# Intelligent Reflecting Surface Optimization for MIMO Communication Using Deep Reinforcement Learning

# Intelligent Reflecting Surface Optimization for MIMO Communication Using Deep Reinforcement Learning

Kenneth Ikeagu
*School of Engineering and Physical Sciences*
*Heriot-Watt University*
Edinburgh, United Kingdom
kci1@hw.ac.uk

Yuan Ding
*School of Engineering and Physical Sciences*
*Heriot-Watt University*
Edinburgh, United Kingdom
yuan.ding@hw.ac.uk

Chaoyun Song
*Department of Engineering*
*King's College London*
London, United Kingdom
chaoyun.song@kcl.ac.uk

Muhammad R. A. Khandaker
*Main Roads Western Australia*
Perth, Australia
ruhul.khandaker@mainroads.wa.gov.au

***Abstract*—This paper focuses on the optimization of the phase shifts of an intelligent reflecting surface (IRS) for an IRS-aided multiple input multiple output (MIMO) communication system. Motivated by the massive success of deep reinforcement learning (DRL) algorithms in handling high-dimensional continuous action spaces and tackling non-convex optimization problems, we propose a deep deterministic policy gradient (DDPG) framework for solving the formulated non-convex optimization problem. Numerical simulations demonstrate the robustness and efficiency of the proposed model in terms of spectral efficiency and algorithm run time when compared to a state-of-the-art scheme.**

***Index Terms*—Deep Reinforcement Learning, Passive Beamforming, Intelligent Reflecting Surfaces, MIMO**

## I. Introduction

INTELLIGENT reflecting surface (IRS) technology has been envisioned as a promising technology for future generations of wireless systems. An IRS is composed of a considerable number of reconfigurable passive elements, which are configured in such a way that they reflect incoming signals with a predetermined phase shift [1]. A vital research direction for IRS-aided systems in the literature is to optimize the IRS reflection coefficients in order to maximize the system's spectral efficiency. Consequently, several conventional algorithms have been proposed for designing the IRS reflection coefficients [2]–[5]. However, a common problem associated with such algorithms e.g. the semidefinite relaxation (SDR) method [3], and most conventional alternating algorithms is that they are computationally demanding.

On the other hand, machine learning (ML) techniques especially deep reinforcement learning (DRL)-based methods, seem to address such problems efficiently [6]. Moreso, DRL algorithms are able to tackle non-convex optimization problems effectively. Thus, they have been applied to a large set of problems in the wireless communications domain including but not limited to hybrid beamforming [7], resource allocation [6], secrecy rate maximization [1], and channel estimation [8]. In [9], the authors proposed a DRL-empowered algorithm to optimize the reflection coefficients of an IRS-aided MISO system. Authors in [10] proposed a DRL-based framework to design the digital beamforming and analog phase shift matrices for an IRS-aided MISO interference channel system aiming to maximize the achievable sum rate using the deep deterministic policy gradient (DDPG) framework. The DDPG framework was also recently adopted in [11] to jointly optimize the precoding matrix and IRS matrix in a multi-user MISO system, and in [12] to jointly design the precoding matrices and the IRS phase shift matrix in a multi-user uplink MIMO communication system to enable single data transmission from the users to a base station.

Motivated by the massive success of DRL techniques in tackling varying degrees of problems in the wireless communications domain, in this work, we investigate an IRS-aided downlink MIMO communication system, where a base station transmits multiple data streams to a user with multiple antennas. As far as the author's knowledge, this system model is yet to be investigated using any known DRL-based method. To achieve this, we employ the well-known DDPG algorithm [13] to optimize the IRS reflection coefficients to enable the transmission of multiple data streams with the goal of maximizing the system's spectral efficiency. We compare the performance of our developed model with a benchmark scheme and a MIMO system implemented without an IRS. Simulation results verify the robustness of our proposed model in terms of spectral efficiency and run time.

## II. System Model and Problem Formulation

We consider an IRS-assisted MIMO communication system where a base station (BS) with $T \geq 1$ antenna elements transmits $D_s$ data streams to a user with $R \geq 1$ antennas. The downlink data transmission is enabled via an IRS with

$M$ reflecting elements. Let $\mathbf{H}_{\text{TM}} \in \mathbb{C}^{M\times T}$, $\mathbf{H}_{\text{MR}}^H \in \mathbb{C}^{R\times M}$ and $\mathbf{H}_{\text{TR}}^H \in \mathbb{C}^{R\times T}$ denote the channel coefficients from the BS to the IRS, from the IRS to the user and from BS to user respectively. The equivalent channel from BS to the user is thus given as $\mathbf{H}^H \triangleq \mathbf{H}_{\text{MR}}^H \mathbf{\Phi} \mathbf{H}_{\text{TM}} + \mathbf{H}_{\text{TR}}^H$ where $\mathbf{\Phi}$ represents the phase shift matrix of the IRS modeled as $\mathbf{\Phi} = \text{diag}([e^{j\phi_1}, e^{j\phi_2}, \cdots, e^{j\phi_M}])$. Here $\phi_m \in [0, 2\pi)$ denotes the phase shift of the $m$-th reflecting element of the IRS for $m = 1, \cdots, M$. Thus, the processed received signal can be expressed as:

$$\hat{\mathbf{y}}_b = \sqrt{P}\mathbf{W}\mathbf{H}^H\mathbf{F}\mathbf{s} + \mathbf{W}\mathbf{n}, \tag{1}$$

where $\mathbf{F}$ and $\mathbf{W}$ are the optimum precoding and combining matrices respectively, $\mathbf{s}$ is the $D_s \times 1$ symbol vector which satisfies $\mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \frac{1}{D_s}\mathbf{I}_{D_s}$, and $\mathbf{n}$ denotes the additive white Gaussian noise (AWGN) with variance $\sigma_n^2$. Under this system model, the spectral efficiency is defined as:

$$R_s = \log_2 |\mathbf{I}_{D_s} + \frac{P}{D_s}\varrho_n^{-1}\mathbf{W}^H\mathbf{H}^H\mathbf{F}\mathbf{F}^H\mathbf{H}\mathbf{W}|, \tag{2}$$

where $\varrho_n^{-1} = \sigma_n^2\mathbf{W}^H\mathbf{W}$. Note that, for a fixed $\mathbf{\Phi}$, the optimal $\mathbf{F}$ and $\mathbf{W}$ can be obtained via the SVD of $\mathbf{H}^H$ as $\mathbf{H}^H = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$ [5] where $\mathbf{U} \in \mathbb{C}^{R\times R}$ and $\mathbf{V} \in \mathbb{C}^{T\times T}$ are unitary matrices, and $\mathbf{\Sigma}$ is a strictly positive diagonal matrix. It is possible to approximate $\mathbf{H}^H$ in such a way that only the $D_s$ strongest components are retained, i.e. $\mathbf{H}^H \approx \mathbf{U}_{D_s}\mathbf{\Sigma}_{D_s}\mathbf{V}_{D_s}^H$ where $\mathbf{U}_{D_s} \triangleq \mathbf{U}(:, 1:D_s)$, $\mathbf{\Sigma}_{D_s} \triangleq \mathbf{\Sigma}(1:D_s, 1:D_s)$, and $\mathbf{V}_{D_s} \triangleq \mathbf{V}(:, 1:D_s)$. Thus $\mathbf{F} = \mathbf{V}_{D_s}$ and $\mathbf{W} = \mathbf{U}_{D_s}$. Our goal is to maximize the spectral efficiency in (2) by optimizing the IRS matrix $\mathbf{\Phi}$ subject to the unit modulus constraint. This optimization problem can be formulated as:

$$\begin{aligned}(\text{P1}): \ &\max_{\mathbf{\Phi}} \|\mathbf{H}_{\text{MR}}^H \mathbf{\Phi} \mathbf{H}_{\text{TM}} + \mathbf{H}_{\text{TR}}^H\|^2, \\ &\text{s.t.} |\mathbf{\Phi}_{i,i}| = 1, \quad \forall i = 1, 2, \ldots, M. \end{aligned} \tag{3}$$

To address problem P1, we employ the DDPG algorithm [13] to obtain an optimal $\mathbf{\Phi}$ which is described in more detail in the next section.

## III. Proposed DRL-Based Solution

In this section, we present the basics of reinforcement learning and DDPG, followed by the DDPG-based algorithm to solve (P1) whose framework is shown in Fig. 1.

### A. Basics of RL and DDPG

Reinforcement learning (RL) is a learning approach whereby an agent learns to interact with an environment through trial and error to achieve a specific goal. At time step $t$ in each learning episode, the agent obtains the current state $s_t$ of the environment and receives a reward $r_t$ when it takes an action $a_t$ (from an action space $\mathcal{A}$). The state of the environment is then updated to reflect the new state $s_{t+1}$. The agent takes its actions based on a policy $\pi(a_t|s_t)$. Note that the agent can be punished for selecting actions that lead to undesirable outcomes. Thus, the overall goal of the agent is to explore the environment, learn the optimal policy $\pi(a_t|s_t)$, and take actions that lead to the highest possible reward. The reward function is therefore considered one of the key components in RL and is pivotal to the success of RL algorithms. Q-learning is one of the key algorithms used

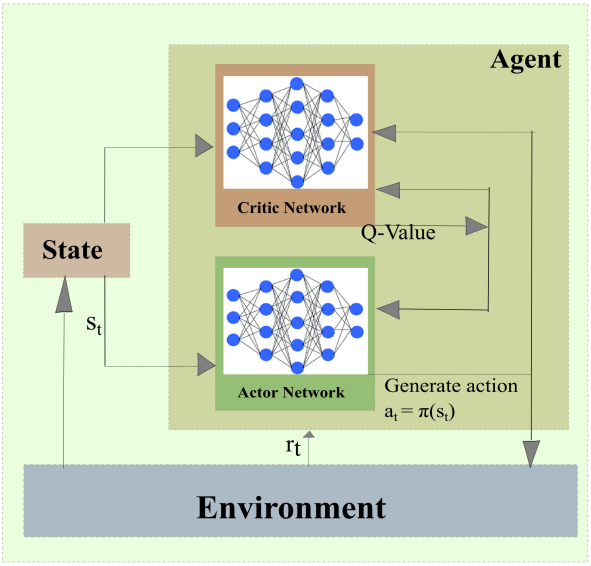


Fig. 1: The DDPG-Based Framework.

in RL. It is a model-free RL algorithm which implies that it does not require the model of the environment but directly learns from experiences. It uses the Bellman equation to update a 'Q-table' [8] which represents the quality of selecting a particular action in a certain state. Significant progress in research has made it possible to combine deep learning with RL resulting in the deep Q network (DQN) [13], which is a value-based algorithm that is capable of tackling problems with high-dimensional observation spaces. However, DQN is only applicable for problems with discrete action spaces and therefore not suitable for addressing the problem (P1) given in (3). Hence, in this letter, we propose to utilize a DDPG-based algorithm to tackle P1.

DDPG is an off-policy model-free RL algorithm that can learn the optimal policy deterministically and is capable of handling high-dimensional continuous action spaces [9]. It can be used for more complex and dynamic environments, unlike DQN. The DDPG algorithm is based on the popular actor-critic algorithm [13] which involves two different neural networks: the actor and critic networks. In a nutshell, the function of the actor network is to select actions $a = \mu(s; \theta_\mu)$ from an action space $\mathcal{A}$, while the critic network uses a Q network, $Q(s, a; \theta_q)$, to evaluate the performance of the selected actions. Here, $\theta_\mu$ and $\theta_q$ are the parameters of the respective networks. DDPG also invokes a target actor network $\mu'(s; \theta_{\mu'})$ and a target critic network $Q'(s, a; \theta_{q'})$ (which are copies of the actor and critic networks) for calculating their target values. Through soft updates [13], the parameters of these target networks are updated according to:

$$\boldsymbol{\theta}_{\mu'} = \tau_\mu \boldsymbol{\theta}_\mu + (1 - \tau_\mu)\boldsymbol{\theta}_{\mu'}, \tag{4}$$

$$\boldsymbol{\theta}_{q'} = \tau_q \boldsymbol{\theta}_q + (1-\tau_q)\boldsymbol{\theta}_{q'}, \tag{5}$$

where $\tau_\mu \ll 1$ and $\tau_q \ll 1$ are the soft update coefficients. To improve its performance over time, DDPG makes use of an experience replay; a technique that allows it to store past experiences in a replay buffer $\mathcal{R}$ as $\{s_t, a_t, r_t, s_{t+1}\}$ which it uses to train the critic network. As a result, the critic network can learn faster and generalize better to new situations. Consequently, the critic network can then sample $B_s$-size random mini-batch $\{s_i, a_i, r_i, s_{i+1}\}$ (where $i = 1, 2, ..., B_s$) from $\mathcal{R}$ to obtain the target Q-Value as [9]:

$$y_i = \begin{cases} r_i, & i = B_s, \\ r_i + \delta Q'\left(s_{i+1}, \mu'\left(s_{i+1}; \boldsymbol{\theta}_{\mu'}\right); \boldsymbol{\theta}_{q'}\right), & i < B_s, \end{cases} \tag{6}$$

where $\delta$ is the discount factor. The loss function of the critic network is obtained as:

$$L\left(\boldsymbol{\theta}_q\right) = \frac{1}{B_s}\sum_{i=1}^{B_s}\left(y_i - Q\left(s_i, a_i; \boldsymbol{\theta}_q\right)\right)^2. \tag{7}$$

### *B. State, Action and Reward function*

The state space $s_t$ at time step $t$ is defined as $s_t = \left[R_s^{(t-1)}, \phi_1^{(t-1)}, \ldots, \phi_M^{(t-1)}\right]$, where $R_s^{(t-1)}$ is the system's spectral efficiency at time step $t-1$, and $\phi_1^{(t-1)}, \ldots, \phi_M^{(t-1)}$ are the IRS phase shifts at time step $t-1$. Under the current channel states, the agent updates the IRS phase shifts using state $s_t$ as inputs. The action space $a_t$ is thus defined as $a_t = \left[\phi_1^{(t)}, \ldots, \phi_M^{(t)}\right]$ where $\phi_1^{(t)}, \ldots, \phi_M^{(t)}$ are the updated IRS phase shifts. Our goal, in this paper, is to maximize the system's spectral efficiency given as $R_s$ in (2). Thus, given the instantenous channels $\mathbf{H}_{\text{MR}}^H, \mathbf{H}_{\text{TM}}$, and $\mathbf{H}_{\text{TR}}^H$, and the action $\phi_1^{(t)}, \ldots, \phi_M^{(t)}$, the reward function is calculated as $r_t = R_s^{(t)}$.

### *C. Algorithm Description*

At the start of the algorithm, the replay buffer $\mathcal{R}$ (with capacity $\mathcal{C}$), the IRS phase shifts, and the parameters of the actor network, critic network, target actor network, and target critic network are all initialized. The IRS phase shifts are initialized by selecting them randomly from 0 to $2\pi$. The algorithm runs for $K$ number of episodes with each episode iterating $T$ number of times. During each episode, the channels $\mathbf{H}_{\text{MR}}^H, \mathbf{H}_{\text{TM}}$, and $\mathbf{H}_{\text{TR}}^H$ are first obtained; the actor network then produces action $a_t$ given state $s_t$, and the agent is rewarded with $r_t$ which is calculated using (2). The state of the environment is thereafter updated to $s_{t+1}$. $s_t$, $a_t$, $r_t$, and $s_{t+1}$ are then stored as one transition into the experience replay buffer $\mathcal{R}$ as $\{s_t, a_t, r_t, s_{t+1}\}$. A $B_s$-size mini-batch is sampled by the critic network from $\mathcal{R}$ to calculate $y_i$ using (6). The critic network's loss function is obtained using (7) and updated via stochastic gradient descent (SGD). Next, the actor network is updated using the policy gradient as [13]:

$$\Delta_{\boldsymbol{\theta}_\mu} J(\mu) = \frac{1}{B_s}\sum_{i=1}^{B_s}\left(\nabla_a Q\left(s_i, \mu\left(s_i; \boldsymbol{\theta}_\mu\right); \boldsymbol{\theta}_q\right) | \nabla_{\boldsymbol{\theta}_\mu} \mu\left(s_i; \boldsymbol{\theta}_\mu\right)\right). \tag{8}$$

Finally, through soft update, the parameters of the target networks (actor and critic networks) are updated according to (4) and (5) respectively. A summary of the algorithm is provided in Algorithm 1.

**Algorithm 1** Proposed DDPG-based Algorithm

**Input:** $\mathbf{H}_{\text{MR}}^H, \mathbf{H}_{\text{TM}}$, $\mathbf{H}_{\text{TR}}^H$, $\delta$, $\tau_\mu$, $\tau_q$, $\mathcal{C}$, $B_s$, and the learning rate $\eta$.

**Output:** The optimal $\boldsymbol{\Phi}$ and $R_s$.

```
Randomly initialize μ(s; θ_μ) and Q(s, a; θ_q).
Initialize μ'(s; θ_μ') and Q'(s, a; θ_q') with θ_μ' = θ_μ and θ_q' = θ_q.
Empty the experience replay buffer R.
for episode i = 1, 2, ..., K do
    Obtain H_MR^H (i), H_TM (i), and H_TR^H (i);
    Randomly choose the elements of matrix Φ^(0) as initial state s_1;
    Adopt the Ornstein-Uhlenbeck process [13] to initialize a noise process N for action exploration;
    for t = 1, 2, ..., T do
        Generate action a_t = μ(s_t; θ_μ) + N;
        Reform a_t into Φ^(t) = diag(e^{jφ_1^(t)}, e^{jφ_2^(t)}, ..., e^{jφ_M^(t)}) to calculate R_s^(t) and acquire the new state s_{t+1};
        Store the experience {s_t, a_t, r_t, s_{t+1}} in R;
        Update target Q-value using (6);
        Calculate the loss function L(θ_q) of the critic network using (7);
        Update μ(s; θ_μ) using (8);
        Update the parameters of the target networks through soft update using (4) and (5);
        Set s_t = s_{t+1}.
    end for
end for
```

## IV. Simulation Results

In this section, we evaluate the performance of our proposed DRL-based algorithm and compare it to the alternating algorithm (AO) [2] and a MIMO system implemented without an IRS. We assume that the BS, the user, and the IRS are located at (0, 0), (12, 0), and (6, 10) in meters (m) in a two-dimensional plane. The channel matrices $\mathbf{H}_{\text{MR}}^H$, $\mathbf{H}_{\text{TM}}$, and $\mathbf{H}_{\text{TR}}^H$ are formulated as the product of the large-scale and small-scale fading as in [4] wherein the path loss is modeled as $P_l = P_{lo} - 10\varepsilon log10(d/d_o)$. We set $P_{lo} = -30dB$, $d_o$ = 1m, and the path loss exponent, $\varepsilon$ = 3. The small-scale fading matrix entries are generated as zero-mean complex Gaussian random variables with unit variance. Unless otherwise specified, we set $T = 8$, $R = 8$, $D_s = 8$, and the noise power at the user as $\sigma_n^2$ = 1. The hyperparameters of the proposed DRL-based algorithm are summarized in Table 1.

Fig. 2 shows the spectral efficiency (SE) versus transmit power for our proposed model in comparison with the AO algorithm in [2]. We demonstrate the performance of the systems by varying the available transmit power at the BS from 20dB to 50dB for cases where $M = 20$ and $M = 150$.

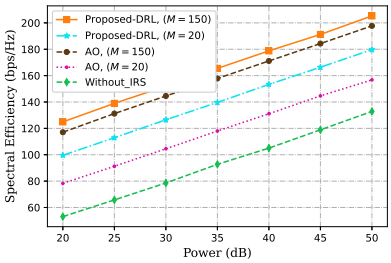


Fig. 2: SE. vs Power.

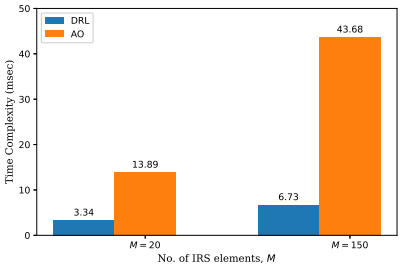


Fig. 3: Average run time for the DRL and AO algorithms for different number of IRS elements.

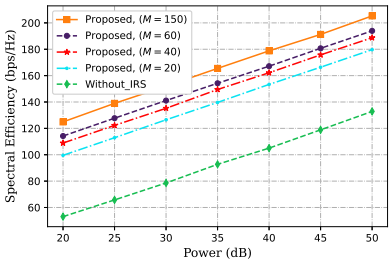


Fig. 4: SE. vs Power for a varying number of IRS reflecting elements, $M$.

TABLE I: The DRL Hyperparameters

| Hyperparameter | Value |
|---|---|
| No. of hidden layers | 2 |
| No. of units in each hidden layer | 300 |
| Activation function of hidden layers | ReLU |
| Output layer activation | tanh(·) |
| Batch size, $B_s$ | 16 |
| No. of episodes, $K$ | 500 |
| No. of iterations per episode, $T$ | 10 |
| Size of Replay Buffer | 100000 |
| Learning rate | 0.001 |
| Optimizer | Adam |
| Discount factor | 0.95 |
| Soft update coefficient | 0.005 |

In the same figure, we show the performance of a MIMO system implemented without an IRS indicated as 'Without-IRS'. We observe that our proposed DRL model achieves the best performance when compared to the other schemes which demonstrates the effectiveness of our model. Besides, we show through Fig. 3 that the average run time of the proposed model is also lower which verifies its robustness and efficiency. Next, we demonstrate the response in performance of the system when implemented with a varying number of IRS reflecting elements as shown in Fig. 4. The results show that the system implemented with the highest number of IRS elements (in this case, when $M = 150$) achieves the best performance which demonstrates the added benefit of implementing such a system with a large number of IRS elements.

## ACKNOWLEDGMENT

This work was supported by the School PhD scholarship from Heriot-Watt University and by the Engineering and Physical Sciences Research Council (EPSRC), UK, under Grant number EP/V002635/1.

## V. CONCLUSION

In this paper, we investigated an IRS-aided downlink MIMO communication system with the aim of maximizing the system's spectral efficiency by optimizing the IRS phase shifts. We proposed a robust and efficient DRL-based approach to optimize the reflecting coefficients subject to the non-convex unit modulus constraint. Our simulation results demonstrated that the proposed model achieves better performance than a state-of-the-art scheme and a scheme implemented without an IRS in terms of spectral efficiency and algorithm run time.

## REFERENCES

[1] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, 2021.

[2] S. Zhang and R. Zhang, "Capacity characterization for intelligent reflecting surface aided MIMO communication," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1823–1838, 2020.

[3] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1410–1414, 2019.

[4] L. Dong and H.-M. Wang, "Secure MIMO transmission via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 787–790, 2020.

[5] S. H. Hong, J. Park, S.-J. Kim, and J. Choi, "Hybrid beamforming for intelligent reflecting surface aided millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7343–7357, 2022.

[6] X. Wang, Y. Zhang, R. Shen, Y. Xu, and F.-C. Zheng, "DRL-based energy-efficient resource allocation frameworks for uplink NOMA systems," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7279–7294, 2020.

[7] Q. Wang, X. Li, S. Jin, and Y. Chen, "Hybrid beamforming for mmwave MU-MISO systems exploiting multi-agent deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 1046–1050, 2021.

[8] K. Kim, Y. K. Tun, M. S. Munir, W. Saad, and C. S. Hong, "Deep reinforcement learning for channel estimation in RIS-aided wireless networks," *IEEE Commun. Lett.*, vol. 27, no. 8, pp. 2053–2057, 2023.

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

[10] J. Zhang, H. Zhang, Z. Zhang, H. Dai, W. Wu, and B. Wang, "Deep reinforcement learning-empowered beamforming design for IRS-assisted MISO interference channels," in *2021 13th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2021, pp. 1–5.

[11] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, 2020.

[12] D. Pereira-Ruisánchez, Fresnedo, D. Pérez-Adán, and L. Castedo, "Joint optimization of IRS-assisted MU-MIMO communication systems through a drl-based twin delayed ddpg approach," in *2022 IEEE Int. Symp. Broadband Multi. Syst. Broadcast. (BMSB)*, 2022, pp. 1–6.

[13] 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," 2019.