

Deep Reinforcement Learning for Optimizing RIS-Assisted HD-FD Wireless Systems

Alice Faisal, *Student Member, IEEE*, Ibrahim Al-Nahhal, *Member, IEEE*,
Octavia A. Dobre, *Fellow, IEEE* and Telex M. N. Ngatched, *Senior
Member, IEEE*

Abstract

This letter investigates the reconfigurable intelligent surface (RIS)-assisted multiple-input single-output (MISO) wireless system, where both half-duplex (HD) and full-duplex (FD) operating modes are considered together, for the first time in the literature. The goal is to maximize the rate by optimizing the RIS phase shifts. A novel deep reinforcement learning (DRL) algorithm is proposed to solve the formulated non-convex optimization problem. The complexity analysis and Monte Carlo simulations illustrate that the proposed DRL algorithm significantly improves the rate compared to the non-optimized scenario in both HD and FD operating modes using a single parameter setting. Besides, it significantly reduces the computational complexity of the downlink HD MISO system and improves the achievable rate with a reduced number of steps per episode compared to the conventional DRL algorithm.

Index Terms

Reconfigurable intelligent surface (RIS), half-duplex full-duplex (HD-FD), deep reinforcement learning (DRL).

I. INTRODUCTION

RECONFIGURABLE intelligent surfaces (RIS) have emerged as a promising paradigm to fulfill the need of a smart and programmable wireless environment, and meet the demands of future wireless networks [1], [2]. RIS consists of a two-dimensional array of low-cost passive electromagnetic (EM) elements [3]. By overcoming the random nature of EM wave propagation,

A. Faisal, I. Al-Nahhal, O. A. Dobre, and T. M. N. Ngatched are with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL, Canada, (e-mail: afaisal, ioalnahhal, odobre@mun.ca; tngatched@grenfell.mun.ca).

RIS enables controlling different characteristics of radio waves, such as scattering, reflection, and refraction. Consequently, it effectively enhances the signal quality and boosts the wireless spectral efficiency by realizing a controllable environment [4].

RIS-assisted multiple-input multiple-output systems have recently drawn significant attention as a cost-effective solution to enhance the wireless transmission in both half-duplex (HD) and full-duplex (FD) operating modes [5]–[10]. In the HD mode, systems require additional resources to receive and forward signals, which results in a decreased spectral efficiency. In contrast, the FD mode has the potential to significantly increase the throughput of wireless systems as it enables simultaneous transmission and reception of signals in the same frequency band. However, this comes at the cost of increased interference and implementation complexity. To this end, some researchers are considering HD-FD transmission schemes that combine the advantages of both HD and FD modes [11]. In [5] and [6], RIS-HD systems are optimized to minimize the total transmit power. In [7], a joint optimization problem is considered to maximize the achievable rate of an RIS-HD system. In [8] and [9], the sum-rate and spectral efficiency of an RIS-FD system is maximized, respectively. In [10], the weighted minimum rate is maximized for a multi-user RIS-FD system. Most of these works decoupled the optimization variables using alternating optimization algorithms, which exhibit both loss of optimality and high computational complexity.

Deep learning has emerged as a powerful approach to optimize the RIS phase shifts by tackling the practical implementation problems of the optimization techniques [12], [13]. In particular, deep reinforcement learning (DRL) is a potential candidate to optimize the RIS phase shifts without the need for offline training with a labeled dataset. A few works have considered DRL approaches to optimize RIS-HD systems [14]–[16]. The authors in [14] proposed an optimization-driven deep deterministic policy gradient (DDPG) to minimize the access point's transmit power. The sum-rate maximization problem of a multi-user RIS-HD system was addressed in [15] using a DRL algorithm. Furthermore, a conventional DRL algorithm is introduced in [16] to maximize the received signal-to-noise ratio of the downlink RIS-HD multiple-input single-output (MISO) system. To the best of the authors' knowledge, utilizing DRL for RIS-FD systems has not yet been discussed in the literature.

In this letter, a novel DRL algorithm is proposed to optimize the phase shifts of an RIS-assisted HD-FD MISO system. The contributions are summarized as follows:

- The proposed DRL algorithm achieves promising results in the HD and FD operating modes

without the need of additional parameters tuning.

- It provides a significant improvement in the rate compared to the non-optimized RIS phase shifts in the HD and FD operating modes.
- It significantly reduces the computational complexity, while providing a considerable rate improvement with a reduced number of required steps for each episode, compared to the conventional DRL in [16] for the HD mode.
- The complexity analysis and Monte Carlo simulations support the findings.

The remainder of this letter is organized as follows: Section II presents the system model and problem formulation for the RIS-assisted HD-FD MISO system. The proposed DRL algorithm is introduced in Section III, and its computational complexity is analyzed in Section IV. Simulation results and conclusions are presented in Sections V and VI, respectively.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Consider an RIS-assisted HD-FD MISO system as illustrated in Fig. 1, where S_1 and S_2 represent the base station (BS) and user equipment (UE), respectively. Both the BS and UE are equipped with M transmit antennas and one receive antenna. The UE sometimes operates in a HD mode, where it only receives information from the BS (i.e., downlink HD mode), while other times the UE and BS transmit and receive information simultaneously in the same frequency band (i.e., FD mode). Henceforth, Ω denotes the operating mode, where $\Omega \in \{\text{HD}, \text{FD}\}$. The RIS is composed of N programmable reflecting elements, which assists the communication between S_1 and S_2 by optimizing the RIS phase shifts through an RIS controller. Given $\bar{i} = 3 - i \forall i = 1, 2$, let $\mathbf{H}_{S_{\bar{i}}R} \in \mathbb{C}^{N \times M}$, $\mathbf{h}_{RS_i}^H \in \mathbb{C}^{1 \times N}$, and $\mathbf{h}_{S_{\bar{i}}S_i}^H \in \mathbb{C}^{1 \times M}$ denote the channel coefficients of the $S_{\bar{i}}$ -RIS, RIS- S_i , and $S_{\bar{i}}$ - S_i links, respectively. The self-interference (SI) channels, which are involved in the FD mode at the BS and UE are denoted by $\mathbf{h}_{S_iS_i}^H \in \mathbb{C}^{1 \times M}$.

At the receiver-side, the signal is received from the direct and reflected links of the BS and RIS, respectively. Thus, the noisy received signals of the downlink HD and FD operating modes are respectively expressed as

$$y_i^\Omega = \underbrace{\left(\mathbf{h}_{RS_i}^H \mathbf{\Theta} \mathbf{H}_{S_{\bar{i}}R} \right)}_{\text{Reflected signal}} + \underbrace{\left(\mathbf{h}_{S_{\bar{i}}S_i}^H \right)}_{\text{Direct signal}} \mathbf{w}_{\bar{i}} x_{\bar{i}} + n, \quad i = 2, \Omega = \text{HD}, \quad (1)$$

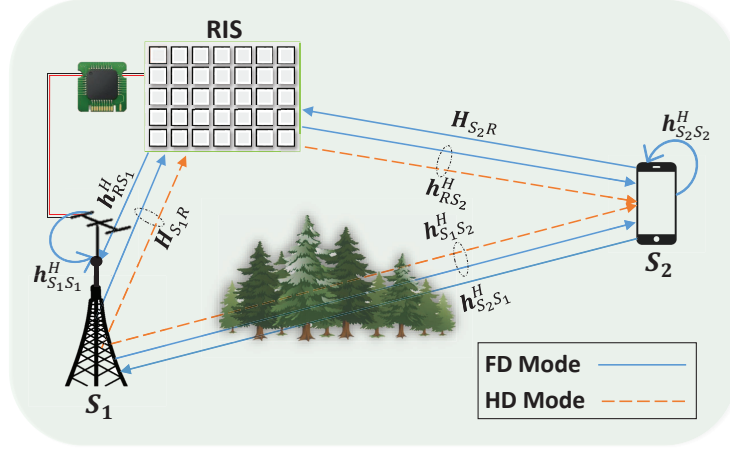


Fig. 1: RIS-assisted HD-FD MISO system.

and

$$y_i^\Omega = \underbrace{\left(\mathbf{h}_{RS_i}^H \mathbf{\Theta} \mathbf{H}_{S_i R} + \mathbf{h}_{S_i S_i}^H \right)}_{\text{Reflected signal}} \underbrace{\mathbf{w}_i x_i}_{\text{Direct signal}} + \underbrace{\mathbf{h}_{S_i S_i}^H \mathbf{w}_i x_i}_{\text{Residual SI}} + n, \quad i = 1, 2, \Omega = \text{FD}, \quad (2)$$

where $n \sim \mathcal{CN}(0, \sigma^2)$ denotes the additive white complex Gaussian noise with zero-mean and variance σ^2 . The diagonal matrix $\mathbf{\Theta} = \text{diag}(e^{j\varphi_1}, \dots, e^{j\varphi_n}, \dots, e^{j\varphi_N}) \in \mathbb{C}^{N \times N}$ represents the phase shifts of the RIS, where $\varphi_n \in [-\pi, \pi)$ is the phase shift introduced by the n -th reflecting element. The source node, S_i , employs an active beamforming $\mathbf{w}_i \in \mathbb{C}^{M \times 1}$ to transmit the information signal, x_i , with $\mathbb{E}\{|x_i|^2\} = 1$, where $\mathbb{E}\{\cdot\}$ denotes the expectation operation. The third term in (2) represents the SI introduced by the FD mode operation.

The achievable rate and sum-rate of the downlink HD and FD operating modes, measured in bit per second per Hertz (bps/Hz), are respectively given as

$$\mathcal{R}^\Omega = \log_2 \left(1 + \frac{\left| \left(\mathbf{h}_{RS_i}^H \mathbf{\Theta} \mathbf{H}_{S_i R} + \mathbf{h}_{S_i S_i}^H \right) \mathbf{w}_i \right|^2}{\sigma^2} \right), \quad i = 2, \Omega = \text{HD}, \quad (3)$$

and

$$\mathcal{R}^\Omega = \sum_{i=1}^2 \log_2 \left(1 + \frac{\left| \left(\mathbf{h}_{RS_i}^H \mathbf{\Theta} \mathbf{H}_{S_i R} + \mathbf{h}_{S_i S_i}^H \right) \mathbf{w}_i \right|^2}{|\mathbf{h}_{S_i S_i}^H \mathbf{w}_i|^2 + \sigma^2} \right), \quad \Omega = \text{FD}. \quad (4)$$

Here, the goal is to maximize the rate of the RIS-assisted HD-FD MISO system by optimizing

the RIS phase shifts. Thus, the resulting optimization problem can be expressed as

$$(P1) \quad \max_{\boldsymbol{\varphi}} \quad \mathcal{R}^{\Omega}, \quad \Omega \in \{\text{HD}, \text{FD}\} \quad (5a)$$

$$\text{s.t.} \quad -\pi \leq \varphi_n \leq \pi, \quad n = 1, \dots, N. \quad (5b)$$

It is worth noting that the conventional DRL algorithm in [16] has been proposed to solve the non-convex problem (P1) only when $\Omega = \text{HD}$, and suffers from high computational complexity. Moreover, the DRL for the FD operating mode has not yet been investigated in the literature.

III. PROPOSED DRL ALGORITHM

This section proposes a novel DRL algorithm to solve (P1) for the RIS-assisted HD-FD MISO system. To deal with (P1), the RIS phase shifts are optimized using the proposed DRL algorithm. Then, for a given optimized $\boldsymbol{\Theta}$, the transmit beamformers, $\mathbf{w}_{\bar{i}}$, are optimized using a closed and semi-closed form solutions for the HD and FD operating modes, respectively. The optimization problem is solved in an iterative fashion until the optimized $\boldsymbol{\Theta}$ and $\mathbf{w}_{\bar{i}}$ converge.

A. Beamforming Design for a Given $\boldsymbol{\Theta}$

The optimal beamforming vector for the HD operating mode is calculated using the maximum ratio transmission approach, whereas a semi-closed optimal solution of the FD beamforming vectors is given in [8]. Consequently, for a given optimized $\boldsymbol{\Theta}$, the optimal beamforming vectors of the HD and FD modes, $\mathbf{w}_{\bar{i}}$, are respectively given as

$$\mathbf{w}_{\bar{i}}^{\dagger} = \sqrt{P_{\max}} \frac{\left(\mathbf{h}_{RS_i}^H \boldsymbol{\Theta} \mathbf{H}_{S_i R} + \mathbf{h}_{S_i S_i}^H \right)^H}{\left\| \left(\mathbf{h}_{RS_i}^H \boldsymbol{\Theta} \mathbf{H}_{S_i R} + \mathbf{h}_{S_i S_i}^H \right) \right\|}, \quad i = 2, \quad \Omega = \text{HD}, \quad (6)$$

and

$$\mathbf{w}_{\bar{i}}^{\dagger} = (\delta \mathbf{h}_{S_i S_i} \mathbf{h}_{S_i S_i}^H + v^{\dagger} \mathbf{I})^{-1} \mathcal{B}, \quad i = 1, 2, \quad \Omega = \text{FD}, \quad (7)$$

where P_{\max} is the maximum transmitted power of $S_{\bar{i}}$, \mathbf{I} is the identity matrix, and v^{\dagger} is the optimal dual Lagrangian variable associated with the power constraint that is found by performing a bisection search over the interval $\left[0, \sqrt{\mathcal{B}^T \mathcal{B}} / \sqrt{P_{\max}} \right]$. Here, \mathcal{B} and δ are given as

$$\mathcal{B} \triangleq \frac{1}{\tilde{b}_i} \left(1 + \frac{b_i}{\|\mathbf{h}_{S_i S_i}^H \tilde{\mathbf{w}}_{\bar{i}}\|^2 + \sigma^2} \right) \mathbf{h}_{\bar{i}} \mathbf{h}_{\bar{i}}^H \tilde{\mathbf{w}}_{\bar{i}}, \quad (8)$$

and

$$\delta \triangleq \frac{b_i \left(|\mathbf{h}_{\tilde{i}}^H \tilde{\mathbf{w}}_{\tilde{i}}|^2 + \tilde{b}_i \right)}{\tilde{b}_i \left(|\mathbf{h}_{S_i S_i}^H \tilde{\mathbf{w}}_{\tilde{i}}|^2 + \sigma^2 \right)^2}, \quad (9)$$

where $b_i \triangleq |\mathbf{h}_i^H \mathbf{w}_i|^2$, $\tilde{b}_i \triangleq |\mathbf{h}_{S_i S_i}^H \mathbf{w}_i|^2 + \sigma^2$, $\mathbf{h}_{\tilde{i}} \triangleq \mathbf{H}_{S_i R}^H \mathbf{\Theta}^H \mathbf{h}_{RS_i} + \mathbf{h}_{S_i S_i}$, and $\tilde{\mathbf{w}}_{\tilde{i}}$ is a given feasible point.

B. Phase Shift Design Based on the Proposed DRL Algorithm

1) *Problem Transformation*: The RIS controller represents the DRL *agent*, while the RIS-assisted HD-FD MISO communication system represents the DRL *environment*. Thus, the *state space*, *action space*, and *reward* for the proposed DRL algorithm are defined as follows:

- **State space**: The state space at time step t , $s_t \in \mathbb{R}^{1 \times (N+1)}$, includes $\varphi_n \forall n = 1, \dots, N$ and the corresponding \mathcal{R}^Ω at time step $t-1$, and is defined as

$$s_t = \left[\mathcal{R}^{\Omega, (t-1)}, \varphi_1^{(t-1)}, \dots, \varphi_n^{(t-1)}, \dots, \varphi_N^{(t-1)} \right]. \quad (10)$$

- **Action space**: Since (P1) aims to optimize the RIS phase shifts, the action space at time step t , $a_t \in \mathbb{R}^{1 \times N}$, is expressed as

$$a_t = \left[\varphi_1^{(t)}, \dots, \varphi_n^{(t)}, \dots, \varphi_N^{(t)} \right]. \quad (11)$$

- **Reward**: As the target of (P1) is to maximize \mathcal{R}^Ω , the reward is expressed as

$$r_t = \mathcal{R}^{\Omega, (t)}, \quad \Omega \in \{\text{HD}, \text{FD}\}. \quad (12)$$

At each time step t , the agent receives the current state s_t from the environment, takes an action a_t based on a *policy* $\tilde{\pi}$, and receives a scalar reward r_t . Then, a new state s_{t+1} is obtained. The return of a state is defined as the *total discounted reward* from time step t onwards, and is given by $R_t = \sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k)$, where $\gamma \in (0, 1]$ is the DRL discount factor. The goal is to learn a policy that maximizes the expected cumulative discounted reward from the start state, as: $J(\tilde{\pi}) = \mathbb{E}[R_1 | \tilde{\pi}]$. The DDPG, which combines the benefits of value-based and policy-based approaches [17], is used to learn the optimal policy for a continuous a_t . In particular, the DDPG algorithm aims at maximizing the Q-value of (s, a) pair by training a deep neural network (DNN), defined as

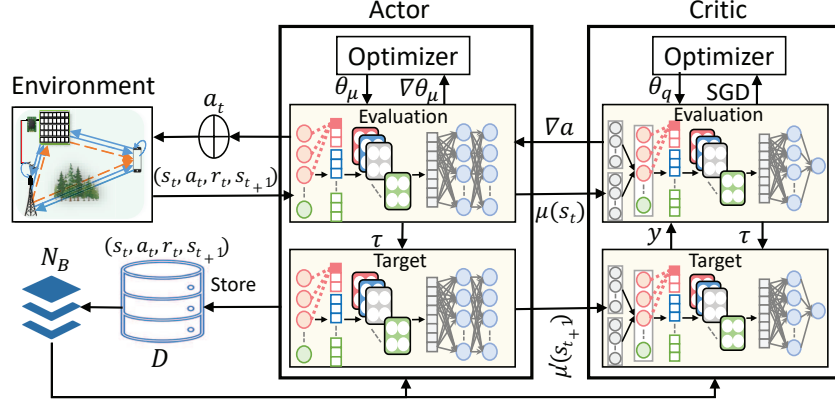


Fig. 2: The proposed DRL algorithm structure.

$$Q^{\tilde{\pi}_{\theta}}(s, a) = \mathbb{E}_{\tilde{\pi}_{\theta}} \left[R_1 | s_1 = s, a_1 = a \right], \quad (13)$$

where θ represents the DNN parameters, as well as finding the optimal policy by performing the gradient ascent of

$$\nabla_{\theta} J(\tilde{\pi}_{\theta}) = \mathbb{E}_{\tilde{\pi}_{\theta}} \left[Q^{\tilde{\pi}_{\theta}}(s, a) \nabla_{\theta} \log \pi_{\theta}(a|s) \right]. \quad (14)$$

The DDPG algorithm is based on the actor-critic technique, which consists of two DNN models: actor and critic. The actor, $\mu(s_t|\theta_{\mu})$, represents the policy network that takes the state as an input for a given θ_{μ} and outputs $a_t = \mu(s_t|\theta_{\mu}) + \xi$, where ξ is a random process that is added to the actions for exploration. ξ is modeled as complex Gaussian process with zero mean and variance 0.1. The critic, $Q(s_t, a_t|\theta_q)$, represents the network that evaluates the actions. It takes s_t and a_t as an input for a given θ_q , and outputs the Q-value. The DDPG algorithm utilizes the concept of experience replay with memory D to reduce the correlation of the training samples by randomly sampling minibatch transitions, N_B . Moreover, target networks are introduced to stabilize the learning process. The target networks are generated by making a copy of the actor and critic evaluation NNs, $\mu'(s_t|\theta_{\mu'})$ and $Q'(s_t, a_t|\theta_{q'})$, and are used to calculate the corresponding target values, y_t in (15). The actor and critic NN parameters, θ_{μ} and θ_q , are updated using the stochastic gradient descent (SGD) from (16) and policy gradient from (17), respectively. Finally, the target NN parameters are updated using a soft update coefficient, τ , based on (18) and (19). After T steps of each episode, the agent's performance saturates and it outputs the optimized Θ . The structure of the proposed DRL algorithm is illustrated in Fig. 2

and summarized in Algorithm 1.

Algorithm 1 Proposed DRL algorithm.

Initialize: θ_μ and θ_q with random weights, D , γ , τ , and learning rate α ;

Set: $\theta_{\mu'} \leftarrow \theta_\mu$ and $\theta_{q'} \leftarrow \theta_q$;

1: **repeat**

2: Collect the channels of the k -th episode based on Ω ;

3: Randomly initialize $\varphi_n \forall n = 1, \dots, N$ to obtain the initial state;

4: **if** $\Omega = \text{HD}$ **then**

5: Calculate $\mathbf{w}_{\bar{i}}$ using (6);

6: **else**

7: Calculate $\mathbf{w}_{\bar{i}}$ using (7);

8: **end if**

9: Initialize $\xi \sim \mathcal{CN}(0, 0.1)$;

10: **repeat**

11: Obtain $a_t = \mu(s_t | \theta_\mu) + \xi$ from the actor network and reshape it;

12: Repeat **Lines** #4-8;

13: Observe the new state, s_{t+1} , given a_t ;

14: Store (s_t, a_t, r_t, s_{t+1}) in D ;

15: When D is full, sample a minibatch of N_B transitions randomly (s_j, a_j, r_j, s_{j+1}) from D ;

16: Compute the target value using target networks:

$$y_j = r_j + \gamma Q'(s_{j+1}, \mu'(s_{j+1} | \theta_{\mu'}) | \theta_{q'}); \quad (15)$$

17: Update the critic by minimizing the loss using SGD:

$$L = \frac{1}{N_B} \sum_j (y_j - Q(s_j, a_j | \theta_q))^2; \quad (16)$$

18: Update the actor using the policy gradient:

$$\nabla_{\theta_\mu} = \frac{1}{N_B} \sum_j \nabla_a Q(s, a | \theta_q) |_{s=s_j, a=\mu(s_j)} \nabla_{\theta_\mu} \mu(s | \theta_\mu) |_{s_j}; \quad (17)$$

19: Update the target NNs through soft update:

$$\theta_{q'} \leftarrow \tau \theta_q + (1 - \tau) \theta_{q'}, \quad (18)$$

$$\theta_{\mu'} \leftarrow \tau \theta_{\mu'} + (1 - \tau) \theta_{\mu'}. \quad (19)$$

20: **until** $t = T$;

21: **until** $k = K$;

Output: Optimal action that corresponds to the optimal Θ .

2) *Proposed DNN Design:* As can be seen from Fig. 2, the proposed DRL algorithm contains four NNs (i.e., two NNs for the actor and two NNs for the critic). A novel design is proposed

for the four NNs, which consists of the input layer, two hidden layers and the output layer. The two hidden layers are a combination of one convolutional layer and one feed-forward (FF) layer with a flatten layer between them. The input layer of the actor and critic networks contains $N + 1$ neurons (i.e., size of s_t) and $2N + 1$ neurons (i.e., concatenation of s_t and a_t), respectively. The output layer of the actor and critic networks contains N neurons (i.e., size of a_t) and one neuron (i.e., scalar Q-value), respectively. The convolutional hidden layer for each of the actor and critic networks uses the *ReLU* activation function since it does not suffer from vanishing or exploding problems. In contrast, the FF hidden layer uses the *softmax* activation function to obtain probabilistic values for all inputs.

IV. COMPLEXITY ANALYSIS

The computational complexity of the conventional DRL algorithm in [16] and the proposed DRL algorithm for $\Omega = \text{HD}$ is derived in terms of the number of NN parameters $C_{\mathcal{P}}$ required to be stored, real additions $C_{\mathcal{A}}$, and real multiplications $C_{\mathcal{M}}$. The conventional DRL algorithm uses two hidden FF layers, and its computational complexity is given as

$$C_{\mathcal{P}} = \sum_{i=1}^3 (\eta_i + 1) \eta_{i+1}, \quad (20)$$

$$C_{\mathcal{M}} = \sum_{i=1}^3 \eta_i \eta_{i+1}, \quad (21)$$

$$C_{\mathcal{A}} = \sum_{i=1}^3 \eta_i \eta_{i+1} + \sum_{i=1}^3 \eta_{i+1}, \quad (22)$$

where η_i is the number of neurons of the i -th layer. It is worth noting that, for simplicity, each activation function is considered to cost one real addition.

Based on the NNs design in Section III-B2, the complexity for the proposed DRL algorithm is given as

$$C_{\mathcal{P}} = (\eta_F \eta_3 + F_z + 1) F_n + (\eta_4 + 1) \eta_3 + \eta_4, \quad (23)$$

$$C_{\mathcal{M}} = (F_z + \eta_3) \eta_F F_n + \eta_3 \eta_4, \quad (24)$$

$$C_{\mathcal{A}} = (F_z + \eta_3 + 1) \eta_F F_n + (\eta_4 + 1) \eta_3 + \eta_4, \quad (25)$$

where $\eta_F = \lfloor \frac{\eta_1 - F_z}{F_s} + 1 \rfloor$, with $\lfloor \cdot \rfloor$ as the floor operation, F_z is the filter size, F_n is the number of filters, and F_s is the stride. The complexity reduction of using the proposed DRL algorithm

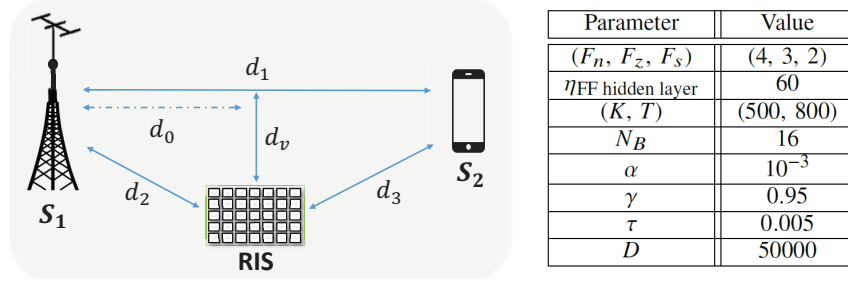


Fig. 3: Simulation setup and DDPG parameters.

over the conventional one for $\Omega = \text{HD}$ is

$$\text{Reduction} = 1 - \frac{\{C_{\chi}^{\text{Actor}} + C_{\chi}^{\text{Critic}}\}_{\text{Proposed}}}{\{C_{\chi}^{\text{Actor}} + C_{\chi}^{\text{Critic}}\}_{\text{Conventional}}}, \chi \in \{\mathcal{P}, \mathcal{A}, \mathcal{M}\}. \quad (26)$$

V. SIMULATION RESULTS

This section evaluates the performance of the proposed DRL algorithm for the RIS-assisted HD-FD MISO system. The simulation setup is shown in Fig. 3, where the considered parameters are $d_v = 2$ m and $d_1 = 50$ m. The distances of the BS-RIS and UE-RIS links are calculated as $d_2 = \sqrt{d_0^2 + d_v^2}$ m and $d_3 = \sqrt{(d_1 - d_0)^2 + d_v^2}$ m, respectively. The path loss (PL) at distance d_j , $\forall j \in \{1, 2, 3\}$ is modeled as $\text{PL} = \text{PL}_0 - 10\zeta \log_{10} \left(\frac{d_j}{D_r} \right)$ [16], where PL_0 is the PL at a reference distance D_r and ζ is the PL exponent, in which $\text{PL}_0 = -30$ dB and $D_r = 1$ m. As in [16], the BS-UE channels are modeled as Rayleigh fading (assuming a blocking element between S_1 and S_2), while the rest of the channels are Rician with a factor of 10. The PL exponents of the BS-UE, BS-RIS, and UE-RIS channels are set to $\zeta_{\text{BU}} = 3$ and $\zeta_{\text{BR}} = \zeta_{\text{UR}} = 2$, respectively. The PL of the SI channels for the FD mode is -95 dB. The total transmit power is $P = 5$ dBm, while the noise power is $\sigma^2 = -80$ dBm. The antenna gain at the BS and UE is 0 dBi, while the RIS gain is 5 dBi. The penetration loss in the BS-UE and RIS-UE links is 10 dB.

The parameters of the proposed DRL algorithm are summarized in Fig. 3. Furthermore, the design of the NNs is explained in Section III-B2, and its parameters are provided in Fig. 3. The Adam optimizer is used to update the parameters of the NNs. To assess the performance of the proposed algorithm, it is compared with the non-optimized scenario, referred to as random phase shifts. The conventional DRL algorithm in [16] with $T = 1000$ is also included to show the superiority of the proposed DRL algorithm in the HD mode. It is worth noting that the current

form of the conventional DRL algorithm can not be used to optimize the RIS phase shifts in the FD mode.

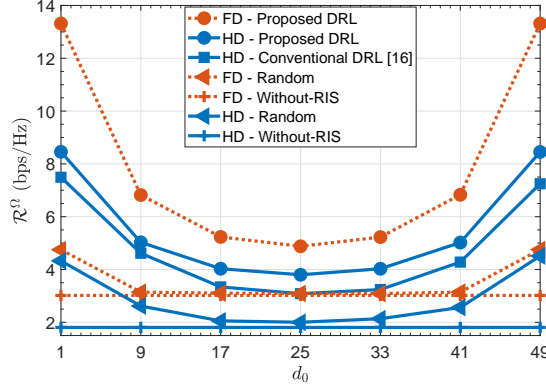


Fig. 4: RIS deployment investigation.

Figure. 4 studies the impact of the RIS location on the system performance. It is shown that the proposed DRL algorithm significantly improves the rate for both operating modes, compared to the random phase shifts and without-RIS scenarios, especially when the RIS is located closer to either the BS or the UE. On the other hand, the random phase shifts scenario does not improve the rate when the RIS is located relatively far from both BS and UE, compared to the scenario without-RIS. Consequently, a proper optimization for the RIS phase shifts is needed to achieve a satisfactory performance. For the rest of the letter, it is considered that $d_0 = 1$ m.

Figure. 5 illustrates the effect of increasing N on the system performance. As can be observed, R^Ω increases as N increases for all algorithms. The proposed DRL algorithm provides an improvement of 4.6 bps/Hz and 8.5 bps/Hz in the achievable rate and sum-rate of the HD and FD modes, respectively, compared to the random phase shifts scenario at $N = 40$. It is worth noting that the gain gap increases as N increases for the proposed DRL algorithm.

In the HD operating mode, the proposed DRL algorithm improves the achievable rate performance by 1.4 bps/Hz and 0.6 bps/Hz at $N = 20$ and $N = 40$, respectively, when compared to [16], as depicted in Fig. 5. Moreover, as shown in Fig. 6a, the proposed DRL algorithm in the HD mode (with $T = 800$ steps) significantly reduces the computational complexity of each NN in the range of 94% to 86% for the practical case of $N = 20$ to 60, respectively, compared to conventional DRL in [16] (with $T = 1000$ steps). Although the complexity reduction seems to decrease as N increases, it saturates at 63% for a certain large value of N , as seen from the asymptotic

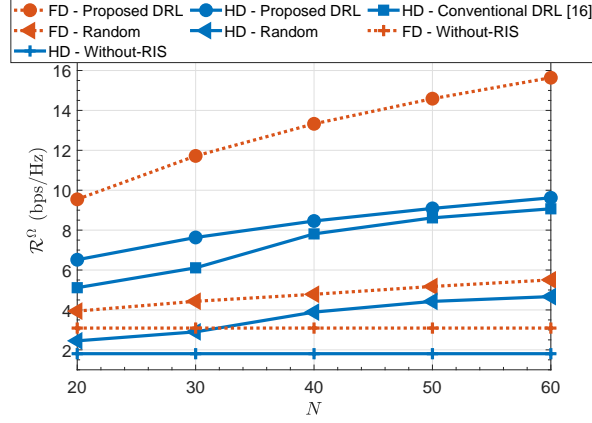


Fig. 5: The impact of varying N on the system performance.

complexity bound in Fig. 6b.

Finally, the proposed DRL algorithm provides a significant improvement in the rate for both operating modes, compared with the random phase shifts scenario. Besides, with a 20% reduction in the number of required steps when compared with the conventional DRL algorithm, the proposed DRL algorithm guarantees a faster convergence and improves the rate with lower computational complexity for each of the four NNs.

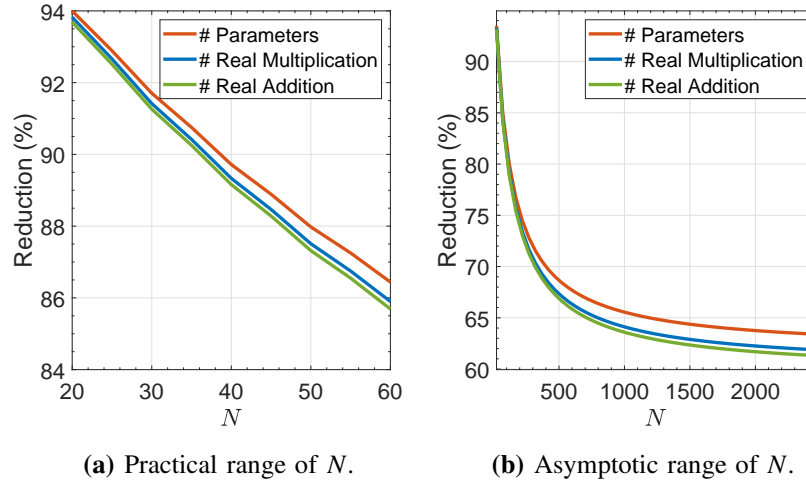


Fig. 6: Complexity reduction percentage versus N .

VI. CONCLUSION

This letter considered DRL for the rate maximization problem of the RIS-assisted HD-FD MISO system, for the first time in the literature. With a single parameter setting, the proposed

DRL algorithm optimized the RIS phase shifts for both HD and FD operating modes. A novel DNN structure was proposed to learn the optimal policy of the proposed DRL algorithm. Compared to the non-optimized scenario, the proposed DRL algorithm significantly improved the rate for the HD and FD operating modes, respectively. Compared to the conventional DRL algorithm in HD mode, the proposed DRL algorithm saved 20% of the required steps per episode and achieved up to 1.4 bps/Hz rate improvement with up to 94% reduction in the computational complexity. Future works can consider extending the proposed DRL algorithm to optimize the multi-user scenario.

REFERENCES

- [1] I. Al-Nahhal *et al.*, "Reconfigurable intelligent surface-assisted uplink sparse code multiple access," *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 2058–2062, Feb. 2021.
- [2] L. Bariah *et al.*, "A prospective look: Key enabling technologies, applications and open research topics in 6G networks," *IEEE Access*, vol. 8, pp. 174792–174820, Aug. 2020.
- [3] E. Basar *et al.*, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, Aug. 2019.
- [4] R. Alghamdi *et al.*, "Intelligent surfaces for 6G wireless networks: A survey of optimization and performance analysis techniques," *IEEE Access*, vol. 8, pp. 202795–202818, Oct. 2020.
- [5] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838–1851, Dec. 2020.
- [6] G. Zhou *et al.*, "Robust beamforming design for intelligent reflecting surface aided MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1658–1662, Jun. 2020.
- [7] N. S. Perović *et al.*, "Achievable rate optimization for MIMO systems with reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3865–3882, Feb. 2021.
- [8] H. Shen *et al.*, "Beamforming design with fast convergence for IRS-aided full-duplex communication," *IEEE Commun. Lett.*, vol. 24, no. 12, pp. 2849–2853, Aug. 2020.
- [9] J. Zhao *et al.*, "Energy efficient full-duplex communication systems with reconfigurable intelligent surface," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Feb. 2020, pp. 1–5.
- [10] Z. Peng *et al.*, "Multiuser full-duplex two-way communications via intelligent reflecting surface," *IEEE Trans. Signal Process.*, vol. 69, pp. 837–851, Jan. 2021.
- [11] M. Elhattab *et al.*, "Reconfigurable intelligent surface enabled full-duplex/half-duplex cooperative non-orthogonal multiple access," Jan. 2021. [Online]. Available: <https://arxiv.org/abs/2101.01307>
- [12] A. Zappone *et al.*, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Jun. 2019.
- [13] Y. Chen *et al.*, "Reinforcement learning meets wireless networks: A layering perspective," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 85–111, Jan. 2021.
- [14] J. Lin *et al.*, "Deep reinforcement learning for robust beamforming in IRS-assisted wireless communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Jan. 2020, pp. 1–6.

- [15] C. Huang *et al.*, “Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Jun. 2020.
- [16] K. Feng *et al.*, “Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 745–749, Jan. 2020.
- [17] T. P. Lillicrap *et al.*, “Continuous control with deep reinforcement learning,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2016, pp. 1–14.