1 - \epsilon_c, 1 + \epsilon_c\right)\hat{A}_t)] \tag{6}$$

where ϵ_c is a hyperparameter. The second term in the $\min$ operator guarantees the probability ratio to be inside the interval $[1-\epsilon_c, 1+\epsilon_c]$ with the help of the $\text{clip}(.,.,.)$ function. The $\text{clip}(.,.,.)$ saturates the variable in the first input between the values of the second and third input. The details of the PPO algorithm can be found in [12].

Generally, PPO is presented as an MDP with observation and action spaces. When solving the joint active and passive beamforming problem, the BS, RIS, and all the UEs in the system are denoted by the environment $\mathcal{E}$, while the agent is BS which is able to control the RIS. The following are the key PPO elements that are employed to solve the joint active and passive beamforming problem.

1) Observation space: At each timestep t, the observation part consists of three parts. At first, it contains the real and imaginary parts of the beamforming vector $\mathbf{f}_u$, i.e., $\mathcal{F} = \{\mathcal{F}_{\text{r}}, \mathcal{F}_{\text{i}}\}$ with $\mathcal{F}_{\text{r}} = \{\text{Re}(\mathbf{f}_1), \ldots, \text{Re}(\mathbf{f}_U)\}$ and $\mathcal{F}_{\text{i}} = \{\text{Im}(\mathbf{f}_1), \ldots, \text{Im}(\mathbf{f}_U)\}$. The second part is the real and imaginary parts of the phase shifts of the RIS, i.e., $\mathcal{R} = \{\text{Re}(\text{diag}(\boldsymbol{\Theta})), \text{Im}(\text{diag}(\boldsymbol{\Theta}))\}$. The third part of the observation is $\mathcal{C} = \{\mathcal{C}_{\text{r}}, \mathcal{C}_{\text{i}}\}$, where $\mathcal{C}_{\text{r}} = \{\text{Re}(\text{vec}(\mathbf{C}_1)), \ldots, \text{Re}(\text{vec}(\mathbf{C}_U))\}$ and $\mathcal{C}_{\text{i}} = \{\text{Im}(\text{vec}(\mathbf{C}_1)), \ldots, \text{Im}(\text{vec}(\mathbf{C}_U))\}$. Finally, the observation vector at timestep t is $s_t = \{\mathcal{F}, \mathcal{R}, \mathcal{M}, \mathcal{N}\}$, and the observation shape is $2NU(U+1) + 2M$.

2) Action space: At each timestep t the action space is the vector containing the real and imaginary parts of the beamforming vectors for all the UEs and the real and imaginary parts of the phase shifts of the RIS. Thus, the action shape is $2UN + 2M$ and the action range is $[0, 2\pi]$.

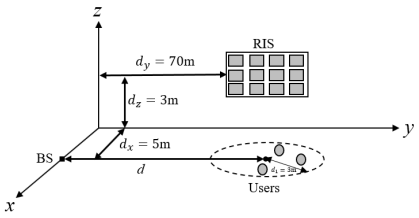

Fig. 2: Environment setup

3) Reward function: At timestep t, the reward is the sum rate of all the users, i.e., $r_t = \sum_{u=1}^{U} R_u(t)$ where $R_u(t)$ is the rate of u-th user at time step t. The details of the proposed PPO algorithm are presented in Algorithm 1.

Algorithm 1: PPO, Actor-Critic Style

Initialisation: Initialise time, states, actions, and replay buffer $\mathcal{D}$. ;
for *episode* $j = 1, \ldots, J$ **do**
  Initialise the environment $\mathcal{E}$ and make the initial state s_0;
  Run the policy $\pi_{\boldsymbol{\theta}_{\text{old}}}$ for T timesteps ;
  Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$;
  Optimise surrogate $\mathcal{L}^{\text{CLIP}}(\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$, with K epochs and minibatch size $M \leq T$ using (6);
  $\boldsymbol{\theta}_{\text{old}} \leftarrow \boldsymbol{\theta}$
end

V. Simulation Results

In this section, simulation results are provided to evaluate the performance of our proposed algorithm. In the simulation, the deployment of the BS and the RIS is shown in Fig. 2. Three users are randomly located at a circle with the radius of 3m. The large-scale path loss is given by $\delta_1 = 10^{-3} d_{\text{BR}}^{-2.5}$, $\delta_{u,2} = 10^{-3} d_{\text{RU}}^{-3}$ and $\delta_{u,0} = 10^{-3} d_{\text{BU}}^{-3}$, where d_{BR}, d_{RU} and d_{BU} are the distances between BS and RIS, RIS and the u-th UE and BS and the u-th UE, respectively. The number of BS antennas at the horizontal and vertical axis is 8 and 4, respectively. The RIS consists of 8 horizontal and 4 vertical elements. The noise power density is set to be -80dBm, and the maximum transmission power at the BS is 10dBm. The Rician factor between BS and UEs are set to be $\kappa_{u,0} = -3\text{dB}, u = 1, \ldots, U$ and that of between BS and RIS and RIS and UEs is $\kappa_1 = \kappa_{u,2} = 10\text{dB}, u = 1, \ldots, U$. A typical realization of the environment is illustrated in Fig. 2. At the beginning of each episode, a new environment is set as Fig. 2 with the fixed locations of the BS and RIS and randomly generated UE positions. The hyperparameters for the PPO algorithm are listed in Table I.

Fig. 3 shows the convergence of the PPO-based algorithm as a function of the number of episodes. The reward curve is obtained by the cumulative rewards obtained from each episode, i.e., $r'_t = \sum_{t=1}^{T'} r_t$, where T' is the total time steps in each episode. It can be seen from Fig. 3 that about 500 episodes are sufficient for the PPO to be converged.

TABLE I: Hyperparameter Values for PPO

Hyperparameter	Typical Value
Learning Rate (α)	0.0001
Epochs	100
Batch Size	32
Clip Parameter (ϵ_c)	0.2
GAE Lambda (λ)	0.95
Discount Factor (γ)	0.995
Exploration Noise Variance	0.001

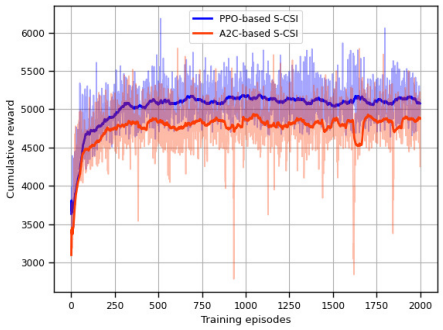

Fig. 3: Reward versus training episodes

Furthermore, the PPO-based approach outperforms the A2C-based method trained with the same reward function.

For comparing the transmit power, the agent is trained separately with different transmit powers ranging from 0dBm to 20dBm with steps of 5dBm. Furthermore, the Rician factors are set to be $\kappa_1 = \kappa_{u,2} = \kappa$ and $\kappa_{u,0} = 0$. The ADMM-based approach proposed in [7] is simulated for performance compression since [7] is one of the first attempts to design the active and passive beamforming based on the S-CSI information. Also, to compare to the I-CSI-based method, the performance of the BCD algorithm proposed in [2] is shown in Fig. 4. It can be seen that as SNR increases, the I-CSI-based algorithm outperforms the others, while the PPO-based S-CSI algorithm outperforms the A2C-based and ADMM-based algorithms. In addition, all algorithms perform significantly better than random phase shifts of RIS and the case in the absence of RIS.

In Fig. 5, the impact of the Rician factor on the average sum rate is investigated by assuming $\kappa_1 = \kappa_{u,2} = \kappa$ and $\kappa_{u,0} = 0$. It is revealed from Fig. 5 that the performance of all the algorithms with S-CSI and I-CSI improves when the Ricisn factor increases. For the S-CSI approach, in particular, this is because as κ increases, the BS-RIS-users link becomes more deterministic, which means the LoS link becomes more dominant. It is also observed that the gap between I-CSI and S-CSI cases eventually reaches a constant. This is because the direct link between the BS and UEs is assumed to be fully Rayleigh, and therefore no statistical information can be extracted to further improve performance in high Racian factors. It is also observed that the performance of the algorithm proposed in [2] is similar to the PPO-based approach when the I-CSI is available at the BS. Finally, in the random phase-shift case and in the case without RIS, the average sum rate is insensitive to the Rician factor.

In practice, the phase shifts of the RIS are quantized, so

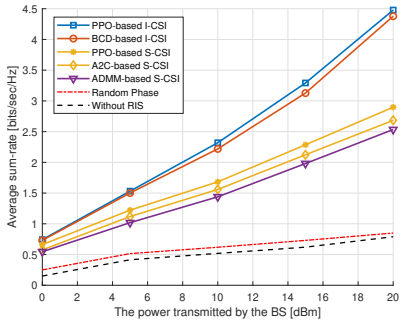

Fig. 4: Average sum-rate versus of BS transmit power

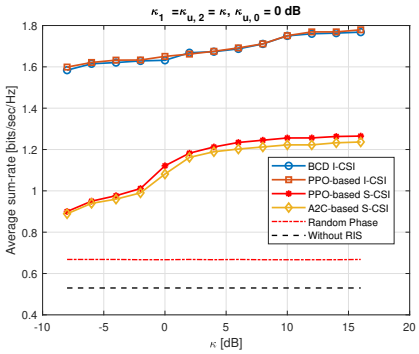

Fig. 5: Average sum-rate versus the Rician factor

the passive beamforming is done with discrete values rather than continuous values for the phase shifters. In particular, if the RIS is quantized with q bits, the values of the phase shifters could just take the values of the set $\mathcal{Q} = \{0, \frac{2\pi}{2^q}, \ldots, \frac{2\pi(2^q-1)}{2^q}\}$, where q is the number of quantization bits. In Fig. 6, the effect of the RIS-UE distance on the average sum rate of the users is shown. For finding the discrete values of the phase shifts of the RIS, we set $\hat{\xi}_i = g(\xi_m)$, where the function $g(\xi_m)$ maps the continuous phase shifts of the RIS, ξ_m, to its nearest point in $\mathcal{Q}$, that is

$$g(\xi_m) = \hat{\xi}_i, \text{ if } |\xi_m - \hat{\xi}_i| \leq |\xi_m - \hat{\xi}_j|, \forall \hat{\xi}_i, \hat{\xi}_j \in \mathcal{Q}, \forall i \neq j. \quad (7)$$

For training the agent, at the environment setup of Fig. 2, the number of users is reduced to one user and the single user is located at the center of the cell. The BS maximum power budget is set to be $P = 5\text{dBm}$ and $\kappa_1 = \kappa_{u,2} = 10\text{dB}$ and $\kappa_{u,0} = -3\text{dB}$ where $u = 1$. Fig. 5 shows how the performance of the RIS is significantly higher when the user is within close proximity of the RIS, even with 2-bit quantized RIS; The 2-bit quantized RIS, on the other hand, performs significantly worse than the 3-bit quantized and the ideal case, however, this gap can be fulfilled with 3-bits quantized RIS which has a close performance in comparison with the ideal case with continuous phase shifters.

TABLE II: The running time

Algorithm	Train time	Test time
PPO-based S-CSI	1.5 hours on GPU	2.45 seconds
ADMM-based S-CSI	-	168 seconds
BCD-based I-CSI	-	345 seconds

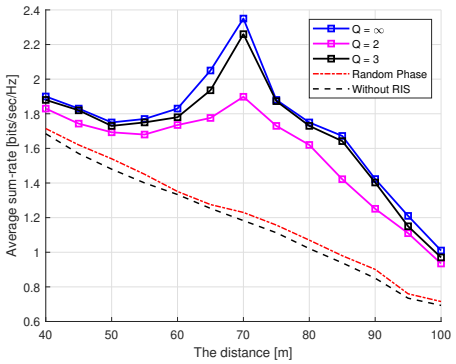

Fig. 6: Average sum-rate versus the RIS-UE distance (d)

Table II displays the runtime performance comparison of the PPO algorithm, accompanied by the ADMM and BCD algorithms. As depicted in II, the PPO algorithm trains once and exhibits rapid responsiveness in designing active and passive beamformers. In contrast, both ADMM and BCD algorithms necessitate a fresh run for each problem configuration, resulting in significantly longer execution times for each channel realization when compared to the efficiency of the PPO approach.

VI. Conclusion

In this letter, we present a PPO-based algorithm for joint active and passive beamforming in RIS-aided multiuser MISO systems. Leveraging S-CSI, we jointly optimized beamforming vectors at the BS and phase shifts at the RIS. Moreover, our proposed algorithm exhibits rapid convergence. Notably, 3-bit quantization of the RIS suffices to achieve performance parity with continuous phase shifts of the RIS. Additionally, the PPO method boasts significantly lower time complexity in comparison to iterative methods.

Appendix A

The calculation of $\mathbb{E}[\mathbf{h}_u^*\mathbf{h}_u^T]$, begins with finding $\mathbf{h}_u^*\mathbf{h}_u^T$ where $\mathbf{h}_u^T \triangleq \mathbf{h}_{u,0}^T + \mathbf{h}_{u,2}^T\boldsymbol{\Xi}\mathbf{H}_1$

$$\begin{aligned}\mathbf{h}_u^*\mathbf{h}_u^T = &\underbrace{\mathbf{h}_{u,0}^*\mathbf{h}_{u,0}^T}_{\mathbf{A}} + \underbrace{\mathbf{h}_{u,0}^*\mathbf{h}_{u,2}^T\boldsymbol{\Xi}\mathbf{H}_1}_{\mathbf{B}} + \underbrace{\mathbf{H}_1^H\boldsymbol{\Xi}^*\mathbf{h}_{u,2}^*\mathbf{h}_{u,0}^T}_{\mathbf{C}} \\ &+ \underbrace{\mathbf{H}_1^H\boldsymbol{\Xi}^*\mathbf{h}_{u,2}^*\mathbf{h}_{u,2}^T\boldsymbol{\Xi}\mathbf{H}_1}_{\mathbf{D}}. \end{aligned} \tag{8}$$

Since $\tilde{\mathbf{h}}_{u,2}$, $\tilde{\mathbf{h}}_{u,0}$ and $\tilde{\mathbf{H}}_1$ are independently distributed complex Gaussian random matrices with zero mean and unit variance, $\mathbb{E}[\tilde{\mathbf{h}}_{u,2}] = \mathbf{0}_{M\times 1}, \mathbb{E}[\tilde{\mathbf{h}}_{u,0}] = \mathbf{0}_{N\times 1}$ and $\mathbb{E}[\tilde{\mathbf{H}}_1] = \mathbf{0}_{M\times N}$. Next, $\mathbb{E}[\mathbf{A}]$ is derived as follows

$$\begin{aligned}\mathbb{E}[\mathbf{A}] = &\mathbb{E}\left[\frac{\delta_{u,0}\kappa_{u,0}}{1+\kappa_{u,0}}\bar{\mathbf{h}}_{u,0}^*\bar{\mathbf{h}}_{u,0}^T\right] + \mathbb{E}\left[\frac{\delta_{u,0}}{1+\kappa_{u,0}}\tilde{\mathbf{h}}_{u,0}^*\tilde{\mathbf{h}}_{u,0}^T\right] \\ &+ \mathbb{E}\left[\sqrt{\frac{\delta_{u,0}\kappa_{u,0}}{1+\kappa_{u,0}}}\sqrt{\frac{\delta_{u,0}}{1+\kappa_{u,0}}}\bar{\mathbf{h}}_{u,0}^*\tilde{\mathbf{h}}_{u,0}^T\right] \\ &+ \mathbb{E}\left[\sqrt{\frac{\delta_{u,0}\kappa_{u,0}}{1+\kappa_{u,0}}}\sqrt{\frac{\delta_{u,0}}{1+\kappa_{u,0}}}\tilde{\mathbf{h}}_{u,0}^*\bar{\mathbf{h}}_{u,0}^T\right]\end{aligned} \tag{9}$$

The second and the third term of (9) is zero, the first term is a constant and the last term is the definition of the covariance matrix, $\mathbb{E}[\mathbf{A}] = \frac{\delta_{u,0}\kappa_{u,0}}{1+\kappa_{u,0}}\bar{\mathbf{h}}_{u,0}^*\bar{\mathbf{h}}_{u,0}^T + \frac{\delta_{u,0}}{1+\kappa_{u,0}}\mathbf{I}_N$. By the same approach and noting that for a Gaussian distributed zero mean and unit variance matrix $\mathbf{X} \in \mathbb{C}^{P\times Q}$, $\mathbb{E}[\mathbf{X}^H\mathbf{Z}\mathbf{X}] = \mathrm{tr}(\mathbf{Z})\mathbf{I}_Q$ and with some simple calculations the remaining terms could be calculated and listed as follows

$$\mathbb{E}[\mathbf{B}] = \sqrt{\frac{\delta_1\delta_{u,0}\delta_{u,2}\kappa_1\kappa_{u,0}\kappa_{u,2}}{(1+\kappa_1)(1+\kappa_{u,0})(1+\kappa_{u,2})}}\bar{\mathbf{h}}_{u,0}^*\bar{\mathbf{h}}_{u,2}^T\boldsymbol{\Xi}\bar{\mathbf{H}}_1 \tag{10}$$

$$\mathbb{E}[\mathbf{C}] = \sqrt{\frac{\delta_1\delta_{u,0}\delta_{u,2}\kappa_1\kappa_{u,0}\kappa_{u,2}}{(1+\kappa_1)(1+\kappa_{u,0})(1+\kappa_{u,2})}}\bar{\mathbf{H}}_1^H\boldsymbol{\Xi}^*\bar{\mathbf{h}}_{u,2}^*\bar{\mathbf{h}}_{u,0}^T \tag{11}$$

$$\begin{aligned}\mathbb{E}[\mathbf{D}] = &\frac{M\delta_1\delta_{u,2}}{1+\kappa_1}\mathbf{I}_N + \frac{M\delta_{u,2}\delta_1\kappa_1}{(1+\kappa_1)(1+\kappa_{u,2})} \\ &\mathbf{a}_{\mathrm{BS}}(\phi^{(\mathrm{BS})},\psi^{(\mathrm{BS})})^H\mathbf{a}_{\mathrm{BS}}(\phi^{(\mathrm{BS})},\psi^{(\mathrm{BS})}) \\ &+ \frac{\delta_1\delta_{u,2}\kappa_1\kappa_{u,2}}{(1+\kappa_1)(1+\kappa_{u,2})}\bar{\mathbf{H}}_1^H\boldsymbol{\Xi}^*\bar{\mathbf{h}}_{u,2}^*\bar{\mathbf{h}}_{u,2}^T\boldsymbol{\Xi}\bar{\mathbf{H}}_1\end{aligned} \tag{12}$$

Finally, by summing up all the terms, the result will be obtained which completes the proof.

References

[1] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2019.

[2] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3064–3076, 2020.

[3] 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 Communications Letters*, vol. 12, no. 3, pp. 431–435, 2022.

[4] R. Liu, M. Li, Y. Liu, Q. Wu, and Q. Liu, "Joint transmit waveform and passive beamforming design for ris-aided dfrc systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 5, pp. 995–1010, 2022.

[5] H. Zhao, F. Wu, W. Xia, Y. Zhang, Y. Ni, and H. Zhu, "Joint beamforming design for ris-aided secure integrated sensing and communication systems," *IEEE Communications Letters*, pp. 1–1, 2023.

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

[7] X. Gan, C. Zhong, C. Huang, and Z. Zhang, "Ris-assisted multi-user miso communications exploiting statistical csi," *IEEE Transactions on Communications*, vol. 69, no. 10, pp. 6781–6792, 2021.

[8] K. Feng, Q. Wang, X. Li, and C.-K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for miso communication systems," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 745–749, 2020.

[9] M. Eskandari, K. Zhi, H. Zhu, C. Pan, and J. Wang, "Two-timescale design for ris-aided cell-free massive mimo systems with imperfect csi," *arXiv preprint arXiv:2304.02606*, 2023.

[10] M.-M. Zhao, Q. Wu, M.-J. Zhao, and R. Zhang, "Intelligent reflecting surface enhanced wireless networks: Two-timescale beamforming optimization," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 2–17, 2020.

[11] P. Liu, K. Luo, D. Chen, T. Jiang, and M. Matthaiou, "Spectral efficiency analysis of multi-cell massive mimo systems with ricean fading," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2018, pp. 1–7.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.