Deep Reinforcement Learning for Maximizing Downlink Spectral Efficiency in Non-Stationary RIS-Aided Multiuser-MISO Systems

Haoze Zhang[*], Xiang Huang[†], Zhangyu Guan[§],
Rong-Rong Chen[†], Arman Farhang[¶], and Mingyue Ji[*]
[*]*Electrical & Computer Eng. Dept., University of Florida*, Gainesville, FL, USA
[†]*Electrical & Computer Eng. Dept., University of Utah*, Salt Lake City, UT, USA
[§]*Computer Science & Eng. Dept., University of Minnesota - Twin Cities*, Minneapolis, MN, USA
[¶]*Electrical & Elect Eng. Dept., Trinity College Dublin*, Dublin, Ireland
Email: [*]{haoze.zhang, mingyueji}@ufl.edu, [†]{eric.xiang.huang, rchen}@utah.edu
Email: [§]zguan@umn.edu, [¶]arman.farhang@tcd.ie

***Abstract*—We investigate the problem of maximizing the overall Spectral Efficiency (SE) in a Reconfigurable Intelligent Surface (RIS)-aided Multi-User Multiple-Input Single-Output (MU-MISO) downlink system by jointly optimizing the beamforming at the Base Station (BS) and the phase shift of the RIS. To address this highly non-convex optimization challenge, we propose a Deep Reinforcement Learning (DRL) framework utilizing the Deep Deterministic Policy Gradient (DDPG) algorithm. The DRL agent interacts with the communication environment through trial-and-error learning, receiving the rewards that reflect the quality of actions under continuously changing states. One advantage of our proposed scheme is its capability to handle the non-stationary conditions of MU-MISO environment efficiently. This capability is achieved through a carefully designed, richly structured state representation, which captures the detailed information from both the current and previous time steps. Additionally, we introduce a dual-normalization network structure to promote stable learning and effective exploration during training. DRL agent is trained with an off-policy actor-critic method that leverages an experience replay buffer and soft-updated target networks to maintain stable convergence in the continuous action space. Simulation results under the 3GPP propagation environment demonstrate that our proposed scheme can achieve better SE performance compared with several state-of-the-art benchmarks.**

I. Introduction

Reconfigurable Intelligent Surface (RIS) / Intelligent Reflection Surface (IRS), has become a revolutionary technology for the next-generation wireless communication [1]–[3]. RIS typically consists of numerous passive reflecting units capable of manipulating the radio propagation environment by adjusting the phase shifts of incident electromagnetic waves. By intelligently steering these signals, RIS can significantly improve the sum Spectral Efficiency (SE) and mitigate interference among all users across various wireless communication scenarios [4]. Furthermore, due to the passive structure, its power consumption can be nearly ignored, and generates no additional thermal noise during signal reflection.

Despite these benefits, RIS technology faces several practical challenges [5], [6]. Among these challenges, one of the most critical applications is to maximize SE among users in MU-MISO downlink communication systems. The joint optimization of passive RIS phase shifts and active beamforming at the Base Station (BS) is a highly non-convex problem [7].

Several approaches have been proposed to address these challenges. The authors in [8] adopt a two-block Block-Coordinate Descent (BCD) framework, alternating between Weighted Minimum Mean Square Error (WMMSE) beamforming optimization and Riemannian Conjugate Gradient (RCG) phase shift updates. Although this method achieves strong performance, it incurs inner-iteration computational cost at each time step. Similarly, [10] introduces the Gradient-based Manifold Meta Learning method (GMML), which also requires extensive inner-loop learning at each time step.

In recent years, Deep Reinforcement Learning (DRL) has emerged as an effective technique for solving RIS-aided wireless problems. In [9], [11], [12], the authors explored DRL approaches, such as Deep Deterministic Policy Gradient (DDPG) and Soft Actor Critic (SAC) algorithms, demonstrating promising performance for MU-MISO systems. However, the limitation is that the performance in a non-stationary environment is not guaranteed. More specifically, the wireless channels change continuously from one transmission step to the next. This limitation highlights the need for a framework that can efficiently and reliably optimize system performance in a non-stationary environment.

Motivated by these limitations, we propose a new DRL-based framework leveraging the DDPG algorithm specifically designed to handle the non-stationary MU-MISO system effectively. The state representation includes detailed environmental information from both the current and previous time step, enabling the DRL agent to accurately capture temporal dynamics. We also introduce a dual-normalization network architecture combined with an adaptive exploration noise schedule to ensure stable training and efficient exploration during learning. We compare with several existing state-of-the-art benchmarks. Simulation results show that our scheme achieves comparable

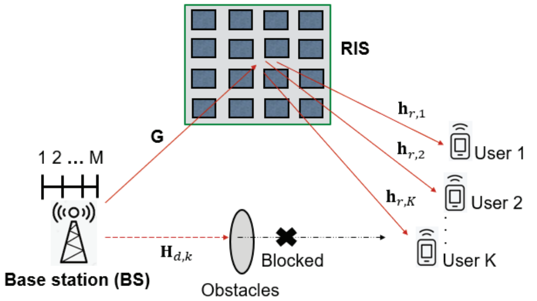


Fig. 1. The RIS MU-MISO downlink communication system.

or even better performance than the baseline methods.

II. System Model and Problem Formulation

We consider a RIS-aided MU-MISO downlink communication system consisting of a BS with M antennas, a RIS with N reflecting units, and K single-antenna user equipments (UEs), where $M \geq K$. As depicted in Fig. 1, the direct link $\mathbf{H}_{\mathrm{d},k}$ between the BS and the UEs is assumed to be blocked. At each time step t, the data stream is transmitted from the BS to the RIS, and reflected toward the UEs by dynamically adjusting the phase shifts according to the communication environment. Let $\mathbf{G} \in \mathbb{C}^{N\times M}$ denote the channel from BS to RIS, and $\mathbf{h}_{\mathrm{r},k} \in \mathbb{C}^{N\times 1}$ denote the channel between RIS to user k. The phase shift matrix is defined as $\mathbf{\Phi} = \mathrm{diag}([e^{j\phi_1}, e^{j\phi_2}, ..., e^{j\phi_N}]) \in \mathbb{C}^{N\times N}$ where ϕ_n corresponds to the phase shift applied by the n-th RIS element. We assume the data stream column vector $\mathbf{s} = (s_1, \ldots, s_K)$ is unit-power transmitted symbols, where $\mathbb{E}[|\mathbf{s}|^2] = 1$. The transmit beamforming matrix at BS is defined as $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{C}^{M\times K}$. The signal received by the k-th user is

$$\begin{aligned} y_k &= \mathbf{h}_{\mathrm{r},k}^H \mathbf{\Phi} \mathbf{G} \mathbf{W} \mathbf{s} + n_k, \\ &= \mathbf{h}_{\mathrm{r},k}^H \mathbf{\Phi} \mathbf{G} \mathbf{w}_k s_k + \sum_{n=1, n\neq k}^{K} \mathbf{h}_{\mathrm{r},k}^H \mathbf{\Phi} \mathbf{G} \mathbf{w}_n s_n + n_k, \end{aligned} \tag{1}$$

where $n_k \sim \mathcal{CN}(0, \sigma^2)$ is the complex Gaussian noise at user k, with zero mean and variance σ^2. $(\cdot)^H$ is the Hermitian operator. By combining the channels from BS to RIS and from RIS to the k-th user, the cascaded channel is given by $\mathbf{H}_k = \mathrm{diag}(\mathbf{h}_{\mathrm{r},k}^H)\mathbf{G}$. We define the phase shift vector $\boldsymbol{\phi} = [e^{j\phi_1}, e^{j\phi_2}, ..., e^{j\phi_N}]^H$ to efficiently represent the overall propagation environment, the received signal y_k becomes

$$y_k = (\boldsymbol{\phi}^H \mathbf{H}_k)\mathbf{w}_k s_k + \sum_{n=1, n\neq k}^{K} (\boldsymbol{\phi}^H \mathbf{H}_k)\mathbf{w}_n s_n + n_k. \tag{2}$$

In (2), the first term is treated as the received signal at user k, the second term is treated as multi-user interference to user k. The signal-to-interference noise ratio (SINR) is defined as

$$\mathbf{SINR}_k = \frac{|(\boldsymbol{\phi}^H \mathbf{H}_k)\mathbf{w}_k|^2}{\sum_{n=1, n\neq k}^{K} |(\boldsymbol{\phi}^H \mathbf{H}_k)\mathbf{w}_n|^2 + \sigma^2}. \tag{3}$$

In this paper, our objective is to maximize the overall SE of all users in the RIS-aided system. To achieve the objective, we jointly optimize two parameters, $\mathbf{W}$ and $\boldsymbol{\phi}$, subject to the transmit power P_t and phase shift constraints, i.e.,

$$\begin{aligned} \underset{\mathbf{W}, \boldsymbol{\phi}}{\mathrm{argmax}} \quad & \mathrm{R}(\mathbf{W}, \boldsymbol{\phi}) = \sum_{k=1}^{K} \log_2(1 + \mathbf{SINR}_k) \\ \textbf{s.t.} \quad & \mathrm{tr}\{\mathbf{W}^H \mathbf{W}\} \leq P_t \\ & |e^{j\phi_n}| = 1, \quad \forall n = 1, \ldots, N. \end{aligned} \tag{4}$$

We assume the setup where all channels are frequency-flat Rician fading with correlated, and all channel state information (CSI) is perfectly known. The block fading channel remains unchanged within a time step t. From BS to RIS channel $\mathbf{G}^{(t)}$ and RIS to UE channels $\mathbf{h}_{\mathrm{r},k}^{(t)}$ are modeled as Rician fading process [13], which is

$$\begin{aligned} \mathbf{G}^{(t)} &= \sqrt{\mathrm{PL}(d_{\mathbf{G}})} \left(\sqrt{\frac{\kappa}{\kappa+1}} \bar{\mathbf{G}} + \sqrt{\frac{1}{\kappa+1}} \tilde{\mathbf{G}}^{(t)} \right), \\ \mathbf{h}_{\mathrm{r},k}^{(t)} &= \sqrt{\mathrm{PL}(d_{\mathbf{h}_{\mathrm{r},k}})} \left(\sqrt{\frac{\kappa}{\kappa+1}} \bar{\mathbf{h}}_{\mathrm{r},k} + \sqrt{\frac{1}{\kappa+1}} \tilde{\mathbf{h}}_{\mathrm{r},k}^{(t)} \right), \end{aligned} \tag{5}$$

and the pathloss of BS to RIS and RIS to UE are denoted by $\mathrm{PL}(d_{\mathbf{G}})$ and $\mathrm{PL}(d_{\mathbf{h}_{\mathrm{r},k}})$, respectively. κ is the Rician factor for both channels $\mathbf{G}^{(t)}$ and $\mathbf{h}_{\mathrm{r},k}^{(t)}$. The Non-Line-of Sight (NLoS) random components $\tilde{\mathbf{G}}^{(t)}$ and $\tilde{\mathbf{h}}_{\mathrm{r},k}^{(t)}$ are formulated by the temporal correlated first-order complex Gauss-Markov block fading model [14], which represent how the channels evolve over time steps,

$$\begin{aligned} \tilde{\mathbf{G}}^{(t)} &= \sqrt{1-\rho^2}\, \mathbf{V}^{(t)} + \rho\, \tilde{\mathbf{G}}^{(t-1)}, \quad \mathbf{V}^{(t)} \sim \mathcal{CN}(0, 1), \\ \tilde{\mathbf{h}}_{\mathrm{r},k}^{(t)} &= \sqrt{1-\rho^2}\, \mathbf{U}_k^{(t)} + \rho\, \tilde{\mathbf{h}}_{\mathrm{r},k}^{(t-1)}, \quad \mathbf{U}_k^{(t)} \sim \mathcal{CN}(0, 1), \end{aligned} \tag{6}$$

where $\mathbf{V}^{(t)}$, $\mathbf{U}_k^{(t)}$ are channels randomly generated with i.i.d, and $\rho \in [0, 1]$ is the temporal correlation coefficient. Furthermore, we assume that the antennas at BS and RIS are arranged in uniform linear arrays with half-wavelength spacing. $\bar{\mathbf{G}}$ and $\bar{\mathbf{h}}_{\mathrm{r},k}$ are defined as

$$\begin{aligned} \bar{\mathbf{G}} &= \mathbf{a}_N(\theta_{\mathrm{ARIS}})\mathbf{a}_M(\theta_{\mathrm{DBS}})^H, \\ \bar{\mathbf{h}}_{\mathrm{r},k} &= \mathbf{a}_N(\theta_{\mathrm{DRIS}}), \end{aligned} \tag{7}$$

where $\mathbf{a}_N$ and $\mathbf{a}_M$ represent the steering vector at RIS and BS. θ_{ARIS} is the angle of arrival direction at RIS, θ_{DBS} is the angle of departure direction at BS, and θ_{DRIS} defines the angle at departure direction at RIS.

III. Proposed Approach

A. Overview of DRL

In RL, the agent learns a policy that maximizes cumulative reward by trial-and-error interaction with the environment. This model is formulated as Markov Decision Process (MDP), defined by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the state transition probability function, and $\mathcal{R}$ is the reward function. At time step t, the agent observes the state $s_t \in \mathcal{S}$, then chooses an action $a_t \sim \mu(\cdot|s_t) \in \mathcal{A}$ based on policy μ which satisfies $\sum_{a_t\in\mathcal{A}} \mu(s_t, a_t) = 1$. Next, the environment moves to the next state $s_{t+1} \sim P(\cdot|s_t, a_t) \in \mathcal{P}$

and the agent returns a scalar reward $r_{t+1} = \mathcal{R}(s_t, a_t, s_{t+1})$, measuring the quality of a_t. The cumulative reward is defined as $G_t = \sum_{i=1}^{\infty} \gamma^i r_{t+i}$, where i is how many steps ahead of the current time step t, and $\gamma \in (0,1]$ is the discount factor. The agent seeks a policy μ that maximizes $J(\mu) = \mathbb{E}_{s\sim p^{\mu}}[Q(s,\mu(s))]$ where Q is the Q-function, which satisfies $Q^{\mu}(s,a) = \mathbb{E}[G_t | s_t = s, a_t = a]$ as the expected return.

DRL is based on the RL algorithm, which exploits the representational advantage of Deep Neural Network (DNN) by using two networks, the actor (policy) and critic (Q-function). DDPG [15] is a DRL algorithm tailored for the continuous action space, where the policy μ is a deterministic mapping $a = \mu(s)$ from state s to action a. DDPG uses the actor-critic architecture, where the actor $\mu(s|\theta^{\mu})$ network is controlled by the parameter θ^{μ}, and the critic $Q(s,a|\theta^{Q})$ network is with the parameter θ^{Q}. To maintain the training stability and reduce the correlation between each experience, a replay buffer $\mathcal{D}$ is reserved to store past experiences. Define state as s_j, action as a_j, and reward as r_j. For every time step, the actor and critic networks are trained by randomly sampling a mini-batch N_{B} of experiences from $\mathcal{D}$, and performing Stochastic Gradient Descent (SGD) updates U times with U different sampling from $\mathcal{D}$. The mini-batch is defined as $\mathcal{B} = \{s_j, a_j, r_j, s'_j\}, j = 1, 2, ..., N_{\mathrm{B}}$, where s'_j is the next state of the environment transition from s_j. The critic network parameter θ^{Q} is updated by minimizing the loss

$$\mathcal{L}_{\mathrm{critic}}(\theta^{Q}) = \frac{1}{N_B} \sum_{j=1}^{N_B} \left(y_j - Q(s_j, a_j|\theta^{Q})\right)^2, \tag{8}$$

where the bootstrapped target y_j is defined as $y_j = r_j + \gamma Q'(s'_j, \mu'(s'_j, \theta^{\mu'})|\theta^{Q'})$. RL aims to approach the optimal Q value, but the actual target value is unknown. Thus, RL defines target actor and target critic networks as $Q'(\cdot|\theta^{Q'})$ and $\mu'(\cdot|\theta^{\mu'})$, which are the copies of the critic and actor network parameters that $\theta^{Q'} \leftarrow \theta^{Q}, \theta^{\mu'} \leftarrow \theta^{\mu}$. Hence, y_j is used as a proxy to approximate the actual target value. After the agent computes the critic loss, the critic network parameter θ^{Q} is updated by $\theta^{Q} \leftarrow \theta^{Q} - \eta^{Q} \nabla_{\theta^{Q}} \mathcal{L}_{\mathrm{critic}}(\theta^{Q})$, where η^{Q} is the learning rate of the critic network. Then, the actor parameter θ^{μ} is trained by minimizing the actor loss

$$\mathcal{L}_{\mathrm{actor}}(\theta^{\mu}) = -\frac{1}{N_B} \sum_{i=1}^{N_B} Q(s_j, \mu(s_j|\theta^{\mu})|\theta^{Q}), \tag{9}$$

which aims to maximize the Q-function. The agent updates the actor network parameter θ^{μ} by the actor loss that $\theta^{\mu} \leftarrow \theta^{\mu} + \eta^{\mu} \nabla_{\theta^{\mu}} \mathcal{L}_{\mathrm{actor}}(\theta^{\mu})$, where η^{μ} is the learning rate for the actor network. To enhance stability, the target network parameters $\theta^{\mu'}$ and $\theta^{Q'}$ are updated slower (i.e. the target network is updated once every N_{target} time steps) than the main networks by the soft-update method, where $\tau \in (0,1)$ is the soft-update coefficient with usually $\tau \ll 1$

$$\begin{aligned} \theta^{Q'} &\leftarrow \tau\,\theta^{Q} + (1-\tau)\,\theta^{Q'}, \\ \theta^{\mu'} &\leftarrow \tau\,\theta^{\mu} + (1-\tau)\,\theta^{\mu'}, \end{aligned} \tag{10}$$

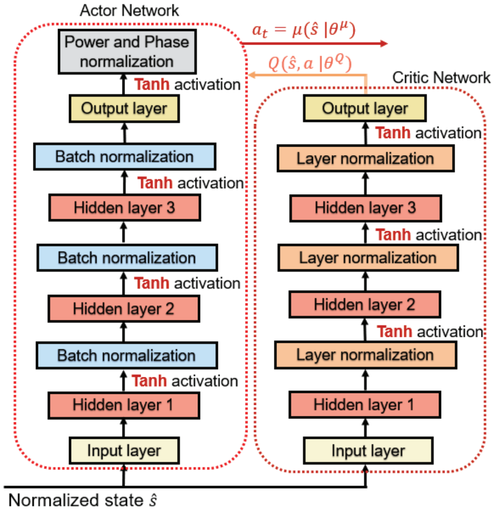


Fig. 2. Proposed DNN structure for both networks.

B. Proposed DDPG Scheme

At each time step t, the temporally structured state s_t is normalized to $\hat{s}_t = \frac{s_t - \bar{s}_t}{\sigma_s}$, where $\bar{s}_t$ and σ_s are the mean and standard deviation of s_t. The $\hat{s}_t$ is fed into the actor network to capture the feature dependencies and output action $a_t = \mu(\hat{s}_t|\theta^{\mu})$, where the policy parameter θ^{μ} is represented by a fully connected DNN. The critic network is parameterized by θ^{Q}, which is a fully connected DNN to estimate the Q value. The structure of DNNs is shown in Fig. 2. Several basic elements of the proposed DDPG are present in this section.

*1) **Dual-Normalization DNN**:* Both the actor and critic networks have three fully connected hidden layers with an input and output layer. We deliberately choose batch normalization (BN) for the actor and layer normalization (LN) for the critic, placing them on opposite sides of Tanh activation function. In the actor, we apply Tanh followed by BN, so the mini-batch N_{B} of experiences introduces slight fluctuations to each forward pass, which prevents the policy from converging too quickly and ensures exploration. In the critic, we perform LN before Tanh to normalize each batch sample independently. LN can effectively suppress internal covariate shift, stabilize gradient flow to produce more stable Q value estimation.

*2) **Action**:* At each time step t, the action is combined by the beamforming matrix $\mathbf{W}$ and RIS phase shift vector $\boldsymbol{\phi}$ as $a_t = \{\mathbf{W}^{(t)}, \boldsymbol{\phi}^{(t)}\}$. During training, a decaying random exploration noise $n_t = \{n_{\mathbf{W}}^{(t)}, n_{\boldsymbol{\phi}}^{(t)}\}$ is added to a_t to ensure sufficient exploration. Define $T_{\max}$ as the longest decaying step. For time step $t \leq T_{\max}$, the beamforming noise is $n_{\mathbf{W}}^{(t)} \sim \mathcal{N}(0, \sigma_{\mathbf{W}}^2(t))$, where $\sigma_{\mathbf{W}}(t) = \alpha_{\mathrm{init}} \frac{\sqrt{P_t}}{K} - \frac{t}{T_{\max}} \left(\alpha_{\mathrm{init}} \frac{\sqrt{P_t}}{K} - \alpha_{\mathrm{final}} \frac{\sqrt{P_t}}{K}\right)$ that K is the number of users. The α_{init} and α_{final} are the initial and final beamforming noise scaling factors. The noisy beamforming matrix $\mathbf{W}^{(t)} + n_{\mathbf{W}}^{(t)}$ is re-normalized to satisfy the transmit power P_t constraint in (4). Simultaneously, each RIS unit's phase shift add the angular noise, where $n_{\boldsymbol{\phi}}^{(t)} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\phi}}^2(t))$ with $\sigma_{\boldsymbol{\phi}}(t) = \beta_{\mathrm{init}} - \frac{t}{T_{\max}}(\beta_{\mathrm{init}} - \beta_{\mathrm{final}})$. Similarly, β_{init} and β_{final} are the

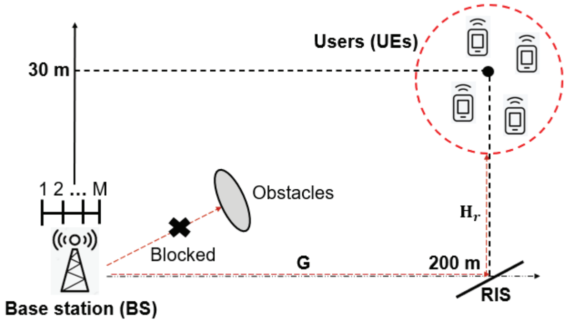


Fig. 3. Overview of MU-MISO scenario.

initial and final phase shift angular noise. The noisy RIS vector $\boldsymbol{\phi}^{(t)}+n_{\boldsymbol{\phi}}^{(t)}$ follows the phase shift constraint in (4). After $T_{\max}$ steps, a small fixed noise $\sigma_{\mathbf{W}} = \alpha_{\text{final}}\frac{\sqrt{P_t}}{K}$ and $\sigma_{\boldsymbol{\phi}} = \beta_{\text{final}}$ are applied to enable fine-tuned policy. This two-stage noise schedule encourages exploration during early learning and stable convergence at the end of training.

*3) **State**:* Since the channel varies in each time step t, the state vector has to include richly structured environmental features [16] with the information of past environmental conditions. The state at $t+1$ is defined as $s_{t+1} \triangleq o_1^{(t+1)} \cup o_2^{(t+1)}$

$$\begin{aligned} o_1^{(t+1)} &\triangleq \{\mathbf{W}^{(t)}, \boldsymbol{\phi}^{(t)}, P_k^{(t)}, \mathbf{H}_k^{(t)}, \mathbf{H}_k^{(t-1)}\}, \\ o_2^{(t+1)} &\triangleq \{P_{\mathrm{r},k}^{(t)}, I_k^{(t)}, \hat{I}_k^{(t)}, R_k^{(t)}, \frac{R_k^{(t)}}{\sum_{k=1}^{K} R_k^{(t)}}\}. \end{aligned} \tag{11}$$

The $o_1^{(t+1)}$ contains the transmission features where $\mathbf{W}^{(t)}$ and $\boldsymbol{\phi}^{(t)}$ are the beamforming and phase shift at step t. The BS's transmit power allocation of each user k at step t is $P_k^{(t)} = [P_1^{(t)}, P_2^{(t)}, \ldots, P_K^{(t)}]$. $\mathbf{H}_k^{(t)}$ and $\mathbf{H}_k^{(t-1)}$ represent the cascaded channels of each user k at step t and $t-1$. Note that it is assumed that the future environment information cannot be accessed at the current step. To predict the optimal action for the next step, the previous step channel observation is included in the state representation, enabling the network to capture the temporal channel pattern.

The $o_2^{(t+1)}$ contains information at the users' end. $P_{\mathrm{r},k}^{(t)}$ is the received signal power of user k at step t. Then $I_k^{(t)} = \sum_{n=1,n\neq k}^{K} |(\boldsymbol{\phi}^{(t)})^H \mathbf{H}_k^{(t)} \mathbf{w}_n^{(t)}|^2 + \sigma^2$ is the total received interference at k-th user in time step t. $\hat{I}_k^{(t)}$ is the total interference of k-th user at the beginning of time step t that beamforming and phase shift changed, but the cascaded channel have not changed yet, denoted as $\hat{I}_k^{(t)} = \sum_{n=1,n\neq k}^{K} |(\boldsymbol{\phi}^{(t)})^H \mathbf{H}_k^{(t-1)} \mathbf{w}_n^{(t)}|^2 + \sigma^2$. This observation provides the effect of action updates, independent of channel dynamics within a step. $R_k^{(t)}$ is the SE of each user at step t, and the last term $R_k^{(t)}/\sum_{k=1}^{K} R_k^{(t)}$, is defined as the ratio of each user's SE, that express the contribution of k-th user to overall SE among all users. Our state contains the previous time step channel to enable the agent to identify the short temporal correlation in the non-stationary environment.

*4) **Reward**:* At each time step t, the reward function is set as the instantaneous system SE, which exactly matches the optimization objective (4). However, since r_t scales with the transmit power, operating at low transmit power produces near-zero rewards, which vanish the gradients, and high transmit power leads to dispersed rewards that destabilize learning. To remove the reward scaling from the transmit power and preserve a consistent gradient magnitude, we perform the batch-level reward normalization. Given a mini-batch size of rewards $\{r_i\}_{i=1}^{B}$, we compute $\bar{r} = \frac{1}{B}\sum_{i=1}^{B} r_i$ and $\sigma_r = \sqrt{\frac{1}{B}\sum_{i=1}^{B}(r_i-\bar{r})^2}$, then replace each r_i by $\hat{r}_i = \frac{r_i-\bar{r}}{\sigma_r}$.

TABLE I
HYPERPARAMETERS FOR THE PROPOSED DDPG SCHEME

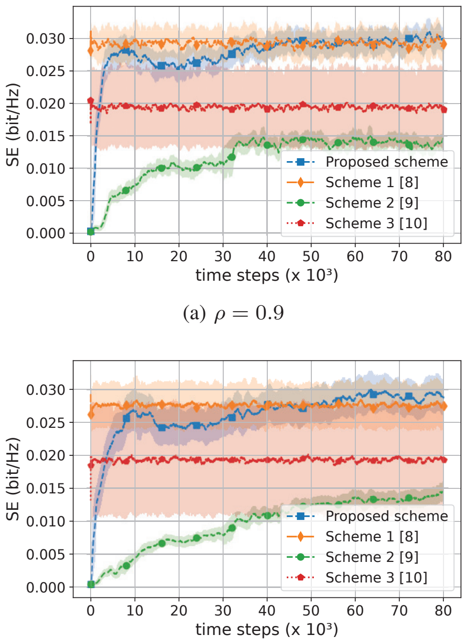
Parameter	Symbol	Value
Actor learning rate	η^{μ}	1.5×10^{-4}
Critic learning rate	η^{Q}	2×10^{-4}
Soft-update coefficient	τ	1.5×10^{-4}
Discount factor	γ	0.99
Batch size	N_{B}	64
Replay buffer size	$\lvert\mathcal{D}\rvert$	1×10^{5}
Number of total time steps	T	8×10^{4}
Number of steps updating target network	N_{target}	9
Beamforming noise scaling factor (init)	α_{init}	0.3
Beamforming noise scaling factor (final)	α_{final}	0.05
Phase-shift noise (init std)	β_{init}	0.15
Phase-shift noise (final std)	β_{final}	0.04
Decaying steps before fixed noise	$T_{\max}$	1×10^{4}
SGD updates per time step	U	5
Weight decay regularization		1×10^{-5}

IV. SIMULATION

A. Simulation setup

In this section, we show our proposed scheme's performance through simulation and compare it with other state-of-the-art benchmarks. We consider a RIS-aided MU-MISO system [17] as shown in Fig. 3. The BS is positioned at (0m, 0m), the RIS at (200m, 0m), and the UEs are randomly distributed within a circular area centered at (200m, 30m) with a radius of 10m. We assume that the direct link $\mathbf{H}_{d,k}$ between the BS and the users is blocked and thus unavailable. Moreover, we set the number of antennas at BS equal to 4, and each of the 4 users has a single-antenna device. The RIS comprises 16 units, and each unit can independently adjust its phase shift. The pathloss component is based on the 3GPP propagation environment [18]. The pathloss of NLoS part is defined as $\mathrm{PL}(d) = 35.6 + 22.0\log_{10}(d)$, where d is the transmission distance. The system is operated at a bandwidth of $B_0 = 180$ kHz with the thermal noise power of $-170 + 10\log_{10}(B_0) \approx -117.45$ dBm, and the Rician factor is set to $\kappa = 10$. The layer dimensions for the actor and critic networks in Fig. 2 are [1112, 2048, 1024, 512, 64] and [1112, 2112, 2048, 1024, 1], respectively. The simulations are implemented in PyTorch and trained using the Adam optimizer. The remaining hyperparameter settings are summarized in Table I. The proposed scheme is evaluated against three benchmark schemes for beamforming and phase-shift optimization.

- Scheme 1: WMMSE followed by RCG algorithm [8].
- Scheme 2: Baseline DDPG [9].
- Scheme 3: Gradient-based manifold meta learning [10].



(a) $\rho = 0.9$

(b) $\rho = 0.5$

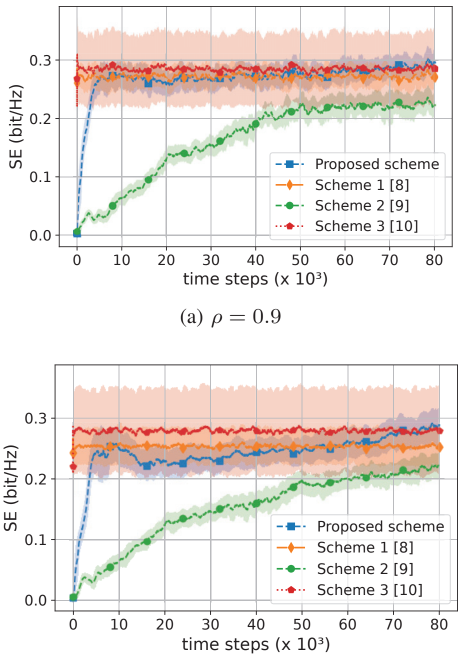
Fig. 4. Different ρ comparison at transmit power = -10dBm.

Schemes 1 and 3 require the inner computation or learning iterations to obtain the optimal $\mathbf{W}$ and ϕ. For all the schemes, we assume that future CSI cannot be accessed at the current step. The beamforming and phase shift inferred from the current step are applied to the environment of the next step to calculate the spectral efficiency.

The simulation is evaluated across 5 transmit power levels $P_t \in \{-10, 0, 10, 20, 30\}$ dBm and 2 temporal channel correlation coefficients ρ, which are set to 0.9 and 0.5 to represent strong and moderate correlation, respectively. Given a (P_t, ρ) pair, within each simulation, all schemes are evaluated on the same set of UE positions and channels. We conduct 5 independent simulations for all schemes. Across different simulations, the UE locations and channels are randomly regenerated. When comparing different ρ cases with a given P_t, UE locations remain fixed while the channels are regenerated. The simulation results are averaged over 5 simulations.

B. Results

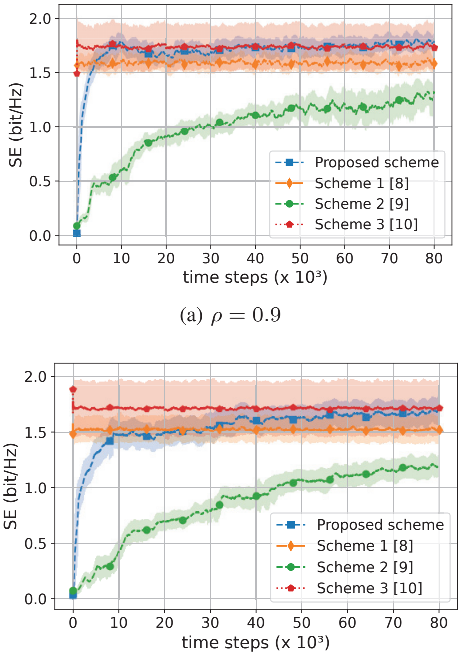
Figs. 4-8 illustrate the spectral efficiency (SE) performance comparison among different schemes, with SE on the y-axis and time step index on the x-axis. The shaded regions in the figures represent the variance around the mean curve. The performance at each time step is computed as a moving average over the previous 1,000 steps. As a result, (i.e., Fig. 4), Scheme 3 initially appears below the average since there is no prior data for smoothing. Our proposed scheme demonstrates significantly faster convergence than Scheme 2, particularly within the first 10,000 steps, which guarantees sufficient exploration during the early stages of training and achieves better performance. When $\rho = 0.9$, the proposed scheme consistently achieves the best performance over all the benchmarks across all transmit power levels. When $\rho = 0.5$, the proposed scheme generally maintains superior performance over the benchmarks, except for cases with $P_t = 10$ and $P_t = 30$ dBm. As shown in Fig. 6, when $P_t = 10$ dBm, our scheme achieves very close performance to Scheme 3, and better than others. Fig. 8 indicates that, at $P_t = 30$ dBm, the proposed scheme outperforms Scheme 3 and is slightly worse than Scheme 1. These results demonstrate the adaptability and robustness of our proposed scheme across various levels of temporal channel correlation and transmit power.



(a) $\rho = 0.9$

(b) $\rho = 0.5$

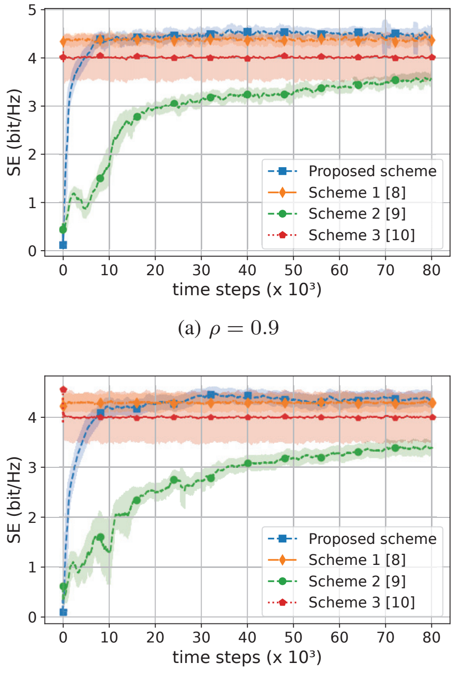
Fig. 5. Different ρ comparison at transmit power = 0dBm.



(a) $\rho = 0.9$

(b) $\rho = 0.5$

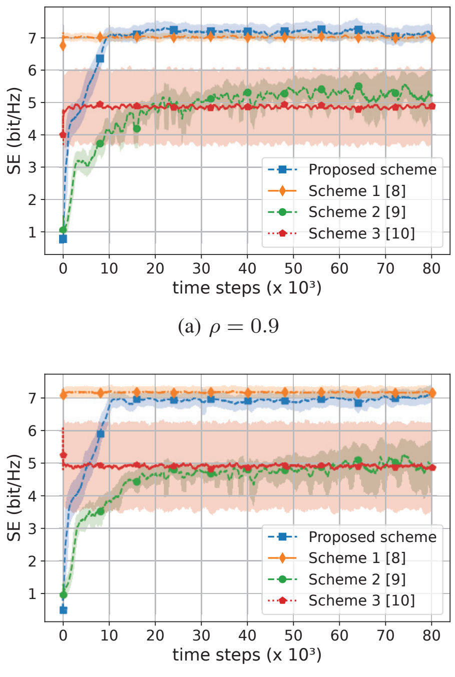
Fig. 6. Different ρ comparison at transmit power = 10dBm.



(a) $\rho = 0.9$

(b) $\rho = 0.5$

Fig. 7. Different ρ comparison at transmit power = 20dBm.



(a) $\rho = 0.9$

(b) $\rho = 0.5$

Fig. 8. Different ρ comparison at transmit power = 30dBm.

V. Conclusion

This work presents a jointly beamforming and phase shift optimization based on Deep Deterministic Policy Gradient in the non-stationary MU-MISO downlink environment. The proposed scheme utilizes the Dual-normalization network structure, a carefully designed state that embeds the detailed environmental features of both the current and previous time steps and the transmit power-related exploration noise. The simulation results indicate that our proposed DDPG scheme outperforms several state-of-the-art approaches in most cases across a wide range of transmit powers.

Acknowledgment

This work was supported in part by the U.S. National Science Foundation (NSF) under Awards 2229563, 2450418, 2549413, 2229562, 2153875, 2450419, and 2516634.

References

[1] Q. Wu and R. Zhang, ”Towards Smart and Reconfigurable Environment: Intelligent Reflecting Surface Aided Wireless Network,” in IEEE Communications Magazine, vol. 58, no. 1, pp. 106-112, January 2020.

[2] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, “Channel Estimation for RIS-Empowered Multi-User MISO Wireless Communications,” IEEE Trans. Commun., vol. 69, no. 6, pp. 4144–4157, Jun. 2021.

[3] W. Tang et al., “Wireless communications with programmable metasurface: New paradigms, opportunities, and challenges on transceiver design,” 2019.

[4] Reconfigurable Intelligent Surfaces: Principles and Opportunities,” IEEE Commun. Surv. Tutor., vol. 23, no. 3, pp. 1546–1577, 2021.

[5] Z. Chen, J. Tang, X. Y. Zhang, Q. Wu, G. Chen, and K.- K. Wong, “Robust Hybrid Beamforming Design for Multi-RIS Assisted MIMO System With Imperfect CSI,” IEEE Trans. Wireless Commun., vol. 22, no. 6, pp. 3913–3926, Jun. 2023.

[6] M. Gao, J. Yang, H. Li, and Y. Wang, “Robust Beamform- ing Optimization Design for RIS-Aided MIMO Systems With Practical Phase Shift Model and Imperfect CSI,” IEEE Internet Things J., vol. 11, no. 1, pp. 958–973, Jan. 2024.

[7] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, “Intelligent Reflecting Surface: Practical Phase Shift Model and Beamforming Optimization,” IEEE Trans. Commun., vol. 68, no. 9, pp. 5849–5863, Sept. 2020.

[8] Guo, H., Liang, Y., Chen, J., & Larsson, E. G. (2019). Weighted Sum-Rate Maximization for Reconfigurable Intelligent Surface Aided Wireless Networks.

[9] C. Huang, R. Mo and C. Yuen, ”Reconfigurable Intelligent Surface Assisted Multiuser MISO Systems Exploiting Deep Reinforcement Learning,” in IEEE Journal on Selected Areas in Communications, vol. 38, no. 8, pp. 1839-1850, Aug. 2020.

[10] Zhu, F., Wang, X., Huang, C., Yang, Z., Chen, X., Alhammadi, A., Zhang, Z., Yuen, C., & Debbah, M. (2024). Robust Beamforming for RIS-aided Communications: Gradient-based Manifold Meta Learning.

[11] K. Feng, Q. Wang, X. Li and C. -K. Wen, ”Deep Reinforcement Learning Based Intelligent Reflecting Surface Optimization for MISO Communication Systems,” in IEEE Wireless Communications Letters, vol. 9, no. 5, pp. 745-749, May 2020.

[12] Saglam, B., Gurgunoglu, D., & Kozat, S. S. (2022). Deep Reinforcement Learning Based Joint Downlink Beamforming and RIS Configuration in RIS-aided MU-MISO Systems Under Hardware Impairments and Imperfect CSI.

[13] Y. Han, W. Tang, S. Jin, C. Wen, and X. Ma, “Large intelligent surface assisted wireless communication exploiting statistical CSI,” IEEE Trans. Veh. Technol., vol. 68, no. 8, pp. 8238–8242, Jun. 2019.

[14] Y. S. Nasir and D. Guo, ”Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks,” in IEEE Journal on Selected Areas in Communications, vol. 37, no. 10, pp. 2239-2250, Oct. 2019.

[15] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,”

[16] X. Zhang, A. Bhuyan, S. K. Kasera and M. Ji, ”Distributed Power Allocation for 6-GHz Unlicensed Spectrum Sharing via Multi-agent Deep Reinforcement Learning,” 2023 IEEE International Conference on Industrial Technology (ICIT), Orlando, FL, USA, 2023, pp. 1-6.

[17] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, “Reconfigurable intelligent surfaces for energy efficiency in wireless communication,” IEEE Trans. Wireless Commun., vol. 18, no. 8, pp. 4157–4170, Jun. 2019.

[18] 3GPP, “Further Advancements for E-UTRA Physical Layer Aspects (Release 9),” TS 36.814 V9.2.0, Mar. 2010.