Artificial Intelligence-Driven Resource Allocation Techniques for Non-orthogonal Multiple Access Systems

Abdulhamed Khaled Waraiet

Doctor of Philosophy

University of York

School of Physics, Engineering and Technology

December, 2023

Abstract

With the unprecedented demand for wireless connectivity, and given the scarce radio resources, the quest for efficient and reliable multiple access (MA) techniques has never been more crucial. Unlike conventional MA techniques, non-orthogonal MA (NOMA) offers superior spectral and energy efficiencies. Particularly, NOMA allows spectrum sharing under controlled circumstances which enables massive connectivity. In addition, by combining NOMA with other techniques such as intelligent reflecting surfaces (IRS), the performance of NOMA is further enhanced. However, combining such sophisticated techniques often leads to highly complex optimization problems given the number of design parameters. Therefore, the conventional optimization-based approach often leads to high computational complexity algorithms that suffer from latency and scalability issues. Therefore, based on the recent advances in machine learning (ML) techniques, this thesis attempts to provide an ML-based alternative for addressing the latency and complexity challenges in the conventional approach. In particular, the reinforcement learning (RL) framework is utilized to solve resource allocation problems in NOMA systems.

Firstly, a robust joint design for an IRS-assisted downlink (DL) NOMA system with imperfect channel state information is considered. To overcome the joint non-convexity of the problem, it is then reformulated as an RL environment, and a twin-delayed deep deterministic policy gradient (TD3) agent is developed to solve the problem. Secondly, to reduce the receiver's complexity, users are clustered in an IRS-assisted DL NOMA system. Next, the beamforming design is proposed through the zero-forcing principle, and a joint robust design of power allocation and IRS phase shifts is proposed based on the TD3 agent. Thirdly, a robust design for energy efficiency (EE) maximization in an IRS-assisted uplink NOMA system is proposed. Moreover, an algorithm is developed based on the soft actor-critic (SAC) agent to jointly optimize power allocation and IRS phase shifts in the long-term EE maximization of the system.

"Read in the name of your Lord who created, created man from a clinging substance, read and your Lord is the most generous, who taught by the pen, taught man that which he knew not"

I would like to dedicate this thesis to my loving parents and family...

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. Some of the work presented in this thesis has been published or planned to be submitted to journals, which are listed as follows:

1. **A. Waraiet**, K. Cumanan, Z. Ding and O. A. Dobre, "Robust Design for IRS-Assisted MISO-NOMA Systems: A DRL-Based Approach," in *IEEE Wireless Communications Letters*, doi: 10.1109/LWC.2023.3335622.

2. **A. Waraiet**, K. Cumanan, Z. Ding and O. A. Dobre, "Deep Reinforcement Learning-based Robust Design for an IRS-assisted MISO-NOMA System," in *IEEE Transactions on Machine Learning in Communications and Networking*, (Accepted).

3. **A. Waraiet**, K. Cumanan, "Outage Constrained Robust Resource Allocation Framework for IRS-empowered NOMA Systems: A DRL-based Joint Design," in *IEEE Open Journal of the Communications Society*, (Under review).

4. **A. Waraiet**, K. Cumanan, "Robust EE Maximization for IRS-Empowered Uplink MIMO-NOMA Systems: A DRL-Based Approach," in *IEEE Transactions on Wireless Communications*, (Under review).

Abdulhamed Khaled Waraiet 31/December/2023

Acknowledgments

First and foremost, I would like to express my sincere and deepest gratitude to my supervisor, Dr. Kanapathippillai Cumanan, for his persistent and invaluable guidance, exceptional encouragement, motivation and constant support. Without his guidance and constant feedback, this PhD would not have been achievable. It is my distinct privilege and honour to have such a selfless supervisor during my PhD journey.

Also, I would like to extend my gratitude to Dr. Hamed Ahmadi for his invaluable suggestions and comments throughout the research assessments which helped me in shaping my research work.

I must also thank Prof. Zhiguo Ding and Prof. Octavia A. Dobre, for their kind support and helpful comments on my research work.

Finally, I am grateful to my entire family, especially my parents for their endless love, prayers, and unbounded support.

Contents

List of figures

List of tables

List of symbols

$(.)^T$	Transpose
$(.)^H$	Hermitian transpose
$\mathbf{x}$	Vector $\mathbf{x}$
$\mathbf{X}$	Matrix $\mathbf{X}$
$\|x\|$	Norm of complex number x
$\|\|\mathbf{x}\|\|_2$	Euclidean norm of vector $\mathbf{x}$
$\|\|\mathbf{X}\|\|_F$	Euclidean norm of matrix $\mathbf{X}$
$\mathbf{Card}(\mathbf{x})$	cardinality of vector $\mathbf{x}$
$\mathbb{R}$	The set of real numbers
$\mathbb{C}$	The set of complex numbers
$\mathbb{E}$	The mathematical expectation
$\mathbf{X} \succeq 0$	Positive semidefinite matrix
$\mathrm{tr}(X)$	Trace of the matrix X
$\mathbf{dom}\ f$	Domain of function f
∇f	First derivative of function f
$\nabla^2 f$	Second derivative of function f
min	Minimum of a function
max	Maximum of a function
$\mathcal{CN} \sim (\mu', \sigma^2)$	Complex Gaussian random variable with mean μ' and variance σ^2
$\mathcal{O}(.)$	Worst-case complexity for an algorithm

List of abbreviations

1G	First Generation
2G	Second Generation
3G	Third Generation
4G	Fourth Generation
5G	Fifth Generation
6G	Sixth Generation
MA	Multiple Access
BS	Base Station
UE	User Equipment
LoS	Line-of-Sight
OMA	Orthogonal Multiple Access
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FDM	Frequency Division Multiplexing
GSM	Global System for Mobile Communication
TDMA	Time Division Multiple Access
DS-CDMA	Direct-Sequence Code Division Multiple Access
MC-CDMA	Multi-Carrier Code Division Multiple Access
OFDMA	Orthogonal Frequency Division Multiple Access
MIMO	Multiple-Input Multiple-Output
URLLC	Ultra-Reliable Low Latency Communications

VR	Virtual Reality
IoT	Internet-Of-Things
mmWave	Millimetre Wave
EH	Energy Harvesting
M2M	Machine-To-Machine
D2D	Device-To-Device
EE	Energy Efficiency
KPI	Key Performance Indicator
Mb/s	Megabits Per Second
Gb/s	Gigabits Per Second
Tb/s	Terabits Per Second
b/J	Bit Per Joule
DFT	Discrete Fourier Transform
IRS	Intelligent Reflecting Surfaces
CoMP	Cooperative Multi-Point
UAV	Unmanned Aerial Vehicles
THz	Terahertz
VLC	Visible Light Communications
QAM	Quadrature Amplitude Modulation
PAPR	Peak-to-Average-Power Ratio
NOMA	Non-Orthogonal Multiple Access
RSMA	Rate-Splitting Multiple Access
SC	Superposition Coding
CDF	Cumulative Distribution Function
PDF	Probability Density Function
MRT	Maximum-Ratio Transmission
ADMM	Alternating Direction Method of Multipliers
SOCP	Second Order Cone Programming

SIC	Successive Interference Cancellation
SINR	Signal-to-Interference-plus-Noise Ratio
AI	Artificial Intelligence
ML	Machine Learning
ReLU	Rectified Linear
Tanh	Hyperbolic Tangent
DL	Downlink
UL	Uplink
MISO	Multiple-Input Single-Output
RB	Resource Block
CSI	Channel State Information
CSIT	Channel State Information at the Transmitter
CSIR	Channel State Information at the Receiver
RL	Reinforcement Learning
DRL	Deep Reinforcement Learning
MDP	Markov Decision Process
SDP	Semidefinite Programming
SDR	Semidefinite Relaxation
DDPG	Deep Deterministic Policy Gradient
TD3	Twin-Delayed Deep Deterministic Policy Gradient
DQN	Deep Q-Network
ZF	Zero-Forcing
ZFBF	Zero-Forcing Beamforming
SAC	Soft Actor-Critic
PPO	Proximal Policy Optimization
MMSE	Minimum Mean Squared Error
BC	Broadcast Channel
AWGN	Additive White Gaussian Noise

QoS	Quality of Service
SR	Sum-Rate
WSR	Weighted Sum-Rate
MMF	Max-Min Fairness
CNR	Channel-to-Noise Ratio
SCA	Successive Convex Approximation
DPC	Dirty Paper Coding
SISO	Single-Input Single-Output
DNN	Deep Neural Network
NN	Neural Network
LSTM	Long Short-Term Memory
MSE	Mean Squared Error
NPCU	Natural log Per Channel Use
MC	Monte Carlo
TD	Temporal Difference
SARSA	State-Action-Reward-State-Action
NFQ	Neural-Fitted Q-network
IID	Independent Identically Distributed
VPA	Vanilla Policy Gradient
SUPA	Successive User-Pairing Algorithm
DC	Difference of Convex
BEM	Bounded Error Model
UEM	Unbounded Error Model

Chapter 1

Introduction

1.1 Overview

The development of more efficient and scalable multiple access (MA) has never been more crucial. With the ever-increasing demand for wireless connectivity, 6G is expected to meet unprecedented spectral and energy efficiency requirements. In addition, given the scarcity of radio resources especially in the lower frequency bands, highly efficient protocols are necessary for realizing the potential of the next generation of wireless networks. Furthermore, the MA technique which is responsible for allocating the network resources plays an important role in enabling future applications for 6G and beyond.

Historically, orthogonal multiple access (OMA) techniques have been used as the standard means of serving multiple users in cellular networks. In the 1980s, the first generation (1G) of wireless technology adopted the orthogonal frequency division multiple access (FDMA) which is based on the frequency division multiplexing (FDM) technique [3]. In FDMA, users are served at the same time, each with a slice of the total available bandwidth of the system. Motivated by the shortcomings of analog modulation in 1G and with the advancement of technology, 2G implemented digital modulation techniques to overcome some of the drawbacks of the previous generation. Hence, the global system for mobile communication (GSM) which employed time division multiple access (TDMA) as its MA technique became the norm in cel-

lular networks during the 1990s [4]. Unlike FDMA, TDMA allocates different time slots for each user in which they can use the network resources. Shortly after, more wireless technologies capitalized on the features of digital communications which resulted in the development of direct sequence code division multiple access (DS-CDMA) and multi-carrier code division multiple access (MC-CDMA) [5]. CDMA was the main MA wireless technology for 3G networks where users allocated unique orthogonal codes while simultaneously sharing time and frequency resources. However, with the increased demand for higher data rates and numbers of devices, it was obvious that CDMA could not meet such requirements due to its scalability issues and tight power control constraints. Orthogonal frequency division multiple access (OFDMA) emerged as the next exciting MA wireless technology for 4G networks offering more flexibility through its multi-carrier solutions. In addition to its scalability, OFDMA's robustness to multipath fading, relaxed power control requirement, and full compatibility with multiple-input-multiple-output (MIMO) setups made it the preferred MA technology [6]. Powered by the ever-escalating demand for connectivity, 5G and beyond networks must be able to support ultra-reliable low latency communications (URLLC) applications such as online healthcare and virtual reality (VR) applications, bandwidth-thirsty and high throughput multimedia systems, high spectral efficiency requirements due to congested and expensive spectrum as well as an escalating number of Internet-of-things (IoT) connected devices. Furthermore, 6G is expected to support emerging technological advancements such as robotics-based industrial automation, smart rail mobility and connectivity in remote areas [7].

To meet such requirements, several key technologies have been proposed by researchers including massive MIMO, millimetre wave (mmWave) communications, wireless energy harvesting (EH), machine-to-machine (M2M), and device-to-device (D2D) communications [8]- [9]. The work in [10] highlights the future of ultra-dense networks beyond 6G as shown in Figure 1.1.

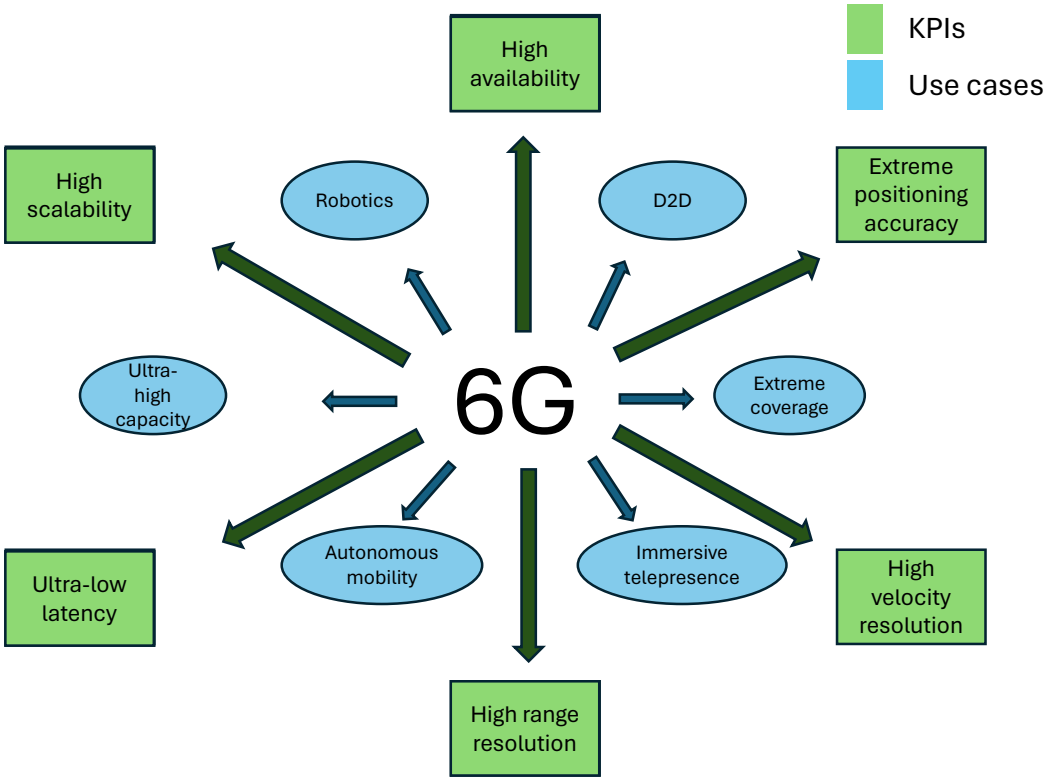

Figure 1.1: 6G and Beyond [1].

1.2 Towards 6G and Beyond

1.2.1 6G Requirements

6G and beyond wireless networks are expected to enhance technologies standardized by previous generations as well as adopt completely new technologies to keep up with the increasing demand for wireless connectivity [7]. Even though 6G requirements are still in their initial stages, both academic and industrial parties actively develop new and enhanced proposals for the next generation of wireless connectivity. The expected 6G requirements can be summarized as follows:

- **Lower latency**: While the 5G standard requires a 1 ms latency for specific applications, it is expected that the next generation will require ultra-low latency of around 0.1 ms. The

use of very wide bandwidths (> 1) GHz is envisioned to be the key to achieving modern latency targets.

- **Peak spectral efficiency**: 6G is expected to double the peak spectral efficiency of the previous generation by setting the target value at 60 Bit/s/Hz. In addition, it is estimated that the experienced spectral efficiency required by 6G is around 10 fold compared to that of 5G.

- **Energy efficiency (EE)**: EE is expected to play a more crucial role in the new standard to ensure that 6G wireless networks are more energy efficient compared to the previous generation.

- **Mobility**: 6G is expected to double the speed at which the user will still be able to have a seamless connection to the network.

- **Connection density**: while 5G requires the support of 10^6 devices/km^2, 6G is expected to support 10^7 devices/km^2 to meet the increasing demand of future wireless networks.

- **High data rates**: 6G will be expected to support extremely high peak data rates of up to 1 Tb/s for both outdoor and indoor environments with user-experienced data rates of around 1 Gb/s.

Table 1.1 summarizes the expected requirements for the 6G standard [7].

1.2.2 6G Enabling Technologies

In order to meet the aforementioned requirements, 6G is expected to enhance immature technologies proposed for 5G as well as utilize new techniques on the infrastructure, spectrum, protocol, and algorithmic levels. Hence, the 6G enabling technologies can be summarized as follows:

- **Ultra-massive MIMO**: while massive MIMO is one of the main features of 5G networks, practical implementations such as 2D discrete Fourier transform (DFT) have limited the

Table 1.1: Expected 6G key performance indicator (KPI) requirements.

KPI	**5G**	**6G**
Latency (ms)	1	0.1
Peak data rate (Gb/s)	20	1000
Experienced data rate (Gb/s)	0.1	1
Peak spectral efficiency (Bit/s/Hz)	30	60
Experienced spectral efficiency (Bit/s/Hz)	0.3	3
Mobility (Km/h)	500	1000
Maximum bandwidth (GHz)	1	100
Energy efficiency (Tb/J)	not specified	1
Connection density (Devices/Km^2)	10^6	10^7

performance of 5G massive MIMO systems to suboptimal performance levels compared to the optimal results in the literature. Therefore, 6G is expected to fully exploit the additional gains brought about by massive MIMO on a very large scale [11].

- **Intelligent reflecting surfaces (IRS)**: with the increasing demand for wider bandwidths to support higher data rates, higher frequency bands have plenty of unused spectrum that can be utilized to achieve such targets. However, the propagation characteristics of higher-frequency carriers are less favourable. The IRS technology can be utilized to combat the adverse channel conditions in higher bands. In principle, the IRS consists of multiple elements with unconventional electromagnetic properties [7]. In addition, these IRS elements are controlled by reprogrammable phase shifters that can be adjusted to enhance the quality of the channel between the transmitter and the receiver as illustrated in Figure 1.2.

- **Cell-free and user-centric networks**: even though the massive MIMO technology adopted by 5G is capable of dealing with the increasing number of active users in the network, the cellular nature of current massive MIMO systems is still unable to realize the full

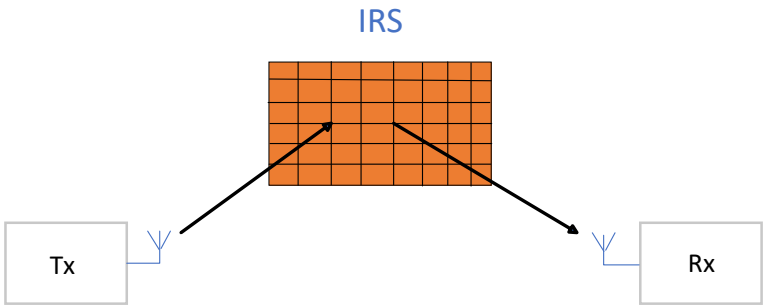

Figure 1.2: An IRS-assisted wireless communications system.

benefits brought about by the massive MIMO technology. This is partly due to the inter-cell interference problems. On the other hand, the inter-cell interference problem can be solved using the so-called cooperative multi-point (CoMP) transmission where two or more base stations (BS)s act as a single coherent transmission/reception unit. However, such technology requires sharing extensive user data and control information which is considerably more expensive in terms of the computational complexity of the setup, raising questions about the scalability of the technique. On the other hand, the cell-free concept is expected to address the inter-cell interference problem by adopting a cell-less network architecture as the name implies. The cell-free concept proposes a network setup where a massive number of antennas are distributed all over the served geographical area instead of being mounted on a single BS tower. Hence, the cell-free massive MIMO is one of the most promising proposals in the cell-free network architecture envisioned for 6G. However, with the cell-free concept emerges new challenges such as the initial/random access procedure which needs to be completely redesigned to support the distributed new architecture [12–14].

- **Integrated space and terrestrial networks**: to address the coverage problems at sea and in rural areas, the integrated space and terrestrial architecture proposes a multi-tier network that consists of three layers: the space-based layer which comprises low and medium earth orbit satellites, the air-based layer which comprises of aircraft, unmanned

aerial vehicles (UAV)s and airships, and the conventional ground-based terrestrial networks. Hence, 6G is expected to adopt a seamless integration across the three layers to address the access coverage problems in previous generations [7].

- **Terahertz (THz) communications**: given the linear relationship between the network capacity and the network bandwidth, the need for a wider spectrum will always exist. Historically, we have not been able to utilize higher frequency bands due to challenges in the electronics design. However, thanks to recent technological advancements, the mmWave frequency which ranges from 24 GHz to 100 GHz has already been included in 5G. Additionally, it is estimated that even with the adoption of mmWave frequencies, the current spectrum is not sufficient to address dense future wireless networks. Therefore, the sub-Terahertz (sub-THz) (above 100 GHz) is envisioned to be the next range that is utilized to meet the unprecedented demand of 6G networks [15]. Despite the featured studies in the literature, there are challenging problems facing the implementation of THz communications in practice such as the range effective serving area which needs addressing before unlocking the full potential of this spectrum [16, 17].

- **Visible light communications (VLC)**: along with the THz communications, both THz and VLC are called "6G enablers on the spectrum level". However, unlike THz, VLC operates in the infrared range ($400-770$ THz) and offers extremely high bandwidth, better security, and robustness to electromagnetic interference. Therefore, researchers have been exploring VLC as a viable alternative for future wireless applications that require large bandwidths and/or dense deployment. However, VLC also suffers from the range problem due to the physical characteristics of the wave in the aforementioned frequency level [18].

- **New waveform, full-duplex, and modulation schemes**: without enhancing lower layers' algorithms and protocols, 6G will not be able to capitalize on the aforementioned new technologies. Therefore, new channel coding and decoding techniques with excellent performance are required to realize the high-reliability requirements in 6G net-

works. Additionally, new modulation schemes are required to address the fading benefits of higher-order quadrature amplitude modulation (QAM) due to hardware imperfections which in turn paves the way for higher spectral efficiency and data rates. Furthermore, a new waveform with a smaller peak-to-average-power ratio (PAPR) is crucial for maintaining the high reliability of low-energy IoT and edge devices. Moreover, a full duplex which implies that the transmission and reception are performed on the same frequency or time resources is another interesting technique that has the potential to drastically enhance the spectral efficiency in current half-duplex wireless systems. However, additional challenges arise when considering the full-duplex architecture such as sub-band interference, which requires more advanced interference mitigation techniques and perhaps additional measurement reports [19, 20].

- **Advanced interference management protocols**: the aim of the wireless communication system is to serve multiple users at the requested QoS level using shared radio resources. This gives rise to the resource allocation problem which can become extremely challenging as the number of network resources grow. Up until the current wireless communication systems, OMA techniques have been used by allocating resources in a non-overlapping manner along a specific dimension. However, this concept scales the required resources with the number of users in the network, leading to various inefficiencies. Therefore, more advanced MA techniques such as the non-orthogonal multiple access (NOMA) and the rate-splitting multiple access (RSMA) have been considered for 6G and beyond networks. In power-domain NOMA, the signals for all users are encoded using the superposition coding (SC) technique at the transmitter with different power levels. Hence, NOMA utilizes the power domain to multiplex users while allowing them to share the time and frequency resources thereby enhancing the spectral efficiency of the system. At the receiver side, the successive interference cancellation (SIC) technique is used to remove more severe interference and enhance the signal-to-interference-plus-noise ratio (SINR). On the other hand, the RSMA technique is based on dividing users' messages into common and private parts and is therefore considered as a generalization

of NOMA [21, 22].

- **Artificial intelligence (AI)-driven algorithms**: the machine learning (ML) field in AI has proven to be extremely useful in wireless communications design. Conventional algorithmic approaches are model-based and are only optimal when the model is accurate and the underlying assumptions hold. However, this is more challenging in practice where the propagation environment changes drastically leading to suboptimal behaviours. On the other hand, ML algorithms are data-based, which means that they can capture unknown system behaviours and take them into account during the training phase. Additionally, conventional optimization algorithms do not scale well with an additional number of decision variables resulting in costly algorithms in terms of computational complexity. However, ML-based algorithms offer much lower computational complexity for a negligible performance loss, allowing fast computations for latency-sensitive applications. Hence, it is expected that 6G will adopt AI-driven approaches to reap the benefits of ML-based algorithms for cross-layer design [23, 24].

1.2.3 6G Use Cases

The next generation of wireless networks will require unprecedented performance requirements. Such requirements are driven by exciting new applications. Hence, the following use cases are expected for 6G networks:

- **Robotics and industrial automation**: the future manufacturing business is envisioned to be robotics-based which requires ultra-high reliability wireless networking with $0.1 - 1$ ms round-trip time. Therefore, 6G is expected to play a crucial role in realizing such a reliability metric.

- **Smart rail mobility**: while 5G has already proposed enhancements to rail connectivity, true seamless networking for data-intensive applications is expected to be one of the main use cases for the next generation of wireless networks.

- **Short-range device-to-device communications**: to reduce the load on the wireless infrastructure, the D2D paradigm allows devices to exchange data at high speeds with minimal involvement from the BS's side. Hence, D2D is expected to play a bigger role in 6G networks.

- **Ultra-high capacity xHaul**: in order to facilitate extreme capacity networks, the midhaul and backhaul networks must be optimized. Therefore, one of the primary use cases for 6G is the seamless integration of fiber and wireless backhaul which has low complexity and high bandwidth.

- **Connectivity in remote areas**: it is estimated that almost half of the world's population is still without internet connectivity, it is expected that one of the key targets of 6G is to guarantee a 10 Mb/s internet connection in every populated area.

- **Autonomous mobility**: with the increasing demand for autonomous transportation, a combination of stringent reliability, latency, and high mobility requirements are required to ensure the safety of the aforementioned applications. Furthermore, high data rates are often required to facilitate seamless data exchange between autonomous vehicles. Hence, 6G is expected to be the enabler for such reliable and interconnected transportation technologies.

1.3 NOMA as The Future MA Technique

Earlier wireless communication systems had two network resources, time and frequency. Hence, FDMA and TDMA were adopted for 1G and 2G to multiplex users in the frequency and time domains, respectively. However, this changed with the adoption of multi-user MIMO in 4G where the spatial dimension was used to multiplex different users while allowing them to share time and frequency resources with great success. With 5G massive MIMO, the spatial multiplexing concept is scaled to an unprecedented level. In order to reduce the resource allocation problem complexity and increase the spectral efficiency, operators are utilizing the spatial mul-

tiplexing technique as the sole interference management technique, i.e., by performing signal processing algorithms for beam design at the BS. However, in the case of far users, spatial multiplexing alone is not sufficient for mitigating interference especially when the ratio of users to antennas is high, resulting in poor experienced QoS.

By utilizing an additional dimension, NOMA preserves the spectral efficiency of sharing frequency and time resources while also addressing interference through SIC. Note that the power domain concept is also applied in the case of multi-user MIMO systems. In general, NOMA outperforms its OMA counterparts in spectral and efficiencies, power consumption, and fairness [21]. In addition, since the SIC is a serial algorithm, the decoding order in the NOMA system is important. Hence, the BS needs to perform joint power allocation and decoding order design to ensure that the benefits of NOMA are realized. Moreover, combining NOMA with multiple antennae, IRS and other techniques results in even better system performance. In particular, downlink (DL) multi-user multiple-input single-output (MISO)-NOMA systems have been studied extensively where NOMA outperforms OMA for various system objectives when the users have distinctively different channels. Additionally, multi-user uplink (UL) NOMA systems have also been studied where the BS schedules more than one user to transmit on the same set of resource blocks (RB) and then decodes each user's signal using SIC. More recently, IRS-assisted MISO-NOMA systems have been considered with the aim of combining the channel-strengthening capabilities of the IRS with the superior spectral efficiency and data rate of MISO-NOMA systems. However, the benefits brought about by NOMA come at the expense of more complex resource allocation algorithms and receiver architectures. Therefore, the resource allocation problem for NOMA systems is more complicated and generally has more constraints than its OMA counterparts. Therefore, low-complexity algorithms for NOMA systems with recent technologies are required to realize the benefits of NOMA and motivate it as a promising candidate to be the main MA technique in 6G and beyond.

1.4 Thesis Outline and Contributions

Future wireless networks are expected to be heterogeneous in nature due to the vast array of current and future applications. Therefore, NOMA is envisioned to be one of the main MA for future networks thanks to its advantageous spectral and energy efficiencies and fairness. In addition, the compatibility of NOMA with other advanced techniques allows for the implementation of high resource allocation techniques algorithms that meet the stringent capacity, latency, and reliability requirements of future networks. However, as the number of decision variables in the resource allocation problems increases, joint design of these coupled variables is required in order to produce efficient and acceptable solutions. Moreover, the conventional convex optimization approach dictates that non-convex joint design problems be solved separately often resulting in expensive algorithms from a computational perspective. Furthermore, by taking into account practical imperfections such as imperfect channel state information (CSI) at the transmitter, the resource allocation problem often becomes NP-hard which cannot be solved directly using the conventional optimization approach. Therefore, this thesis develops an alternative framework for solving such complicated resource allocation problems using ML-based algorithms. In particular, the reinforcement learning (RL) paradigm is utilized to propose robust and competitive policies for various resource allocation problems using NOMA with recent technologies.

Chapter 2 provides the fundamentals of power-domain NOMA and the details of the associated techniques. In addition, the resource allocation problems for two users using user-fairness maximization, system sum-rate maximization, and EE maximization and their solutions using conventional optimization methods are reviewed. Moreover, these resource allocation problems are also reviewed for the multi-user case along with their solution using the conventional approaches. To ensure the simplicity of the problems, they are all considered using a single antenna BS. Then, the combination of NOMA with other wireless techniques is briefly discussed.

Chapter 3 reviews the mathematical background and the methodology used in the subsequent chapters. In particular, the semi-definite programming (SDP) framework from convex optimization theory is briefly explained. Then, the RL framework is introduced which is the main

methodology used to solve the optimization problems in this thesis.

In chapter 4, the problem of maximizing the long-term sum-rate maximization for an IRS-assisted MISO-NOMA system is considered. In particular, the sum-rate maximization objective is considered subject to QoS constraints under channel uncertainty, maximum transmit power, and IRS phase shift constraints. The imperfect CSI is assumed to belong to a norm-bounded region and therefore, the problem is called worst-case joint beamforming and IRS phase shifts design. Furthermore, since the optimization variables are tightly coupled together in the objective, the problem is NP-hard and cannot be solved directly using convex optimization techniques. Additionally, the existence of the expectation operator which signifies the long-term aspect further complicates the problem by not allowing approximation methods to be applied to approximate the problem. Therefore, the RL framework is utilized to tackle these issues. In particular, the problem is reformulated into an RL environment where the state, action, and reward functions are defined. Then, a solution is developed based on the twin-delayed deep deterministic policy gradient (TD-DDPG) (or TD3 for short) agent. The TD3-based proposed algorithm is capable of producing robust beamforming and IRS solutions that satisfy the QoS of the users thanks to the selected actions and state spaces. Additionally, the computational complexity of the proposed algorithm is much lower than that of the iterative conventional algorithms which makes it more attractive to latency-sensitive applications. Furthermore, the simulation results demonstrate the robustness and competitive performance of the proposed TD3-based design. In addition, to demonstrate the applicability of the developed robust design, the proposed TD3-based algorithm is trained and tested for both fixed and dynamic-channels scenarios.

Chapter 5 dives more into the practicality of IRS-assisted MISO-NOMA systems. In particular, the system model considered in chapter 4 is useful only when the number of users is small since the number of required SIC operations scales linearly with the number of active users in the system. Hence, this chapter considers the user clustering problem first, i.e., pairing users into clusters to reduce the number of SIC operations required by the strong user. Therefore, a correlation-based algorithm is proposed to pair the users based on their channel coefficients. Then, the long-term sum-rate maximization problem is formulated. Particularly, the objective

is to maximize the sum rate for an IRS-assisted MISO-NOMA. Furthermore, the unbounded channel uncertainty model is considered in this chapter where the channel errors are assumed to be the result of imperfect channel estimation. Hence, the outage-based sum-rate maximization problem is formulated subject to QoS constraints, user and cluster power allocation, beamforming, and IRS phase shift design. The formulated robust design problem is challenging and NP-hard which means that conventional optimization methods cannot be directly applied. The zero-forcing (ZF) principle is utilized to tackle the beamforming design. Unfortunately, the problem is still non-convex and difficult to solve. Hence, the outage-based robust design problem is reformulated as an RL environment through careful selection of the state and actions space, and the reward function design. Then, a TD3-based algorithm is developed to jointly optimize user and cluster power allocation and the IRS phase shifts. The simulation results demonstrate the superior and robust performance of the proposed algorithm compared to different benchmark schemes in the literature. In addition, to demonstrate the applicability of the developed robust design, the proposed TD3-based algorithm is trained and tested for both fixed and dynamic-channels scenarios.

Chapter 6 discusses the EE-based design in IRS-assisted multi-user UL NOMA systems. In particular, a generalized robust framework for joint optimization of IRS and user power allocation based on the soft actor-critic (SAC) deep reinforcement learning (DRL) agent. The original robust design problem is formulated as the long-term EE maximization problem for an IRS-assisted UL NOMA system subject to QoS, maximum transmit power per user, and IRS phase shift constraints. Two problems are formulated as a result of considering both the bounded error model (BEM) and the unbounded error model (UEM) in the framework. The robust design for EE maximization is extremely challenging to solve directly given the intricate structure of the objective function. Therefore, the RL framework is applied to reformulate the original robust design problem into an RL environment where the state, actions, and reward functions are defined. To address the receive beamforming design at the BS, the minimum mean squared error (MMSE) with a SIC (MMSE-SIC) combiner is utilized at the BS due to its compact closed-form solution. However, since the available CSI at the BS is imperfect, the robust design is

realized through accurate user power allocation and phase shift design. Then, the SAC-based algorithm is developed which takes into account the assumed channel uncertainty model. Furthermore, the simulation results demonstrate that the developed algorithm outperforms existing algorithms in the literature in terms of the average achieved EE of the system by a significant margin. Note that in chapters $4, 5,$ and6, the proposed algorithms are trained and tested for fixed and dynamic-channel scenarios to demonstrate their adaptive capabilities. In addition, to demonstrate the applicability of the developed robust design, the proposed SAC-based algorithm is trained and tested for both fixed and dynamic-channels scenarios.

Finally, chapter 7 concludes the thesis and discusses possible future research directions.

Chapter 2

Literature Review

2.1 Overview

Despite being the most widely used MA technology in 4G and 5G, OFDMA struggles to meet the requirements of the next generations of wireless networks due to its high PARP, strict synchronization requirements and the inability to support massive number of connections due to its orthogonal design structure. Motivated by the shortcomings of OMA techniques, NOMA has received a great deal of attention as the main enabling MA technology to meet the unprecedented requirements in future generations of wireless systems. This is because unlike OMA methods which allocate orthogonal RBs for different users, NOMA can serve more than one user in the same RB, a feature that makes it more spectrum-efficient than any of the OMA techniques [25]. According to [26], theoretically, NOMA is the *optimal* technique for using the spectrum in both UL and DL scenarios. It is important to highlight that the basic concept of multiplexing more than one user's signals at the transmitter and detecting them at different receivers, is not new to the research community. In 1972, Cover introduced the mentioned problem and found the theoretical upper and lower bounds of the capacity region for the proposed system [27]. In 1973, Bergmans showed in [28] that SC is theoretically capable of approaching the capacity region for the Gaussian broadcast channel (BC). In 2015, an important milestone for NOMA was reached as a software-defined radio-based NOMA prototype was

proposed by Xion *et al.* in [29] for the two-user case. NOMA techniques can be classified into two main categories; power-domain NOMA and code-domain NOMA. Each has received a lot of attention from the research community. Several contributions have been made to the field of code-domain NOMA including low-density signature-based CDMA [30], trellis-coded multiple access [31], interleave division multiple access [32] and the pattern division multiple access [33]. In this thesis, the focus will be on power-domain NOMA. Instead of using codes as the means of multiplexing users, power-domain NOMA exploits the power domain to multiplex users (power-domain multiplexing) at the transmitter and employs SIC at the receiver ends to decode different users' signals. The power levels and decoding orders are usually selected based on the strengths of the channels. Note that from this point onwards, the term NOMA will refer to power-domain NOMA unless stated otherwise.

2.2 Fundamentals of DL NOMA

The key enabling techniques for NOMA are SC at the transmitter and SIC at the receiver. In a DL transmission of a NOMA system, the BS exploits power domain multiplexing based on users' channel strengths, where weaker users are allocated with more power levels to guarantee fairness [26]. Hence, NOMA uses the power domain to multiplex multiple users' signals into a single superimposed signal for transmission. The resource allocation problems in DL NOMA are examined with the two-user case and multiple-user case separately. Note that both cases are considered with a single antenna BS in this literature review.

2.2.1 Two-User DL NOMA

Consider a two-user DL NOMA scenario as illustrated in Figure 2.1 in which user 1 has stronger channel $|h_1|^2$ and user 2 has weaker channel $|h_2|^2$, i.e., $|h_1|^2 \geq |h_2|^2$ [2]. The superimposed transmitted signal from the BS can be expressed as [26]

$$y = \sqrt{p_1}x_1 + \sqrt{p_2}x_2, \tag{2.1}$$

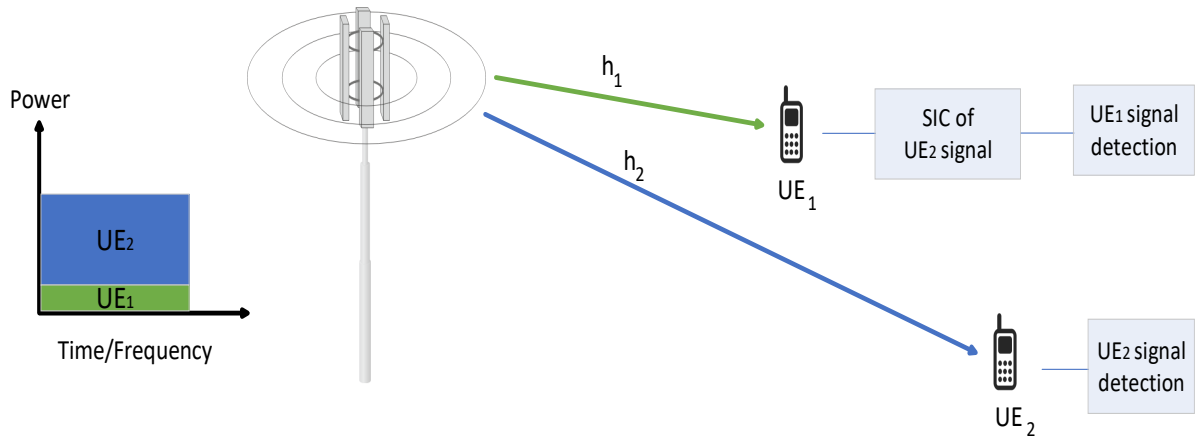

Figure 2.1: Two User DL NOMA scenario [2].

where p_1, p_2 represent the allocated power for users 1 & 2 respectively. x_1, x_2 are the intended message symbols to user 1 and user 2, respectively. The received signal at user n, where $n = 1, 2$, can be expressed as [26]

$$y_n = h_n(\sqrt{p_1}x_1 + \sqrt{p_2}x_2) + z_n, \tag{2.2}$$

where y_n is the signal received by user n, h_n is the channel coefficient from the BS to user n, where $h_n = h_n' d_n^{-\iota}$, h_n' is a random variable with Rayleigh distribution, ι is the path-loss exponent and d_n is the distance between user n and the BS. The term z_n represents the additive white Gaussian noise (AWGN) with zero mean and variance σ_n^2. To maintain fairness between users and facilitate SIC at the receiver ends, NOMA dictates that the weaker user should be allocated more power [34], That is $p_1 \leq p_2$. Therefore, according to [35], we can write the SINR for the two users as follows

$$\gamma_1 = \frac{p_1|h_1|^2}{\sigma_1^2}, \tag{2.3}$$

$$\gamma_2 = \frac{p_2|h_2|^2}{|h_2|^2 p_1 + \sigma_2^2}. \tag{2.4}$$

From (2.3) and (2.4), it can be seen that the stronger user does not suffer from any interference which allows it to decode its signal without any interference. This is possible for user 1 through the SIC, as it can decode the weaker user's message and subtract it from the received

superimposed signal to extract its signal. User 2, however, treats user 1 signal as interference and decodes its signal. This is the reason weaker users are allocated more power in NOMA, to enhance their SINR so that they can decode their messages in the presence of interference from other users in the cell. Having written the SINR expressions, the achievable rates for each user can be expressed as [35]

$$R_1 = B log_2(1+\gamma_1), \tag{2.5}$$

$$R_2 = B log_2(1+\gamma_2), \tag{2.6}$$

where B is the system bandwidth. To prove the superiority of NOMA over OMA, Chen *et al.* provided a rigorous mathematical proof in [36]. The work in [37] proved that NOMA can strictly increase the achievable rate region in the two user BC cases with different power levels. It is worth mentioning that a larger gap in power levels is desired in NOMA and leads to more gain over OMA. The effect of having distinctly different power levels will become clearer in the resource allocation techniques presented in the next section.

2.2.2 Multiple Users DL NOMA

The multiple-user DL NOMA is an extension of the two-user case, typically with 3 or more users, i.e., $n = 1,2,...,N$. Similar to the two user cases, the superimposed signal transmitted from the BS can be expressed as [2]

$$y = \sum_{n=1}^{N} \sqrt{p_n} x_n, \tag{2.7}$$

The received signal at any user n can be expressed as [2]

$$y_n = h_n \sum_{n=1}^{N} (\sqrt{p_n} x_n) + z_n. \tag{2.8}$$

Also, similar to the two user cases, user ordering is vital in NOMA to decide on a power allocation strategy. More on this will be discussed in the next section. Suppose that users are ordered according to their channel strengths such that:

$$|h_1|^2 \geq |h_2|^2 \geq |h_3|^2 \geq ..... \geq |h_N|^2. \tag{2.9}$$

According to (2.9), user 1 is the strongest user, and user N is the weakest user. Therefore, user N should be allocated with the largest amount of power as per NOMA standards to guarantee fairness between users. To formulate the SINR expressions for the N users, it is important to highlight the SIC and the decoding order among users as this will affect the SINR and consequently the achievable rate of each user. Suppose that user k, where $1 < k < N$, user k will have the ability to perform SIC and decode the messages of all weaker users, i.e., $n > k$ users. However, user k will experience interference from all stronger users $n < k$ for decoding its signal. The SINR expressions for the strongest, weakest and k-th users, respectively, can be written as [34]

$$\gamma_1 = \frac{p_1|h_1|^2}{\sigma_1^2}, \tag{2.10}$$

$$\gamma_2 = \frac{p_2|h_2|^2}{|h_2|^2 p_1 + \sigma_2^2}, \tag{2.11}$$

$$\gamma_k = \frac{p_k|h_k|^2}{|h_k|^2 \sum_{n=1}^{k-1} p_n + \sigma_k^2}. \tag{2.12}$$

Based on (2.10), (2.11), (2.12), the general SINR equation for the multi-user DL NOMA can be written as

$$\gamma_k = \begin{cases} \frac{p_k|h_k|^2}{\sigma_k^2}, & k = 1 \\ \frac{p_k|h_k|^2}{|h_k|^2 \sum_{n=1}^{k-1} p_n + \sigma_k^2}, & k \neq 1. \end{cases} \tag{2.13}$$

In the next section, the resource allocation problem in DL NOMA systems is discussed for both single and multiple antenna BS setups.

2.3 Resource Allocation in DL NOMA Systems

In the previous section, DL SINR and achievable rate expressions have been presented. These expressions will serve as the foundation for this section. Resource allocation can be defined as the selection of a transmit strategy that satisfies the constraints on the available resources in the system [13]. Since NOMA exploits the power domain to multiplex users' signals, power allocation plays a pivotal role in achieving the additional gains brought about by the NOMA

principle, particularly in single antenna NOMA systems.

To enhance the readability of this section we have divided the topic of resource allocation in NOMA systems into Four categories; single antenna systems, and multiple-antenna systems in which the resource allocation for both MISO and MIMO NOMA systems is discussed. Then, the literature on resource allocation for IRS-assisted NOMA systems is presented. Finally, ML-based resource allocation techniques for NOMA systems are discussed.

2.3.1 Single Antenna NOMA Systems

In single antenna NOMA systems, both the BS and the UEs are equipped with single antenna. Particular emphasis has been placed on power allocation for different system objectives [38–42]. The work in [43] proposed a power allocation scheme for two-user DL NOMA, while authors in [44] proposed a generalized power allocation scheme for both DL and UL in a two-user NOMA system. There have been several contributions for the multi-user NOMA systems, in [45], Wang *et al.* investigated the convexity of the weighted sum rate maximization problem in DL NOMA systems and provided an analytical solution in the absence of the quality of service (QoS) constraints. In [46], Timotheou *et al.* studied the power allocation in DL NOMA systems in terms of fairness and provided a low-complexity algorithm for the non-convex problem considered in the paper. EE optimization for DL NOMA has been studied as well, authors in [47] proposed an EE-based transmission strategy and showed that NOMA is more energy efficient compared to OMA techniques. Multiple channel or multiple carrier NOMA resource allocation has also been subject to several contributions, in [48], Cejudo *et al.* proposed a user pairing algorithm based on optimal channel gain ratio concept presented in the same paper. In [49], the problem of joint power and channel allocation in DL NOMA systems is presented as a combinatorial optimization problem and three heuristic solutions.

2.3.2 Multiple-Antenna NOMA Systems

Despite using a sophisticated scheme such as SIC to null interference, NOMA users in the DL still suffer from inter-user interference except for the strongest user in the cell which can decode

its signal interference-free. It has been well established that the use of multiple antennas -either at the transmitter or at the receiver or both- is an effective technique to combat interference in wireless systems [50], [51]. Even though the first implementation of multiple antenna schemes was applied to OMA systems, the basic concept can be extended to NOMA to yield optimized results in terms of spectral efficiency [38,42,52–62]. In [63], Ding *et al.* proposed a framework for MIMO-NOMA systems where they studied the effect of user pairing on the performance of NOMA systems and compared it to OMA-based MIMO results. MISO-NOMA design as illustrated in Figure 2.2 has also been studied extensively, in [64], beamforming-based NOMA design for maximizing the sum rate of stronger users in the system was proposed. In [65], a QoS-based design for sum rate maximization in DL MISO-NOMA systems was proposed where the resulting problem was solved via successive convex approximation (SCA) to obtain feasible power allocation schemes. None of these studies proposed an optimal solution to the MISO-NOMA DL power allocation. In [66], an important step towards MISO-NOMA optimality was addressed. The authors provided closed-form solutions for the optimal power allocation scheme in a two-user MISO-NOMA DL scenario given that the *quasi-degradation* condition is satisfied. They proved that if the channels are quasi-degraded, then their proposed algorithm can achieve the same performance as the dirty-paper coding (DPC), which is considered the optimal theoretical pre-coding algorithm [67]. In [68], Zhu *et al.* proposed an optimal beamforming design by extending the two-user case in [66] to multiple users under the quasi-degradation condition and quantized the performance gap between DPC versus the optimal performance of NOMA. Unlike the works in [67, 68], the authors in [69] considered the sum-rate maximization instead of the transmit power minimization objective which is more challenging since the objective function is non-convex and therefore cannot be efficiently solved directly. EE maximization in MISO-NOMA systems has also been subject to extensive studies. The work in [58] considered the beamforming design for global EE maximization in MISO-NOMA systems with minimum user rate requirement and transmit power constraint. Moreover, maximizing the user-fairness has also been considered for multiple antenna NOMA systems. The work in [70] considered the beamforming design problem in mmWave DL NOMA system. In particular, the authors

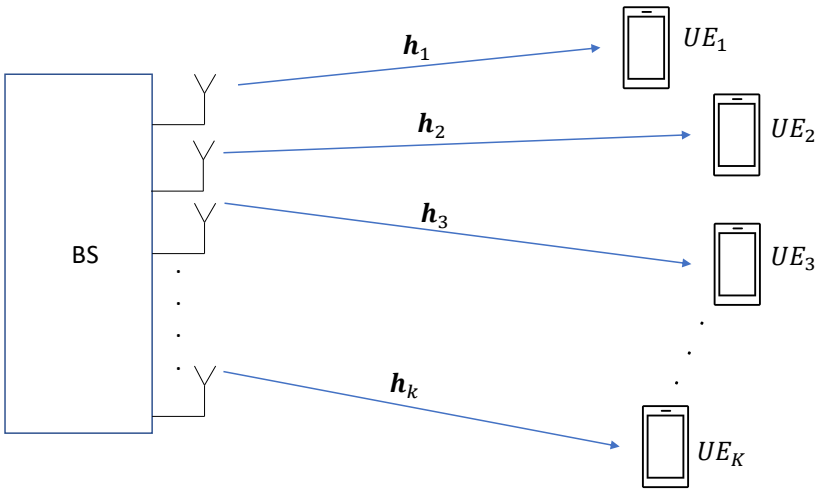

Figure 2.2: DL transmission of a MISO-NOMA system.

proposed a beamforming design with an optimal power allocation algorithm.

The challenges associated with multiple antenna NOMA design are greater than those for the single antenna. This is mainly because, in single antenna NOMA, the problem of power allocation is based on a single variable. However, in multiple antenna settings, more variables are involved due to the additional degrees of freedom known as spatial multiplexing. Therefore, the complexity of the multi-antenna NOMA resource allocation problem far exceeds that of the single-antenna case.

The MIMO case is considered to be an extension for the MISO case where the BS as well as the user it serves may be equipped with multiple antennas. In [63], a setup of a BS with M antennas serving different users, each with N antennas was considered. In [71], the authors assumed a BS with M antennas serving users each with $N > \frac{M}{2}$ antennas aiming to implement the concept of signal alignment. They justified this assumption that in small cells, heterogeneous users might be served with low-cost BSs in which user devices might be equipped with more

antennas than the BS. In MIMO-NOMA, optimally allocating network resources becomes even more challenging than in the MISO-NOMA case as user clustering is often added to the problem. One of the most common frameworks for MIMO-NOMA is the cluster-based design [72]. In cluster-based design, users are grouped into several clusters to strike a balance between system complexity and performance. Users can be allocated into clusters based on many criteria. Ali *et al.* in [73] selected users that are quasi-orthogonal to form a cluster. The work in [74] clustered users based on the SINR maximization criterion. In [26], the spatial direction is used as the basis for the clustering of the users. The cluster-based framework of MIMO-NOMA can be further divided into two main design approaches; the inter-cluster interference-free design and the inter-cluster interference-tolerant design. In inter-cluster interference-free design such as in [63], [71], the MIMO-NOMA system is decomposed into multiple SISO-NOMA channels. Two main advantages to such a design, first, when the MIMO-NOMA system is decomposed to separate SISO-NOMA channels, the challenging multi-antenna user ordering constraint can be circumvented which significantly lowers the system complexity. The second advantage is that ZF beamforming can be utilized to completely cancel the inter-cluster interference. However, the main practical challenge with this design is that it requires the user equipment to have at least the same number of antennas as the BS in [63] or more than half the number of antennas at the BS in [71], which is generally impractical.

The other type of cluster-based MIMO-NOMA setup is the inter-cluster interference-tolerant design which allows the existence of interference between the clusters in the system. It was proposed in [73], they used *decoding scaling weight* to increase the strength of the desired signals. The idea behind this design is that it allows for making channel gains of the users more distinctive which is highly desirable in any NOMA system as it helps achieve smooth SIC at the receivers. The main advantage of this design over the inter-cluster interference-free design is that it does not employ the ZF techniques, therefore, there are no constraints imposed either on the BS or the users in regards to the minimum number of antennas which makes it a more practical approach that can be employed to mmWave NOMA or massive-MIMO-NOMA as highlighted in [75]. So far the implementation of multiple antenna design is shown to in-

crease the performance of NOMA networks either in terms of energy efficiency or in optimized system throughput and interference cancellation. However, these promising features come at the expense of more complex resource allocation problems that are non-convex in general. In cluster-based MIMO-NOMA, balancing the energy and spectral efficiency is a trade-off where we can only settle for Pareto-optimal [76] solutions based on some design criterion. With any cluster-based MIMO-NOMA setup, resource allocation challenges arise in the following: The number of clusters in the system; the number of users allocated to each cluster; which users are allocated to which clusters; power allocation for each cluster and power allocation for each user in each cluster. It could be concluded that solving the resource allocation problem in cluster-based MIMO-NOMA optimally is a challenging task. Therefore, many studies in the literature deal with one or two of these challenges while assuming the ideal implementation of the others. While useful in analysing the underlying structure of the problem, they mostly end up with a sub-optimal solution due to the high computational complexity of optimal solutions.

Research contributions to solve resource allocation problems in cluster-based MIMO-NOMA systems can be generally grouped into two main categories [72]; solve the resource allocation problem jointly or solve a subset of them separately in a decoupled manner. In [77], Sun *et al.* proposed a solution that optimizes both user scheduling and power allocation via Monotonic optimization. Though it can obtain optimality, they mention that it is rather for benchmarking purposes than a practical one due to its high computational complexity. In [73], [78], the problem of user scheduling is investigated separately, solutions based on heuristic methods were proposed, while in [79], the same problem was considered while matching theory-based solutions were proposed instead. Several contributions studied power allocation as well. In [78], Liu *et al.* proposed a low-complexity power allocation algorithm that employs the bi-section method for MIMO-NOMA systems. In [79], authors proposed a low-complexity geometric programming-based algorithm to solve the power allocation in DL MIMO-NOMA networks. In [78], perhaps a more practical solution to the power allocation problem than previous efforts is presented in a closed-from using Lagrangian algorithms. Finally, it can be concluded that if convex optimization is considered to be the only method for solving the resource allo-

cation problem in MIMO-NOMA or any other NOMA scenario for that matter, then a solution that is both optimal and computationally viable cannot be obtained with the currently available computational power. This is one of the main motivations to consider alternative methods that can solve such complex problems with optimal or near-optimal performance, one of which is ML-based methods which will be discussed in detail in the following section.

2.3.3 IRS-assisted NOMA Systems

The IRS technology has been subject to extensive studies recently, thanks to the benefits it adds to conventional wireless communication systems. By controlling the phase shifts of the IRS elements, system designers can enhance the channel quality between the BS and the UEs. Hence, countless studies have considered the resource allocation problem in IRS-assisted NOMA systems [80, 81]. The work in [82] proposed a spatial division multiple access-based design for an IRS-assisted DL NOMA system. In particular, the BS station designs orthogonal beams to serve the near UEs. Then, IRS-assisted NOMA is used to ensure that edge UEs can also be served using the designed beams. The work in [83] considered an EE design of IRS-assisted DL NOMA networks. More specifically, the authors considered a resource allocation problem in which the beamforming vectors at the BS and the passive IRS elements are alternatively optimized to maximize the EE of the system. The work in [84] considered a similar problem with the objective of minimizing the total transmit power at the BS. The authors proposed a method to optimize the IRS to tune the channels in the MISO-NOMA system such that the quasi-degradation condition is satisfied. In terms of the system sum-rate maximization, the work in [85] proposed an SDP-based algorithm for optimizing the IRS phase shifts in a DL MISO-NOMA systems while utilizing the maximum-ratio transmission for the beamforming design subproblem. Moreover, the work in [86] considered the resource allocation problem in a multi-cluster DL MISO-NOMA system with the aim of minimizing the transmit power at the BS. The authors proposed a SOCP-ADMM-based algorithm to optimize the beamforming vectors and the IRS elements.

Overall, the resource allocation problem in IRS-assisted becomes more challenging as the num-

ber of design variables is increased. Therefore, ML-based methods have been utilized to address the complexity issue as discussed next.

2.3.4 ML-Based Resource Allocation for NOMA Systems

In this section, a brief introduction to different types of ML is presented. Then we focus on RL which is the building block for the advanced DRL as this kind of learning is often considered to hold the most potential for solving problems in wireless communications. ML refers to the field where computers can *learn* some features about a dataset without being explicitly instructed how to do so [87]. ML is generally classified into three main sub-fields; supervised learning, unsupervised learning and reinforcement learning. In supervised learning, the agent is provided the dataset along with the desired outcomes. The goal is that after it is well trained, it would be able to classify or predict new data "of similar type" correctly. Examples of supervised learning include linear and non-linear regression normally used in prediction models, and classification where the model is taught to classify data into two or more categories. Unsupervised learning is when the agent is fed a dataset without any labels, the goal of unsupervised learning is for the agent to group similar data points with common features. Examples of unsupervised learning include K-mean clustering, K-nearest neighbour and principle component analysis. Reinforcement learning is different from the other two, it learns by trial and error. The RL framework consists of an agent and an environment, if the agent takes the correct action, then it is given a reward (by the environment) for the correct choice while if the action taken is not the desired one, it is given a punishment in the form of a negative reward, and through a memory, the agent starts distinguishing the good choices from the bad ones by trying to maximize its reward. It can be seen that an RL agent unlike supervised and unsupervised agents, does not require a predefined dataset as it learns simultaneously while interacting with the environment in an "online" fashion. This is one of the practical advantages of RL over the other two forms of ML and the one that makes it attractive for wireless communication applications as we will see later in this section.

The fast-paced development of deep neural networks (DNNs) including convolutional NNs and

recurrent NNs which are said to mimic the learning mechanism of the human brain led to interesting applications of ML to almost all disciplines, wireless communications is no exception [88, 89]. Deep learning which is a method that employs DNNs to learn features of input data has been applied to many applications in wireless communications. Zhang *et al.* presented a survey on the application of deep learning for wireless communications in [90]. In [91], authors presented an overview of the applications of DRL in wireless communications. DRL is a combination of deep learning and RL which has shown great potential in solving some of the challenging problems in the wireless communication domain [92]. In [93], authors proposed a multi-agent DRL-based algorithm with centralized training and distributed execution to maximize the system throughput by coordinating interference between multiple cells. In terms of application to NOMA, [91] used deep learning to estimate the NOMA channel and learn the CSI using the long short-term memory (LSTM) nets. Kim *et al.* used deep learning in code-domain NOMA in [94] to carry out mapping and decoding of received data in NOMA systems. To solve the problem of user pairing in multi-carrier NOMA networks, authors in [95] proposed a multi-agent RL design. Xiao *et al.* proposed a fast Q-learning-based power allocation scheme to prevent MIMO-NOMA systems from jamming attacks. In [96], authors proposed a deep Q-network (DQN) algorithm to obtain near-optimal solutions to the joint problem of power and channel allocation in DL NOMA systems. Kang *et al.* designed a deep learning-based algorithm to minimize the mean squared error (MSE) by optimizing both precoding and SIC decoding in MIMO-NOMA systems. In [97], the authors proposed a Q-learning-based algorithm to solve the problem of joint user scheduling and power allocation in DL MIMO-NOMA systems. Ding *et al.* applied the DDPG agent, an advanced continuous action space DRL agent to UL Cognitive-Radio CR-NOMA to maximize the long-term throughput for an energy-constrained secondary user [98]. Recently, Xie *et al.* applied DDPG to an IRS-assisted DL MISO-NOMA system in [99] for both varying and fixed channels. [100] used a similar setup with single antenna BS and UAV-mounted IRS where a DDPG agent is used to optimize the optimal location for the UAV jointly with the optimal power allocation for the DL NOMA system.

The motivations for using ML algorithms as a means to solve resource allocation problems in NOMA systems can be summarised as follows:

- Models are not always accurate as they suffer from practical imperfections and since ML algorithms are data-driven, not model-based, they can be more accurate when trained properly.
- Model-based resource allocation problems cannot accommodate future heterogeneous requirements while ML designs can be more flexible.
- Model-based solutions suffer from severe performance degradation when trying to solve a non-convex problem, which is the case in the majority of resource allocation problems in NOMA systems. ML on the other hand can learn the underlying structure of the problem, especially with the powerful function approximation feature of DNNs, and be able to provide optimal or near-optimal solutions if care is taken when designing and training the algorithms.

Additionally, two problems from the literature are briefly discussed to illustrate how RL agents are applied to problems in the wireless communications domain.

2.4 Summary

This chapter summarizes the relevant literature on resource allocation in DL NOMA systems. In particular, the conventional optimization techniques applied for different system objectives with single and multiple antennas as well as IRS-assisted NOMA systems are shown. Then, the literature on ML-based resource allocation in NOMA systems was reviewed, focusing on the RL-based approaches with lower computational complexity than the conventional schemes. The next chapter lays the mathematical background for the conventional convex optimization technique and the RL framework.

Chapter 3

Mathematical Background and Methodology

3.1 Overview

The resource allocation problem in a wireless communication system generally consists of allocating a finite amount of network resources to either maximize a utility or minimize a cost function subject to some performance constraints. Therefore, resource allocation problems are often formulated as mathematical optimization problems to enhance their readability and allow for the application of mathematical tools to solve the formulated problems. Generally, optimization problems are categorized into convex and non-convex optimization problems. Convex optimization refers to the case where the objective function and all the constraints are convex. Moreover, this class of optimization problems can be solved efficiently using off-the-shelf solvers such as CVX [101]. Unfortunately, most practical problems are non-convex due to their complicated structure. Despite the existence of extensive literature on the subject of approximation and relaxation of non-convex functions, most of the relevant resource allocation problems in the wireless communications domain are still non-convex. One of the main drawbacks of non-convex optimization is the lack of a generalized framework which leads to the development of problem-specific algorithms. However, in the constantly evolving world of today, this

approach is not particularly useful as there are countless applications created every day. ML-based algorithms hold a promising answer due to their inherently different way of solving the problem. The non-convexity of the problem has no impact on ML-based methods rendering them extremely useful for solving large-scale problems. In this chapter, a brief overview of convex optimization techniques focusing on SDP is presented followed by the RL frameworks.

3.2 Convex Optimization

3.2.1 Convex Sets

Let $\mathcal{C} \in \mathbb{R}^n$ be a convex set. This entails that the line segment between any two elements in $\mathcal{C}$ lies in $\mathcal{C}$, i.e., $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$ [102]. This is mathematically expressed as

$$\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathcal{C}, \tag{3.1}$$

where $\theta \in [0,1]$. Therefore, $\mathcal{C}$ must contain no empty regions and be a solid body. In addition, the convex set $\mathcal{C}$ remains convex after performing the following operations [103]:

- $\mathcal{C}$ remains convex under any intersection operations.
- The affine transformation set $\mathbf{A}\mathcal{C}+\mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$, preserves its convexity as

$$\mathbf{A}\mathcal{C}+\mathbf{b} = \{\mathbf{A}\mathbf{x}+\mathbf{b} | \mathbf{x} \in \mathcal{C}\}. \tag{3.2}$$

- $\mathcal{C} \subseteq \mathbf{dom}\, P = \mathbb{R}^n \times \mathbb{R}_{++}$ entails that its perspective transform which is expressed as

$$P(\mathcal{C}) = \{(\mathbf{x},t) \in \mathbb{R}^{n+1} | \frac{\mathbf{x}}{t} \in \mathcal{C}, t > 0\}, \tag{3.3}$$

  is also convex, where $\mathbb{R}_{++}$ denotes the set of all positive numbers [102, 103].

3.2.2 Convex Cones

In addition to the aforementioned general convex sets, there is a special type of convex sets called convex cones. A convex cone $\mathcal{C}$ is a set in which for each $\mathbf{x} \in \mathcal{C}$ and $\theta \geq 0, \theta\mathbf{x} \in \mathcal{C}$. This

can be expressed as

$$\theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 \in \mathcal{C}, \forall \theta_1 \geq 0, \forall \theta_2 \geq 0, \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}. \tag{3.4}$$

Different types of convex cones are being extensively used in the wireless communications domain including the second-order cone $\mathcal{C} = \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} | ||\mathbf{x}||_2 \leq t, \}$ where $||.||_2$ is the Euclidean norm, the nonnegative orthant $\mathbb{R}^n_+$, and the positive semidefinite cone $\mathcal{C} = \{\mathbf{X} \in \mathbb{S}^n_+ | \mathbf{X} \succcurlyeq 0\}$, where $\mathbb{S}^n_+$ denotes the set of symmetric positive $n \times n$ matrices.

3.2.3 Convex Functions

Let $f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function. Given $\forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}\, f(\mathbf{x})$, and $\forall \theta \in [0,1]$, then the following inequality must hold [102]:

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}). \tag{3.5}$$

The function $f(\mathbf{x})$ is called strictly convex if strict inequality holds in (3.5). Additionally, suppose that $f(\mathbf{x})$ is continuously differentiable for all $x, y \in \mathbf{dom}\, f(\mathbf{x})$, then $f(\mathbf{x})$ is convex if and only if $\mathbf{dom}\, f(\mathbf{x})$ is convex and the following inequality holds

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}). \tag{3.6}$$

where the inequality in (3.6) is the first-order Taylor series approximation of f near $\mathbf{x}$. This entails that if the first-order Taylor series expansion of a function is always a global of f, then the function $f(\mathbf{x})$ is convex. Moreover, if $f(\mathbf{x})$ is twice differentiable, then $f(\mathbf{x})$ is convex if:

$$\nabla^2 f(\mathbf{x}) \succcurlyeq 0, \forall \mathbf{x} \in \mathbb{R}^n, \tag{3.7}$$

where $\nabla^2 f(\mathbf{x})$ is the Hessian function of the second derivative. Hence, expression in (3.7) implies that for $f(\mathbf{x})$ to be convex, its curvature at $\mathbf{x}$ must be positive [102, 103].

3.2.4 Convex Optimization Problem

Now that convex optimization preliminaries are explained, the general convex optimization problem can be expressed as

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \quad f_0(\mathbf{x}) \\ &\text{subject to} \\ &f_i(\mathbf{x}) \leq 0, \ i = 1, ..., m, \\ &h_i(\mathbf{x}) = 0, \ i = 1, ..., p, \end{aligned} \tag{3.8}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the vector containing the optimization or the decision variables, $f_0(\mathbf{x})$ is the objective or the cost function, the convex functions $f_i(\mathbf{x}), i = 1, ..., m$ and the linear functions $h_i(\mathbf{x}), i = 1, ..., p$ are called the inequality and equality constraints, respectively. The aim of a convex optimization program is to find the optimal vector $\mathbf{x}^*$ that minimizes $f_0(\mathbf{x})$ while satisfying the equality and inequality constraints. Note that in order for (3.8) to be classified as convex problem, the following three conditions must hold [102, 103]:

- The objective $f_0(\mathbf{x})$ must be convex.
- The functions in the inequality constraints $f_i(\mathbf{x}), i = 1, ..., m$, must be convex.
- The functions in the equality constraints $h_i(\mathbf{x}), i = 1, ..., p$, must be affine.

3.2.5 Semidefinite Programming

One of the most recent and interesting class of convex optimization problems is called SDP. The general form of SDP can be expressed as [102]

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \quad \mathbf{c}^T \mathbf{x} \\ &\text{subject to} \\ &\mathbf{x}_1 \mathbf{F}_1 + \mathbf{x}_2 \mathbf{F}_2 + ... + \mathbf{x}_n \mathbf{F}_n + \mathbf{G} \preccurlyeq 0, \\ &\mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned} \tag{3.9}$$

where $\mathbf{G}, \mathbf{F}_1, ..., \mathbf{F}_n \in \mathbb{S}^{k\times k}$, and $\mathbf{A} \in \mathbb{R}^{p\times n}$. Furthermore, a standard form of SDP can be written as [102]:

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}} \quad \mathbf{tr}(\mathbf{CX}) \\ &\text{subject to} \\ &\mathbf{tr}(\mathbf{A}_i\mathbf{X}) = b_i, \ i = 1, ..., p, \\ &\mathbf{X} \succcurlyeq 0, \end{aligned} \tag{3.10}$$

$\mathbf{C}, \mathbf{A}_1, ..., \mathbf{A}_n \in \mathbb{S}^{k\times k}$. Note that the SDP form in (3.10) has linear equality constraints and a positive semi-definite constraint imposed on the variable $\mathbf{X} \in \mathbb{S}^{k\times k}$.

3.3 Reinforcement Learning

Reinforcement learning is an ML technique where an agent learns through trial and error [104]. It does so by trying to maximize its long-term reward. The two main elements in RL are the agent itself and the environment. The agent is the hardware in which the algorithm is deployed, while the environment is everything else. The learning elements for an RL system are a policy, a reward signal and a value function. The goal of the designer is to formulate the reward function so that the agent reaches an optimal policy, which is the optimal mapping function that maps the inputs to the best actions that yield the highest long-term reward for the agent. Figure 3.1 shows the agent-environment interaction of RL agents. As shown in Figure 3.1, at time step t and current state s^t, the agent takes an action a^t based on some policy π, and as a result, the environment produces a reward r^{t+1} and a new state s^{t+1}. If the reward is positive, the agent will realize that the action a^t taken when the system was at state s^t was a good choice as it yielded a positive reward. The agent keeps doing so until it reaches a level where for any system state, it takes the action that has the highest long-term reward based on previous interactions. The advantage of RL can be seen here as the agent does not need to know any information about the system it is optimizing except for good/bad actions. However, there is a fundamental challenge in RL agents known as the *exploration-exploitation* problem, which refers to the dilemma when the agent chooses to limit its actions to only those taken from past interactions with the

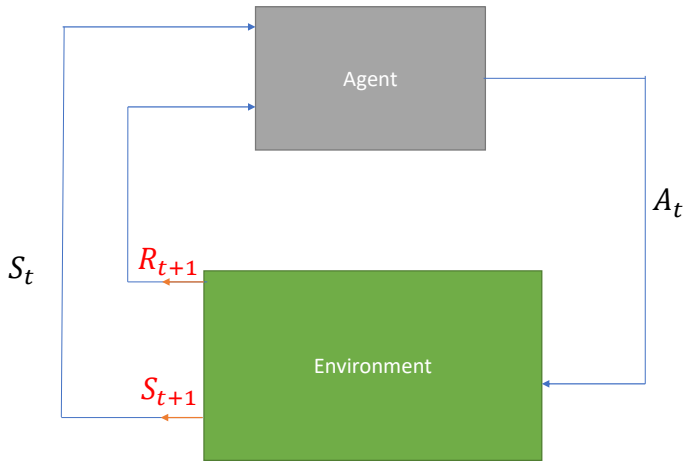

Figure 3.1: RL agent components.

environment "*exploit*", it will not be able to explore new actions which might provide higher rewards. On the other hand, if the agent chooses to always "*explore*", it will never converge to an optimal or near-optimal policy which will render the agent useless.

RL algorithms can be categorized in several different ways depending on perspective, a useful categorization divides RL methods into model-based and model-free methods. Model-based methods require a model of the environment for effective learning which makes it less appealing in the case of wireless communications where random variables play an essential role. Model-free methods are more interesting and applicable to many wireless communication scenarios as they do not require a model to start learning. An overview of model-free RL algorithms is illustrated in Figure 3.2. As Figure 3.2 shows, there are prediction methods such as Monte Carlo (MC) and temporal-difference (TD), these methods are used in the RL algorithms to estimate the value function of states and actions. They use different approaches to estimate the value function where TD updates the value after a defined number of time steps using bootstrapping, while MC methods only provide an estimate after the whole episode is terminated. Such dif-

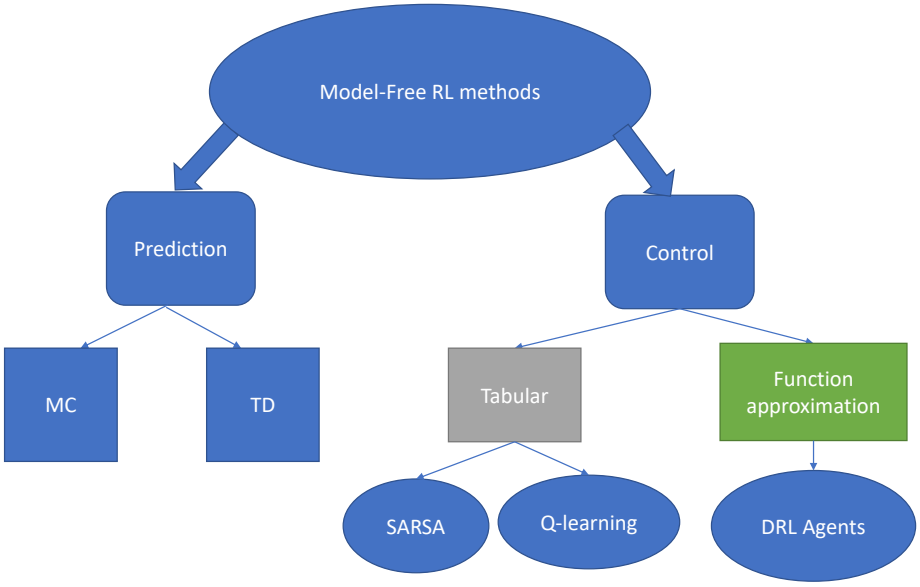

Figure 3.2: Model-free RL algorithms.

ferent approaches lead to different estimates of the value function, MC methods are prone to higher variance while TD methods suffer from bias.

In general, however, TD methods are more practical and more frequently result in better outcomes [104]. RL control methods are decisive factors as to whether the algorithm is suitable for certain applications or not. Control methods can be further divided into two sub-domains: On-Policy agents and Off-Policy agents. On-policy agents use a single policy, which means the policy used to generate data samples by interacting with the environment is the same policy that is being optimized during training. Off-policy agents on the other hand use two policies, one to generate samples by interacting with the environment while the other is optimized by experiences obtained from the first policy. In terms of learning optimal policies, there are two main methods used by all RL algorithms; tabular methods and function approximation methods. Tabular RL agents like Q-learning and state-action-reward-state-action (SARSA) use tables to save the Q value, which is the value of taking an action a^t from state s^t following some policy π afterwards. Since they use tables to store the Q values, this limits the table entries to a finite

number which may be considered a disadvantage in some applications. Therefore, Q-learning and SARSA algorithms are limited to finite-space problems, which means the state space set as well as the action space set must be of a finite length to be incorporated into memory. The update of the action-value function in the SARSA "on-policy" algorithm is given by the following [105]

$$Q(s^t,a^t) \leftarrow Q(s^t,a^t) + \alpha^{'''}\left[r^{t+1} + \delta Q(s^{t+1},a^{t+1}) - Q(s^t,a^t)\right], \tag{3.11}$$

while the update of the action-value function for the "off-policy" Q-learning algorithm is expressed as [106]

$$Q(s^t,a^t) \leftarrow Q(s^t,a^t) + \alpha'''\left[r^{t+1} + \delta \max_a Q(s^{t+1},a) - Q(s^t,a^t)\right], \tag{3.12}$$

where $\alpha^{'''}$ is the learning rate and δ is the discount factor which determines the current value of future rewards. The difference between on-policy and off-policy RL agents can be seen from (3.11) and (3.12) where Q-learning takes the action that has the highest Q value based on a greedy policy, while SARSA has a more conservative choice as it selects the next action based on the current policy. In general, tabular RL algorithms use the algorithm illustrated in Figure 3.3. However, due to the limited number of table entries, tabular RL methods cannot be applied to continuous action and state space environments which severely limits their application in modern wireless communication systems. Hence, using a DNN as an RL agent to approximate an infinite table is discussed next.

3.4 Deep Learning

Deep learning is a sub-field of ML which deals with DNNs. Unlike shallow NNs, DNNs imply that the deep learning model comprises at least two hidden layers. Deep learning has proved to be one of the most successful ML techniques thanks to its superior learning capabilities compared to other ML algorithms.

Deep learning uses DNNs to learn the relationship between the features in the input layer and the labels in the output layer. Unlike conventional optimization methods, ML algorithms learn

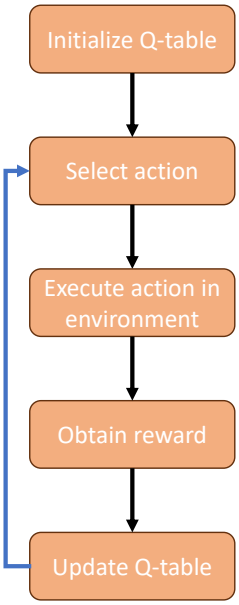

Figure 3.3: The general algorithm for tabular RL methods.

autonomously using the backpropagation algorithm to adjust the DNN weights to achieve the desired outcome [107]. Deep learning is often applied for supervised learning problems where the deep learning model is trained using solved examples such that after training, the model can predict the outcome for never-seen-before inputs. There are different deep learning models optimized for specific applications including convolutional neural networks for image processing, recurrent neural networks for time-series prediction and others. The most important feature of DNNs is that they are considered to be universal function approximators [108]. Therefore, deep learning models will be utilized as agents in the RL framework which allows them to be applied to a much wider array of applications in the wireless communications domain.

3.5 Deep Reinforcement Learning

Tabular RL algorithms such as SARSA and Q-learning struggle with three main issues:

- Very large state space environments like in the case of Go Board which has a state space cardinality of 10^{170}.

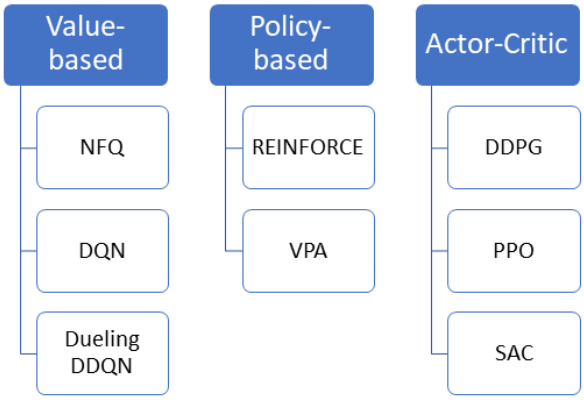

Figure 3.4: DRL algorithms summary.

- Continuous state and/or action space(s).
- A combination of both.

Unfortunately, almost every target application has at least one of these features. DRL agents were designed to solve these problems by employing DNNs. It has been established that DNNs are excellent function approximators [109]. Therefore, instead of using tables to store the Q value for each action-state pair, DRL agents use DNNs to approximate the value function itself and in this way, a well-trained DRL agent will always be able to reach at least a near-optimal policy of mapping input states to optimal actions. There is more than one type of DRL agents as Figure 3.4 shows. Value-based methods use DNNs to approximate the value function which allows them to handle large state space problems. In fact, Google's DeepMind showed that DQN -which is a value-based DRL agent- was able to master the Atari Breakout game to the level of a professional human player. Neural Fitted Q-iteration or (NFQ) was the first attempt to approximate the action-value function with a NN [110]. However, since the input data in RL environments is not stationary and constantly changing depending on the policy of the agent, NFQ had stability issues during training. To overcome these stability issues which were the results of data being neither stationary nor independent identically distributed (IID), research studies led to one of the most important DRL agents of all time, the DQN algorithm [111]. DQN solved NFQ issues by using target networks, a delayed copy of the main NN to fix the target

for several training steps to help stabilise the training process. DQN also adds the experience replay feature where the agent uses a replay buffer to save a large amount of past experiences (between 10,000-1,000,000) to make training data more diverse and closer to an IID setting. Figure 3.5 shows the DRL framework in which the agent is a DNN.

Even though they can deal with continuous and very large state spaces, value-based DRL meth-

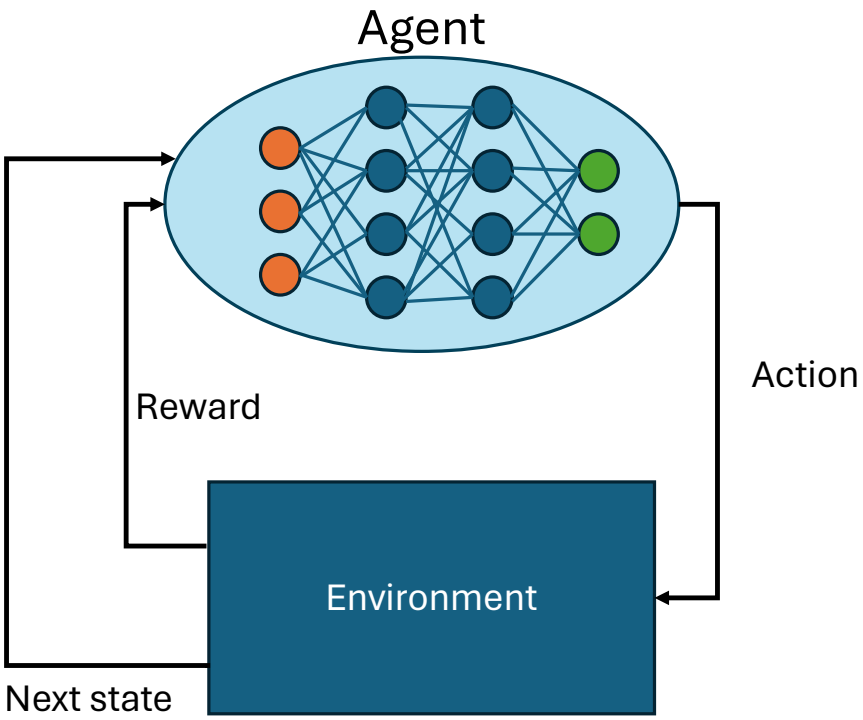

Figure 3.5: The DRL framework.

ods are restricted to discrete action spaces. Policy-gradient-based methods are DRL algorithms that optimize the policy of the agent directly and therefore, they can deal with continuous action spaces. Examples of such agents are Reinforce [112] and Vanilla policy gradient (VPA) [113]. Despite being able to directly optimize policies, pure policy-gradient-based DRL algorithms suffer from high variance during training which is highly undesirable for most applications. Actor-critic agents are the state-of-the-art in DRL methods [114]. Such agents emerged by combining both value-based and policy-based methods into a single agent as shown in Figure 3.6. Note that the architecture shown in Figure 3.6 is only one way of implementing the A2C

agent and by no means the only way. As the name implies, there are two main networks in actor-critic agents, one that takes actions called the actor, while the critic network provides the value that results from the action taken by the Actor. DDPG was introduced in [115] and it

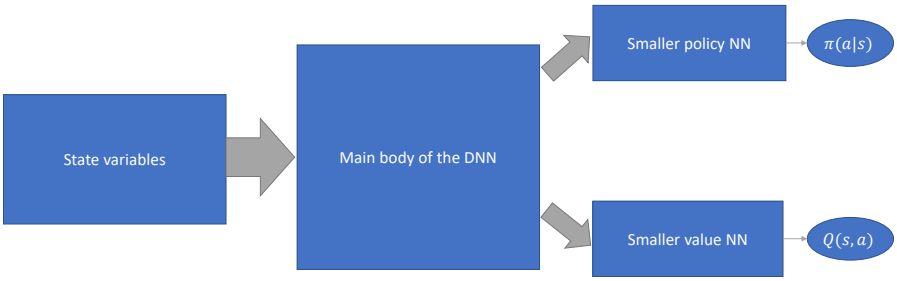

Figure 3.6: Shared-DNN architecture for actor-critic DRL agents.

can handle continuous state as well as action spaces which makes it a perfect candidate for many applications in wireless communications as highlighted in [98] and [100]. DDPG uses a deterministic policy, which means that it generates a single deterministic action for an input state. Because DDPG is a deterministic agent, Gaussian noise is added to the actions taken by the agent to encourage exploration. SAC [116] is a state-of-the-art DRL algorithm which was introduced in 2018. It is similar to DDPG being an off-policy algorithm, but it optimizes a stochastic policy instead of a deterministic one to encourage exploration. The SAC can do so because the entropy of the stochastic policy is part of the value function that the agent is trying to optimize. This way, the SAC adds a bonus reward to the agent for getting into high entropy situations. Proximal Policy Optimization (PPO) was introduced by Schulman *et al.* in 2017 [117]. Unlike SAC and DDPG, PPO is an on-policy algorithm, which means it cannot use previous experiences as it optimizes the same policy being used to take action. PPO uses a clipped objective which makes for smoother and more stable training as well as reduced variance. PPO is faster than SAC while SAC is more sample efficient as shown in [118]. Therefore, the choice of a better method depends on the application.

Finally, continuous action space-capable DRL algorithms are currently a hot research topic in

wireless communications as shown by the examples above. In the next chapters, DRL-based resource allocation techniques for IRS-assisted NOMA systems are presented.

3.6 Summary

In this chapter, the mathematical background and methodology used in the contributions of this thesis are presented. A brief introduction to convex optimization is discussed first. In particular, the details of SDP class of convex optimization is provided as it is used as a benchmark scheme in the following chapters. Then, the RL framework is presented focusing on table-based RL algorithms. Moreover, the shortcomings of tabular RL methods are discussed and a brief introduction to deep learning is then presented. Finally, the DRL framework which uses DNN as agents in the RL framework is discussed focusing on the actor-critic agents which are used extensively in the upcoming chapters.

Chapter 4

Worst-Case Robust Design for an IRS-Assisted DL MISO-NOMA System

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

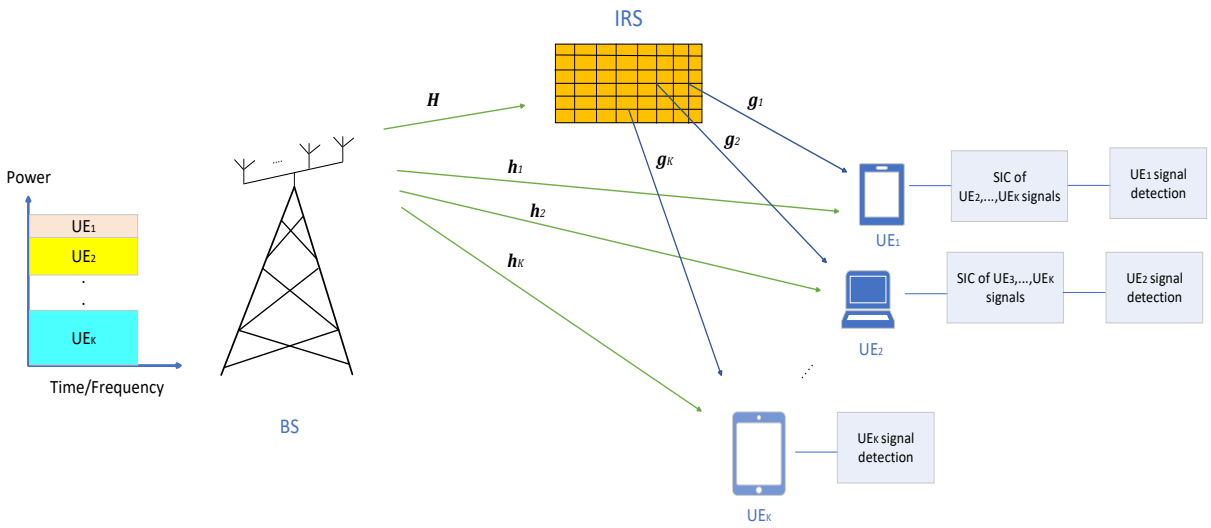

Figure 4.1: IRS-assisted Downlink MISO-NOMA system.

4.1 System Model and Problem Formulation

A DL transmission of an IRS-assisted MISO-NOMA system is considered, in which a BS equipped with N transmit antennas serves K single antenna UEs. The IRS consists of M reflecting elements. Furthermore, the effects of inter-cell interference are assumed to be either absent or accounted for in the noise at the receiver end. Such a system model setup can be utilized for various wireless communication systems in future wireless networks [119–121]. As shown in Figure 4.1, the BS establishes communications with UEs through a direct link and an indirect link through the IRS. In this NOMA system, the transmitted signal from the BS can be written as

$$\mathbf{x} = \sum_{i=1}^{K} \mathbf{w}_i s_i, \forall i \in \mathcal{K}, \tag{4.1}$$

where s_i is the information bearing symbol for UE$_i$, $\mathbf{w}_i \in \mathbb{C}^{N\text{x}1}$ is the beamforming vector designed for UE$_i$, and $\mathcal{K} = \{1, ..., K\}$ is the set of all active UEs in the system. The power of the symbol is assumed to be 1, i.e., $E\{s_i s_i^*\} = 1$. Assuming flat fading channel conditions, the received signal at UE$_i$ can be represented as

$$y_i = \mathbf{h}_i^{\text{H}} \mathbf{x} + \mathbf{g}_i^{\text{H}} \Upsilon \mathbf{H} \mathbf{x} + z_i, \forall i \in \mathcal{K}, \tag{4.2}$$

where $\mathbf{h}_i \in \mathbb{C}^{N\text{x}1}$ is the direct link channel vector between the BS and the UE$_i$. $\mathbf{g}_i \in \mathbb{C}^{M\text{x}1}$ represents the channel between the IRS and UE$_i$ and $\Upsilon = \text{diag}(v_1, ..., v_M) \in \mathbb{C}^{M\text{x}M}$ is a diagonal matrix that represents the phase shifts of IRS elements. The phase shift of each IRS element is modelled by $v_m = \zeta_m e^{j\theta_m}$, $m \in \mathcal{M}$, where $\mathcal{M}$ is the set of all IRS elements, $\zeta_m \in [0,1]$ and $\theta_m \in [0, 2\pi]$, represent the amplitude and the phase shift of the m-th IRS element, respectively. Furthermore, an ideal reflection is assumed with no energy losses by considering only the first-order reflection, i.e., $|v_m|^2 = 1, \forall m \in M$. The phase shift values are determined at the BS and then communicated to the IRS through a feedback link [122]. $\mathbf{H} \in \mathbb{C}^{M\text{x}N}$ is the channel matrix between the BS and the IRS. Note that it is assumed that the IRS is located on a fixed base (on top of a building for example) and therefore, the distance between the BS and IRS is a constant. It is further assumed that there exist LoS paths from the BS to the IRS, as well as from the IRS to the K UEs [123]. The z_i is the noise experienced by UE$_i$ and is modelled as an AWGN with zeros mean and variance σ_i^2. The received signal in (4.2) can be written in a more compact form as follows:

$$y_i = \left(\mathbf{h}_i^{\text{H}} + \mathbf{v}^{\text{H}}\mathbf{Q}_i\right)\mathbf{x} + z_i, \forall i \in \mathcal{K}, \tag{4.3}$$

$$y_i = \tilde{\mathbf{h}}_i\mathbf{x} + z_i, \forall i \in \mathcal{K}, \tag{4.4}$$

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

Since NOMA utilizes SIC at the receiver end in the DL [68] [69], determining an adequate decoding order is crucial in order to unlock the full potential benefits of NOMA. Channel strength is usually used as the criterion for deciding a decoding order that is optimal in the single antenna case, which is not the case for the multiple-antenna NOMA systems [68] [72]. Nevertheless, the channel strength-based decoding order is adopted here, as optimal decoding order design is beyond the scope of this work. According to channel strength-based decoding order, the UE with the strongest channel (referred to as the strongest UE), will be able to successively decode and subtract other UEs' signals, then proceed to decode its own signal. The UE with the weakest channel (referred to as the weakest UE), will directly decode its signal while considering interference from other UEs' signals as noise. To further clarify this decod-

ing order, suppose that there are K users in the system and their estimated channels at the BS are $||\hat{\tilde{\mathbf{h}}}_1||_2^2 \geq ||\hat{\tilde{\mathbf{h}}}_2||_2^2 \geq ... \geq ||\hat{\tilde{\mathbf{h}}}_K||_2^2$, where $\hat{\tilde{\mathbf{h}}}_i$ is the estimated version of $\tilde{\mathbf{h}}_i$ at the BS; then, the decoding order set is $\varsigma = \{1, 2, ..., K\}$ where UE_1 decodes $\text{UE}_2, ..., \text{UE}_K$ signals before decoding its own, UE_2 decodes $\text{UE}_3, ..., \text{UE}_K$ signals before decoding its own signal while treating UE_1's signal as noise, and so on. The weakest user, UE_K, will not carry out any SIC operations and will directly decode its own signal while treating interference from other UEs as noise [68] [69] [72].

4.1.1 Channel Uncertainty Model

Channel uncertainties are inevitable in wireless communications due to channel estimation and quantization errors. These two main sources of imperfect CSIT are, in fact, modelled differently. Channel estimation errors are unbounded and normally expressed using statistical models [124]. The error vectors from this type of error form a normal distribution with a known mean and covariance matrix. Quantization errors, on the other hand, originate from imperfect CSI reporting from the receiver side. A good example where quantization errors are encountered is in FDD systems where the receiver uses a rate-limited feedback channel to report its channel information after quantization. However, given the constrained resolution quantizers used in UEs, additional errors are introduced in the estimated signal during quantization. The quantized channel coefficients transmitted by the UE in the UL feedback link are affected by some quantization errors. Assuming the UE is using a uniform quantizer, the quantization errors can be modelled using a bounded error model [125–129]. In this work, the aim is to study the effects of imperfect CSIT due to quantization errors on the beamforming design at the BS, and consequently, on the achievable system sum rate. In particular, a worst-case beamforming design approach is developed that guarantees the minimum rates requested by the UEs for any value of errors within the bounded region. Furthermore, since there are two links from the BS to the UEs, namely, a direct link and a reflected link through the IRS elements, the following two error models are considered:

1. *Partial error model*: In this error model, it is assumed that the direct link between the BS

and UE$_i, \forall i$, has negligible quantization error effects, while the reflected link is plagued by quantization errors. This scenario is motivated by the fact that the reflected channel is more challenging to obtain than the direct channel due to the passive elements of the IRS [129] [130]. The true reflected channel $\mathbf{Q}_i$, can be modelled as

$$\mathbf{Q}_i = \hat{\mathbf{Q}}_i + \Delta\mathbf{Q}_i, \forall i \in \mathcal{K}, \tag{4.5}$$

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

2. *Full error model*: In this model, a full uncertainty scenario is considered where both the direct and the reflected links are plagued by quantization errors. The true reflected channel expression is the same as in (4.5), while the true direct channel can be expressed as [61] [129]

$$\mathbf{h}_i = \hat{\mathbf{h}}_i + \Delta\mathbf{h}_i, \forall i \in \mathcal{K}, \tag{4.6}$$

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

The unknown errors are norm-bounded such that $||\Delta\mathbf{Q}_i||_F \leq e_{i,r}$, $||\Delta\mathbf{h}_i||_2 \leq e_{i,d}$, for the reflected and the direct channels, respectively. The error bounds $e_{i,r}$, $e_{i,d}$ of UE$_i$ are known at the BS and expressed as [129]

$$e_{i,r} = \sqrt{\frac{\beta_{i,r}^2 \Gamma_{2MN}^{-1}}{2}}, \forall i \in \mathcal{K}, \tag{4.7}$$

$$e_{i,d} = \sqrt{\frac{\beta_{i,d}^2 \Gamma_{2N}^{-1}}{2}}, \forall i \in \mathcal{K}, \tag{4.8}$$

where $\beta_{i,r}^2 = \lambda_r^2 ||\mathbf{q}_i||_2^2$, $\mathbf{q}_i = \text{vec}(\hat{\mathbf{Q}}_i) \in \mathbb{C}^{MN\text{x}1}$ and $\beta_{i,d}^2 = \lambda_d^2 ||\hat{\mathbf{h}}_i||_2^2$ are the variances of ΔQ_i and Δh_i, respectively. $\lambda_r, \lambda_d \in (0,1]$ are scalars that indicate the relative value of the error boundaries. $\Gamma_{2MN}^{-1}, \Gamma_{2N}^{-1}$ are the inverse of the CDF for the Chi-square distribution with $2MN, 2N$ degrees of freedom for the reflected and the direct links, respectively. As seen from (6.4), the error boundary of the reflected channel $e_{i,r}$ is a function of the number of transmit antennas N, the number of IRS elements M, and the quality of the reflected CSI feedback represented by λ_r. According to (4.8), the error boundary of the direct channel $e_{i,d}$ is only related to the number

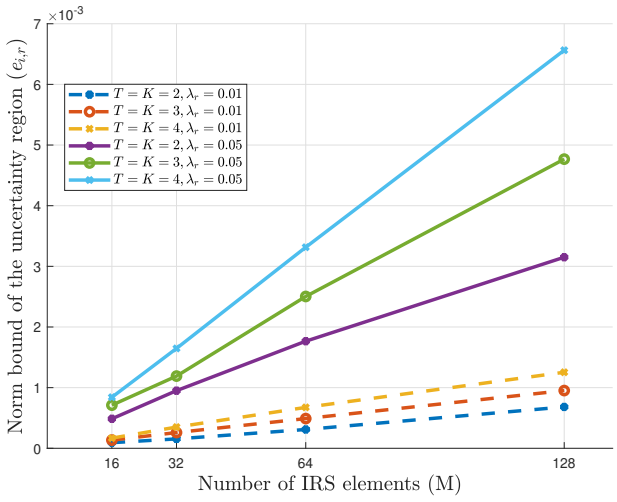

Figure 4.2: Norm bound of uncertainty region versus the number of IRS elements for different system parameters.

of transmit antennas N and λ_d. Figure 4.2 illustrates how different system parameters of (6.4) have an impact on the error bounds of the uncertainty region.

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

4.1.2 SINR and Achievable Rate Expressions

Taking into account the error model and the decoding order discussed in the previous subsections, the SINR expressions are now considered. Without loss of generality, the SINR of UE$_i$'s signal at UE$_j$ is expressed as [68]

$$\gamma_i^j = \frac{|\tilde{\mathbf{h}}_j^{\mathrm{H}} \mathbf{w}_i|^2}{\sum_{j=1}^{i-1} |\tilde{\mathbf{h}}_j^{\mathrm{H}} \mathbf{w}_j|^2 + \sigma_j^2}, \forall j \in \mathcal{IN}_i, \tag{4.9}$$

where $\mathcal{IN}_i$ is the set of interfering users with higher decoding order ranks than UE$_i$ according to their channel strengths. Therefore, the received SINR of UE$_i$ when decoding its own signal

can be expressed as [69]

$$\gamma_i^j = \frac{|\tilde{\mathbf{h}}_i^{\mathrm{H}}\mathbf{w}_i|^2}{\sum_{j=1}^{i-1}|\tilde{\mathbf{h}}_i^{\mathrm{H}}\mathbf{w}_j|^2+\sigma_i^2}, \forall j \in \mathcal{IN}_i. \tag{4.10}$$

To guarantee the smoothness of the SIC operation at stronger UEs, UE_i's SINR is [68]

$$\gamma_i = \min\left(\gamma_i^j, ..., \gamma_i^j\right), \forall j \in \mathcal{IN}_i. \tag{4.11}$$

As a result, the achievable rate at UE_i can be written as

$$R_i = log_2\left(1+\gamma_i\right), \forall i \in \mathcal{K}. \tag{4.12}$$

Note that despite the beamforming vectors and the phase shifts of the IRS elements being designed at the BS based on the estimated channel $\hat{\tilde{\mathbf{h}}}_i$, the SINR expressions in (4.9) and (4.10) are evaluated using the true channel $\tilde{\mathbf{h}}_i$, which contains the unknown norm-bounded error elements [61, 129]. Hence, the considered robust beamforming design is more challenging to the BS in this case due to the unknown errors. The next subsections discuss the robust design problem in detail.

4.1.3 Implications of Error Model on NOMA Systems

In the previous section, the bounded error model considered in this work is explained. However, it is worthwhile to explain the implications of using bounded and unbounded error models on the SINR expressions. In the case of a bounded error model, the CSIT imperfection is caused by the quantization errors in the UL CSI report transmitted by the UE, not channel estimation errors. The quantization error region can therefore be approximated by a ball [125] [131]. Channel estimation error, on the other hand, is modelled statistically where the error vector is drawn from a complex Gaussian distribution with a known mean vector and covariance matrix [124] [129]. Therefore, considering a channel estimation error model leads to taking into consideration imperfect SIC as well, since there is going to be an SIC residual when the stronger UE is trying to decode the weaker UE's signal. Hence, the assumption of a bounded error model because of channel uncertainty is inconsistent for NOMA systems, as channel estimation and

SIC errors are described using an unbounded error model [132]. In this work, however, the focus is on imperfect CSIT due to quantization errors. Therefore, the assumptions of CSIR and ideal SIC do not conflict with the adopted channel uncertainty model.

4.1.4 Problem Formulation

In this work, a robust design to maximize the long-term sum rate of an IRS-assisted MISO-NOMA system under minimum QoS requirements is considered. This robust design is developed based on the worst-case performance approach. In other words, the robust design should meet the required QoS regardless of the experienced channel uncertainties. The beamforming matrix is defined as $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K]$, where $\mathbf{W} \in \mathbb{C}^{N\text{x}K}$, which contains the beamforming vectors of all UEs. The original long-term robust design can be formulated as the following optimization problem:

$$\underset{\Upsilon, \mathbf{W}}{\text{maximize}} \qquad \mathbb{E}\Big\{ \sum_{t=1}^{\infty} \sum_{i=1}^{K} \delta^{t-1} R_i^t \,\Big|\, \pi_t, s_t \Big\} \tag{4.13a}$$

subject to

$$\frac{\Big|\big(\hat{\mathbf{h}}_j^{\text{H}} + \Delta\mathbf{h}_j^{\text{H}} + \mathbf{v}^{\text{H}}(\hat{\mathbf{Q}}_j + \Delta\mathbf{Q}_j)\big)\mathbf{w}_i\Big|^2}{\sum_{j=1}^{i-1}\Big|\big(\hat{\mathbf{h}}_j^{\text{H}} + \Delta\mathbf{h}_j^{\text{H}} + \mathbf{v}^{\text{H}}(\hat{\mathbf{Q}}_j + \Delta\mathbf{Q}_j)\big)\mathbf{w}_j\Big|^2 + \sigma_j^2} \geq 2^{R_i^{min}} - 1, \forall ||\Delta\mathbf{U}_i||_l \leq e_{i,k}, \forall i \in \mathcal{K}, \tag{4.13b}$$

$$\sum_{i=1}^{K} ||\mathbf{w}_i||_2^2 \leq P_{max}, \tag{4.13c}$$

$$|v_m|^2 = 1, \ \forall m \in \mathcal{M}, \tag{4.13d}$$

$$0 \leq \theta_m \leq 2\pi, \ \forall m \in \mathcal{M}. \tag{4.13e}$$

where $\mathbb{E}\Big\{ \sum_{t=1}^{\infty} \sum_{i=1}^{K} \delta^{t-1} R_i^t \,\Big|\, \pi_t, s_t \Big\}$ denotes the expected value of long-term system sum rate, given the policy and the state of the agent, and δ is the discount factor. These entities are explained in the next section. The constraint in (4.13b) ensures the successful implementation of SIC and that the required minimum QoS at UE_i is achieved regardless of the channel uncertainties, where $\mathbf{U}_i \in \{\mathbf{Q}_i, \mathbf{h}_i\}$, $l \in \{F, 2\}$ and $k \in \{r, d\}$ [133]. The constraint in (4.13c) takes into account the available maximum transmit power at the BS, while constraints (4.13d) and

(4.13e) is related to the IRS elements to guarantee ideal reflection and appropriate phase shifts, respectively.

The above optimization problem is non-convex in terms of the beamforming vectors $\mathbf{W}$ and phase shifts Υ. In addition, it is an NP-hard problem in general due to the coupled optimization variables in (6.12a) and (4.13b). Note that the problem is still non-convex even in the absence of (4.13d) and (4.13e) as highlighted by [69]. Therefore, solving this problem using a convex optimization approach will require transforming the problem into convex form using different approximation methods and obtaining solutions based on iterative algorithms. Such iterative algorithms are highly complex in general. In particular, the algorithm should be executed for each new set of channels. In other words, the optimization problem needs to be solved for each new set of channels. To further demonstrate the complexity of the optimization problem in (6.12a), the work in [85] which solved the WSR problem by proposing a centralized solution based on SDP for optimizing the IRS phase shifts, and using the MRT for beamforming design. However, the existing work does not consider the power allocation problem in MRT, which is non-trivial and challenging to optimize optimally [13,134]. The same work proposed an iterative algorithm in an alternating manner to optimize the IRS phase shifts and the beamforming vectors. The work in [135] proposed a distributed solution based on fractional programming and the ADMM algorithm to iteratively solve the WSR optimization problem. However, both the centralized methods which utilize the SDP and the iterative methods are still expensive in terms of latency and computational complexity, especially when the number of inputs is high. It is also worth mentioning that such algorithms are hand-crafted for OMA, and not for NOMA systems. It is well-known that NOMA introduces additional constraints to the optimization problem to ensure the smoothness of the SIC operation at the receivers which is an essential part of the NOMA principle [69]. Therefore, the aforementioned conventional optimization approaches cannot be applied directly to the problem considered in this work.

To address these issues with iterative solution approaches, a DRL-based robust design is proposed. Since RL agents are designed to optimize a long-term objective in a given environment, the problem is reformulated as an RL environment and an RL-based algorithm where the agent

solves the challenging optimization problem. In particular, an approach to solve this robust design is developed using the TD3 agent, which is an enhanced version of DDPG. There are mainly three main motivations for considering this DRL-based approach. First, using a DRL-based algorithm allows for solving the original problem, not an approximated version of it, which means that any feasible solution is guaranteed to solve the problem with no additional assumptions or conditions. This holds for both fixed and varying channels. The second relates to the computational complexity of trained DRL models. As shown in the next section, the time complexity of obtaining a feasible solution from the trained network is almost trivial, which makes it more attractive to latency-sensitive applications. Thirdly, unlike the DQN agent, the TD3 agent is capable of handling continuous action spaces which is required in solving this problem due to the length of the optimization variables vector. Finally, the fact that TD3 converges to a deterministic policy which is also the case for DDPG. However, TD3 is more stable and robust against policy-breaking issues found in the baseline DDPG as explained in the next section.

4.2 Problem Reformulation As An RL Environment

In this section, the basic concepts of RL are briefly summarized focusing on the TD3 agent. Then, the original optimization problem in (6.12a)-(4.13e) is reformulated as an appropriate RL environment to efficiently solve by a TD3 agent.

4.2.1 RL and DRL

Tabular RL methods like Q-learning and SARSA are limited to solving problems with discrete action and state spaces [136]. DRL methods, on the other hand, utilize the function approximation capabilities of DNN, which makes them applicable to a wider variety of problems. DRL methods can be classified primarily into three categories; value-based methods, such as DQN [111] which can handle continuous state space but only support discrete action space. Policy-based methods such as the Reinforce algorithm [112] which optimize the policy directly

through an actor network. Actor-critic methods such as DDPG and TD3 [115] [137], are recent off-policy agents that train deterministic policies. The actor takes actions and optimizes the policy of the agent while the critic evaluates the action taken by the actor with regards to the current state and returns a Q-value. Through these interactions, actor-critic agents optimize the policy of the agent until it converges to an optimal or near-optimal policy. Furthermore, actor-critic agents can handle continuous action and state spaces which widens their applicability to a larger set of problems in wireless communications. Note that any actor-critic agent with continuous actions and state spaces can be applied to solve the robust design problem using the reformulation provided. However, the TD3 agent is utilized because it is an off-policy agent with higher sample efficiency due to the use of a replay buffer which allows for reusing past experiences. Furthermore, the TD3 agent optimizes a deterministic policy which is generally easier to implement compared to stochastic policies.

4.2.2 Brief Overview of TD3

TD3 is an off-policy actor-critic DRL agent that is capable of handling continuous action and state spaces. A TD3 agent consists of two main parts, an actor and a critic. The actor is a DNN responsible for generating actions. It takes in the current state as input and generates an action based on its current policy. The critic's DNN is responsible for generating the corresponding Q-value for the action taken by the actor. As a result, the critic's DNN has two inputs, the current state and the current action taken by the actor. Note that training in the context of RL is not the same as in deep learning. In the case of RL, the agent learns in an online fashion, which has two important implications; training-data generation and learning is carried out simultaneously, and training targets are constantly changing according to the agent's current policy. In order to stabilise learning, both the actor and the critic use a delayed copy of their current DNNs called target networks. Target networks stabilise learning by fixing the target value when optimizing actor and critic DNNs. Experience replay buffer is utilized by the majority of off-policy DRL agents and TD3 is no exception [138]. Previous interactions with the environment defined as tuples of $\{s, a, r, s^{t+1}\}$, are saved in the replay buffer $\mathcal{D}$. The buffer is then sampled to obtain

training data. Replay buffer with larger memory makes data more IID, which reduces the DNN variance during training. The critic of the DDPG agent can be considered as a modified DQN that takes in the action performed by the actor and outputs a scalar Q-value. To mitigate the problem of overestimating the Q-value in DDPG, TD3 uses two (or more) critics and selects the smallest estimate of the target Q-value. Given that the next state s^{t+1} is not the terminal state, the target can be expressed as [137]

$$y'(r, s^{t+1}) = r + \delta \min_{i=1,2} \mathcal{Q}_{\phi_i'}(s^{t+1}, \mu_\psi'(s^{t+1})), \tag{4.14}$$

where $\mathcal{Q}_{\phi_i'}$ is the target network for the critic's DNN $\phi_i', i = 1, 2$, δ is the discount factor (current value) for future rewards, and μ_ψ' is the actor's target network which provides the next action a^{t+1} given a next state s^{t+1}. Then, the two critics learn the Q-function by minimizing their respective objectives as follows [137]:

$$\begin{aligned} L(\phi_1, \mathcal{D}) &= \mathop{\mathbb{E}}_{(a,s,r,s^{t+1})\sim\mathcal{D}} \left[\left(\mathcal{Q}_{\phi_1}(s,a) - y(r, s^{t+1}) \right)^2 \right], \\ L(\phi_2, \mathcal{D}) &= \mathop{\mathbb{E}}_{(a,s,r,s^{t+1})\sim\mathcal{D}} \left[\left(\mathcal{Q}_{\phi_2}(s,a) - y(r, s^{t+1}) \right)^2 \right]. \end{aligned} \tag{4.15}$$

The actor in TD3 aims to optimize the policy. This is achieved by adjusting the weights of its DNN μ_ψ to maximize the corresponding Q-value, which is defined by optimizing the following objective [115]:

$$\max_{\psi} \mathop{\mathbb{E}}_{s\sim\mathcal{D}} \left[\mathcal{Q}_{\phi_1}\left(s, \mu_\psi(s)\right) \right], \tag{4.16}$$

which is identical to the DDPG actor. Unlike DDPG, TD3 updates its policy using (4.16) less frequently than its Q-values to reduce variance during the training. Hence, the policy update in (4.16) is not executed in each training step. When it does, the policy, however, gets updated by (4.16). The target networks for both the critics and the actor are updated at a much slower rate than their main counterparts using

$$\begin{aligned} \phi_i' &= \kappa\phi_i + (1-\kappa)\phi_i', \, i = 1, 2, \\ \psi' &= \kappa\psi + (1-\kappa)\psi', \end{aligned} \tag{4.17}$$

where $0 < \kappa \leq 1$ is the target network smoothing factor. Algorithm 1 summarizes the key steps of how the TD3's actor and critics process one experience. Note that in practice, these steps are carried out in batches instead of single experiences to increase computational efficiency. Overall, TD3 theoretically outperforms DDPG by utilizing double Q-learning to reduce over-

Algorithm 1 TD3 actor and critic training

1: A tuple $\{s, a, r, s^{t+1}\}$ is randomly sampled from the replay buffer $\mathcal{D}$
2: The current state s is fed to actor's DNN μ_ψ to generate current action a
3: Both s and a are fed to the critics' DNNs to generate $\mathcal{Q}_{\phi_1}(s,a)$ and $\mathcal{Q}_{\phi_2}(s,a)$
4: The next state s^{t+1} is fed to the actor's target DNN μ'_ψ to generate the next action a^{t+1}
5: The critics' target DNNs $\mathcal{Q}_{\phi'_i}(s,a), i = 1,2$, are fed with s^{t+1} and a^{t+1} to calculate the target using (4.14)
6: The critics are trained using (4.15)
7: **if** Time to update policy **then**
8: The actor is trained using (4.16)
9: **end if**
10: Target networks are updated using (4.17)

estimation effects and updating its policy less frequently to reduce variance. Furthermore, it employs target policy smoothing by adding noise to actor actions, and target actions as well to prevent the agent from exploiting errors in Q-value estimations [137]. Figure 4.3 shows the interactions between the internal components of the agent interact with each other to produce an optimal or near-optimal policy that maps states to the best possible actions. Despite that these upgrades may seem simple, combined together with hyperparameter tuning, they are the driving factor for any additional gain of TD3 over DDPG. Simulation results presented in section IV confirm the additional gain of TD3.

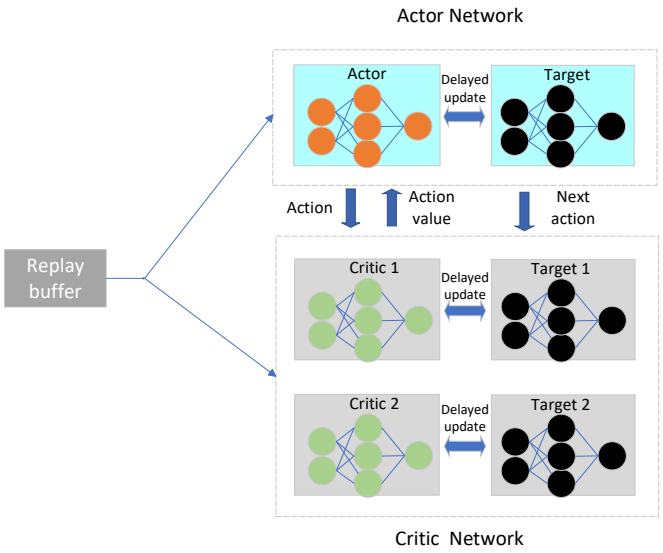

Figure 4.3: TD3 agent blocks.

4.2.3 Robust Design Problem As TD3 Environment

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

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

$$\mathbf{a}^t = \left[\mathbf{w}_1^t, \ldots, \mathbf{w}_K^t, \vec{v}_1^t, \ldots, \vec{v}_M^t\right]^N. \tag{4.18}$$

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

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

vector for the TD3 agent can be expressed as

$$\begin{aligned}\mathbf{s}^t = \Big[&||\mathbf{w}_1^{t-1}||_2^2, .., ||\mathbf{w}_K^{t-1}||_2^2, e_1, .., e_K, \\ &R_1^{t-1}, R_2^{1,t-1}, .., R_K^{K-1,t-1}, R_K^{K,t-1}, \mathbf{a}^{t-1}\Big]^N,\end{aligned} \tag{4.19}$$

where the error values in the state vector are directly mapped to the reflected error bound in the case of the partial error model, while the error bounds correspond to the sum of the direct and reflected error bounds in the case of the full uncertainty error model. Therefore, $\mathbf{s}^t \in \mathbb{C}^{2K+\frac{K(K+1)}{2}+KN+M}, K \geq 2$, where $\frac{K(K+1)}{2}$ determines the number of all possible rates in the considered MISO-NOMA system.

Note that both beamforming vectors and phase shifts are complex-valued design parameters and they are part of the action and state spaces. However, since real-valued neural networks are used for building the DRL agent, each complex vector is mapped to two separate real vectors where one represents the real values while the other represents the imaginary values of the original complex-valued vector [139] [134]. Therefore, the beamforming vector (or any complex vector for that matter) $\mathbf{w}_i \in \mathbb{C}^{N\text{x}1}$ is mapped to $\mathbf{Re}(\mathbf{w}_i) \in \mathbb{R}^{N\text{x}1}$ representing the real part of $\mathbf{w}_i$, and $\mathbf{Im}(\mathbf{w}_i) \in \mathbb{R}^{N\text{x}1}$ representing the imaginary part of $\mathbf{w}_i$. This is also true for the complex value phase shifts of the IRS elements, where each scalar complex phase shift value is mapped to two real scalars representing the real and complex parts of the original element. Note that this technique basically doubles the size of input and output layers for the critic and the actor DNNs. However, it unlocks the potential for using neural networks to deal with a wider range of problems such as the one considered in this work. To reconstruct the complex-valued beamformers and IRS phase shift elements obtained from the action vector, the mapping process explained earlier is reversed. Therefore, the $\mathbf{a}^t \in \mathbb{R}^{2KN+2M}$, $\mathbf{s}^t \in \mathbb{R}^{2K+\frac{K(K+1)}{2}+2KN+2M}$ are corresponding real-only action and state space vectors, respectively.

- Finally, as the objective is to maximize the long-term sum rate of the system, the sum rate at time-step t is chosen as the reward. Thus, the reward can be expressed as

$$r^t = \sum_{i=1}^{K} R_i^t, \forall i \in \mathcal{K}. \tag{4.20}$$

It is important to highlight that the agent will only be rewarded the sum rate of the step if its action satisfies all constraints of the original optimization problem. However, since RL agents are only interested in maximizing their rewards, they cannot solve convex optimization problems directly. For this reason, the agent is forced to meet the constraints by normalizing its actions to fall within the feasible region. First, the maximum transmit power constraint is considered. Since the objective is an increasing function of the transmit power, at the optimal conditions, the transmitter will use all the available transmit power (i.e., $P_{\max}$). Therefore, the transmit power constraint (4.13c) is rewritten as follows:

$$\sum_{i=1}^{K} ||\mathbf{w}_i||_2^2 = P_{\max}, \forall i \in \mathcal{K}. \tag{4.21}$$

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

$$P_{\text{total}}^t = \sum_{i=1}^{K} ||\mathbf{w}_i^t||_2^2, \forall i \in \mathcal{K}. \tag{4.22}$$

Then, the normalization coefficient can be expressed as

$$\bar{\iota}^t = \sqrt{\frac{P_{\max}}{P_{\text{total}}^t}}. \tag{4.23}$$

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

$$\mathbf{f}_i^t = \bar{\iota}^t \mathbf{w}_i^t, \forall i \in \mathcal{K}. \tag{4.24}$$

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

$$v_m = \frac{\bar{v}_m^t}{|\bar{v}_m^t|}, \forall m \in \mathcal{M}. \tag{4.25}$$

With the normalized action, the agent is either rewarded the sum rate in (4.20) if the QoS requirements are satisfied under the channel uncertainty, otherwise, the agent is punished with a negative reward. Any negative reward will work as the agent will try to avoid such action in the future. The sum of the rate deficit across all users is used as the negative reward [99]. The

set ε' contains users $j = 1,..,J, \varepsilon' \in \mathcal{K}$ whose QoS are not satisfied at time-step t. Thus, the sum of the rate deficit across all users is defined as

$$r_d^t = \sum_{j=1}^{J} \left(R_j^t - R_j^{min}\right), \ \forall j \in \varepsilon'. \tag{4.26}$$

Therefore, if $\mathbf{a}^t$ satisfies the QoS constraints under some bounded error region, the agent will be given a positive reward according to (4.20), otherwise, it will be punished with the negative reward in (4.26). Algorithm 1 summarizes the proposed TD3-based algorithm for solving the original robust design problem. Note that Algorithm 2 summarizes the training process for the proposed agent. However, once the agent has been trained successfully, the actor network is the one deployed in practice. The trained actor network can then be integrated into the BS hardware to be used to generate the solutions. To implement the proposed solution, in a practical IRS-assisted MISO-NOMA system, the BS receives the CSI reports in the UL band. The BS then queries the trained actor network by using the obtained channels, i.e., executing steps $7-11$. The resulting IRS vector is transmitted to the IRS via a feedback link, while the beamforming vectors are used for transmission.

4.2.4 Computational Complexity Analysis

In this subsection, the computational complexity of the proposed TD3-based algorithm is defined. Similar to other deep learning models, the complexity of the proposed DRL framework can be divided into two categories: offline complexity, which is associated with training the actor network by plugging in critics and the replay buffer, and online complexity which is associated with inference or deployment of the actor's network. Calculating the best and the worst case run times for offline training of neural networks is still an open issue due to the complexity associated with the implementation of backpropagation and other hyperparameters in DNNs [134] [140]. Furthermore, it is assumed that the offline complexity of this model can be afforded. Nevertheless, empirical comparisons for four different profiles with different hardware specifications are included in Table 4.1. The specification of each hardware platform and the system parameters used for each case are provided in Tables 4.2, 4.3.

Algorithm 2 TD3-based Robust Beamforming and Phase Shift Design

Initialize TD3 target and training parameters, empty replay buffer $\mathcal{D}$ and initialize the Gaussian random process $\mathcal{A}$
Set $\phi_1' \leftarrow \phi_1, \phi_2' \leftarrow \phi_2, \psi' \leftarrow \psi$
while Episode $\leq$ Total Episodes **do**
    Acquire training channels based on the system parameters K, M, N
    Calculate $\Delta\mathbf{Q}_i, \forall i$, according to (6.4) for the partial error model, adding $\Delta\mathbf{h}_i, \forall i$, according to (4.8) for the full error model
    Initialize the beamforming vectors and the phase shift elements randomly
    while t $\leq$ Time steps **do**
        Observe the current state s^t and obtain an action from the actor network using $\mathbf{a}^t = \text{clip}(\mu_\psi(s) + \varepsilon, a_{low}, a_{high}), \varepsilon \in \mathcal{A}$, normalize action values using (4.23), (4.24) and (4.25)
        Recover the complex value beamforming vectors and the IRS elements from step 6
        Using vector $\mathbf{v}$ generated in the previous step, build the final estimated channels $\hat{\tilde{\mathbf{h}}}_i, \forall i$, according to (4.3)
        Decide a descending decoding order ς such that $||\hat{\tilde{\mathbf{h}}}_1||_2^2 \geq ||\hat{\tilde{\mathbf{h}}}_2||_2^2 \geq ... \geq ||\hat{\tilde{\mathbf{h}}}_K||_2^2$, based on the estimated channels $\hat{\tilde{\mathbf{h}}}_i, \forall i$
        Build the true channels $\tilde{\mathbf{h}}_i, \forall i$, using vector $\mathbf{v}$ and random errors based on (4.3), (4.5) and (4.6)
        Evaluate the SINR values and calculate the corresponding rates $R_i, \forall i$
        if $R_i \geq R_i^{min}, e_i, \forall i \in \mathcal{K}$ **then**
            Use reward in (4.20)
        else
            Use reward in (4.26)
        end if
        Obtain next state $\mathbf{s}^{t+1}$. Save tuple $\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\}$ to replay buffer $\mathcal{D}$
        Randomly sample replay buffer using a batch of size b to calculate the target according to (4.14) and train the two critic networks ϕ_1, ϕ_2 using (4.15)
        if time to update policy **then**
            Update policy with one step using (4.16)
        end if
        Update target networks using (4.17)
        $t = t + 1$
        Set $\mathbf{s}^t = \mathbf{s}^{t+1}$
    end while
    $Episode = Episode + 1$
end while
Output: $\{\mathbf{f}_1^*, ..., \mathbf{f}_K^*, v_1^*, ..., v_m^*\}$

Table 4.1: Numerical time-complexity.

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

For estimating the time complexity of inference, which is the cost of a feed-forward pass

Table 4.2: Hardware profiles.

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

Table 4.3: System parameters for run time testing.

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

through the trained actor DNN, big $\mathcal{O}$ notation is a common method of measuring the worst-case run time of an algorithm. Since all modern libraries and deep learning frameworks use matrix notation to perform calculations through DNNs, it is straightforward to conclude that a matrix-vector multiplication operation, $\mathbf{z}_l = \Psi \mathbf{c}_l$, where Ψ is the weights matrix, $\mathbf{c}_l$ is the input vector, and $\mathbf{z}_l$ is the output vector from the l-th hidden layer, is performed for each hidden

layer. The output vector $\mathbf{z}$ is then passed through an activation layer as $\mathbf{b}^l = g(\mathbf{z}^l)$, where $\mathbf{b}^l$ is the activated vector that is fed to the next hidden layer in the DNN. Since the activation is an element-wise operation, it has a time complexity of $\mathcal{O}(\aleph_l)$, where $\aleph_l$ is the number of neurons in the l-th hidden layer. According to the proposed actor's architecture shown in Figure 4.4, there are three weight matrices in total, $\Psi_1 \in \mathbb{R}^{\aleph \text{ x} \mathbf{Card}(\mathbf{s}^t)}$, linking the input to the first hidden layer, $\Psi_2 \in \mathbb{R}^{\aleph^2}$, between the two hidden layers, assuming $\aleph_1 = \aleph_2 = \aleph$, and $\Psi_3 \in \mathbb{R}^{\mathbf{Card}(\mathbf{a}^t) \text{x} \aleph}$, linking the second hidden layer to the output layer. Therefore, the total run-time can be expressed as $\mathcal{O}\Big(T\big(\aleph \cdot \mathbf{Card}(\mathbf{s}^t) + \aleph^2 + \mathbf{Card}(\mathbf{a}^t) \cdot \aleph + 2\aleph + \mathbf{Card}(\mathbf{a}^t)\big)\Big)$, where T is added as an implication of using the action space as part of the state space. Also, since the action vector is part of the state vector, then $\mathbf{Card}(\mathbf{s}^t) > \mathbf{Card}(\mathbf{a}^t)$ always holds. Therefore, the worst-case run time for evaluating the actor's DNN can be approximated as $\mathcal{O}\Big(\aleph \cdot max\big(\aleph, \mathbf{Card}(\mathbf{s}^t)\big)\Big)$ for single feed-forward pass. To define the complexity of the proposed DRL algorithm in context, a complexity review for related works in the literature is provided. The worst-case complexity for the iterative algorithm proposed in [69], which only solves the beamforming design problem, is $\mathcal{O}(K^7)$ per iteration. The SDP-based algorithm for optimizing the IRS phase shifts proposed in [85] has a worst-case complexity of $\mathcal{O}(M^6)$, while the iterative algorithm proposed in [135] reduced the IRS phase shifts optimization complexity to $\mathcal{O}(M^3)$ using ADMM. Furthermore, the worst-case run-time for the proposed algorithm scales linearly with the system parameters for a fixed number of neurons, while the worst-case run-time of the model-based algorithms is cubic at best. Therefore, compared to the complexities of the existing methods, the proposed algorithm has a significant advantage in terms of run times, while still maintaining competitive performance.

4.3 Training, Simulation and Numerical Results

In this section, the performance of the proposed TD3-based algorithm is evaluated with different system parameters.

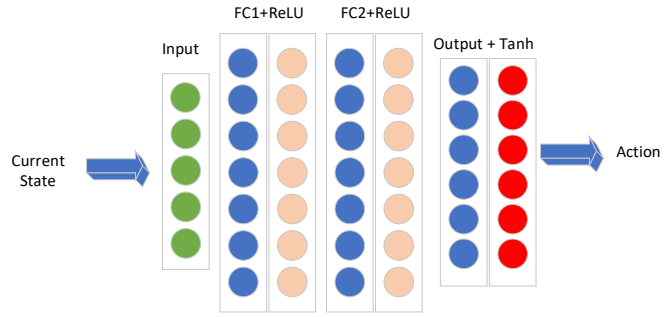

Figure 4.4: TD3 actor DNN.

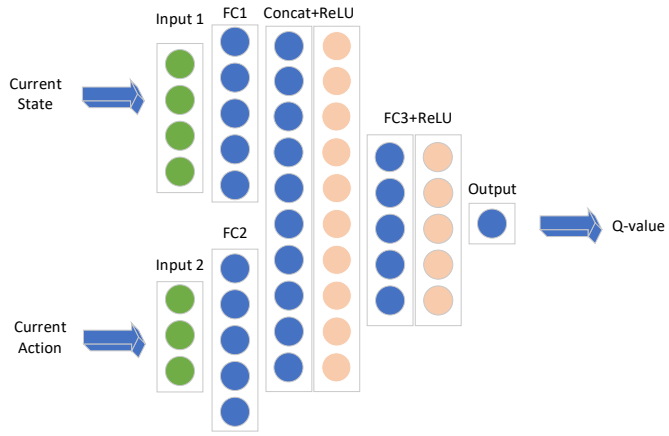

Figure 4.5: TD3 critic DNN.

4.3.1 Agents Structure and Hyperparameters

To evaluate the performance of the proposed robust design, a TD3 agent with one actor and two identical critics is trained. Note that despite the two critics being identical in terms of layer type and size, the random initialization of their respective DNNs makes them behave differently, and therefore, produce different Q-value estimates. The architecture of the actor and critics DNNs are shown in Fig. 4.4 and 4.5, respectively. Table 4.4 describes the structure and size of the actor and critics networks. The number of hidden nodes is set to 300 for each hidden layer, irrespective of the input and output sizes, the *ReLU* activation function, $f(x) = max(0,x)$, is used for activating the hidden layers in both actor and critics' networks. The *Tanh* function, $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, is used as an activation function for the output in the actor's network. The

Table 4.4: Actor and critic layers.

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

ADAM optimizer is used for both actor's and critics' DNNs as it is more robust than other optimizers, and more appropriate for non-stationary objectives [141]. Table 5.1 provides a summary of hyperparameters used to train the agent for both fixed and dynamic channel cases. The reward discount factor is set to 0.99 to steer the agent towards a long-term optimal reward policy. Generally, the hyperparameters chosen for the TD3 agent in this work are more on the conservative side. Such an approach favours training stability over faster convergence, which is recommended for the agent to form a more robust policy against channel uncertainties.

4.3.2 System Parameters

In terms of system parameters, an IRS-assisted, DL MISO-NOMA system is considered where $N = K = 2,3,4$, which is one of the cases where NOMA has the most advantage over OMA [69]. Table 4.6 summarizes the system parameters used in the simulations. Because of the high computational complexity associated with SIC receivers, the maximum number of UEs is limited to $K = 4$ where the strongest UE will perform 3 SIC operations. Increasing the number of UEs requires pairing the UEs into clusters, which is beyond the scope of this work. For the channel model, both small-scale and large-scale fading are taken into account. The large-scale

Table 4.5: Hyperparameters of the TD3 agent.

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

fading is a function of the distance from the BS and the IRS, for the direct and the reflected channels, respectively. The small-scale fading is modelled by Rician and Rayleigh fading for the reflected and direct channels, respectively. The channel coefficients for direct and reflected paths are drawn from a complex Gaussian distribution with zero mean and unit variance. The first part of the reflected channels from the BS to the IRS is modelled as

$$\mathbf{H} = \frac{1}{\sqrt{d_{irs}^{\iota_{b\to irs}}}} \left(\sqrt{\frac{K'}{1+K'}} \mathbf{H}_{LoS} + \sqrt{\frac{1}{1+K'}} \mathbf{H}_{nLoS} \right), \tag{4.27}$$

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

$$\mathbf{g}_i = \frac{1}{\sqrt{d_i^{\iota_{irs\to u}}}} \left(\sqrt{\frac{K'}{1+K'}} \mathbf{g}_{LoS} + \sqrt{\frac{1}{1+K'}} \mathbf{g}_{nLoS} \right), \tag{4.28}$$

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

To fairly assess the performance of the proposed algorithm, the following benchmark algorithms are used

- **DDPG**: The DDPG agent has been widely adopted in the DRL literature. DDPG is included as a DRL benchmark to showcase the performance gain of the proposed TD3-based design in terms of convergence, system sum rate, and robustness.

- **Baseline** 1: This benchmark scheme is based on SDP. More specifically, an SDP is used to solve the IRS optimization subproblem [85], and then the best possible rates are achieved for the given maximum available power through solving the transmit power minimization problem [13, 134]. Note that this scheme has prohibitively high complexity and is therefore used as an analytical benchmark.

- **Baseline** 2: This scheme is based on the well-known ZF principle as a solution to the beamforming design subproblem. However, since the multi-user power allocation problem is non-trivial in the ZF beamforming case, a fixed power allocation strategy is assumed for this scheme. Therefore, this is a non-robust scheme. The IRS optimization subproblem is solved using SDP [85].

- **Baseline** 3: This is a random benchmarking scheme, i.e., the IRS phase shifts and the beamforming vectors are randomly generated. Such a scheme is included to show that the agent has derived a competitive policy that adapts to the environment.

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

Table 4.6: Summary of system parameters.

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

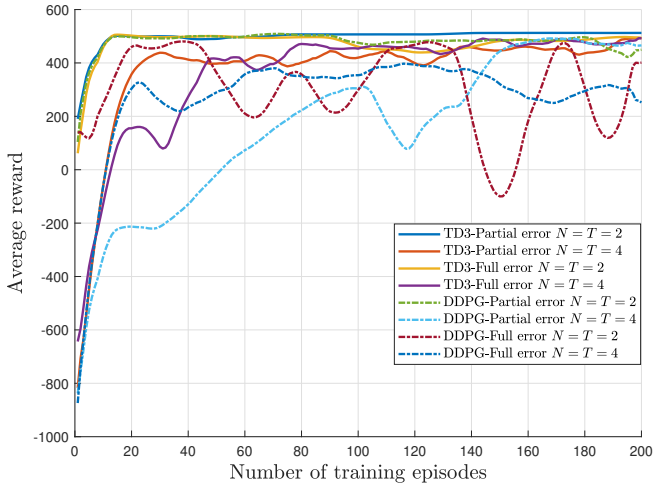

Figure 4.6: The reward of the proposed robust TD3, and DDPG agents for 200 training episodes, with fixed channels, $M = 16$, $R^{min} = 1$ b/s/Hz.

4.3.3 Fixed-Channel Senario

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

Figures 4.6 and 4.7 present the convergence of the agent during training for the two extreme cases of IRS elements, $M = 16$ and $M = 128$, respectively. These convergence plots suggest that both agents can converge faster in the case of $M = 16$, compared to the other case with $M = 128$. This is expected, as M is directly related to the length of the state and the action vectors, and the error bound, making faster convergence in the case of $M = 128$ more challenging for the agents. Note that in both cases, the TD3 agent shows a more stable and consistent behaviour compared to that of the DDPG agent, thanks in part to the additional critic used by TD3. As seen in Figures 4.6 and 4.7, the TD3 agent requires around 40 episodes of training to reach an average reward level of greater than 400 in the first case, while other case requires around 130 episodes to achieve the same reward. The DDPG shows a similar performance in the case $M = 16$. However, Figure 4.7 shows the DDPG requires much higher training episodes

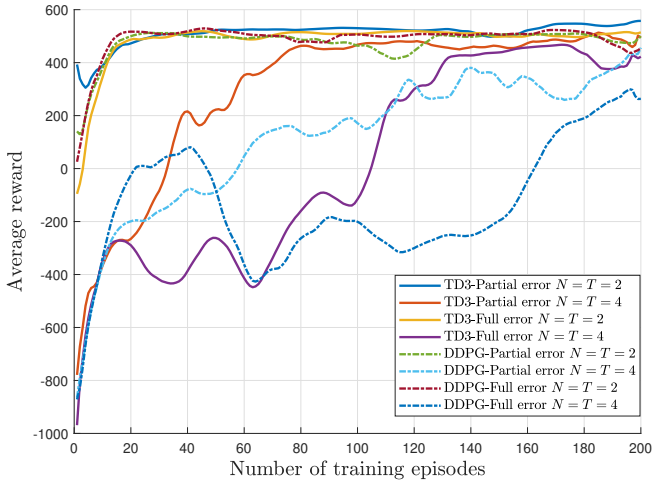

Figure 4.7: The reward of the proposed robust TD3, and DDPG agents for 200 training episodes, 200 time-steps per episode with fixed channels, $M = 128$, $R^{min} = 1$ b/s/Hz.

to determine a high reward policy when $K = 2, 4$. Overall, both agents require more training episodes to achieve convergence in the case of the full error model than in the partial error model. This is expected, as the robust beamforming design with a larger error bound is more challenging than the one with a small error bound. To demonstrate the potential capabilities of the TD3 agent in maximizing system sum rate, Figures 4.8, 4.9 and 4.10 show the performance gains of the proposed TD3 agent. These simulation results are generated by taking the average rates of the agents when they are tested for a total of $1,000$ episodes, with 10 steps per episode. The achievable system sum rates are higher in the partial error case across the three plots. The proposed TD3 agent outperforms the DDPG benchmark and random schemes with variable margins. The most significant TD3 gains over DDPG are achieved in the cases of $K = N = 4, M = 64$ and $K = N = 3, M = 128$, with 3.2 b/s/Hz, 5.4 b/s/Hz, for the partial and full error cases, respectively. This clearly shows that the proposed TD3 agent is able to derive a more accurate and higher rewarding policy than the DDPG agent. Another interesting observation from the achieved system sum rates is that there are different peak rates for different numbers of UEs. In Figure 4.8, where $K = N = 2$, the maximum system sum rate is achieved with $M = 64$, while in the case of $K = N = 3$, the sum rate is achieved with $M = 128$, and in the case

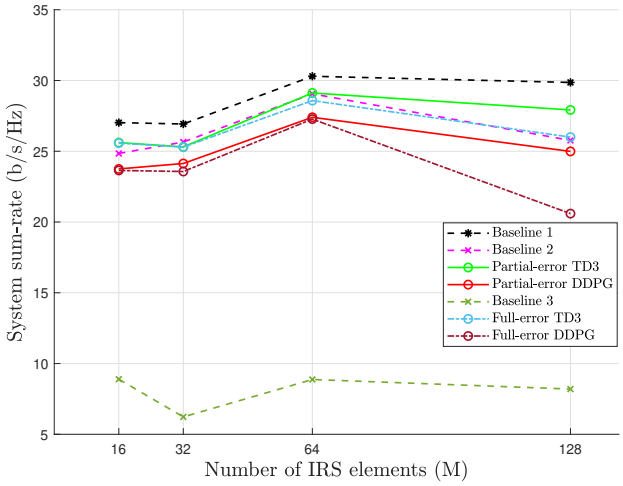

Figure 4.8: The achieved system sum rate of the proposed robust design versus the number of IRS elements for $K = N = 2$, $R^{min} = 1$ b/s/Hz.

$K = N = 4$ it reaches with $M = 32$. This suggests that in each case, there is a sweet spot between having the ideal number of IRS elements to maximize the sum rate and having a manageable error region. It also suggests that, unlike many studies in the literature, increasing the number of IRS elements does not always result in an increased system sum rate. In fact, when considering a robust design, increasing the number of IRS elements beyond a certain number may result in a degraded performance for the fixed channel case. Compared to the benchmark schemes, the TD3 agent generally outperforms the ZF baseline, even when the full error model is used. The performance gap in terms of the achieved system sum rates between the proposed TD3-based design and the upper-bound baseline is marginal at best, with 1.9 b/s/Hz and 2.5 b/s/Hz for the partial and full error models, respectively. In terms of achieved rates of UEs, Figure 4.11 presents UE_1 and UE_4 rates for both error models achieved by both agents, which represent the strongest and the weakest UEs in the system, respectively. The figure shows that UE_1 achieves higher rates when using the TD3 agent's policy. As for UE_4, both agents were able to consistently achieve the target rate required by the weakest UE for both error models. The apparent high variance in UE_1's rate for baseline 2 is caused by channel errors during testing since it is a non-robust scheme. This is also evident by the casual dips in UE_1's rate as shown

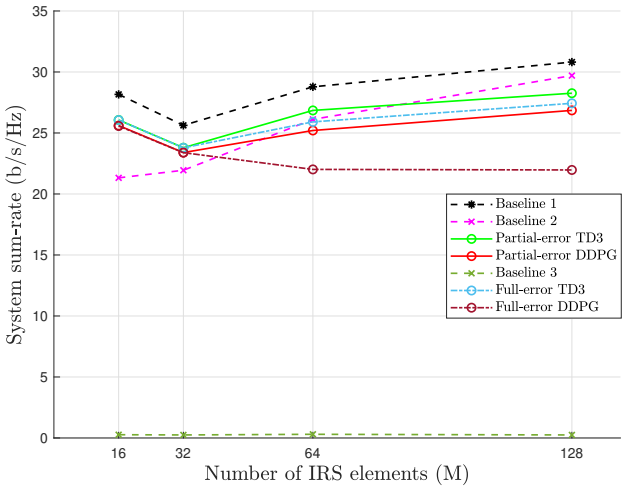

Figure 4.9: The achieved system sum rate of the proposed robust design versus the number of IRS elements for $K = N = 3$, $R^{min} = 1$ b/s/Hz.

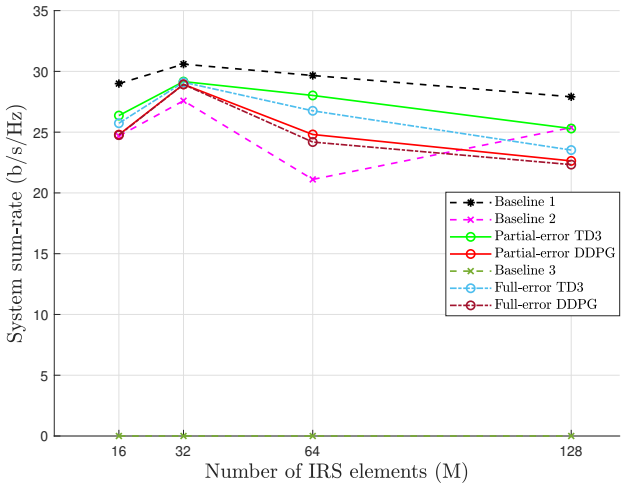

Figure 4.10: The achieved system sum rate of the proposed robust design versus the number of IRS elements for $K = N = 4$, $R^{min} = 1$ b/s/Hz.

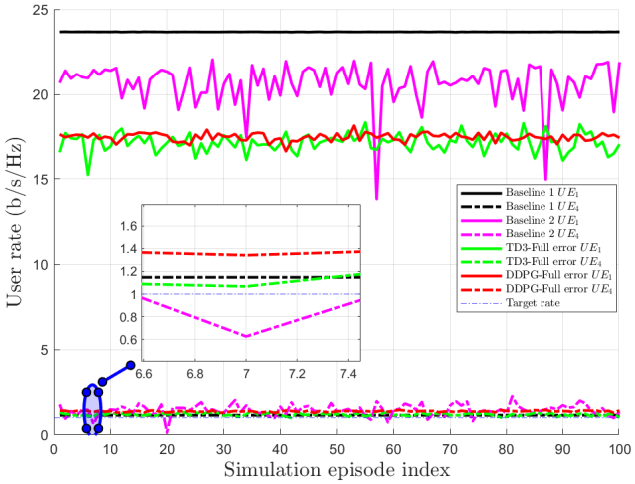

Figure 4.11: The achieved individual user rate of the proposed robust design across 100 testing episodes for $K = N = 4$, $R^{min} = 1$ b/s/Hz.

in the same figure. Furthermore, to rigorously assess the robustness of both agents, Figure 4.12 demonstrates the performance of the agents for different target rates. The figure shows that the TD3 agent is able to achieve a perfect score up to the training target rate, and after. In particular, the TD3 agent with $M = 128$ for the partial error model is able to attain a target rate of 1.5 b/s/Hz with a robustness score of 88%, which is impressive considering it was trained on a lower target rate of 1 b/s/Hz. The performance of the DDPG agent, on the other hand, is degraded in the case of full channel uncertainty, achieving a score of 89% with $M = 16$ as its worst case.

4.3.4 Dynamic-Channel Scenario

In the previous scenario, the channels were assumed to be fixed. While this may be the case for stationary devices or low-mobility UEs, fixed channel models cannot be used for high-mobility situations where channels change drastically. To solve this dynamic channel problem, the TD3 agent is trained on a small dataset of distinctively different channels. Also, the full error model is used for the varying channel case as it focuses more on the practical implementation aspects of this design. Therefore, the TD3 agent is trained for a total of $2,000$ episodes and 300 steps per

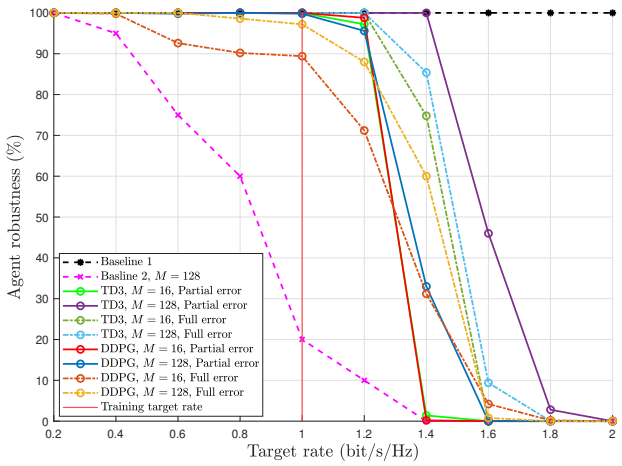

Figure 4.12: The robustness performance of the proposed agent versus the target rate with fixed channels, for $K = N = 4$, $R^{min} = 1$ b/s/Hz.

episode. At the beginning of each episode, a different set of channels randomly sampled from a dataset of 10 channels is selected. These training channels are generated based on a uniform sampling of the distance between the BS and the maximum cell radius. This uniform sampling is chosen to ensure that the training channels reflect the variance of the channels across the entire cell. Corresponding error bounds for direct and reflected links are also fed to the agent for each new episode during training as part of the state vector. Furthermore, to prevent the optimization problem from becoming infeasible due to higher channel variations, the target rate is reduced to 0.3 b/s/Hz for the dynamic channels scenario. To evaluate the performance of the agent in a dynamic-channel environment, a total of 250 randomly generated channels with $d_{id} \in [10, 200]$ m are used as a testing set. Also, the agent is simulated for $1,000$ episodes, with 10 steps per episode for testing, to determine the average achieved sum rates. The convergence of the agent is shown in Figure 4.13 for the two extreme cases $K = N = 2, 4, M = 128$, where relatively higher training variance is apparent. This is expected since the channels are inherently different, and consequently, the reward will also have a higher variance. From Figure 4.13, it can be seen that there is a significant difference in terms of stability and consistency between the TD3 and the DDPG agents, where TD3 shows superior convergence properties. This is

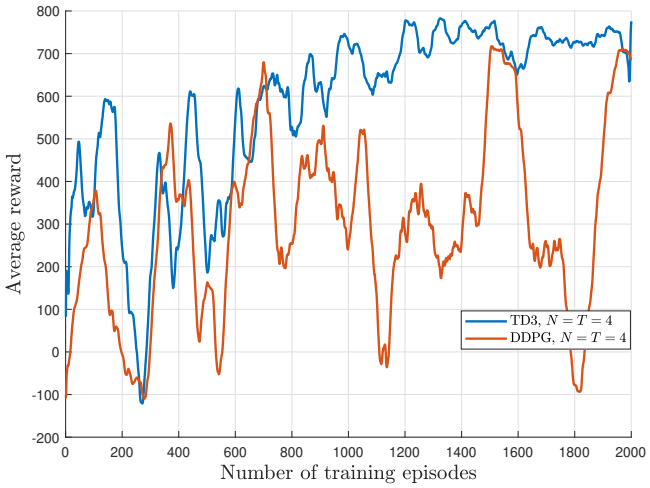

Figure 4.13: The reward of the proposed robust TD3, and DDPG agents for 2,000 training episodes, with dynamic channels, $M = 128$, $R^{min} = 0.3$ b/s/Hz.

further evident by the relatively lower variance of the TD3 agent compared to the higher training variance of DDPG. Instability during training often leads to performance degradation due to the inadequately derived policy. Figures 4.14, 4.15 and 4.16 illustrate the achieved system sum rates for different system parameters. The TD3 agent shows marginal gains compared to the DDPG agent, with the most significant gain being 2.14 b/s/Hz, achieved in the case $K = N = 3, M = 64$. For the dynamic channel case, it can be seen that increasing the number of IRS elements is exploited by both agents, leading to a slight increase in terms of the sum rate. The TD3 agent is able to achieve a gain of 2.1 b/s/Hz in the system sum rate for the case $K = N = 3, M = 64$. However, despite the addition of 64 IRS elements, the system sum rate has not increased as much between $M = 64$ and $M = 128$, which further proves the point that the number of IRS elements may be utilized by the agent up to a certain number before starting to degrade the performance. Compared to the benchmarking schemes, the proposed TD3 agent achieves a similar sum rate performance to the ZF baseline scheme on average, while the sum rate gap between the upper-bound baseline and the proposed agent has increased in the varying channels case with an average gap of 3.3 b/s/Hz. In terms of achieved individual rates, Figure 4.17 illustrates the rate for each UE for the dynamic channels case, with $K = N = 4, M = 128$. This

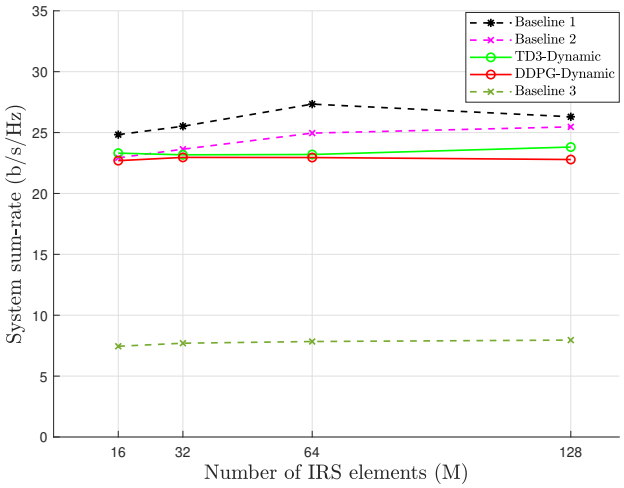

Figure 4.14: The achieved system sum rate of the proposed robust design versus the number of IRS elements with dynamic channels, for $K = N = 2$, $R^{min} = 0.3$ b/s/Hz.

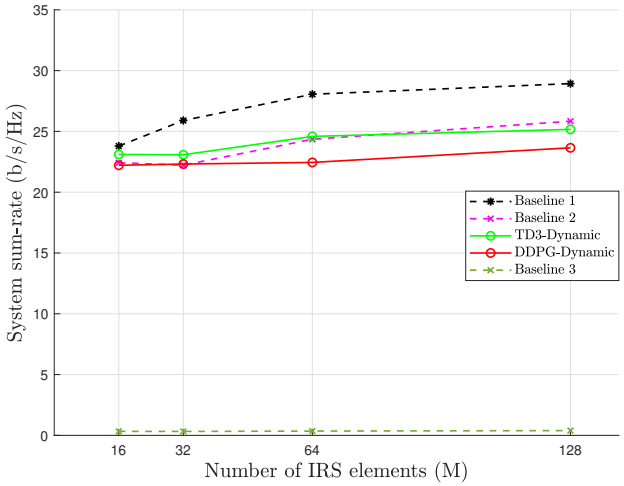

Figure 4.15: The achieved system sum rate of the proposed robust design versus the number of IRS elements with dynamic channels, for $K = N = 3$, $R^{min} = 0.3$ b/s/Hz.

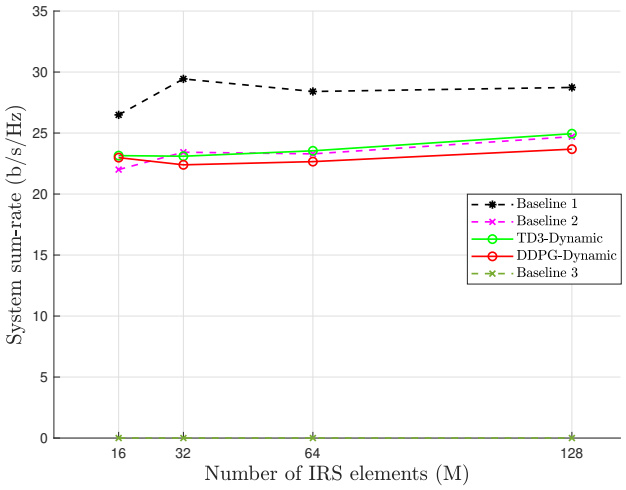

Figure 4.16: The achieved system sum rate of the proposed robust design versus the number of IRS elements with dynamic channels, for $K = N = 4$, $R^{min} = 0.3$ b/s/Hz.

Figure shows some casual drops of UE_4's rate below the 0.3 b/s/Hz mark by both the TD3 and the DDPG agents. This is expected due to the dynamic channels used for testing. Another observation is that DDPG achieved a higher rate for UE_1 at the expense of not satisfying the target rate required by UE_4, which is the result of converging to a non-optimal policy.

Finally, to evaluate the limits of the TD3 agent's derived policy in terms of robustness, the trained agent is tested for a set of target rates for $K = N = 4$. Figure 4.18 shows the robustness of the agent in satisfying each of the target rates. As expected, there is a trade-off between target rates and the robustness of the agent. Despite the dynamic channels used for testing, TD3 is able to maintain a robustness performance of at least 65%. Furthermore, with $M = 64$; the agent maintained a competitive score up to 0.5 b/s/Hz, which is 66% higher than the target rate used during training. While both agents achieve similar system sum rates as highlighted by Figures 4.14, 4.15 and 4.16, DDPG is less robust to channel uncertainties. The seemingly enhanced robustness score for baseline 2 is not related to the algorithm itself. Instead, it is due to the lower target rates used for dynamic-channels testing.

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

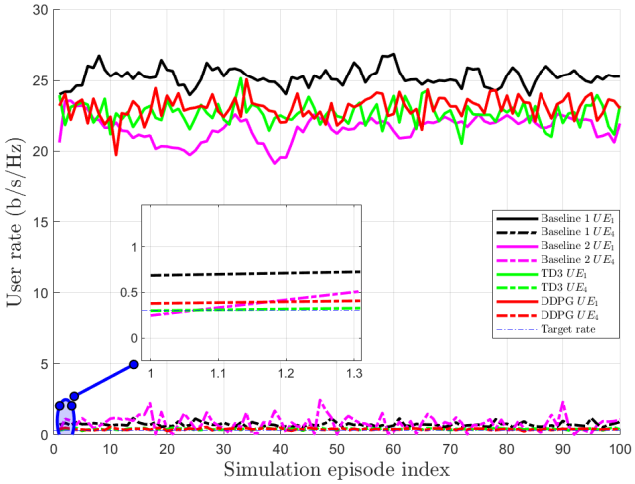

Figure 4.17: The achieved individual user rate of the proposed robust design across 100 testing episodes, with dynamic channels for $K = N = 4$, $R^{min} = 0.3$ b/s/Hz.

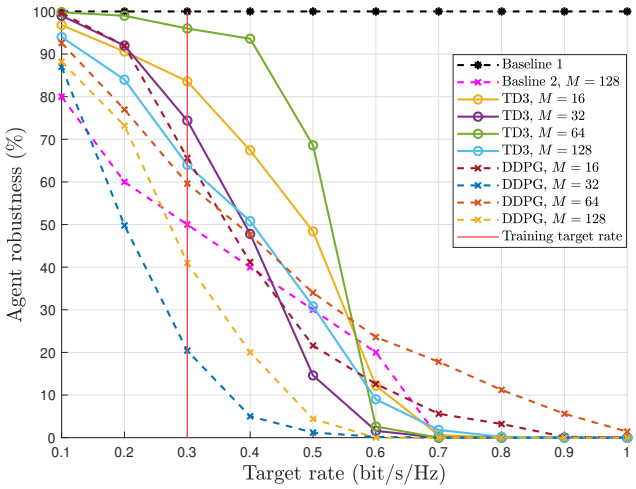

Figure 4.18: The robustness performance of the proposed agent versus the target rate with dynamic channels, for $K = N = 4$, $R^{min} = 0.3$ b/s/Hz.

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

4.4 Summary

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

Chapter 5

Outage-Constrained Resource Allocation for an IRS-Assisted DL MISO-NOMA System

The previous chapter proposed a joint phase shifts and beamforming design for an IRS-assisted DL MISO-NOMA system. However, according to the previous chapter's design, the number of required SIC operations by the strongest UE scales linearly with the number of active UEs using the same RBs resulting in a prohibitively complex receiver architecture. To address this issue, this chapter proposes a more generalized DRL-based framework that utilizes the user-clustering approach. In particular, a robust resource allocation framework is proposed for an IRS-assisted MISO-NOMA system is proposed in this chapter. In particular, a long-term system sum rate maximization objective is considered. The impacts of imperfect channel estimation on both the transmitter and the receiver are taken into account. 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 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.

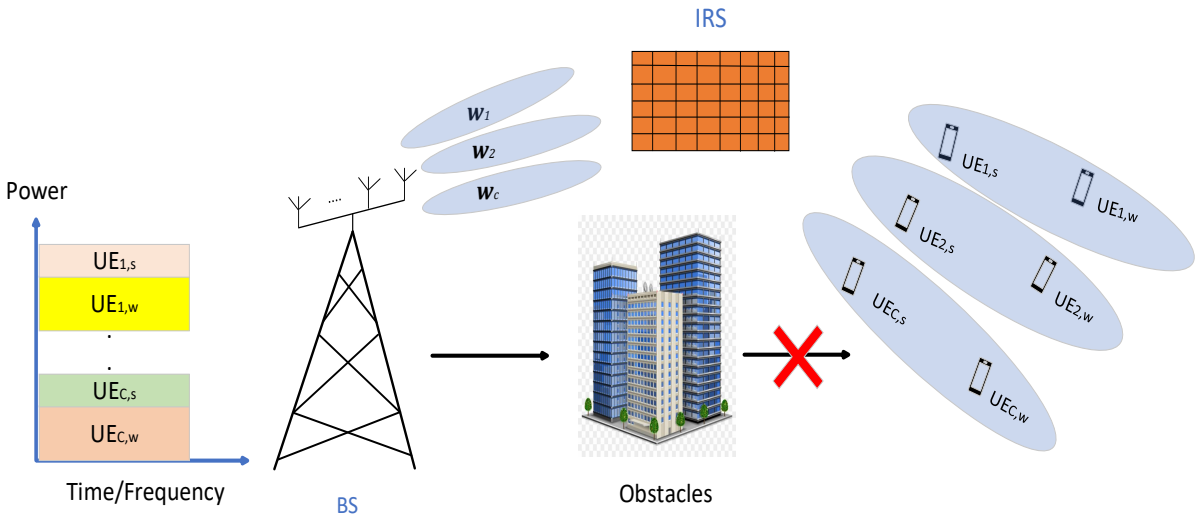

Figure 5.1: Cluster-based IRS-assisted DL MISO-NOMA system.

5.1 System and Channel Uncertainty Models

A DL transmission of an IRS-assisted MISO-NOMA system is considered in which the BS is equipped with N transmit antenna serves $2K$ single antenna UEs as shown in Figure 6.1. To increase the system capacity, the UEs are paired into $\mathcal{C} = \{1,...,C\}$ 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, the number of UEs in each cluster is limited to 2 [142, 143]. 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 $\mathcal{W}$. $\text{UE}_{c,s}$ and $\text{UE}_{c,w}$ are used 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 [122]. In addition, it is assumed 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 $\mathrm{UE}_{c,i}$ can be expressed as

$$y_{c,i} = \mathbf{g}_{c,i}^{\mathrm{H}} \Upsilon \mathbf{G} \sum_{c=1}^{C} \mathbf{w}_c x_c + z_{c,i}, \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \tag{5.1}$$

where $\mathbf{g}_{c,i} \in \mathbb{C}^{M\mathrm{x}1}$ represents the channel between $\mathrm{UE}_{c,i}$ and the IRS, $\mathbf{G} \in \mathbb{C}^{M\mathrm{x}N}$ denotes the channel between the BS and the IRS, and $\Upsilon = \mathrm{diag}(v_1, ..., v_M) \in \mathbb{C}^{M\mathrm{x}M}$ is the diagonal IRS phase shifts matrix, and $v_m = \zeta_m e^{j\theta_m}$. In this work, an ideal reflection is assumed at the IRS elements, i.e., $|v_m|^2 = 1, m = 1, ..., M$. $\mathbf{w}_c \in \mathbb{C}^{N\mathrm{x}1}$ 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 $\mathrm{UE}_{c,i}$ can be expressed in a more compact form as

$$y_{c,i} = \mathbf{h}_{c,i} \sum_{c=1}^{C} \mathbf{w}_c x_c + z_{c,i}, \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \tag{5.2}$$

where $\mathbf{h}_{c,i} = \mathbf{v}^{\mathrm{H}} \mathbf{Q}_{c,i} \in \mathbb{C}^{1\mathrm{x}N}$ is the final channel vector, $\mathbf{v} = \mathrm{vec}(\Upsilon) \in \mathbb{C}^{M\mathrm{x}1}$, and $\mathbf{Q}_{c,i} = \mathrm{diag}(\mathbf{g}_{c,i}^{\mathrm{H}})\mathbf{G} \in \mathbb{C}^{M\mathrm{x}N}$ is the cascaded channel for $\mathrm{UE}_{c,i}$. 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 work, and given that $||\mathbf{h}_{s,i}||_2 \gg ||\mathbf{h}_{w,i}||_2$, a fixed decoding order is assumed 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 [68, 144].

5.1.1 Channel Uncertainty Model

Due to the random nature of the wireless transmissions, uncertainties in the wireless channel estimations are inevitable. Furthermore, with the introduction of the IRS, accurate channel estimation becomes even more challenging due to the passive elements in the IRS [129, 145]. Channel estimation and quantization errors are two of the main contributors to the imperfect channel estimation in wireless communication systems [127, 128]. 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 [129, 146]. 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 work, the channel uncertainties are assumed to be the result of the imperfect channel estimation. Note that in NOMA systems, channel imperfections at the receiver leads to SIC degradation which is also taken into account. In particular, the aim of this work is to propose a robust resource allocation strategy that takes into account the imperfect CSI in the system. Therefore, the following error model is considered for the cascaded channel [129]:

$$\mathbf{Q}_{c,i} = \hat{\mathbf{Q}}_{c,i} + \Delta\mathbf{Q}_{c,i}, \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \tag{5.3}$$

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 is 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

$$\beta_{c,i}^2 = \lambda^2 ||\hat{\mathbf{q}}_{c,i}||_2^2, \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \tag{5.4}$$

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 [129]. 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 SINR and the corresponding achievable rates.

5.1.2 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

$$\gamma_{c,s} = \frac{|\mathbf{h}_{c,s}\mathbf{w}_c|^2 P_c \alpha_{c,s}}{|(\mathbf{v}^{\text{H}}\Delta\mathbf{Q}_{c,s})\mathbf{w}_c|^2 P_c \alpha_{c,w} + \sum_{\substack{k=1 \\ k\neq c}}^{C} |\mathbf{h}_{c,s}\mathbf{w}_k|^2 P_k + \sigma_{c,s}^2}, \quad \forall s \in \{\mathcal{S}\}, c \in \mathcal{C}, \tag{5.5}$$

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

$$\gamma_{c,w}^{c,w} = \frac{|\mathbf{h}_{c,w}\mathbf{w}_c|^2 P_c \alpha_{c,w}}{|\mathbf{h}_{c,w}\mathbf{w}_c|^2 P_c \alpha_{c,s} + \sum_{\substack{k=1 \\ k\neq c}}^{C} |\mathbf{h}_{c,w}\mathbf{w}_k|^2 P_k + \sigma_{c,w}^2}, \quad \forall w \in \{\mathcal{W}\}, c \in \mathcal{C}. \tag{5.6}$$

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

$$\gamma_{c,w}^{c,s} = \frac{|\mathbf{h}_{c,s}\mathbf{w}_c|^2 P_c \alpha_{c,w}}{|\mathbf{h}_{c,s}\mathbf{w}_c|^2 P_c \alpha_{c,s} + \sum_{\substack{k=1 \\ k\neq c}}^{C} |\mathbf{h}_{c,s}\mathbf{w}_k|^2 P_k + \sigma_{c,s}^2}, c \in \mathcal{C}. \tag{5.7}$$

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

$$\gamma_{c,w} = \Big(1 + min(\gamma_{c,w}^{c,s}, \gamma_{c,w}^{c,w})\Big), c \in \mathcal{C}. \tag{5.8}$$

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

$$\begin{aligned} R_{c,s} &= log_2(1+\gamma_{c,s}), \\ R_{c,w} &= log_2(1+\gamma_{c,w}), \forall s \in \{\mathcal{S}\}, w \in \{\mathcal{W}\}, c \in \mathcal{C}. \end{aligned} \tag{5.9}$$

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

5.2 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, the objective of maximizing the ergodic system sum rate is considered under channel uncertainties while taking into account the dynamics of the system over multiple time slots [147–149]. Therefore, the long-term outage-constrained joint robust design problem with the sum rate maximization objective can be formulated as

$$\underset{\mathbf{w}_c, \mathbf{v}, P_c, \alpha_{c,i}, b_{s,w}}{\text{maximize}} \quad \mathbb{E}\Bigg\{ \sum_{t=1}^{\infty} \delta^{t-1} \sum_{c=1}^{C} [R_{c,s}^t + R_{c,w}^t] b_{s,w}^t \Bigg\} \tag{5.10a}$$

subject to

$$p_i \triangleq \Pr\{\gamma_{c,i} \geq 2^{R_{c,i}^{min}} - 1\} \geq \Gamma, \forall i \in \{\mathcal{S}, \mathcal{W}\}, c \in \mathcal{C}, \tag{5.10b}$$

$$||\mathbf{w}_c||_2^2 = 1, c \in \mathcal{C}, \tag{5.10c}$$

$$\sum_{c=1}^{C} P_c \leq P_{max}, c \in \mathcal{C}, \tag{5.10d}$$

$$\alpha_{c,s}^t + \alpha_{c,w}^t = 1, c \in \mathcal{C}, s \in \mathcal{S}, w \in \mathcal{W} \tag{5.10e}$$

$$\sum_{c=1}^{C} b_{s,w}^t \leq 1, b \in \{0,1\}, c \in \mathcal{C}, \tag{5.10f}$$

$$|\mathbf{v}_m|^2 = 1, 0 \leq \theta_m \leq 2\pi, m = 1, ..., M, \tag{5.10g}$$

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 QoS constraint for each UE, and $b^t_{s,w} \in \{0,1\}$ is the binary UE pairing coefficient. The outage constraint in (5.10b) guarantees that the QoS requirements of the UEs are achieved with probability Γ, while the constraint in (5.10c) ensures normalized power for all the beamforming vectors. The constraints in (5.10d) and (5.10e) represent the maximum available transmit power for all clusters and the UEs power allocation coefficients within each cluster, respectively. The pairing constraint in (5.10f) guarantees that each stronger UE is only paired with a single weaker UE and vice versa. Finally, the constraints in (5.10g) 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 [150]. Note that even without considering the binary constraint, the problem in (6.12a) is still non-convex and NP-hard [69, 92, 151, 152], 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 (6.12a) since approximation methods cannot be directly applied.

- The outage constraints in (5.10b) do not admit closed-form solutions [153].

- The UE pairing variable in (5.10f) 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.

5.2.1 User Pairing

UE pairing is considered one of the enabling techniques in multi-user NOMA systems for future wireless networks [142, 143, 154]. In addition, it has been shown that pairing a stronger UE with a weaker UE leads to enhanced overall performance in NOMA systems [155, 156]. 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 [157, 158]. Since each cluster is served with a single beam, higher UEs 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 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, the correlation coefficient between two UEs in the system is defined as [158]

$$\varepsilon_{i,j} = \frac{||\hat{\mathbf{h}}_i.\hat{\mathbf{h}}_j||_2}{||\hat{\mathbf{h}}_i||_2||\hat{\mathbf{h}}_j||_2}, \forall i \in \mathcal{S}, \forall j \in \mathcal{W}, \tag{5.11}$$

where $\hat{\mathbf{h}}_k, k \in \{i, j\}$, is the estimated final channel for UE_k and is known at the BS. Algorithm 3 provides the key steps for the proposed UE pairing design. Therefore, executing Algorithm 3 will eliminate the binary constraint in (5.10f). The next section presents the robust resource allocation framework for a given UE pairing configuration.

Algorithm 3 SUPA

Initialise: UEs sets $\mathcal{S}$,$\mathcal{W}$, initial IRS vector $\mathbf{v}_{init}$, and UE clusters $c \in \mathcal{C}$
Calculate the final estimated channels at the BS using $\hat{\mathbf{h}}_{c,i} = \mathbf{v}_{init}^{\mathrm{H}} \mathbf{Q}_{c,i}, \forall c \in \mathcal{C}, \forall i \in \mathcal{S}, \mathcal{W}$
Sort all $\mathrm{UE}_i, \forall i \in \mathcal{S}$, according to their channel norms such that $||\hat{\mathbf{h}}_1||_2 \geq ||\hat{\mathbf{h}}_2||_2 \geq ... \geq ||\hat{\mathbf{h}}_K||_2$
for $i = 1 : K, i \in \mathcal{S}$ **do**
  for $j = 1 : K, j \in \mathcal{W}$ **do**
    Calculate the correlation coefficent between UE_i and UE_j according to (5.11)
  end for
  Find $j' = \mathrm{argmax}(Corr_{i,j}), \forall j \in \mathcal{W}$
  Assign UE_i and $\mathrm{UE}_{j'}$ to cluster $c(i)$
  Set $\hat{\mathbf{h}}_{j'} \leftarrow \mathbf{0}$, $j' \in \mathcal{W}$
end for
Output: $\{\mathrm{UE}_{1,s}, \mathrm{UE}_{1,w}\}$,...,$\{\mathrm{UE}_{C,s}, \mathrm{UE}_{C,w}\}$

5.3 RL Framework For Robust Resource Allocation

With given UE pairs using Algorithm 3, the remaining resource allocation problem is expressed as

$$\underset{\mathbf{w}_c, \mathbf{v}, P_c, \alpha_{c,i}}{\text{maximize}} \quad \mathbb{E}\left\{ \sum_{t=1}^{\infty} \delta^{t-1} \sum_{c=1}^{C} [R_{c,s}^t + R_{c,w}^t] \right\} \tag{5.12a}$$

$$\text{subject to} \quad (5.10b), (5.10c), (5.10d), (5.10e), (5.10g). \tag{5.12b}$$

Unfortunately, the optimization problem in (5.12a) 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 (5.10c) [159].

5.3.1 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 as-

sumption, the ZFBF provides a closed-form solution to the beamforming design problem with a reasonable trade-off between complexity and performance [160]. In addition, the ZFBF has been extensively used in the literature as one of the beamforming designs for sum rate maximization [158, 160, 161]. 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

$$\frac{\mathbf{h}_i}{||\mathbf{h}_i||_2}\mathbf{w}_k = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{if } k \neq i. \end{cases} \tag{5.13}$$

However, since a multi-cluster NOMA system is considered, the ZFBF vector can only be designed based on a single channel for each cluster, not both. Hence, in this work 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 from the imperfect beamforming design based on the estimated channel. Thereby, the expression in (5.13) can written as

$$\frac{\hat{\mathbf{h}}_i}{||\hat{\mathbf{h}}_i||_2}\mathbf{w}_k = \begin{cases} 1 & \text{if } k = i \\ > 0 & \text{if } k \neq i. \end{cases} \tag{5.14}$$

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, $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_C]$ is defined as the matrix that contains the ZFBF vectors for all clusters, and $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_{1,s}^T, ..., \hat{\mathbf{h}}_{C,s}^T]^T$ 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 [158]:

$$\mathbf{W} = (\hat{\mathbf{H}})^{\dagger} \tag{5.15}$$

where $(\hat{\mathbf{H}})^{\dagger} = \hat{\mathbf{H}}^{\text{H}}\left(\hat{\mathbf{H}}\hat{\mathbf{H}}^{\text{H}}\right)^{-1}$ 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.

5.3.2 Problem Reformulation

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

$$\underset{\mathbf{v}, P_c, \alpha_{c,i}}{\text{maximize}} \quad \mathbb{E}\left\{ \sum_{t=1}^{\infty} \delta^{t-1} \sum_{c=1}^{C} [R_{c,s}^t + R_{c,w}^t] \right\} \tag{5.16a}$$

$$\text{subject to} \quad (5.10b), (5.10d), (5.10e), (5.10g). \tag{5.16b}$$

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, to develop a joint robust design, the problem in (5.16a) 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 an MDP [162]. 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 [104]. 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 $\pi^*(a|s)$ 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.

In order 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

$$\mathbf{a}^t = [P_1^t, ..., P_C^t, \alpha_{1,w}^t, ..., \alpha_{C,w}^t, \mathbf{v}^t]^{\mathrm{T}}. \tag{5.17}$$

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 the used DNN architecture is only compatible with real numbers, complex vectors are represented using real values in this work. 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 [163]. Therefore, $\mathbf{a}^t \in \mathbb{R}^{(2K+2M)\mathrm{x}1}$ is written 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, the previous action is included 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 (5.11), 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

$$\rho_{i,j} = 10 log_{10}\big(\frac{||\hat{\mathbf{h}}_i||_2}{||\hat{\mathbf{h}}_j||_2}\big), \forall i \in \mathcal{S}, j \in \mathcal{W}. \tag{5.18}$$

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

$$\mathbf{s}^t = [\mathbf{a}^{t-1}, \varepsilon_1^{t-1}, ..., \varepsilon_C^{t-1}, \rho_1^{t-1}, ..., \rho_C^{t-1}, R_{1,s}^{t-1}, ..., R_{C,w}^{t-1}]^{\mathrm{T}}, \tag{5.19}$$

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

$$\begin{aligned}\mathbf{s}^t_{\text{dyn}} = [&\beta^2_{1,s},...,\beta^2_{C,w},\mathbf{a}^{t-1},\varepsilon^{t-1}_1,...,\varepsilon^{t-1}_C,\\ &\rho^{t-1}_1,...,\rho^{t-1}_C,R^{t-1}_{1,s},...,R^{t-1}_{C,w}]^{\mathrm{T}},\end{aligned} \tag{5.20}$$

where $\mathbf{s}^t \in \mathbb{R}^{(8K+2M)\text{x}1}$. Note that since the variance of the estimated channel is closely related to the estimation error according to (5.4), 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 (6.12a) 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

$$r^t = \sum_{c=1}^{C}(R^t_{c,s}+R^t_{c,w})+\sum_{c=1}^{C}\varepsilon^t_c+\sum_{c=1}^{C}\rho^t_c, c\in\mathcal{C}. \tag{5.21}$$

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

$$r^t = \sum_{k=1}^{2K}\min(R^t_k - R^{min}_k, 0), \tag{5.22}$$

where $r^t < 0$ always hold in (5.22). Therefore, after each action taken by the agent, the environment uses the positive reward function in (5.21) in case the action satisfies the QoS constraints, otherwise, the environment uses the negative reward function in (5.22). The details of how the reward function is utilized by the agent during training is discussed in 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 is 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 (5.10d), the feasible cluster power vector is expressed

$$\bar{\mathbf{P}}^t = \frac{P_{max}}{\sum_{c=1}^{C} P_c^t}\mathbf{P}^t, \tag{5.23}$$

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

$$\bar{v}_m^t = \frac{v_m^t}{|v_m^t|}, m = 1, ..., M. \tag{5.24}$$

Note that the angle θ_m can be directly mapped to the feasible region.

5.3.3 The Robust TD3-based Algorithm

The RL agents like the Q-learning and the SARSA are called tabular methods because they use tables to keep track of the Q-values for each state-action pair [105, 164]. 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 [91].

In this work, the proposed robust resource allocation framework is developed based on the TD3 agent [137]. 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 [115], 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 $\mathcal{D}$ 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_ψ is the one responsible for taking action. 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

$$\mu_\psi(\mathbf{s}) = \mathbf{a}^*, \tag{5.25}$$

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_ψ. 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

$$\phi_i(\mathbf{s}, \mathbf{a}) = Q^*, i = 1,2. \tag{5.26}$$

where Q^* is the optimal Q-value for the state-action pair. Note that (5.26) 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. $\mu'_\psi{}'$ and $\phi'_i, i = 1,2$, represent the actor's and the critics' target networks. To elaborate, the training starts by sampling a batch of experiences $\mathcal{B}$ from the replay buffer. However, the focus is placed 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:

$$y'(r^t, \mathbf{s}^{t+1}) = r^t + \delta \min_{i=1,2} \phi_i'\left(\mathbf{s}^{t+1}, \mu_\psi'(\mathbf{s}^{t+1})\right), i = 1,2, \tag{5.27}$$

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 δ value that is closer to 1 means that the agent is interested in maximizing its long-term reward. Note that according to (5.27), both the actor's target and critics' target networks are used to calculate $y'(r^t, \mathbf{s}^t)$. After obtaining the target using the minimum Q-value, both critics are trained by minimizing their respective MSE objectives. This is expressed as [137]

$$L(\phi_i, \mathcal{D}) = \mathop{\mathbb{E}}_{\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\} \sim \mathcal{D}} \left[\left(Q_{\phi_i}(\mathbf{s}^t, \mathbf{a}^t) - y'(r^t, \mathbf{s}^t) \right)^2 \right], i = 1,2. \tag{5.28}$$

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 (5.28), 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 [115]

$$\max_{\psi} \mathop{\mathbb{E}}_{\mathbf{s}^t \sim \mathcal{D}} \left[\mathcal{Q}_\phi \left(s, \mu_\psi(s) \right) \right], \tag{5.29}$$

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:

$$\begin{aligned} \phi_i' &= \kappa \phi_i + (1-\kappa)\phi_i', \, i = 1,2, \\ \psi' &= \kappa \psi + (1-\kappa)\psi', \end{aligned} \tag{5.30}$$

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

$$\mathbf{a}^t = \text{clip}\big(\mu_\psi(\mathbf{s}^t) + \mathbf{n}, a_{high}, a_{low}\big), \tag{5.31}$$

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'.

So far, the problem reformulation into an RL environment is discussed and the inner workings of the TD3 agent are explained. Hence, the developed TD3-based algorithm for robust resource allocation is explained in Algorithm 6.

Note that unlike conventional optimization algorithms, the outage probability during the training and learning stage is not explicitly considered in the TD3-based robust design, however, it is included implicitly through the random errors as explained in Algorithm 6. The first motivation for the proposed approach is that since the TD3 agent is initialized with a random policy, basing the reward function 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 aim 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 6. 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.

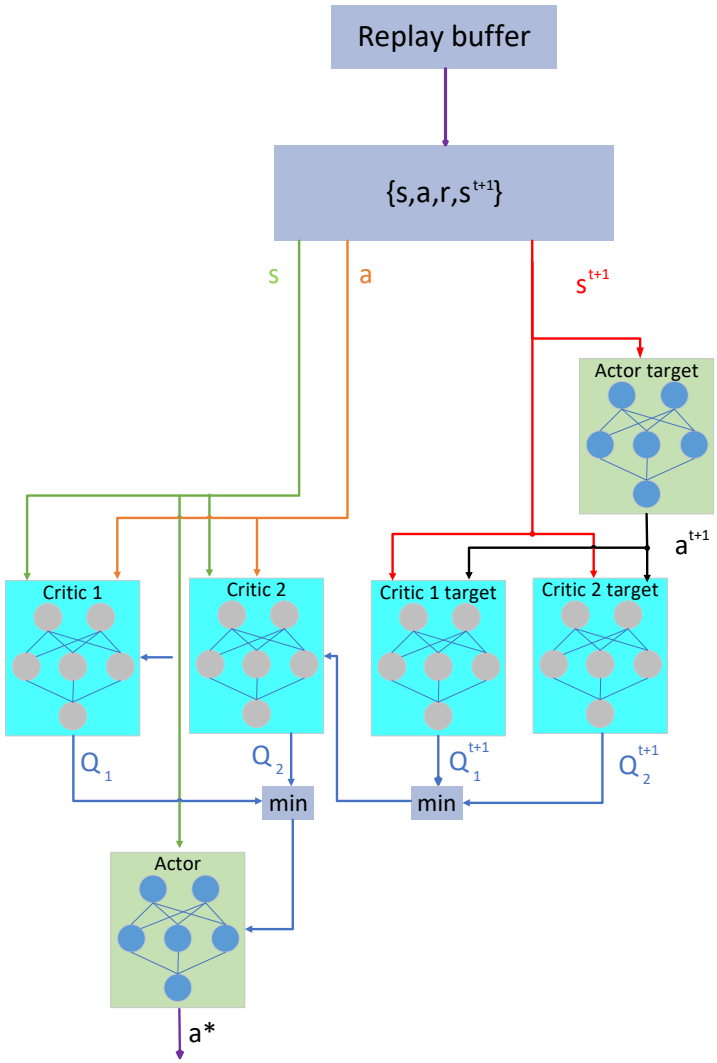

Figure 5.2: The actor-critic interactions in the proposed TD3 agent.

Algorithm 4 TD3-based robust resource allocation

Initialise: agent's hyperparameters $\mu_\psi, \phi_1, \phi_2, \mathcal{D}, \mathcal{N}, b$, and the IRS vector $\mathbf{v}_{init}$
Set $\phi_i' \leftarrow \phi_i, i = 1,2$, and $\mu_\psi' \leftarrow \mu_\psi$
while $episode \leq Episodes$ **do**
  Obtain the estimated channels for all UEs, $\hat{\mathbf{h}}_k, k = 1,...,2K$
  Execute algorithm 3 to obtain the UE pairs.
  Calculate the ZFBF matrix $\mathbf{W}$ according to (5.15)
  Obtain the channel error samples $\Delta\mathbf{Q}_1,...,\Delta\mathbf{Q}_{2K}$ according to (5.3)
  while $step \leq Steps$ **do**
    Get the actions vector $\mathbf{a}^t$ by evaluating the actor's DNN using the current state according to (5.31)
    Extract $\bar{\mathbf{v}}^t, \bar{\mathbf{P}}^t$ according to (5.23) and (5.24)
    Add the random channel error terms according to (5.3) to create the final true channels
    Evaluate the SINR equations for all UEs according to (5.5) and (5.8) using the true channels
    Calculate the achieved rates for all UEs according to (5.9)
    if $R_k^t \geq R_k^{min}, k = 1,...,2K$: **then**
      Use the reward function in (5.21)
    else
      Use the reward function in (5.22)
    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 $\mathcal{B}$ experiences randomly from $\mathcal{D}$
    Calculate the targets for the sampled experiences according to (5.27)
    Train the two critics using (5.28)
    if $update\_policy == True$: **then**
      Train the actor network using (5.29)
    end if
    Update the target networks using (5.30)
    $step = step + 1$
    Set $\mathbf{s}^t = \mathbf{s}^{t+1}$
  end while
  $episode = episode + 1$
end while
Output : $[\mathbf{w}_1,...,\mathbf{w}_C, \bar{\mathbf{v}}^*, \bar{\mathbf{P}}^*, \alpha_{1,s}^*,..., \alpha_{C,w}^*]$

5.3.4 Complexity Analysis

In this section, the computational complexity for the developed TD3-based algorithm is defined. In particular, since DRL agents are only trained once, it is assumed that the offline training complexity can be afforded [163]. Hence, the focus is placed 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 Ψ hidden layers, with each consisting of $\aleph$ neurons, then it is straightforward to conclude that there are $\Psi+1$ 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 $\mathcal{O}\Big(T\left(\aleph\cdot\mathbf{Card}(\mathbf{s}^t)+\Psi\cdot\aleph^2+\mathbf{Card}(\mathbf{a}^t)\cdot\aleph+\Psi\cdot\aleph+\mathbf{Card}(\mathbf{a}^t)+CN^2\right)\Big)$, where $\mathbf{Card}(\mathbf{s}^t)=8K+2M$ for the dynamic-channels case as highlighted by (5.20), $\mathbf{Card}(\mathbf{a}^t)=2K+2M$, the term $CN^2, C\geq N$, represents the complexity for calculating the pseudoinverse in (5.15), while the terms $\Psi\cdot\aleph$ 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 $\aleph\gg\mathbf{Card}(\mathbf{s}^t)$, and $\aleph\gg CN^2$, then, the worst-case run-time for the actor's DNN is reduced to $\approx\mathcal{O}(\aleph^2)$, 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 the previous action is considered 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.

To compare the analytical complexity of the proposed TD3-based algorithm to existing convex optimization algorithms, three widely adopted conventional optimization approaches for solving the static version of the considered optimization problem are briefly reviewed. In [165], a SOCP-ADMM-based algorithm was developed 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 [69]. The proposed iterative algorithm solves a SOCP optimization problem with a worst-case complexity of $\mathcal{O}\big((2K)^7\big)$ per iteration. For IRS-aided MISO systems, the work in [152] proposed an 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 work, it is assumed 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 $\mathcal{O}(K^2)$.

5.4 Training, Simulation and Numerical Results

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

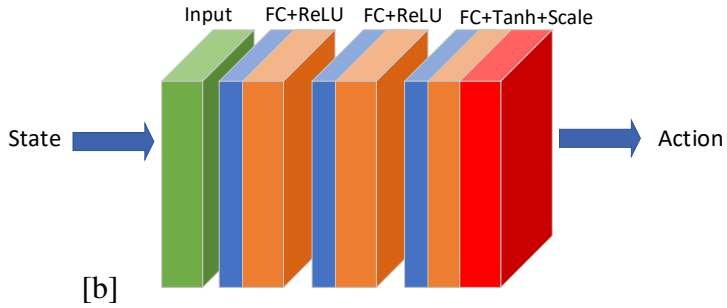

Figure 5.3: Actor's DNN architecture

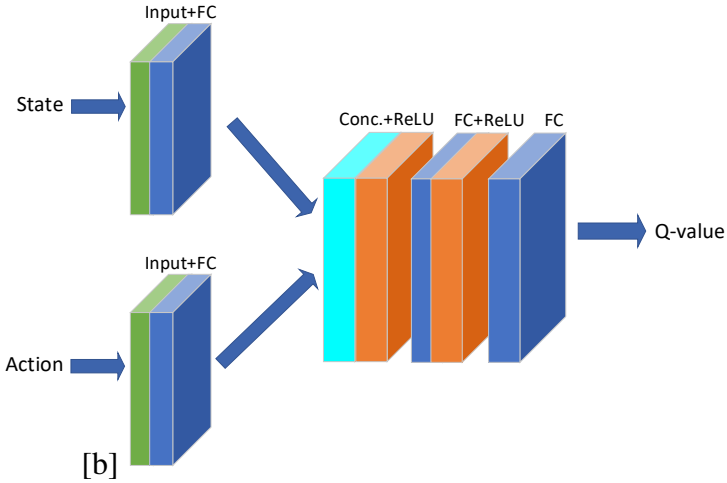

Figure 5.4: Critic's DNN architecture.

5.4.1 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 **??** and 6.3, 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 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

Table 5.1: 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 ($\mathcal{D}$)	$100,000$
Minibatch size ($\mathcal{D}$)	128
Number of episodes, time steps (fixed channels)	$300, 1000$
Number of episodes, time steps (dynamic channels)	$800, 1000$

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 [141]. Note the number of neurons in each hidden layer is identical for both DNNs. Table 5.1 lists the TD3 agent's hyperparameters and the training parameters used in this work.

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 [166], two different neuron values for each channel case are used. In particular, for the fixed-channels case, one set of simulation results is generated 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 is also generated. Similarly, the same process is replicated for the dynamic-channels case with 256 and 512 neurons for each set of simulation results.

5.4.2 System Parameters

A DL transmission for a clustered and IRS-assisted MISO-NOMA system that is identical to the one illustrated in Figure 6.1 is considered. In addition, the channel between the BS and the IRS is assumed to have both an LoS and non-LoS components, and therefore, modelled using the Rician fading coefficients. In particular, the BS-IRS link is expressed as

$$\mathbf{G} = \frac{1}{\sqrt{d_{irs}^{\iota_{b\to irs}}}}\left(\sqrt{\frac{K'}{1+K'}}\mathbf{G}_{LoS} + \sqrt{\frac{1}{1+K'}}\mathbf{G}_{nLoS}\right), \tag{5.32}$$

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 $K' = 1$ 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

$$\mathbf{g}_k = \frac{\tilde{g}}{\sqrt{d_k^{\iota_{irs\to u}}}}, k = 1, ..., 2K, \tag{5.33}$$

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, it is assumed that the UEs are located between $[50-100]$ m away from the BS. Table 6.2 lists all the system parameters used to generate the simulation results.

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

- **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 [152], then, the non-robust ZFBF with fixed power allocation is used for the beamforming design.

Table 5.2: 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

- **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.

5.4.3 Fixed-Channels Case

To evaluate the performance of the proposed algorithm against channel errors, the case where the channels are fixed throughout the training process is considered first. 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 5.5 illustrates the convergence of the TD3 and DDPG agents. With two clusters (i.e., $C = 2$), both agents develop a highly rewarding policy after a few training episodes. However, when the number of users in the system increases, both agents require at least 150 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. 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 6.8. The rates provided represent the average system sum rate over 1000 testing episodes. Across the three cluster settings, the TD3 agent outperforms the benchmark schemes. In particular, the TD3-256 agent achieves the highest average sum rate of approximately 18 Bit/s/Hz, when $C = 4$, with 3.5 Bit/s/Hz gap compared to the benchmark schemes.

Note that Figure 6.8 only shows partial information about the agent's performance. To gain a better insight, Figure 6.8 is interpreted in the context of the outage performance of the agent illustrated by Figures 5.8 and 5.9. However, since the outage performance of the agent is related to the weakest UE's achieved rate, Figure 5.7 depicts the achieved rates PDF for the weakest UEs in the system. Based on the weakest UE rate for each setting, it can be inferred 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 5.7 are for the weakest UEs in each category, this represents the worst-case performance of the agent.

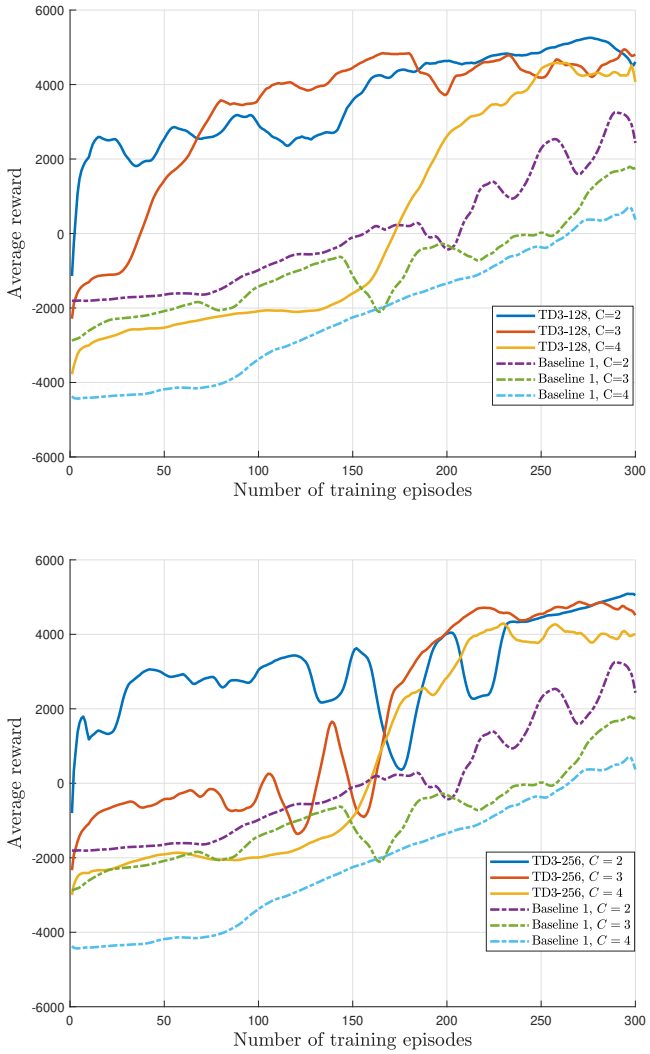

Figure 5.5: Convergence of the proposed TD3 agent for the fixed-channels case.

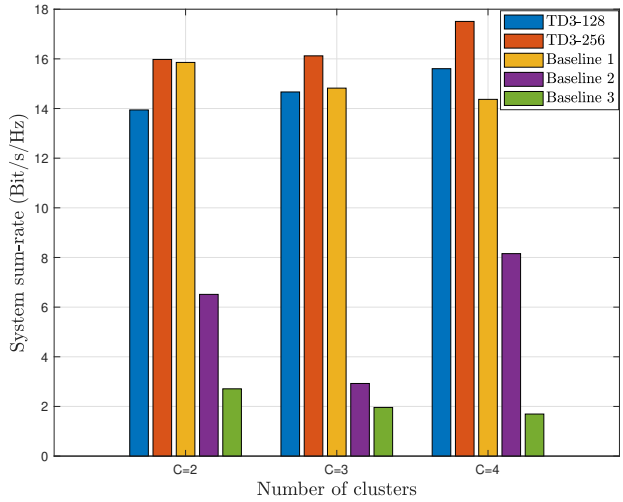

Figure 5.6: The average system sum rates for the fixed-channels case with various number of UEs.

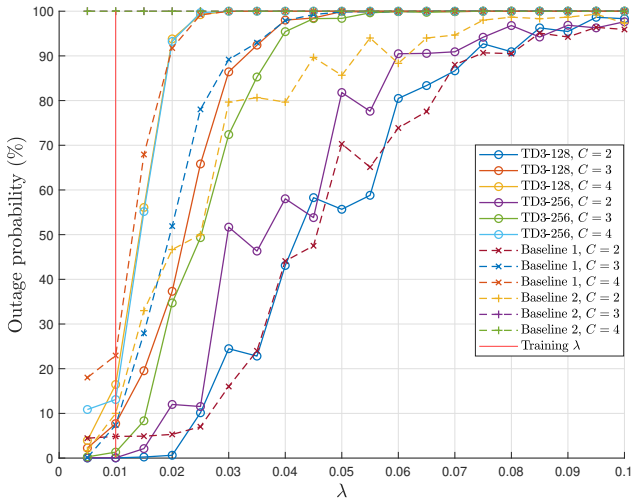

Figure 5.8: The average outage probability versus the estimation quality factor λ.

To assess the outage performance of the proposed agent against the relative channel estimation quality λ, Figure 5.8 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

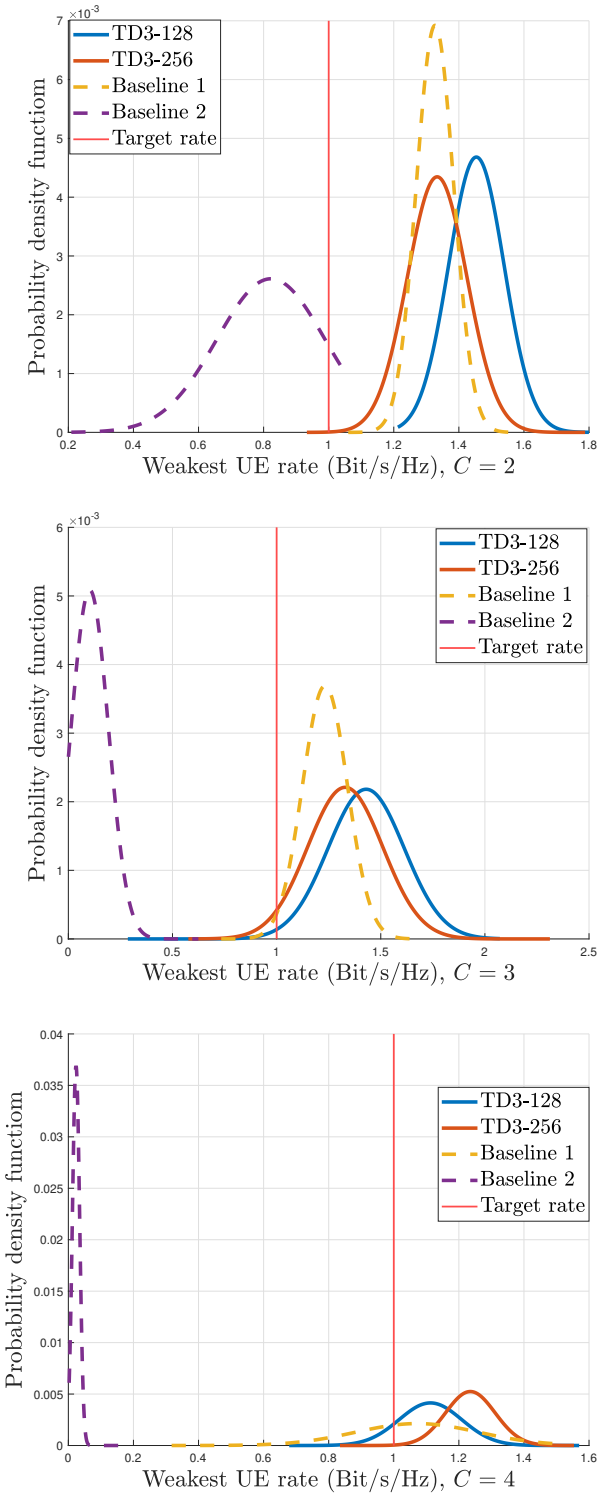

Figure 5.7: The PDFs for the weakest UE's achieved rate in the system.

probability of 84% and 87% for TD3-500 and TD3-1000, respectively, 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 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 a 93% 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 5.9 illustrates the non-outage probability versus different target rates. The agent's performance generally follows the same pattern as in Figure 5.8, 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 6% performance gap. In addition, the TD3 agent's policy is able to sustain a 50% increase in the target rate while still achieving a non-outage probability of 90% on average, which proves that the agent has developed a solid robust policy.

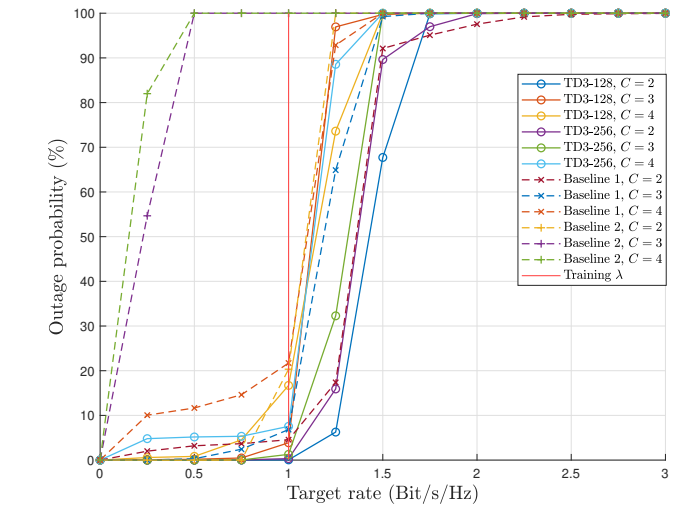

Figure 5.9: The average outage probability versus the target rate R_k^{min}.

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 hyper parameterized in the proposed robust design, it is impacted by the selected learning parameters of the TD3 agent.

5.4.4 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, the developed algorithm is extended 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

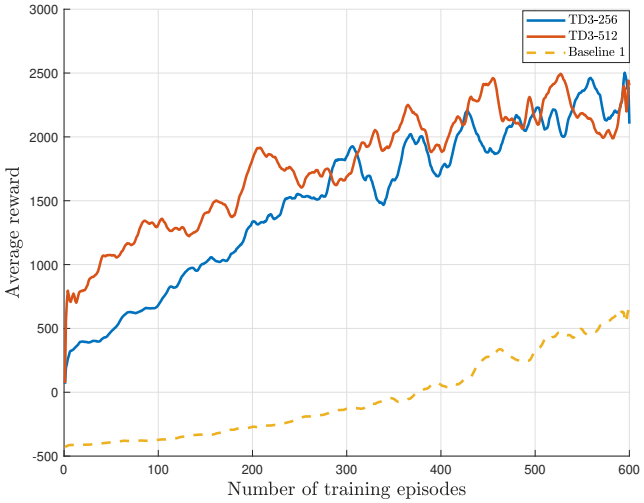

Figure 5.10: Convergence of the TD3 agent for the dynamic-channels case, $C = 2$.

can be generalized to never-seen-before channels. Hence, after training the agent once, it could be deployed to any channel condition afterwards.

Figure 5.10 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 5.11. The average sum rates figure shows that baseline 1 achieves the highest rate, which is explained by the worse outage performance illustrated in Figure 5.12. 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 17% at the 0.2 Bit/s/Hz target rate. On the other hand, the DDPG agent has an average outage probability of around 23% at the same target rate. In addition, the average outage performance gap between the TD3 agent and baseline 2 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. Fur-

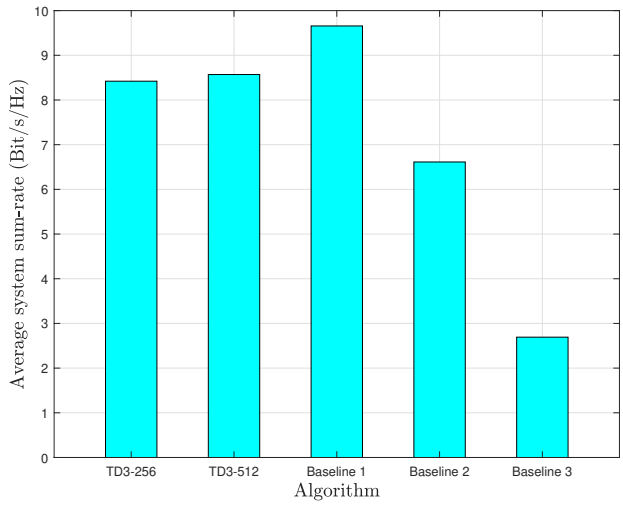

Figure 5.11: The average system sum rates for the dynamic-channels case, $C = 2$.

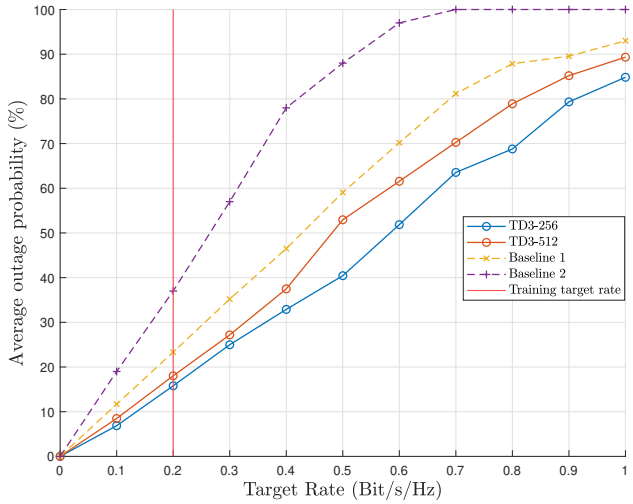

Figure 5.12: The average outage probability of the TD3 agent versus the target rate for the dynamic-channels case, $C = 2$.

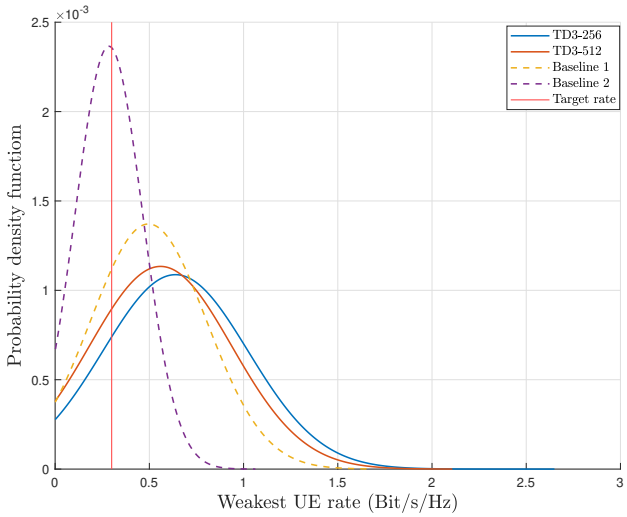

Figure 5.13: The PDFs for the weakest UE's achieved rate in the system, $C = 2$.

thermore, the PDFs of the average rate achieved by the weakest UE in the system are illustrated in Figure 5.13. 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 achieves both the lowest outage probability and the highest sum rate for the fixed-channels case, while the agent sacrifices some of the average system sum rate for an additional non-outage gain. This shows that the proposed TD3-based algorithm is capable of converging to adaptive policies that suit the problem requirements.

5.5 Summary

The resource allocation problem for an IRS-assisted MISO-NOMA system is considered in this chapter. 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 is formulated. A correlation-based UE clustering algorithm is proposed to pair the UEs into clusters. Then, the challenging robust design problem is reformulated into an RL environment since it cannot be solved directly using conventional optimization techniques. Subsequently,

a DRL-based framework is developed to solve the reformulated problem using the TD3 agent. The simulation results demonstrate 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 shows 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.

In the next chapter, a DRL-based novel framework for maximizing the EE in IRS-assisted NOMA systems is proposed.

Chapter 6

Robust EE Maximization for an IRS-Assisted UL NOMA System

In the previous two chapters, DRL-based robust resource allocation strategies were developed with the aim to maximize the long-term system sum rate under channel uncertainties. In this chapter, the EE objective is considered. In particular, this chapter proposes a joint robust design for an IRS-assisted UL NOMA system. In particular, the imperfect CSI at the BS is taken into account. Then, the robust design problem is formulated as a long-term EE maximization problem subject to quality-of-service constraints. Both bounded and unbounded channel uncertainty models are considered in the formulated robust design. Moreover, the formulated robust problem is not jointly convex in the optimization variables and cannot be solved directly using conventional optimization methods. Therefore, the MMSE-SIC receiver is utilized to deal with the robust beamforming design. Subsequently, a novel DRL-based framework to jointly optimize the users' power allocation and the IRS phase shifts using the SAC DRL agent is developed. In particular, the non-convex joint design problem is reformulated as an RL environment and a SAC-based agent is developed to solve the reformulated problem. Through extensive simulation results, the robust and competitive performance of the proposed algorithm for both the fixed and the dynamic channels and compare the performance with that of the benchmark schemes in the literature are demonstrated. The simulation results show that the proposed algo-

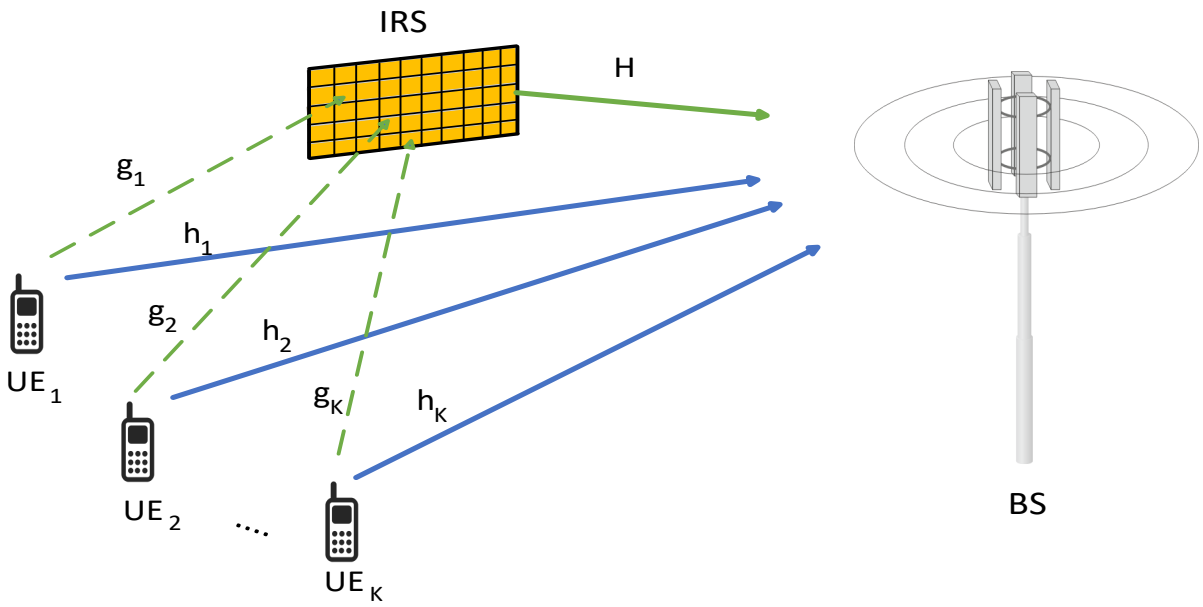

Figure 6.1: Multi-user IRS-assisted UL NOMA system.

rithm outperforms the benchmark schemes by achieving significantly higher EE levels for both fixed and dynamic channels.

6.1 System and Channel Models

In this work, an UL transmission of an IRS-assisted NOMA system as shown in Figure 6.1 is considered where K single-antenna UE communicates with a multi-antenna BS. It is assumed that the BS is equipped with N antennas. The IRS consists of M passive phase shifters that provide an additional communication link between the UEs and the BS besides the direct link. Since a NOMA UL transmission is considered, all UEs are scheduled to transmit on the same RB to enhance the spectral efficiency of the system. 6.1. In addition, $\mathbf{g}_i \in \mathbb{C}^{M\text{x}1}$ represents the channel between UE$_i$ and the IRS, $\mathbf{h}_i \in \mathbb{C}^{N\text{x}1}$ is the direct channel between UE$_i$ and the BS, while $\mathbf{H} \in \mathbb{C}^{N\text{x}M}$ is the channel between the IRS and the BS. $\Upsilon \in \mathbb{C}^{M\text{x}M}$ is the diagonal IRS phase shifts matrix and is expressed as $\Upsilon = \text{diag}(v_1, ..., v_M)$. Moreover, $v_m = \zeta_m e^{j\theta_m}$, where ζ_m and θ_m represent the amplitude and the angle of m-th IRS element, respectively. Note that in this work, an ideal reflection is assumed, i.e., $|\zeta_m|^2 = 1$. Furthermore, the phase shifts of the

IRS elements are computed at the BS and fed back to the IRS through a feedback link [122]. Therefore, the received signal at the BS can be expressed as

$$\mathbf{y} = \sum_{i=1}^{K} (\mathbf{H}\Upsilon\mathbf{g}_i + \mathbf{h}_i)\sqrt{p_i}s_i + \mathbf{z}, \forall i \in \mathcal{K}, \tag{6.1}$$

where $s_i\,(\mathbb{E}[|s_i|^2] = 1)$ and p_i are the normalized information symbol and the transmit power of UE_i, respectively. $\mathbf{z} \sim \mathcal{CN}(0, N_0\mathbf{I}_N)$ is the additive white Gaussian noise at the BS, N_0 is the noise spectral density, and $\mathcal{K}$ is the set of all active UEs sharing the same RB in the system. In order to represent the channel more compactly, the variable $\mathbf{v} = \text{vec}(\Upsilon) \in \mathbb{C}^{M\text{x}1}$ that captures the diagonal elements of Υ is introduced. Then, the final effective channel for UE_i can be written as $\tilde{\mathbf{h}}_i = \mathbf{Q}_i\mathbf{v} + \mathbf{h}_i$, where $\mathbf{Q}_i = \text{diag}(\mathbf{g}^{\text{H}}{}_i)\mathbf{H}, \mathbf{Q}_i \in \mathbb{C}^{N\text{x}M}$. Hence, the received signal at the BS can be expressed as

$$\mathbf{y} = \sum_{i=1}^{K} \tilde{\mathbf{h}}_i\sqrt{p_i}s_i + \mathbf{z}, \forall i \in \mathcal{K}, \tag{6.2}$$

The next section provides the details of the channel uncertainty models considered in this work.

6.1.1 Channel Uncertainty Models

Distortion in the channel estimate is inevitable in wireless communications [167]. As a result, it is extremely challenging for the BS to obtain the accurate channel at the receiver. In order to reduce the negative impact introduced by channel errors, the robust design approach takes such imperfections into account in order to improve the reliability of the wireless communication system. Since the reflected (cascaded) channel through the IRS is more challenging to estimate, it is assumed that the direct channel can be perfectly estimated at the BS while the channel uncertainty in the reflected link is taken into account [145]. Different approaches for the robust design have been proposed in the literature [168–170]. In general, the robust design mainly depends on the channel uncertainty model. The two most widely adopted uncertainty models are briefly explained in the following subsections.

6.1.1.1 The Bounded Error Model

One of the often-used channel uncertainty models is the BEM [57, 171]. According to this channel error model, the CSI is assumed to be within a well-defined bounded region (for example, multi-dimensional ellipsoids). Such an uncertainty model is often associated with the FDD systems where the channel report is plagued with quantization errors and provided through a rate-limited feedback link [129]. Therefore, since the error bound is known, the robust design developed to mitigate this type of channel uncertainty is called the worst-case design. Hence, according to the BEM, the true reflected channel can be expressed as

$$\mathbf{Q}_i = \hat{\mathbf{Q}}_i + \Delta\mathbf{Q}_i, \forall i \in \mathcal{K}, \tag{6.3}$$

where $\hat{\mathbf{Q}}_i$ is the estimated channel known at BS, and $\Delta\mathbf{Q}_i$ is the unknown error that belongs to a bounded region. In particular, the unknown error term is norm-bounded such that $||\Delta\mathbf{Q}_i||_F \leq e_i, \forall i \in \mathcal{K}$. Furthermore, the error bound e_i is known at the BS and is expressed as [129]

$$e_i = \sqrt{\frac{\beta_i^2 \Gamma_{2MN}^{-1}}{2}}, \forall i \in \mathcal{K}, \tag{6.4}$$

where $\beta_i^2 = \lambda^2 ||\hat{\mathbf{q}}_i||_2^2$, $\hat{\mathbf{q}}_i = \text{vec}(\hat{\mathbf{Q}}_i) \in \mathbb{C}^{MN\text{x}1}$ is the variance of $\Delta\mathbf{Q}_i$. $\lambda \in (0, 1]$ is a scalar that indicates the relative value of the error boundary. The term Γ_{2MN}^{-1} represents the inverse of the cumulative distribution function (CDF) for the Chi-square distribution with $2MN$ degrees of freedom. It is obvious from (6.4) that the error region is a function of the system parameters M and N.

6.1.1.2 The Unbounded Error Model

The other widely adopted uncertainty model is the unbounded error model. The channel error under this model is assumed to be independent and Gaussian distributed and hence, unbounded. This channel uncertainty model is often associated with channel estimation errors due to insufficient pilots among other imperfections [125, 129]. Because of the unbounded nature of this error model, even the robust design approach cannot guarantee that the QoS constraints are always satisfied. Hence, the outage probability-based robust design is often exploited to

guarantee reliability up to a certain probability. Therefore, the reflected channel according to the UEM can be expressed as

$$\mathbf{Q}_i = \hat{\mathbf{Q}}_i + \Delta\mathbf{Q}_i, \forall i \in \mathcal{K}, \tag{6.5}$$

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

$$\beta_i^2 = \lambda^2 ||\hat{\mathbf{q}}_i||_2^2, \forall i \in \mathcal{K}, \tag{6.6}$$

where $\hat{\mathbf{q}}_i = \text{vec}(\hat{\mathbf{Q}}_i) \in \mathbb{C}^{MN\text{x}1}$.

Note that the variance of the error term is identical under both error models. However, the error is bounded by e_i in the BEM while the channel uncertainty is unbounded in the case of the UEM. Moreover, these aforementioned differences will have a significant impact on the optimization problem as explained in later sections of the work. Next, the SINR and achievable rates are defined for the considered system model.

6.1.2 Achievable Rates

At the receiver side, the BS uses the receive beamforming technique to decode the signal of each UE by enhancing the quality of the received signal. Therefore, the extracted signal of UE_i with receiver beamforming $\mathbf{u}_i$ can be expressed as

$$\begin{aligned} \mathbf{u}_i^{\text{H}}\mathbf{y} &= \mathbf{u}_i^{\text{H}} \sum_{k=1}^{K} \tilde{\mathbf{h}}_k \sqrt{p_k} s_k + \mathbf{u}_i^{\text{H}}\mathbf{z}, \forall i,k \in \mathcal{K}, \\ \mathbf{u}_i^{\text{H}}\mathbf{y} &= \mathbf{u}_i^{\text{H}}\tilde{\mathbf{h}}_i \sqrt{p_i} s_i + \sum_{k\neq i}^{K} \mathbf{u}_i^{\text{H}}\tilde{\mathbf{h}}_k \sqrt{p_k} s_k + \mathbf{u}_i^{\text{H}}\mathbf{z}, \forall i,k \in \mathcal{K}. \end{aligned} \tag{6.7}$$

In order to remove the interference from other UEs, the BS performs $K-1$ SIC operations according to the NOMA principle. Therefore, deciding a decoding order is crucial in NOMA

systems [172]. Throughout this work, the UEs are assumed to be ordered based on their direct channel strengths between the BS and UEs. Therefore, the UEs are ordered as follows:

$$||\mathbf{h}_1||_2 \geq ||\mathbf{h}_2||_2 \geq ... \geq ||\mathbf{h}_K||_2, \tag{6.8}$$

where UE_1, UE_K are the UEs with the strongest and weakest direct channels, respectively. Therefore, the variable $\varsigma = \{1,2,...,K\}$ is defined as the decoding order where the BS decodes UE_1's signal while treating the other $K-1$ signals as interference. The BS then subtracts UE_1's signal before repeating the same procedure for UE_2's signal and so on. Hence, in the ideal case of $K-1$ successful SIC operations, the BS decodes UE_k's signal free from any interference. Hence, the selected decoding embraces fairness between the UEs in the system. Note that the decoding order does not affect the sum rate in UL NOMA systems, it does however impact individual UE rates [173–175]. After deciding a decoding order and without the loss of generality, the SINR expression for UE_i's signal can written as

$$\gamma_i = \frac{|\mathbf{u}_i^{\text{H}}\tilde{\mathbf{h}}_i|^2 p_i}{\sum_{k=1}^{i-1} \left|\mathbf{u}_i^{\text{H}}(\Delta\mathbf{Q}_k\mathbf{v})\right|^2 p_k + \sum_{k=i+1}^{K} |\mathbf{u}_i^{\text{H}}\tilde{\mathbf{h}}_k|^2 p_k + |\mathbf{u}_i^{\text{H}}\mathbf{z}|^2}, \tag{6.9}$$

where the term $\sum_{k=1}^{i-1} \left|\mathbf{u}_i^{\text{H}}(\Delta\mathbf{Q}_k\mathbf{v})\right|^2 p_k$ is the SIC residual that results from imperfect CSI at the BS. Note that unlike in [174, 176] where perfect CSI is assumed, the SINR expression in (6.9) has significant implications on the resource allocation strategy and the resulting system performance as explained in the next sections.

Based on the SINR expressions, the achievable rate for UE_i's signal is expressed as

$$R_i = B log_2(1+\gamma_i), \forall i \in \mathcal{K}, \tag{6.10}$$

where B is the channel bandwidth.

6.2 Problem Formulation

The aim of this work is to develop a resource allocation framework for maximizing the long-term EE of IRS-assisted UL NOMA systems under channel uncertainty. The EE of the consid-

ered system can be defined as

$$EE = \frac{B\sum_{i=1}^{K} log_2(1+\gamma_i)}{\sum_{i=1}^{K} \zeta' p_i + P_c}, \forall i \in \mathcal{K}, \tag{6.11}$$

where ζ' is a scalar that represents the energy inefficiency of the power amplifier at the UE, and P_c represents the power loss at the BS and the IRS which is from the circuit hardware power consumption.

Based on the definition in (6.11), the long-term EE maximization problem under the BEM can be written as follows:

$$\underset{\mathbf{U},\mathbf{v},\mathbf{p}}{\text{maximize}} \quad \mathbb{E}\left\{ \sum_{t=1}^{\infty} \delta^{t-1} \frac{\sum_{i=1}^{K} R_i^t(\mathbf{U},\mathbf{v},\mathbf{p})}{\sum_{i=1}^{K} \zeta' p_i^t + P_c} \Big| \pi, s^t \right\} \tag{6.12a}$$

subject to

$$\frac{|\mathbf{u}_i^{\mathrm{H}} \tilde{\mathbf{h}}_i|^2 p_i}{\sum_{k=1}^{i-1} \left|\mathbf{u}_i^{\mathrm{H}} (\Delta \mathbf{Q}_k \mathbf{v})\right|^2 p_k + \sum_{k=i+1}^{K} |\mathbf{u}_i^{\mathrm{H}} \tilde{\mathbf{h}}_k|^2 p_k + \sigma_i^2} \geq 2^{R_i^{min}} - 1, \forall ||\Delta \mathbf{Q}_i||_F \leq e_i, \forall i \in \mathcal{K}, \tag{6.12b}$$

$$p_i \leq P_{max}, \forall i \in \mathcal{K}, \tag{6.12c}$$

$$|\Upsilon_m|^2 = 1, 0 \leq \theta_m \leq 2\pi, m = 1, ..., M, \tag{6.12d}$$

where the expectation operator in (6.12a) indicates that the aim is to maximize the long-term EE of the system over multiple time-steps given a policy π and a state s^t. The policy and the state concepts are explained in detail in the DRL section discussed next. $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_K], \mathbf{p} = [p_1, ..., p_K]$ represent the beamforming matrix and the UEs power allocation vector, respectively. The constraint in (6.12b) represents the QoS requirement under channel uncertainties according to the BEM, while the constraint in (6.12c) limits the maximum transmit power at each UE. Finally, the constraints in (6.12d) represent the amplitude and phase shift feasible region.

Similarly, the long-term EE maximization under the UEM can be expressed as follows:

$$\underset{\mathbf{U},\mathbf{v},\mathbf{p}}{\text{maximize}} \quad \mathbb{E}\left\{ \sum_{t=1}^{\infty} \delta^{t-1} \frac{\sum_{i=1}^{K} R_i^t(\mathbf{U},\mathbf{v},\mathbf{p})}{\sum_{i=1}^{K} \zeta' p_i^t + P_c} \Big| \pi, s^t \right\} \tag{6.13a}$$

subject to

$$p_i' \triangleq \Pr\{ \frac{|\mathbf{u}_i^{\mathrm{H}} \tilde{\mathbf{h}}_i|^2 p_i}{\sum_{k=1}^{i-1} \left|\mathbf{u}_i^{\mathrm{H}} (\Delta \mathbf{Q}_k \mathbf{v})\right|^2 p_k + \sum_{k=i+1}^{K} |\mathbf{u}_i^{\mathrm{H}} \tilde{\mathbf{h}}_k|^2 p_k + \sigma_i^2} \geq 2^{R_i^{min}} - 1 \} \geq \eta, \ \forall i \in \mathcal{K}, \tag{6.13b}$$

$(6.12c), (6.12d),$

where $\eta \in (0, 1]$ is the non-outage probability. Note that since channel error is unbounded under UEM, it is not possible to determine the design parameters to always meet the QoS constraints, and hence, the outage-constrained problem in (6.13a). Note that the considered optimization problem under both channel uncertainty models is non-trivial and extremely challenging to determine a feasible solution efficiently due to the following factors

- The objective function is not jointly convex in the optimization variables. Moreover, the optimization variables are coupled in the SINR expressions.
- The expectation operator prevents the approximation of the objective function using a closed-form expression making it more challenging to optimize efficiently.
- The problems in (6.12a) and (6.13a) are NP-hard [177] and therefore cannot be solved directly using conventional optimization methods.
- Under the UEM, it is well-known that the outage constraint in (6.13b) does not admit a closed-form solution [153].

Note that most of the EE problems studied in the literature consider the static optimization problem where the EE is maximized for a particular set of channels [174, 178]. However, this does not accurately reflect the average system performance under channel uncertainties [147]. Moreover, it is much more challenging to solve the dynamic EE maximization problem using conventional optimization approaches. Therefore, in the next section, a novel DRL-based framework that learns to efficiently solve the joint robust design problem is proposed.

6.3 Proposed Solution

In the considered robust design problem, there are three design parameters, namely the receive beamforming vectors, the UE power allocation and the IRS phase shifts. Since the objective

function is not jointly convex in terms of the optimization variables, the conventional approach is to divide the original robust design problem into three subproblems, wherein each subproblem an optimization variable is optimized while fixing the other two design variables leading to an iterative algorithm and often a suboptimal solution [174, 176]. In addition, this conventional approach often results in high-complexity algorithms that require at least multiple iterations to converge to an acceptable solution. This in turn impacts the latency of the system. Therefore, such high-complexity algorithms are impractical for latency-sensitive applications. Moreover, the conventional approaches cannot be directly applied to the considered robust design problem since the objective function does not have a closed-form expression.

To reduce the complexity of the problem, the beamforming design is tackled using the well-established MMSE-SIC receiver which is the optimal receiver under the perfect CSI assumption [179]. Using the closed-form expression of the MMSE-SIC, the achievable system sum rate is maximized. However, since perfect CSI is not available at the BS in the considered problem, the optimality of the MMSE-SIC receiver can no longer be guaranteed. Furthermore, the remaining power allocation and IRS phase shifts are optimized jointly using the RL framework. In particular, since the beamforming vectors $\mathbf{U}$ are designed based on the estimated channels, the robust design problem is realized through the other two design variables (phase shifts and power allocation).

6.3.1 The MMSE-SIC Receiver

The MMSE-SIC receiver utilizes the MMSE detector followed by SIC to completely eliminate the interference in an ideal system [180]. One of the main advantages of the MMSE-SIC is the closed-form expression of receive beamforming vectors for maximizing the system rate which is expressed as [35, 179]

$$\mathbf{u}_i' = \left(N_0\mathbf{I}_N + \sum_{k=i+1}^{K} p_k \hat{\tilde{\mathbf{h}}}_k \hat{\tilde{\mathbf{h}}}_k^{\mathrm{H}}\right)^{-1} \hat{\tilde{\mathbf{h}}}_i, \forall i,k \in \mathcal{K}, \tag{6.14}$$

where $\hat{\tilde{\mathbf{h}}}_i$ is the estimated channel of the true channel $\hat{\mathbf{h}}_i$. Then, the resulting receive beamforming vectors are normalized as follows:

$$\mathbf{u}_i = \frac{\mathbf{u}'_i}{||\mathbf{u}'_i||_2}, \forall i \in \mathcal{K}, \tag{6.15}$$

which ensures that $||\mathbf{u}_i||_2 = 1$. Note that since the error bound is known under the BEM, this information can be utilized to enhance the accuracy of (6.14). In particular, the true and estimated channels for UE_i, respectively, are expressed as

$$\tilde{\mathbf{h}}_i = (\hat{\mathbf{Q}}_i + \Delta\mathbf{Q}_i)\mathbf{v} + \mathbf{h}_i, \forall i \in \mathcal{K}, \tag{6.16}$$

$$\hat{\tilde{\mathbf{h}}}_i = \hat{\mathbf{Q}}_i\mathbf{v} + \mathbf{h}_i, \forall i \in \mathcal{K}. \tag{6.17}$$

However, since $||\Delta\mathbf{Q}_i||_F \leq e_i$, where e_i is known at the BS, $||\Delta\mathbf{Q}'_i||_F = e_i$ is set, which corresponds to the worst-case error. Then, the worst-case estimated effective channel is expressed as

$$\hat{\tilde{\mathbf{h}}}'_i = (\hat{\mathbf{Q}}_i + \Delta\mathbf{Q}'_i)\mathbf{v} + \mathbf{h}_i, \forall i \in \mathcal{K}, \tag{6.18}$$

which is known at the BS under the BEM assumption. Therefore,(6.14) is rewritten as

$$\mathbf{u}'_{\text{BEM},i} = \left(N_0\mathbf{I}_N + \sum_{k=i+1}^{K} p_k \hat{\tilde{\mathbf{h}}}'_k \hat{\tilde{\mathbf{h}}}'^{\text{H}}_k\right)^{-1} \hat{\tilde{\mathbf{h}}}'_i, \forall i,k \in \mathcal{K}, \tag{6.19}$$

$$\mathbf{u}_{\text{BEM},i} = \frac{\mathbf{u}'_{\text{BEM},i}}{||\mathbf{u}'_{\text{BEM},i}||_2}, \forall i \in \mathcal{K}, \tag{6.20}$$

where $\mathbf{u}_{\text{BEM},i}$ is the normalized worst-case receive beamforming vector for UE_i's signal.

With the derived receive beamforming vectors, the remaining robust design problem can be reformulated as

$$\begin{aligned} \underset{\mathbf{v},\mathbf{p}}{\text{maximize}} \quad & \mathbb{E}\left\{\sum_{t=1}^{\infty} \delta^{t-1} \frac{\sum_{i=1}^{K} R_i^t(\mathbf{U},\mathbf{v},\mathbf{p})}{\sum_{i=1}^{K} \zeta' p_i^t + P_c} \middle| \pi, s^t\right\} \\ \text{subject to} \quad & \\ & (6.12b) \text{ or } (6.13b), (6.12c), (6.12d). \end{aligned} \tag{6.21}$$

Unfortunately, the optimization problem is still non-convex due to the coupled optimization variables. Therefore, a DRL-based framework for joint optimization of $\mathbf{v}$ and $\mathbf{p}$ is proposed in the following section.

6.3.2 DRL-Based Joint Design Approach

It is well-established that the problem of optimizing a system objective under uncertainty can be formulated as an MDP [162]. Moreover, the RL framework is one of the useful techniques that has been developed to solve problems formulated as MDPs, especially in model-free systems [91]. The RL framework consists of two entities; the agent and the environment. Given a system state $\mathbf{s}$, the agent takes an action $\mathbf{a}$ and the environment provides a reward signal r and a next state $\mathbf{s}'$. Moreover, the agent aims to maximize its reward by taking actions that yield higher rewards. Using this simple but highly effective concept, DRL agents have been applied to solve challenging problems, especially joint design problems, in the wireless communications domain. In particular, DRL agents with the actor-critic architecture have been able to solve extremely complicated problems with some interesting results [91].
In order to develop a DRL-based agent that learns how to efficiently solve the robust design problem in (6.21), the problem must be recast into an RL environment. Moreover, to reformulate the problem into a standard RL environment format, three entities must be defined, namely the state space, the actions space and the reward function. Therefore, those entities are defined as follows:

- The actions space $\mathbf{a}^t$: the actions space defines the decision variables in which the DRL agent is applied to optimize. Therefore, the actions space of the RL environment is selected to be the decision variables of the optimization problem in (6.21). Hence, this is formally expressed as

$$\mathbf{a}^t = [\bar{p}_1^t, ..., \bar{p}_K^t, \bar{v}_1^t, ..., \bar{v}_M^t]^T, \tag{6.22}$$

  Since fully connected DNNs that support real numbers only are being used, each IRS phase shift element is represented by two real values which capture the real and the imaginary parts [134]. Therefore, the dimension of the action space $\mathbf{a} \in \mathbb{R}^{K+2M}$.

- The state space $\mathbf{s}^t$: the state space defines the important features of the considered problem. Therefore, the previous action taken by the agent is selected as part of the state space to help the agent evaluate itself. Also, since the problem in (6.21) is subject to QoS

constraints, the achieved rates of the previous time-step are selected as part of the state space as well. Finally, the previous reward (which is defined next) is the added to state space. Therefore, the state space with the above elements can be formalized as follows:

$$\mathbf{s}^t = [\mathbf{a}^{t-1}, R_1^{t-1}, ..., R_K^{t-1}, r^{t-1}]^T, \tag{6.23}$$

and hence, the dimension of the state space is $\mathbf{s}^t \in \mathbb{R}^{2K+2M+1}$. However, since the dynamic-channels scenario will be considered where the channel condition between the UE and the BS changes drastically, the state space is modified by adding the following additional elements according to the assumed error model:

$$\begin{aligned} \mathbf{s}^t_{\text{dy,BEM}} &= [\mathbf{a}^{t-1}, R_1^{t-1}, ..., R_K^{t-1}, r^{t-1}, e'_1, ..., e'_K]^T, \\ \mathbf{s}^t_{\text{dy,UEM}} &= [\mathbf{a}^{t-1}, R_1^{t-1}, ..., R_K^{t-1}, r^{t-1}, \varepsilon_1, ..., \varepsilon_K]^T, \end{aligned} \tag{6.24}$$

where $e'_i, \varepsilon_i, \forall i \in \mathcal{K}$ are the normalized error bounds and the normalized estimated channel variances for the BEM and UEM, respectively. Therefore, the state space in (6.23) is used when the fixed-channels scenario is considered irrespective of the assumed error model, while the state spaces in (6.24) are used when the dynamic-channels scenario is considered. A more elaborate explanation on the differences between the two scenarios is provided in the simulation section. Therefore, the dimension of the state space for dynamic channel scenarios $\mathbf{s}^t_{\text{dy},\tau} \in \mathbb{R}^{3K+2M+1}, \tau \in \{\text{BEM, UEM}\}$.

- The reward function r^t: The reward function is the only feedback signal that evaluates the utility/cost of the agent's action. Therefore, designing an accurate and robust reward signal is vital in the RL framework. The logic behind the reward function design is as follows; the reward function must return a positive value in case the action taken by the agent maximizes the objective function and does not violate any of the constraints. On the other hand, the reward function should return a negative value in case the agent's action violates any of the constraints. Therefore, the agent is rewarded positively according to the following function:

$$r^t = \varpi E E^t, \tag{6.25}$$

where $\varpi > 0$ is used to scale the positive reward by the preferred value. On the other hand, the agent is punished by the following negative reward function:

$$r^t = \varpi' \sum_{i=1}^{K} min(R_i^t - R_i^{min}, 0), \forall i \in \mathcal{K}, \tag{6.26}$$

where (6.26) is always negative, and $\varpi' > 0$ is used to scale the negative reward value.

Note that since RL agents are inherently oblivious to the constraints of the optimization problem, actions normalization is often utilized to ensure that all actions taken by the agent fall within the feasible region. Therefore, the transmit powers vector generated by the agent is limited to the range $(0, 1]$, and the output powers vector is normalized as follows:

$$\mathbf{p}^t = P_{max}\bar{\mathbf{p}^t}, \tag{6.27}$$

where $\bar{\mathbf{p}^t} = [\bar{p}_1^t, ..., \bar{p}_K^t]^T$ is the powers vector obtained from the policy network while $\mathbf{p^t} = [p_1^t, ..., p_K^t]^T$ is the feasible powers vector that satisfies (6.12c). Similarly, the amplitude of the IRS phase shift elements are normalized as

$$\Upsilon_m = \frac{\bar{v}_m}{|\bar{v}_m|}, m = 1, .., M, \tag{6.28}$$

where $|v_m| = 1$ is guaranteed according to (6.12d).

Another important aspect that relates to the robust design strategy in the proposed reformulation is that, unlike conventional optimization algorithms in which the worst-case/outage constraints in (6.12b) and (6.13b) are explicitly enforced, these constraints are implicitly included in the environment. In particular, the receive beamforming vectors are designed based on the IRS phase shifts and transmit power values generated by the agent based on the estimated channels, not the true channels. Furthermore, the SINR expressions and the achievable rates are evaluated using the true channels. Therefore, if the agent's policy is not robust, it will fail to satisfy the required QoS constraints and will be punished with a negative reward. Hence, the agent learns to move away from non-robust policies since it aims to maximize its long-term reward. Moreover, explicitly including the aforementioned constraints assumes that the DRL agent has some kind of expert knowledge on the problem such that it is capable of generating competitive

solutions from the start which is not the case. Consequently, the derived policy's robustness is also hyperparameterized in the proposed design.

6.3.3 The SAC DRL Agent

In this work, a robust joint design algorithm based on the SAC DRL agent is proposed [116]. The SAC is the state-of-the-art in DRL which combines the stability and inherent exploration ability of on-policy actor-critic agents with the sample efficiency and accuracy of off-policy actor-critic agents. Therefore, the SAC is an off-policy DRL agent that optimizes a stochastic policy. Moreover, the SAC has been shown to outperform the PPO, the DDPG, and the TD3 thanks to its superior exploration policy [116].

The SAC agent comprises two entities; the actor or the policy $\mu_{\psi}(\mathbf{a}|\mathbf{s})$ DNN which is responsible for taking the actions, and the critic DNN $\mathcal{Q}_{\phi}(\mathbf{s},\mathbf{a})$. Note that in this work, the version of the SAC which uses two critic networks is adopted to reduce the impact of the overestimation bias problem in off-policy methods. As the name implies, the critic network criticizes the actions taken by the policy network. The goal is to train the SAC policy network until it converges to the optimal policy such that

$$\pi^{*}(\mathbf{s}) = \mathbf{a}^{*}, \tag{6.29}$$

where for any given system state $\mathbf{s}$, the agent generates the optimal action a^{*}. However, since the critic DNN is the one responsible for estimating the value of the action in the SAC agent, the critic must be trained first. Therefore, the $\mathcal{Q}$-value $\mathcal{Q}(\mathbf{s},\mathbf{a})$ update for being in a state $\mathbf{s}$ and taking action $\mathbf{a}$, which is known as the Bellman equation, is expressed as [116]

$$q_{\pi}(\mathbf{s},\mathbf{a}) = r + \delta\left(q_{\pi}(\mathbf{s},\mathbf{a}) + \kappa'\mathcal{H}\big(\pi(\mathbf{a}|\mathbf{s})\big)\right), \tag{6.30}$$

where r is the reward, δ is the discount factor which represents the current value of the future rewards, $\mathcal{H}\big(\pi(\mathbf{a}|\mathbf{s})\big)$ is the entropy term and κ' is the entropy coefficient. Note that including the entropy term in the Bellman equation is one of the novelties of the SAC agent. This is because maximizing the $\mathcal{Q}$-value now requires the joint optimization of both the reward and the entropy term. In practical terms, this encourages the SAC agent to sufficiently explore the

environment before converging to the final policy. This is crucial in DRL agents since unlike the deep learning framework, the DRL agent does not know the optimal solution to the problem it is trying to solve. The critic DNN in the SAC agent is similar to that of the TD3 agent. Therefore, the SAC critic DNN also uses the replay buffer and the target network concepts to stabilize the learning and training process. The replay buffer $\mathcal{D}$ is a memory in which previous experiences (called tuples) are saved. In particular, the replay buffer consists of a large number of tuples in the form $\{\mathbf{s}, \mathbf{a}, r, \mathbf{s}^{t+1}\}$. The replay buffer is then sampled during training using a mini-batch of size $\mathcal{B}$ to decorrelate the training data. In addition, the target critic network $\mathcal{Q}_{\phi'}(\mathbf{s}^{t+1}, \mathbf{a}^{t+1})$ which is a delayed version of the main critic network is used to stabilize the learning process by fixing the target when training the critic. Therefore, given a single tuple, the critic's target is calculated as [116]

$$y'(r, \mathbf{s}^{t+1}) = r + \delta \left[\min_i \mathcal{Q}_{\phi_i'}\left(\mathbf{s}^{t+1}, \mu_\psi(\mathbf{s}^{t+1})\right) - \kappa' log \mu_\psi(\mathbf{a}^{t+1} | \mathbf{s}^{t+1}) \right], \tag{6.31}$$

where $i = 1, 2$ refers to the first and the second target critic DNN, respectively. Note that, unlike the TD3 and DDPG agents, the SAC agent uses the same current policy network to generate $\mathbf{a}^{t+1}$ since the SAC agent does USE a target actor-network. After calculating the target, the SAC critics are trained by minimizing an MSE objective as follows:

$$L(\phi_i, \mathcal{B}) = \mathop{\mathbb{E}}_{\{\mathbf{s}, \mathbf{a}, r, \mathbf{s}^{t+1}\} \sim \mathcal{B}} \left[\left(Q_{\phi_i}(\mathbf{s}, \mathbf{a}) - y(r, \mathbf{s}^{t+1}) \right)^2 \right], i = 1, 2. \tag{6.32}$$

where the expectation operator indicates that optimizing the objective is carried out across the whole batch of samples instead of a single tuple.

Now that the critics' training step is completed, the policy network will be discussed. The SAC agent uses a stochastic actor, i.e., it uses a Gaussian policy that generates the mean μ_ϕ and the standard deviation σ'_ϕ. Hence, to get the final action values, the aforementioned distribution is sampled. However, since the Gaussian distribution is unbounded, a *Tanh* function is used to squash or bound the sampled value and then rescale it if needed. Therefore, the SAC agent's action can be expressed as

$$\mathbf{a} = tanh(\varepsilon), \varepsilon \sim \mathcal{N}\left(\mu_\psi(\mathbf{s}), \sigma'_\psi(\mathbf{s})\right). \tag{6.33}$$

Algorithm 5 The proposed SAC algorithm

Initialise: actor and main critic DNNs ψ, ϕ_1, ϕ_2, empty replay buffer $\mathcal{D}$
Set critic target DNNs equal to the main critic networks: $\psi' \leftarrow \psi, \phi_1' \leftarrow \phi_1, \phi_2' \leftarrow \phi_2$,
repeat
    Observe initial or current state s and feed it to the policy network to obtain the action a according to (6.33)
    Execute a in the environment
    Obtain reward r and next state s^{t+1} from the environment, save the tuple s, a, r, s^{t+1} to $\mathcal{D}$
    if $time\_to\_update == True$ **then**
        while $j \leq Updates$ **do**
            Randomly sample $\mathcal{D}$ using a batch of size $|\mathcal{B}|$
            Compute the corresponding target values for the experiences in $\mathcal{B}$ using (6.31)
            Train the two critics using the stochastic gradient descend algorithm to minimize the MSE objective according to (6.32)
            Train the policy network using (6.34)
            Update the critic target DNNs according to (6.35)
            $j = j + 1$
        end while
    end if
until Consistent high average reward is achieved
Output: Optimal policy $\pi^*(s; \phi)$

Moreover, the SAC agent's policy network is trained by adjusting its hyperparameters to maximize the $\mathcal{Q}$-values for the sampled states. This concept is formalized as follows:

$$J(\phi,\mathcal{B}) = \underset{\{\mathbf{s}\sim\mathcal{B},\mathbf{a}\sim\pi\}}{\mathbb{E}}\Big[Q_\phi(\mathbf{s},\mathbf{a}) - \kappa' log\mu_\psi(\mathbf{a}|\mathbf{s})\Big], i = 1,2. \tag{6.34}$$

After training the policy network, the critics' target networks are then updated. In the DRL literature, there are two ways to update the target networks in off-policy actor-critic agents, namely smoothing and periodic updates. In this work, the smoothing targets update method is adopted to stabilize the learning process. In particular, the target networks are updated according to the following expression

$$\phi_i' = \kappa\phi_i + (1-\kappa)\phi_n', \, i = 1,2, \tag{6.35}$$

where $0 < \kappa \leq 1$ is the target smoothing factor. It is obvious from (6.35) that larger κ values correspond to faster updates of the target DNNs which is not recommended, especially when the agent is applied to a complicated environment.

Note that despite optimizing a stochastic policy, the SAC agent does converge to an approximately deterministic policy given enough training steps [116]. Therefore, the SAC agent employs a superior exploration strategy to prevent the agent from getting stuck in a bad policy during the earlier training episodes. However, after sufficient exploration of the environment, the agent starts converging towards the optimal policy based on its previous experiences. Finally, Algorithm 5 summarizes the interactions between the actor and the critics during the SAC agent's training.

Now that the inner workings of the SAC agent are explained, Algorithm 6 summarizes the important step for the developed SAC-based joint design. Note that Algorithm 6 presents the process of a single training step to explain how different parts of the joint design are combined. Practical agent training is always carried out on a batch of samples as highlighted in Algorithm 5. In addition, despite using the closed-form expression to calculate $\mathbf{u}_i, \forall i \in \mathcal{K}$, Algorithm 6 indirectly contributes to the enhancement of the receive beamforming performance through the optimization variables. Therefore, the output of the proposed algorithm also includes $\mathbf{U}$.

Algorithm 6 The SAC-based Joint-Design Algorithm

Input: Estimated channels $\hat{\mathbf{Q}}_i, \mathbf{h}_i, B, P_{max}, M, N_0, \lambda$, and probability value for Γ_{2MN}^{-1}
Initialise: Agent parameters $\phi, \varphi_n, \varphi_n^-, \mathcal{D}, \mathcal{B}$ and the environment
Set: $\phi_1' \leftarrow \phi_1, \phi_2' \leftarrow \phi_2$
while $Episode \leq Total\_Episodes$ **do**
    Set initial state such that $\mathbf{s}_{initial}^t = \mathbf{1}$
    Randomly sample the training channels set and compute $\Delta\mathbf{Q}_i$ according to the adopted error model
    while $Step \leq Total\_Steps$ **do**
        feed initial state to the policy network to obtain action $\mathbf{a}^t$ according to (6.33)
        Execute $\mathbf{a}^t$ in the environment
        Extract feasible vectors $\mathbf{v}^t$ and $\mathbf{p}^t$ from the actions vector according to (6.27) and (6.28)
        Compute the final estimated channel $\hat{\tilde{\mathbf{h}}}_i, \forall i \in \mathcal{K}$ using the optimized IRS vector $\mathbf{v}^t$ according to (6.17) or (6.18) for the UEM and BEM, respectively
        Compute $\mathbf{u}_i$ or $\mathbf{u}_{\text{BEM},i}$, $\forall i \in \mathcal{K}$, using the optimized $\mathbf{v}^t, \mathbf{p}^t$ according to (6.15) or (6.20)
        Build the effective true channels $\tilde{\mathbf{h}}_i, \forall i \in \mathcal{K}$ by adding a random error term according to the appropriate error model using (6.16)
        Evaluate the SINRs in (6.9) using the true channels and calculate the achieved rates according to (6.10)
        if $R_i^t \geq R_i^{min}, \forall i \in \mathcal{K}$ **then**
            Use reward function in (6.25)
        else
            Use reward function in (6.26)
        end if
        Save the tuple $\{\mathbf{a}^t, \mathbf{s}^t, r, \mathbf{s}^{t+1}\}$ to the replay buffer $\mathcal{D}$
        Train the critic DNNs using (6.31) and (6.32)
        Train the policy network using (6.34)
        Update the critic target networks using (6.35)
        $Step = Step + 1$
        Set $\mathbf{s}^t = \mathbf{s}^{t+1}$
    end while
    Set $Episode = Episode + 1$
end while
Output: $\mathbf{U}^*, \mathbf{p}^*, \mathbf{v}^*$

6.3.4 Computational Complexity Analysis

There are two stages in which the computational complexity of machine learning algorithms can be evaluated, namely the offline training complexity and the online deployment complexity. However, since the agent only needs to be trained once, it is assumed that the offline training complexity can be afforded in practical implementations and focus mainly on the deployment complexity.

The big $\mathcal{O}(.)$ notation is one of the most widely adopted measures for the worst-case complexity of an algorithm as a function of its input size. This notation will be used to characterize the worst-case complexity of the proposed algorithm. For the trained SAC agent, only the policy network will be used to take action during the deployment stage. Therefore, the complexity of the trained SAC agent is equal to the complexity of a single forward pass through the policy network. Moreover, the forward pass can be mathematically formulated as $\Psi+1$ vector-matrix multiplications where the size of these product operations are determined by the input layer and output layer sizes, and the number of neurons in each hidden layer $\aleph$. In addition, Each layer in the network requires an activation, except for the input layer. Therefore, the total worst-case complexity can be written as $\mathcal{O}\Big(T\big(\aleph\cdot\mathbf{Card}(\mathbf{s}^t)+\aleph^2+\mathbf{Card}(\mathbf{a}^t)\cdot\aleph+\Psi\cdot\aleph+\mathbf{Card}(\mathbf{a}^t)+(KN)^4\big)\Big)$, where T is included to highlight that the action space is part of the state space. Note that $T=2$ is considered in this work to keep the computational complexity of the proposed algorithm to a minimum. In addition, $\mathbf{Card}(\mathbf{s}^t),\mathbf{Card}(\mathbf{a}^t)$ represent the cardinality of the state and actions spaces, respectively, and the linear terms represent the activation operations for the hidden and output layers, respectively. The term $(KN)^4$ represents the complexity of calculating the receive beamforming vectors in $\mathbf{U}$. Moreover, since $\mathbf{Card}(\mathbf{s}^t)>\mathbf{Card}(\mathbf{a}^t)$ always hold in the proposed algorithm, and that $\mathbf{Card}(\mathbf{s}^t)=3K+2M+1$, then, the worst-case complexity can be approximated to $\mathcal{O}\big(\aleph\cdot(3K+2M+1)+(KN)^4\big)$ as a function of the two dominant terms. Therefore, the final worst-case complexity of the proposed algorithm is $\mathcal{O}\big((KN)^4\big)$, which is completely determined by the MMSE-SIC expression.

Comparing the derived complexity of the proposed SAC-based design to other algorithms in the literature, the complexity of the algorithm in [174] is $\mathcal{O}\big(T((KN)^4+T_1K^2+T_2(K+M^2)^{3.5}+$

$JM)\big)$ with the worst-case complexity of $\mathcal{O}\big(T(T_2M^7)\big)$ which is completely determined by the number of IRS elements M and the number of iterations T and T_2. Consequently, it is evident that such complexity is relatively expensive, particularly for a large number of IRS elements. The proposed SAC-based algorithm on the other hand scales linearly with M, making it much faster compared to the expensive SDP-based algorithm [174].

6.4 Agent Architecture and Simulation Results

In this section, the details of the agent architecture of the proposed agent and the hyperparameters used to train it are proposed. In addition, the system parameters are highlighted and simulation results that demonstrate the competitive performance of the proposed SAC-based robust design algorithm are presented.

6.4.1 Agent Architecture

The SAC agent uses the actor-critic architecture illustrated in Figs. 6.2 and 6.3. In this work, the developed SAC agent comprises two critics and one actor DNNs. In addition, the *ReLU* function $f(x) = \max(0,x)$ is used to activate the hidden layers in both the actor and the critics DNNs. The actor's DNN has two outputs for each action which correspond to a distribution with a mean and a standard deviation values. In order to bound the actions generated by the actor-network, a *Tanh* layer is used. Moreover, the ADAM optimizer is used to optimize the DNNs during the training [141]. Furthermore, the number of hidden nodes is set to 128 and 256 for the fixed and dynamic channels, respectively. Table 6.1 summarizes the hyperparameters used to train the SAC agent.

Note that the SAC agent's performance is a function of the chosen hyperparameters, therefore, the results presented in the simulation section reflect the SAC agent's performance with the selected hyperparameters, not the optimal performance of the SAC agent in general.

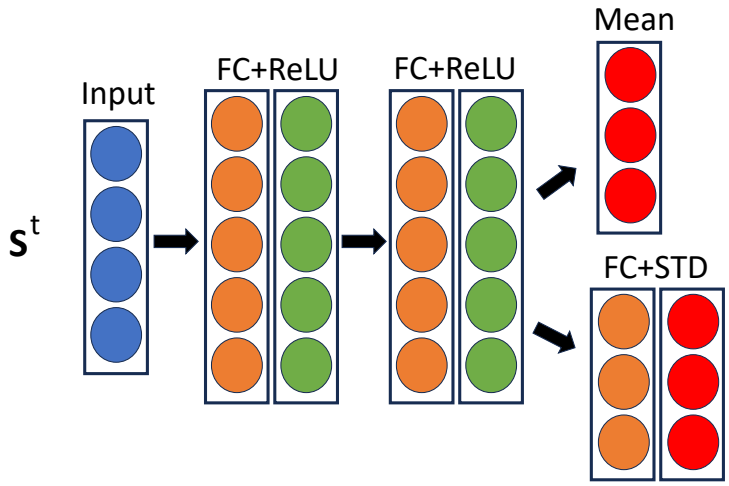

Figure 6.2: The actor's DNN.

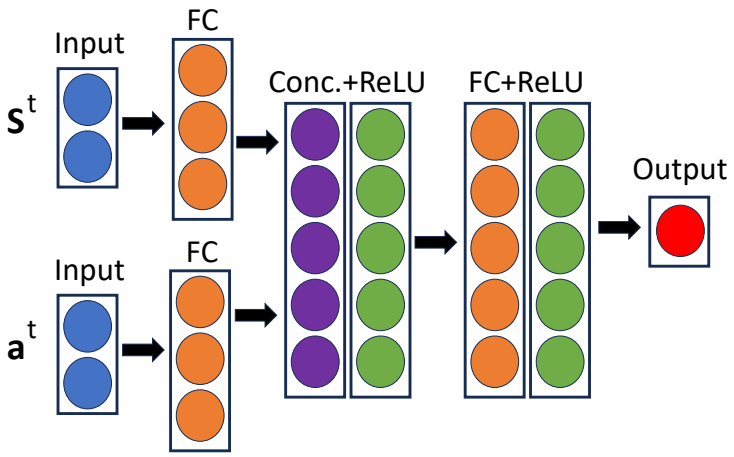

Figure 6.3: The critic's DNN.

Table 6.1: Hyperparameters of the SAC agent.

Hyperparameter	Value
Actor learning rate (fixed/dynamic channels)	0.0005/0.0001
Critics learning rate (fixed/dynamic channels)	0.0007/0.0004
Entropy coefficient (κ')	0.05
Discount factor (δ)	0.99
Policy update frequency	1
Replay buffer size ($\mathcal{D}$)	$100,000$
Minibatch size ($\mathcal{B}$)	128
Smoothness factor (κ) (fixed/dynamic channels)	$0.0002/0.00005$
Number of episodes, time-steps (fixed-channels)	$200,300$
Number of episodes, time-steps (dynamic-channels)	$600,500$

6.4.2 System Parameters

An IRS-assisted UL NOMA system as illustrated in Figure 6.1 is considered. In particular, both the large and small-scale fading of the channels are taken into account. Moreover, both $\mathbf{g}_i$ and $\mathbf{H}$ are assumed to have line-of-sight components and therefore modelled using the Rician distribution as follows:

$$\mathbf{g}_i = \frac{1}{\sqrt{d_i^{\iota_{UE\to IRS}}}}\left(\sqrt{\frac{K'}{1+K'}}\mathbf{g}_{LoS} + \sqrt{\frac{1}{1+K'}}\mathbf{g}_{nLoS}\right), \tag{6.36}$$

$$\mathbf{H} = \frac{1}{\sqrt{d_{irs}^{\iota_{IRS\to BS}}}}\left(\sqrt{\frac{K'}{1+K'}}\mathbf{H}_{LoS} + \sqrt{\frac{1}{1+K'}}\mathbf{H}_{nLoS}\right), \tag{6.37}$$

where $\iota_{UE\to IRS}$ and $\iota_{IRS\to BS}$ are the path-loss exponents between the UE and the IRS, and between the IRS and the BS, respectively, and $K' = 1$ is the Rician factor. Moreover, d_i and d_{irs} represent the distance between the UE and the IRS and the distance between the IRS and the BS, respectively. On the other hand, the direct links between the UEs and the BS are assumed

to be Rayleigh fading and are expressed as

$$\mathbf{h}_i = \frac{h'_i}{\sqrt{d_{id}^{\iota_{UE\rightarrow BS}}}}, \tag{6.38}$$

where d_{id} represents the distance between the UE_i and the BS, while $h' \sim \mathcal{CN}(0,1)$. Table 6.2 summarizes the system parameters used to generate the simulation results.
In order to benchmark the proposed algorithm, the following baselines are used:

- **Baseline** 1: This is a convex optimization-based approach in which the IRS phase shifts are optimized using the SDP/SDR approach in [174]. In addition, this scheme uses the MMSE-SIC receiver where all UEs are transmitting at their maximum transmit power.
- **Baseline** 2: This is also a convex optimization-based approach which uses SDP/SDR for the IRS optimization while using the MMSE receiver at the BS.
- **Baseline** 3: This is a low-complexity naive approach which uses $\mathbf{U} = \mathbf{1}$ as the receive beamforming matrix while using random values for the IRS phase shifts and for UEs transmit power.

6.4.3 Fixed-Channels Scenario

For the fixed-channel scenario, it is assumed that the channels remain constant throughout the training period while the error samples are changing in each episode. In particular, it is assumed that the UEs are uniformly distributed in the cell. Furthermore, to ensure the statistical relevance of the obtained results, the trained agents are tested for 2000 episodes using 250 error samples.
Figs. 6.4 and 6.5 illustrate the convergence of the agents for $M = 10$ and $M = 20$, respectively. An interesting observation from the convergence figures is that the agent converges to a higher reward when the number of IRS elements is smaller. In addition, there is a negative relationship between the average reward obtained by the agent and the number of antennas under both error models.
Since the reward function is based on the achieved EE, a higher reward is linked to higher EE in

Table 6.2: Summary of the system parameters.

System parameter	**Value**
Cell radius	100 m
Number of UEs (K)	4
Number of antennas at the BS (N)	$4, 6, 8, 10$
Number of IRS elements (M)	$10, 20$
Maximum transmit power	23 dBm
Bandwidth (B)	1 MHz
UE power amplifier inefficiency factor (ζ')	2.5
Power loss P_c (per antenna)	30 dBm
Noise power spectral density	-174 dBm/Hz
Probability value for Γ^{-1}_{2MN}	0.95
Relative value for the error boundary λ	0.03
Path-loss exponent (UE-IRS) $\iota_{UE \to IRS}$	2.5
Path-loss exponent (UE-IRS) $\iota_{IRS \to BS}$	1.8
Path-loss exponent (UE-BS) $\iota_{UE \to BS}$	3.5
Target rate R_i^{min}	1 MBit/s

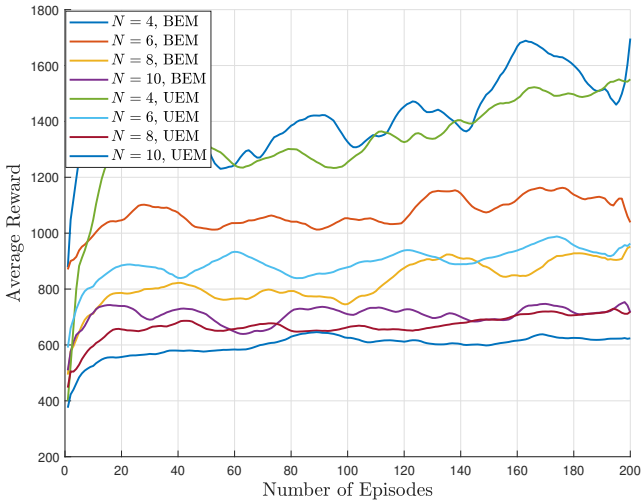

Figure 6.4: The convergence of the SAC agent for the fixed-channels scenario, $M = 10$.

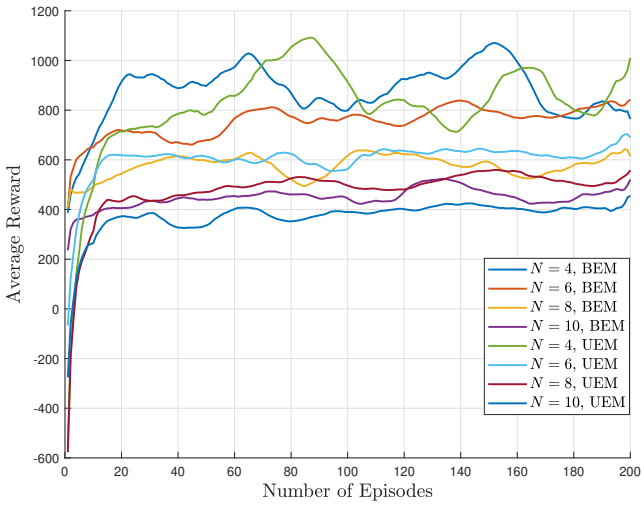

Figure 6.5: The convergence of the SAC agent for the fixed-channels scenario, $M = 20$.

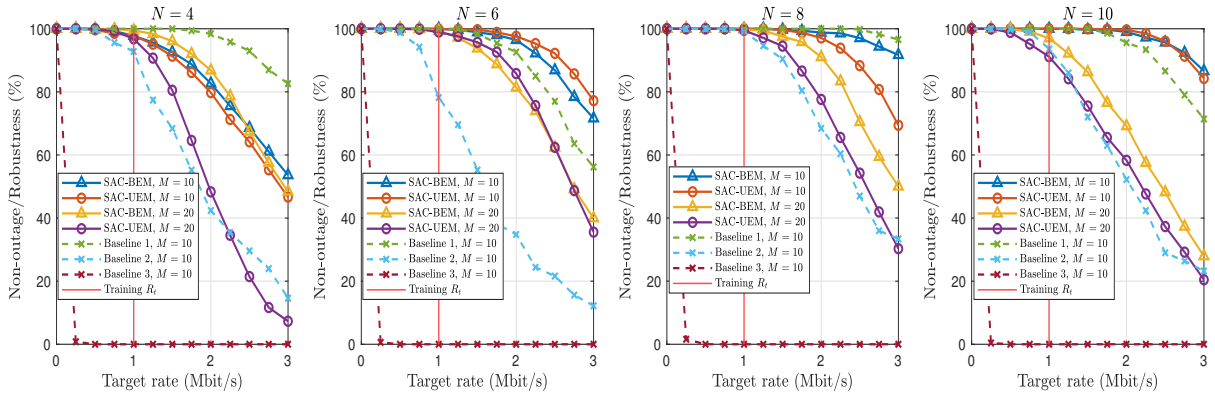

Figure 6.6: The non-outage probability of the proposed algorithm versus the target rate with different N and M.

the system. Figure 6.7 shows the EE achieved by the agent for different system parameters under both error models. Moreover, the highest EE is obtained by the SAC agent under the BEM, achieving an EE of 5.1 Mbit/s/Joule when $M = 10, N = 4$. As the number of antennas increases, the achieved EE decreases accordingly as a result of the losses brought about by the additional antennas. In addition, the same figure shows that increasing the number of IRS elements in fact leads to a decreased EE, even for a fixed number of antennas. This is an interesting result since the understanding in the literature is exactly the opposite when considering a perfect CSI in the system. Furthermore, the EE figure illustrates that the proposed algorithm outperforms the benchmark schemes by significant margins, thanks to its developed adaptive policy.

Figure 6.8 shows the average system sum rate achieved by the agent as a result of its policy. The figure shows an interesting disparity between the agent's achieved sum rate for different M values. In particular, the agent is able to utilize the additional number of antennas in the system to increase the achieved sum rate when $M = 10$, while it is not always the case when $M = 20$. This behaviour can be demonstrated by the fact that since the considered error models are a function of the system parameters, the agent prefers a more conservative and robust policy when the number of IRS elements is doubled at the expense of a reduced system sum rate. Note that this behaviour of the agent is inline with our expectations since the aim of the agent is to robustly increase the average EE of the system. Consequently, Figure 6.6 depicts the non-outage

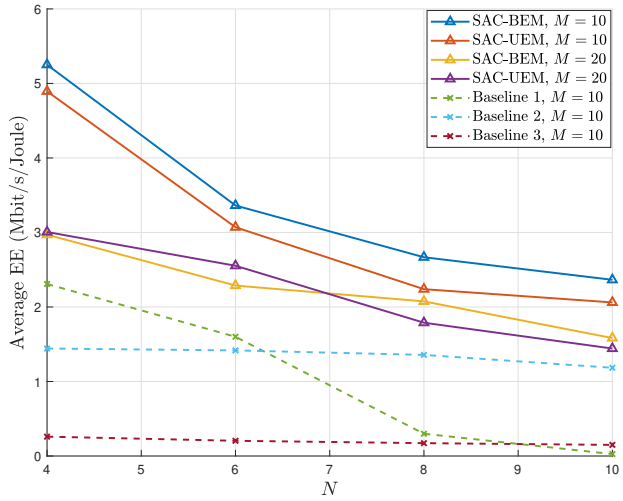

Figure 6.7: The achieved EE for $M = 10, M = 20$ under the bounded and the unbounded error models.

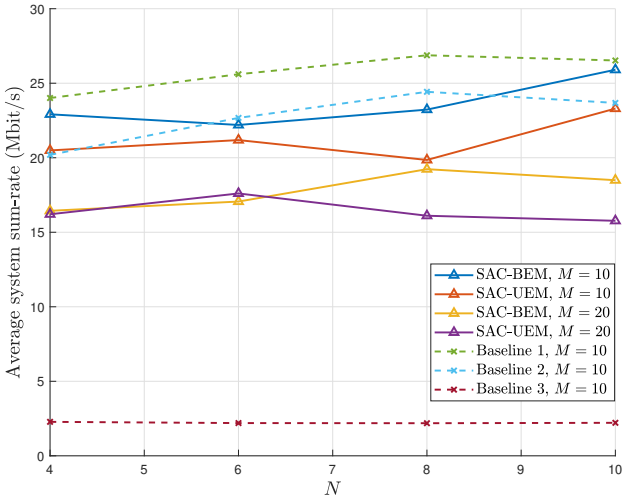

Figure 6.8: The achieved system sum rate for $M = 10, M = 20$ under the bounded and the unbounded error models.

probability of the proposed agent. The figure shows that the agent capitalizes on the additional number of transmit antennas to enhance the robustness of its policy. In particular, the agent significantly enhances the robustness of its policy for $M = 10$ between the cases $N = 4$ and $N = 10$, where the agent is able to sustain almost a 100% non-outage score when the target rate is double what is used during training. On the other hand, the agent performs slightly lower for the case of $N = 10$ compared to $N = 4$ when the number of IRS elements is set to 20. This shows that increasing both the number of antennas and the IRS elements makes it more challenging for the agent to guarantee the required QoS from the UEs for arbitrary errors, particularly under the UEM.

In order to quantify the power consumption by the UEs as a result of the agent's policy, Figure 6.9 illustrates the average power consumption by all UEs for various number of antennas in the system. It can be seen that on average, the agent's policy leads to a higher power consumption

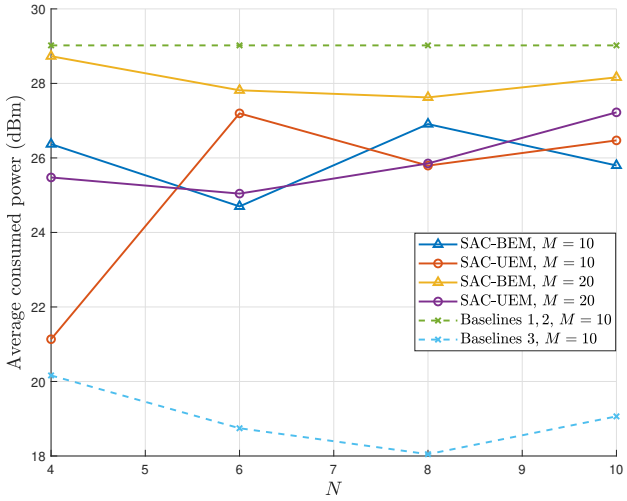

Figure 6.9: The average power consumption versus the number of antennas for the fixed-channels scenario.

by the UEs when $M = 20$ compared to $M = 10$. In addition, the agent's developed policy for $M = 20$ under the BEM leads to the highest power consumption in order to achieve the superior robustness performance shown in Figure 6.6.

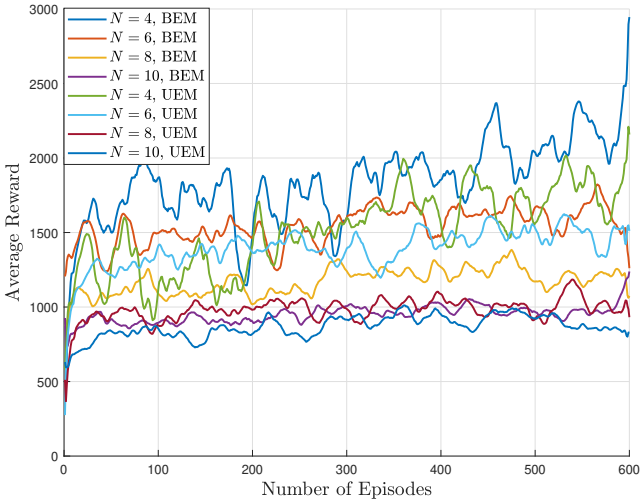

Figure 6.10: The SAC agent's convergence for the dynamic-channels scenario.

6.4.4 Dynamic-Channels Scenario

The fixed-channels scenario is useful for analysing the performance of the agent for a single channel realization. However, since the channel is constantly changing in practice, the dynamic-channels scenario is considered. In particular, a different set of channels is introduced to the agent in each new training episode. The channel set is randomly sampled from a pool of 30 channel sets generated uniformly to incorporate all possible distances between the UE and the BS. In addition, each channel has 30 corresponding error samples randomly selected in each new episode. Furthermore, the dynamic-channels scenario is restricted to $M = 10$. Moreover, to ensure the statistical relevance of the results, the trained agent is tested for 2000 episodes using 250 new set of (never-seen-before) channels.

Figure 6.10 shows the convergence of the developed SAC agent for different N. Similar to the fixed-channels scenario, the agent is able to achieve the highest average reward when $N = 4$ under the BEM. Moreover, the variance in the reward curves is higher than in the fixed-channels scenario due to the different channel samples used for the training. Nevertheless, the SAC agent was able to converge to a relatively high rewarding policy which suggests that the agent has developed a competitive policy.

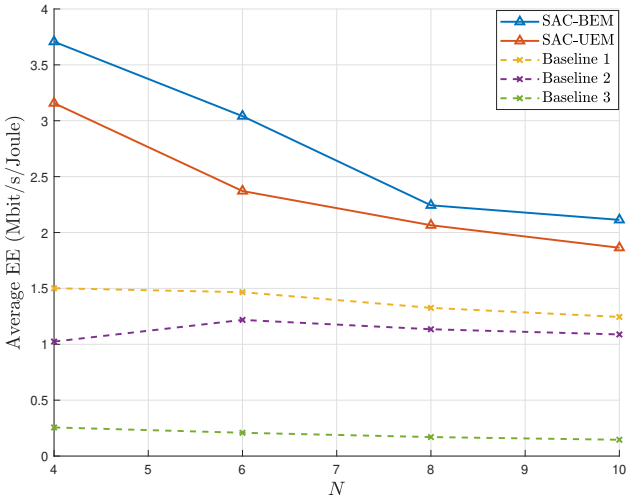

Figure 6.11: The average EE achieved by the SAC agent for the dynamic-channels scenario.

In order to demonstrate the performance of the trained SAC agent, Figure 6.11 shows the average EE achieved by the agent for different N. The figure shows a similar pattern to that of the fixed-channels scenario where the average EE decreases as the number of antennas at the BS is increased. It is worth noting that the average achieved EE is understandably smaller than that of the fixed channels scenario due to the different channels used for testing. Nevertheless, the SAC agent outperforms the benchmark schemes by a significant margin achieving an average EE value of 3.7 Mbit/s/Joule at $N = 4$ which is around 2.5 folds of the EE achieved by the best benchmark scheme. This shows that even for dynamic-channel scenarios the SAC agent was able to achieve considerably higher EE.

In terms of the average system sum rate, Figure 6.13 illustrates the agent's performance. Both curves show a positive slop which indicates that the agent utilizes the additional number of antennas to increase the average system sum rate increasing from 18 Mbit/s for $N = 4$ to 22 Mbit/s for $N = 10$ under the BEM. This positive trend between the average system sum rate and the number of antennas persists under the UEM as well. To benchmark the robustness achieved by the agent's policy, Figure 6.12 depicts the average non-outage probability versus the target rate for different N. The figure shows a consistent pattern in which the agent's robustness improves

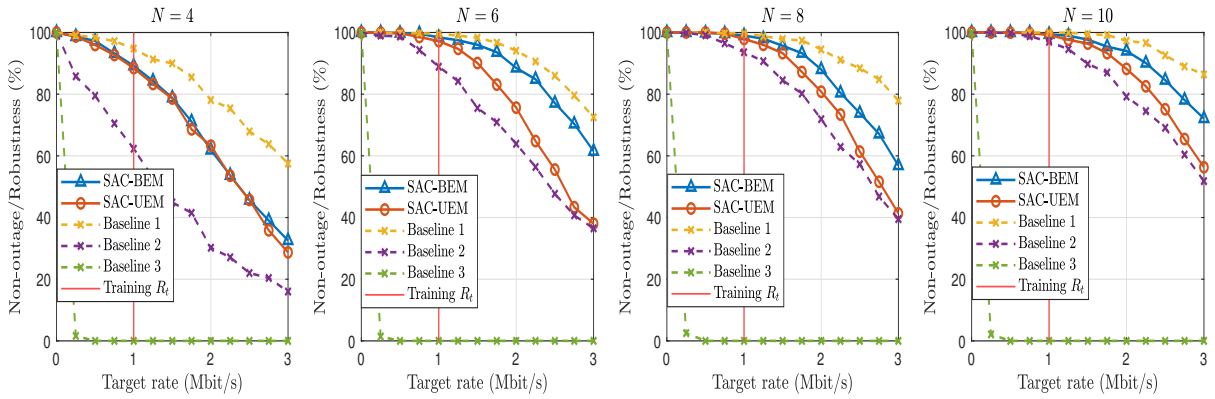

Figure 6.12: The average non-outage probability of the proposed algorithm versus the target rate for different N.

as the number of N increases. In particular, the average non-outage probability is around 90% for $N = 4$, while it is almost 100% when $N = 10$. This shows that the agent was able to utilize the additional number of antennas to strengthen its policy in terms of robustness against channel uncertainties. Furthermore, the same figure also shows that the BEM-based SAC agent outperforms the UEM-based one due to the bounded nature of the BEM.

The average power consumption by the UEs is illustrated in Figure 6.14. Unlike the fixed-channels scenario, the agent exploits the additional number of antennas to reduce the average transmit power in the system. In particular, the figure shows that the average power consumption is less under the BEM for $N = 4, 6$, while the opposite is true for $N = 8, 10$.

Overall, the SAC agent is able to converge to a robust policy that maximizes the EE of the system while satisfying the QoS requirements with a relatively small outage probability. In addition, the SAC agent exploits the additional information about the BEM and achieves marginally better results across all performance metrics compared to the more challenging UEM. Nevertheless, the UEM-based SAC agent shows competitive performance in terms of both EE and robustness compared to that of the benchmark schemes.

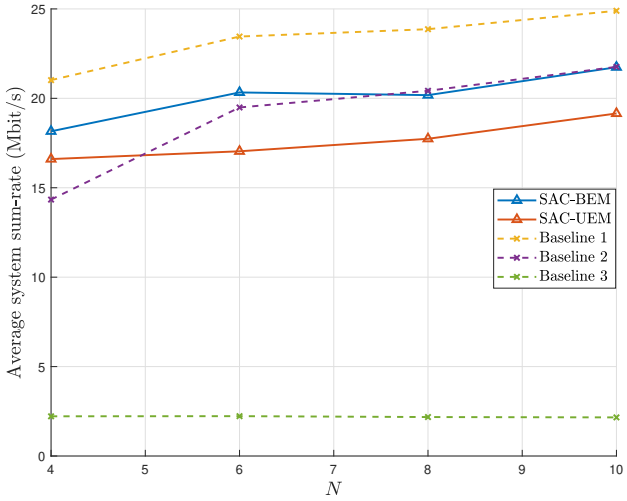

Figure 6.13: The average system sum rate for $M = 10$ under the bounded and the unbounded error models.

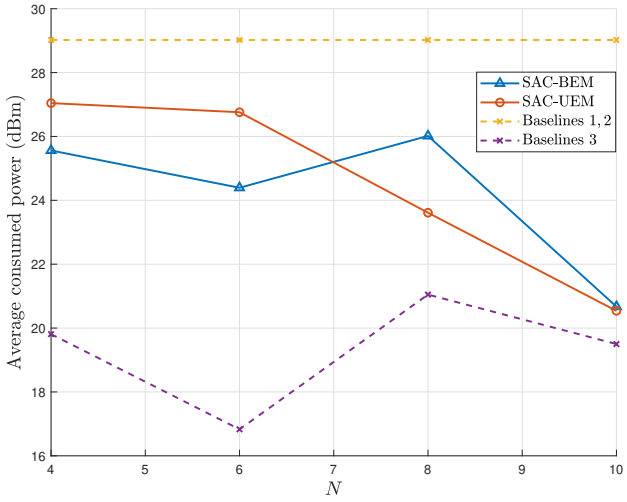

Figure 6.14: The average power consumption versus the number of antennas for the dynamic-channels scenario.

6.5 Summary

In this work, the long-term robust EE maximization problem for an IRS-assisted UL multi-user NOMA system was considered. In particular, by taking into account the channel uncertainties in the system, the robust design problem was formulated as an optimization problem subject to QoS and maximum transmit power constraints. Then, the MMSE-SIC receiver was utilized to determine the combining matrix at the BS. In addition, the joint robust design problem for optimizing the UEs transmit power and the IRS phase shifts was recast as an RL environment. Moreover, an algorithm based on the SAC DRL agent was developed to jointly optimize the design parameters. Through simulation results, the competitive performance of the proposed SAC-based algorithm was demonstrated. Furthermore, the results show that the proposed algorithm outperforms the benchmark schemes in terms of the achieved average EE of the system by a significant margin for both fixed and dynamic-channel scenarios.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

NOMA has been envisioned as a promising MA technique for future wireless networks thanks to its superior spectral and energy efficiencies, fairness and support for massive connectivity. Unlike its OMA counterparts, NOMA supports the incorporation of more than one user's signal in the same RB by exploiting the power-domain multiplexing at the transmitter through SC and the SIC at the receiver. Therefore, the BS can schedule more than one user in the same time slot and on the same frequency while still meeting the QoS requirements. Moreover, it has been shown that NOMA is compatible with other enabling technologies such as MIMO and IRS which allows it to achieve even greater spectral and energy efficiencies. However, as more sophisticated techniques are incorporated into wireless communication systems, designing an efficient and competitive resource allocation strategy becomes more challenging due to the number of design parameters and the resulting constraints. The conventional optimization approach of approximation and relaxation of non-convex functions often results in iterative and computationally complex algorithms. Consequently, these algorithms tend to have a non-negligible latency which severely limits their application in practice. On the other hand, ML-based algorithms only require intensive computational complexity during the training stage. However, once trained, their deployment complexity is significantly lower than their conventional coun-

terparts. Therefore, ML-based algorithms are envisioned to dominate the resource allocation area for latency-sensitive applications. Hence, this thesis focuses on the development of robust ML-based resource allocation for future NOMA systems.

In chapter 4, a DRL-based robust design for an IRS-assisted MISO-NOMA system is developed. In addition, by taking the channel uncertainty into account, the formulated long-term system sum-rate maximization was NP-hard which cannot be solved using conventional optimization techniques. Hence, the robust design problem was reformulated into an RL environment, then, a TD3-based algorithm was developed to solve the reformulated problem. The simulation results demonstrated the superior and robust performance of the proposed algorithm for both fixed and dynamic-channel scenarios compared to the benchmark algorithms in the literature.

Chapter 5 addressed the receiver complexity issue in which the number of SIC operations at the strongest user scales with the number of active users in the same RB. In particular, a correlation-based algorithm was proposed to pair users into clusters. This in turn relaxes the number of SIC operations required by the cluster heads at the expense of an additional inter-cluster interference. Moreover, the ZFBF was utilized to determine the beamforming vectors. However, since the perfect CSI is not available in the system, the ZFBF beamformers cannot eliminate the inter-cluster interference in the system. Hence, the outage-based resource allocation problem for sum-rate maximization is reformulated into an RL environment since the original robust design problem is non-convex and cannot be solved efficiently using conventional optimization techniques. Then, a TD3-based algorithm is developed to jointly optimize the phase shifts, and user and cluster power allocation. The simulation results demonstrated the robust performance of the developed design. In addition, the proposed algorithm was able to outperform conventional and other DRL benchmark schemes in the literature for both fixed and dynamic-channel scenarios.

Finally, chapter 6 considered the robust design for EE maximization in an IRS-assisted UL NOMA system. The long-term robust design problem with EE maximization objective is formulated for the BEM and the UEM subject to QoS and other relevant constraints. Since

the optimization problem is not jointly convex in the design parameters and therefore cannot be solved using conventional optimization methods, the problem is reformulated into an RL environment. To address the beamforming design problem, the MMSE-SIC receiver is adopted due to its attractive closed-form expression. However, while the MMSE-SIC is optimal under perfect CSI conditions, the perfect CSI is not available in the considered system model and therefore, optimality cannot be guaranteed. Therefore, the robust resource allocation strategy is realized through the joint optimization of the phase shifts and users' transmit power. Hence, an algorithm based on the SAC DRL agent is developed to optimize the design parameters with the aim of maximizing the long-term EE of the system. The simulation results demonstrated the superior performance of the proposed algorithm compared to the conventional benchmark schemes in the literature. In particular, the proposed algorithm outperformed the benchmark schemes by a significant margin in terms of EE for both fixed and dynamic-channel scenarios.

7.2 Future Work

This section reviews the potential extensions of the current works of this thesis.

7.2.1 Intelligent Resource Allocation for NOMA-Empowered Integrated Sensing and Communications

The use of higher frequency bands in 6G and beyond systems will be utilized for high-resolution and high-accuracy sensing, which will enable the integration of both wireless sensing and communications in a single system for their mutual benefits. However, combining the constraints from both communication and sensing aspects into a single optimization problem leads to tractability issues when considering the increased number of users in future wireless networks. Therefore, the development of efficient and high-performance ML algorithms will be crucial in realizing the required spectral and energy efficiencies in integrated wireless sensing and communication systems [181].

7.2.2 Age of Information (AoI)-Based Resource Allocation Algorithms for NOMA Systems

AoI refers to the freshness of the information at the receiver. AoI has been subject to extensive study recently due to its utility for IoT and UAV-based networks. In one way, the AoI measures the average network access time which is crucial for time-critical applications. From the resource allocation perspective, the aim of AoI-based design is to minimize the average AoI over multiple time slots subject to relevant power and QoS constraints. Hence, AoI-based resource allocation problems can be solved efficiently by developing DRL-based algorithms similar to problems considered in chapters 4, 5 and 6.

7.2.3 Expert-aided DRL algorithms for Resource Allocation

One of the inherent downsides of DRL-based methods is the number of samples required for training. Since the agent starts with a random policy, it needs to sufficiently sample the environment of the problem space before converging to a highly rewarding policy. However, the number of training samples required by the agent scales with the number of design parameters and the complexity of the problem environment leading to infeasible training-time requirements. Hence, one way to speed up and enhance the sample efficiency of DRL-based resource allocation algorithms is to use an expert-knowledge-based reward function where additional information about the direction of the optimal or near-optimal solution can be exploited. Reduced training time and training samples will be crucial in future wireless networks which are ML-driven such as the radio access network intelligent controller (RIC) unit in the open RAN (O-RAN) architecture where faster training leads to a more responsive RAN.

7.2.4 DRL-based Resource Allocation for Cell-free Systems

One of the most promising concepts for 6G and beyond is cell-free wireless networks. The major advantage of such a concept is the elimination of inter-cell interference and thereby, enhancing the cell-edge user experience. However, there are still practical challenges that need

to be addressed before the realization of the cell-free concept. One of such challenges is the initial access procedure in a cell-less infrastructure. Additionally, the signalling involved in moving users is much higher. Hence, more flexible and adaptive resource allocation techniques are required to address these challenges and realize the additional gains of cell-free architecture. DRL-based algorithms are expected to play a crucial role in developing such dynamic resource allocation techniques which is one possible extension of the works proposed in this thesis.

References

[1] A. Behravan, V. Yajnanarayana, M. F. Keskin, H. Chen, D. Shrestha, T. E. Abrudan, T. Svensson, K. Schindhelm, A. Wolfgang, S. Lindberg *et al.*, "Positioning and sensing in 6g: Gaps, challenges, and opportunities," *IEEE Vehicular Technology Magazine*, 2022.

[2] Z. Wei, J. Yuan, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," *arXiv preprint arXiv:1609.01856*, 2016.

[3] S. Weinstein and P. Ebert, "Data transmission by frequency-division multiplexing using the discrete fourier transform," *IEEE Transactions on Communication Technology*, vol. 19, no. 5, pp. 628–634, 1971.

[4] L. Szczecinski and F. Piera, "Multi-user detection in tdma systems using decision feedback equalizers: case of is-136 and gsm," in *Vehicular Technology Conference. IEEE 55th Vehicular Technology Conference. VTC Spring 2002 (Cat. No.02CH37367)*, vol. 4, 2002, pp. 1975–1979 vol.4.

[5] L. Hanzo, L.-L. Yang, E. Kuan, and K. Yen, *Single-and multi-carrier DS-CDMA: multi-user detection, space-time spreading, synchronisation, standards and networking*. John Wiley & Sons, 2003.

[6] H. Yin and S. Alamouti, "Ofdma: A broadband wireless access technology," in *2006 IEEE sarnoff symposium*. IEEE, 2006, pp. 1–4.

[7] N. Rajatheva, I. Atzeni, E. Bjornson, A. Bourdoux, S. Buzzi, J.-B. Dore, S. Erkucuk, M. Fuentes, K. Guan, Y. Hu *et al.*, "White paper on broadband connectivity in 6g," *arXiv preprint arXiv:2004.14247*, 2020.

[8] E. Hossain and M. Hasan, "5G cellular: key enabling technologies and research challenges," *IEEE Instrumentation & Measurement Magazine*, vol. 18, no. 3, pp. 11–21, 2015.

[9] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE communications magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[10] S. Andreev, V. Petrov, M. Dohler, and H. Yanikomeroglu, "Future of ultra-dense networks beyond 5G: Harnessing heterogeneous moving cells," *IEEE Communications Magazine*, vol. 57, no. 6, pp. 86–92, 2019.

[11] I. F. Akyildiz and J. M. Jornet, "Realizing ultra-massive mimo (1024× 1024) communication in the (0.06–10) terahertz band," *Nano Communication Networks*, vol. 8, pp. 46–54, 2016.

[12] A. Osseiran, J. F. Monserrat, and W. Mohr, "Coordinated multipoint (comp) systems," 2011.

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

[14] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-centric 5g cellular networks: Resource allocation and comparison with the cell-free massive mimo approach," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1250–1264, 2019.

[15] W. K. Alsaedi, H. Ahmadi, Z. Khan, and D. Grace, "Spectrum options and allocations for 6G: A regulatory and standardization review," *IEEE Open Journal of the Communications Society*, 2023.

[16] I. F. Akyildiz, J. M. Jornet, and C. Han, "Teranets: Ultra-broadband communication networks in the terahertz band," *IEEE Wireless Communications*, vol. 21, no. 4, pp. 130–135, 2014.

[17] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless communications and applications above 100 ghz: Opportunities and challenges for 6g and beyond," *IEEE access*, vol. 7, pp. 78 729–78 757, 2019.

[18] P. H. Pathak, X. Feng, P. Hu, and P. Mohapatra, "Visible light communication, networking, and sensing: A survey, potential and challenges," *IEEE communications surveys & tutorials*, vol. 17, no. 4, pp. 2047–2077, 2015.

[19] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Transactions on communications*, vol. 63, no. 12, pp. 4651–4665, 2015.

[20] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE Journal on selected areas in communications*, vol. 32, no. 9, pp. 1637–1652, 2014.

[21] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Non-orthogonal multiple access for 5g and beyond," *arXiv preprint arXiv:1808.00277*, 2018.

[22] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for mimo wireless networks: A promising phy-layer strategy for lte evolution," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 98–105, 2016.

[23] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 648–664, 2018.

[24] T. O'shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, 2017.

[25] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *2013 IEEE 77th vehicular technology conference (VTC Spring)*. IEEE, 2013, pp. 1–5.

[26] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple access techniques for 5G wireless networks and beyond*. Springer, 2019, vol. 159.

[27] T. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, 1972.

[28] P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Transactions on Information Theory*, vol. 19, no. 2, pp. 197–207, 1973.

[29] X. Xiong, W. Xiang, K. Zheng, H. Shen, and X. Wei, "An open source sdr-based noma system for 5G networks," *IEEE Wireless Communications*, vol. 22, no. 6, pp. 24–32, 2015.

[30] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous cdma systems over awgn channel," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1616–1626, 2008.

[31] F. Brannstrom, T. M. Aulin, and L. K. Rasmussen, "Iterative detectors for trellis-code multiple-access," *IEEE Transactions on Communications*, vol. 50, no. 9, pp. 1478–1485, 2002.

[32] L. Liu, J. Tong, and L. Ping, "Analysis and optimization of cdma systems with chip-level interleavers," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 1, pp. 141–150, 2005.

[33] X. Dai, Z. Zhang, B. Bai, S. Chen, and S. Sun, "Pattern division multiple access: A new multiple access technology for 5G," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 54–60, 2018.

[34] L. Zhang, M. Xiao, G. Wu, M. Alam, Y.-C. Liang, and S. Li, "A survey of advanced techniques for spectrum sharing in 5G networks," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 44–51, 2017.

[35] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[36] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "A mathematical proof of the superiority of noma compared to conventional oma," *arXiv preprint arXiv:1612.01069*, 2016.

[37] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge university press, 2011.

[38] H. Al-Obiedollah, H. B. Salameh, A. Gharaibeh, K. Cumanan, Z. Ding, and O. A. Dobre, "Harvested power fairness-based multi-carrier NOMA IoT networks with SWIPT," *IEEE Wireless Communications Letters*, vol. 12, no. 2, pp. 381–385, 2022.

[39] X. Wei, H. Al-Obiedollah, K. Cumanan, Z. Ding, and O. A. Dobre, "Energy efficiency maximization for hybrid TDMA-NOMA system with opportunistic time assignment," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 8, pp. 8561–8573, 2022.

[40] X. Wei, H. Al-Obiedollah, K. Cumanan, W. Wang, Z. Ding, and O. A. Dobre, "Spectral-energy efficiency trade-off based design for hybrid TDMA-NOMA system," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 3377–3382, 2022.

[41] H. Al-Obiedollah, K. Cumanan, H. B. Salameh, G. Chen, Z. Ding, and O. A. Dobre, "Downlink multi-carrier NOMA with opportunistic bandwidth allocations," *IEEE wireless communications letters*, vol. 10, no. 11, pp. 2426–2429, 2021.

[42] P. Xu and K. Cumanan, "Optimal power allocation scheme for non-orthogonal multiple access with

\

alpha $-fairness,''$ *IEEE Journal on Selected Areas in Communications*, $vol.35, no.10, pp.2357 - -2369, 2017.$

[43] C.-L. Wang, J.-Y. Chen, and Y.-J. Chen, "Power allocation for a downlink non-orthogonal multiple access system," *IEEE wireless communications letters*, vol. 5, no. 5, pp. 532–535, 2016.

[44] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE transactions on wireless communications*, vol. 15, no. 11, pp. 7244–7257, 2016.

[45] J. Wang, Q. Peng, Y. Huang, H.-M. Wang, and X. You, "Convexity of weighted sum rate maximization in NOMA systems," *IEEE signal processing letters*, vol. 24, no. 9, pp. 1323–1327, 2017.

[46] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1647–1651, 2015.

[47] Y. Zhang, H.-M. Wang, T.-X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2852–2857, 2017.

[48] E. C. Cejudo, H. Zhu, and J. Wang, "Resource allocation in multicarrier NOMA systems based on optimal channel gain ratios," *IEEE Transactions on Wireless Communications*, 2021.

[49] Z. Xu, I. Petrunin, T. Li, and A. Tsourdos, "Efficient allocation for downlink multi-channel NOMA systems considering complex constraints," *Sensors*, vol. 21, no. 5, p. 1833, 2021.

[50] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE transactions on information theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[51] M. A. Maddah-Ali, M. A. Sadrabadi, and A. K. Khandani, "Broadcast in MIMO systems based on a generalized QR decomposition: Signaling and performance analysis," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1124–1138, 2008.

[52] Y. Huang, C. Zhang, J. Wang, Y. Jing, L. Yang, and X. You, "Signal processing for MIMO-NOMA: Present and future challenges," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 32–38, 2018.

[53] C. Rao, Z. Ding, K. Cumanan, and X. Dai, "A GSVD-based precoding scheme for MIMO-NOMA relay transmission," *IEEE Internet of Things Journal*, 2023.

[54] D. Lin, K. Cumanan, and Z. Ding, "Beamforming design for BackCom assisted NOMA systems," *IEEE wireless communications letters*, 2023.

[55] Z. Cao, X. Ji, J. Wang, W. Wang, K. Cumanan, Z. Ding, and O. A. Dobre, "Artificial noise aided secure communications for cooperative NOMA networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 946–963, 2021.

[56] H. Al-Obiedollah, K. Cumanan, H. B. Salameh, S. Lambotharan, Y. Rahulamathavan, Z. Ding, and O. A. Dobre, "A joint beamforming and power-splitter optimization technique for SWIPT in MISO-NOMA system," *IEEE Access*, vol. 9, pp. 33 018–33 029, 2021.

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

[58] 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 Transactions on Communications*, vol. 67, no. 6, pp. 4117–4131, 2019.

[59] H. M. Al-Obiedollah, K. Cumanan, J. Thiyagalingam, J. Tang, A. G. Burr, Z. Ding, and O. A. Dobre, "Spectral-energy efficiency trade-off-based beamforming design for MISO non-orthogonal multiple access systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6593–6606, 2020.

[60] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "On the performance of cell-free massive MIMO relying on adaptive NOMA/OMA mode-switching," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 792–810, 2019.

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

[62] ——, "Robust beamforming techniques for non-orthogonal multiple access systems with bounded channel uncertainties," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2033–2036, 2017.

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

[64] J. Cui, Z. Ding, and P. Fan, "Beamforming design for MISO non-orthogonal multiple access systems," *IET Communications*, vol. 11, no. 5, pp. 720–725, 2017.

[65] F. Zhu, Z. Lu, J. Zhu, J. Wang, and Y. Huang, "Beamforming design for downlink non-orthogonal multiple access systems," *IEEE Access*, vol. 6, pp. 10 956–10 965, 2018.

[66] Z. Chen, Z. Ding, P. Xu, and X. Dai, "Optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1263–1266, 2016.

[67] U. Erez and S. ten Brink, "A close-to-capacity dirty paper coding scheme," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3417–3432, 2005.

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

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

[70] Z. Xiao, L. Zhu, Z. Gao, D. O. Wu, and X.-G. Xia, "User fairness non-orthogonal multiple access (NOMA) for millimeter-wave communications with analog beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 7, pp. 3411–3423, 2019.

[71] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4438–4454, 2016.

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

[73] S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE access*, vol. 5, pp. 565–577, 2016.

[74] N. Nonaka, Y. Kishiyama, and K. Higuchi, "Non-orthogonal multiple access using intra-beam superposition coding and SIC in base station cooperative MIMO cellular downlink," *IEICE Transactions on Communications*, vol. 98, no. 8, pp. 1651–1659, 2015.

[75] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2370–2382, 2017.

[76] F. Mornati, "Pareto optimality in the work of pareto," *Revue européenne des sciences sociales. European Journal of Social Sciences*, no. 51-2, pp. 65–82, 2013.

[77] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1077–1091, 2017.

[78] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," *IEEE Communications Letters*, vol. 20, no. 7, pp. 1465–1468, 2016.

[79] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7686–7698, 2016.

[80] H. Al-Obiedollah, H. B. Salameh, K. Cumanan, Z. Ding, and O. A. Dobre, "Competitive IRS assignment for IRS-based NOMA system," *IEEE wireless communications letters*, 2023.

[81] ——, "Self-sustainable multi-IRS-aided wireless powered hybrid TDMA-NOMA system," *IEEE Access*, 2023.

[82] Z. Ding and H. V. Poor, "A simple design of IRS-NOMA transmission," *IEEE Communications Letters*, vol. 24, no. 5, pp. 1119–1123, 2020.

[83] F. Fang, Y. Xu, Q.-V. Pham, and Z. Ding, "Energy-efficient design of IRS-NOMA networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 14 088–14 092, 2020.

[84] J. Zhu, Y. Huang, J. Wang, K. Navaie, and Z. Ding, "Power efficient IRS-assisted NOMA," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 900–913, 2020.

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

[86] Y. Li, M. Jiang, Q. Zhang, and J. Qin, "Joint beamforming design in multi-cluster MISO NOMA intelligent reflecting surface-aided downlink communication networks," *arXiv preprint arXiv:1909.06972*, 2019.

[87] M. Svensén and C. M. Bishop, "Pattern recognition and machine learning," 2007.

[88] T. Wang, C.-K. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," *China Communications*, vol. 14, no. 11, pp. 92–111, 2017.

[89] Y. Chu, K. Cumanan, S. Smith, O. Dobre *et al.*, "Deep learning based fall detection using WiFi channel state information," *IEEE Access*, 2023.

[90] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Communications surveys & tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.

[91] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[92] A. Waraiet, K. Cumanan, Z. Ding, and O. A. Dobre, "Robust design for irs-assisted miso-noma systems: A drl-based approach," *IEEE Wireless Communications Letters*, 2023.

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

[94] M. Kim, N.-I. Kim, W. Lee, and D.-H. Cho, "Deep learning-aided SCMA," *IEEE Communications Letters*, vol. 22, no. 4, pp. 720–723, 2018.

[95] S. Wang, T. Lv, and X. Zhang, "Multi-agent reinforcement learning-based user pairing in multi-carrier NOMA systems," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2019, pp. 1–6.

[96] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2200–2210, 2019.

[97] J. Lee and J. So, "Reinforcement learning-based joint user pairing and power allocation in MIMO-NOMA systems," *Sensors*, vol. 20, no. 24, p. 7094, 2020.

[98] Z. Ding, R. Schober, and H. V. Poor, "No-pain no-gain: DRL assisted optimization in energy-constrained CR-NOMA networks," *IEEE Transactions on Communications*, 2021.

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

[100] S. Jiao, X. Xie, Z. Ding *et al.*, "Deep reinforcement learning based optimization for IRS based UAV-NOMA downlink networks," *arXiv preprint arXiv:2106.09616*, 2021.

[101] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[102] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[103] X. Wei, "Resource allocation techniques for non-orthogonal multiple access in beyond 5g," September 2022. [Online]. Available: https://etheses.whiterose.ac.uk/31771/

[104] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[105] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems*. Citeseer, 1994, vol. 37.

[106] C. Watkins, "Learning from delayed rewards, king's college, cambridge, 1989 ph. d," Ph.D. dissertation, thesis.

[107] S. Agarwal, "Data mining: Data mining concepts and techniques," in *2013 international conference on machine intelligence and research advancement*. IEEE, 2013, pp. 203–207.

[108] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[109] M. A. Nielsen, *Neural networks and deep learning*. Determination press San Francisco, CA, 2015, vol. 25.

[110] M. Riedmiller, "Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method," in *European conference on machine learning*. Springer, 2005, pp. 317–328.

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

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

[113] M. Morales, *Grokking deep reinforcement learning*. Simon and Schuster, 2020.

[114] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1928–1937.

[115] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[116] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[117] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[118] M. Lapan, *Deep reinforcement learning hands-on*. Packt publishing, 2020.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[142] M. B. Shahab, M. Irfan, M. F. Kader, and S. Young Shin, "User pairing schemes for capacity maximization in non-orthogonal multiple access systems," *Wireless Communications and Mobile Computing*, vol. 16, no. 17, pp. 2884–2894, 2016.

[143] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 328–331, 2018.

[144] 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 Transactions on Communications*, vol. 67, no. 6, pp. 4117–4131, 2019.

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

[146] A. Agrawal, J. G. Andrews, J. M. Cioffi, and T. Meng, "Iterative power control for imperfect successive interference cancellation," *IEEE Transactions on wireless communications*, vol. 4, no. 3, pp. 878–884, 2005.

[147] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4847–4861, 2016.

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

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

[150] A. Bulut and T. K. Ralphs, "On the complexity of inverse mixed integer linear optimization," *SIAM Journal on Optimization*, vol. 31, no. 4, pp. 3014–3043, 2021.

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

[152] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE transactions on wireless communications*, vol. 18, no. 11, pp. 5394–5409, 2019.

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

[154] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, 2015.

[155] A. Kumar *et al.*, "User pairing and power allocation for IRS-assisted NOMA systems with imperfect phase compensation," *IEEE Wireless Communications Letters*, vol. 11, no. 12, pp. 2492–2496, 2022.

[156] 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 Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2413–2424, 2017.

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

[158] B. Kimy, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *MILCOM 2013-2013 IEEE Military Communications Conference*. IEEE, 2013, pp. 1278–1283.

[159] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on selected areas in communications*, vol. 24, no. 3, pp. 528–541, 2006.

[160] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Transactions on Signal Processing*, vol. 56, no. 9, pp. 4409–4418, 2008.

[161] M. Sharif and B. Hassibi, "A comparison of time-sharing, DPC, and beamforming for MIMO broadcast channels with many users," *IEEE Transactions on Communications*, vol. 55, no. 1, pp. 11–15, 2007.

[162] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[163] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1866–1880, 2020.

[164] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

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

[166] M. Uzair and N. Jamil, "Effects of hidden layers on the efficiency of neural networks," in *2020 IEEE 23rd international multitopic conference (INMIC)*. IEEE, 2020, pp. 1–6.

[167] Y. Sun, C. E. Koksal, and N. B. Shroff, "Capacity of compound MIMO Gaussian channels with additive uncertainty," *IEEE transactions on information theory*, vol. 59, no. 12, pp. 8267–8274, 2013.

[168] Y. Xu, R. Q. Hu, and G. Li, "Robust energy-efficient maximization for cognitive NOMA networks under channel uncertainties," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8318–8330, 2020.

[169] H. Sun, F. Zhou, R. Q. Hu, and L. Hanzo, "Robust beamforming design in a noma cognitive radio network relying on swipt," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 1, pp. 142–155, 2018.

[170] A. Zakeri, A. Khalili, M. R. Javan, N. Mokari, and E. Jorswieck, "Robust energy-efficient resource management, sic ordering, and beamforming design for mc miso-noma enabled 6g," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2481–2498, 2021.

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

[172] Z. Ding, R. Schober, and H. V. Poor, "Unveiling the importance of SIC in NOMA systems—part 1: State of the art and recent findings," *IEEE Communications Letters*, vol. 24, no. 11, pp. 2373–2377, 2020.

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

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

[175] H. Tabassum, M. S. Ali, E. Hossain, M. J. Hossain, and D. I. Kim, "Uplink vs. downlink noma in cellular networks: Challenges and research directions," in *2017 IEEE 85th vehicular technology conference (VTC Spring)*. IEEE, 2017, pp. 1–7.

[176] M. Zeng, E. Bedeer, O. A. Dobre, P. Fortier, Q.-V. Pham, and W. Hao, "Energy-efficient resource allocation for IRS-assisted multi-antenna uplink systems," *IEEE Wireless Communications Letters*, vol. 10, no. 6, pp. 1261–1265, 2021.

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

[178] M. R. Zamani, M. Eslami, M. Khorramizadeh, H. Zamani, and Z. Ding, "Optimizing weighted-sum energy efficiency in downlink and uplink noma systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 11 112–11 127, 2020.

[179] C. Windpassinger, "Detection and precoding for multiple input multiple output channels," Ph.D. dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 2004.

[180] M.-G. Berceanu, C. Voicu, and S. Halunga, "The performance of an uplink large scale mimo system with MMSE-SIC detector," in *2019 International Conference on Military Communications and Information Systems (ICMCIS)*. IEEE, 2019, pp. 1–4.

[181] Z. Wang, Y. Liu, X. Mu, Z. Ding, and O. A. Dobre, "NOMA empowered integrated sensing and communication," *IEEE Communications Letters*, vol. 26, no. 3, pp. 677–681, 2022.