
Article:
Waraiet, Abdulhamed Khaled E orcid.org/0000-0001-8818-6935, Cumanan, Kanapathippillai orcid.org/0000-0002-9735-7019, Ding, Zhiguo et al. (1 more author) (2024) Robust Design for IRS-assisted MISO-NOMA Systems:A DRL-Based Approach. IEEE wireless communications letters. pp. 592-596. ISSN 2162-2345

https://doi.org/10.1109/LWC.2023.3335622

Robust Design for IRS-assisted MISO-NOMA Systems: A DRL-Based Approach

Abdulhamed Waraiet, *Graduate Student Member, IEEE*, Kanapathippillai Cumanan, *Senior member, IEEE*, Zhiguo Ding, *Fellow, IEEE*, and Octavia A. Dobre, *Fellow, IEEE*

***Abstract*—In this paper, we propose a robust design for an intelligent reflecting surface (IRS)-aided multiple-input single-output (MISO) non-orthogonal multiple access (NOMA) system. In particular, the ergodic sum-rate maximization problem is formulated by taking into account the channel uncertainties of both direct links and the reflected links through IRS elements. The unbounded channel uncertainties with imperfect channel estimation are mathematically modelled based on the statistical channel state information (CSI) error model. However, the formulated ergodic sum-rate maximization problem with the outage-constraints is not jointly convex in terms of the beamforming vectors and the phase shifts of IRS elements, and hence it cannot be solved with the conventional optimization algorithms. To address the non-convexity issues and develop a joint design, the challenging robust design is reformulated as a reinforcement learning (RL) environment. Two deep RL agents are developed to jointly optimize the beamforming vectors and phase shifts of the IRS elements with the channel uncertainties and quality of service constraints. Simulation results are provided to validate the performance of the proposed agents for both fixed and dynamic channels.**

***Index Terms*—DRL, MISO-NOMA, IRS, Imperfect CSI.**

I. Introduction

Non-orthogonal multiple access (NOMA) has been considered as one of the promising multiple access techniques for 6G and beyond. By utilizing superposition coding (SC) at the transmitter and successive interference cancellation (SIC) at the receiver, NOMA enables encoding more than one user in the same resource block. This leads to higher spectral and energy efficiencies compared to its orthogonal counterparts and also enables massive connectivity [1]. Numerous studies in the literature demonstrate the superiority of NOMA over conventional orthogonal multiple access (OMA) techniques [2].

Recently, the intelligent reflecting surface (IRS) technology has shown great potential in enhancing the quality of the communication links. Therefore, IRS-aided multiple antenna NOMA systems have also been subject to extensive studies, as they offer enhanced link reliability with interference mitigation [3]–[5]. However, the resource allocation problem becomes more challenging with such advancements, and often requires problem-specific hand-crafted algorithms with higher computational complexities. To address these complexity issues, deep learning (DL)-based approaches have been considered as a viable alternative for solving the resource allocation problems, which is proved to be particularly useful for latency-constrained applications, thanks to their low deployment complexity. In [6], a DL framework for the beamforming design of a MISO system was proposed. However, DL models require labelled data. This is the main drawback of DL, which is addressed by combining deep learning with reinforcement learning (RL) into a single framework: deep reinforcement learning (DRL).

DRL combines the function approximation capabilities of deep neural networks with the sequential decision making framework of RL. Different DRL techniques have been exploited to solve a variety of resource allocation problems in IRS-assisted and cognitive NOMA systems [7]–[10]. However, in all aforementioned studies, perfect channel state information knowledge at the transmitter (CSIT) and receiver (CSIR) is assumed, which is not the case in practice. This often leads to unrealistic results and, in some cases, inaccurate conclusions. This work is motivated by the fact that the model-based robust design algorithms often suffer from exponential computational complexities, which render them impractical for latency-sensitive applications. In addition, another motivation is the lack of DRL-based robust designs in the literature. Therefore, the aim of this work is to propose a joint robust design framework which takes into account both the imperfect CSI at the transmitter and SIC at the receiver with unbounded channel uncertainties [11], [12]. In particular, we propose a DRL framework to jointly optimize the beamforming vectors and the IRS phase shifts. Unlike conventional optimization methods, the proposed robust DRL model has much lower computational complexity. This widens the applicability of the proposed approach to communication systems with stringent latency requirements. To the best of the authors' knowledge, this is the first work to address the joint robust design problem for IRS-assisted MISO-NOMA systems with unbounded channel uncertainties using actor-critic DRL agents. The contributions of this work are summarized as follows: 1) The non-convex long-term system sum-rate optimization problem with outage constraints is reformulated as an RL environment. 2) Then, two actor-critic DRL agents, namely proximal policy optimization (PPO) and twin-delayed deep deterministic policy gradient (TD3), are developed to solve the robust design problem. 3) Through simulation results, we show the convergence properties of the agents and the achieved system sum-rates, as well as their robust performance, for both fixed and dynamic channels.

A. Waraiet and K. Cumanan are with the University of York, UK. Z. Ding is with the University of Manchester, UK. O. A. Dobre is with Memorial University, Canada. The work of K. Cumanan was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/X01309X/1. The work of O. A. Dobre was supported in part by the Natural Sciences and Engineering research Council of Canada (NSERC) through its Discovery program.

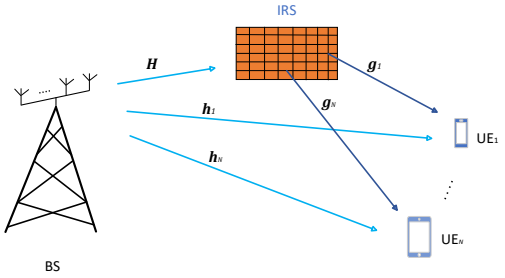

Fig. 1: IRS-assisted MISO-NOMA system model.

II. System Model and Problem Formulation

We consider a single-cell, IRS-aided MISO-NOMA downlink system as illustrated in Fig. 1, where a base station (BS) equipped with N antennas, serves K single antenna user equipment (UE). The M passive IRS elements provide a reflected path to the signal in the downlink, in addition to the direct channel between the BS and the UEs. The phase shifts of the IRS elements are designed at the BS and transmitted to the IRS hardware through a feedback link [13]. The BS transmits the precoded signal as $\mathbf{x} = \sum_{i=1}^{N} \mathbf{w}_i s_i$, where $\mathbf{w}_i \in \mathbb{C}^{N\text{x}1}$ is the beamforming vector for UE_i, s_i is the information bearing symbol for UE_i. The received signal at UE_i is expressed as

$$y_i = \mathbf{h}_i^{\text{H}}\mathbf{x} + \mathbf{g}_i^{\text{H}}\boldsymbol{\Upsilon}\mathbf{H}\mathbf{x} + z_i, \forall i \in \{1, ..., K\}, \tag{1}$$

where $\mathbf{h}_i \in \mathbb{C}^{N\text{x}1}$ is the Rayleigh fading channel vector modelled as $\mathbf{h}_i = \frac{h_i}{\sqrt{d_{id}^{\alpha_{b\to u}}}}$, where $\alpha_{b\to u}$ and d_{id} are the path-loss exponent and the distance between UE_i and the BS, respectively. $\mathbf{g}_i \in \mathbb{C}^{M\text{x}1}$ is the channel vector between the IRS and UE_i, modelled as Rician fading and expressed as $\mathbf{g}_i = \frac{1}{\sqrt{d_i^{\alpha_{irs\to u}}}}\big(\sqrt{\frac{\xi}{1+\xi}}\mathbf{g}_{LoS} + \sqrt{\frac{1}{1+\xi}}\mathbf{g}_{nLoS}\big)$, with $\xi = 1$ as the Rician factor. $\boldsymbol{\Upsilon} \in \mathbb{C}^{M\text{x}M}$ is the phase shifts matrix of the IRS elements, $\mathbf{H} \in \mathbb{C}^{M\text{x}N}$ is the channel matrix between the BS and the IRS, which is also assumed to be Rician fading channel expressed as $\mathbf{H} = \frac{1}{\sqrt{d_i^{\alpha_{irs\to u}}}}\big(\sqrt{\frac{\xi}{1+\xi}}\mathbf{H}_{LoS} + \sqrt{\frac{1}{1+\xi}}\mathbf{H}_{nLoS}\big)$. The received signal at UE_i can be written in a more compact form as follows:

$$y_i = \big(\mathbf{h}_i^{\text{H}} + \mathbf{v}^{\text{H}}\mathbf{Q}_i\big)\mathbf{x} + z_i, \forall i \in \mathcal{K}, \tag{2}$$

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

For the channel uncertainty model, we consider two cases. One considers a partial uncertainty case where the direct channel is assumed to be perfectly known at the BS, while the reflected channel is imperfectly estimated. This is motivated by the fact that the reflected channel is more challenging to estimate accurately than the direct channel due to the passive IRS elements [14]. The other case is the full uncertainty model where both direct and cascaded channels are imperfect. The true channels can be expressed as

$$\begin{aligned} \mathbf{Q}_i &= \hat{\mathbf{Q}}_i + \Delta\mathbf{Q}_i, \forall i \in \mathcal{K}, \\ \mathbf{h}_i &= \hat{\mathbf{h}}_i + \Delta\mathbf{h}_i, \forall i \in \mathcal{K}, \end{aligned} \tag{3}$$

where $\hat{\mathbf{Q}}_i$, $\hat{\mathbf{h}}_i$ are the estimated channels available at the BS, and $\Delta\mathbf{Q}_i$, $\Delta\mathbf{h}_i$ are the unknown, unbounded errors for the cascaded and direct channels, respectively. The considered error model encompasses channel estimation errors due to the white Gaussian noise and insufficient pilot sequences in practical wireless communication systems. Therefore, the unknown errors are drawn from a circularly symmetric complex Gaussian distribution and expressed as $\Delta\mathbf{q}_i \sim \mathcal{CN}(\mathbf{0}, \Lambda_r)$, $\Delta\mathbf{h}_i \sim \mathcal{CN}(\mathbf{0}, \Lambda_d)$, where $\Delta\mathbf{q}_i = \text{vec}(\Delta\mathbf{Q}_i)$, $\Lambda_r \in \mathbb{C}^{MN\text{x}MN}$ and $\Lambda_d \in \mathbb{C}^{N\text{x}N}$ are the positive semidefinite error covariance matrices for the reflected and the direct channels, respectively [12]. Furthermore, the variances of the unknown error terms are functions of their corresponding estimated channels and are expressed as $\beta_{i,r}^2 = \lambda_r^2||\mathbf{q}_i||_2^2$, $\mathbf{q}_i = \text{vec}(\hat{\mathbf{Q}}_i) \in \mathbb{C}^{MT\text{x}1}$ and $\beta_{i,d}^2 = \lambda_d^2||\hat{\mathbf{h}}_i||_2^2$ for the reflected and direct channels, respectively. λ_r and λ_d relate to the uncertainty of the CSI measurement for the reflected and cascaded channels, respectively, and both are in the range $(0, 1]$.

Before proceeding to the signal-to-interference-plus-noise ratio (SINR) expressions, deciding a decoding order is crucial for NOMA systems. In this paper, we adopt a channel-strength based decoding order, i.e., $||\hat{\tilde{\mathbf{h}}}_1||_2^2 \geq ||\hat{\tilde{\mathbf{h}}}_2||_2^2 \geq ... \geq ||\hat{\tilde{\mathbf{h}}}_K||_2^2$, where $\hat{\tilde{\mathbf{h}}}_i = (\hat{\mathbf{h}}_i^{\text{H}} + \mathbf{v}^{\text{H}}\hat{\mathbf{Q}}_i)$ is the estimated version of the final combined channel $\tilde{\mathbf{h}}_i$ at the BS. Then, the decoding order set is $\zeta = \{1, 2, ..., K\}$, where UE_1 is the strongest UE that will perform SIC to decode and eliminate $\text{UE}_2, ..., \text{UE}_K$ signals before decoding its own. Therefore, the SINR expression of the UE_i's signal at UE_l, under the full channel uncertainty model can be expressed as

$$\gamma_l^i = \frac{|\tilde{\mathbf{h}}_l\mathbf{w}_i|^2}{\sum_{j=i+1}^{K}|(\Delta\mathbf{h}_l^{\text{H}} + \mathbf{v}^{\text{H}}\Delta\mathbf{Q}_l)\mathbf{w}_j|^2 + \sum_{j=1}^{i-1}|\tilde{\mathbf{h}}_l\mathbf{w}_j|^2 + \sigma_l^2}, \tag{4}$$

where $\sum_{j=i+1}^{K}|(\Delta\mathbf{h}_l^{\text{H}} + \mathbf{v}^{\text{H}}\Delta\mathbf{Q}_l)\mathbf{w}_j|^2$ is the sum of the SIC residuals due to imperfect channel estimation at UE_l, and $\sum_{j=1}^{i-1}|\tilde{\mathbf{h}}_l\mathbf{w}_j|^2$ is the sum of the interference caused by stronger UEs. When only the partial uncertainty model is considered, $\Delta\mathbf{h}$ is removed from (4). Therefore, the achievable rate for UE_i is expressed as

$$R_i = log_2\big(1 + \min_{l\in\{i,i+1,..,K\}}(\gamma_l^i)\big). \tag{5}$$

Since the considered uncertainty model is unbounded, it is challenging to design a set of beamformers that achieve the required rates regardless of the channel uncertainties. Therefore, the aim of this work is to maximize the long-term system sum-rate under the outage constraints as follows:

$$\underset{\mathbf{w}_i, \mathbf{v}}{\text{maximize}} \quad \mathbb{E}\Big\{\sum_{t=1}^{\infty}\sum_{i=1}^{K}\delta^{t-1}R_i^t\Big\} \tag{6a}$$

$$\text{subject to} \quad p_i \triangleq \Pr\{\gamma_l^i \geq 2^{R_i^{min}} - 1\} \geq \Gamma, \forall i \in \mathcal{K}, \tag{6b}$$

$$\sum_{i=1}^{N}||\mathbf{w}_i||_2^2 \leq P_{max}, \ \forall i \in \mathcal{K}, \tag{6c}$$

$$|v_m|^2 = 1, \ 0 \leq \theta_m \leq 2\pi, \ \forall m \in \mathcal{M}, \tag{6d}$$

where (6a) is the expected value of the discounted cumulative system sum-rate, $\Gamma \in (0, 1]$ is the non-outage probability, P_{max} is the maximum transmit power budget at the BS, and v_m and θ_m correspond to the amplitude and the phase shift for

the m-th IRS element, respectively. This optimization problem is not jointly convex in terms of the beamforming vectors and the IRS phase shifts. This is due to the coupled nature of the optimization variables in (6a) and (6b). Such optimization problem is generally NP-hard, which makes it more challenging to solve using conventional optimization techniques. Furthermore, since RL agents aim to maximize their long-term reward, they can be utilized to solve this challenging problem. In this paper, we propose a DRL-based approach to solve the robust design problem. This approach is motivated by the low deployment complexity and the generalized solutions generated by the DRL agents.

III. Proposed DRL Based Algorithm

In order to develop DRL agents to solve the original optimization problem, we need to reformulate it into an RL environment. There are three features that define any RL environment, namely the state vector $\mathbf{s}^t$, the actions vector $\mathbf{a}^t$, and the reward function r^t. A DRL agent aims to maximize its reward through interactions with the environment. At a current system state $\mathbf{s}^t$, the agent takes an action $\mathbf{a}^t$, and the environment provides a new state $\mathbf{s}^{t+1}$ and a reward r^t based on the utility of the action taken by the agent. Therefore, the agent tries to maximize its reward by taking actions that yield higher rewards. Hence, the design of an accurate reward function in RL is crucial, being the only performance indicator the agent understands.

In this paper, we use the variables of the optimization problem (6a) as the actions vector of the agent, i.e., $\mathbf{a}^t = \left[\mathbf{w}_1^t, .., \mathbf{w}_K^t, v_1^t, .., v_M^t\right]^T$. The state vector is defined as the power of each beamforming vector of the previous time-step, the normalized variances of the channel uncertainty terms, the achieved rates for the previous time-step, and the actions vector of the previous time-step. Hence, $\mathbf{s}^t = \left[||\mathbf{w}_1^{t-1}||_2^2, .., ||\mathbf{w}_K^{t-1}||_2^2, \beta_1^{2,t}, .., \beta_K^{2,t}, R_1^{1,t-1}, .., R_K^{K,t-1}, \mathbf{a}^{t-1}\right]^T$. Finally, there are two reward functions depending upon the action taken by the agent. In case the action satisfies the quality of service (QoS) for all UEs, then, the agent is rewarded by the system sum-rate at that time-step, expressed as $r^t = \sum_{i=1}^K R_i^t$; otherwise, the agent is punished with the following negative reward, $r^t = \sum_{i=1}^K min(0, R_i^t - R_i^{min})$, which is the sum of the rate deficit across all UEs. This reward function deters the agent from taking actions that result in a negative reward, and provides incentive for the agent to increase the system sum-rate to maximize its reward. Since we work with neural networks that do not support complex numbers, the actions and state vectors can only contain real numbers. One of the proposed solutions to this problem is to represent each complex vector by two real vectors, which will be adopted in this work [6]. As a result, each beamforming vector $\mathbf{w}_i \in \mathbb{C}^{N\text{x}1}$, will be mapped to two real vectors, and therefore, will be $\mathbf{w}_i \in \mathbb{R}^{2N\text{x}1}$. This is also true for the complex phase shift values. Therefore, $\mathbf{a}^t \in \mathbb{R}^{2(NK+M)\text{x}1}$, while $\mathbf{s}^t \in \mathbb{R}^{2N+\frac{N(N+1)}{2}+2(NK+M)}, N \geq 2$, where the expression $\frac{N(N+1)}{2}$ calculates the number of all possible rates in the MISO-NOMA system. Another issue we need to address when using DRL agents is to ensure that the actions vector values fall within the feasible region. Therefore, normalization and scaling by the maximum power is necessary to ensure optimal agent performance. Hence, $P_{\text{total}}^t = \sum_{i=1}^K ||\mathbf{w}_i^t||_2^2$ is the unconstrained total power at time-step t; then, the feasible scaling factor can be written as $\kappa^t = \sqrt{\frac{P_{\max}}{P_{\text{total}}^t}}$, which is used to scale the beamforming vectors. Furthermore, to ensure that the IRS phase shifts satisfy the amplitude constraint in (6d), they are normalized as $\frac{v_m^t}{|v_m^t|}$, while θ_m can be directly mapped to a feasible angle.

The proximal policy optimization (PPO) is an on-policy, actor-critic, DRL agent which optimizes a stochastic policy [15]. The PPO agent was proposed mainly to address the slow training and low sample efficiency issues of the trust region policy optimization agent. It utilizes the action-advantage function to improve its policy. Therefore, the objective of the actor network can be expressed as [15]

$$L(\Phi) = \mathbb{E}\Big[\min(RA_t(\Phi)\hat{\mathbf{A}}_\mathbf{t}, \text{clip}(RA_t(\Phi), 1-\epsilon, 1+\epsilon)\hat{\mathbf{A}}_\mathbf{t})\Big], \tag{7}$$

where Φ is the parameterized policy of the actor network, $RA_t(\Phi) = \frac{\pi_\Phi(\mathbf{a}^t|\mathbf{s}^t)}{\pi_{\Phi_{old}}(\mathbf{a}^t|\mathbf{s}^t)}$ is the ratio between the new and the old policies, and $\hat{\mathbf{A}}_\mathbf{t}$ is the advantage function at time step t. This clipped objective keeps the new policy from deviating too far from the old policy to prevent policy breaking issues during training. On the other hand, the twin-delayed deep deterministic policy gradient (TD-DDPG) or (TD3) for short, is an off-policy, actor-critic, DRL agent which optimizes a deterministic policy [16]. The TD3 agent was proposed to address the overestimation problems in the baseline DDPG agent by utilizing two critic networks instead of one, among other enhancements. The actor network of the TD3 agent optimizes the following objective [16]:

$$\nabla_\Phi J = \frac{1}{B}\sum_{i=1}^{B} G_{ai} G_{\pi a}, \tag{8}$$

where B is the mini-batch size, G_{ai} is the gradient of the critic with the minimum value with respect to the action taken by the actor, and $G_{\pi a}$ is the gradient of the actor output with respect to its network parameters.

Algorithm 1 summarizes the steps of obtaining DRL-based robust beamforming and IRS solutions. Note that Algorithm 1 highlights the essential steps taken by each agent to solve the problem. Agent-specific steps are omitted from the Algorithm for simplicity.

TABLE I: System parameters summary.

System parameter	Value
Cell radius	200 m
Transmit power	30 dbm
Noise power	-90 dbm
λ_r	0.02
λ_d	0.03
$\alpha_{b\to irs}, \alpha_{irs\to u}$	2
$\alpha_{b\to u}$	2.5
Target rate R_i^{min} (fixed channels)	1 bit/s/Hz
Target rate R_i^{min} (dynamic channels)	0.3 bit/s/Hz

Another important aspect of the proposed framework is mapping the non-outage probability Γ to the number of training episodes and time-steps. It is up to the agent designer to select adequate values so that the agent has formed a policy that is robust against channel uncertainties. This is determined by the number of error observations considered during training, given a proper selection of hyperparameters. In this work, we introduce a new set of error values for each new episode during training for both fixed and dynamic channels. Therefore, the more episodes the agent is trained for, the more robust its policy becomes.

In terms of computational complexity, we assume that the offline training can be afforded and focus on the deployment (online) complexity. Therefore, the complexity can be described as a feed-forward pass through the agent's actor network. Hence, there are $L+1$ matrix-vector multiplications, where L represents the number of hidden layers in the actor network. Also, there are $L+1$ activation operations, including output activation. Since layer activation is an element-wise operation, it has a complexity of $\mathcal{O}(n)$, where n is the number of neurons in the layer. Assuming that all hidden layers have the same number of neurons n, then, the model complexity can be expressed as $\mathcal{O}(In+Ln^2+nO+Ln)$, where I and O represent the size of the input and output vectors, respectively. Therefore, the final worst run-time complexity expression can be reduced to $\mathcal{O}\big(max(In, n^2, nO)\big)$. Since the previous action is part of the state vector, multiple steps might be required to achieve satisfactory results.

Algorithm 1 DRL-based Robust Design

Initialise: agent's actor, critic and their target networks
while $Episode \leq TotalEpisodes$ **do**
  Reset environment and obtain an initial state
  while $Step \leq TotalSteps$ **do**
    Take action $\mathbf{a}^t$
    Recover complex-valued beamforming vectors $\mathbf{w}_i, \forall i$ and IRS phase shifts vector $\mathbf{v}$
    Evaluate UE SINRs according to (4)
    Calculate the corresponding reward
    Set $\mathbf{s}^{t+1} = \mathbf{s}^t$
    $Step = Step + 1$
  end while
  if Time to update policy **then**
    Copy main network parameters to target networks
  end if
  $Episode = Episode + 1$
end while
Output: $\mathbf{w}_1^*, .., \mathbf{w}_K^*, v_1^*, .., v_M^*$.

IV. Simulation and Numerical Results

We consider the system model mentioned in Section II, in which a BS equipped with $N = 4$ antennas serves $K = 4$ single antenna UEs. The numerical values for the system parameters are summarized in Table I. As for the agents, we train a TD3 agent with a single actor network and two critic networks. Furthermore, we train a PPO agent with one actor and one critic. The number of neurons, $n = 128$, is set for the actor networks, while $n = 300$ is set for the critic networks, for both agents. Table II summarizes the hyperparameters and training parameters used to train the two agents. Both partial and full channel uncertainty models are considered for the fixed channel case, while only the full channel uncertainty model is considered for dynamic channels. A benchmark scheme based on the semidefinite programming and the zero-forcing beamforming is used as a baseline.

TABLE II: Summary of the agents hyperparameters.

Hyperparameter	Value
Critics learning rate	0.001
Actor learning rate	0.0007
Policy update frequency (TD3)	2
Discount factor	0.99
Smoothness factor (TD3)	0.0002
Replay buffer size (TD3)	$1e05$
Minibatch size	128
Clip factor (PPO)	$0.07 - 0.12$
Entropy loss weight (PPO)	$0.005 - 0.007$
Number of Episodes, Time-steps (fixed channels)	$700, 500$
Number of Episodes, Time-steps (dynamic channels)	$1500, 500$

Fig. 2 shows the convergence of the agents throughout the training period for both fixed and dynamic channels. It can be observed that both agents have similar convergence properties with different variances. In the dynamic channels case, the agents are trained on a set of 10 channels, sampled to reflect the distance between the BS and the maximum cell radius. As expected, the convergence curves for both agents show higher variance, due to the fact that the channels used for training are inherently different, and therefore, the rewards obtained by the agents vary accordingly. The TD3 agent achieves higher average reward than PPO in both cases, while the PPO agent shows more stable convergence to a lower average reward level.

To assess the performance of both agents in terms of robustness, each agent is tested for 1000 episodes, with 10 steps per episode. The average robustness performance for both agents for fixed and dynamic channels is illustrated in Fig. 3. The top figure shows the robustness performance against the reflected-channel estimation quality λ_r. TD3 marginally outperforms PPO for both channel scenarios in the case $M = 32$, achieving a 99% robustness score at $\lambda_r = 0.02$, compared to PPO's 93% for the fixed-channel full uncertainty model. On the other hand, PPO performs better than TD3 for fixed-channels scenario with a 10% margin for the case $M = 128$ under full uncertainty model. The bottom figure shows the robustness against channel uncertainties for higher and lower target rates than those used for training. TD3 yields a higher robustness score for the dynamic-channels case, achieving an average non-outage performance of 90% at 0.4 bit/s/Hz compared to PPO's 45%, for both IRS elements cases. However, for the fixed-channels scenario with a higher training target rate, PPO scores better than TD3 across all categories where the target rate is greater than 1 bit/s/Hz. PPO is able to maintain a robustness score over 70% when the requested target rate is 20% higher than that used for training. Overall, the PPO agent's performance is more consistent, which suggests that it

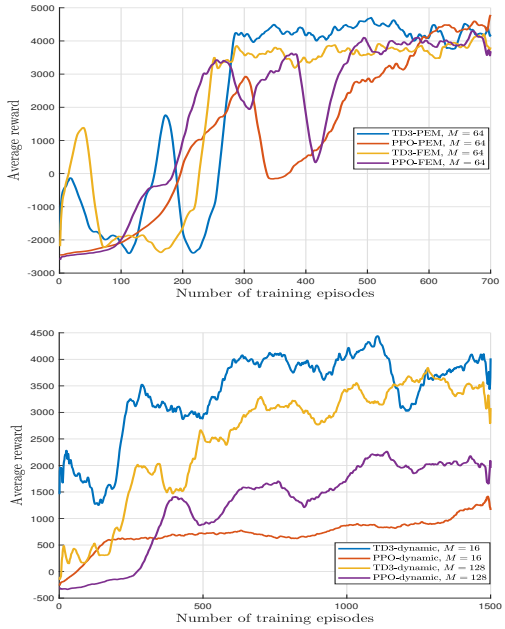

Fig. 2: PPO and TD3 convergence for fixed (top) and dynamic (bottom) channels.

is more suitable for fixed-channel scenarios with a wider range of IRS elements. On the other hand, the TD3 agent performs better in the dynamic-channels case, especially with a smaller number of IRS elements. Furthermore, PPO is much faster to train and requires less hyperparameter tuning than TD3.

V. Conclusion

In this paper, we proposed a DRL-based robust beamforming design for a downlink, IRS-assisted MISO-NOMA system. In particular, an outage-constrained robust design problem with unbounded channel uncertainties was considered. The non-convex optimization problem was reformulated into an RL environment. The PPO and TD3 agents were then developed to efficiently solve the robust design problem jointly in terms of the beamforming vectors and phase shifts of IRS elements. Both agents were able to achieve robust performance for both fixed and dynamic channels. The agents were capable of generalizing their robust policies to any set of channels within the cell radius in the case of dynamic channels. Furthermore, the computational complexity of the trained actor network for both agents is considered very low for the robust design problem, which makes DRL methods more attractive for latency-sensitive applications.

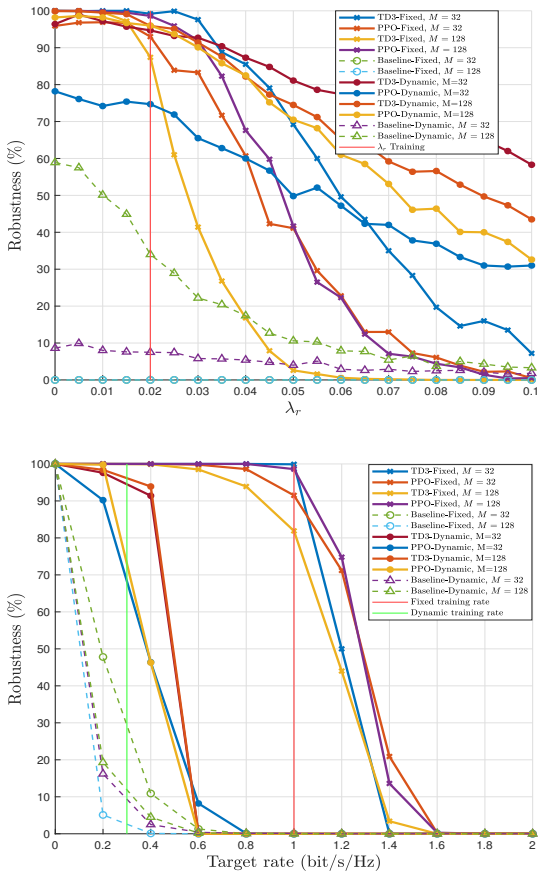

Fig. 3: PPO and TD3 non-outage probability versus estimation quality (top) and target rate (bottom) values.

References

[1] Y. Liu *et al.*, "Nonorthogonal Multiple Access for 5G and Beyond," Proc. IEEE, vol. 105, no. 12, pp. 2347-2381, Dec. 2017.

[2] Q. Sun *et al.*, "On the Ergodic Capacity of MIMO NOMA Systems," IEEE Wireless Commun. Lett., vol. 4, no. 4, pp. 405-408, Aug. 2015.

[3] Y. Li *et al.*, "Joint Beamforming Design in Multi-Cluster MISO NOMA Reconfigurable Intelligent Surface-Aided Downlink Communication Networks," IEEE Trans. Commun., vol. 69, no. 1, pp. 664-674, Jan. 2021.

[4] J. Zhu *et al.*, "Power Efficient IRS-Assisted NOMA," IEEE Trans. Commun., vol. 69, no. 2, pp. 900-913, Feb. 2021.

[5] Z. Ding et al., "A State-of-the-Art Survey on Reconfigurable Intelligent Surface-Assisted Non-Orthogonal Multiple Access Networks," Proc. IEEE, vol. 110, no. 9, pp. 1358-1379, Sept. 2022.

[6] W. Xia *et al.*, "A Deep Learning Framework for Optimization of MISO Downlink Beamforming," IEEE Trans. Commun., vol. 68, no. 3, pp. 1866-1880, Mar. 2020.

[7] X. Xie *et al.*, "A Reinforcement Learning Approach for an IRS-assisted NOMA Network," arXiv preprint arXiv:2106.09611 (2021).

[8] X. Gao *et al.*, "Machine Learning Empowered Resource Allocation in IRS Aided MISO-NOMA Networks," IEEE Trans. Wireless Commun., vol. 21, no. 5, pp. 3478-3492, May 2022.

[9] Z. Ding *et al.*, "No-Pain No-Gain: DRL Assisted Optimization in Energy-Constrained CR-NOMA Networks," IEEE Trans. Commun., vol. 69, no. 9, pp. 5917-5932, Sept. 2021.

[10] M. Shehab *et al.*, "Deep Reinforcement Learning Powered IRS-Assisted Downlink NOMA," IEEE Open J. Commun. Soc., vol. 3, pp. 729-739, 2022.

[11] A. Agrawal *et al.*, "Iterative Power Control for Imperfect Successive Interference Cancellation," IEEE Trans. Wireless Commun., vol. 4, no. 3, pp. 878-884, May 2005.

[12] G. Zhou *et al.*, "A Framework of Robust Transmission Design for IRS-Aided MISO Communications With Imperfect Cascaded Channels," IEEE Trans. Signal Process., vol. 68, pp. 5092-5106, Aug. 2020.

[13] C. Pan *et al.*, " aided MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer." IEEE J. Sel. Areas Commun., vol. 38, no. 8, pp. 1719-1734, Aug. 2020.

[14] N. Kundu *et al.*, "A Deep Learning-based Channel Estimation Approach for MISO Communications With Large Intelligent Surfaces." 2020 IEEE 31st Annual PIMRC.

[15] J. Schulman *et al.*, "Proximal Policy Optimization Algorithms." arXiv preprint arXiv:1707.06347 (2017).

[16] S. Fujimoto *et al.*. "Addressing Function Approximation Error in Actor-Critic Methods." ICML, vol. 80, pp. 1587-1596, 2018.