IEEE Open Journal of the Communications Society

Received 11 February 2022; revised 24 March 2022; accepted 2 April 2022. Date of publication 7 April 2022; date of current version 21 April 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3165590

Deep Reinforcement Learning Powered IRS-Assisted Downlink NOMA

MUHAMMAD SHEHAB[1] **(Member, IEEE), BEKIR S. CIFTLER**[2] **(Member, IEEE), TAMER KHATTAB**[1] **(Senior Member, IEEE), MOHAMED M. ABDALLAH**[2] **(Senior Member, IEEE), AND DANIELE TRINCHERO**[3]

[1]Department of Electrical Engineering, Qatar University, Doha, Qatar

[2]Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

[3]Dipartimento di Elettronica, Politecnico di Torino, 10129 Torino, Italy

CORRESPONDING AUTHOR: M. J. SHEHAB (e-mail: muhammadjshehab@gmail.com)

This research work was made possible by grant number AICC03-0530-200033 from the Qatar National Research Fund (QNRF). Open access funding was provided by the Qatar National Library (QNL). The statements made herein are the sole responsibility of the authors.

ABSTRACT In this work, we examine an intelligent reflecting surface (IRS) assisted downlink non-orthogonal multiple access (NOMA) scenario intending to maximize the sum-rate of users. The optimization problem at the IRS is quite complicated, and non-convex since it requires the tuning of the phase shift reflection matrix. Driven by the rising deployment of deep reinforcement learning (DRL) techniques that are capable of coping with solving non-convex optimization problems, we employ DRL to predict and optimally tune the IRS phase shift matrices. Simulation results reveal that the IRS-assisted NOMA system based on our utilized DRL scheme achieves a high sum-rate compared to OMA-based one, and as the transmit power increases, the capability of serving more users increases. Furthermore, results show that imperfect successive interference cancellation (SIC) has a deleterious impact on the data rate of users performing SIC. As the imperfection increases by ten times, the rate decreases by more than 10%.

INDEX TERMS Intelligent reflecting surfaces (IRS), non-orthogonal multiple access (NOMA), deep reinforcement learning (DRL), 5G and beyond, 6G, phase shift design.

I. INTRODUCTION

As WIRELESS technologies have grown exponentially over the last few decades, wireless communications are promising to meet the demand for the enormous number of connections. The next-generation networks will be an end-to-end ecosystem to enable a fully connected and sustainable community. The main purpose of these networks is to provide seamless and ubiquitous communications for users with higher throughput, low latency, low energy consumption and support the escalation in mobile data consumption for hundreds of thousands of connections. As 5G networks are being deployed, technologies for 6G networks are being researched and examined to attain more reliable and faster communication systems [1].

Among these technologies, are the intelligent reflecting surfaces (IRS)s, which are considered as key enablers for the next generation networks including beyond 5G (B5G) and 6G communication networks. IRS regulates the wireless environment to boost the energy and spectral efficiencies [2]. IRS consists of a huge number of passive elements or IRS units where each unit can passively reflect the incident electromagnetic wave signal and manipulate it in terms of phase, frequency, amplitude, or polarization [3]. Thus, it enables the electromagnetic waves to be propagated and controlled in an energy-efficient manner [2]. IRS is capable of reconfiguring the wireless propagation environment via a programmable controller that electronically controls the reflective elements to modify the phase of the reflected signals which allows for either constructive or destructive addition of the reflected signals [4], [5]. Most of the research papers in the literature [6]–[12] are considering passive IRS where only a phase shift to the incident signal is applied. Thus, the IRS will not consume any transmit power. Consequently, the IRS optimization problem is focused on optimizing the phase shift matrix.

Furthermore, the IRS serves as a relay between the transmitters and receivers, especially where there is no line of sight (LOS) between the transmitting antenna and the receiving antenna, or if the direct link suffers from shadowing and deep fading rendering the quality of the channel for direct communications unreliable. Thus, deploying IRS aids the BS to serve multiple users, and improves the system sum-rate significantly [13]. IRS is a transformational relaying technology that solely relies on large number of signal reflecting elements that are able to collect wireless signals from the transmitter and passively reflect these incident signals independently with an adjustable phase shift [14]. IRS received a lot of attention because of its ability to act as a controller of the wireless propagation environment as opposed to the classical approach where the wireless channel was an imposed component in the communication system. Every IRS element can be intelligently controlled to modify the phase of the incident electromagnetic wave signal. Therefore, it renders the strength and direction of the signal highly controllable at the receiver side. This merit can be employed to add various signals constructively at the receiver to improve the sum-rate. Therefore, IRS can be utilized to optimize the signal to interference plus noise ratio (SINR), coverage probability, and data rate [15]. Compared to decode and forward and amplify and forward, IRS demands less energy and power consumption because of its passive features. Thus, the energy and rate efficiency of the wireless communication channels can be optimized significantly [16]. Hence, IRS is anticipated to be a promising solution for future communication systems [17]. Compared to massive or large multiple input multiple output (M-MIMO) antenna networks that utilize a large number of antennas to increase the energy and spectrum efficiency, IRS is considered a potential energy-efficient component for 6G networks by regulating the propagation in the wireless environment [7].

Recently, non-orthogonal multiple access (NOMA) has drawn a great amount of attention in many scenarios in 5G wireless networks because of the high spectrum efficiency it provides in addition to its support for massive connectivity [18]–[20]. In the previous cellular systems, many multiple access technologies were adopted such as the time division multiple access (TDMA), frequency division multiple access (FDMA), spatial division multiple access (SDMA), and orthogonal frequency division multiple access (OFDMA). Based on their design, these technologies are considered orthogonal multiple access (OMA) techniques, since the wireless resources are allocated to multiple users orthogonally. The users are separated in the chosen access domain whether it is in frequency, time, or space. If orthogonality is violated, the users will suffer from interference, and the quality of communication links will degrade leading to loss of information and/or inefficient resources utilization. Nonetheless, OMA schemes cannot satisfy the massive connectivity requirements for future communication systems which causes the need for NOMA [21]. NOMA achieves a high sum-rate capacity as compared to the traditional orthogonal multiple access (OMA) techniques. The reason is that it enables multiple users to transmit simultaneously in the same set of shared resources. This results in an interference, but NOMA utilizes a method called successive interference cancellation (SIC) to eliminate the resulting interference [18].

A. RELATED WORK

Many research studies based on IRS in 5G and 6G networks are being conducted [22]–[26], but all of these studies are based on OMA scheme. However, several recent research studies [27]–[32] started investigating IRS for NOMA networks to increase the system performance. The authors in [31] examined an uplink scenario for IRS NOMA to maximize the sum-rate for all users taking into consideration the power constraint for each user. The non-convex problem was solved using semi-definite relaxation (SDR) that provides a near-optimal solution. In this study, the authors assumed that the channel state information (CSI) of all channels is perfectly known at the BS and IRS. Furthermore, the authors in [32] maximized the total signal power received at the user side by optimizing the transmit beamforming at the access point (AP) and reflect beamforming at the IRS. They proposed an algorithm based on SDR to solve the non-convex problem, which obtains a sub-optimal solution and assumes that the CSI is known at the IRS. Moreover, machine learning techniques were employed in [24] and specifically DRL, which have much lower complexity than SDR, to solve the non-convex problem. They inspected the scenario of IRS for multiple-input single-output (MISO) systems to optimize the phase shift matrix at the IRS to maximize the signal-to-noise ratio (SNR), and the DRL-based scheme almost achieved the upper bound of the received SNR. However, the scheme did not include NOMA, and the channels were assumed to be available at both the BS and IRS.

Nonetheless, the above studies [27]–[32] utilized mathematical methods to solve their problems, and they assumed that the channels between the IRS and users are known. Such assumption contradicts the practical case, where IRSs are passive elements incapable of perfectly estimating channels. In the case where we have limited channel state information (CSI), the use of machine learning techniques can add value to the problem.

B. CONTRIBUTIONS

In our research work, we address the aforementioned gap in the surveyed literature by leveraging reinforcement learning (RL), in particular, deep reinforcement learning (DRL), to optimize the sum-rate of a NOMA downlink system utilizing IRS. The major challenge in our scenario while optimizing the phase shifts at the IRS lies in the unit modulus constraints (because the IRS can reflect the signal without amplifying it), which are fundamentally non-convex. Therefore, the problem is an NP-hard problem, and it is not easy to obtain an optimal solution in closed form [12], [24], [31]. Moreover, in our system we take into consideration the practical situation where the instantaneous CSI of the channel between the IRS and the users is unknown, while its long-term average is known. In particular, we exploit Deep Deterministic Policy Gradient (DDPG) due to its suitability for our scenario to investigate the phase shift design and to tackle the non-convexity caused by the constant modulus constraints. DDPG algorithm, which is a DRL technique, is very efficient when coping with complex non-convex, and intractable optimization problems. It eliminates the need for gathering a large dataset for training, and it provides a solid and robust performance. Simulation results reveal that DDPG powered IRS NOMA outperforms the IRS OMA-based one in terms of sum-rate and approaches the upper bound.

Our contributions in this work can be summarized as follows:

- First, we formulate the IRS NOMA downlink phase shift optimization problem to maximize the sum-rate for NOMA users taking into consideration the limited CSI knowledge (we only assume knowledge of first and second order statistics of CSI) between the IRS and users, which prevents SIC order adaptation and NOMA power adaptation.
- Second, we incorporate imperfect interference cancellation in practical NOMA within our system model formulation.
- Third, DDPG based solution is proposed for predicting the best phase shift matrix in which the IRS learns the best way for reflecting the incident signals by modifying the phase.
- Fourth, numerical results reveal the effectiveness of the DDPG algorithm since the sum-rate value for the DDPG based IRS-assisted NOMA system produces sum-rates almost close to a discretized exhaustive search upper-bound approximation. Moreover, the studied DDPG based IRS-assisted NOMA outperforms the IRS OMA scheme system with minimum training overhead.

C. PAPER ORGANIZATION

The rest of the paper is structured as follows. Section II introduces the problem formulation and outlines the system model along with practical considerations. Section III explains the proposed solution using DDPG-based phase control for IRS. Section IV discusses the performance benchmarking. In Section V the numerical results are presented, while the paper is concluded in Section VI.

II. PROBLEM STATEMENT

A. SYSTEM MODEL

We consider the downlink of an IRS-assisted NOMA system with K users as shown in Fig. 1. All users have a single antenna each as well as the BS. Without loss of generality, the users are ordered according to their distance from the IRS such that user 1 is the farthest user from the IRS and user k is the nearest user to the IRS. Consequently, the users can be considered as ordered based on the expected value of their channel gains assuming $|\mathbf{h}_{r,1}|^2 < |\mathbf{h}_{r,2}|^2 < \cdots < |\mathbf{h}_{r,k}|^2$. The users are assumed to be moderate to slowly move such that their ordering (ordered distances from the IRS) is quasi-static (i.e., changes slowly with time) and as this happens the system will trigger power allocation reconfiguration and adapt accordingly. There is no direct LOS link between the users and the BS. Thus, the communication between the BS and users is performed through the IRS which is deployed with $M = M_x M_y$ reflecting elements, where M_x and M_y represent the number of passive elements in the IRS in every row and column, respectively. Further, in the considered system, the following practical aspects are assumed:

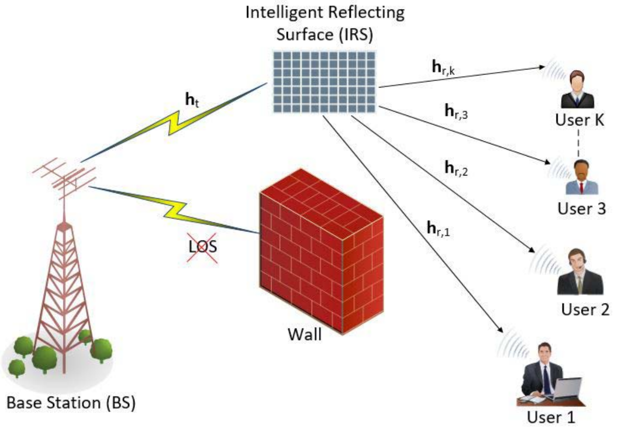

FIGURE 1. IRS assisted downlink NOMA, where the wireless communication between the BS and the endusers is accomplished through the IRS.

1) We are using DRL with IRS for both equal power allocation OMA and average-based power allocation NOMA while instantaneous CSI is difficult to attain with the existence of IRS. The power allocation factors for NOMA, β_i, are allocated such that $\beta_1 > \beta_2 > \cdots > \beta_K$, where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_k]$ is the vector of coefficients of the users' power allocated such that $\beta_1 + \beta_2 + \cdots + \beta_k = 1$. When we mention power allocation in NOMA configuration, we are referring to power allocation to the different users according to their distances from the IRS, which depends on the long-term average of the random channel gains instead of using instantaneous CSI.
2) Users perform SIC based on the long-term channel statistics since we have limited knowledge about the instantaneous CSI between the IRS and users.
3) The end-to-end channel (equivalent base station to user channel combining the effects of $\mathbf{h}_t$ and $\mathbf{h}_{r,k}$) channel can still be estimated by the base station[1] [33].

Equivalently, our system model operation can be thought of as using static power allocation while optimizing the NOMA sum-rate performance through adapting the channel

[1] The overall channel estimation can be performed through pilot transmission by BS/user since both are active nodes.

(using IRS phases) to compensate for random channel variations as well as compensating for using long-term SIC ordering instead of the classical approach which changes the SIC order with every channel realization.

The transmitted signal at the BS is

$$x = \sum_{k=1}^{K} \sqrt{P_k} s_k, \tag{1}$$

where s_k represents the signal for user k with unit power (i.e., $\mathbb{E}[|s_k|^2] = 1$, $k \in \{1, \dots, K\}$, and $\mathbb{E}[.]$ denotes the expectation). The power allocated to each user is $P_k = \beta_k P$, where P is the BS total transmit power. The power allocation satisfies the following relationship: $P_1 > P_2 > \cdots > P_K$ [34].

The received signal for user k can be written as:

$$\begin{aligned} y_k &= \mathbf{h}_{r,k}^H \Phi \mathbf{h}_t x + n, \\ y_k &= \mathbf{h}_{r,k}^H \Phi \mathbf{h}_t \sum_{k=1}^{K} \sqrt{P_k} s_k + n, \end{aligned} \tag{2}$$

where $\mathbf{h_t} \in \mathbb{C}^{M\times 1}$ is the channel between the BS and the IRS, $\mathbf{h}_{r,k} \in \mathbb{C}^{M\times 1}$ is the channel between the IRS and users, and $\Phi = \text{diag}(e^{j\theta_1}, e^{j\theta_2}, \dots, e^{j\theta_M})$ is the phase shift reflection matrix that satisfies the constant modulus constraint $|\phi_i|^2 = |e^{j\theta_i}|^2 = 1$, $\forall i \in \{1, 2, \dots, M\}$, because the IRS reflects the signal without amplifying it, and diag(.) denotes the diagonal matrix. Further, the phase shift of the i^{th} passive reflecting element is denoted by θ_i, where the value of θ_i is between 0 and 2π, and $n \sim \mathcal{CN}(0, \sigma^2)$ represents the additive white Gaussian noise (AWGN). Both channels follow the Rician fading model:

$$\mathbf{h_t} = \sqrt{\frac{K_1}{K_1+1}}\bar{\mathbf{ht}} + \sqrt{\frac{1}{K_1+1}}\tilde{\mathbf{ht}}, \tag{3}$$

$$\mathbf{h_{r,k}} = \sqrt{\frac{K_2}{K_2+1}}\bar{h}_{\mathbf{r,k}} + \sqrt{\frac{1}{K_2+1}}\tilde{\mathbf{h}}_{\mathbf{r,k}}, \tag{4}$$

where K_1 is the rician factor of $\mathbf{h_t}$, $\bar{\mathbf{ht}} \in C^{M\times 1}$ and $\tilde{\mathbf{ht}} \in C^{M\times 1}$ are the LoS component and non-LoS (NLoS) component, respectively. Similarly, K_2 is the rician factor of $\mathbf{h_{r,k}}$, $\bar{h}_{\mathbf{r,k}} \in C^{M\times 1}$ and $\tilde{\mathbf{h}}_{\mathbf{r,k}} \in C^{M\times 1}$ are the LoS component and NLoS component, respectively.

Hence, the received SINR at user k can be represented by the following equation:

$$\gamma_k = \left(\frac{\left|\mathbf{h}_{r,k}^H \Phi \mathbf{h}_t\right|^2 P_k}{\left|\mathbf{h}_{r,k}^H \Phi \mathbf{h}_t\right|^2 \sum_{i=k+1}^{K} P_i + \sigma^2} \right). \tag{5}$$

We note that when $k = K$, the term $|\mathbf{h}_{r,k}^H \Phi \mathbf{h}_t|^2 \sum_{i=k+1}^{K} P_i$ is equal to 0.

Furthermore, the data rate of user k is represented by:

$$R_k = \log_2(1 + \gamma_k). \tag{6}$$

The feedback and signaling model for our IRS NOMA system is shown in Fig. 2, where the BS allocates power $P_k = [P_1, P_2, P_3, \dots, P_k]$ for users. The users estimate and send the rates to the IRS which in turn act as the DRL agent in our system. The DRL agent in turn will calculate the reward and adjust the phase accordingly. In fact, our DRL model is tracing the variation of the magnitude of the channel by continuously adjusting the phases of the IRS to maximize the sum-rate. Thus, the system model is based on utilizing DRL to adjust the phases of the IRS according to the total sum-rate fed back from the users to learn and reach optimal sum-rate tuning with limited CSI.

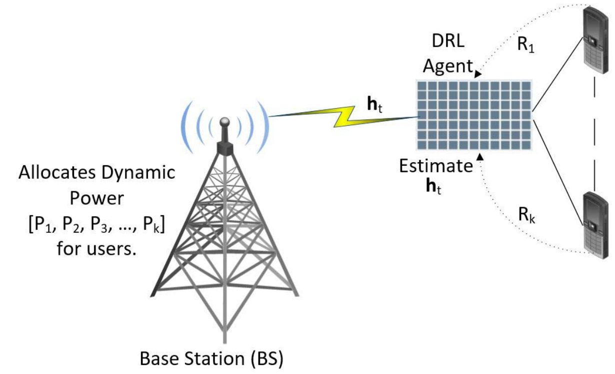

FIGURE 2. IRS NOMA feedback and signaling model.

B. PROBLEM FORMULATION

Our objective in this work is to find the values of the phase shifts of the IRS elements that maximize the sum-rate of all users given by

$$R_{sum} = \sum_{k=1}^{K} \log_2(1 + \gamma_k). \tag{7}$$

However, equation (7) is valid for perfect successive interference cancellation (SIC), which is an ideal case. Perfect SIC can happen only under two assumptions: (i) perfect channel state information is available and (ii) perfect decoding of information is possible. Both assumptions are impractical. A more practical scenario is to assume that each stage of interference cancellation leaves a residual fraction, $0 \le \epsilon \ll 1$, of the interfering signal after cancellation [35]. In our system, NOMA power allocation and SIC rely on the distance (long term channel average) while the DRL based tuning of IRSs is used to reconfigure the channel to satisfy the optimal NOMA performance. Thus, the received SINR in (5) at user k can be rewritten as

$$\tilde{\gamma}_k = \frac{|\mathbf{h}_{r,k}^H \Phi \mathbf{h}_t|^2 P_k}{\left|\mathbf{h}_{r,k}^H \Phi \mathbf{h}_t\right|^2 \left(\epsilon \sum_{j=1}^{k-1} P_j + \sum_{i=k+1}^{K} P_i\right) + \sigma^2}. \tag{8}$$

When $k = 1$, the term $|\mathbf{h}_{r,k}^H \Phi \mathbf{h}_t|^2 \sum_{j=1}^{k-1} P_j = 0$, and when $\epsilon = 0$, $\tilde{\gamma}_k = \gamma_k$ which is the ideal case.

Accordingly, the formulated problem at the IRS is to obtain the phase shift reflection matrix Φ that maximizes

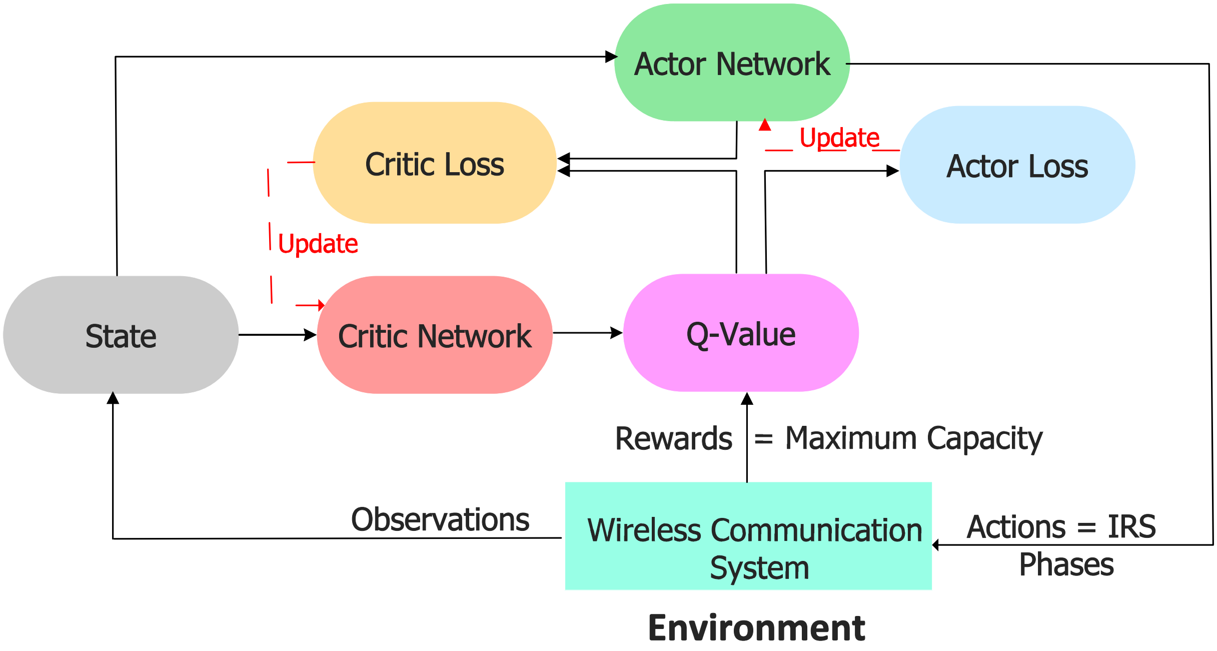

FIGURE 3. DDPG model.

R_{sum}, can be written as

$$\begin{aligned} &\max_{\Phi} \sum_{k=1}^{K} \log_2(1+\tilde{\gamma}_k), \\ &s.t. \;\; \text{C1: } |\phi_i|^2 = 1, \forall i \in \{1,2,\ldots,M\}, \\ &\qquad \text{C2: } \sum_{k=1}^{K} \beta_k = 1, \\ &\qquad \text{C3: } \beta_k \geq 0, \end{aligned} \tag{9}$$

where constraint C1 represents the use of IRS with phase adaptation only, C2 represents the fact that the summation of the power of all users equals the power transmitted by the BS, and C3 represents the fact that the powers cannot be negative. As stated earlier in the system model, the considered system adopts static power allocation factors based on long term channel statistics. Therefore, constraints C2 and C3 are redundant and can be removed from the optimization problem.

The optimization problem in (9) can be alternatively viewed as finding the phase shift reflection matrix Φ that maximizes the SINR $\tilde{\gamma}_k$ as follows [31]:

$$\begin{aligned} &\max_{\Phi} \sum_{k=1}^{K} \left|\mathbf{h}_{r,k}^{H} \Phi \mathbf{h}_t\right|^2 P_k, \\ &s.t. \;\; |\phi_i|^2 = 1, \forall i \in \{1,2,\ldots,M\}. \end{aligned} \tag{10}$$

This is an NP-hard problem because of the non-convexity of the constant modulus constraint and the objective function. The authors in [31] and [32] employed the SDR method to solve this problem which provides a near-optimal solution. However, the complexity of the SDR is of $\mathcal{O}(M+1)^6$ which is prohibitively expensive [12]. Thus, taking into consideration the non-convexity of the optimization problem, the practical assumptions that we have limited knowledge about the CSI for the channels between the IRS and the users, $\mathbf{h}_{r,k}$, as well as the imperfect interference cancellation, rendering solving the optimization problem in (10) analytically non-tractable. Considering the dynamics of the system, we propose the use of reinforcement learning to find the optimal values of the phase shifts of the IRS which maximizes the overall sum-rate.

III. PROPOSED DDPG-BASED IRS PHASE CONTROL

In this section, we propose a DDPG-based IRS phase control method (see Fig. 3), considering the optimization problem in (9). DDPG is a model-free reinforcement learning technique that combines the advantages of policy gradients and Q-learning. Considering that the states in our system are mainly dependent on channel gains and output sum rate, while the actions are the IRS phase shift, we are considering a continuous sate and continuous action system. DDPG's main advantage lies in the fact that it uses both the continuous action and state spaces [36].

DDPG consists of four neural networks; one for actor-network, one for critic network, one for target actor-network, and one for target critic network, which ensures stability. The optimization problem in (9) can be solved using DDPG by learning the policy.

A. RL SYSTEM MAPPING

The first step in solving a problem using RL is to map the problem into the key components of an RL system; namely, state-space, action-space, and reward function. In the following, we discuss this mapping as well as the general behavior of the RL method using DDPG.

1) STATE-SPACE

The state-space of the DDPG agent at timestep (t) can be defined as follows

$$\mathbf{s}^{(t)} = \left[\mathbf{h}_t^{(t)},\ \mathbf{\Phi}^{(t-1)},\ \hat{\boldsymbol{\gamma}}^{(t-1)}\right], \tag{11}$$

where $\mathbf{h}_t$ represents the channel gain between the source which is the BS and the IRS which acts as the agent in the environment, $\mathbf{\Phi}$ is the last phase action taken by IRS, and $\hat{\boldsymbol{\gamma}}$ is the estimated SINR values of the users based on their data rates for that action (i.e., $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_k, \ldots, \gamma_K]$).

2) ACTION-SPACE

The actions are the IRS phase shift values. The real values of the actions coming from the neural network are used as the argument value of a complex exponential representing the actual phase. The output is an array that defines the phase of each element in the IRS. Thus, the action-space is defined by the following policy function:

$$\mathbf{a}^{(t)} = \mu\left(\mathbf{s}^{(t)}|\theta^{\mu}\right) + \mathbf{n}(t) \tag{12}$$

where μ is the policy function and θ^{μ} is parameters (i.e., weights of neural network), and $\mathbf{n}(t)$ is the Ornstein-Uhlenbeck (OU) process-based action noise [37].

3) REWARD FUNCTION

The reward function is defined based on the current channel capacity and the maximum capacity ever reached as follows:

$$r^{(t)} = R_{sum}^{(t)} - R_{sum,max}, \tag{13}$$

where $R_{sum}^{(t)}$ is the actual sum-rate of the users, while $R_{sum,max}$ is the maximum sum-rate achieved.

4) EXPLORATION VS. EXPLOITATION

Since the action-space of the DDPG is continuous, the exploration of action-space is handled with noise generated by the OU process. OU process samples noise from a correlated normal distribution.

5) DDPG ALGORITHM

The agent acquires the current CSI of the transmitter to IRS channel via feedback of estimated channels by the transmitter (base station) to the agent residing at the IRS. It does not have access to CSI for the IRS to receiver channel (unknown instantaneous CSI) and compensates this using the fed back SNR as an indirect indicator. The goal of the DDPG algorithm is to train the agent IRS to take actions that maximizes the long-term average reward (sum rate) coping with changes of unknown environments. The agent basically learns how to adjust its randomized policy such that it copes with the random statistical behaviour of the environment to maintain a long-term average reward (equivalently sum-rate) maximization. It is not meant to provide an optimal instantaneous response to the instantaneous random changes of the channel. In other words, the agent tries to learn the statistical model of the random environment and adjust its own statistical model of the response policy to provide long-term reward maximization.

Algorithm 1 DDPG-Based IRS Phase Control Training

```
1: Initialization: Set $t = 0$ and initialize reply buffer of DDPG agent $\mathcal{D}$ with capacity M.
2: Randomly initializes the weights of actor networks $\theta^{\mu}$ and critic networks $\theta^{Q}$.
3: Initialize target networks: $\theta^{\mu'} \leftarrow \theta^{\mu}$ and $\theta^{Q'} \leftarrow \theta^{Q}$.
4: for $t = 1$ to $\infty$ do
5:     Observe state $\mathbf{s}^{(t)}$ and select an action with exploration OU noise $\mathbf{a}^{(t)} = \mu(\mathbf{s}^{(t)}|\theta^{\mu}) + \mathbf{n}_t$
6:     Execute action $\mathbf{a}^{(t)}$ at IRS.
7:     Receive the immediate reward $r^{(t)}$, and observe next state $s^{(t+1)}$, store transition $(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}, r^{(t)}, \mathbf{s}^{(t+1)})$ in $D$.
8:     Randomly sample mini-batch transitions from $\mathcal{D}$:
         $B \leftarrow \{(\mathbf{s}^{(i)}, \mathbf{a}^{(i)}, r^{(i)}, \mathbf{s}^{(i+1)})\} \in \mathcal{D}$ .
9:     Compute the targets:
         $\tilde{Q}(\mathbf{s}^{(i)}, \mathbf{a}^{(i)}|\theta^{Q'}) = r^{(i)} + \Gamma Q(\mathbf{s}^{(i+1)}, \mu(\mathbf{s}^{(i)}|\theta^{\mu'})|\theta^{Q'})$
10:    Update the $\theta^{Q}$ in critic network by minimizing the loss:
         $L = \frac{1}{|B|}\sum_{i=1}^{|B|}\left(\tilde{Q}(\mathbf{s}^{(i)}, \mathbf{a}^{(i)}|\theta^{Q'}) - Q(\mathbf{s}^{(i)}, \mathbf{a}^{(i)}|\theta^{Q})\right)^2$
11:    Update the $\theta^{\mu}$ in actor network according to the sampled policy gradient:
         $\nabla_{\theta^{\mu}}\boldsymbol{J} \approx \frac{1}{|B|}\sum_{i=1}^{|B|}\nabla_{a}Q(\mathbf{s}^{(i)}, \mathbf{a}^{(i)}|\theta^{Q})\nabla_{\theta^{\mu}}\mu(\mathbf{s}^{(i)}|\theta^{\mu})$
12:    Update the target networks:
         $\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}$
         $\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}$
13: end for
```

For every iteration, the agent IRS observes the state which includes the transmitter channel $\mathbf{h}_t^{(t)}$, the last action $\mathbf{\Phi}^{(t-1)}$, and the last estimated SINR $\hat{\boldsymbol{\gamma}}^{(t-1)}$, and calculates the action $\mathbf{\Phi}^{(t)}$ which maximizes the long term reward as stated in (13); therefore, the total sum-rate. This is done by the actor network, while the critic network accepts the state and the action and generates the prediction of the total sum-rate. After the total sum-rate is received from the users, a new state is observed, and the IRS will adjust the phases accordingly. Based on the total sum-rate fed back from the users, the IRS will modify the policy parameters (θ^{μ}) used to calculate the action till the system learns how to reach the optimal sum-rate tuning with limited CSI. The actor network, critic network, target actor network, and target critic network have the same structure and parameterization. To improve stability, the target networks will be updated periodically according to the newest actor and critic parameter values.

As shown in Algorithm 1, we begin initializing the replay buffer D of the agent with transaction capacity M in step 1. In step 2, we initialize the weights of actor and critic networks for the agent. The target networks are initialized by copying the same weights in step 3. The steps from 4 to 13 are repeated for every iteration (i.e., timestep t). Thus, in each iteration (i.e., timestep t), we observe the state $\mathbf{s}$ for the agent (IRS) and determine an action (i.e., phase shift value)

Algorithm 2 Exhaustive Search for the Phase Shift Matrix

```
Initialize M = 4, $\Delta\Phi = \frac{2\pi}{30}$,
for $\phi_1 = 0 : \frac{2\pi}{30} : 2\pi$; do
    for $\phi_2 = 0 : \frac{2\pi}{30} : 2\pi$; do
        for $\phi_3 = 0 : \frac{2\pi}{30} : 2\pi$; do
            for $\phi_4 = 0 : \frac{2\pi}{30} : 2\pi$; do
                Calculate and store $R_{sum}(\phi_1, \phi_2, \phi_3, \phi_4)$
            end for
        end for
    end for
end for
Find $\Phi^* = \text{argmax}_{\phi_1,\phi_2,\phi_3,\phi_4} R_{sum}$
```

with exploration noise based on the OU process as shown in step 5. In step 6, the agent determines and executes the action. After that, a reward $r^{(t)}$ is received, a new state $\mathbf{s}^{(t+1)}$ is observed, and the transactions are stored in respective replay memories as explained in step 7. Then a random mini-batch of transitions is sampled in step 8. By using Bellman's equation, the actor and critic networks' targets are computed as shown in step 9. Using the computed targets, the critic network weights are updated in step 10 by minimizing the loss. The actor-network weights are updated for the sampled policy gradient in step 11. Finally, in step 12, the agent's target networks are updated using the update rate (τ) to increase the stability in the learning process.

6) NEURAL NETWORK ARCHITECTURE

DDPG agent's architecture consists of 4 neural networks including actor and critic networks and target actor and critic for stability. Both actor and critic networks consist of 2 hidden layers, with 256 hidden nodes in each layer. The actor-network input has the size of $2M + K$, and the output is M, thanks to the continuous definition of DDPG. As can be seen from DDPG agents' structure, it allows scalability to a much larger extent with linearly increasing complexity.

B. DISCUSSION ON COMPLEXITY

To reveal the value of DDPG, we provide a quantitative analysis of the exhaustive search algorithm complexity $\mathcal{N}_E$, versus the complexity of the proposed DDPG based algorithm $\mathcal{N}_\mathcal{D}$. The complexities can be easily deduced from the description of the algorithms given in Algorithm 1 and Algorithm 2. For exhaustive search assuming K users, M IRS elements and $N = \frac{2\pi}{\Delta\Phi}$ phase change steps, we can write the complexity as

$$\mathcal{N}_E = O\Big(K \times N^M\Big). \tag{14}$$

For the DDPG based system, the complexity for the trained network (steady-state complexity) depends mainly on the forward network architecture (actor-network). Assume the number of states (size of (actor-network) input) is S, the number of hidden layers is n, the number of neurons in each hidden layer is U, the number of actions (i.e., phase of each IRS element) which is the size of the output layer A, and the DDPG algorithm will always provide the action of the highest reward for the A distinct actions as output. Therefore, the complexity of the DDPG can be written as

$$\mathcal{N}_D = O(S \times n \times U \times A). \tag{15}$$

Thus, the complexity of DDPG is much lower than that of the exhaustive search as the number of users or the number of IRS elements increases.

IV. PERFORMANCE BENCH MARKING

To measure the performance of our proposed scheme, we need reference systems to compare with their performance. To the best of our knowledge, there is no existing tight upper bound theoretical limit on the considered system in literature and the mathematical derivation of an upper bound is cumbersome. Therefore, we provide two benchmarking reference models; one is based on using exhaustive search across a discretized grid of the IRS phases, which acts as an approximated upper bound, while the other is considering orthogonal multiple access which can act as a lower bound on performance.

A. UPPER BOUND ON PERFORMANCE

To measure the performance of the DDPG algorithm and to verify that our sum-rate values approach the upper bound, a discretized exhaustive search method is used to search for the optimum phase shift matrix that results in an approximation to the maximum sum-rate as shown in Algorithm 2. Further, to avoid the huge complexity of the exhaustive search scheme, we consider a limited number of IRS reflecting elements as case proof that our DDPG algorithm can track the upper bound. For every IRS element, we consider the phases between 0 and 2π with a step size of $2\pi/30$, this will give us 30^M combinations of phase shift matrices. Then we calculate the sum-rates accordingly for K users.

B. OMA BASELINE SCHEME

The signal model for OMA is assumed such that the resources (frequency/time) are divided equally between the K users. This enables OMA users to receive the signal with free interference, whereas the merit of NOMA is the simultaneous transmission and the interference can be controlled. However, to serve K OMA users, FDMA / TDMA requires K time slots. The first user will use the first frequency/time slot, the second user will use the second frequency/time slot, and user K will use the K^{th} frequency/time slot accordingly.

The transmitted signal by the BS is given by:

$$x_k^{OMA} = \sqrt{P} s_k, \tag{16}$$

The received signal at the user side can be expressed as:

$$y_k^{OMA} = \sqrt{P} s_k \mathbf{h}_{r,k}^H \Phi \mathbf{h}_t + n, \tag{17}$$

Hence, the received SNR at user k can be represented as:

$$\gamma_k = \left(\frac{\left| \mathbf{h}_{r,k}^H \Phi \mathbf{h}_t \right|^2 P}{\sigma^2} \right), \tag{18}$$

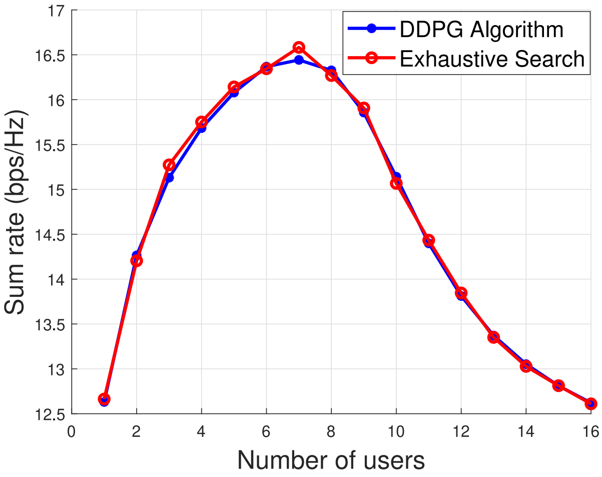

FIGURE 4. Upper bound on performance vs proposed DDPG algorithm. M = 4, K = 16, and $\Delta\Phi = \frac{2\pi}{30}$.

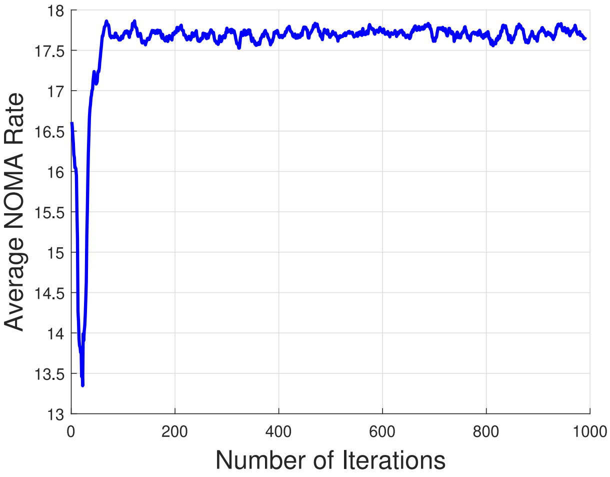

FIGURE 5. NOMA sum-rate vs iteration plots.

Further, data rate of user k is represented by:

$$R_k^{OMA} = \frac{1}{K} log_2\left(1 + \gamma_k^{OMA}\right), \tag{19}$$

Therefore, the sum-rate of OMA can be expressed as:

$$R_{sum}^{OMA} = \sum_{k=1}^{K} R_k^{OMA}, \tag{20}$$

$$R_{sum}^{OMA} = \frac{1}{K} \sum_{k=1}^{K} \log_2\left(1 + \frac{\left|\mathbf{h}_{r,k}^H \Phi \mathbf{h}_t\right|^2 P}{\sigma^2}\right). \tag{21}$$

V. NUMERICAL RESULTS

In this section, we first evaluate the performance of the DDPG algorithm to make sure that the sum-rate values calculated are close to the upper bound. By using the exhaustive search algorithm, we calculate the maximum sum-rate by obtaining the optimum phase shift matrix assuming that the channel is known. The exhaustive search scheme is highly complex, so the number of IRS reflecting element used is $M = 4$ rather than $M = 16$. This is to verify that our DDPG algorithm can approach the upper bound. For each element, we consider the phases between 0 and 2π with a step size of $\frac{2\pi}{30}$. Thus, the total number of combinations of phase shift matrices is 30^4. The sum-rates are calculated for 16 users and Monte-Carlo simulations equal to 1000.

Fig. 4 reveals that the NOMA sum-rate generated by the DDPG algorithm approaches the upper bound and it is close to optimal. The complexity of the exhaustive search algorithm can be calculated as 1000 x 16 x 30^4 which is equal to 1.2960×10^{10} iterations with an elapsed time equal to 124.86 hours.

Moreover, the result in Fig. 5 verifies the convergence of our DRL algorithm. It shows the average NOMA rate versus the iteration plots. The average rate is increasing with time, which means that the training process is conducted successfully. Further, the simulation results below reveal the performance of our DRL-based IRS NOMA system with IRS reflecting elements $M = 16$. The default parameters used in the simulation are shown in Table 1. The number of BS antennas is $N_t = 1$, the number of antennas per user is $N_r = 1$, the distance between the BS and the IRS is 50 m and the distances between the IRS and the users are randomly generated between 200 and 1500 m. The channel between the BS and the IRS and the channel between the IRS and users follow the Rician fading model with rician factor $K1 = K2 = 10$. However, the channel between the BS and the IRS is assumed to be perfectly estimated, whereas the channels between the IRS and users are assumed to have limited CSIs. The bandwidth is 10 MHz, the BS transmit power Pt is 40 dBm, and the noise power spectral density equals -174 dBm/Hz. Simulation results are generated using 10^3 Monte-Carlo runs.

TABLE 1. Parameters used in simulation.

Simulation Parameters	Values
Number of Reflecting Elements (M)	16
Number of BS antennas N_t	1
Number of antennas per user N_r	1
Distance between BS and IRS	50
Distance between the IRS and the users	200 - 1500
BS transmit power	40 dBm
Bandwidth	10 MHz
Noise power spectral density	-174 dBm/Hz
BS to IRS Path loss exponent	2
IRS to users Path loss exponent	2.8
Rician Factor	10
Critic learning rate	0.001
Actor learning rate	0.0005
Discount factor Γ	0.99
Coefficient of Soft Updates τ	0.05
Batch size	64
Buffer Capacity C	10000

In the proposed DDPG algorithm, the actor and critic networks are both dense neural networks (DNN). The input of the actor-network is the number of states that contains 128 neurons, while the output is the number of actions that contains 16 neurons. The hidden layers in the actor-network are two layers that contain 256 neurons each, followed by

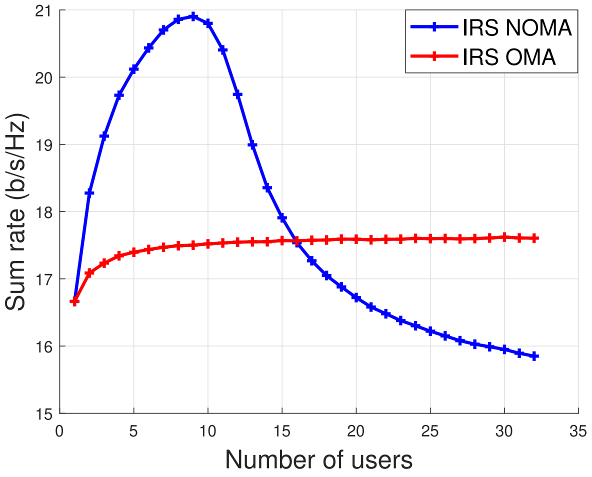

FIGURE 6. Comparison between IRS NOMA and IRS OMA sum-rates vs a different number of users. M = 16, and Pt = 40 dbm.

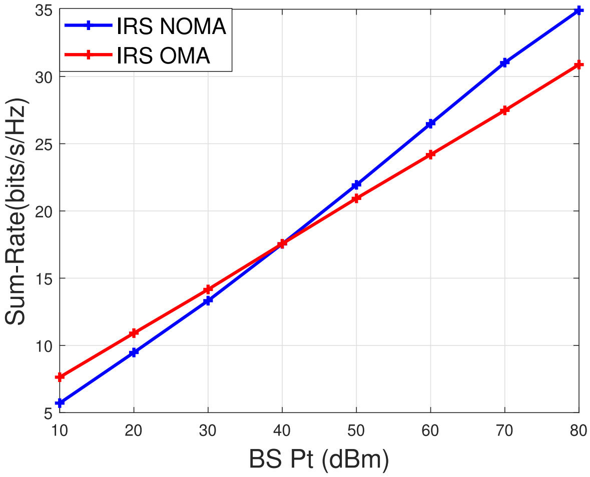

FIGURE 7. Comparison between IRS NOMA and IRS OMA sum-rates vs different power levels. M = 16, and K = 16.

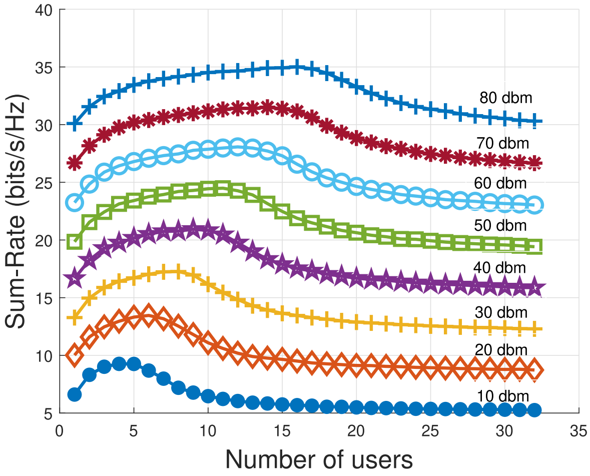

FIGURE 8. NOMA sum-rate for various power levels. M = 16.

the ReLU activation function. The output layer of the actor-network uses the tanh(·) function to provide enough gradient. For the critic network, the input layer is the number of states and the number of actions. The state input is followed by two dense layers of 128, and 256 neurons respectively with ReLU activation functions, and the action input is followed by one dense layer of 128 neurons. Both outputs are passed via a separate layer before concatenating to represent the input of the critic network. After that, two hidden layers are added each of 256 neurons with ReLU activation functions. This is pursued by the output layer of the critic network which contains 16 neurons. Both actor and critic main networks use Adam optimizer to update parameters. Moreover, we produce average results by considering the average sum-rate over 1000 channel realizations, we set the actor learning rate = 0.0005, the critic learning rate = 0.001, the coefficient of soft updates $\tau = 0.05$, the discount factor $\Gamma = 0.99$, the buffer capacity = 100 000. The noise is complex additive white Gaussian with mean equal to zero and variance equal to 0.1.

A comparison between IRS NOMA and IRS OMA sum-rates versus the number of users is shown in Fig. 6, where the transmit power Pt is 40 dBm. It is realized that NOMA performs better than OMA since it provides a higher sum rate for several users less than 16. The reason is that in NOMA there is resource sharing among users since NOMA multiplexes users in the power domain, and thus there is no bandwidth division. Therefore the rate and spectral efficiency are higher. However, in OMA there is no resource sharing and thus the bandwidth is divided among users. Further, when the number of users increases above 16, interference between users increases, and thus OMA performs better in this case and provides a higher sum-rate than NOMA. From Fig. 7, we notice that at low transmit power levels and for 16 users, IRS OMA performs slightly better than IRS NOMA. The reason is that, at low SINR, IRS NOMA users suffer from the interference caused by the simultaneous transmission, while OMA users do not experience any such interference. The NOMA system, at low SINR, will not have sufficient power to ensure a significant channel disparity among the users which restricts the potential gain brought by the NOMA system. When we increase the power. However, at high power levels, the sum-rate of IRS NOMA system is better than the IRS OMA system.

Moreover, Fig. 8 demonstrates the sum-rate of NOMA vs the number of users for different power levels starting from 10 up to 80 dBm. The lower curve represents the sum-rate generated at transmit power equals 10 dBm, and the highest curve depicts the sum-rate generated at transmit power equals 80 dBm. It is realized that as power increases the sum-rate increases and thus our IRS NOMA system can serve more users. Fig. 9 reveals the scalability of our DDPG algorithm to a larger number of IRS elements. However, we cannot show the exhaustive search approximate upperbound in these cases due to the computationally prohibitive complexity. It needs to be noted that the sum-rate performance is enhanced with a larger number of IRS elements due to the added degrees of

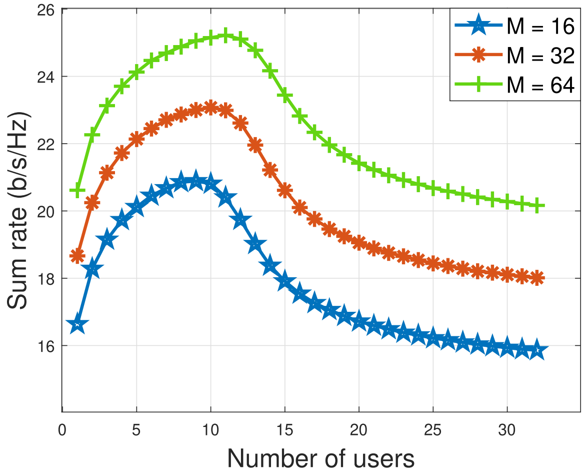

FIGURE 9. NOMA sum-rate for various number of reflecting elements. Pt = 40 dbm, and K = 32.

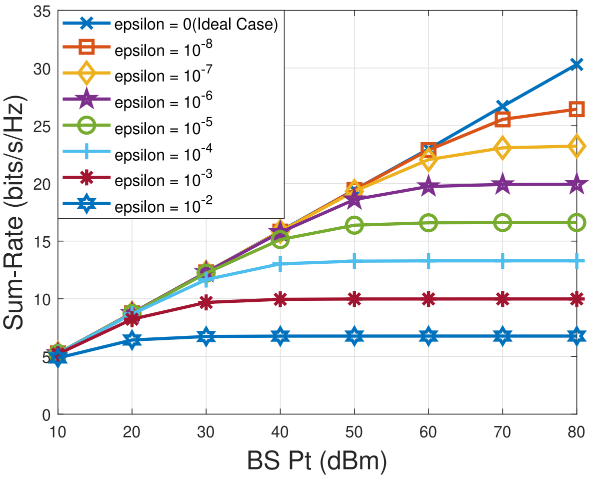

FIGURE 10. Achievable Sum-rate at nearest user with imperfect SIC. M = 16, and K = 32.

freedom and the increased ability to further focus the signal at the destination.

Furthermore, figures 10 and 11 shows the sum-rate of users during imperfect SIC where we have residual interference of all users' power in the denominator. Fig. 10 shows the rate for user K, the nearest user to the BS, for different power levels and ϵ values. It is well known that user 1, the farthest user from the BS does not perform SIC and thus our focus will be on the rate for user K. It is obvious from Fig. 10 that as the imperfection increases, the rate for user k decreases. The curves are plotted for different values of ϵ which represents the fraction of residual interference. When ϵ equals 0, SIC is perfect, and thus the rate for user k is the highest compared to other rates when $\epsilon > 0$. As ϵ value increases the rate decreases due to increasing the fraction of imperfectness. Therefore, imperfect SIC has a deleterious impact on the rate of the users performing SIC. Fig. 11 reveals the performance of the DDPG algorithm compared to exhaustive search scheme during imperfect SIC. We calculated the optimum phase shift which maximizes the sum-rate taking into consideration that the channel is known, and ϵ is greater than zero. It is clear from Fig. 11 that DDPG algorithm can approach the upper bound even during imperfect SIC.

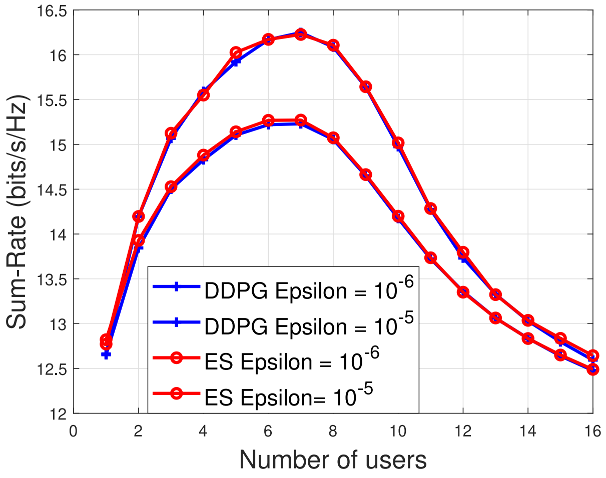

FIGURE 11. Upper bound on performance vs proposed DDPG algorithm with imperfect SIC. M = 4, K = 16, and $\Delta\Phi = \frac{2\pi}{30}$.

VI. CONCLUSION

In this paper, we considered the downlink scenario of the IRS NOMA system. Our main goal was to maximize the sum-rate of NOMA users. The formulated problem is non-convex since it involves the constant modulus constraint and the objective function which is also non-convex. Thus, the problem is suitable for DRL learning techniques. In particular, we have used the DDPG which is a DRL algorithm to solve the sum-rate maximization problem for our IRS NOMA scenario. Simulation results generated revealed that the sum-rate for NOMA can track the upper bound obtained through an exhaustive search, and it is superior to OMA for a specific number of users and predefined transmit power. Moreover, increasing the transmit power results in increasing the number of users served by the IRS NOMA system since NOMA multiplexes users in the power domain. Further, when considering the imperfect SIC scenario, which is more realistic, results showed that as the imperfection factor increases, the sum-rate of users decreases. This reveals the significance of performing SIC perfectly.Further, during imperfection, DDPG can still approach the upper bound. Thus, DDPG is a powerful algorithm, which enables us to include the optimization of dynamic power allocation in similar problems in the future.

REFERENCES

[1] X. H. You *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, no. 1, 2021, Art. no. 110301. [Online]. Available: https://doi.org/10.1007/s11432-020-2955-6

[2] H. Hashida, Y. Kawamoto, and N. Kato, "Intelligent reflecting surface placement optimization in air-ground communication networks toward 6G," *IEEE Wireless Commun.*, vol. 27, no. 6, pp. 146–151, Dec. 2020, doi: 10.1109/MWC.001.2000142.

[3] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019, doi: 10.1109/ACCESS.2019.2935192.

[4] P. Wang, J. Fang, X. Yuan, Z. Chen, and H. Li, "Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14960–14973, Dec. 2020, doi: 10.1109/TVT.2020.3031657.

[5] W. Tang *et al.*, "Wireless communications with reconfigurable intelligent surface: Path loss modeling and experimental measurement," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 421–439, Jan. 2021, doi: 10.1109/TWC.2020.3024887.

[6] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019, doi: 10.1109/TWC.2019.2922609.

[7] S. Hu, F. Rusek, and O. Edfors, "Beyond massive MIMO: The potential of data transmission with large intelligent surfaces," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2746–2758, May 2018, doi: 10.1109/TSP.2018.2816577.

[8] M. Jung, W. Saad, Y. Jang, G. Kong, and S. Choi, "Performance analysis of large intelligent surfaces (LISs): Asymptotic data rate and channel hardening effects," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2052–2065, Mar. 2020, doi: 10.1109/TWC.2019.2961990.

[9] S. Hu, F. Rusek, and O. Edfors, "The potential of using large antenna arrays on intelligent surfaces," in *Proc. IEEE 85th Veh. Technol. Conf. (VTC Spring)*, 2017, pp. 1–6, doi: 10.1109/VTCSpring.2017.8108330.

[10] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for intelligent reflecting surface-enhanced wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6, doi: 10.1109/GLOBECOM38437.2019.9013288.

[11] X. Yu, D. Xu, and R. Schober, "MISO wireless communication systems via intelligent reflecting surfaces: (Invited paper)," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2019, pp. 735–740, doi: 10.1109/ICCChina.2019.8855810.

[12] M. Jung, W. Saad, and G. Kong, "Performance analysis of active large intelligent surfaces (LISs): Uplink spectral efficiency and pilot training," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3379–3394, May 2021, doi: 10.1109/TCOMM.2021.3056532.

[13] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting intelligent reflecting surfaces in NOMA networks: Joint beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6884–6898, Oct. 2020, doi: 10.1109/TWC.2020.3006915.

[14] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Reconfigurable intelligent surfaces: Three myths and two critical questions," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 90–96, Jan. 2021, doi: 10.1109/MCOM.001.2000407.

[15] A. Almohamad *et al.*, "Smart and secure wireless communications via reflecting intelligent surfaces: A short survey," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1442–1456, 2020, doi: 10.1109/OJCOMS.2020.3023731.

[16] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?" *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 244–248, Feb. 2020, doi: 10.1109/LWC.2019.2950624.

[17] M. W. Akhtar, S. A. Hassan, R. Ghaffar, H. Jung, S. Garg, and M. S. Hossain, "The shift to 6G communications: Vision and requirements," *Human Centric Comput. Inf. Sci.*, vol. 10, no. 1, p. 53, 2020, doi: 10.1186/s13673-020-00258-2.

[18] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017, doi: 10.1109/COMST.2016.2621116.

[19] *Initial Views and Evaluation Results on Non-Orthogonal Multiple Access for NR Uplink*, document 3GPP R1-163111, 3GPP, Valbonne, France, Apr. 2016.

[20] P. Swami and V. Bhatia, "Impact of distance on outage probability in IRS-NOMA for beyond 5G networks," in *Proc. IEEE 18th Annu. Consumer Commun. Netw. Conf. (CCNC)*, 2021, pp. 1–2, doi: 10.1109/CCNC49032.2021.9369548.

[21] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018, doi: 10.1109/COMST.2018.2835558.

[22] 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, Jan. 2021, doi: 10.1109/TWC.2020.3024860.

[23] 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, Aug. 2020, doi: 10.1109/JSAC.2020.3000835.

[24] K. Feng, Q. Wang, X. Li, and C. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 745–749, May 2020, doi: 10.1109/LWC.2020.2969167.

[25] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb, "Deep reinforcement learning for intelligent reflecting surfaces: Towards standalone operation," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2020, pp. 1–5, doi: 10.1109/SPAWC48557.2020.9154301.

[26] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Deep learning for large intelligent surfaces in millimeter wave and massive MIMO systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6, doi: 10.1109/GLOBECOM38437.2019.9013256.

[27] Z. Ding and H. V. Poor, "A simple design of IRS-NOMA transmission," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1119–1123, May 2020, doi: 10.1109/LCOMM.2020.2974196.

[28] Z. Zhang, L. Lv, Q. Wu, H. Deng, and J. Chen, "Robust and secure communications in intelligent reflecting surface assisted NOMA networks," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 739–743, Mar. 2021, doi: 10.1109/LCOMM.2020.3039811.

[29] G. Yang, X. Xu, and Y. Liang, "Intelligent reflecting surface assisted non-orthogonal multiple access," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2020, pp. 1–6, doi: 10.1109/WCNC45663.2020.9120476.

[30] J. Zuo, Y. Liu, E. Basar, and O. A. Dobre, "Intelligent reflecting surface-enhanced millimeter-wave NOMA systems," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2632–2636, Nov. 2020, doi: 10.1109/LCOMM.2020.3009158.

[31] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for IRS-assisted uplink NOMA," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 234–238, Jan. 2021, doi: 10.1109/LCOMM.2020.3025978.

[32] Q. Wu and R. Zhang, "Intelligent reflecting surface-enhanced wireless network: Joint active and passive beamforming design," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–6, doi: 10.1109/GLOCOM.2018.8647620.

[33] T. Xu, L. Sun, S. Yan, J. Hu, and F. Shu, "Pilot-based channel estimation design in covert wireless communication," Aug. 2019, *arXiv:1908.00226*.

[34] T. Hou, X. Sun, and Z. Song, "Outage performance for non-orthogonal multiple access with fixed power allocation over Nakagami-*m* fading channels," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 744–747, Apr. 2018, doi: 10.1109/LCOMM.2018.2799609.

[35] I. A. Mahady, E. Bedeer, S. Ikki, and H. Yanikomeroglu, "Sum-rate maximization of NOMA systems under imperfect successive interference cancellation," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 474–477, Mar. 2019, doi: 10.1109/LCOMM.2019.2893195.

[36] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Found. Trends Mach. Learn.*, vol. 11, nos. 3–4, pp. 219–354, 2018. doi: 10.1561/2200000071.

[37] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the Brownian motion," *Phys. Rev.*, vol. 36, no. 5, pp. 823–841, 1930. doi: 10.1103/PhysRev.36.823.