Deep Reinforcement Learning-based Robust Design for an IRS-assisted MISO-NOMA System

Abdulhamed Waraiet[1], Graduate Student Member, IEEE, Kanapathippillai Cumanan[1],Senior Member, IEEE, Zhiguo Ding[2], Fellow, IEEE, and Octavia A. Dobre[3], Fellow, IEEE

[1]School of Physics, Engineering and Technology, University of York, York, UK
[2]Department of Computer and Communication Engineering, Khalifa University, Abu Dhabi, UAE
[3]Faculty of Engineering and Applied Science, Memorial University, St. John's, NL, Canada

abdulhamed.waraiet@york.ac.uk

The work of A. Waraiet and K. Cumanan were supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/X01309X/1. The work of O. Dobre is funded by the Canada Research Chair CRC-2022-00187.

ABSTRACT In this paper, we propose a robust design for an intelligent reflecting surface (IRS)-assisted multiple-input single output non-orthogonal multiple access (NOMA) system. By considering channel uncertainties, the original robust design problem is formulated as a sum-rate maximization problem under a set of constraints. In particular, the uncertainties associated with reflected channels through IRS elements and direct channels are taken into account in the design and they are modelled as bounded errors. However, the original robust problem is not jointly convex in terms of beamformers at the base station and phase shifts of IRS elements. Therefore, we reformulate the original robust design as a reinforcement learning problem and develop an algorithm based on the twin-delayed deep deterministic policy gradient agent (also known as TD3). In particular, the proposed algorithm solves the original problem by jointly designing the beamformers and the phase shifts, which is not possible with conventional optimization techniques. Numerical results are provided to validate the effectiveness and evaluate the performance of the proposed robust design. In particular, the results demonstrate the competitive and promising capabilities of the proposed robust algorithm, which achieves significant gains in terms of robustness and system sum-rates over the baseline deep deterministic policy gradient agent. In addition, the algorithm has the ability to deal with fixed and dynamic channels, which gives deep reinforcement learning methods an edge over hand-crafted convex optimization-based algorithms.

INDEX TERMS MISO-NOMA, power allocation, non-convex optimization, reinforcement learning, robust design.

I. INTRODUCTION

Non-orthogonal multiple access (NOMA) has been identified as one of the promising multiple access (MA) techniques for future wireless communications. This novel multiple access technique has the ability to support more than one user in the same resource block [1]. In addition, NOMA utilizes superposition coding (SC) at the transmitter and successive interference cancellation (SIC) at the receiver. This enables NOMA to offer higher spectral and energy efficiencies, massive connectivity, and better fairness while meeting the unprecedented requirements of future wireless networks. It has been demonstrated that NOMA can achieve superior performance over orthogonal multiple access (OMA) by efficiently utilizing the available radio resources [2], [3].

NOMA systems with multiple antennas have been subject to extensive studies recently, thanks to their additional degrees of freedom over single antenna systems [4]–[8]. Several contributions have been made for various system objectives including transmit power minimization [9], max-min rate optimization and sum-rate maximization [10]. In [10], Hanif *et al.* proposed an iterative algorithm to solve the sum-rate maximization problem for downlink multiple-input single-output (MISO)-NOMA system.

Recently, intelligent reflecting surfaces (IRS) have been

identified as another promising technology to combat the effects of channel fading, which improves the reliability of wireless systems [11]. The IRS consists of multiple passive elements with programmable phase-shift surfaces which can redirect incoming signals towards the desired direction [12]. The beamforming design for IRS-assisted multiple-antenna NOMA systems with various objectives has been developed in [13] [14] [15].

The non-convex nature of many resource allocation problems in multiple antenna NOMA systems makes conventional convex optimization approaches less attractive, especially for real-time applications with stringent delay requirements.

Artificial intelligence-driven algorithms, on the other hand, have shown great potential in solving various challenging problems in wireless communications. A deep learning-based beamforming framework was proposed in [16] which can be applied to ultra-low latency communication systems. However, since deep learning models require training data and labelled solutions to effectively learn the problem, they are restricted to problems solved a priori, using hand-crafted optimization algorithms. Deep reinforcement learning (DRL) which combines RL with deep learning, on the other hand, can be leveraged to solve hard optimization problems that have not been solved beforehand, i.e., it does not require labelled data for training and learning. Instead, it generates its own policy and training data by interacting with the environment. Therefore, DRL methods are not simply mimicking agents, but active agents which aim to maximize their reward in a given environment through trial and error. In [17], Meng *et al.* proposed a DRL-based solution for sum-rate maximization in multi-cell networks. Xiao *et al.* proposed a deep deterministic policy gradient (DDPG) based solution to jointly optimize the beamforming and phase shifts of IRS elements for sum-rate maximization in an IRS-assisted MISO-NOMA system [18]. In [19], Gao *et al.* proposed a deep Q-network (DQN) based algorithm to jointly optimize IRS phase shifts and cluster power allocation in a NOMA system using the zero forcing approach. Multi-agent DRL-based design was proposed in [20] for solving the resource allocation problem in IRS-assisted semi-grant-free NOMA transmissions. Furthermore, Benfaid *et al.* proposed a resource allocation framework for unmanned aerial vehicles (UAV)-NOMA systems based on DQN [21]. In [22], Ding *et al.* applied a DDPG agent to maximize the long-term sum-rate for energy-constrained cognitive radio NOMA networks by optimizing the transmit power and the time-sharing coefficient of the system. More recently, authors in [23] proposed a DDPG-based solution to jointly optimize the IRS phase shifts and power allocation for a single antenna NOMA system with the assumption of imperfect SIC at the receivers. However, in all aforementioned studies, it is assumed a perfect channel state information at the transmitter (CSIT), which is seldom the case in practice. While the assumption of perfect CSIT is useful to derive upper bounds on the performance of different schemes, it often leads to overly optimistic results.

In this paper, we propose a robust design for the downlink of an IRS-assisted MISO NOMA system. By taking into account the channel uncertainties, the beamformers at the base station (BS) and phase shifts at IRS elements are jointly designed based on the twin delayed DDPG (TD3). This robust design is developed based on the worst-case approach. Both partial and full uncertainty models are considered. In the partial model, the errors are only considered for the links through IRS elements (cascaded channels) whereas the full uncertainty model considers the errors in both the direct and the cascaded channels. To the best of the authors' knowledge, this is the first work on the TD3-based robust design for a downlink MISO-NOMA system. The contributions of this work are summarized as follows:

- We consider the partial and full channel uncertainty models, where the true channels lie within a bounded error region around the estimated CSIT. This type of channel uncertainty is due to quantization errors. We then formulate the original robust design as an optimization problem. The objective of the optimization problem is to maximize the long-term system sum-rate under a set of QoS, total power, IRS amplitude and phase shift constraints.
- The original robust design problem is not convex jointly in terms of the beamformers and the phase shifts. Therefore, we reformulate the problem into an RL environment such that a TD3 agent can learn the environment and solve the original robust design problem. This reformulation allows for utilizing DRL agents to solve the challenging non-convex problem. Since RL agents cannot perform constrained optimization, we use normalization to ensure that actions taken by the agent fall within the feasible region of the original problem. In order to formulate the problem as an RL environment, the state, action and reward functions are defined appropriately. Then, we propose a TD3-based algorithm to solve the original non-convex joint robust optimization problem for the IRS-assisted MISO-NOMA system. By incorporating multiple error bounds within the original worst-case bound during training, the agent learns to design robust beamforming and IRS phase shifts for any error bound within the bounded error region. This enhances the sum-rate performance in the case of changing feedback quality during deployment.
- Unlike the conventional optimization approaches, the proposed novel design distributes the computational complexity of solving the joint design problem between the training and learning stages. Therefore, the result is a trained agent that generates competitive solutions with much lower complexity compared to the conventional optimization techniques. Such an advantage becomes particularly important in the case of highly dynamic-channels environments where conventional schemes need to execute the whole algorithm for each new

channel realization, leading to increased system latency and computational overhead at the BS.

- We provide extensive numerical simulation results to demonstrate the performance of the proposed TD3-based robust design. These results confirm the superior convergence properties of the proposed TD3-based algorithm. In addition, we benchmark the proposed agent against the baseline DDPG used in the literature, and the random algorithm for fixed and dynamic channel scenarios. The proposed agent outperforms the benchmark schemes in terms of achieved system sum-rates and robustness for both fixed and dynamic channel cases.

The rest of the paper is organized as follows. Section II presents the system model and the channel uncertainty model, and formulates the original robust design into an optimization problem. In Section III, brief overviews of RL and DRL agents are provided focusing on the TD3 agent. In addition, the original problem is reformulated as a DRL environment and an algorithm is developed to solve the original robust design problem. Section IV presents numerical results to demonstrate the superior performance of the proposed TD3 algorithm. Section V concludes this paper.

Bold upper case and lower case letters are used to represent matrices and vectors, respectively. Standard normal letters denote scalar quantities. $\boldsymbol{x^H}$ is the hermitian transpose of vector $\boldsymbol{x}$. $\|.\|_2$ and $|\ .\ |$ represent the Euclidean norm of a vector and the absolute value of a complex number, respectively. $||.||_F$ and $||.||_2$ denote the Frobenius norm and the L2 norm, respectively. $\mathbf{Card}(\mathbf{x})$ refers to the cardinality of vector $\mathbf{x}$. $\mathbb{R}$ denotes the set of real numbers, whereas $\mathbb{C}$ represents the set of complex numbers.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a downlink transmission of an IRS-assisted MISO-NOMA system, in which a BS equipped with T transmit antennas serves N single antenna user equipment (UEs). The IRS consists of M reflecting e lements. Furthermore, the effect of inter-cell interference is assumed to be either absent or accounted for in the noise at the receiver end. Such a system model setup can be utilized for various wireless communication systems in future wireless networks [24]–[26]. As shown in Figure 1, the BS establishes communications with UEs through a direct link and an indirect link through the IRS. In this NOMA system, the transmitted signal from the BS can be written as

$$\mathbf{x} = \sum_{i=1}^{N} \mathbf{w}_i s_i, \forall i \in \mathcal{N}, \tag{1}$$

where s_i is the information-bearing symbol for UE$_i$, $\mathbf{w}_i \in \mathbb{C}^{T\text{x}1}$ is the beamforming vector designed for UE$_i$, and $\mathcal{N} = \{1, ..., N\}$ is the set of all active UEs in the system. The power of the symbol is assumed to be 1, i.e., $E\{s_i s^*\} \overline{=} 1$.

Assuming flat fading channel conditions, the received signal at UE$_i$ can be represented as

$$y_i = \mathbf{h}_i^{\text{H}}\mathbf{x} + \mathbf{g}_i^{\text{H}}\boldsymbol{\Upsilon}\mathbf{H}\mathbf{x} + z_i, \forall i \in \mathcal{N}, \tag{2}$$

where $\mathbf{h}_i \in \mathbb{C}^{T\text{x}1}$ is the direct link channel vector between the BS and the UE$_i$. $\mathbf{g}_i \in \mathbb{C}^{M\text{x}1}$ represents the channel between the IRS and UE$_i$ and $\Upsilon = \text{diag}(v_1, ..., v_M) \in \mathbb{C}^{M\text{x}M}$ is a diagonal matrix that represents the phase shifts of IRS elements. The phase shift of each IRS element is modelled by $v_m = \alpha_m e^{j\theta_m}$, $m \in \mathcal{M}$, where $\mathcal{M}$ is the set of all IRS elements, $\alpha_m \in [0, 1]$ and $\theta_m \in [0, 2\pi]$, represent the amplitude and the phase shift of the m-th IRS element, respectively.Furthermore, $m \in \mathcal{M}$ We assume an ideal reflection with no energy losses by considering only the first-order reflection, i.e., $|v_m|^2 = 1, \forall m \in M$. The phase shift values are determined at the BS and then communicated to the IRS through a feedback link [27]. $\mathbf{H} \in \mathbb{C}^{M\text{x}T}$ is the channel matrix between the BS and the IRS. Note that we assume that the IRS is located on a fixed base (on top of a building for example) and therefore, the distance between the BS and IRS is a constant. We further assume that there exist line-of-sight (LoS) paths from the BS to the IRS, as well as from the IRS to the *N* UEs [28]. The z_i is the noise experienced by UE$_i$ and is modelled as an additive white Gaussian noise (AWGN) with zeros mean and variance σ_i^2. The received signal in (2) can be written in a more compact form as follows:

$$y_i = \left(\mathbf{h}_i^{\text{H}} + \mathbf{v}^{\text{H}}\mathbf{Q}_i\right)\mathbf{x} + z_i, \forall i \in \mathcal{N}, \tag{3}$$

$$y_i = \tilde{\mathbf{h}}_i\mathbf{x} + z_i, \forall i \in \mathcal{N}, \tag{4}$$

where $\mathbf{v} = \text{vec}(\Upsilon) \in \mathbb{C}^{M\text{x}1}$ and $\mathbf{Q}_i = \text{diag}(\mathbf{g}_i^{\text{H}})\mathbf{H} \in \mathbb{C}^{M\text{x}T}$ is the reflected (cascaded) channel matrix for UE$_i$.

Since NOMA utilizes SIC at the receiver end in the downlink [9] [10], determining an adequate decoding order is crucial in order to unlock the full potential benefits of NOMA. Channel strength is usually used as the criterion for deciding a decoding order that is optimal in the single antenna case, which is not the case for the multiple-antenna NOMA systems [9] [29]. Nevertheless, we will adopt the channel strength-based decoding order here, as optimal decoding order design is beyond the scope of this paper. According to channel strength-based decoding order, the UE with the strongest channel (referred to as the strongest UE), will be able to successively decode and subtract other UEs' signals, then proceed to decode its own signal. The UE with the weakest channel (referred to as the weakest UE), will directly decode its signal while considering interference from other UEs' signals as noise. To further clarify this decoding order, suppose that there are N users in the system and their estimated channels at the BS are $||\hat{\tilde{\mathbf{h}}}_1||_2^2 \geq ||\hat{\tilde{\mathbf{h}}}_2||_2^2 \geq ... \geq ||\hat{\tilde{\mathbf{h}}}_N||_2^2$, where $\hat{\tilde{\mathbf{h}}}_i$ is the estimated version of $\tilde{\mathbf{h}}_i$ at the BS; then, the decoding order set is $\zeta = \{1, 2, ..., N\}$ where UE$_1$ decodes UE$_2$, ..., UE$_N$ signals before decoding its own, UE$_2$ decodes UE$_3$, ..., UE$_N$ signals before decoding its own signal while treating UE$_1$'s signal as noise, and so on. The weakest

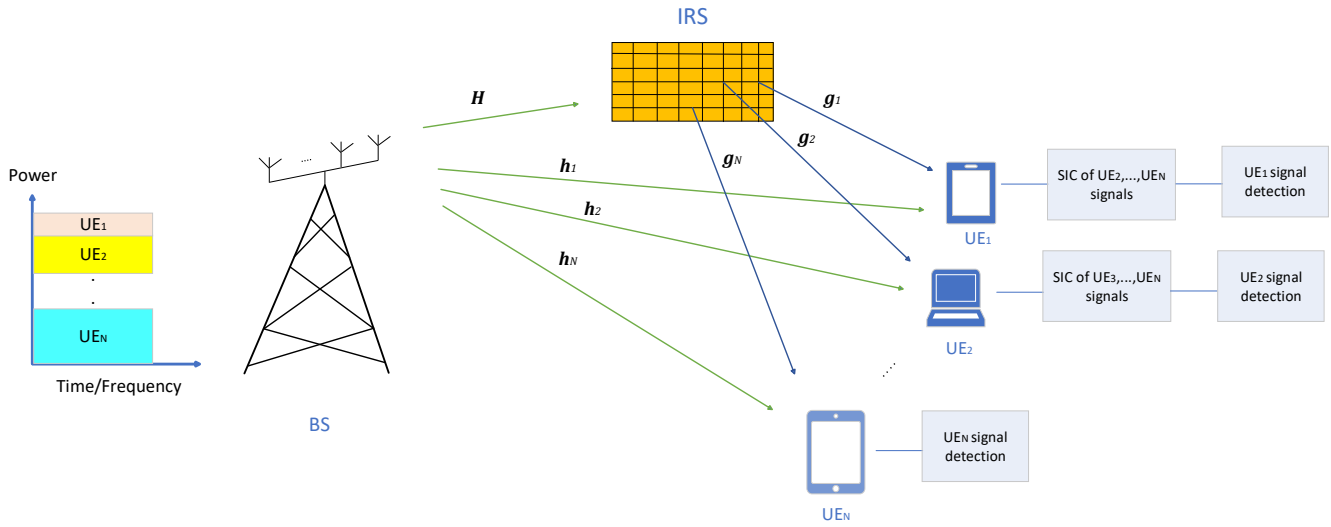

FIGURE 1. IRS-assisted Downlink MISO-NOMA system.

user, UE_N, will not carry out any SIC operations and will directly decode its own signal while treating interference from other UEs as noise [9] [10] [29].

A. CHANNEL UNCERTAINTY MODEL

Channel uncertainties are inevitable in wireless communications due to channel estimation and quantization errors. These two main sources of imperfect CSIT are, in fact, modelled differently. Channel estimation errors are unbounded and normally expressed using statistical models [30]. The error vectors from this type of error form a normal distribution with a known mean and covariance matrix. Quantization errors, on the other hand, originate from imperfect CSI reporting from the receiver side. A good example where quantization errors are encountered is in frequency division duplex (FDD) systems where the receiver uses a rate-limited feedback channel to report its channel information after quantization. However, given the constrained resolution quantizers used in UEs, additional errors are introduced in the estimated signal during quantization. The quantized channel coefficients transmitted by the UE through the uplink feedback link are affected by some quantization errors. Assuming the UE is using a uniform quantizer, the quantization errors can be modelled using a bounded error model [31]–[35]. In this paper, we aim to study the effects of imperfect CSIT due to quantization errors on the beamforming design at the BS, and consequently, on the achievable system sum-rate. In particular, we develop a worst-case beamforming design approach that guarantees the minimum rates requested by the UEs for any value of errors within the bounded region. Furthermore, since there are two links from the BS to the UEs, namely, a direct link and a reflected link through the IRS elements, we consider the following two error models:

1) *Partial error model*: In this error model, we assume that the direct link between the BS and $\text{UE}_i, \forall i$, has negligible quantization error effects, while the reflected link is plagued by quantization errors. This scenario is motivated by the fact that the reflected channel is more challenging to obtain than the direct channel due to the passive elements of the IRS [35] [36]. The true reflected channel $\mathbf{Q}_i$, can be modelled as

$$\mathbf{Q}_i = \hat{\mathbf{Q}}_i + \Delta\mathbf{Q}_i, \forall i \in \mathcal{N}, \tag{5}$$

where $\hat{\mathbf{Q}}_i$ is the reflected CSI estimate at the BS and $\Delta\mathbf{Q}_i$ is the unknown error.

2) *Full error model*: In this model, we consider a full uncertainty scenario where both the direct and the reflected links are plagued by quantization errors. The true reflected channel expression is the same as in (5), while the true direct channel can be expressed as [5] [35]

$$\mathbf{h}_i = \hat{\mathbf{h}}_i + \Delta\mathbf{h}_i, \forall i \in \mathcal{N}, \tag{6}$$

where $\hat{\mathbf{h}}_i$ is the estimate of direct CSI at the BS and $\Delta\mathbf{h}_i$ is the unknown error.

The unknown errors are norm-bounded such that $||\Delta\mathbf{Q}_i||_F \leq e_{i,r}$, $||\Delta\mathbf{h}_i||_2 \leq e_{i,d}$, for the reflected and the direct channels, respectively. The error bounds $e_{i,r}$, $e_{i,d}$ of UE_i are known at the BS and expressed as [35]

$$e_{i,r} = \sqrt{\frac{\beta_{i,r}^2 \, \Gamma_{2MT}^{-1}}{2}}, \forall i \in \mathcal{N}, \tag{7}$$

$$e_{i,d} = \sqrt{\frac{\beta_{i,d}^2 \, \Gamma_{2T}^{-1}}{2}}, \forall i \in \mathcal{N}, \tag{8}$$

where $\beta_{i,r}^2 = \lambda_r^2 ||\mathbf{q}_i||_2^2$, $\mathbf{q}_i = \text{vec}(\hat{\mathbf{Q}}_i) \in \mathbb{C}^{MT\text{x}1}$ and $\beta_{i,d}^2 = \lambda_d^2 ||\hat{\mathbf{h}}_i||_2^2$ are the variances of ΔQ_i and Δh_i, respectively. $\lambda_r, \lambda_d \in (0, 1]$ are scalars that indicate the relative

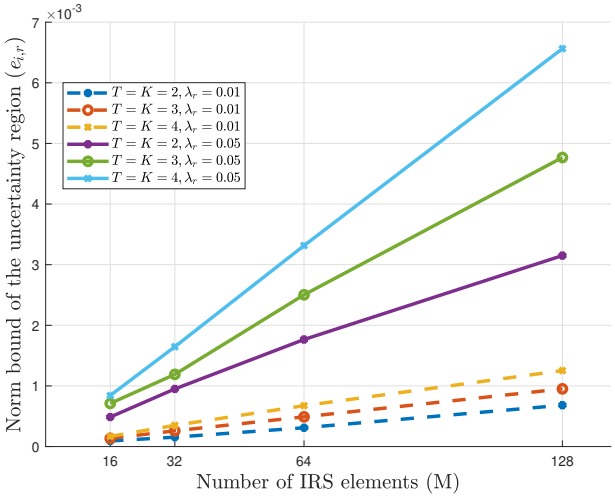

FIGURE 2. Norm bound of uncertainty region versus the number of IRS elements for different system parameters.

value of the error boundaries. $\Gamma^{-1}_{2MT}, \Gamma^{-1}_{2T}$ are the inverse of the cumulative distribution function (CDF) for the Chi-square distribution with $2MT, 2T$ degrees of freedom for the reflected and the direct links, respectively. As seen from (7), the error boundary of the reflected channel $e_{i,r}$ is a function of the number of transmit antennas T, the number of IRS elements M, and the quality of the reflected CSI feedback represented by λ_r. According to (8), the error boundary of the direct channel $e_{i,d}$ is only related to the number of transmit antennas T and λ_d. Figure 2 illustrates how different system parameters of (7) have an impact on the error bounds of the uncertainty region.

Note that we assume perfect channel state information at the receiver (CSIR), and thus, ideal SIC at the receivers, there is no contradiction between these assumptions and the error model considered in this work. To elaborate, we consider the imperfect CSIT to be due to feedback errors, not due to channel estimation errors, as we show in the next subsection. Therefore, the SINR expressions above do not account for any SIC residuals.

B. SINR AND ACHIEVABLE RATE EXPRESSIONS

Taking into account the error model and the decoding order discussed in the previous subsections, we can now proceed to the signal-to-interference-plus-noise (SINR) expressions. Without loss of generality, the SINR of UE$_i$'s signal at UE$_j$ is expressed as [9]

$$\gamma_i^j = \frac{|\tilde{\mathbf{h}}_j^{\mathrm{H}} \mathbf{w}_i|^2}{\sum_{j=1}^{i-1} |\tilde{\mathbf{h}}_j^{\mathrm{H}} \mathbf{w}_j|^2 + \sigma_j^2}, \forall j \in \mathcal{B}_i, \tag{9}$$

where $\mathcal{B}_i$ is the set of interfering users with higher decoding order ranks than UE$_i$ according to their channel strengths. Therefore, the received SINR of UE$_i$ when decoding its own signal can be expressed as [10]

$$\gamma_i^i = \frac{|\tilde{\mathbf{h}}_i^{\mathrm{H}} \mathbf{w}_i|^2}{\sum_{j=1}^{i-1} |\tilde{\mathbf{h}}_i^{\mathrm{H}} \mathbf{w}_j|^2 + \sigma_i^2}, \forall j \in \mathcal{B}_i. \tag{10}$$

To guarantee the smoothness of the SIC operation at stronger UEs, UE$_i$'s SINR is [9]

$$\gamma_i = \min\left(\gamma_i^j, ..., \gamma_i^i\right), \forall j \in \mathcal{B}_i. \tag{11}$$

As a result, the achievable rate at UE$_i$ can be written as

$$R_i = \log_2\left(1 + \gamma_i\right), \forall i \in \mathcal{N}. \tag{12}$$

Note that despite the beamforming vectors and the phase shifts of the IRS elements being designed at the BS based on the estimated channel $\hat{\tilde{\mathbf{h}}}_i$, the SINR expressions in (9) and (10) are evaluated using the true channel $\tilde{\mathbf{h}}_i$, which contains the unknown norm-bounded error elements [5], [35]. Hence, the considered robust beamforming design is more challenging to the BS in this case due to the unknown errors. The next subsections discuss the robust design problem in detail.

C. IMPLICATIONS OF ERROR MODEL ON NOMA SYSTEMS

In the previous section, we explained the bounded error model we consider in this work. However, it is worthwhile to explain the implications of using bounded and unbounded error models on the SINR expressions. In the case of a bounded error model, the CSIT imperfection is caused by the quantization errors in the uplink CSI report transmitted by the UE, not channel estimation errors. The quantization error region can therefore be approximated by a ball [31] [37]. Channel estimation error, on the other hand, is modelled statistically where the error vector is drawn from a complex Gaussian distribution with a known mean vector and covariance matrix [30] [35]. Therefore, considering a channel estimation error model leads to taking into consideration imperfect SIC as well, since there is going to be an SIC residual when the stronger UE is trying to decode the weaker UE's signal. Hence, the assumption of a bounded error model because of channel uncertainty is inconsistent for NOMA systems, as channel estimation and SIC errors are described using an unbounded error model [38]. In this work, however, we focus on imperfect CSIT due to quantization errors. Therefore, the assumptions of CSIR and ideal SIC do not conflict with the channel uncertainty model we use.

D. PROBLEM FORMULATION

In this paper, we consider a robust design to maximize the long-term sum-rate of an IRS-assisted MISO-NOMA system under minimum QoS requirements. This robust design is developed based on the worst-case performance approach. In other words, the robust design should meet the required QoS regardless of the experienced channel uncertainties. We define the beamforming matrix $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_N]$, where $\mathbf{W} \in \mathbb{C}^{T \mathrm{x} N}$, which containts the beamforming vectors of all

UEs. The original long-term robust design can be formulated as the following optimization problem:

$$\max_{\boldsymbol{\Upsilon},\mathbf{W}} \quad \mathbb{E}\Big\{\sum_{t=1}^{\infty}\sum_{i=1}^{N}\delta^{t-1}R_i^t \,\big|\, \pi_t, s_t\Big\} \tag{13a}$$

$$\text{s.t.} \quad \frac{\left|\big(\hat{\mathbf{h}}_j^{\mathrm{H}} + \Delta\mathbf{h}_j^{\mathrm{H}} + \mathbf{v}^{\mathrm{H}}(\hat{\mathbf{Q}}_j + \boldsymbol{\Delta}\mathbf{Q}_j)\big)\mathbf{w}_i\right|^2}{\sum_{j=1}^{i-1}\left|\big(\hat{\mathbf{h}}_j^{\mathrm{H}} + \Delta\mathbf{h}_j^{\mathrm{H}} + \mathbf{v}^{\mathrm{H}}(\hat{\mathbf{Q}}_j + \boldsymbol{\Delta}\mathbf{Q}_j)\big)\mathbf{w}_j\right|^2 + \sigma_j^2}$$

$$\geq 2^{R_i^{min}} - 1, \forall ||\Delta\mathbf{U}_i||_l \leq e_{i,k}, \forall i \in \mathcal{N}, \tag{13b}$$

$$\sum_{i=1}^{N} ||\mathbf{w}_i||_2^2 \leq P_{max}, \tag{13c}$$

$$|v_m|^2 = 1, \ \forall m \in \mathcal{M}, \tag{13d}$$

$$0 \leq \theta_m \leq 2\pi, \ \forall m \in \mathcal{M}. \tag{13e}$$

where $\mathbb{E}\Big\{\sum_{t=1}^{\infty}\sum_{i=1}^{N}\delta^{t-1}R_i^t \,\big|\, \pi_t, s_t\Big\}$ denotes the expected value of long-term system sum-rate, given the policy and the state of the agent, and δ is the discount factor. These entities are explained in the next section. The constraint in (13b) ensures the successful implementation of SIC and that the required minimum QoS at UE_i is achieved regardless of the channel uncertainties, where $\mathbf{U}_i \in \{\mathbf{Q}_i, \mathbf{h}_i\}$, $l \in \{F, 2\}$ and $k \in \{r, d\}$ [39]. The constraint in (13c) takes into account the available maximum transmit power at the BS, while constraints (13d) and (13e) are related to the IRS elements to guarantee ideal reflection and appropriate phase shifts, respectively.

The above optimization problem is non-convex in terms of the beamforming vectors $\mathbf{W}$ and phase shifts Υ. In addition, it is an NP-hard problem in general due to the coupled optimization variables in (13a) and (13b). Note that the problem is still non-convex even in the absence of (13d) and (13e) as highlighted by [10]. Therefore, solving this problem using a convex optimization approach will require transforming the problem into convex form using different approximation methods and obtaining solutions based on iterative algorithms. Such iterative algorithms are highly complex in general. In particular, the algorithm should be executed for each new set of channels. In other words, the optimization problem needs to be solved for each new set of channels. To further demonstrate the complexity of the optimization problem in (13a), the work in [40] which solved the weighted sum-rate maximization (WSR) problem by proposing a centralized solution based on semidefinite programming (SDP) for optimizing the IRS phase shifts, and using the maximum-ratio transmission (MRT) for beamforming design. However, the existing work in the literature does not consider the power allocation problem in MRT, which is non-trivial and challenging to optimize optimally [16], [41]. The same work proposed an iterative algorithm in an alternating manner to optimize the IRS phase shifts and the beamforming vectors. The work in [42] proposed a distributed solution based on fractional programming and the alternating direction method of multipliers (ADMM) algorithm to iteratively solve the WSR optimization problem. However, both the centralized methods which utilize the SDP and the iterative methods are still expensive in terms of latency and computational complexity, especially when the number of inputs is high. It is also worth mentioning that such algorithms are hand-crafted for OMA, and not for NOMA systems. It is well-known that NOMA introduces additional constraints to the optimization problem to ensure the smoothness of the SIC operation at the receivers which is an essential part of the NOMA principle [10]. Therefore, the aforementioned conventional optimization approaches cannot be applied directly to the problem considered in this work.

To address these issues with iterative solution approaches, we propose a DRL-based robust design. Since RL agents are designed to optimize a long-term objective in a given environment, we can reformulate the problem as an RL environment and develop an RL-based algorithm where the agent solves the challenging optimization problem. In particular, we develop an approach to solve this robust design using the TD3 agent, which is an enhanced version of DDPG. There are mainly three main motivations for considering this DRL-based approach. First, using a DRL-based algorithm allows for solving the original problem, not an approximated version of it, which means that any feasible solution is guaranteed to solve the problem with no additional assumptions or conditions. This holds for both fixed and varying channels. The second relates to the computational complexity of trained DRL models. As we will see in the next section, the time complexity of obtaining a feasible solution from the trained network is almost trivial, which makes it more attractive to latency-sensitive applications. Finally, the fact that TD3 converges to a deterministic policy which is also the case for DDPG. However, TD3 is more stable and robust against policy-breaking issues found in the baseline DDPG as we explain in the next section.

III. PROBLEM REFORMULATION AS A RL ENVIRONMENT

In this section, we briefly summarize the basic concepts of RL focusing on the TD3 agent. Then, we reformulate the original optimization problem in (13a)-(13e) as an appropriate RL environment to efficiently solve by a TD3 agent.

A. RL AND DRL

Tabular RL methods like Q-learning and SARSA are limited to solving problems with discrete action and state spaces [43]. DRL methods, on the other hand, utilize the function approximation capabilities of deep neural networks (DNN), which makes them applicable to a wider variety of problems. DRL methods can be classified primarily into three categories; value-based methods, such as DQN [44] which can handle continuous state space but only support discrete action space. Policy-based methods such as the Reinforce algorithm [45] which optimize the policy directly through an actor network. Actor-critic methods such as DDPG and

TD3 [46] [47], are recent off-policy agents that train deterministic policies. The actor takes actions and optimizes the policy of the agent while the critic evaluates the action taken by the actor with regards to the current state and returns a Q-value. Through these interactions, actor-critic agents optimize the policy of the agent until it converges to an optimal or near-optimal policy. Furthermore, actor-critic agents can handle continuous action and state spaces which widens their applicability to a larger set of problems in wireless communications. Note that any actor-critic agent with continuous actions and state spaces can be applied to solve the robust design problem using the reformulation provided. However, we utilize the TD3 agent because it is an off-policy agent with higher sample efficiency due to the use of a replay buffer which allows for reusing past experiences. Furthermore, the TD3 agent optimizes a deterministic policy which is generally easier to implement compared to stochastic policies.

B. BRIEF OVERVIEW OF TD3

TD3 is an off-policy actor-critic DRL agent that is capable of handling continuous action and state spaces. A TD3 agent consists of two main parts, an actor and a critic. The actor is a DNN responsible for generating actions. It takes in the current state as input and generates an action based on its current policy. The critic's DNN is responsible for generating the corresponding Q-value for the action taken by the actor. As a result, the critic's DNN has two inputs, the current state and the current action taken by the actor. Note that training in the context of RL is not the same as in deep learning. In the case of RL, the agent learns in an online fashion, which has two important implications; training-data generation and learning are carried out simultaneously, and that training targets are constantly changing according to the agent's current policy. In order to stabilise learning, both the actor and the critic use a delayed copy of their current DNNs called target networks. Target networks stabilise learning by fixing the target value when optimizing actor's and critics' DNNs. Experience replay buffer is utilized by the majority of off-policy DRL agents and TD3 is no exception [48]. Previous interactions with the environment defined as tuples of $\{s, a, r, s'\}$, are saved in the replay buffer $\mathcal{D}$. The buffer is then sampled to obtain training data. Replay buffer with larger memory makes data more independent and identically distributed (iid), which reduces the DNN variance during training. The critic of the DDPG agent can be considered as a modified DQN that takes in the action performed by the actor and outputs a scalar Q-value. To mitigate the problem of overestimating the Q-value in DDPG, TD3 uses two (or more) critics and selects the smallest estimate of the target Q-value. Given that the next state s' is not the terminal state, the target can be expressed as [47]

$$y(r, s') = r + \delta \min_{i=1,2} \mathcal{Q}_{\phi_{i,\eta}}(s', \mu_{\psi_\eta}(s')), \tag{14}$$

where $\mathcal{Q}_{\phi_{i,\eta}}$ is the target network for the critic's DNN $\phi_i, i = 1, 2$, δ is the discount factor (current value) for future rewards, and μ_{ψ_η} is the actor's target network which provides the next action a' given a next state s'. Then, the two critics learn the Q-function by minimizing their respective objectives as follows [47]:

$$\begin{aligned} L(\phi_1, \mathcal{D}) &= \mathop{\mathbb{E}}_{(a,s,r,s')\sim\mathcal{D}} \left[\Big(\mathcal{Q}_{\phi_1}(s, a) - y(r, s') \Big)^2 \right], \\ L(\phi_2, \mathcal{D}) &= \mathop{\mathbb{E}}_{(a,s,r,s')\sim\mathcal{D}} \left[\Big(\mathcal{Q}_{\phi_2}(s, a) - y(r, s') \Big)^2 \right]. \end{aligned} \tag{15}$$

The actor in TD3 aims to optimize the policy. This is achieved by adjusting the weights of its DNN μ_ψ to maximize the corresponding Q-value, which is defined by optimizing the following objective [46]:

$$\max_{\psi} \mathop{\mathbb{E}}_{s\sim\mathcal{D}} \Big[\mathcal{Q}_{\phi_1}\big(s, \mu_\psi(s)\big) \Big], \tag{16}$$

which is identical to the DDPG actor. Unlike DDPG, TD3 updates its policy using (16) less frequently than its Q-values to reduce variance during the training. Hence, the policy update in (16) is not executed in each training step. When it does, the policy, however, gets updated by (16). The target networks for both the critics and the actor are updated at a much slower rate than their main counterparts using

$$\begin{aligned} \phi_{\eta,i} &= \rho\phi_i + (1-\rho)\phi_{\eta,i},\ i = 1, 2, \\ \psi_\eta &= \rho\psi + (1-\rho)\psi_\eta, \end{aligned} \tag{17}$$

where $0 < \rho \leq 1$ is the target networks' smoothing factor. Algorithm 1 summarizes the key steps of how the TD3's actor and critics process one experience. Note that in practice, these steps are carried out in batches instead of single experiences to increase computational efficiency.

Algorithm 1: TD3 Actor and Critic training

1. A tuple $\{s, a, r, s'\}$ is randomly sampled from the replay buffer $\mathcal{D}$;
2. The current state s is fed to actor's DNN μ_ψ to generate current action a;
3. Both s and a are fed to the critics' DNNs to generate $\mathcal{Q}_{\phi_1}(s, a)$ and $\mathcal{Q}_{\phi_2}(s, a)$;
4. The next state s' is fed to the actor's target DNN μ_{ψ_η} to generate the next action a';
5. The critics' target DNNs $\mathcal{Q}_{\phi_{i,\eta}}(s, a), i = 1, 2$, are fed with s' and a' to calculate the target using (14);
6. The critics are trained using (15);
7. The actor is trained using (16);
8. Target networks are updated using (17);

Overall, TD3 theoretically outperforms DDPG by utilizing double Q-learning to reduce overestimation effects and updating its policy less frequently to reduce variance. Furthermore, it employs target policy smoothing by adding

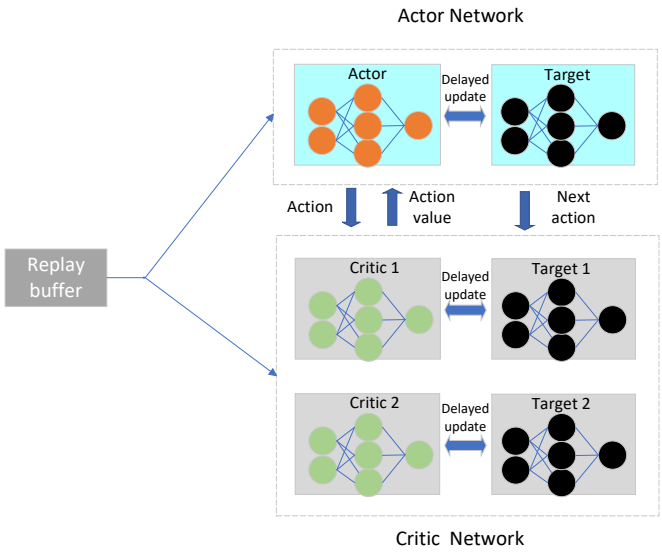

FIGURE 3. TD3 agent blocks.

noise to actor actions, and target actions as well to prevent the agent from exploiting errors in Q-value estimations [47]. Figure 3 shows the interactions between the internal components of the agent interact with each other to produce an optimal or near-optimal policy that maps states to the best possible actions. Despite that these upgrades may seem simple, combined together with hyperparameter tuning, they are the driving factor for any additional gain of TD3 over DDPG. Simulation results presented in section IV confirm the additional gain of TD3.

C. ROBUST DESIGN PROBLEM AS TD3 ENVIRONMENT

In order to solve the original robust problem using TD3, three entities must be clearly defined, namely, action space, state space, and reward. In this work, we define these entities as follows

- Since the optimization variables are the beamforming vectors and the phase shifts of IRS elements, these will be chosen as the agent's action. Therefore, the action vector of the agent at time-step t during training is expressed as

$$\mathbf{a}^t = \left[\mathbf{w}_1^t, ..., \mathbf{w}_N^t, v_1^t, ..., v_M^t\right]^T. \tag{18}$$

  where $\mathbf{a}^t \in \mathbb{C}^{NT+M}$.

- The state vector is defined with four important pieces of information about the environment, the power of the beamforming vectors from the previous time-step, the achieved rates including rates at which stronger UEs decode weaker UEs' signals, and random error bounds within the maximum error bound. Furthermore, to assist the agent in evaluating itself, we include the previous action $\mathbf{a}^{t-1}$ as part of the state. Therefore, we can express the state vector for our TD3 agent as follows:

$$\mathbf{s}^t = \Big[||\mathbf{w}_1^{t-1}||_2^2, .., ||\mathbf{w}_N^{t-1}||_2^2, e_1, .., e_N, R_1^{t-1}, R_2^{1,t-1}, .., R_N^{N-1,t-1}, R_N^{N,t-1}, \mathbf{a}^{t-1}\Big]^T, \tag{19}$$

  where the error values in the state vector are directly mapped to the reflected error bound in the case of the partial error model, while the error bounds correspond to the sum of the direct and reflected error bounds in the case of the full uncertainty error model. Therefore, $\mathbf{s}^t \in \mathbb{C}^{2N+\frac{N(N+1)}{2}+NT+M}, N \geq 2$, where $\frac{N(N+1)}{2}$ determines the number of all possible rates in the considered MISO-NOMA system.
  Note that both beamforming vectors and phase shifts are complex-valued design parameters and they are part of the action and state spaces. However, since we will be using real-valued neural networks for building the DRL agent, each complex vector is mapped to two separate real vectors where one represents the real values while the other represents the imaginary values of the original complex-valued vector [49] [16]. Therefore, the beamforming vector (or any complex vector for that matter) $\mathbf{w}_i \in \mathbb{C}^{T\text{x}1}$ is mapped to $\mathbf{Re}(\mathbf{w}_i) \in \mathbb{R}^{T\text{x}1}$ representing the real part of $\mathbf{w}_i$, and $\mathbf{Im}(\mathbf{w}_i) \in \mathbb{R}^{T\text{x}1}$ representing the imaginary part of $\mathbf{w}_i$. This is also true for the complex value phase shifts of the IRS elements, where each scalar complex phase shift value is mapped to two real scalars representing the real and complex parts of the original element. Note that this technique basically doubles the size of input and output layers for the critic and the actor DNNs. However, it unlocks the potential for using neural networks to deal with a wider range of problems such as the one considered in this paper. To reconstruct the complex-valued beamformers and IRS phase shift elements obtained from the action vector, we simply reverse the mapping process explained earlier. Therefore, the $\mathbf{a}^t \in \mathbb{R}^{2NT+2M}$, $\mathbf{s}^t \in \mathbb{R}^{2N+\frac{N(N+1)}{2}+2NT+2M}$ are corresponding real-only action and state space vectors, respectively.
- Finally, as the objective is to maximize the long-term sum-rate of the system, we choose the sum-rate at time-step t as the reward. Thus, the reward can be expressed as

$$r^t = \sum_{i=1}^{N} R_i^t, \forall i \in \mathcal{N}. \tag{20}$$

It is important to highlight that the agent will only be rewarded the sum-rate of the step if its action satisfies all constraints of the original optimization problem. However, since RL agents are only interested in maximizing their rewards, they cannot solve convex optimization problems directly. For this reason, we force the agent to meet the constraints by normalizing its actions to fall within the

feasible region. First, we start with the maximum transmit power constraint. Since the objective is an increasing function of the transmit power, at the optimal conditions, the transmitter will use all the available transmit power (i.e., $P_{\max}$). Therefore, we can rewrite the transmit power constraint (13c) as follows:

$$\sum_{i=1}^{N} ||\mathbf{w}_i||_2^2 = P_{\max}, \forall i \in \mathcal{N}. \tag{21}$$

The total power at time-step t can be expressed as

$$P_{\text{total}}^t = \sum_{i=1}^{N} ||\mathbf{w}_i^t||_2^2, \forall i \in \mathcal{N}. \tag{22}$$

We can then write the normalization coefficient as

$$\kappa^t = \sqrt{\frac{P_{\max}}{P_{\text{total}}^t}}. \tag{23}$$

Finally, the constraint-satisfying beamforming vectors can be written as

$$\mathbf{f}_i^t = \kappa^t \mathbf{w}_i^t, \forall i \in \mathcal{N}. \tag{24}$$

A similar process is carried out for the IRS elements. Since the angle θ can be mapped directly to a value in the feasible region, only amplitudes of the IRS elements need to be normalized as

$$\frac{v_m^t}{|v_m^t|}, \forall m \in \mathcal{M}. \tag{25}$$

With the normalized action, we then decide to either reward the agent with the sum-rate in (20) if the QoS requirements are satisfied under the channel uncertainty, otherwise, the agent is punished with a negative reward. Any negative reward will work as the agent will try to avoid such action in the future. We will use the sum of the rate deficit across all users as the negative reward [18]. The set ε contains users $j = 1, .., J, \varepsilon \in \mathcal{N}$ whose QoS are not satisfied at time-step t. Thus, we define the sum of the rate deficit across all users as

$$r_d^t = \sum_{j=1}^{J} \left(R_j^t - R_j^{min}\right), \ \ \forall j \in \varepsilon. \tag{26}$$

Therefore, if $\mathbf{a}^t$ satisfies the QoS constraints under some bounded error region, the agent will be given a positive reward according to (20), otherwise, it will be punished with the negative reward in (26). Algorithm 1 summarizes the proposed TD3-based algorithm for solving the original robust design problem. Note that Algorithm 2 summarizes the training process for the proposed agent. However, once the agent has been trained successfully, the actor network is the one we deploy in practice. The trained actor network can then be integrated into the BS hardware to be used to generate the solutions. To implement the proposed solution, in a practical IRS-assisted MISO-NOMA system, the BS receives the CSI reports in the uplink band. The BS then queries the trained actor network by using the obtained channels, i.e., executing steps $7-11$. The resulting IRS

Algorithm 2: TD3-based Robust Beamforming and Phase Shift Design

Initialize TD3 target and training parameters, empty replay buffer $\mathcal{D}$ and initialize the Gaussian random process $\mathcal{A}$;
Set $\phi_{\eta,1} \leftarrow \phi_1, \phi_{\eta,2} \leftarrow \phi_2, \psi_\eta \leftarrow \psi$;
while *Episode* $\leq$ *Total Episodes* **do**
  Acquire training channels based on the system parameters N, M, T;
  Calculate $\Delta\mathbf{Q}_i, \forall i$, according to (7) for the partial error model, adding $\Delta\mathbf{h}_i, \forall i$, according to (8) for the full error model;
  Initialize the beamforming vectors and the phase shift elements randomly;
  while $t \leq$ *Time steps* **do**
    Observe the current state s^t and obtain an action from the actor network using $\mathbf{a}^t = \text{clip}(\mu_\psi(s) + \epsilon, a_{low}, a_{high}), \epsilon \in \mathcal{A}$, normalize action values using (23), (24) and (25);
    Recover the complex value beamforming vectors and the IRS elements from step 6;
    Using vector $\mathbf{v}$ generated in the previous step, build the final estimated channels $\hat{\tilde{\mathbf{h}}}_i, \forall i$, according to (3);
    Decide a descending decoding order ζ such that $||\hat{\tilde{\mathbf{h}}}_1||_2^2 \geq ||\hat{\tilde{\mathbf{h}}}_2||_2^2 \geq ... \geq ||\hat{\tilde{\mathbf{h}}}_N||_2^2$, based on the estimated channels $\hat{\tilde{\mathbf{h}}}_i, \forall i$;
    Build the true channels $\tilde{\mathbf{h}}_i, \forall i$, using vector $\mathbf{v}$ and random errors based on (3), (5) and (6);
    Evaluate the SINR values and calculate the corresponding rates $R_i, \forall i$;
    if $R_i \geq R_i^{min}, e_i, \forall i \in \mathcal{N}$ **then**
      Use reward in (20);
    else
      Use reward in (26);
    end
    Obtain next state $\mathbf{s}^{t+1}$. Save tuple $\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\}$ to replay buffer $\mathcal{D}$;
    Randomly sample replay buffer using a batch of size b to calculate the target according to (14) and train the two critic networks ϕ_1, ϕ_2 using (15);
    if *time to update policy* **then**
      Update policy with one step using (16);
    end
    Update target networks using (17);
    $t = t + 1$;
    Set $\mathbf{s}^t = \mathbf{s}^{t+1}$;
  end
  $Episode = Episode + 1$;
end
Output: Obtain $\{\mathbf{f}_1^*, ..., \mathbf{f}_N^*, v_1^*, ..., v_m^*\}$

TABLE 1. Numerical time-complexity.

Profile	case 1	case 2	case 3	case 4
1	0.1667 h	0.1667 h	2.45 h	2.65 h
2	0.3167 h	0.30 h	4.283 h	5.067 h
3	0.3167 h	0.3833 h	4.45 h	4.783 h
4	1 h	1.283 h	12.95 h	14.15 h

vector is transmitted to the IRS via a feedback link, while the beamforming vectors are used for transmission.

D. COMPUTATIONAL COMPLEXITY ANALYSIS

In this subsection, we define the computational complexity of the proposed TD3-based algorithm. Similar to other deep learning models, the complexity of the proposed DRL framework can be divided into two categories: offline complexity, which is associated with training the actor network by plugging in critics and the replay buffer, and online complexity which is associated with inference or deployment of the actor's network. Calculating the best and the worst case run times for offline training of neural networks is still an open issue due to the complexity associated with the implementation of backpropagation and other hyper-parameters in DNNs [16] [50]. Furthermore, we assume that the offline complexity of this model can be afforded. Nevertheless, we include empirical comparisons for four different profiles with different hardware specifications in Table 1. The specification of each hardware platform and the system parameters used for each case are provided in Tables 5, 6 in the appendix.

For estimating the time complexity of inference, which is the cost of a feed-forward pass through the trained actor DNN, big $\mathcal{O}$ notation is a common method of measuring the worst-case run time of an algorithm. Since all modern libraries and deep learning frameworks use matrix notation to perform calculations through DNNs, it is straightforward to conclude that a matrix-vector multiplication operation, $\mathbf{z}_l = \mathbf{\Psi}\mathbf{c}_l$, where $\mathbf{\Psi}$ is the weights matrix, $\mathbf{c}_l$ is the input vector, and $\mathbf{z}_l$ is the output vector from the l-th hidden layer, is performed for each hidden layer. The output vector $\mathbf{z}$ is then passed through an activation layer as $\mathbf{b}^l = g(\mathbf{z}^l)$, where $\mathbf{b}^l$ is the activated vector that is fed to the next hidden layer in the DNN. Since the activation is an element-wise operation, it has a time complexity of $\mathcal{O}(\aleph_l)$, where $\aleph_l$ is the number of neurons in the l-th hidden layer. According to the proposed actor's architecture shown in Figure 4, there are three weight matrices in total, $\mathbf{\Psi}_1 \in \mathbb{R}^{\aleph \mathrm{x} \mathbf{Card}(\mathbf{s}^t)}$, linking the input to the first hidden layer, $\mathbf{\Psi}_2 \in \mathbb{R}^{\aleph^2}$, between the two hidden layers, assuming $\aleph_1 = \aleph_2 = \aleph$, and $\mathbf{\Psi}_3 \in \mathbb{R}^{\mathbf{Card}(\mathbf{a}^t) \mathrm{x} \aleph}$, linking the second hidden layer to the output layer. Therefore, we can write the total run-time as $\mathcal{O}\Big(T'\big(\aleph \cdot \mathbf{Card}(\mathbf{s}^t) + \aleph^2 + \mathbf{Card}(\mathbf{a}^t) \cdot \aleph + 2\aleph + \mathbf{Card}(\mathbf{a}^t)\big)\Big)$, where T' highlights the fact that the action space is part of the state space. Moreover, since the action vector is part of the state vector, then $\mathbf{Card}(\mathbf{s}^t) > \mathbf{Card}(\mathbf{a}^t)$ always holds. Therefore, we can approximate the worst-case run time for evaluating the actor's DNN as $\mathcal{O}\Big(\aleph \cdot max\big(\aleph, \mathbf{Card}(\mathbf{s}^t)\big)\Big)$. To define the complexity of the proposed DRL algorithm in context, we provide a complexity review for related works in the literature. The worst-case complexity for the iterative algorithm proposed in [10], which only solves the beamforming design problem, is $\mathcal{O}(N^7)$ per iteration. The SDP-based algorithm for optimizing the IRS phase shifts proposed in [40] has a worst-case complexity of $\mathcal{O}(M^6)$, while the iterative algorithm proposed in [42] reduced the IRS phase shifts optimization complexity to $\mathcal{O}(M^3)$ using ADMM. Furthermore, the worst-case run-time for the proposed algorithm scales linearly with the system parameters for a fixed number of neurons, while the worst-case run-time of the model-based algorithms is cubic at best. Therefore, compared to the complexities of the existing methods, the proposed algorithm has a significant advantage in terms of run times, while still maintaining competitive performance.

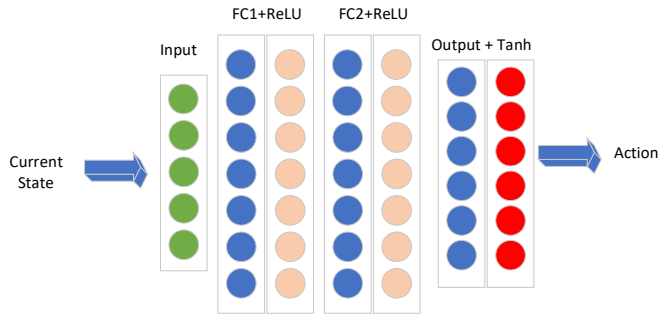

FIGURE 4. TD3 Actor DNN.

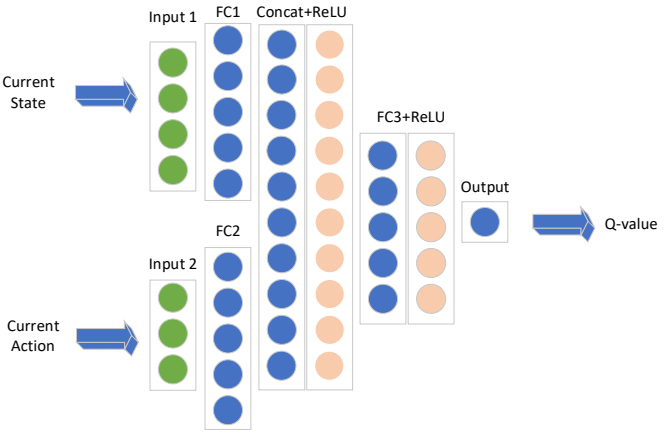

FIGURE 5. TD3 Critic DNN.

IV. TRAINING, SIMULATION AND NUMERICAL RESULTS

In this section, we evaluate the performance of the proposed TD3-based algorithm with different system models.

TABLE 2. Actor and critic layers.

Layer name	Layer size	Actor	Critic
Input layer 1	$\mathbf{Card}(\mathbf{s}^t)$	1	1
Fc1+ReLU	300	1	1
Input layer 2	$\mathbf{Card}(\mathbf{a}^t)$	-	1
Fc2+ReLU	300	1	1
Concat.+ReLU	300+300	-	1
Fc3+Tanh	$\mathbf{Card}(\mathbf{a}^t)$	1	–
Fc3+ReLU	300	-	1
Fc4	1	-	1

TABLE 3. Hyperparameters of the TD3 agent.

Hyperparameter	Value
Critics learning rate	0.001
Actor learning rate	0.0007
Policy update frequency	2
Discount factor	0.99
Smoothness factor (fixed channels),$N = 2, 3, N = 4$	$0.0007, 0.0002$
Smoothness factor (varying channels)	0.0005
Replay buffer size $(\mathcal{D})$	$100,000$
Minibatch size (b)	128
Number of Episodes, Time-steps (fixed channels)	$200, 200$
Number of Episodes, Time-steps (varying channels)	$2000, 300$

A. AGENTS STRUCTURE AND HYPERPARAMETERS

To evaluate the performance of the proposed robust design, we train a TD3 agent with one actor and two identical critics. Note that despite the two critics being identical in terms of layer type and size, the random initialization of their respective DNNs makes them behave differently, and therefore, produce different Q-value estimates. The architecture of the actor and critics DNNs are shown in Fig. 4 and 5, respectively. Table 2 describes the structure and size of the actor and critics networks. We set the number of hidden nodes to 300 for each hidden layer, irrespective of the input and output sizes, the *ReLU* activation function, $f(x) = max(0, x)$, is used for activating the hidden layers in both actor and critics' networks. The *Tanh* function, $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, is used as an activation function for the output in the actor's network. The ADAM optimizer is used for both actor's and critics' DNNs as it is more robust than other optimizers, and more appropriate for non-stationary objectives [51]. Table 3 provides a summary of hyperparameters used to train the agent for both fixed and dynamic channel cases. The reward discount factor is set to 0.99 to steer the agent towards a long-term optimal reward policy. Generally, the hyperparameters chosen for the TD3 agent in this paper are more on the conservative side. Such an approach favours training stability over faster convergence, which is recommended for the agent to form a more robust policy against channel uncertainties. Furthermore, since the optimal hyperparameter selection is an exhaustive search problem, the performance of the proposed algorithm can be considered the average performance in the context of the selected hyperparameters.

B. SYSTEM PARAMETERS

In terms of system parameters, we consider an IRS-assisted, downlink MISO-NOMA system, where $T = N = 2, 3, 4$, which is one of the cases where NOMA has the most advantage over OMA [10]. Table 4 summarizes the system parameters used in the simulations. Because of the high computational complexity associated with SIC receivers, the maximum number of UEs is limited to $N = 4$ where the strongest UE will perform 3 SIC operations. Increasing the number of UEs requires pairing the UEs into clusters, which is beyond the scope of this paper. For the channel model, both small-scale and large-scale fading are taken into account. The large-scale fading is a function of the distance from the BS and the IRS, for the direct and the reflected channels, respectively. The small-scale fading is modelled by Rician and Rayleigh fading for the reflected and direct channels, respectively. The channel coefficients for direct and reflected paths are drawn from a complex Gaussian distribution with zero mean and unit variance. The first part of the reflected channels from the BS to the IRS is modelled as

$$\mathbf{H} = \frac{1}{\sqrt{d_{irs}^{\alpha_{b\to irs}}}} \left(\sqrt{\frac{K}{1+K}} \mathbf{H}_{LoS} + \sqrt{\frac{1}{1+K}} \mathbf{H}_{nLoS} \right), \tag{27}$$

where K is the Rician factor that indicates the strength of the LoS component and is assumed to be 1, d_{irs} is the distance between the BS and the IRS and is fixed to 70 m. Similarly, the channel coefficients from the IRS to UE$_i$ are expressed as

$$\mathbf{g}_i = \frac{1}{\sqrt{d_i^{\alpha_{irs\to u}}}} \left(\sqrt{\frac{K}{1+K}} \mathbf{g}_{LoS} + \sqrt{\frac{1}{1+K}} \mathbf{g}_{nLoS} \right), \tag{28}$$

where d_i is the distance between the IRS and UE$_i$. The direct channels $\mathbf{h}_i$ between the BS and the UE$_i$ are modelled as $\mathbf{h}_i = \frac{h_i}{\sqrt{d_{id}^{\alpha_{b\to u}}}}$, where d_{id} is the distance between the BS and UE$_i$.

To fairly assess the performance of the proposed algorithm, we use the following benchmark algorithms

- **DDPG**: The DDPG agent has been widely adopted in the DRL literature. DDPG is included as a DRL benchmark to showcase the performance gain of the proposed TD3-based design in terms of convergence, system sum-rate, and robustness.
- **Baseline** 1: This benchmark scheme is based on SDP. More specifically, an SDP is used to solve the IRS optimization subproblem [40], and then the best possible rates are achieved for the given maximum available power through solving the transmit power minimization problem [16], [41]. Note that this scheme has pro-

TABLE 4. Summary of system parameters.

System parameter	Value
Cell radius	200 m
Number of UEs (N)	$2, 3, 4$
Number of antennas at the BS (T)	$2, 3, 4$
Number of IRS elements (M)	$16, 32, 64, 128$
Transmit power	30 dbm
Noise power	-90 dbm
Relative value for reflected error boundary λ_r	0.01
Relative value for direct error boundary λ_d	0.03
Probability value for $\Gamma_{2MT}^{-1}, \Gamma_{2T}^{-1}$	0.95
Path-loss exponent (BS-IRS) $\alpha_{b\to irs}$	2
Path-loss exponent (IRS-UEs) $\alpha_{irs\to u}$	2
Path-loss exponent (BS-UEs) $\alpha_{b\to u}$	2.5
Target rate R_i^{min} (fixed channels)	1 b/s/Hz
Target rate R_i^{min} (varying channels)	0.3 b/s/Hz

hibitively high complexity and is therefore used as an analytical benchmark.

- **Baseline** 2: This scheme is based on the well-known zero-forcing (ZF) principle as a solution to the beamforming design subproblem. However, since the multi-user power allocation problem is non-trivial in the ZF beamforming case, a fixed power allocation strategy is assumed for this scheme. Therefore, this is a non-robust scheme. The IRS optimization subproblem is solved using SDP [40].
- **Baseline** 3: This is a random benchmark scheme, i.e., the IRS phase shifts and the beamforming vectors are randomly generated. Such a scheme is included to show that the agent has derived a competitive policy that adapts to the environment.

In the following subsections, we provide simulation results generated by the agent for two system scenarios. The first is a fixed-channel scenario, where the channels are assumed to be fixed throughout the training period. The other scenario is a more realistic one where the channels are assumed to be dynamic, i.e., the UEs are randomly deployed such that $d_{id} \in [10, 200]$ m changes during both training and testing. Note that this translates to varying large-scale fading for each UE, which is more practical and more challenging to solve.

C. FIXED-CHANNEL SCENARIO

For the fixed-channel case, both partial and full error models are considered. The agent is trained for 200 episodes, with 200 time-steps per episode. The UEs are assumed to be separated by a distance of at least 30 m from each other. In each new episode, the agent is fed with new error values within their error bounds as part of the state vector.

Figures 6 and 7 present the convergence of the agent during training for the two extreme cases of IRS elements, $M = 16$ and $M = 128$, respectively. These convergence

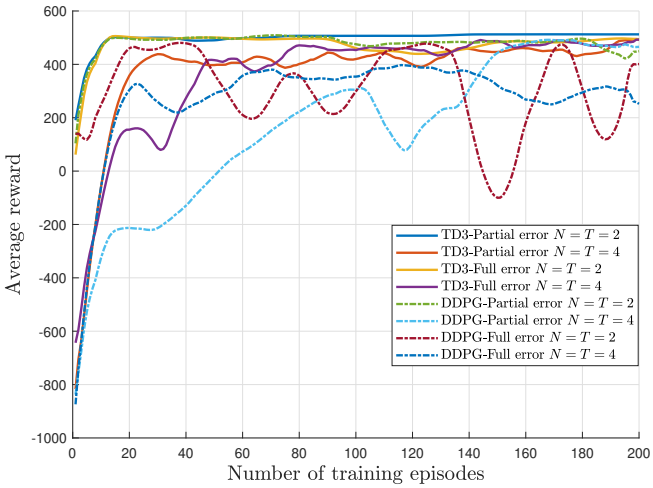

FIGURE 6. The reward of the proposed robust TD3, and DDPG agents for 200 training episodes, with fixed channels, $M = 16$, $R^{min} = 1$ b/s/Hz.

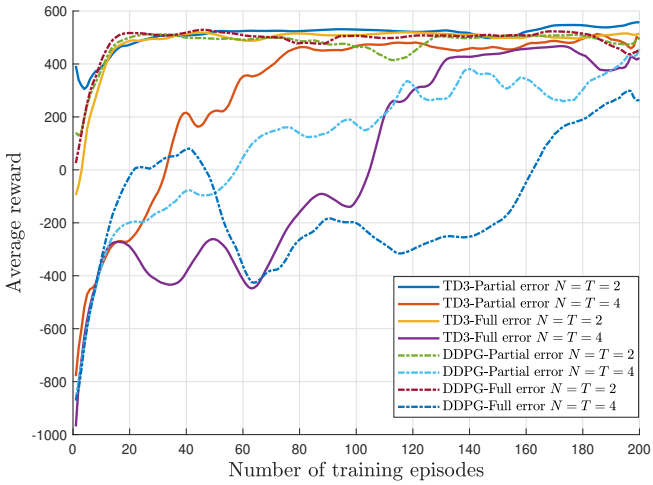

FIGURE 7. The reward of the proposed robust TD3, and DDPG agents for 200 training episodes, 200 time-steps per episode with fixed channels, $M = 128$, $R^{min} = 1$ b/s/Hz.

plots suggest that both agents are able to converge faster in the case of $M = 16$, compared to the other case with $M = 128$. This is expected, as M is directly related to the length of the state and the action vectors, and the error bound, making faster convergence in the case of $M = 128$ more challenging for the agents. Note that in both cases, the TD3 agent shows a more stable and consistent behaviour compared to that of the DDPG agent, thanks in part to the additional critic used by TD3. As seen in Figures 6 and 7, the TD3 agent requires around 40 episodes of training to reach an average reward level of greater than 400 in the first case, while other case requires around 130 episodes to achieve the same reward. The DDPG shows a similar performance in the case $M = 16$. However, Figure 7 shows the DDPG requires much higher training episodes to determine a high

reward policy when $N = 2, 4$. Overall, both agents require more training episodes to achieve convergence in the case of the full error model than in the partial error model. This is expected, as the robust beamforming design with a larger error bound is more challenging than the one with a small error bound. To demonstrate the potential capabilities of the TD3 agent in maximizing system sum-rate, Figures 8, 9 and 10 show the performance gains of the proposed TD3 agent. These simulation results are generated by taking the average rates of the agents when they are tested for a total of $1,000$ episodes, with 10 steps per episode. The achievable system sum-rates are higher in the partial error case across the three plots. The proposed TD3 agent outperforms the benchmarking DDPG and random schemes with variable margins. The most significant T D3 g ains o ver D DPG are achieved in the cases of $N = T = 4, M = 64$ and $N = T = 3, M = 128$, with 3.2 b/s/Hz, 5.4 b/s/Hz, for the partial and full error cases, respectively. This clearly shows that the proposed TD3 agent is able to derive a more accurate and higher rewarding policy than the DDPG agent. Another interesting observation from the achieved system sum-rates is that there are different peak rates for different numbers of UEs. In Figure 8, where $N = T = 2$, the maximum system sum-rate is achieved with $M = 64$, while in the case of $N = T = 3$, the sum-rate is achieved with $M = 128$, and in the case $N = T = 4$ it reaches with $M = 32$. This suggests that in each case, there is a sweet spot between having the ideal number of IRS elements to maximize the sum-rate, and having a manageable error region. It also suggests that, unlike many studies in the literature, increasing the number of IRS elements does not always result in an increased system sum-rate. In fact, when considering a robust design, increasing the number of IRS elements beyond a certain number may result in a degraded performance for the fixed c hannel c ase. C ompared t o the benchmark schemes, the TD3 agent generally outperforms the ZF baseline, even when the full error model is used. The performance gap in terms of the achieved system sum-rates between the proposed TD3-based design and the upper-bound baseline is marginal at best, with 1.9 b/s/Hz and 2.5 b/s/Hz for the partial and full error models, respectively.

In terms of achieved rates of UEs, Figure 11 presents UE_1 and UE_4 rates for both error models achieved by both agents, which represent the strongest and the weakest UEs in the system, respectively. The figure shows that U E_1 achieves higher rates when using the TD3 agent's policy. As for UE_4, both agents were able to consistently achieve the target rate required by the weakest UE for both error models. The apparent high variance in UE_1's rate for baseline 2 is caused by channel errors during testing since it is a non-robust scheme. This is also evident by the casual dips in UE_1's rate as shown in the same figure. Furthermore, to rigorously assess the robustness of both agents, Figure 12 demonstrates the performance of the agents for different target rates. The figure shows that the TD3 agent is able to achieve a perfect

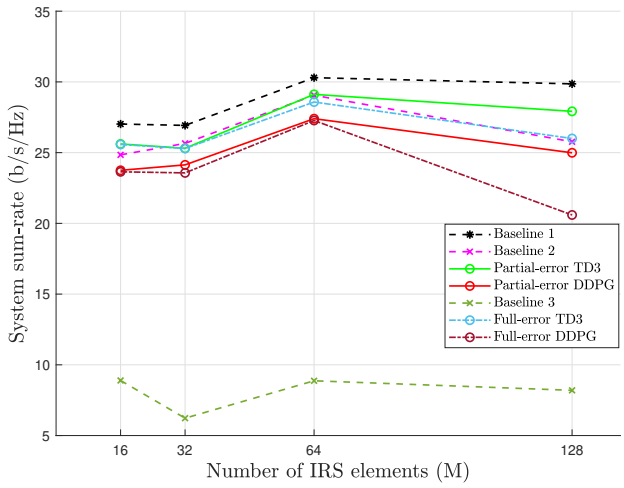

FIGURE 8. The achieved system sum-rate of the proposed robust design versus the number of IRS elements for $N = T = 2$, $R^{min} = 1$ b/s/Hz.

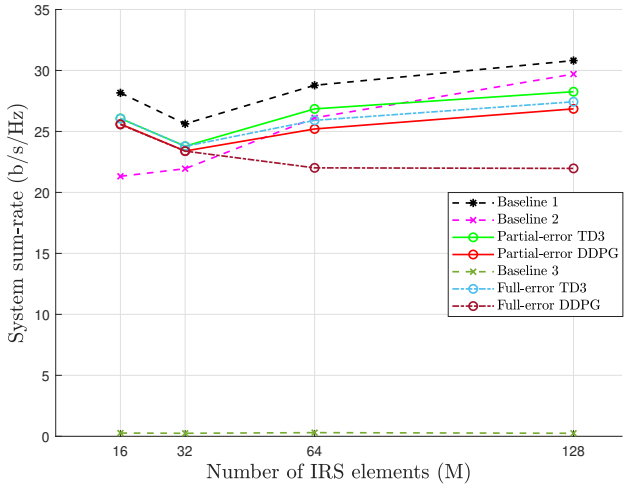

FIGURE 9. The achieved system sum-rate of the proposed robust design versus number of IRS elements for $N = T = 3$, $R^{min} = 1$ b/s/Hz.

score up to the training target rate, and after. In particular, the TD3 agent with $M = 128$ for the partial error model is able to attain a target rate of 1.5 b/s/Hz with a robustness score of 88%, which is impressive considering it was trained on a lower target rate of 1 b/s/Hz. The performance of the DDPG agent, on the other hand, is degraded in the case of full channel uncertainty, achieving a score of 89% with $M = 16$ as its worst case.

D. DYNAMIC-CHANNEL SCENARIO

In the previous scenario, the channels were assumed to be fixed. While this may be the case for stationary devices or low-mobility UEs, fixed channel models cannot be used for high-mobility situations where channels change drastically. To solve this dynamic channel problem, we train the TD3 agent on a small dataset of distinctively different channels.

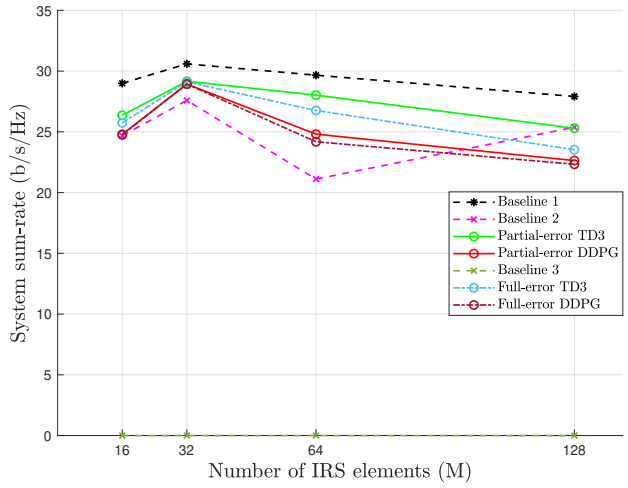

FIGURE 10. The achieved system sum-rate of the proposed robust design versus number of IRS elements for $N = T = 4$, $R^{min} = 1$ b/s/Hz.

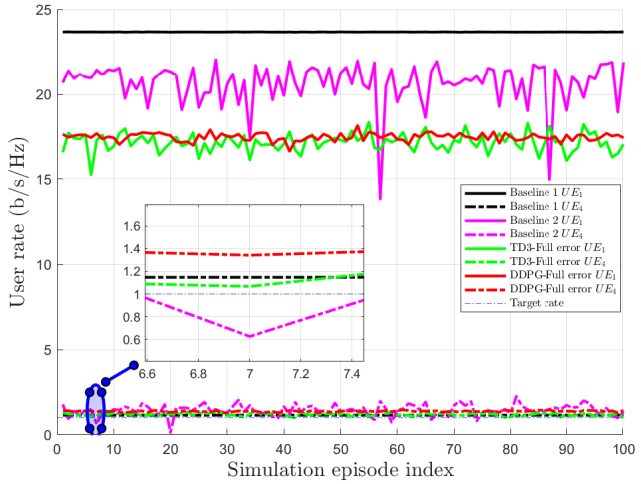

FIGURE 11. The achieved individual user rate of the proposed robust design across 100 testing episodes for $N = T = 4$, $R^{min} = 1$ b/s/Hz.

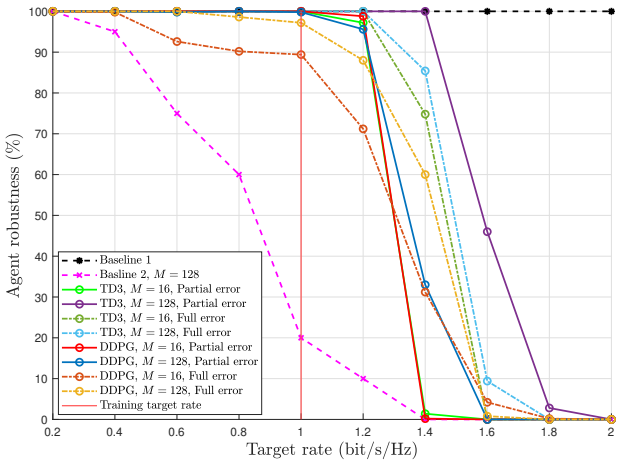

FIGURE 12. The robustness performance of the proposed agent versus the target rate with fixed channels, for $N = T = 4$, $R^{min} = 1$ b/s/Hz.

FIGURE 13. The reward of the proposed robust TD3, and DDPG agents for $2,000$ training episodes, with dynamic channels, $M = 128$, $R^{min} = 0.3$ b/s/Hz.

Also, we use the full error model for the varying channel case as we focus more on the practical implementation aspects of this design. Therefore, the TD3 agent is trained for a total of $2,000$ episodes and 300 steps per episode. At the beginning of each episode, a different set of channels randomly sampled from a dataset of 10 channels is selected. These training channels are generated based on a uniform sampling of the distance between the BS and the maximum cell radius. This uniform sampling is chosen to ensure that the training channels reflect the variance of the channels across the entire cell. Corresponding error bounds for direct and reflected links are also fed to the agent for each new episode during training as part of the state vector. Furthermore, to prevent the optimization problem from becoming infeasible due to higher channel variations, we reduce the target rate to 0.3 b/s/Hz for the dynamic channels scenario. To evaluate the performance of the agent in a dynamic-channel environment, we use a total of 250 randomly generated channels with $d_{id} \in [10, 200]$ m as a testing set. Also, the agent is simulated for $1,000$ episodes, with 10 steps per episode for testing, to determine the average achieved sum-rates. The convergence of the agent is shown in Figure 13 for the two extreme cases $N = T = 2, 4, M = 128$, where relatively higher training variance is apparent. This is expected since the channels are inherently different, and consequently, the reward will also have a higher variance. From Figure 13, we can see that there is a significant difference in terms of stability and consistency between the TD3 and the DDPG agents, where TD3 shows superior convergence properties. This is further evident by the relatively lower variance of the TD3 agent compared to the higher training variance of DDPG. Instability during training often leads to performance degradation

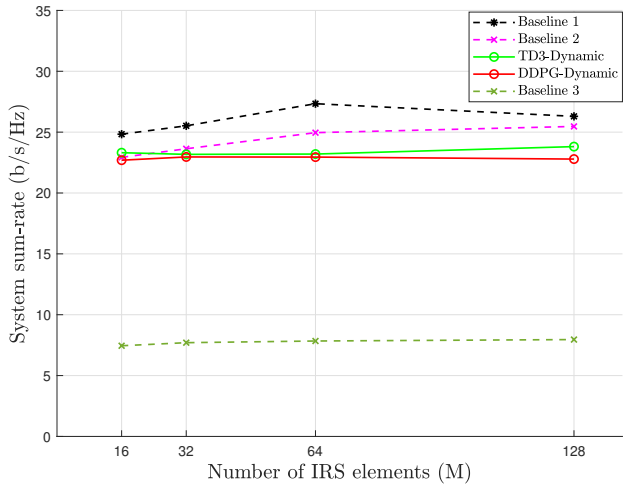

FIGURE 14. The achieved system sum-rate of the proposed robust design versus the number of IRS elements with dynamic channels, for $N = T = 2$, $R^{min} = 0.3$ b/s/Hz.

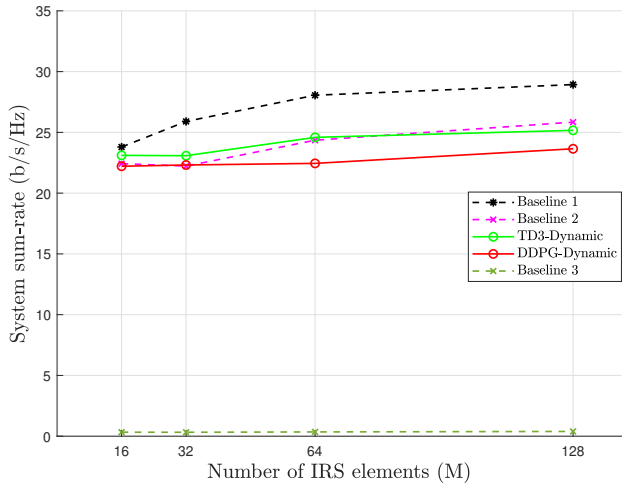

FIGURE 15. The achieved system sum-rate of the proposed robust design versus the number of IRS elements with dynamic channels, for $N = T = 3$, $R^{min} = 0.3$ b/s/Hz.

due to the inadequately derived policy. Figures 14, 15 and 16 illustrate the achieved system sum-rates for different system parameters. The TD3 agent shows marginal gains compared to the DDPG agent, with the most significant gain being 2.14 b/s/Hz, achieved in the case $N = T = 3, M = 64$. For the dynamic channel case, we can see that increasing the number of IRS elements is exploited by both agents, leading to a slight increase in terms of sum-rate. The TD3 agent is able to achieve a gain of 2.1 b/s/Hz in the system sum-rate for the case $N = T = 3, M = 64$. However, despite the addition of 64 IRS elements, the system sum-rate has not increased as much between $M = 64$ and $M = 128$, which further proves the point that the number of IRS elements may be utilized by the agent up to a certain number before starting to degrade the performance. Compared to the benchmark schemes, the proposed TD3 agent achieves a similar sum-rate performance to the ZF baseline scheme on average, while the sum-rate gap between the upper-bound baseline and the proposed agent has increased in the varying channels case with an average gap of 3.3 b/s/Hz. In terms of achieved individual rates, Figure 17 illustrates the rate for each UE for the dynamic channels case, with $N = T = 4, M = 128$. This Figure shows some casual drops of UE_4's rate below the 0.3 b/s/Hz mark by both the TD3 and the DDPG agents. This is expected due to the dynamic channels used for testing. Another observation is that DDPG achieved a higher rate for UE_1 at the expense of not satisfying the target rate required by UE_4, which is the result of converging to a non-optimal policy.

Finally, to evaluate the limits of the TD3 agent's derived policy in terms of robustness, we tested the trained agent for a set of target rates for $N = T = 4$. Figure 18 shows the robustness of the agent in satisfying each of the target rates. As expected, there is a trade-off between target rates and the

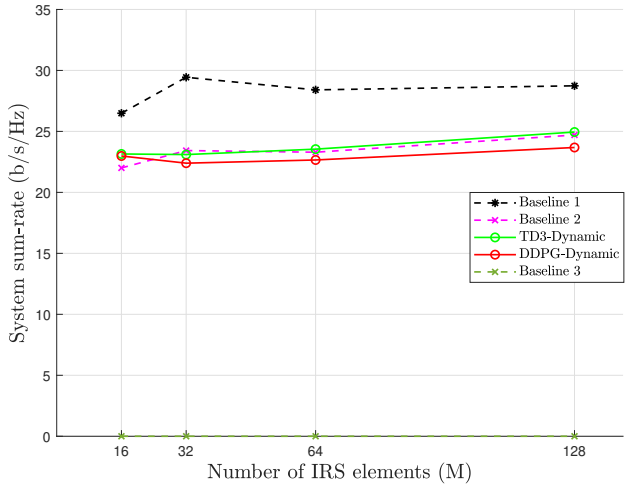

FIGURE 16. The achieved system sum-rate of the proposed robust design versus the number of IRS elements with dynamic channels, for $N = T = 4$, $R^{min} = 0.3$ b/s/Hz.

robustness of the agent. Despite the dynamic channels used for testing, TD3 is able to maintain a robustness performance of at least 65%. Furthermore, with $M = 64$; the agent maintained a competitive score up to 0.5 b/s/Hz, which is 66% higher than the target rate used during training. While both agents achieve similar system sum-rates as highlighted by Figures 14, 15 and 16, DDPG is less robust to channel uncertainties. The seemingly enhanced robustness score for baseline 2 is not related to the algorithm itself. Instead, it is due to the lower target rates used for dynamic-channels testing.

Overall, the TD3 agent outperforms the DDPG agent in every category, with marginal gain in some cases and significant in others. Furthermore, the results from the dynamic

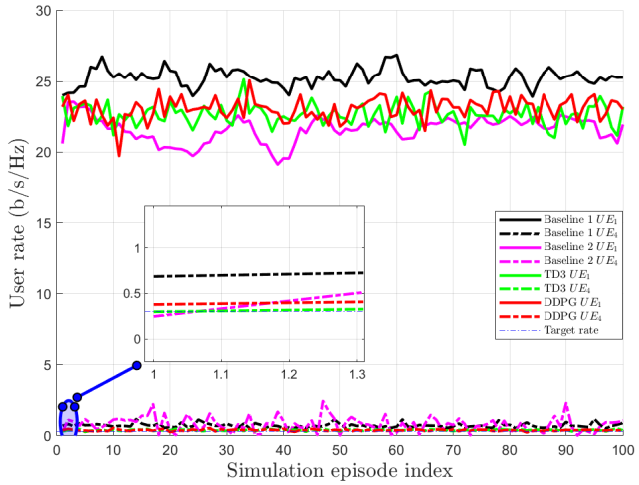

FIGURE 17. The achieved individual user rate of the proposed robust design across 100 testing episodes, with dynamic channels for $N = T = 4$, $R^{min} = 0.3$ b/s/Hz.

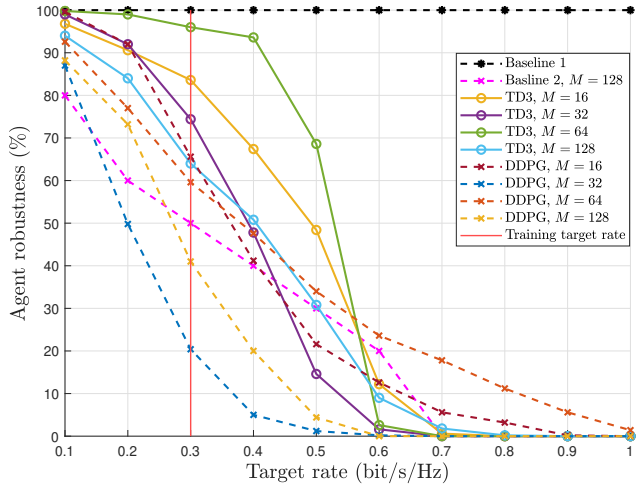

FIGURE 18. The robustness performance of the proposed agent versus the target rate with dynamic channels, for $N = T = 4$, $R^{min} = 0.3$ b/s/Hz.

channels scenario suggest that the TD3 agent is more robust to channel uncertainties.

V. CONCLUSION

In this paper, we proposed a DRL-based robust design for an IRS-assisted downlink MISO-NOMA system with imperfect channel feedback. In particular, a TD3 agent is developed to jointly optimize the beamforming vectors and the phase shifts of IRS elements to satisfy the required QoS with channel uncertainties. Through numerical simulations, we have shown that the proposed robust TD3 agent was able to maintain its robustness against channel uncertainties and achieved competitive performance in both fixed and dynamic channel cases. We showed that, unlike conventional convex optimization methods, the proposed robust TD3-based design solved the original non-convex problem, not an approximation of it. Furthermore, the agent only needed to converge to a good policy once. After being trained successfully, the agent was able to generate robust vectors and IRS phase shifts by performing a simple forward pass through its actor network, which was shown to have a low time complexity. This drastically reduces the latency in DRL-based designs and expands their applicability to low-latency systems. Conventional algorithmic methods, on the other hand, need to solve the problem each time a change occurs in the system state, causing higher system latency. We also showed that while additional IRS elements may improve the system sum-rate, it is not always the case that a higher number of IRS elements leads to sum-rate gains, especially when channel uncertainty is taken into account.

VI. APPENDIX

To ensure that MATLAB is able to exploit the maximum amount of computational resources on each of these hardware platforms, no other applications were running in the background during the testing period. Therefore, the empirical results provided in Table 1 reflect the best performance that these machines can sustain.

Profile 1 is equipped with state-of-the-art CPU, GPU and RAM units, which demonstrates the superior performance of this platform.

TABLE 5. Hardware profiles.

Profile	CPU	GPU	RAM size	RAM speed
1	13900KF	RTX4080	64GB	5200 MHz
2	10920X	A5000	128GB	2933 MHz
3	Xeon 6138	Tesla V100	40GB	2666 MHz
4	Xeon 6138	None	40GB	2666 MHz

TABLE 6. System parameters for run-time testing.

Case No.	$N = T$	M	Channel type	Episodes, steps
1	2	16	Fixed	200,200
2	4	128	Fixed	200,200
3	2	16	Varying	2000,300
4	4	128	Varying	2000,300

REFERENCES

[1] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347–2381, Nov. 2017.

[2] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proceedings of the IEEE 77th Vehicular Technology Conference (VTC Spring)*, 2013, pp. 1–5.

[3] Q. Sun, S. Han, C.-L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 405–408, Apr. 2015.

[4] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 537–552, Sep. 2015.

[5] F. Alavi, K. Cumanan, Z. Ding, and A. G. Burr, "Beamforming techniques for nonorthogonal multiple access in 5G cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9474–9487, Jul. 2018.

[6] K. Cumanan, Z. Ding, Y. Rahulamathavan, M. M. Molu, and H.-H. Chen, "Robust MMSE beamforming for multiantenna relay networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 3900–3912, 2017.

[7] L. Ni, X. Da, H. Hu, M. Zhang, and K. Cumanan, "Outage constrained robust secrecy energy efficiency maximization for EH cognitive radio networks," *IEEE Wireless Communications Letters*, vol. 9, no. 3, pp. 363–366, 2020.

[8] W. Wang, X. Li, R. Wang, K. Cumanan, W. Feng, Z. Ding, and O. A. Dobre, "Robust 3D-trajectory and time switching optimization for dual-UAV-enabled secure communications," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 11, pp. 3334–3347, 2021.

[9] J. Zhu, J. Wang, Y. Huang, K. Navaie, Z. Ding, and L. Yang, "On optimal beamforming design for downlink MISO NOMA systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 3008–3020, Jan. 2020.

[10] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 76–88, Sep. 2015.

[11] E. Basar, "Transmission through large intelligent surfaces: A new frontier in wireless communications," in *Proceedings of the IEEE European Conference on Networks and Communications (EuCNC)*, 2019, pp. 112–117.

[12] W. Hao, G. Sun, M. Zeng, Z. Chu, Z. Zhu, O. A. Dobre, and P. Xiao, "Robust design for intelligent reflecting surface-assisted MIMO-OFDMA terahertz IoT networks," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 13 052–13 064, Mar. 2021.

[13] Y. Li, M. Jiang, Q. Zhang, and J. Qin, "Joint beamforming design in multi-cluster MISO NOMA reconfigurable intelligent surface-aided downlink communication networks," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 664–674, Oct. 2020.

[14] G. Li, M. Zeng, D. Mishra, L. Hao, Z. Ma, and O. A. Dobre, "Energy-efficient design for IRS-empowered uplink MIMO-NOMA systems," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 9490–9500, May 2022.

[15] J. Zuo, Y. Liu, E. Basar, and O. A. Dobre, "Intelligent reflecting surface enhanced millimeter-wave NOMA systems," *IEEE Communications Letters*, vol. 24, no. 11, pp. 2632–2636, Jul. 2020.

[16] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1866–1880, Dec. 2019.

[17] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6255–6267, Jun. 2020.

[18] X. Xie, S. Jiao, and Z. Ding, "A reinforcement learning approach for an IRS-assisted NOMA network," *arXiv preprint arXiv:2106.09611*, Jun. 2021.

[19] X. Gao, Y. Liu, X. Liu, and L. Song, "Machine learning empowered resource allocation in IRS aided MISO-NOMA networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3478–3492, May 2022.

[20] J. Chen, L. Guo, J. Jia, J. Shang, and X. Wang, "Resource allocation for irs assisted SGF NOMA transmission: A MADRL approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 4, pp. 1302–1316, Jan. 2022.

[21] A. Benfaid, N. Adem, and B. Khalfi, "Adaptsky: A DRL based resource allocation framework in NOMA-UAV networks," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 01–07.

[22] Z. Ding, R. Schober, and H. V. Poor, "No-pain no-gain: DRL assisted optimization in energy-constrained CR-NOMA networks," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5917–5932, Jun. 2021.

[23] M. Shehab, B. S. Ciftler, T. Khattab, M. M. Abdallah, and D. Trinchero, "Deep reinforcement learning powered IRS-assisted downlink NOMA," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 729–739, Apr. 2022.

[24] L. Subrt and P. Pechac, "Intelligent walls as autonomous parts of smart indoor environments," *IET communications*, vol. 6, no. 8, pp. 1004–1010, 2012.

[25] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE wireless communications*, vol. 24, no. 4, pp. 72–80, 2017.

[26] Q. Wu, W. Chen, D. W. K. Ng, and R. Schober, "Spectral and energy-efficient wireless powered IoT networks: NOMA or TDMA?" *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6663–6667, 2018.

[27] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, and L. Hanzo, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1719–1734, Jun. 2020.

[28] Y. Han, S. Zhang, L. Duan, and R. Zhang, "Cooperative double-IRS aided communication: Beamforming design and power scaling," *IEEE Wireless Communications Letters*, vol. 9, no. 8, pp. 1206–1210, Apr. 2020.

[29] Y. Liu, H. Xing, C. Pan, A. Nallanathan, M. Elkashlan, and L. Hanzo, "Multiple-antenna-assisted non-orthogonal multiple access," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 17–23, Apr. 2018.

[30] J. Zhang, M. Kountouris, J. G. Andrews, and R. W. Heath, "Multi-mode transmission for the MIMO broadcast channel with imperfect channel state information," *IEEE Transactions on Communications*, vol. 59, no. 3, pp. 803–814, Dec. 2010.

[31] M. B. Shenouda and T. N. Davidson, "Convex conic formulations of robust downlink precoder designs with quality of service constraints," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 714–724, Dec. 2007.

[32] K.-Y. Wang, A. M.-C. So, T.-H. Chang, W.-K. Ma, and C.-Y. Chi, "Outage constrained robust transmit optimization for multiuser MISO downlinks: Tractable approximations by conic optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 21, pp. 5690–5705, 2014.

[33] F. Alavi, K. Cumanan, Z. Ding, and A. G. Burr, "Robust beamforming techniques for non-orthogonal multiple access systems with bounded channel uncertainties," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2033–2036, 2017.

[34] F. Alavi, K. Cumanan, M. Fozooni, Z. Ding, S. Lambotharan, and O. A. Dobre, "Robust energy-efficient design for MISO non-orthogonal multiple access systems," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7937–7949, 2019.

[35] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "A framework of robust transmission design for IRS-aided MISO communications with imperfect cascaded channels," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5092–5106, Aug. 2020.

[36] N. K. Kundu and M. R. Mckay, "A deep learning-based channel estimation approach for MISO communications with large intelligent surfaces," *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–6, 2020.

[37] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5060, Oct. 2006.

[38] A. Agrawal, J. Andrews, J. Cioffi, and T. Meng, "Iterative power control for imperfect successive interference cancellation," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 878–884, May 2005.

[39] G. Zhou, C. Pan, H. Ren, K. Wang, M. D. Renzo, and A. Nallanathan, "Robust beamforming design for intelligent reflecting surface aided MISO communication systems," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1658–1662, 2020.

[40] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.

[41] E. Björnson, E. Jorswieck *et al.*, "Optimal resource allocation in coordinated multi-cell systems," *Foundations and Trends® in Communications and Information Theory*, vol. 9, no. 2–3, pp. 113–381, 2013.

[42] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for intelligent reflecting surface enhanced wireless networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[43] W. Qiang and Z. Zhongli, "Reinforcement learning model, algorithms and its application," in *Proceedings of the IEEE International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)*, 2011, pp. 1143–1146.

[44] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, Dec. 2013.

[45] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.

[46] 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," *arXiv preprint arXiv:1509.02971*, Sep. 2015.

[47] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1587–1596.

[48] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Machine learning*, vol. 8, no. 3, pp. 293–321, May 1992.

[49] M. Kulin, T. Kazaz, I. Moerman, and E. De Poorter, "End-to-end learning from spectrum data: A deep learning approach for wireless signal identification in spectrum monitoring applications," *IEEE Access*, vol. 6, pp. 18 484–18 501, Mar. 2018.

[50] B. Matthiesen, A. Zappone, K.-L. Besser, E. A. Jorswieck, and M. Debbah, "A globally optimal energy-efficient power control framework and its efficient implementation in wireless interference networks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3887–3902, Jun. 2020.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, Dec. 2014.