Received 21 March 2024; accepted 11 April 2024. Date of publication 19 April 2024; date of current version 9 May 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3391658

Outage-Constrained Robust Resource Allocation Framework for IRS-Empowered NOMA Systems: A DRL-Based Joint Design

ABDULHAMED WARAIET (Graduate Student Member, IEEE), AND KANAPATHIPPILLAI CUMANAN (Senior Member, IEEE)

School of Physics, Engineering and Technology, University of York, YO10 5DD York, U.K.

CORRESPONDING AUTHOR: A. WARAIET (e-mail: abdulhamed.waraiet@york.ac.uk)

This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) underx Grant EP/X01309X/1.

ABSTRACT In this paper, we propose a robust resource allocation framework for an intelligent reflecting surface (IRS)-assisted multiple-input single-output (MISO) non-orthogonal multiple access (NOMA) system. In particular, a long-term robust sum-rate maximization problem is considered. The impacts of imperfect channel estimation on both the transmitter and the receiver are taken into account with an outage-constrained robust design approach. More specifically, the statistical error model is used to model the unbounded channel uncertainty in the system. However, the joint robust resource allocation problem is a mixed-integer optimization problem, which cannot be solved directly using conventional optimization algorithms. A correlation-based user pairing algorithm is proposed to group the users into clusters. Furthermore, the resource allocation problem with clustered users is reformulated as a reinforcement learning environment. Subsequently, a twin-delayed deep deterministic policy gradient (TD3) agent is developed to solve the outage-constrained robust resource allocation problem. Extensive simulation results are provided to demonstrate the superior performance of the developed TD3 agent over existing algorithms in the literature.

INDEX TERMS IRS-assisted MISO-NOMA systems, robust resource allocation techniques, deep reinforcement learning, imperfect CSI.

I. INTRODUCTION

NON-ORTHOGONAL multiple access (NOMA) has been proposed as one of the promising multiple access (MA) techniques for next-generation wireless networks. By utilizing the superposition coding (SC) at the transmitter and the successive interference cancellation (SIC) at the receiver, NOMA offers better spectral and energy efficiencies as well as user-fairness compared to conventional orthogonal multiple access (OMA) techniques [1], [2]. This enables NOMA as a promising MA candidate to realize massive connectivity in 6G and beyond by efficiently allocating scarce radio resources. Instead of multiplexing users in time or frequency, NOMA utilizes the power domain multiplexing to multiplex users. Therefore, NOMA systems generally require more efficient and accurate power allocation algorithms to mitigate the interference levels and enable smooth practical implementations of the SIC at the receiver.

Multiple antenna communications with their additional degrees of freedom have proven to be an effective interference-mitigation technique. Hence, multiple antenna NOMA systems have been studied extensively and demonstrated significant performance gains over OMA-based multiple antenna systems due to their combined spectral efficiency and interference-suppression capabilities [3], [4], [5]. In [6], Hanif et al. proposed an iterative algorithm for a multiple-input single-output (MISO)-NOMA system with the sum-rate maximization objective. The authors in [7] proposed a semidefinite relaxation (SDR)-based approach for the optimal beamforming design problem in MISO-NOMA systems with the transmit power minimization objective. In addition, the authors in [8] proposed a sequential convex

optimization solution for MISO-NOMA systems with the aim of maximizing global energy efficiency.

More recently, intelligent reflecting surface (IRS)-assisted multiple antenna systems have received significant attention from both industry and academia, thanks to the additional link reliability they introduce to the conventional wireless communication systems [9]. In IRS-assisted systems, the phase shifts of the passive IRS elements can be programmed to steer the incoming signal to the desired direction, hence, increasing the channel strength between the transmitter and the receiver(s). The IRS-assisted MISO-NOMA system model has been subject to extensive studies recently to reap the combined benefits of the NOMA, multiple antennas, and IRS techniques. In particular, the work in [10] considered the multi-cluster beamforming and IRS phase shifts design for the transmit power minimization objective, while the energy efficiency objective was considered in [11]. Xie et al. proposed a solution for the max-min fairness system objective. However, combining such sophisticated techniques often leads to tractability problems. Hence, model-based approaches break down the joint-design optimization problem into several subproblems, then, each problem is solved separately in an iterative manner. However, the downside of such approaches is that the overall computational complexity of the proposed solution is often prohibitively high, which severely limits their practical utility, especially for latency-sensitive future wireless networks [12], [13], [14], [15].

Machine learning-based methods have proved to be a viable alternative to model-based solutions for highly complex resource allocation problems in wireless communication systems. The deep learning framework has been applied to channel tracking and estimation, and beamforming design [16], [17], [18], [19]. However, since supervised deep learning requires labelled data for training, it can only be applied to problems solved a priori, which restricts the deep learning-based algorithms to problems that are already solved, albeit not on a large scale. Deep reinforcement learning (DRL) -which combines deep learning and reinforcement learning (RL) into a single framework- addresses the shortcomings of deep learning as an optimization tool. In RL, an active agent learns how to solve the problem through trial and error without any human supervision, and therefore, does not require labels for training and learning [20].

Recently, DRL has been applied to a wide variety of problems in the wireless communications domain. The work in [21] proposed a deep deterministic policy gradient (DDPG)-based design to maximize the sum-rate in cognitive-radio NOMA systems. Meng *et al*. also applied DDPG to solve the downlink dynamic power control problem for maximizing the system sum-rate. The application of the DRL framework has also been extended to IRS-aided NOMA systems. The work in [22] adopted the zero-forcing beamforming (ZFBF) technique while utilizing a deep Q-network (DQN) agent for optimizing phase shifts of the IRS elements. Xie *et al*. used DDPG to jointly optimize the beamforming vectors and IRS phase shifts for the sum-rate maximization problem [23]. The work in [24] proposed a multi-agent DRL-based design that jointly optimizes the subcarrier assignment, power allocation, and IRS phase shifts in NOMA-assisted semi-grant-free systems, while the resource allocation problem for NOMA-unmanned aerial vehicle system was considered in [25].

However, there are still practical issues facing the aforementioned works. First, all of these works assume perfect channel state information (CSI) at the base station (BS) which is extremely challenging in practice. Furthermore, the imperfect CSI at the transmitter and the receiver have severe implications in NOMA systems since the receivers utilize SIC to unlock the additional gains of NOMA. In addition, providing some guarantees of performance under channel uncertainties leads to a more complicated optimization problem that is more challenging to solve in a reasonable time, especially for latency-sensitive applications. Therefore, the performance of DRL-based methods for clustered IRS-assisted MISO-NOMA systems with imperfect CSI and SIC remains an open issue. The second challenge is that most of the literature focuses on a simplified version of the system objective. The work in [22] while considering a cluster-based IRS-assisted MISO-NOMA system, does not take into account cluster power allocation nor the quality-of-service requirements in the proposed design, both of which have a significant impact on the agent selection and the problem environment design. Furthermore, the DQN agent utilized to solve the problem cannot be applied to problems with large continuous action spaces as DQN is restricted to discrete action space problems. The work in [24] uses a DQN agent to solve the discrete channel-assignment problem, while a DDPG agent is utilized to solve the power allocation problem. However, since the BS and the user equipment units (UEs) are assumed to be equipped with a single antenna, no beamforming design is considered. Additionally, while the work in [26] considered a DRL-based approach to solve the sum-rate maximization problem through joint active and passive beamforming design, the number of SIC operations required by the strongest UE grows linearly with the number of UEs in the systems, leading to a practically unscalable and highly complex receiver.

Motivated by the impractical assumptions and the lack of a unified and scalable framework in the DRL literature, we propose a DRL-based joint design framework to solve an outage-constrained robust resource allocation problem in an IRS-assisted MISO-NOMA system. In particular, a correlation-based user-pairing algorithm is developed to limit the number of UEs in each cluster leading to a more scalable implementation of SIC-based receivers. Then, the NOMA principle is applied in each cluster to increase the spectral efficiency of the system. Moreover, the proposed DRL-based design jointly optimizes the clusters and UEs power allocation, and IRS phase shifts, while taking into account the outage-constrained QoS requirements. The ergodic sum-rate

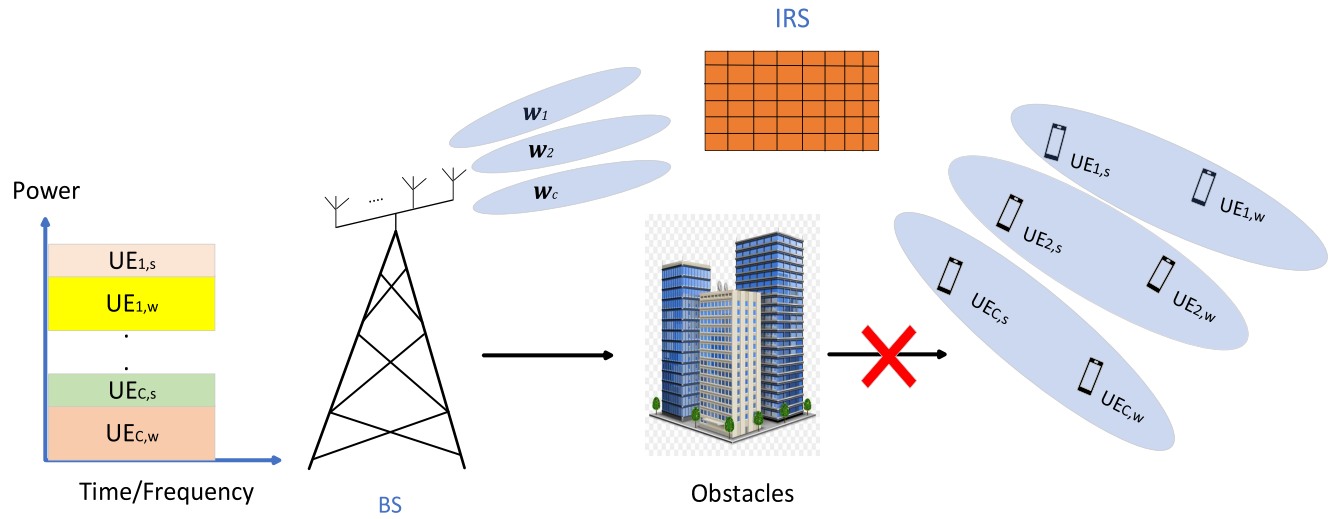

FIGURE 1. Cluster-based IRS-assisted Downlink MISO-NOMA system.

maximization is used as the objective function for the considered system. In addition, the statistical error model is used to describe the channel uncertainty which leads to an outage-constrained robust design. Furthermore, the proposed DRL-based design has much lower deployment computational complexity compared to the conventional optimization methods in the literature while still achieving competitive performance. To the best of the authors' knowledge, this is the first work that proposes a framework for clustering and actor-critic-based resource allocation in IRS-assisted MISO-NOMA systems. The contributions of this work are summarized as follows:

- By assuming a blocked direct path between the BS and the UEs due to obstacles, the BS communicates with the UEs through the IRS. In addition, the statistical error model is to used express the channel uncertainty in the system. However, the formulated robust design problem with the ergodic sum-rate maximization objective is a mixed-integer optimization problem which is challenging to solve. The user-pairing problem is isolated and solved first to reduce the complexity of the problem. Then, the zero-forcing (ZF) principle is adopted to design the beamforming vectors.
- The robust resource allocation problem is still non-convex due to the coupled optimization variables. Therefore, the problem is reformulated into an RL environment. Then, a twin-delayed deep deterministic policy gradient (TD3)-based algorithm is developed to solve the reformulated joint resource allocation problem.
- By providing the complexity analysis for the proposed DRL-agent's architecture, we show that the deployment computational complexity of the proposed algorithm is much less than existing conventional optimization algorithms, which makes the DRL-based design more attractive for latency-stringent applications in future wireless networks.
- The competitive performance of the proposed algorithm is illustrated through extensive simulation results for both fixed and dynamic-channel scenarios. Furthermore, the results show the TD3-based design outperforms existing conventional and other DRL-based benchmark schemes in the literature.

A. ORGANIZATION

The rest of the paper is organized as follows. Section II presents the system and channel uncertainty models. The joint robust design problem is formulated in Section III. In addition, the user-clustering algorithm is also developed. In Section IV, the problem is reformulated into an RL environment and a TD3-based algorithm is developed to solve the reformulated problem. The simulation results are presented in Section V. Finally, Section VI concludes this work.

B. NOTATION

Bold lowercase and uppercase letters are used to represent vectors and matrices, respectively, while standard normal letters denote scalar quantities. $\mathbf{Y}^{\dagger}$ and $\mathbf{y}^{\mathrm{H}}$ denote the pseudoinverse of the matrix $\mathbf{Y}$ and the hermitian transpose of the vector $\mathbf{y}$, respectively. $|.|$ and $||.||$ refer to the absolute value and the Euclidean norm of a vector, respectively. $||.||_2$ and $||.||_F$ represent the L_2 and the Frobenius norms, respectively. $\mathbf{Card}(\mathbf{y})$ denotes the cardinality of the vector $\mathbf{y}$. $\mathbb{C}$ and $\mathbb{R}$ refer to the sets of complex and real numbers, respectively. $\mathbb{E}$ represents the expectation operator.

II. SYSTEM AND CHANNEL UNCERTAINTY MODELS

We consider a downlink transmission of an IRS-assisted MISO-NOMA system in which the BS is equipped with N transmit antenna and serves $2K$ single antenna UEs as shown in Figure 1. To increase the system capacity, the UEs are paired into $\mathcal{C} = \{1, \ldots, C\}$ clusters, and the

NOMA principle is applied in each cluster to mitigate the impact of intra-cluster interference and increase the overall spectral efficiency. Furthermore, to reduce the number of SIC operations carried out by each receiver, we limit the number of UEs in each cluster to 2 [27], [28]. Since the additional gains of NOMA require distinctively different channel conditions, the UEs are divided into two sets, namely the stronger UEs set $\mathcal{S}$, and the weaker UEs set $\mathcal{W}$. We use $\text{UE}_{c,s}$ and $\text{UE}_{c,w}$ to denote the stronger and the weaker UE with the better and the worse channel condition in the c-th cluster, respectively. The IRS consists of M passive elements which are controlled by the BS through a feedback link [29]. In addition, we assume that the direct links between the BS and the UEs are blocked due to obstacles, and therefore, the BS communicates with the UEs only through the IRS link. Hence, the received signal at $\text{UE}_{c,i}$ can be expressed as

$$y_{c,i} = \mathbf{g}_{c,i}^{\text{H}} \Phi \mathbf{G} \sum_{c=1}^{C} \mathbf{w}_c x_c + z_{c,i}, \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \quad (1)$$

where $\mathbf{g}_{c,i} \in \mathbb{C}^{M\text{x}1}$ represents the channel between $\text{UE}_{c,i}$ and the IRS, $\mathbf{G} \in \mathbb{C}^{M\text{x}N}$ denotes the channel between the BS and the IRS, and $\Phi = \text{diag}(v_1, \ldots, v_M) \in \mathbb{C}^{M\text{x}M}$ is the diagonal IRS phase shifts matrix, and $v_m = \zeta_m e^{j\theta_m}$. In this paper, we assume an ideal reflection at the IRS elements, i.e., $|v_m|^2 = 1, m = 1, \ldots, M$. $\mathbf{w}_c \in \mathbb{C}^{N\text{x}1}$ is the beamforming vector for cluster c, while $x_c = \sqrt{\alpha_{c,s}} s_{c,s} + \sqrt{\alpha_{c,w}} s_{c,w}$ is the superposition coded signal transmitted by the BS to the UEs in the c-th cluster. In addition, $s_{c,s}$ and $s_{c,w}$ are the normalized information symbols for the stronger and weaker UEs in the c-th cluster, respectively. The $\alpha_{c,s}$ and $\alpha_{c,w}$ are the power allocation coefficients for the stronger and the weaker UEs in the c-th cluster, respectively. The $z_{c,i}$ is the additive white Gaussian noise with zero mean and variance $\sigma_{c,i}^2$. The received signal at $\text{UE}_{c,i}$ can be expressed in a more compact form as

$$y_{c,i} = \mathbf{h}_{c,i} \sum_{c=1}^{C} \mathbf{w}_c x_c + z_{c,i}, \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \quad (2)$$

where $\mathbf{h}_{c,i} = \mathbf{v}^{\text{H}} \mathbf{Q}_{c,i} \in \mathbb{C}^{1\text{x}N}$ is the final channel vector, $\mathbf{v} = \text{vec}(\Phi) \in \mathbb{C}^{M\text{x}1}$, and $\mathbf{Q}_{c,i} = \text{diag}(\mathbf{g}_{c,i}^{\text{H}}) \mathbf{G} \in \mathbb{C}^{M\text{x}N}$ is the cascaded channel for $\text{UE}_{c,i}$. To unlock the additional gains of NOMA, the receivers need to perform one or more SIC operations. Therefore, designing a decoding order is crucial in NOMA systems. Since the number of UEs is limited to two per cluster in this paper, and given that $||\mathbf{h}_{s,i}||_2 \gg ||\mathbf{h}_{w,i}||_2$, we assume a fixed decoding order in which the stronger UE carries out a single SIC operation to eliminate the weaker UE's signal, then proceeds to decode its own signal. Hence, the total number of SIC operations required in the system is equal to C. Therefore, non-SIC receivers can be admitted to the considered system if they have moderate to weaker channel conditions. Note that in general, however, the process of designing optimal decoding order in NOMA systems is non-trivial [7], [13].

A. CHANNEL UNCERTAINTY MODEL

Due to the random nature of the wireless transmissions, uncertainties in the wireless channel estimation are inevitable. Furthermore, with the introduction of the IRS, accurate channel estimation becomes even more challenging due to the passive elements in the IRS [30], [31]. Channel estimation and quantization errors are two of the main contributors to the imperfect channel estimation in wireless communication systems [14], [32]. However, the two are often modelled differently with the quantization errors considered to belong to a norm-bounded region, while channel estimation errors are modelled statistically using unbounded error models [31], [33]. On the other hand, multiple antenna communication systems make use of the beamforming principle to enhance the system performance by exploiting the CSI at the transmitter. However, to achieve the optimal beamforming gains, perfect CSI is required at the transmitter. Unfortunately, having perfect CSI at the transmitter is extremely challenging to obtain in practical settings due to the aforementioned channel uncertainties. Therefore, robust design algorithms that take into account channel imperfections are more suitable for studying and analysing the system performance under practical conditions. In this paper, we assume that the channel uncertainties are the result of the imperfect channel estimation. Note that in NOMA systems, channel imperfections at the receiver lead to SIC degradation which is also taken into account. In particular, this paper aims to propose a robust resource allocation strategy that takes into account the imperfect CSI in the system.

The statistical error model has been extensively used to describe distortions in the acquired channel due to thermal noise, estimation errors, and insufficient pilot samples [34], [35], [36]. If the channel statistics are not known, the least square estimator is typically used to estimate the channel coefficients at the receiver. Alternatively, when the channel statistics are available, the linear minimum mean square error estimator is normally used to exploit the additional information and obtain more accurate channel estimates. Therefore, if the noise is assumed to be an additive and white Gaussian process, then, it is straightforward to interpret that the difference between the estimated and actual channels can be expressed statistically [37], [38]. Therefore, the following error model is considered for the cascaded channel [31]:

$$\mathbf{Q}_{c,i} = \hat{\mathbf{Q}}_{c,i} + \Delta \mathbf{Q}_{c,i}, \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \quad (3)$$

where $\hat{\mathbf{Q}}_{c,i}$ is the estimated channel known at the BS, while $\Delta \mathbf{Q}_{c,i}$ is an additive, unknown, and unbounded error. The unknown errors are drawn from a circularly symmetric complex Gaussian distribution and are expressed as $\Delta \mathbf{q}_{c,i} \sim \mathcal{CN}(\mathbf{0}, \Lambda)$, where $\Delta \mathbf{q}_{c,i} = \text{vec}(\Delta \mathbf{Q}_{c,i})$, and $\Lambda \in \mathbb{C}^{MN\text{x}MN}$ is the positive semidefinite error covariance matrix for the cascaded channel. In addition, the variance of the unknown term is a function of the estimated cascaded channel and is

expressed as

$$\beta_{c,i}^2 = \lambda^2 \left\| \hat{\mathbf{q}}_{c,i} \right\|_2^2, \ \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \tag{4}$$

where $\hat{\mathbf{q}}_{c,i} = \text{vec}(\hat{\mathbf{Q}}_{c,i}) \in \mathbb{C}^{MN\text{x}1}$, and $\lambda \in (0, 1]$ relates to the uncertainty of the CSI estimate [31]. Therefore, the unbounded error is related to the system parameters through the size of the cascaded channel matrix and the estimation quality. Based on these assumptions, the next section defines the signal-to-interference-plus-noise ratio (SINR) and the corresponding achievable rates.

B. SINR AND ACHIEVABLE RATES

SINR is one of the most widely used metrics for measuring the performance of wireless communication systems. For the considered cluster-based design, the SINR of the stronger UE in the c-th cluster can be defined as

$$\gamma_{c,s} = \frac{\left|\mathbf{h}_{c,s}\mathbf{w}_c\right|^2 P_c \alpha_{c,s}}{\left|\left(\mathbf{v}^{\text{H}} \Delta\mathbf{Q}_{c,s}\right)\mathbf{w}_c\right|^2 P_c \alpha_{c,w} + \sum_{\substack{k=1 \\ k\neq c}}^{C} \left|\mathbf{h}_{c,s}\mathbf{w}_k\right|^2 P_k + \sigma_{c,s}^2}, \quad \forall s \in \{\mathcal{S}\}, c \in \mathcal{C}, \tag{5}$$

where P_c is the allocated power for the c-th cluster. The term $|(\mathbf{v}^{\text{H}}\Delta\mathbf{Q}_{c,s})\mathbf{w}_c|^2 P_c \alpha_{c,w}$ represents the SIC residual and is the result of the imperfect channel estimation at the receiver side, while $\sum_{\substack{k=1 \\ k\neq c}}^{C} |\mathbf{h}_{c,s}^{\text{H}}\mathbf{w}_k|^2 P_k$ is the inter-cluster interference experienced at $\text{UE}_{s,c}$, and $\sigma_{c,s}^2$ is the noise power. Similarly, the SINR of the weaker UE in the c-th cluster when decoding its own signal is defined as

$$\gamma_{c,w}^{c,w} = \frac{\left|\mathbf{h}_{c,w}\mathbf{w}_c\right|^2 P_c \alpha_{c,w}}{\left|\mathbf{h}_{c,w}\mathbf{w}_c\right|^2 P_c \alpha_{c,s} + \sum_{\substack{k=1 \\ k\neq c}}^{C} \left|\mathbf{h}_{c,w}\mathbf{w}_k\right|^2 P_k + \sigma_{c,w}^2}, \quad \forall w \in \{\mathcal{W}\}, c \in \mathcal{C}. \tag{6}$$

Note that since $\text{UE}_{c,w}$ does not carry out any SIC operations, it experiences both intra-cluster and inter-cluster interference. Furthermore, the SINR of $\text{UE}_{c,s}$ for decoding $\text{UE}_{c,w}$'s signal can be expressed as

$$\gamma_{c,w}^{c,s} = \frac{\left|\mathbf{h}_{c,s}\mathbf{w}_c\right|^2 P_c \alpha_{c,w}}{\left|\mathbf{h}_{c,s}\mathbf{w}_c\right|^2 P_c \alpha_{c,s} + \sum_{\substack{k=1 \\ k\neq c}}^{C} \left|\mathbf{h}_{c,s}\mathbf{w}_k\right|^2 P_k + \sigma_{c,s}^2}, \quad c \in \mathcal{C}. \tag{7}$$

Therefore, the achieved SINR of $\text{UE}_{c,w}$ is defined as

$$\gamma_{c,w} = \left(1 + min\left(\gamma_{c,w}^{c,s}, \gamma_{c,w}^{c,w}\right)\right) \quad c \in \mathcal{C}. \tag{8}$$

The achievable rates of both stronger and weaker UEs in the c-th cluster can be expressed as

$$\begin{aligned} R_{c,s} &= log_2\left(1 + \gamma_{c,s}\right), \\ R_{c,w} &= log_2\left(1 + \gamma_{c,w}\right), \forall s \in \{\mathcal{S}\}, w \in \{\mathcal{W}\}, c \in \mathcal{C}. \end{aligned} \tag{9}$$

In the next section, the problem formulation of the robust design for the considered system is provided with details.

III. PROBLEM FORMULATION

The aim of this work is to propose a joint robust design framework for a long-term performance-based resource allocation in IRS-assisted MISO-NOMA systems. In particular, we consider the objective of maximizing the ergodic system sum-rate under channel uncertainties while taking into account the dynamics of the system over multiple time-slots [3], [21], [39]. Therefore, the long-term outage-constrained joint robust design problem with the sum-rate maximization objective can be formulated as

$$\max_{\mathbf{w}_c, \mathbf{v}, P_c, \alpha_{c,i}, b_{s,w}} \mathbb{E}\left\{\sum_{t=1}^{\infty} \delta^{t-1} \sum_{c=1}^{C} \left[R_{c,s}^t + R_{c,w}^t\right] b_{s,w}^t\right\} \tag{10a}$$

$$p_i \triangleq \Pr\left\{\gamma_{c,i} \geq 2^{R_{c,i}^{min}} - 1\right\} \geq \Gamma, \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \tag{10b}$$

$$||\mathbf{w}_c||_2^2 = 1, c \in \mathcal{C}, \tag{10c}$$

$$\sum_{c=1}^{C} P_c \leq P_{max}, c \in \mathcal{C}, \tag{10d}$$

$$\alpha_{c,s}^t + \alpha_{c,w}^t = 1, c \in \mathcal{C}, s \in \mathcal{S}, w \in \mathcal{W} \tag{10e}$$

$$\sum_{c=1}^{C} b_{s,w}^t \leq 1, b \in \{0, 1\}, c \in \mathcal{C}, \tag{10f}$$

$$|\mathbf{v}_m|^2 = 1, \ 0 \leq \theta_m \leq 2\pi, m = 1, \ldots, M, \tag{10g}$$

where $\mathbb{E}$ is the expectation operator, δ^{t-1} is the discount factor which is explained in the problem reformulation section, $\Gamma \in (0, 1]$ is the non-outage probability that the resource allocation strategy satisfies the quality-of-service (QoS) constraint for each UE, and $b_{s,w}^t \in \{0, 1\}$ is the binary UE pairing coefficient. The outage constraint in (10b) guarantees that the QoS requirements of the UEs are achieved with probability Γ, while the constraint in (10c) ensures normalized power for all the beamforming vectors. The constraints in (10d) and (10e) represent the maximum available transmit power for all clusters and the UEs power allocation coefficients within each cluster, respectively. The pairing constraint in (10f) guarantees that each stronger UE is only paired with a single weaker UE and vice versa. Finally, the constraints in (10g) guarantee a unit modulus and a feasible phase shift for the IRS elements.

The joint design problem in (10) is a mixed-integer optimization problem and is known to be NP-hard [40]. Note that even without considering the binary constraint, the problem in (10a) is still non-convex and NP-hard [6], [41], [42], and therefore, cannot be solved directly using conventional optimization methods. The formulated optimization problem is non-trivial and challenging to solve efficiently for the following reasons:

- The objective function is not jointly convex in terms of the optimization variables.
- The expectation operator prevents defining a closed-form expression for the objective function in (10a) since approximation methods cannot be directly applied.

- The outage constraints in (10b) do not admit closed-form solutions [34].
- The UE pairing variable in (10f) is restricted to a binary set, resulting in a mixed-integer optimization problem.

To reduce the complexity of the proposed solution, the user clustering subproblem is tackled first. Then, the rest of the variables are optimized to maximize the system sum-rate.

A. USER PAIRING

UE pairing is considered one of the enabling techniques in multi-user NOMA systems for future wireless networks [27], [28], [43]. In addition, it has been shown that pairing a stronger UE with a weaker UE leads to enhanced overall performance in NOMA systems [44], [45]. Hence, there are two design criteria for UE pairs selection that directly affect the system sum-rate performance in NOMA networks, correlation and channel-gain difference between the paired UEs in a cluster [22], [46]. Since each cluster is served with a single beam, a higher UE correlation within the cluster translates to a lower level of intra-cluster interference experienced by the weaker UE, while sufficient channel-gain difference ensures smooth SIC operation at the stronger UE. However, since the IRS phase shifts are designed at the BS, the phase shifts could be tuned to adjust the channel-gain differences after the cluster design. Therefore, the proposed algorithm is solely based on the initial correlation between the UEs.

The basic premise of the proposed successive UE pairing algorithm (SUPA) is to pair each UE in $\mathcal{S}$ with a single UE from $\mathcal{W}$ to form a cluster, assuming that there are $2K$ UEs in total. Furthermore, since the IRS phase shift values have a direct impact on the channel coefficients, the UE pairing is carried out with a fixed IRS vector, i.e., the initial phase shift values stay constant during the pairing process. To this end, we define the correlation coefficient between two UEs in the system as [46]

$$\epsilon_{i,j} = \frac{\left\| \hat{\mathbf{h}}_i . \hat{\mathbf{h}}_j \right\|_2}{\left\| \hat{\mathbf{h}}_i \right\|_2 \left\| \hat{\mathbf{h}}_j \right\|_2}, \forall i \in \mathcal{S}, \forall j \in \mathcal{W}, \tag{11}$$

where $\hat{\mathbf{h}}_k, k \in \{i, j\}$, is the estimated final channel for UE_k and is known at the BS. Algorithm 1 provides the key steps for the proposed UE pairing design. Therefore, executing Algorithm 1 will eliminate the binary constraint in (10f). The next section presents the robust resource allocation framework for a given UE pairing configuration.

IV. RL FRAMEWORK FOR ROBUST RESOURCE ALLOCATION

With given UE pairs using Algorithm 1, the remaining resource allocation problem is expressed as

$$\max_{\mathbf{w}_c, \mathbf{v}, P_c, \alpha_{c,i}} \mathbb{E}\left\{ \sum_{t=1}^{\infty} \delta^{t-1} \sum_{c=1}^{C} \left[R_{c,s}^t + R_{c,w}^t \right] \right\} \tag{12a}$$

$$\text{s.t.} (10b), (10c), (10d), (10e), (10g). \tag{12b}$$

Algorithm 1 Successive User Pairing Algorithm

1: **Initialise:** UEs sets $\mathcal{S}$,$\mathcal{W}$, initial IRS vector $\mathbf{v}_{init}$, and UE clusters $c \in \mathcal{C}$
2: Calculate the final estimated channels at the BS using $\hat{\mathbf{h}}_{c,i} = \mathbf{v}_{init}^{\text{H}} \mathbf{Q}_{c,i}, \forall c \in \mathcal{C}, \forall i \in \mathcal{S}, \mathcal{W}$
3: Sort all $\text{UE}_i, \forall i \in \mathcal{S}$, according to their channel norms such that $||\hat{\mathbf{h}}_1||_2 \geq ||\hat{\mathbf{h}}_2||_2 \geq \ldots \geq ||\hat{\mathbf{h}}_K||_2$
4: **for** $i = 1 : K, i \in \mathcal{S}$ **do**
5: **for** $j = 1 : K, j \in \mathcal{W}$ **do**
6: Calculate the correlation coefficient between UE_i and UE_j according to (11)
7: **end for**
8: Find $j' = \text{argmax}(Corr_{i,j}), \forall j \in \mathcal{W}$
9: Assign UE_i and $\text{UE}_{j'}$ to cluster $c(i)$
10: Set $\hat{\mathbf{h}}_{j'} \leftarrow \mathbf{0}, j' \in \mathcal{W}$
11: **end for**
12: **Output:** $\{\text{UE}_{1,s}, \text{UE}_{1,w}\}, \ldots, \{\text{UE}_{C,s}, \text{UE}_{C,w}\}$

Unfortunately, the optimization problem in (12a) is still non-convex and there is no standard approach to solve it efficiently. To further simplify the problem, the ZFBF is utilized to tackle the beamforming design constraint in (10c) [47].

A. THE ZERO-FORCING BEAMFORMING

The ZFBF is a low-complexity technique in which the channel knowledge at the transmitter is exploited to design the beamforming vectors. More importantly, under the perfect CSI assumption, the ZFBF provides a closed-form solution to the beamforming design problem with a reasonable trade-off between complexity and performance [48]. In addition, the ZFBF has been extensively used in the literature as one of the beamforming designs for sum-rate maximization [46], [48], [49]. The basic principle behind the ZFBF is to design a beamforming vector $\mathbf{w}_k$ that achieves zero interference to all other $\text{UE}_i, k \neq i$. This is formalized as

$$\frac{\mathbf{h}_i}{||\mathbf{h}_i||_2} \mathbf{w}_k = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{if } k \neq i. \end{cases} \tag{13}$$

However, since we consider a multi-cluster NOMA system, the ZFBF vector can only be designed based on a single channel for each cluster, not both. Hence, in this paper, the ZFBF vectors are designed based on the stronger UE's channel in each cluster to reduce the inter-cluster interference in the system. Furthermore, since the perfect CSI is not available at the BS for the considered robust design, the true channels are replaced with their estimated counterparts. Therefore, there will be an interference leakage as a result of the imperfect beamforming design based on the estimated channel. Thereby, the expression in (13) can written as

$$\frac{\hat{\mathbf{h}}_i}{||\hat{\mathbf{h}}_i||_2} \mathbf{w}_k = \begin{cases} 1 & \text{if } k = i \\ > 0 & \text{if } k \neq i. \end{cases} \tag{14}$$

Note that the fact that $\frac{\hat{\mathbf{h}}_i}{||\hat{\mathbf{h}}_i||_2}\mathbf{w}_k > 0$, for $k \neq i$, is unavoidable due to the imperfect CSI available at the BS. Furthermore, this leakage term is the source of the inter-cluster interference experienced by the stronger UEs in each cluster. Hence, we define $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_C]$ as the matrix that contains the ZFBF vectors for all clusters, and $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_{1,s}^T, \ldots, \hat{\mathbf{h}}_{C,s}^T]^T$ as the estimated channel matrix that contains the stronger UEs' channel vectors, where $\hat{\mathbf{h}}_{c,s}$ is a row vector. Then, the ZFBF matrix is calculated as follows [46]:

$$\mathbf{W} = \left(\hat{\mathbf{H}}\right)^{\dagger}, \tag{15}$$

where $(\hat{\mathbf{H}})^{\dagger} = \hat{\mathbf{H}}^{\mathrm{H}}(\hat{\mathbf{H}}\hat{\mathbf{H}}^{\mathrm{H}})^{-1}$ is Pseudo-inverse of the stronger UEs estimated channel matrix $\hat{\mathbf{H}}$.

Therefore, in this work, the robust resource allocation is realized through the accurate and joint optimization of the IRS phase shifts, cluster and UE power allocation as explained in the next section.

B. PROBLEM REFORMULATION

By tackling the UE pairing and beamforming design problems, the robust resource allocation problem is reduced to the following optimization problem

$$\max_{\mathbf{v}, P_c, \alpha_{c,i}} \mathbb{E}\left\{\sum_{t=1}^{\infty} \delta^{t-1} \sum_{c=1}^{C}\left[R_{c,s}^t + R_{c,w}^t\right]\right\} \tag{16a}$$

$$\text{s.t.} (10b), (10d), (10e), (10g). \tag{16b}$$

Unfortunately, the problem is still non-convex due to the coupled optimization variables and the outage constraint and hence, cannot be optimized jointly using conventional optimization algorithms. Therefore, in order to develop a joint robust design, the problem in (16a) is reformulated into a reinforcement learning environment.

It is well-known that optimizing a system objective under uncertainty or stochastic environment can be modelled as a Markov decision process (MDP) [50]. The RL framework is one of the most effective methods to solve the control problem in MDPs, especially in model-free systems where the transition probability between the states is unknown [51]. The RL framework consists of two entities, the agent which is the active entity that takes actions, and the environment which encloses everything else except the agent. At time-step t, given a state s^t, the agent takes an action a^t. Based on the action taken by the agent, the environment provides the next state s^{t+1}, and the reward r^t which can either be positive or negative, depending on the utility of the taken action. Therefore, through trial and error, the agent aims to maximize its reward by forming an optimal policy $\pi^*(s, a)$ that maps any state to the best action that yields the highest reward. Hence, the RL framework transforms the optimization problem into a series of sequential decision-making steps in which the optimization variables are updated to maximize some utility function.

To reformulate the robust design problem into an RL environment, the state, action and reward entities must be clearly defined.

- The action space $\mathbf{a}^t$: Since the value of the objective is a function of the optimization variables, they are intuitively selected as the actions space of the RL environment. In particular, the actions space vector at time-step t is expressed as

$$\mathbf{a}^t = \left[P_1^t, \ldots, P_C^t, \alpha_{1,w}^t, \ldots, \alpha_{C,w}^t, \mathbf{v}^t\right]^{\mathrm{T}}. \tag{17}$$

Note that since $\alpha_{c,s}^t = 1 - \alpha_{c,w}^t, \forall c \in \mathcal{C}$, only the power allocation coefficients for the weaker UEs are included in the actions vector. Furthermore, since we will be using a deep neural network (DNN) architecture that is only compatible with real numbers, complex vectors are represented using real values in this paper. In particular, and without the loss of generality, since $\mathbf{v} \in \mathbb{C}^{M\mathrm{x}1}$, then, $\mathbf{v} \in \mathbb{R}^{2M\mathrm{x}1}$, where $Re\{\mathbf{v}\} \in \mathbb{R}^{M\mathrm{x}1}$ and $Im\{\mathbf{v}\} \in \mathbb{R}^{M\mathrm{x}1}$ are the real and the imaginary parts of the IRS vector $\mathbf{v}$, respectively [19]. Therefore, we can write $\mathbf{a}^t \in \mathbb{R}^{(2K+2M)\mathrm{x}1}$ as a vector with only real values.

- The state space $\mathbf{s}^t$: To ensure that the state space of the environment includes the necessary information from the original robust design problem, we include the previous action as part of the state vector. Furthermore, since the correlation coefficient between the paired UEs is affected by the IRS phase shifts as highlighted by (12), the correlation coefficients vector is also included in the state space. Additionally, the channel gain between each UE pair is included in the state vector. The channel gain difference defined as the dB ratio between the two channels is used and can be expressed as

$$\rho_{i,j} = 10\, log_{10}\left(\frac{||\hat{\mathbf{h}}_i||_2}{||\hat{\mathbf{h}}_j||_2}\right), \forall i \in \mathcal{S}, j \in \mathcal{W}. \tag{18}$$

Finally, to help the agent evaluate itself during training, the achieved rates of the previous time-step are also taken into account as part of the state space, Therefore, the state space is expressed as

$$\mathbf{s}^t = \left[\mathbf{a}^{t-1}, \epsilon_1^{t-1}, \ldots, \epsilon_C^{t-1}, \rho_1^{t-1}, \ldots, \rho_C^{t-1}, R_{1,s}^{t-1}, \ldots, R_{C,w}^{t-1}\right]^{\mathrm{T}}, \tag{19}$$

where $\mathbf{s}^t \in \mathbb{R}^{(6K+2M)\mathrm{x}1}$. Furthermore, when training for the dynamic-channels environment, the variances of the estimated channels are also included as part of the state space. Therefore, the state vector for the dynamic-channels case is expressed as

$$\mathbf{s}_{\mathrm{dyn}}^t = \left[\beta_{1,s}^2, \ldots, \beta_{C,w}^2, \mathbf{a}^{t-1}, \epsilon_1^{t-1}, \ldots, \epsilon_C^{t-1}, \rho_1^{t-1}, \ldots, \rho_C^{t-1}, R_{1,s}^{t-1}, \ldots, R_{C,w}^{t-1}\right]^{\mathrm{T}}, \tag{20}$$

where $\mathbf{s}^t \in \mathbb{R}^{(8K+2M)\mathrm{x}1}$. Note that since the variance of the estimated channel is closely related to the estimation error according to (4), including this information in the state space helps the agent in forming a more robust policy under the dynamic-channels environment.

- The reward function r^t: Defining an appropriate reward function is crucial in the RL framework as it is the only feedback that indicates the utility of the actions taken by the agent at any time-step t during training. In addition, since the objective in the original robust design problem (10a) is to maximize the long-term system sum-rate, the system sum-rate at time-step t is selected as the reward. In addition, the sum of the correlation coefficients and the channel gain ratios are added to the system sum-rate to incentivise the agent to increase the correlation and the channel gain difference between the stronger and the weaker UEs in each cluster. Therefore, the reward function is expressed as

$$r^t = \sum_{c=1}^{C}\left(R_{c,s}^t + R_{c,w}^t\right) + \sum_{c=1}^{C}\epsilon_c^t + \sum_{c=1}^{C}\rho_c^t, c \in \mathcal{C}. \quad (21)$$

  Furthermore, to discourage the agent from taking actions that do not satisfy the QoS constraints, the following reward function is used to punish the agent:

$$r^t = \sum_{k=1}^{2K}\min\left(R_k^t - R_k^{min}, 0\right), \quad (22)$$

  where $r^t < 0$ always hold in (22). Therefore, after each action taken by the agent, the environment uses the positive reward function in (21) in case the action satisfies the QoS constraints, otherwise, the environment uses the negative reward function in (22). The details of how the reward function is utilized by the agent during training are discussed in the agent's architecture section.

Since RL agents in general cannot directly solve optimization problems, scaling and normalization of the actions space is often required to ensure that the actions taken by the agent are within the feasible region of the optimization variables. Therefore, to guarantee that the cluster power allocation strategy selected by the agent at time-step t adheres to the maximum power constraint in (10d), the feasible cluster power vector is expressed

$$\bar{\mathbf{P}}^t = \frac{P_{max}}{\sum_{c=1}^{C} P_c^t}\mathbf{P}^t, \quad (23)$$

where $\mathbf{P}^t = [P_1^t, \ldots, P_C^t]^T$ is the cluster power allocation vector generated by the agent and $\bar{\mathbf{P}}^t = [\bar{P}_1^t, \ldots, \bar{P}_C^t]^T$ is the scaled clusters power allocation vector. Similarly, to ensure the unit modulus for each IRS element, the feasible value is expressed as

$$\bar{v}_m^t = \frac{v_m^t}{|v_m^t|}, m = 1, \ldots, M. \quad (24)$$

Note that the angle θ_m can be directly mapped to the feasible region. Therefore, the IRS vector recovery process involves obtaining the $2M$ elements from the "real-only" actions vector, and then reorganizing them into a single complex-valued vector, i.e., $\mathbf{v} \in \mathbb{C}^{M\times 1}$. Additionally, after normalization, the optimized IRS vector can be directly applied to calculate the final UE channels.

C. THE ROBUST TD3-BASED ALGORITHM

The RL agents like the Q-learning and the state-action-reward-state-action (SARSA) are called tabular methods because they use tables to keep track of the Q-values for each state-action pair [52], [53]. However, since these agents are only capable of handling discrete state and action spaces, their practical utility is severely limited as most practical problems have continuous state and action spaces.

Actor-critic agents which are state-of-the-art in DRL can handle continuous action and state spaces, and therefore, eliminate the tabular requirement which restricted the earlier RL agents. Consequently, actor-critic DRL agents have been applied to a much wider set of problems in the wireless communications domain [20].

In this paper, the proposed robust resource allocation framework is developed based on the TD3 agent [54]. The TD3 agent is an off-policy actor-critic DRL agent which optimizes a deterministic policy. To address the policy break issue in the baseline DDPG agent [55], the TD3 agent uses two critics instead of one, among other enhancements. Furthermore, since off-policy agents are more sample efficient than their on-policy counterparts, thanks to the replay buffer $\mathcal{B}$ which is used to save and reuse past training samples. This translates to faster learning during training. Finally, unlike stochastic agents, the TD3 optimizes a deterministic policy which is easier to implement.

The TD3 agent consists of two main parts: the actor or the policy DNN and the critic DNN. As the name implies, the actor DNN denoted μ is the one responsible for taking actions. The input to the actor's DNN is the state vector. Therefore, for a trained TD3 agent, the actor's DNN can be expressed mathematically as

$$\mu(\mathbf{s}) = \mathbf{a}^*, \quad (25)$$

where $\mathbf{s}$ is an arbitrary state vector and $\mathbf{a}^*$ is the optimal actions vector. However, since the actor network is initialized randomly at the beginning of the training, the actor DNN cannot evaluate itself. Hence, the critic DNN is used to assess the performance of the actor's network during the training phase. The critic DNNs $\phi_i, i = 1, 2$, are responsible for criticizing the actions taken by the policy network μ. In particular, the critic DNNs predict how good/bad the action taken by the agent is through the Q-value. Hence, each critic DNN takes in the current action which is generated by the actor network and the current state as inputs and generates a corresponding Q-value which is then passed to the actor's DNN. Therefore, the mathematical expression for the critic DNNs is expressed as

$$\phi_i(\mathbf{s}, \mathbf{a}) = Q^*, i = 1, 2. \quad (26)$$

where Q^* is the optimal Q-value for the state-action pair. Note that (26) highlights the importance of the critic DNNs. Therefore, training the critic DNNs is discussed next.

Similar to the DQN and the DDPG agents, the TD3 agent uses target networks to generate the training targets. Target networks are delayed copies of the actor's and the critics'

DNNs. Furthermore, the TD3 agent also utilizes a replay buffer which stores past experiences to further stabilise the learning process. μ' and ϕ_i', $i = 1, 2$, represent the actor's and the critics' target networks. To elaborate, the training starts by sampling a batch of experiences $\mathcal{L}$ from the replay buffer. However, we focus on the process of a single experience for the sake of simplicity. A single experience $\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\}$, also called a tuple, is randomly sampled from the replay buffer. Then, the target for the selected tuple is calculated as follows:

$$\zeta\left(r^t, \mathbf{s}^t\right) = r^t + \delta \min_{i=1,2} \phi_i'\left(\mathbf{s}^{t+1}, \mu'\left(\mathbf{s}^{t+1}\right)\right), i = 1, 2, \quad (27)$$

where $\delta \in (0, 1]$ is the discount factor that determines the current value of future rewards. Therefore, selecting a smaller δ value implies that the agent is myopic, i.e., only cares about short-term reward. On the other hand, selecting a δ value that is closer to 1 means that the agent is interested in maximizing its long-term reward. Note that according to (27), both the actor's target and critics' target networks are used to calculate $\zeta(r^t, \mathbf{s}^t)$. After obtaining the target using the minimum Q-value, both critics are trained by minimizing their respective mean squared error (MSE) objectives. This is expressed as [54]

$$L(\phi_i, \mathcal{B}) = \mathop{\mathbb{E}}_{\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\} \sim \mathcal{B}} \left[\left(Q\left(\mathbf{s}^t, \mathbf{a}^t; \phi_i\right) - \zeta\left(r^t, \mathbf{s}^t\right)\right)^2\right], \quad i = 1, 2. \quad (28)$$

where the expectation operator indicates that this operation is performed over a batch of samples as the MSE objective implies. After training the critics using (28), the minimum Q-values for the state-action pairs generated by the critic DNNs are used to train the actor's DNN. In particular, the actor network adjusts its parameters to maximize the Q-values. Hence, the actor's maximization objective is expressed as [55]

$$\max_{\psi} \mathop{\mathbb{E}}_{\mathbf{s}^t \sim \mathcal{B}} \left[\mathcal{Q}_\phi(s, \mu(s))\right], \quad (29)$$

where ψ is the actor's DNN parameters, and ϕ is the critic's DNN that generates the minimum Q-value prediction. Note that, unlike DPPG, the TD3 agent does not update the policy in each time-step which further stabilises learning. The target networks are then partially updated as follows:

$$\begin{aligned} \phi_i' &= \kappa\phi_i + (1-\kappa)\phi_i', \ i = 1, 2, \\ \psi' &= \kappa\psi + (1-\kappa)\psi', \end{aligned} \quad (30)$$

where $0 < \kappa \leq 1$ is the smoothing factor for the target networks. Hence, κ is one of the most important hyperparameters that have a significant impact on the convergence of the TD3 agent. Another important aspect for DRL agents is exploration. Since the TD3 agent optimizes a deterministic policy, it has no means of exploring other actions. Furthermore, since the agent is initialized randomly, the initial policy is equivalent to that of a random process. Therefore, to address this issue, random noise samples are added to the actions taken by the agent which serve as an exploration strategy. A Gaussian random process $\mathcal{N}$ is often used as a source for the noise samples added to the agent's actions. Therefore, the clipped TD3 action is expressed as

$$\mathbf{a}^t = \text{clip}\left(\mu\left(\mathbf{s}^t\right) + \mathbf{n}, a_{high}, a_{low}\right), \quad (31)$$

where $\mathbf{n} \sim \mathcal{N}(0, \sigma'\mathbf{I})$ is the noise vector obtained from a normally distributed process with zero mean and standard deviation σ'.

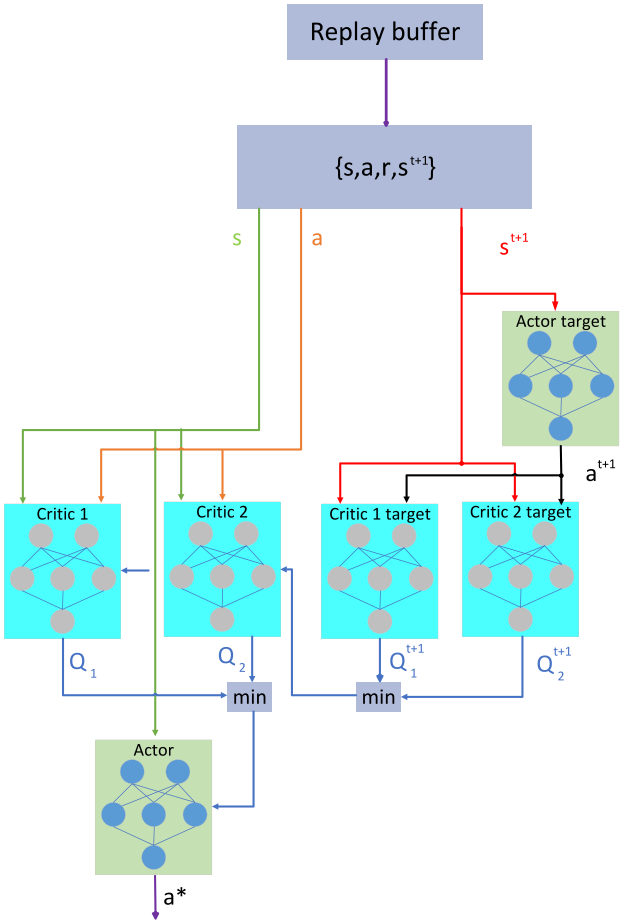

FIGURE 2. The actor-critic interactions in the proposed TD3 agent.

So far, we have discussed the problem reformulation into an RL environment and explained the inner workings of the TD3 agent. Hence, the developed TD3-based algorithm for robust resource allocation is explained in Algorithm 2.

To show how the proposed algorithm is implemented after training, Figure 3 illustrates the integration of the trained TD3 model into the BS of the considered IRS-assisted MISO-NOMA system.

Note that unlike conventional optimization algorithms, we do not explicitly consider the outage probability during the training and learning stage in the TD3-based robust design, however, it is included implicitly through the random errors as explained in Algorithm 2. The first motivation for the proposed approach is that since the TD3 agent is initialized with a random policy, basing the reward function

Algorithm 2 TD3-Based Robust Resource Allocation

Initialise: agent's hyperparameters $\mu, \phi_1, \phi_2, \mathcal{D}, \mathcal{N}, b$, and the IRS vector $\mathbf{v}_{init}$
Set $\phi_i' \leftarrow \phi_i, i = 1, 2$, and $\mu' \leftarrow \mu$
while *episode* $\leq$ *Episodes* **do**
  Obtain the estimated channels for all UEs, $\hat{\mathbf{h}}_k, k = 1, \ldots, 2K$
  Execute algorithm 1 to obtain the UE pairs.
  Calculate the ZFBF matrix $\mathbf{W}$ according to (15)
  Obtain the channel error samples $\Delta\mathbf{Q}_1, \ldots, \Delta\mathbf{Q}_{2K}$ according to (3)
  while *step* $\leq$ *Steps* **do**
    Get the actions vector $\mathbf{a}^t$ by evaluating the actor's DNN using the current state according to (31)
    Extract $\bar{\mathbf{v}}^t$,$\bar{\mathbf{P}}^t$ according to (23) and (24)
    Add the random channel error terms according to (3) to create the final true channels
    Evaluate the SINR equations for all UEs according to (5) and (8) using the true channels
    Calculate the achieved rates for all UEs according to (9)
    if $R_k^t \geq R_k^{min}, k = 1, \ldots, 2K$: **then**
      Use the reward function in (21)
    else
      Use the reward function in (22)
    end if
    Obtain the next $\mathbf{s}^{t+1}$; save the the tuple $\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\}$ to $\mathcal{D}$
    Sample a batch of $\mathcal{L}$ experiences randomly from $\mathcal{D}$
    Calculate the targets for the sampled experiences according to (27)
    Train the two critics using (28)
    if *update_policy* == *True*: **then**
      Train the actor network using (29)
    end if
    Update the target networks using (30)
    step = *step* + 1
    Set $\mathbf{s}^t = \mathbf{s}^{t+1}$
  end while
  episode = *episode* + 1
end while
Output: $[\mathbf{w}_1, \ldots, \mathbf{w}_C, \bar{\mathbf{v}}^*, \bar{\mathbf{P}}^*, \alpha_{1,s}^*, \ldots, \alpha_{C,w}^*]$

on the non-outage probability leads to extremely sparse reward in the initial training steps which eventually leads to divergence. The other motivation is that by basing the reward function on the true achieved rates, the agent always aims for a non-outage probability of 1, which leads to an inherently robust policy. Therefore, the implications of the outage constraints are included implicitly in Algorithm 2. Hence, the non-outage probability of the agent's policy is hyperparameterized in the proposed design. Consequently, the robustness of the agent's policy is a function of the hyperparameters of the TD3 agent.

Note that even though the agent is rewarded by the achieved true sum-rates, this does not imply that the agent has access to the true channels. In particular, since the reward is determined by the environment in the RL framework and the UEs are part of the environment, the true channels are still unknown to the agent.

D. COMPLEXITY ANALYSIS

In this section, we provide the computational complexity for the developed TD3-based algorithm. In particular, since DRL agents are only trained once, we assume that the offline training complexity can be afforded [19]. Hence, we focus on analysing the online or inference complexity during deployment.

The big $\mathcal{O}$ notation is one of the most widely adopted methods that provides an upper bound for the worst-case run-time for a given algorithm with respect to its parameters. Since the trained actor's network is the one that is used to carry out the inference, the deployment complexity of the proposed agent is based on the feed-forward pass through the actor's DNN. In addition, since DNN models are vector-friendly, the worst-case run-time is expressed as a combination of matrix-vector multiplication. Assuming that the actor's network has $\mathcal{I}$ hidden layers, with each consisting of Ω neurons, then it is straightforward to conclude that there are $\mathcal{I}+1$ matrix-vector multiplications in the feed-forward pass. In addition, the hidden and output layers require one activation each using an activation function. Therefore, the computational-complexity is written as $\mathcal{O}(T(\Omega \cdot \mathbf{Card}(\mathbf{s}^t) + \mathcal{I} \cdot \Omega^2 + \mathbf{Card}(\mathbf{a}^t) \cdot \Omega + \mathcal{I} \cdot \Omega + \mathbf{Card}(\mathbf{a}^t) + CN^2))$, where $\mathbf{Card}(\mathbf{s}^t) = 8K + 2M$ for the dynamic-channels case as highlighted by (20), $\mathbf{Card}(\mathbf{a}^t) = 2K + 2M$, the term $CN^2, C \geq N$, represents the complexity for calculating the pseudoinverse in (15), while the terms $\mathcal{I} \cdot \Omega$ and $\mathbf{Card}(\mathbf{a}^t)$ refer to the element-wise activation operations for the hidden and output layers, respectively. Note that since the actions vector is part of the state vector, and assuming that $\Omega \gg \mathbf{Card}(\mathbf{s}^t)$, and $\Omega \gg CN^2$, then, the worst-case run-time for the actor's DNN is reduced to $\approx \mathcal{O}(\Omega^2)$, which implies that the complexity of the algorithm becomes completely dependent on the number of neurons in the hidden layers. Such a case is particularly useful for problems with relatively small state spaces. The term T is specific to the proposed algorithm since we consider the previous action as part of the state vector. Therefore, the actor network is evaluated T times to guarantee competitive performance. Nevertheless, a small T value is often adopted to minimize the latency of the algorithm. Moreover, to keep the latency of the proposed algorithm to a minimum, $T = 2$ is used in the simulation results section unless stated otherwise.

In order to compare the analytical complexity of the proposed TD3-based algorithm to existing convex optimization algorithms, we briefly review three widely adopted conventional optimization approaches for solving the static version of the considered optimization problem. In [10], a SOCP-ADMM-based algorithm was developed

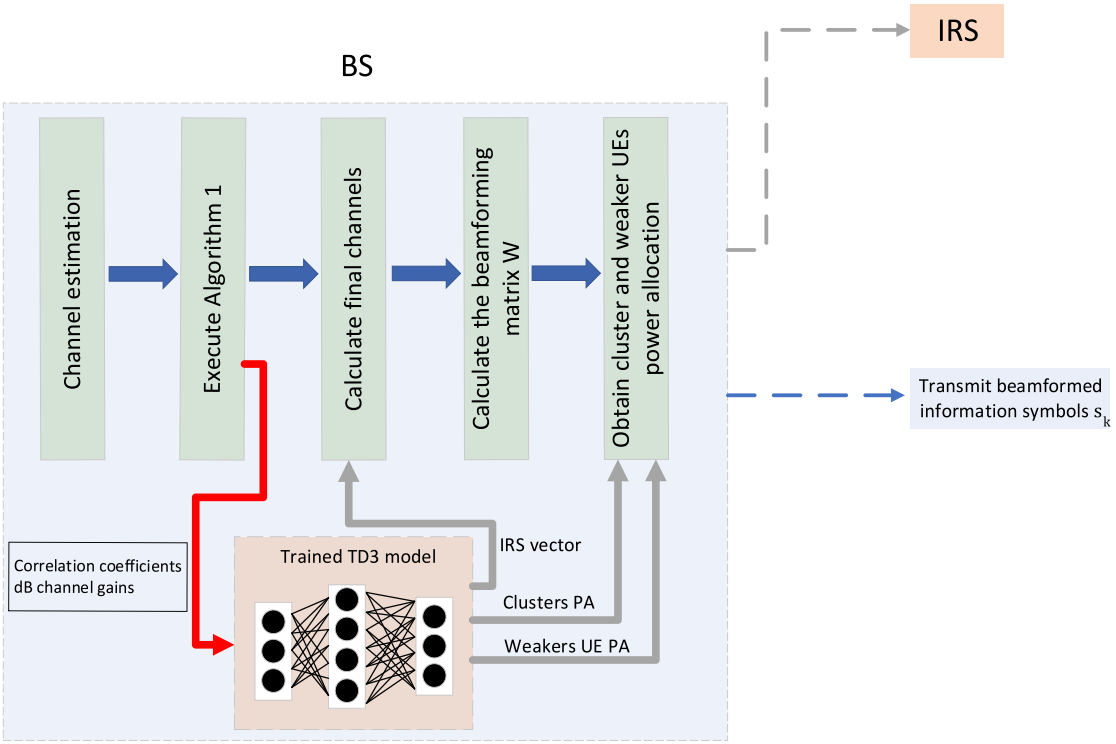

FIGURE 3. The implementation of the proposed algorithm within the BS of the considered IRS-assisted MISO-NOMA system.

to iteratively solve the transmit power minimization problem. The derived algorithm has a worst-case complexity of $\mathcal{O}(K^{1.5}M^3 + K^{4.5}N^3)$. In addition, the non-IRS and non-clustered MISO-NOMA beamforming design was considered for the system sum-rate maximization objective in [6]. The proposed iterative algorithm solves a SOCP optimization problem with a worst-case complexity of $\mathcal{O}((2K)^7)$ per iteration. For IRS-aided MISO systems, the work in [42] proposed a semidefinite programming (SDP) solution for the relaxed IRS optimization subproblem, while utilizing a closed-form solution based on the maximal ratio combining (MRT) for the beamforming design subproblem. The SDP's worst-case complexity is $\mathcal{O}(M^6)$, while the optimal power allocation subproblem is still non-trivial.

While both algorithms provide solid performance and interesting results, it is obvious that they do not scale well in practical scenarios, let alone latency-sensitive applications. Furthermore, the aforementioned algorithms are derived under the assumption that the global CSI is available system-wide, and therefore, cannot be directly extended to the robust design case. On the other hand, the proposed TD3-based algorithm can be utilized to generate competitive and robust joint solutions while keeping the complexity to a minimum. Note that in this paper, we assume that the SUPA is executed in the higher layers which are more latency-tolerant compared to the physical layer. Nevertheless, it is straightforward to conclude that the worst-case run-time for the SUPA is $\mathcal{O}(K^2)$.

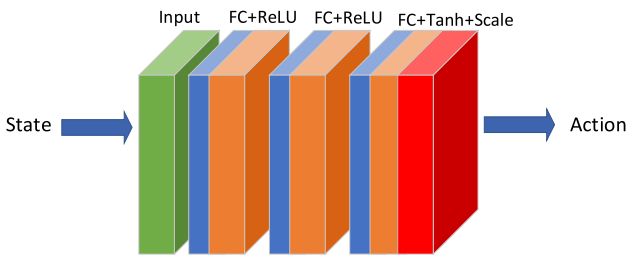

FIGURE 4. Actor's DNN architecture.

V. TRAINING, SIMULATION AND NUMERICAL RESULTS

In this section, we provide the details of the TD3 agent's structure, hyperparameters and training. In addition, the system parameters and the simulation results for both the fixed and the dynamic-channel cases are presented.

A. AGENT STRUCTURE AND HYPERPARAMETERS

The developed TD3 agent consists of one actor and two critic networks. Note that the two critic networks are identical in terms of the architecture, however, they are initialized randomly. The DNN structures for both the actor and the critic networks are illustrated in Figures 4 and 5, respectively. For the actor's DNN, the rectified linear unit ($ReLU$) activation function $f(x) = max(0, x)$, is used to activate the fully connected hidden layers. In addition, the $Tanh$ function $f(y) = \frac{e^y - e^{-y}}{e^y + e^{-y}}$, is utilized to activate the output layer. Furthermore, the scaling layer maps the values of the actions vector to the appropriate levels. Similarly, the $ReLU$ function is also used to activate the hidden layers

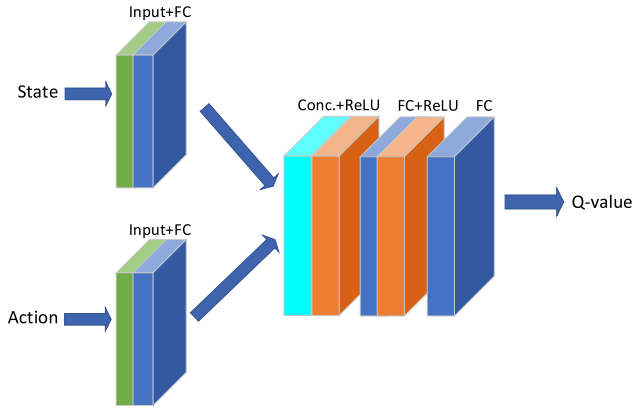

FIGURE 5. Critic's DNN architecture.

TABLE 1. Hyperparameters of the TD3 agent.

Hyperparameter	Value
Actor learning rate (fixed/dynamic channels)	0.0007/0.0001
Critics learning rate (fixed/dynamic channels)	0.0009/0.0003
Discount factor (δ)	0.99
Policy update frequency	2
Smoothness factor (fixed-channels), $C = 2, 3, C = 4$	$0.0005, 0.0001$
Smoothness factor (Dynamic-channels)	0.00001
Replay buffer size ($\mathcal{B}$)	$100,000$
Minibatch size ($\mathcal{L}$)	128
Number of episodes, time-steps (fixed-channels)	$500, 1000$
Number of episodes, time-steps (dynamic-channels)	$800, 1000$

of the critic's DNNs. However, since each critic network takes in both the state and the actions separately, it needs a concatenation layer to merge these two inputs. Note that, unlike the actor's DNN, the critic's network outputs a scalar Q-value which indicates the quality of the state-action pair. Furthermore, a relatively high δ value is selected to drive the agent towards developing a long-term robust policy. In terms of DNNs optimization, the Adam optimizer is utilized for both the actor and the critic networks [56]. Note the number of neurons in each hidden layer is identical for both DNNs. Table 1 lists the TD3 agent's hyperparameters and the training parameters used in this paper.

Since the number of neurons is the dominant factor that determines the learning capability of a DNN with a fixed number of layers, and consequently, the developed TD3 agent [57], we use two different neuron values for each channel case. In particular, for the fixed-channels case, we generate one set of simulation results for a TD3 agent configured with 128 neurons in each hidden layer, and another set for the same agent configured with 256 neurons in each hidden layer. Similarly, the same process is replicated for the dynamic-channels case with 256 and 512 neurons for each set of simulation results.

B. SYSTEM PARAMETERS

We consider a downlink transmission for a clustered and IRS-assisted MISO-NOMA system that is identical to the one illustrated in Figure 1. In addition, the channel between the BS and the IRS is assumed to have both a line-of-sight (LoS) and non-LoS components, and therefore, modelled using the Rician fading coefficients. In particular, the BS-IRS link is expressed as

$$\mathbf{G} = \frac{1}{\sqrt{d_{irs}^{\iota_{b\to irs}}}}\left(\sqrt{\frac{L}{1+L}}\mathbf{G}_{LoS} + \sqrt{\frac{1}{1+L}}\mathbf{G}_{nLoS}\right), \quad (32)$$

where $d_{irs} = 50$ m is the distance between the BS and the IRS and is assumed to be fixed throughout the simulation. $\iota_{b\to irs}$ refers to the path-loss exponent representing the large-scale fading between the BS and the IRS, and $L = 1$ is the Rician factor. On the other hand, the channel between the IRS and the UEs is assumed to experience Rayleigh fading and is expressed as

$$\mathbf{g}_k = \frac{\tilde{g}}{\sqrt{d_k^{\iota_{irs\to u}}}}, k = 1, \ldots, 2K, \quad (33)$$

where d_k is the distance between the IRS and UE$_k$, $\iota_{b\to irs}$ is the path-loss exponent between the IRS and UE$_k$, and $\tilde{g} \sim \mathcal{CN}(0, 1)$. Furthermore, we assume that the UEs are located between $[50-100]$ m away from the BS. Table 2 lists all the system parameters used to generate the simulation results.

TABLE 2. Summary of system parameters.

System parameter	Value
Cell radius	100 m
Number of UEs ($2K$)	$4, 6, 8$
Number of clusters (C)	$2, 3, 4$
Number of antennas at the BS (N)	$2, 3, 4$
Number of IRS elements (M)	16
Transmit power	36 dBm
Noise power	-90 dBm
Relative value for the error boundary λ	0.01
Path-loss exponent (BS-IRS) $\iota_{b\to irs}$	2
Path-loss exponent (IRS-UEs) $\iota_{irs\to u}$	3
Target rate R_k^{min} (fixed channels)	1 Bit/s/Hz
Target rate R_k^{min} (dynamic channels)	0.3 Bit/s/Hz

To compare the performance of the proposed algorithm to existing algorithms in the literature, we use the following benchmark schemes:

- *Baseline 1:* a DDPG agent which has been one of the most widely adopted DRL agents in the literature. This benchmark scheme is included to provide a baseline for convergence and policy robustness testing.
- *Baseline 2:* a convex optimization-based scheme which represents the conventional optimization approach where the IRS optimization subproblem is solved using SDP [42], then, the non-robust ZFBF with fixed power allocation is used for the beamforming design.
- *Baseline 3:* a random algorithm which has an almost negligible complexity is used to benchmark the quality of the policy derived by the proposed agent. In this

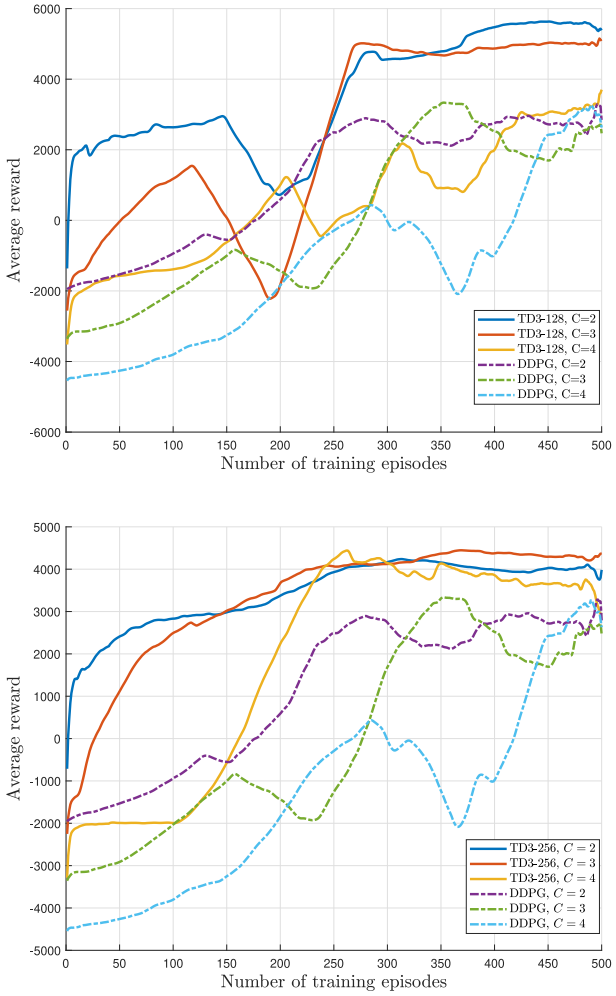

FIGURE 6. Convergence of the proposed TD3 agent for the fixed-channels case.

benchmark, all of the design variables are randomly selected.

C. FIXED-CHANNELS CASE

To evaluate the performance of the proposed algorithm against channel errors, we first consider the case where the channels are fixed throughout the training process. However, a new set of errors is introduced in each training episode. Furthermore, the UEs are assumed to be uniformly distributed in the fixed-channels case.

The convergence plot is a useful measure that indicates the quality of the derived policy by the agent. Figure 6 illustrates the convergence of the TD3 and DDPG agents. With two clusters (i.e., $C = 2$), both agents are able to develop a highly rewarding policy after a few training episodes. However, when the number of users in the system increases, both agents require more training episodes to start forming a high-reward policy.

In the two extreme cases, however, the average reward sustained by the TD3 agent is significantly higher than that for the baseline DDPG agent. Moreover, the TD3-256 shows more stable and consistent convergence compared to both

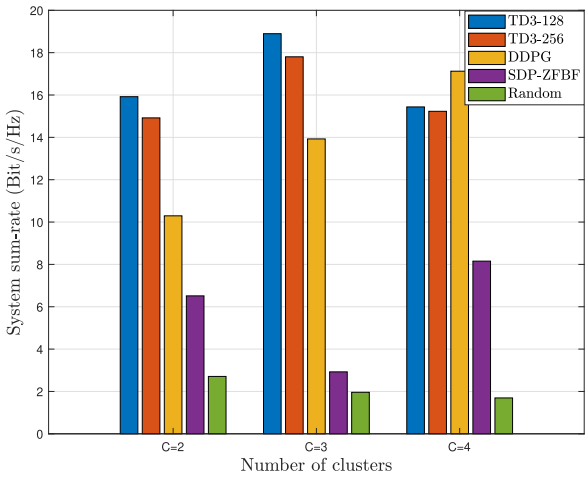

FIGURE 7. The average system sum-rates for the fixed-channels case with various number of UEs.

TD3-128 and DDPG. In order the show the implications of converging to a higher reward policy, the achieved system sum-rates for the trained TD3 agent are shown in Figure 7. The rates provided represent the average system sum-rate over 1000 testing episodes.

The TD3 agent outperforms the benchmark schemes for both $C = 2$ and $C = 3$ scenarios. In particular, the TD3-128 agent achieves the highest average sum-rate of approximately 18.5 Bit/s/Hz, when $C = 3$, with 4.5 Bit/s/Hz gap compared to the DDPG agent. Additionally, the TD3 agents trade-off higher system sum-rate performance when $C = 4$ for improved robustness as explained next.

Note that Figure 7 only shows partial information about the agent's performance. To gain a better insight, Figure 7 is interpreted in the context of the outage performance of the agent illustrated by Figures 9 and 10. However, since the outage performance of the agent is related to the weakest UE's achieved rate, Figure 8 depicts the achieved rates probability density function (PDF) for the weakest UEs in the system.

Based on the weakest UE rate for each setting, we can infer that the TD3 agent has formed an outage-aware policy which results in the least outage across the three different system settings. Note that since the PDFs in Figure 8 are for the weakest UEs in each category, this represents the worst-case performance of the agent.

To assess the outage performance of the proposed agent against the relative channel estimation quality λ, Figure 9 shows the robustness of the agent's policy against different values for λ. The Figure shows that for all system parameters, the TD3 agent has a worst-case non-outage probability of 88% for the TD3 agents when $C = 4$ at $\lambda = 0.01$, compared to DDPG's worst-case of 77% at the same λ value. On the other hand, the best-case performance is sustained when $C = 2$, where the TD3 agent achieves a non-outage probability of 100%, outperforming the DDPG's best-case performance by a margin of 5%. In all cases, the TD3 agents' policies

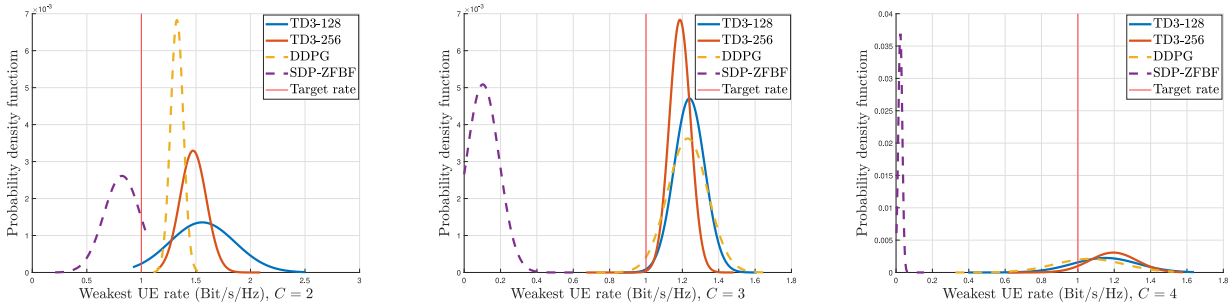

FIGURE 8. The PDFs for the weakest UE's achieved rate in the system.

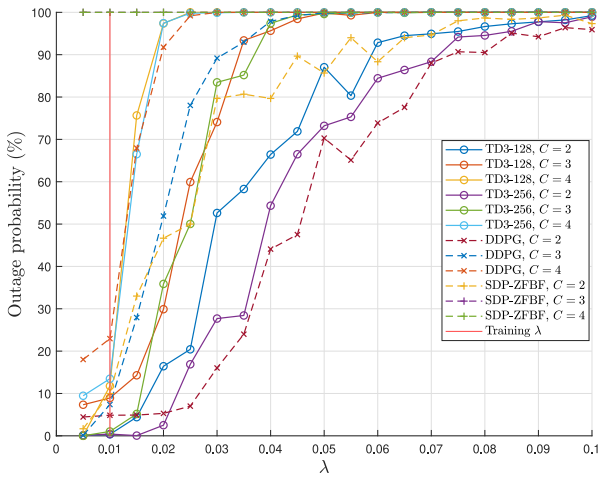

FIGURE 9. The average outage probability versus the estimation quality factor λ.

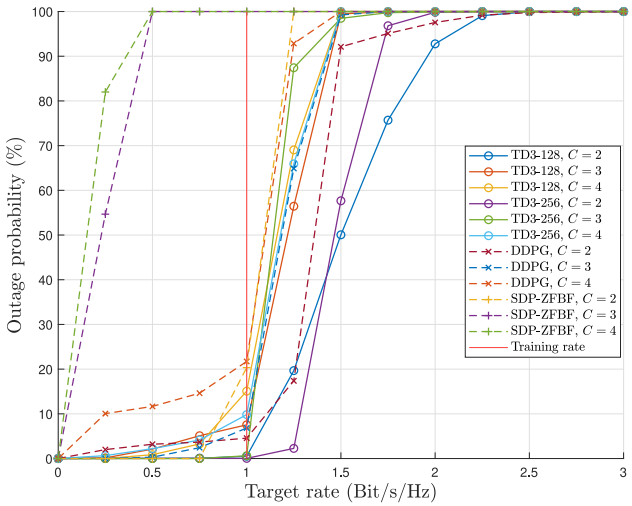

FIGURE 10. The average outage probability versus the target rate R_k^{min}.

perform well in terms of generalization over larger λ values than the one used for training. In particular, the higher number of neurons in the TD3-256 agent pays off in terms of the non-outage probability at $\lambda = 0.01$ where it achieves 98% and 88% scores for the three and four clusters, respectively. This suggests that the agent's derived policy is robust against variations in the estimation error factor. Another practical benchmark for measuring the agent's policy robustness is the outage performance against target rates. Figure 10 illustrates the non-outage probability versus different target rates. The agent's performance generally follows the same pattern as in Figure 9, where the best-case outage performance is achieved when $C = 2$ with 100% non-outage probability at the training target rate of 1 Bit/s/Hz which is around 7% better than that for the DDPG agent. As for the more challenging case when $C = 4$, the TD3 agent still outperforms the DDPG agent with a 6% performance gap. In addition, the TD3 agent's policy is able to sustain a 25% increase in the target rate while still achieving a non-outage probability of 97% on average, which proves that the agent has developed a solid robust policy.

Another important observation is the impact of the number of neurons on the outage probability of the TD3 agent. The simulation results suggest that the TD3-256 outperforms the TD3-128 in the more challenging cases with a higher number of UEs. This further proves our claim that since the outage constraint is hyperparameterized in the proposed robust design, it is impacted by the selected learning parameters of the TD3 agent.

D. DYNAMIC-CHANNELS CASE

The fixed-channels case is useful for rigorous analysis of the agent's developed policy as the channels are considered static. In practice, however, the channel is frequently changing especially when the UEs are moving. Therefore, we extend the developed algorithm to the dynamic-channels case in this subsection. Unlike the fixed-channels case, the users are assumed to be randomly distributed within the cell radius to make the design more practical. In this case, new channels are introduced in each new training episode. Furthermore, the channels are assumed to be quasi-static, i.e., the channels remain constant during each training episode and change afterwards. Moreover, 24 different channel sets are used for training. The aim of the dynamic-channels case is to train the agent to develop a comprehensive robust policy that can be generalized to never-seen-before channels. Hence, after training the agent once, it could be deployed to any channel condition afterwards.

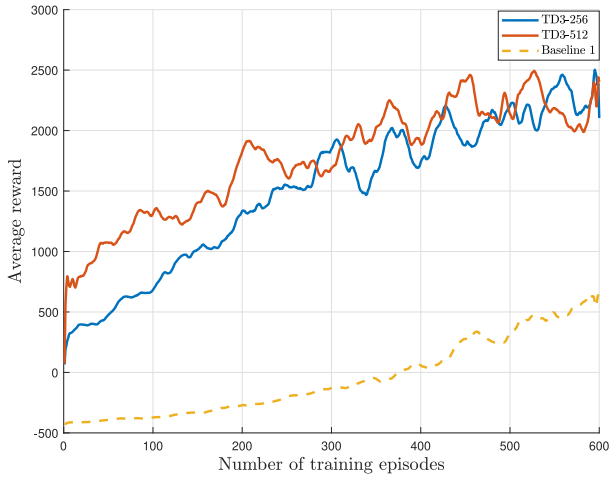

FIGURE 11. Convergence of the TD3 agent for the dynamic-channels case, $C = 2$.

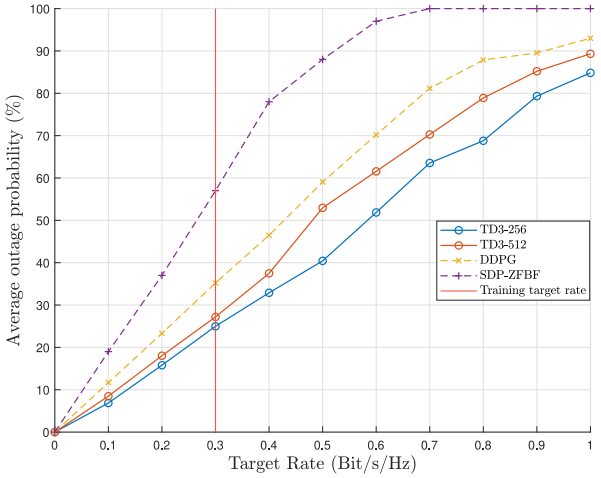

FIGURE 13. The average outage probability of the TD3 agent versus the target rate for the dynamic-channels case, $C = 2$.

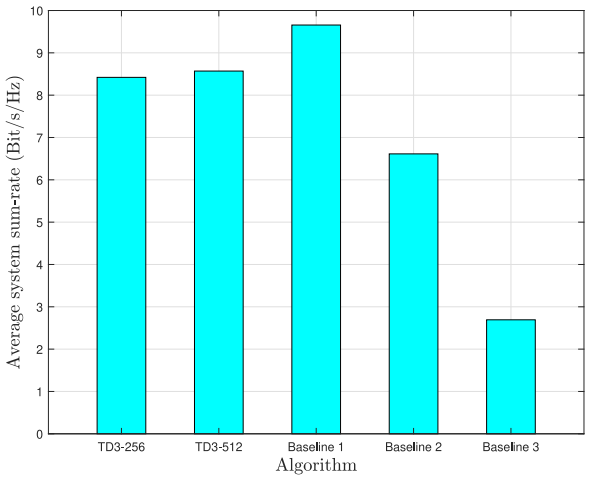

FIGURE 12. The average system sum-rates for the dynamic-channels case, $C = 2$.

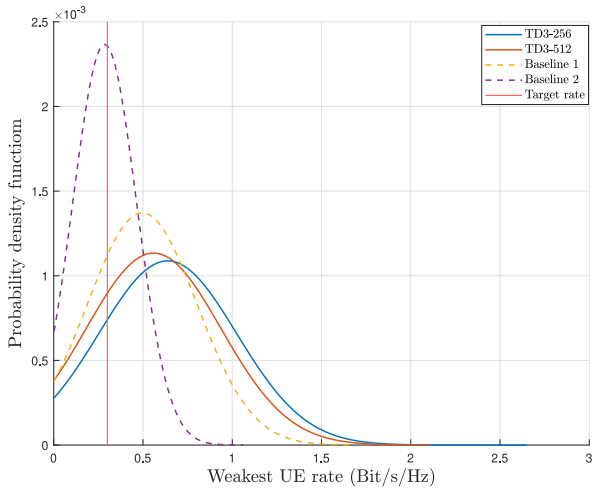

FIGURE 14. The PDFs for the weakest UE's achieved rate in the system, $C = 2$.

Figure 11 illustrates the superior performance of the TD3 agent over the DDPG baseline in developing a highly rewarding policy.

In order to generate statistically meaningful results, a set of 100 channels and 10 error samples per channel are used for testing to generate the average performance results.

The average system sum-rates achieved by the proposed agent are shown in Figure 12.

The average sum-rates figure shows that baseline 1 achieves the highest rate, which is explained by the worse outage performance illustrated in Figure 13. The two figures suggest that there is a trade-off between achieving a higher system sum-rate and a higher non-outage probability. The TD3 agents, for example, achieve an average sum-rate of around 8.5 Bit/s/Hz with an average outage probability of 24% at the 0.3 Bit/s/Hz target rate. On the other hand, the DDPG agent has an average outage probability of around 35% at the same target rate. In addition, the average outage performance gap between the TD3 agent and the SDP-ZFBF baseline widens significantly as the target rate increases. This clearly shows that the TD3 agent has developed a robust policy that is capable of withstanding the channel uncertainty for different channel conditions.

Furthermore, the PDFs of the average rate achieved by the weakest UE in the system are illustrated in Figure 14.

The PDFs figure shows that the TD3 agents achieve the highest mean of around 0.6 Bit/s/Hz, outperforming the other benchmark schemes.

Overall, the TD3 agent outperforms all benchmark algorithms in terms of outage performance. In particular, the TD3 agent shows more adaptive and robust behaviour by trading off higher sum-rates for better outage performance when it is challenging to maximize both. This shows that the proposed TD3-based algorithm is capable of converging to adaptive policies that suit the problem requirements.

VI. CONCLUSION

The resource allocation problem for an IRS-assisted MISO-NOMA system was considered in this paper. In particular,

by taking the imperfect channel estimation at the BS and the UEs into account, the outage-constrained robust design with an ergodic sum-rate maximization objective was formulated. A correlation-based UE clustering algorithm was proposed to pair the UEs into clusters. Then, the challenging robust design problem was reformulated into an RL environment since it cannot be solved directly using conventional optimization techniques. Subsequently, a DRL-based framework was developed to solve the reformulated problem using the TD3 agent. The simulation results demonstrated that the TD3 agent outperforms conventional and other DRL algorithms in terms of generating robust resource allocation strategies for the considered system model under different system parameters. In addition, the performance of the developed TD3-based algorithm in the dynamic-channels case showed that the proposed framework can be implemented in practical scenarios. Furthermore, the competitive performance achieved by the proposed TD3-based algorithm has a much lower computational complexity compared to conventional optimization algorithms, making it a more sensible option for latency-stringent applications.

REFERENCES

[1] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. 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 *Proc. IEEE 77th Veh. Technol. Conf.*, 2013, pp. 1–5.

[3] Q. Sun, S. Han, I. Chin-Lin, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Aug. 2015.

[4] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

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

[6] 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 Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.

[7] J. Zhu, J. Wang, Y. Huang, K. Navaie, Z. Ding, and L. Yang, "On optimal beamforming design for downlink MISO NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3008–3020, Mar. 2020.

[8] H. M. Al-Obiedollah, K. Cumanan, J. Thiyagalingam, A. G. Burr, Z. Ding, and O. A. Dobre, "Energy efficient beamforming design for MISO non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4117–4131, Jun. 2019.

[9] E. Basar, "Transmission through large intelligent surfaces: A new frontier in wireless communications," in *Proc. IEEE Eur. Conf. Netw. Commun. (EuCNC)*, 2019, pp. 112–117.

[10] 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 Trans. Commun.*, vol. 69, no. 1, pp. 664–674, Jan. 2021.

[11] F. Fang, Y. Xu, Q.-V. Pham, and Z. Ding, "Energy-efficient design of IRS-NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 14088–14092, Nov. 2020.

[12] K. Cumanan, R. Krishna, V. Sharma, and S. Lambotharan, "Robust interference control techniques for multiuser cognitive radios using worst-case performance optimization," in *Proc. 42nd Asilomar Conf. Signals, Syst. Comput.*, 2008, pp. 378–382.

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

[14] F. Alavi, K. Cumanan, Z. Dingl, and A. G. Burr, "Outage constraint based robust beamforming design for non-orthogonal multiple access in 5G cellular networks," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, 2017, pp. 1–5.

[15] M. Zhang, K. Cumanan, L. Ni, H. Hu, A. G. Burr, and Z. Ding, "Robust beamforming for AN aided MISO SWIPT system with unknown eavesdroppers and non-linear EH model," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–7.

[16] J. Yu, X. Liu, Y. Gao, C. Zhang, and W. Zhang, "Deep learning for channel tracking in IRS-assisted UAV communication systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7711–7722, Sep. 2022.

[17] S. Liu, M. Lei, and M.-J. Zhao, "Deep learning based channel estimation for intelligent reflecting surface aided MISO-OFDM systems," in *Proc. IEEE 92nd Veh. Technol. Conf.*, 2020, pp. 1–5.

[18] Y. Ahn and B. Shim, "Deep learning-based beamforming for intelligent reflecting surface-assisted mmWave systems," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, 2021, pp. 1731–1734.

[19] 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 Trans. Commun.*, vol. 68, no. 3, pp. 1866–1880, Mar. 2020.

[20] N. C. Luong et al., "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

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

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

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

[24] J. Chen, L. Guo, J. Jia, J. Shang, and X. Wang, "Resource allocation for IRS assisted SGF NOMA transmission: A MADRL approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1302–1316, Apr. 2022.

[25] A. Benfaid, N. Adem, and B. Khalfi, "AdaptSky: A DRL based resource allocation framework in NOMA-UAV networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2021, pp. 01–07.

[26] A. Waraiet, K. Cumanan, Z. Ding, and O. A. Dobre, "Robust design for IRS-assisted MISO-NOMA systems: A DRL-based approach," *IEEE Wireless Commun. Lett.*, vol. 13, no. 3, pp. 592–596, Mar. 2024.

[27] M. B. Shahab, M. Irfan, M. F. Kader, and S. Y. Shin, "User pairing schemes for capacity maximization in non-orthogonal multiple access systems," *Wireless Commun. Mob. Comput.*, vol. 16, no. 17, pp. 2884–2894, 2016.

[28] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 328–331, Apr. 2019.

[29] C. Pan et al., "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1719–1734, Aug. 2020.

[30] N. K. Kundu and M. R. McKay, "A deep learning-based channel estimation approach for MISO communications with large intelligent surfaces," in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mob. Radio Commun.*, 2020, pp. 1–6.

[31] 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 Trans. Signal Process.*, vol. 68, pp. 5092–5106, Aug. 2020.

[32] 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 Trans. Commun.*, vol. 67, no. 11, pp. 7937–7949, Nov. 2019.

[33] A. Agrawal, J. G. Andrews, J. M. Cioffi, and T. Meng, "Iterative power control for imperfect successive interference cancellation," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 878–884, May 2005.

[34] 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 Trans. Signal Process.*, vol. 62, no. 21, pp. 5690–5705, Nov. 2014.
[35] M. B. Shenouda and T. N. Davidson, "Probabilistically-constrained approaches to the design of the multiple antenna downlink," in *Proc. 42nd Asilomar Conf. Signals, Syst. Comput.*, 2008, pp. 1120–1124.
[36] M. Payaro, A. Pascual-Iserte, and M. A. Lagunas, "Robust power allocation designs for multiuser and multiantenna downlink communication systems through convex optimization," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1390–1401, Sep. 2007.
[37] Y. Liu, Z. Tan, H. Hu, L. J. Cimini, and G. Y. Li, "Channel estimation for OFDM," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1891–1908, 4th Quart., 2014.
[38] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, Oct. 2020.
[39] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4847–4861, Nov. 2016.
[40] A. Bulut and T. K. Ralphs, "On the complexity of inverse mixed integer linear optimization," *SIAM J. Optim.*, vol. 31, no. 4, pp. 3014–3043, 2021.
[41] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for intelligent reflecting surface enhanced wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6.
[42] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.
[43] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.
[44] A. Kumar et al., "User pairing and power allocation for IRS-assisted NOMA systems with imperfect phase compensation," *IEEE Wireless Commun. Lett.*, vol. 11, no. 12, pp. 2492–2496, Dec. 2022.
[45] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.
[46] B. Kimy et al., "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Mil. Commun. Conf.*, 2013, pp. 1278–1283.
[47] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Select. areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.
[48] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, Sep. 2008.
[49] M. Sharif and B. Hassibi, "A comparison of time-sharing, DPC, and beamforming for MIMO broadcast channels with many users," *IEEE Trans. Commun.*, vol. 55, no. 1, pp. 11–15, Jan. 2007.
[50] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2014.
[51] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT press, 2018.
[52] C. J. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, pp. 279–292, May 1992.
[53] G. A. Rummery and M. Niranjan, "On-line Q-learning using connectionist systems," Dept. Eng., Univ. Cambridge, Cambridge, U.K., Rep. TR-166, 1994.
[54] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
[55] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[57] M. Uzair and N. Jamil, "Effects of hidden layers on the efficiency of neural networks," in *Proc. IEEE 23rd Int. Multitopic Conf. (INMIC)*, 2020, pp. 1–6.

ABDULHAMED WARAIET (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from Near East University, Northern Cyprus, in 2017, and the M.Sc. degree in signal processing and communications from the University of Edinburgh, U.K., in 2019. He is currently pursuing the Ph.D. degree under the supervision of K. Cumanan. His current research interests include machine learning-based resource allocation algorithms for IRS-assisted non-orthogonal multiple access systems.

KANAPATHIPPILLAI CUMANAN (Senior Member, IEEE) received the B.Sc. degree (with First-Class Hons.) in electrical and electronic engineering from the University of Peradeniya, Sri Lanka, in 2006, and the Ph.D. degree in signal processing for wireless communications from Loughborough University, Loughborough, U.K., in 2009.

He is currently a Senior Lecturer with the School of Physics, Engineering and Technology, University of York, U.K. From March 2012 to November 2014, he was working as a Research Associate with the School of Electrical and Electronic Engineering, Newcastle University, U.K. Prior to this, he was with the School of Electronic, Electrical and System Engineering, Loughborough University. In 2011, he was an Academic Visitor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. From January 2006 to August 2006, he was a Teaching Assistant with the Department of Electrical and Electronic Engineering, University of Peradeniya. He was a research student with Cardiff University, Wales, U.K., from September 2006 to July 2007. He has published more than 100 journal articles and conference papers which attracted more than 4000 Google scholar citations. His research interests include nonorthogonal multiple access, cell-free massive MIMO, physical layer security, cognitive radio networks, convex optimization techniques, and resource allocation techniques.

Dr. Cumanan was the recipient of an Overseas Research Student Award Scheme from Cardiff University. He is currently serving as an Associate Editor for IEEE WIRELESS COMMUNICATIONS LETTERS and IEEE OPEN JOURNAL OF COMMUNICATIONS SOCIETY.