Deep Reinforcement Learning Based Joint Downlink Beamforming and RIS Configuration in RIS-aided MU-MISO Systems Under Hardware Impairments and Imperfect CSI

Baturay Saglam*, *Student Member, IEEE*, Doga Gurgunoglu†, *Student Member, IEEE,*
Suleyman S. Kozat*, *Senior Member, IEEE*
*Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey
†Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm 100 44, Sweden

***Abstract*—We introduce a novel deep reinforcement learning (DRL) approach to jointly optimize transmit beamforming and reconfigurable intelligent surface (RIS) phase shifts in a multiuser multiple input single output (MU-MISO) system to maximize the sum downlink rate under the phase-dependent reflection amplitude model. Our approach addresses the challenge of imperfect channel state information (CSI) and hardware impairments by considering a practical RIS amplitude model. We compare the performance of our approach against a vanilla DRL agent in two scenarios: perfect CSI and phase-dependent RIS amplitudes, and mismatched CSI and ideal RIS reflections. The results demonstrate that the proposed framework significantly outperforms the vanilla DRL agent under mismatch and approaches the golden standard. Our contributions include modifications to the DRL approach to address the joint design of transmit beamforming and phase shifts and the phase-dependent amplitude model. To the best of our knowledge, our method is the first DRL-based approach for the phase-dependent reflection amplitude model in RIS-aided MU-MISO systems. Our findings in this study highlight the potential of our approach as a promising solution to overcome hardware impairments in RIS-aided wireless communication systems.**

***Index Terms*—reconfigurable intelligent surface, sum rate, multiuser multiple input single output, hardware impairment, phase-dependent amplitude, deep reinforcement learning**

I. Introduction

RIS is among the emerging technologies explored for next-generation wireless communication systems [1]. An RIS consists of multiple reflecting elements with sub-wavelength spacing whose impedances are adjusted to induce desired phase shifts on incident waves before they are reflected. This enables the manipulation of multipath interference at the receiver [2]. However, depending on the RIS hardware, the incident wave is attenuated depending on the applied phase shifts to the individual elements, resulting in *phase-dependent reflection amplitudes* [1]. Such phenomenon causes significant performance losses [2].

The non-linear model in [1] renders the already complex optimization-based approaches impractical [3]. An alternative to such methods, deep reinforcement learning (DRL), has become a widely-studied machine learning (ML) approach for RIS-aided wireless systems such as non-orthogonal multiple access (NOMA) downlink systems [4], millimeter wave communications [5], vehicular communications and trajectory optimization [6]–[8], and the transmit beamforming and phase shifts design [6], [9]–[15]. In [4], RIS phase shifts are adjusted using the Deep Deterministic Policy Gradient (DDPG) algorithm [16]. In [5], the joint design of the downlink beamforming matrix and the RIS phase shifts are considered under ideal RIS reflections. In [15], the aforementioned optimization is performed under individual users' signal-to-interference-plus-noise ratio (SINR) constraints, as opposed to maximizing the downlink sum rate, which is susceptible to maximizing the sum rate by significantly lowering certain users' individual rates. In [17], a Deeq Q-Network (DQN)-based framework is proposed to maximize the spectral efficiency (SE) of the downlink of an orthogonal frequency division multiplexing (OFDM) communication system with a low-resolution RIS. While the prior applications of DRL to RIS-aided systems assumed ideal reflections and perfect CSI, a DRL application considering RIS hardware impairments does not exist to the best of our knowledge.

In this paper, we study the joint design of transmit beamforming and phase shifts for RIS-aided multi-user multiple input single output (MU-MISO) systems through a DRL approach. Our objective is to maximize the sum downlink rate of the users under the phase-dependent amplitude model. Since phase-dependent amplitudes make the system more complex, we only consider DRL-based approaches in our study. The main contributions of this study can be summarized as follows:

- We present novel modifications for the application of DRL to RIS-aided systems, which address two critical aspects of *non-episodic* tasks, e.g., the joint design of transmit beamforming and phase shifts, which have been passed over the existing works.
- To the best of our knowledge, we devise the first DRL-based approach for the phase-dependent reflection amplitude model in RIS-aided MU-MISO systems to provide

This study is supported by Turk Telekom within the framework of the 5G and Beyond Joint Graduate Support Programme coordinated by Information and Communication Technologies Authority and the EU Horizon 2020 MSCA-ITN-METAWIRELESS, Grant Agreement 956256.

an alternative ML-based framework for the suboptimal iterative algorithms proposed in [1].
- Although the agent is unaware of the phase-dependent reflections and the presence of channel estimation error, the presented method achieves sum rates close to the existing DRL agent operating with perfect CSI and full awareness of phase-dependent RIS amplitudes.
- To ensure reproducibility and support further research on DRL-based RIS systems, we provide our source code and results in the GitHub repository[1].

II. System Model

We consider the downlink of a narrow-band RIS-aided MU-MISO system consisting of K single-antenna users, M base station (BS) antennas, and L RIS elements. The transmit beamforming matrix $\mathbf{G} \in \mathbb{C}^{M\times K}$ maps K data streams denoted by $\mathbf{x} \in \mathbb{C}^{K\times 1}$ for K users onto M transmit antennas. $\mathbf{H} \in \mathbb{C}^{L\times M}$, $\mathbf{\Phi} \triangleq \text{diag}(\phi_1, \ldots, \phi_L) \in \mathbb{C}^{L\times L}$, and $\mathbf{h}_k \in \mathbb{C}^{L\times 1}$ denote the base station (BS)-RIS channel, the diagonal reflection matrix at the RIS, and RIS-user k channel respectively. In the following subsections, we explain two different environment models: *true environment* with phase-dependent RIS amplitude and perfect CSI, and the *mismatch environment* with ideal reflection assumption and imperfect CSI.

A. True Environment Model

The received signal at user k can be expressed as:

$$z_k = \mathbf{h}_k^\top \mathbf{\Phi}\mathbf{H}\mathbf{G}\mathbf{x} + w_k, \tag{1}$$

where the complex scalars z_k and w_k denote the received signal and the additive receiver noise at the k'th user, respectively, and we assume that $w_k \sim \mathcal{CN}(0, \sigma_w^2)$ for all k. The RIS follows the phase-dependent amplitude model in [1], with entries $\phi_l = \beta(\varphi_l)e^{j\varphi_l}$ for $\varphi \in [0, 2\pi)$, resulting in:

$$\beta(\varphi_l) = (1 - \beta_{\min})\left(\frac{\sin(\varphi_l - \mu) + 1}{2}\right)^{\kappa} + \beta_{\min}, \tag{2}$$

where $\beta_{\min} \in [0, 1]$, $\mu \geq 0$, and $\kappa \geq 0$ are constants that depend on the hardware implementation of the RIS. In the golden standard scenario, the BS knows the individual cascaded channels to each user, denoted by:

$$\mathbf{D}_k \triangleq \text{diag}(\mathbf{h}_k)\mathbf{H} \in \mathbb{C}^{L\times M}, \quad \forall k = 1, \ldots, K. \tag{3}$$

Hence (1) can be rewritten as:

$$z_k = \boldsymbol{\phi}^\top \mathbf{D}_k \mathbf{G}\mathbf{x} + w_k, \tag{4}$$

where $\boldsymbol{\phi} \in \mathbb{C}^{L\times 1}$ denotes the column vector consisting of the diagonal entries of $\mathbf{\Phi}$.

[1]https://github.com/baturaysaglam/RIS-MISO-PDA-Deep-Reinforcement-Learning

B. Mismatch Environment Model

In this simplified model, the RIS reflections are assumed to be lossless, i.e., $\hat{\boldsymbol{\phi}} \triangleq [e^{j\varphi_1}, \ldots, e^{j\varphi_L}]^\top$. Moreover, the agent has access to only an imperfect estimate of the cascaded channels, namely:

$$\hat{\mathbf{D}}_k \triangleq \mathbf{D}_k + \mathbf{E}_k, \quad \forall k = 1 \ldots, K, \tag{5}$$

where $\mathbf{E}_k \in \mathbb{C}^{L\times M}$ denotes the channel estimation error matrix of the cascaded channel of each user, with independent and identically distributed (i.i.d.) entries $e_{l,m}^{(k)} \sim \mathcal{CN}(0, \sigma_e^2)$.

C. Problem Formulation

1) The Golden Standard Objective: Our emphasis is to utilize the DRL agent to maximize the sum downlink rate in the system, which is denoted as:

$$R_\Sigma \triangleq \sum_{k=1}^{K} \log\left(1 + \frac{\|\boldsymbol{\phi}^\top \mathbf{D}_k \mathbf{G}\|^2}{\sum_{j\neq k} \|\boldsymbol{\phi}^\top \mathbf{D}_j \mathbf{G}\|^2 + \sigma_w^2}\right). \tag{6}$$

The BS aims to maximize (6) by adjusting $\mathbf{G}$ and $\boldsymbol{\phi}$. Under the transmission power constraint P_t and the domain restriction of phase shifts, the optimization problem is expressed as:

$$\begin{aligned} \underset{\boldsymbol{\phi},\mathbf{G}}{\text{maximize}} \quad & R_\Sigma \\ \text{subject to} \quad & \varphi_l \in [0, 2\pi), \quad \forall\, l = 1, \ldots, L, \\ & \text{tr}(\mathbf{G}\mathbf{G}^H) \leq P_t. \end{aligned} \tag{7}$$

where $\boldsymbol{\phi}$ depends on $\varphi_1, \ldots, \varphi_L$ and $\beta(\varphi_l)$ according to (2) when the BS agent is aware of the true environment model.

2) The Mismatch Objective: The optimization problem to be solved in the mismatch scenario is defined as:

$$\hat{R}_\Sigma \triangleq \sum_{k=1}^{K} \log\left(1 + \frac{\|\hat{\boldsymbol{\phi}}^\top \hat{\mathbf{D}}_k \mathbf{G}\|^2}{\sum_{j\neq k} \|\hat{\boldsymbol{\phi}}^\top \hat{\mathbf{D}}_j \mathbf{G}\|^2 + \sigma_w^2}\right). \tag{8}$$

Consequently, the BS agent considers the following optimization problem:

$$\begin{aligned} \underset{\hat{\boldsymbol{\phi}},\mathbf{G}}{\text{maximize}} \quad & \hat{R}_\Sigma \\ \text{subject to} \quad & \varphi_l \in [0, 2\pi), \quad \forall\, l = 1, \ldots, L, \\ & \text{tr}(\mathbf{G}\mathbf{G}^H) \leq P_t. \end{aligned} \tag{9}$$

The objective in (8) uses ideal RIS amplitudes and noisy channel estimates instead of the phase-dependent amplitude and the true channel in (6). Hence, they have different forms in terms of the transmit beamformer and the RIS phase shifts. Consequently, the seemingly similar (7) and (9) have different solutions. In other words, the BS is trying to solve a different optimization problem than the actual sum rate in the environment, resulting in inferior transmit beamforming and RIS configuration designs. In Section III, we propose a DRL framework that overcomes this phenomenon. Section III.

III. The Deep Reinforcement Learning Framework

A. Overview

At each discrete time step t, the agent observes a state $s \in \mathcal{S}$ and takes an action $a \in \mathcal{A}$, and observes a next state $s' \in \mathcal{S}$ and receives a reward r, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively. In fully observable environments, the reinforcement learning (RL) problem is usually represented by a Markov decision process, a tuple $(\mathcal{S}, \mathcal{A}, P, \gamma)$, where P is the transition dynamics such that $s', r \sim P(s, a)$ and $\gamma \in [0, 1]$ is a constant discount factor.

The objective in RL is to find an optimal policy π that maximizes the *value* defined by $V_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i+1}$, where the discount factor γ prioritizes the short-term rewards. The policy of an agent is regarded as stochastic if it maps states to action probabilities $\pi : \mathcal{S} \to p(\mathcal{A})$, or deterministic if it maps states to unique actions $\pi : \mathcal{S} \to \mathcal{A}$. The performance of a policy is assessed under the action-value function (Q-function or critic) that represents V_t while following the policy π after acting a in state s: $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s, a_0 = a]$. The Q-function is learned through the Bellman equation [18]:

$$Q^\pi(s, a) = \mathbb{E}_{r, s' \sim P, a' \sim \pi}[r + \gamma Q^\pi(s', a')], \tag{10}$$

where a' is the action selected by the policy on the observed next state s'.

In deep RL, the critic is approximated by a deep neural network Q_θ with parameters θ, i.e., the Deep Q-learning algorithm [19]. Given a transition tuple (s, a, r, s'), the Q-network is trained by minimizing a loss $J(\theta)$ on the temporal-difference (TD) error δ corresponding to Q_θ [20], the difference between the output of Q_θ and learning target y:

$$y \triangleq r + \gamma Q_{\theta'}(s', a'), \tag{11}$$

$$\delta \triangleq y - Q_\theta(s, a); \tag{12}$$

$$\theta \leftarrow \theta - \eta \nabla_\theta J(\theta), \tag{13}$$

where $J(\theta) = |\delta|^2$, $\nabla_\theta J(\theta)$ is the gradient of the loss $J(\theta)$ with respect to θ, and η is the learning rate. The target y in (11) utilizes a separate target network with parameters θ' that maintains stability and fixed objective in learning the optimal Q-function [19]. The target parameters are updated to copy the parameters θ after a number of learning steps.

B. The Soft Actor-Critic Algorithm

We use the state-of-the-art Soft Actor-Critic (SAC) algorithm [21] in our work, outperforming prior suboptimal DRL algorithms [4], [9], [10], [13] in most DRL benchmarks. SAC is an actor-critic, off-policy algorithm that operates in continuous action spaces. It uses a separate actor network to choose actions and stores experiences in the experience replay memory [22]. Unlike on-policy algorithms, SAC samples transitions from the replay memory for training. Our initial simulations showed that SAC was the only actor-critic algorithm that could converge for the problem of interest despite intensive hyperparameter tuning.

A SAC agent maintains three networks: two Q-networks and a single stochastic policy network (or actor network), each being a multi-layer perceptron (MLP). Using two Q-networks is to reduce the overestimation of Q-value estimates [23]. The Q-networks take the states provided by the environment and actions produced by the actor network as inputs and produce Q-value estimates, which are scalar values. Given the actor network π_ψ parameterized by ψ, the Q-networks are jointly trained in the SAC algorithm as:

$$\hat{\mathbf{y}} \triangleq \mathbf{r} + \gamma \min_{i=1,2} Q_{\theta_i'}(\mathbf{s}', \mathbf{a}')|_{\mathbf{a}' \sim \pi_\psi(\cdot|\mathbf{s}')} - \alpha \log(\mathbf{a}'|\mathbf{s}'), \tag{14}$$

$$J(\theta_i) = \frac{1}{N} \|\hat{\mathbf{y}} - Q_{\theta_i}(\mathbf{s}, \mathbf{a})\|_2^2, \tag{15}$$

$$\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} J(\theta_i), \tag{16}$$

where $(\mathbf{s}, \mathbf{a}, \mathbf{r}, \mathbf{s}')_{i=1}^N$ is the mini-batch of transitions sampled from the experience replay buffer and N is the mini-batch size, θ_i are the parameters corresponding to the ith Q-network, α is the entropy regularization term, and $\|\cdot\|_2$ is the L_2 norm. Note that we denote state and action vectors by s and a, respectively, while $\mathbf{s}$ and $\mathbf{a}$ represent the batch of state and action vectors.

Similarly, the policy network takes state vectors from the environment as inputs and produces numerical action vectors. The loss for the policy network in the SAC algorithm is expressed by:

$$J(\psi) = \frac{1}{N} \sum_i^N \alpha \log \pi_\psi(\hat{\mathbf{a}}_i|\mathbf{s}_i) - \min_{j=1,2} Q_{\theta_j}(\mathbf{s}_i, \hat{\mathbf{a}}_i)|_{\hat{\mathbf{a}} \sim \pi_\psi(\cdot|\mathbf{s})}. \tag{17}$$

Then, the policy gradient $\nabla_\psi J(\psi)$ is computed by the stochastic policy gradient algorithm [21] and used to update the parameters through gradient ascent:

$$\psi \leftarrow \psi + \eta \nabla_\psi J(\psi). \tag{18}$$

Lastly, the entropy regularization term α controls exploration, with higher values corresponding to more exploration. While a DRL algorithm with a deterministic policy can be used, it requires additive noise for exploration. In contrast, entropy regularization in SAC considers the current policy's knowledge, making it a more effective solution for the challenging transmit beamforming and phase shift design.

C. Construction of the Environment

RL distinguishes environments into *episodic* and *non-episodic* tasks. An episode ends when a terminal condition is met. In contrast, non-episodic (continuing) tasks have no specific endpoint. In our case, the task is considered continuing because the BS continuously performs beamforming and configures RIS elements. Terminal conditions can be set, as in [9], but they might introduce bias and mislead the learning agents [24]. Therefore, we adopt a continuing task framework in developing our approach.

1) Action: The policy network outputs the flattened concatenation of $\mathbf{G}$ and $\boldsymbol{\phi}$ as the action vector. However, neural networks cannot process complex numbers. Therefore, the actor network produces the real and imaginary parts separately

and constructs $\mathbf{G}$ and $\boldsymbol{\phi}$. To satisfy the transmit power and phase domain constraints, i.e., (7) and (9), the agent normalizes the output of actions. Consequently, an action vector consists of $2MK + 2L$ elements.

2) State: The state vector consists of transmission and reception powers for each user, the previous action, and the cascaded channel matrices $\mathbf{D}_k$ or their estimates ($\hat{\mathbf{D}}_k$) for $k = 1, \ldots, K$, depending on whether the BS has perfect or imperfect CSI. Similarly, these matrices are flattened and the number of elements is doubled due to the real and imaginary parts except for powers. We consider the transmission powers allocated to each data stream at the BS and the reception power at each user. Consequently, we obtain $2K$ power-related entries. $2KLM$ entries come from the cascaded channel estimates for each user, and $2MK+2L$ entries come from the previous action vector, resulting in a $2KLM+2MK+2L+2K$-dimensional state vector. Furthermore, the correlation between state dimensions degrades the performance of learning RL agents [24]. Hence, we whiten state vectors after each environment step. Finally, the initial state of the training still requires the action in the previous step. Thus, we initialize $\mathbf{G}$ as an identity matrix and $\boldsymbol{\phi}$ as a vector of ones to constitute the initial environment state.

3) Reward: At every time step, the reward is determined by the sum downlink rate expressed by either (6) or (8), depending on the considered objective function.

IV. Methodology

A. Adapting the Deep Reinforcement Learning Framework to Non-Episodic Tasks

While $\gamma < 1$ prioritizes short-term rewards, the agent must be equally concerned with instantaneous and future rewards since the reward (sum rate) should always be kept maximum [24]. Therefore, we set $\gamma = 1$. Moreover, agents must carefully remember the outcomes of past actions to compute future actions, which cannot be achieved by learning only from instant rewards [24]. To solve this, the DRL framework should be adapted to non-episodic tasks by considering the *average reward* concept. Therefore, the reward in the current step used to train the agent is modified as follows:

$$\tilde{r} \triangleq r - \bar{r}, \tag{19}$$

where r is the instantaneous reward computed by the environment in the current state and $\bar{r}$ is the average of the rewards collected till the current state. Recall that the definition of $r + Q_{\theta'}(s, a)$ (the estimate of the value $Q_\theta(s, a)$ when $\gamma = 1$) corresponds to the rewards that the agent will collect till the terminal state. However, there is no terminal state in continuing tasks. Hence, the sum of collected rewards could go to infinity. The average reward overcomes this by restraining the estimation value. Hence, the agent should learn only from $\tilde{r}$ instead of r or $\bar{r}$.

B. Maximizing the True Sum Rate Under the Mismatch Environment Model

To maximize (6) while learning from (8), we leverage a recent work proposed for the exploration of continuous action spaces, the Deep Directed Intrinsically Motivated Exploration (DISCOVER) algorithm [25]. Motivated by the animal psychological systems, DISCOVER utilizes a separate *explorer network* ξ_ω with parameters ω that represents a deterministic *exploration policy*. Its objective is to perturb the actions selected by the policy so that the prediction error by the Q-network is constantly maximized. Consequently, this leads agents to state-action spaces where Q-value prediction is difficult, allowing them to correct the prediction error of unknown or less selected actions.

However, we do not directly use the DISCOVER algorithm since SAC already explores the action space by utilizing a stochastic policy with the entropy parameter, i.e., the term α in (14) and (17). Instead, we slightly modify the DISCOVER algorithm such that the explorer network predicts $\beta(\varphi_l)$ for $l = 1, \ldots, L$ on the observed states:

$$\xi_\omega(s) = \big[\underbrace{1 \quad \ldots \quad 1}_{2MK} \quad \underbrace{\hat{\beta}_1 \quad \hat{\beta}_1 \quad \hat{\beta}_2 \quad \ldots \quad \hat{\beta}_L \quad \hat{\beta}_L}_{2L}\big]^\top, \tag{20}$$

where $\hat{\beta}_l \in [\beta_{\min}, 1]$. The number of ones in the latter equation is the number of elements included by $\mathbf{G}$ to the action vector. This is feasible since the transmit beamforming produced by the agent does not affect the RIS reflection loss in the phase-dependent amplitude model. In addition, there are two entries for each $\hat{\beta}_l$ to scale both the real and imaginary parts of $\hat{\boldsymbol{\phi}}$. Thus, the prediction of $\xi_\omega(s)$ is used to perturb only the phase part of the actions:

$$a_{\hat{\beta}} \triangleq a \odot \lambda \cdot \xi_\omega(s) \implies \hat{\boldsymbol{\phi}}_{\hat{\beta}} \triangleq \big[\hat{\beta}_1 e^{j\varphi_1} \ldots \hat{\beta}_L e^{j\varphi_L}\big]^\top, \tag{21}$$

where $\odot$ is the Hadamard product, and the hyperparameter $\lambda \in (0, 1]$ restricts the explorer network not to perturb the actions chosen by the actor detrimentally, similar to DISCOVER. The environment takes $a_{\hat{\beta}}$ from the agent and computes the next state and reward with respect to $a_{\hat{\beta}}$. Therefore, the perturbed actions are sampled from the experience replay buffer instead of the raw ones produced by the policy network. Accordingly, the losses for the Q- and actor networks are modified as follows:

$$\hat{\mathbf{y}}_{\hat{\beta}} \triangleq \tilde{\mathbf{r}} + \min_{i=1,2} Q_{\theta'_i}(\mathbf{s}', \mathbf{a}' \odot \lambda \cdot \xi_{\omega'}(\mathbf{s})) - \alpha \log(\mathbf{a}'|\mathbf{s}'); \tag{22}$$

$$J_{\hat{\beta}}(\theta_i) \triangleq \frac{1}{N} \|\hat{\mathbf{y}}_{\hat{\beta}} - Q_{\theta_i}(\mathbf{s}, \mathbf{a}_{\hat{\beta}})\|_2^2, \tag{23}$$

$$J_{\hat{\beta}}(\psi) \triangleq \frac{1}{N} \sum_{i}^{N} \alpha \log \pi_\psi(\hat{\mathbf{a}}_i|\mathbf{s}_i) - \min_{j=1,2} Q_{\theta_j}(\mathbf{s}_i, \hat{\mathbf{a}}_{\hat{\beta},i})|_{\hat{\mathbf{a}} \sim \pi_\psi(\cdot|\mathbf{s})}, \tag{24}$$

where $\hat{\mathbf{a}}_{\hat{\beta},i} = \hat{\mathbf{a}} \odot \lambda \cdot \xi_\omega$. Notice that a target explorer network also perturbs the next action a', as in DISCOVER. Also, we use $\gamma = 1$ and average reward in (22). The explorer network is optimized such that the sum of absolute TD errors is maximized:

$$\tilde{\delta}_{\hat{\beta}_i}(\mathbf{s}, \mathbf{a} \odot \lambda \cdot \xi_\omega(\mathbf{s})) \triangleq \frac{1}{N} \|\hat{\mathbf{y}}_{\hat{\beta}} - Q_{\theta_i}(\mathbf{s}, \mathbf{a} \odot \lambda \cdot \xi_\omega(\mathbf{s}))\|_2^2, \tag{25}$$

$$J(\omega) = \tilde{\delta}_{\hat{\beta}_1}(\mathbf{s}, \mathbf{a} \odot \lambda \cdot \xi_\omega(\mathbf{s})) + \tilde{\delta}_{\hat{\beta}_2}(\mathbf{s}, \mathbf{a} \odot \lambda \cdot \xi_\omega(\mathbf{s})). \tag{26}$$

The deterministic exploration network is then updated through the Deterministic Policy Gradient algorithm [26]:

$$\nabla_\omega J(\omega) = \sum_{i=1}^{2} \mathbb{E}[\nabla_\zeta \tilde{\delta}_{\hat{\beta}_i}(\mathbf{s}, \mathbf{a} \odot \zeta)|_{\zeta=\lambda\cdot\xi_\omega(\mathbf{s})} \nabla_\omega \xi_\omega(\mathbf{s})], \quad (27)$$

$$\omega \leftarrow \omega + \eta \nabla_\omega J(\omega). \quad (28)$$

This forms our framework to solve the downlink RIS-aided MU-MISO system under the phase-dependent amplitude model. Overall, the explorer network predicts $\beta(\varphi_l)$ and scales the actions selected by the policy using $\hat{\beta}_l$. Then, the scaled action is fed to the environment. Notice that the reward (sum rate) computed by the environment is altered with respect to the scaled actions $a_{\hat{\beta}}$ (or $\hat{\boldsymbol{\phi}}_{\hat{\beta}}$):

$$\hat{R}_{\Sigma,\hat{\beta}} \triangleq \sum_{k=1}^{K} \log\left(1 + \frac{\|\hat{\boldsymbol{\phi}}_{\hat{\beta}}^\top \hat{\mathbf{D}}_k \mathbf{G}\|^2}{\sum_{j\neq k} \|\hat{\boldsymbol{\phi}}_{\hat{\beta}}^\top \hat{\mathbf{D}}_j \mathbf{G}\|^2 + \sigma_w^2}\right). \quad (29)$$

Hence, the agent observes the effect of the explorer network's $\beta(\varphi_l)$ prediction through the reward it receives, which is equivalent to implicitly learning the true environment model. By maximizing the TD error, the exploration policy further forces the Q-networks to learn from its prediction mistakes since now $\hat{\beta}$-altered actions $a_{\hat{\beta}}$ and rewards $\hat{R}_{\Sigma,\hat{\beta}}$ are included in the loss of Q-networks, i.e., (23). Ultimately, the exploration policy learns the true environment model by considering the current knowledge of the Q-networks and policy [25]. We refer to the resulting algorithm as *β-Space Exploration* and provide the pseudocode in our repository[1].

Complexity Analysis: DISCOVER adds another neural network to the training process, slightly increasing computational complexity by less than 33% compared to SAC's three networks (policy and two critics). The introduced complexity can never be 33% since the input dimensions of the explorer and actor networks, i.e., only the state dimension, is always less than the input dimension of the critic network, i.e., the aggregated state and action dimensions.

V. Results

A. Simulation Setup

To test the effectiveness of β-Space Exploration, we compare it against two vanilla SAC agents corresponding to two scenarios: golden standard and mismatch. In the golden standard scenario, the agent knows the true environment model and is trained using the rewards computed according to (6). Moreover, the agent has perfect CSI. In the mismatch case, however, the BS tries to solve (9), learns from the rewards computed according to (8), and has imperfect CSI. While the vanilla SAC agent is tested under both scenarios, the SAC agent combined with β-Space Exploration is tested under the mismatch scenario.

Our simulations follow the well-known DRL benchmarking standards [29], that is, each experiment runs over ten random seeds for a fair comparison with the baselines. Furthermore, the implementation of the SAC algorithm follows the structure outlined in the original paper [21]. We performed an extensive hyperparameter tuning starting from the hyperparameter setting provided by [21]. The tuned hyperparameter setting is outlined in Table I along with the chosen environment parameter values. We also linearly decay the exploration regularization term λ such that it becomes zero at the end of the training. Highly perturbed actions (i.e., large λ values) in the final steps may degrade the performance of a SAC agent that learned to control the environment sufficiently well. Precise experimental setup and implementation can be found in the code of our repository[1].

B. Discussion

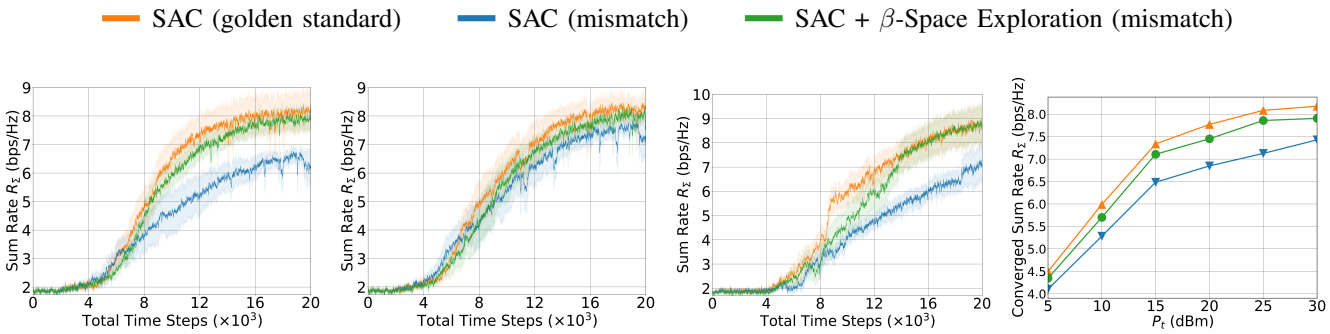
In Fig. 1, we report the instantaneous sum rates computed according to (6), averaged over ten random seeds. While the agents are trained using the average reward $\tilde{r}$ in (19), the performance is assessed under the instantaneous rewards r. Also, Table II reports the converged sum rates, being the average of the last 1000 instant rewards over ten trials, per the DRL benchmarking standards [29].

From the evaluation results, we infer that β-Space Exploration attains near-optimal results in all of the settings tested. Specifically, when $\beta_{\min} = 0.3$, the SAC agent under the mismatch environment shows considerably worse performance than the golden standard. The resulting

TABLE I: The hyperparameter setting used to produce the reported results. No tuning was performed on the environment parameters.

Hyperparameter	**Value**
# of hidden layers†	2
# of units in each hidden layer†	256
Hidden layers activation†	ReLU
Final layer activation (Q-networks)	Linear
Final layer activation (actor, explorer)	tanh
Learning rate η†	10^{-3}
Weight decay†	None
Weight initialization†	Xavier uniform [27]
Bias initialization†	constant
Optimizer†	Adam [28]
Total time steps per training	20000
Experience replay buffer size	20000
Experience replay sampling method	uniform
Mini-batch size	16
Discount term γ	1
Learning rate for target networks τ†	10^{-3}
Network update interval†	after each environment step
Initial α	0.2
Entropy target	`-action dimension`
SAC log standard deviation clipping	$(-20, 2)$
SAC ϵ	10^{-6}
Initial β-Space Exploration λ	0.3
μ‡	0
κ‡	1.5
Channel noise variance σ_e^2‡	10^{-2}
AWGN channel variance σ_w^2 ‡	10^{-2}
Channel matrix initialization (Rayleigh)‡	$\mathcal{CN}(0, 1)$

† Applies to all neural networks
‡ Environment hyperparameter



(a) $\beta_{\min}$ = 0.3, P_t = 30 dBm, K = 4, M = 4, L = 16

(b) $\beta_{\min}$ = 0.6, P_t = 30 dBm, K = 4, M = 4, L = 16

(c) $\beta_{\min}$ = 0.6, P_t = 30 dBm, K = 4, M = 4, L = 64

(d) $\beta_{\min}$ = 0.6, P_t = $\{5, 10, 15, 20, 25, 30\}$ dBm, K = 4, M = 4, L = 16

Fig. 1: Learning curves for the tested settings. Shaded regions represent 95% confidence intervals over 10 random seeds for each result. A sliding window of size 25 smooths the curves for visual clarity.

performance approaches the golden standard when $\beta_{\min}$ is increased to 0.6. This is expected since the interval for possible RIS loss factors shrinks as $\beta(\varphi_l) \in [\beta_{\min}, 1]$. However, our method is not affected by the $\beta_{\min}$ value. For each value of $\beta_{\min}$, it exhibits a robust performance, achieving high sum downlink rates slightly lower than the golden standard. Furthermore, β-Space Exploration regards no issues with the convergence rate, that is, learning curves are practically parallel to the golden standard. This implies that the exploration policy can implicitly learn how its action selections affect the loss in the RIS reflections, which is done in a negligible amount of time compared to the total training duration.

When the number of RIS elements L is increased to 64, the sum rate achieved by the golden standard increases slightly due to additional degrees of freedom to control the propagation environment. On the other hand, the vanilla SAC agent cannot benefit from this due to the increased number of misspecified RIS amplitudes. In contrast, our proposed method converges to the sum rates achieved by the golden standard with a slight delay, despite being trained according to (8). As shown in Table II, 99% of the sum rate loss caused by the mismatch is compensated by β-Space Exploration when $L = 64$. We also observe that β-Space Exploration offers consistent performance gains over the vanilla SAC agent under mismatch for different transmit power levels. Additionally, the resulting confidence intervals of our algorithm are usually tighter than the ones corresponding to the golden standard. This suggests that our framework improves credibly over the baseline due to the structure of the introduced method rather than unintended consequences or any exhaustive hyperparameter tuning.

Lastly, the computational cost of the Q- and actor networks of the SAC algorithm in the considered environment depend on the values of the environment setting parameters M, K, and L. Increasing these parameters increases the number of state and action dimensions, which in turn increases the number of parameters and operations involved in the forward pass of the networks, leading to a higher computational cost. Therefore, the values of M, K, and L should be chosen carefully to balance the performance of the selected DRL algorithm with its computational efficiency.

TABLE II: Average of last 1000 instant rewards achieved by the SAC agents, computed according to (6), over 10 trials of 20000 time steps. $\pm$ captures a 95% confidence interval over the trials. The performance increase denotes the percentage of mean sum rate improvement obtained by β-Space Exploration over the vanilla SAC agent in the mismatch environment with respect to the difference between the golden standard and mismatch scenarios.

Setting	**Golden Standard**	**Mismatch**	**β-Space Exploration**	**Performance Increase**
$\beta_{\min}$ = 0.3, P_t = 30 dBm, K = 4, M = 4, L = 16	8.16 $\pm$ 1.24	6.36 $\pm$ 0.67	7.88 $\pm$ 0.69	84%
$\beta_{\min}$ = 0.6, P_t = 30 dBm, K = 4, M = 4, L = 16	8.18 $\pm$ 0.77	7.43 $\pm$ 0.78	7.91 $\pm$ 0.43	64%
$\beta_{\min}$ = 0.6, P_t = 30 dBm, K = 4, M = 4, L = 64	8.71 $\pm$ 0.84	7.00 $\pm$ 0.57	8.70 $\pm$ 0.94	99%
P_t = 5 dBm, $\beta_{\min}$ = 0.6, K = 4, M = 4, L = 16	4.50 $\pm$ 0.45	4.11 $\pm$ 0.29	4.35 $\pm$ 0.36	62%
P_t = 10 dBm, $\beta_{\min}$ = 0.6, K = 4, M = 4, L = 16	5.99 $\pm$ 0.41	5.28 $\pm$ 0.42	5.70 $\pm$ 0.39	59%
P_t = 15 dBm, $\beta_{\min}$ = 0.6, K = 4, M = 4, L = 16	7.34 $\pm$ 0.83	6.49 $\pm$ 0.61	7.11 $\pm$ 0.54	73%
P_t = 20 dBm, $\beta_{\min}$ = 0.6, K = 4, M = 4, L = 16	7.77 $\pm$ 0.60	6.85 $\pm$ 0.77	7.45 $\pm$ 0.50	65%
P_t = 25 dBm, $\beta_{\min}$ = 0.6, K = 4, M = 4, L = 16	8.08 $\pm$ 0.83	7.13 $\pm$ 0.71	7.86 $\pm$ 0.57	77%
P_t = 30 dBm, $\beta_{\min}$ = 0.6, K = 4, M = 4, L = 16	8.18 $\pm$ 0.77	7.43 $\pm$ 0.78	7.91 $\pm$ 0.43	64%

VI. Concluding Remarks

In this paper, we present a novel DRL-based approach, β-Space Exploration, to address the three critical aspects of non-episodic tasks, imperfect CSI, and hardware impairments in RIS-aided MU-MISO systems represented by the phase-dependent reflection amplitude model [1]. Our method jointly designs transmit beamforming and phase shifts to maximize the sum downlink rate of the users. The empirical studies show that β-Space Exploration attains near-optimal results, is robust to various settings, and compensates for the sum rate loss caused by hardware impairments in the RIS. Consequently, our findings highlight the potential of our approach as a promising solution to overcome hardware impairments in RIS-aided wireless communication systems. In addition, while the current work considers slow-fading channels, channel aging models can easily be added to our environment code although there exist many opportunities to improve the DRL agent design with channel aging in mind.

References

[1] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5849–5863, 2020.

[2] C. Ozturk, M. F. Keskin, H. Wymeersch, and S. Gezici, "On the impact of hardware impairments on RIS-aided localization," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 2846–2851.

[3] J. Wang, W. Tang, Y. Han, S. Jin, X. Li, C.-K. Wen, Q. Cheng, and T. J. Cui, "Interplay between RIS and AI in wireless communications: Fundamentals, architectures, applications, and open research problems," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2271–2288, 2021.

[4] Z. Yang, Y. Liu, Y. Chen, and N. Al-Dhahir, "Machine learning for user partitioning and phase shifters design in RIS-aided NOMA networks," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7414–7428, 2021.

[5] Q. Zhang, W. Saad, and M. Bennis, "Millimeter wave communications with an intelligent reflector: Performance optimization and distributional reinforcement learning," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1836–1850, 2022.

[6] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3978–3983, 2021.

[7] A. Al-Hilo, M. Samir, M. Elhattab, C. Assi, and S. Sharafeddine, "Reconfigurable intelligent surface enabled vehicular communication: Joint user scheduling and passive beamforming," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 2333–2345, 2022.

[8] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in uav-ris wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2042–2055, 2021.

[9] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1839–1850, 2020.

[10] 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.

[11] X. Liu, Y. Liu, Y. Chen, and H. V. Poor, "RIS enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1057–1071, 2021.

[12] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *Trans. Wireless. Comm.*, vol. 20, no. 1, p. 375388, jan 2021. [Online]. Available: https://doi.org/10.1109/TWC.2020.3024860

[13] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, and Z. Zhang, "Hybrid beamforming for RIS-empowered multi-hop terahertz communications: A drl-based method," in *2020 IEEE Globecom Workshops (GC Wkshps*, 2020, pp. 1–6.

[14] J. Kim, S. Hosseinalipour, T. Kim, D. J. Love, and C. G. Brinton, "Multi-irs-assisted multi-cell uplink MIMO communications under imperfect csi: A deep reinforcement learning approach," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2021, pp. 1–7.

[15] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019.

[16] 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." in *ICLR (Poster)*, 2016. [Online]. Available: http://arxiv.org/abs/1509.02971

[17] P. Chen, X. Li, M. Matthaiou, and S. Jin, "DRL-based RIS phase shift design for OFDM communication systems," *IEEE Wireless Communications Letters*, pp. 1–1, 2023.

[18] R. E. Bellman, *Dynamic Programming*. USA: Dover Publications, Inc., 2003.

[19] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013.

[20] R. Sutton, "Learning to predict by the method of temporal differences," *Machine Learning*, vol. 3, pp. 9–44, 08 1988.

[21] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870. [Online]. Available: https://proceedings.mlr.press/v80/haarnoja18b.html

[22] L. ji Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," in *Machine Learning*, 1992, pp. 293–321.

[23] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1587–1596. [Online]. Available: https://proceedings.mlr.press/v80/fujimoto18a.html

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[25] B. Saglam and S. S. Kozat, "Deep intrinsically motivated exploration in continuous control," 2022. [Online]. Available: https://arxiv.org/abs/2210.00293

[26] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 1. Bejing, China: PMLR, 22–24 Jun 2014, pp. 387–395. [Online]. Available: https://proceedings.mlr.press/v32/silver14.html

[27] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. [Online]. Available: https://proceedings.mlr.press/v9/glorot10a.html

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[29] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11694