Machine Learning Empowered Reconfigurable Intelligent Surfaces

by

Ruikang Zhong
Supervisors : Prof. Yue Chen, Dr. Yuanwei Liu
Independent Assessor: Prof. Kok Keong Chai

Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary University of London
United Kingdom

May 2023

Acknowledgments

Foremost, I would like to thank Prof. Yue Chen, Dr. Yuanwei Liu, and independent assessor Prof. Kok Keong Chai for their unwavering support of my Ph.D study. With patience and vast expertise, they not only gave me helpful technical advice and constructive comments on my academic undertakings and directions, but they also offered meindispensable advice for daily life. Their expertise and encouragement have been crucial in shaping my research and helping me overcome obstacles along the way.

I would like to thank all my collaborators: Prof. Lingyang Song (Peking University), Prof. Xianbin Wang (Western University), Prof. Zhu Han (University of Houston), Prof. Lajos Hanzo (University of Southampton), Prof. Ping Zhang (Beijing University of Posts and Telecommunications), Prof. Jianhua Zhang (Beijing University of Posts and Telecommunications), and Dr. Mona Jaber (QMUL) for their helpful suggestions and comments on my research.

I would also like to thank Dr. Xiao Liu, Dr. Xidong Mu, Dr. Gui Zhou, Dr. Zhong Yang, Dr. Tianwei Hou, Dr. Wenqiang Yi, Dr. Syed Khurram Mahmud, Dr. Yixuan Zou, Dr. Zhishu Qu, Yanling Hao, Chao Zhang, Jiaqi Xu, Xinyu Gao, Na Xue, Ziyi Xie, Yuqin Liu, Zhixiong Chen, Yimeng Zhang, Haochen Li, Zhaolin Wang, Zheng Zhang, Meng Zhang, ZhaomingHu, Suyu Lyu, QianGao, Kangda Zhi, Na Yan, Tuo Wu and all my colleges and friends in the communication systems research group and antenna group at the Queen Mary University of London, for their constant encouragement and kind help. I really have had wonderful memories in my Ph.D life and study.

I would like to express my deepest gratitude to my beloved parents. I would also like to express my pure love to my wife, Qilei Wang.

Abstract

Reconfigurable intelligent surfaces (RISs) or known as intelligent reflecting surfaces (IRSs) have emerged as potential auxiliary equipment for future wireless networks, which attracts extensive research interest in their characteristics, applications, and potential. RIS is a panel surface equipped with a number of reflective elements, which can artificially modify the propagation environment of the electrogenic signals. Specifically, RISs have the ability to precisely adjust the propagation direction, amplitude, and phase-shift of the signals, providing users with a set of cascaded channels in addition to direct channels, and thereby improving the communication performances for users. Compared with other candidate technologies such as active relays, RIS has advantages in terms of flexible deployment, economical cost, and high energy efficiency. Thus, RISs have been considered a potential candidate technique for future wireless networks.

In this thesis, a wireless network paradigm for the sixth generation (6G) wireless networks is proposed, where RISs are invoked to construct smart radio environments (SRE) to enhance communication performances for mobile users. In addition, beyond the conventional reselecting-only RIS, a novel model of RIS is originally proposed, namely, simultaneous transmitting and reflecting reconfigurable intelligent surface (STAR-RIS). The STAR-RIS splits the incident signal into transmitted and reflected signals, making full utilization of them to generate 360° coverage around the STAR-RIS panel, improving the coverage of the RIS. In order to fully exert the channel domination and beamforming ability of the RISs and STAR-RSIs to construct SREs, several machine learning algorithms, including deep learning (DL), deep reinforcement learning (DRL), and federated learning (FL) approaches are developed to optimize the communication performance in respect of sum data rate or energy efficiency for the RIS-assisted networks.

Specifically, several problems are investigated including 1) the passive beamforming prob-

lem of the RIS with consideration of configuration overhead is resolved by a DL and a DRL algorithm, where the time overhead of configuration of RIS is successfully reduced by the machine learning algorithms. Consequently, the throughput during a time frame improved 95.2% by invoking the proposed algorithms; 2) a novel framework of mobile RISs-enhanced indoor wireless networks is proposed, and a FL enhanced DRL algorithm is proposed for the deployment and beamforming optimization of the RIS. The average throughput of the indoor users severed by the mobile RIS is improved 15.1% compared to the case of conventional fixed RIS; 3) A STAR-RIS assisted multi-user downlink multiple-input single-output (MISO) communication system is investigated, and a pair of hybrid reinforcement learning algorithms are proposed for the hybrid control of the transmitting and reflecting beamforming of the STAR-RIS, which ameliorate 7% of the energy efficiency of the STAR-RIS assisted networks; 4) A tile-based low complexity beamforming approach is proposed for STAR-RISs, and the proposed tile-based beamforming approach is capable of achieving homogeneous data rate performance with element-based beamforming with appreciable lower complexity.

By designing and operating the computer simulation, this thesis demonstrated 1) the performance gain in terms of sum data rate or energy efficiency by invoking the proposed RIS in the wireless communication networks; 2) the data rate or energy efficient performance gain of the proposed STAR-RIS compared to the existing reflecting-only RIS; 3) the effect of the proposed machine learning algorithms in terms of convergence rate, optimality, and complexity compared to the benchmarks of existing algorithms.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3D	Three-Dimensional
3GPP	Third Generation Partnership Project
5G	Fifth-Generation
6G	Sixth-Generation
AFFL	Access-free Federated Learning
AOA	Angle of Arrival
AWGN	Additive White Gaussian Noise
BS	Base Station
CSI	Channel System Information
DDPG	Deep deterministic policy gradient
DL	Deep learning
DNN	Deep neural network
DoF	Degrees-of-freedom
DQN	Deep Q network
DRL	Deep Reinforcement learning
ES	Energy splitting
ETDL	Environment Trained Deep Learning
FL	Federated Learning
FPGA	Field Programmable Gate Arrays

IRS	Intelligent Reflecting Surfaces
ITU	International Telecommunication Union
LoS	Line of Sight
LTE	Long Term Evolution
MIMO	Multiple Input Multiple Output
MISO	Multiple Input Single Output
ML	Machine Learning
MS	Model Switching
NLOS	Non-line-of-sight
NOMA	Non-orthogonal Multiple Access
PG	Policy gradient
PPO	Proximal policy optimization
QMUL	Queen Mary University of London
RC	Reflection Coefficient
RF	Radio Frequency
RIS	reconfigurable intelligent surface
SIC	Successive interference cancelation
SNR	Signal-to-noise ratio
SRE	Smart Radio Environments
STAR-RIS	simultaneous transmitting and reflecting reconfigurable intelligent surface
TC	Transmission coefficient
TRC	Transmission and reflection Coefficient
WLAN	Wireless local area network
ZF	Zero forcing

Chapter 1

Introduction

In the past three decades, wireless communication technology has made revolutionary progress, but most wireless technologies focus on transceiver sides, such as coding, modulation, etc. Although the fading channel has been identified as the bottleneck that restricts improving the communication quality for decades, the propagation environment having fading and noise was consistently considered to be uncontrollable until the emergence of reconfigurable electromagnetic metamaterials [1]. This advancement in materials technology proffers scholars working in wireless communication an opportunity to artificially modify the electromagnetic propagation environment. As a consequence, after the fifth-generation (5G) mobile communication system has entered the commercial stage, a major innovation for the future communication system, namely the concept of the smart radio environment (SRE) has emerged [2]. The main function of the SRE is to offer propagation environments with less path loss and interference for mobile users.

Two important pillars that support accomplishing the ambitious blueprint of SRE are the reconfigurable intelligent surfaces (RISs) and the machine learning (ML) toolbox. On one hand, RISs provide passive means to change the propagation environments for the electromagnetic signal [3], which laid a physical foundation for achieving SRE. On the other hand, ML approaches play the role of the brain for RIS, thereby controlling

Figure 1.1: Application scenarios of the RIS.

RIS to generate SRE that meets user demands.

RISs [4, 5], also known as intelligent reflecting surfaces (IRSs) [6], have been anticipated as a neonatal component of future communication systems [7]. RISs are two-dimensional (2D) panels and are comprised of a massive number of low-cost reconfigurable elements. By employing reflecting element arrays, the incident signal can be manipulated by RISs. The propagation direction, amplitude, and phase shift of the signals can be precisely adjusted, providing users with a set of cascaded channels in addition to direct channels [8, 9]. Through a smart controller (e.g., a field-programmable gate array (FPGA) attached to the RIS), the phase and the amplitude of these reconfigurable elements can be beneficially controlled, thus reconfiguring the propagation of the incident wireless signals and realizing SRE.

The RIS is a game-changing technology which can be applied in a variety of appli-

cation scenarios. As shown in Fig. 1.1, it can be employed in various communication systems, including but not limited to cellular networks, Internet of Things (IoT) networks, indoor WiFi, and vehicle-to-everything (V2X) communications.

1.1 Research Motivations

Although RIS provides the possibility to change the signal propagation for the users, this function does not necessarily results in any data rate or other gain for mobile users. Aimlessly changing the phase or amplitude of the signal is likely to cause severe inter-user interference. Therefore, it is necessary to concentrate the signal energy to the designated user terminal by using the passive beamforming at RIS and the joint active & passive beamforming at BS and RIS. Some existing research contributions studied the RIS beamforming, such as reflective beamforming for RIS [10], joint active & passive beamforming for RIS-BS [11], bidirectional beamforming for simultaneous transmitting and reflecting RIS (STAR-RIS) [1] [12], and multi-RIS beamforming [13]. Unfortunately, existing research ignores some practical issues that may cause severe signal to the interference-and-noise ratio (SINR) degradation.

First, most existing studies assume an absolutely ideal RIS that responds perfectly to all commands in real time. This assumption simplifies the problem and makes it less challenging. However, as pointed out in [14], the time required to configure all RIS array elements can be significant, causing a reduction in effective transmission time and thus degradation on data throughput. In addition, existing papers generally ignore possible occlusions around the RIS. In fact, losing LoS can cause significantly larger path loss, resulting in a decrease in SINR at the receiving side. Finally, complexity is an important issue in RIS beamforming which is usually ignored in the current research. Since the RIS element generally does not have the power amplification function, in order to ensure signal strength, the RIS needs to have a large number of elements. Therefore,

[1] The transmitting refers to a refraction in a passive manner, which does not require any active radio frequency (RF) signal chain.

computing the phase shift for each array element using traditional schemes such as convex optimization algorithms would generate impractical complexity. These practical problems need to be solved to improve the efficiency of using RIS, which motivate the author to investigate these practical problems and propose several ML algorithms to resolve them, respectively.

1.2 Methodologies

In order to resolve the beamforming problem of the RISs in the wireless networks, in this thesis, several machine learning algorithms are proposed, including deep learning (DL), reinforcement learning (RL), and federated learning (FL) to optimise the deployment, active beamforming of the base station (BS), passive beamforming of the RIS, and resource allocation for RIS-assisted NOMA networks, improving key performance indicators including sum data rate and power efficiency with respect to the quality of service (QoS).

1.2.1 Why Machine Learning

- ***Dynamic and non-functional environment:*** In practical scenarios, the wireless networks have to operate in non-ideal scenarios, where the obstacles, propagation environments, and user distributions are likely to be non-analytical. Moreover, the mobility of the users and surroundings (e.g. vehicles) will lead to dynamic optimisation problems. Due to the aforementioned two challenges, using conventional optimisation techniques for optimising mobile networks is intractable. However, for such dynamic and complex problems, machine learning is capable of indicating a near-optimal solution through an experience-based method but not a functional expression. The agent accumulates certain experiences through continuous exploration of the current environment and obtains high-quality solutions by learning and remembering these fruitful or dreadful experiences [15].

- ***Long-term optimisation problem:*** Since wireless networks have to provide continuous services for different users, the optimisation for a period of service time is a long-term and cumulative optimisation problem [16]. The long-term nature of the problem was ignored in the previous research since the state of transceivers (e.g. modulation scheme, transmit power) can be arbitrarily changed in any time slot, such as modulation mode switching, resource allocation, etc. However, when a kind of physical property cannot be arbitrarily changed, the long-term nature of the problem can have a significant impact. For example, the position of the mobile RIS is limited by the moving speed, thus, the trajectory needs a long-term optimisation when a flexible deployment plan is designed for the RIS. Similar problems requiring long-term optimisation include RIS beamforming problems with configuration overhead or maximum phase-shift change constraints.

- ***Massive elements:*** The last reason is the complexity of controlling RISs. Since elements on RISs are not equipped with active power amplifiers, the power of the transmitted or reflected signal produced by each element is limited. Hence, in order to ensure sufficient signal power at the receiving terminal, RISs require a massive number of elements, and the required number is estimated to start from hundreds to thousands. Since solving the beamforming problem requires calculating the reflecting coefficient (RC) or transmitting coefficient (TC) for each element, the massive number of elements leads to a high-complexity problem. Paradoxically, beamforming needs to be given on the fly based on CSI, which indicates that the system has a low tolerance for time complexity. Well-trained artificial intelligence algorithms are capable of giving a solution spending insignificant time consumption [17], and this real-time prompt feature fits the requirement of the communication system since it needs to make an immediate decision according to the real-time CSI.

1.2.2 DL & RL for the RIS Control

In this thesis, the research aim is to develop ML algorithms to intelligently control the RIS in order to improve the aforementioned communication performances of RIS-assisted wireless networks. Several practical but challenging problems are considered, including the configuration overhead, mobility of RISs and users, non-analytical surroundings, and the non-ideal phase response of the array elements. These issues further complicate the optimisation problem in RIS-assisted networks. As a consequence of reasons summarized in 1.2.1, DL and deep reinforcement learning (DRL) algorithms are developed to solve the formulated problem, and the benefits and drawbacks of DL and DRL are investigated.

Assisted by the fitting function of deep neural networks (DNNs), DL algorithms are indeed conducive to solving complex and non-functional optimisation problems. Meanwhile, DRL algorithms are considered to be a more competent methodology for long-term optimisation problems since DRL is able to formulate a Markov decision process (MDP) to maximise the accumulated reward for a time frame [18, 19]. Therefore, DRL algorithms are employed for the long-term optimisation such as the flexible deployment of the mobile RIS.

1.2.3 Comparing DL & RL in RIS Control

However, due to the participation of DNNs, DL and DRL algorithms are also endowed with Achilles' Heel. Poor interpretability is the first criticism. Even if the solution provided by the agent is manually observed, its principle and the logic behind the obtained solutions can only be inferred with the lack of interpretability [20]. Moreover, since the optimality of DL and RL algorithms is difficult to be analytically proved, this has led to a question. For a specified circumstance, which algorithm has salient performance and applicability? Although a number of artificial intelligence algorithms have been proposed for the communication field and are claimed to be effective, their performance in communication systems is still inconclusive. Thus, it is still problematic to determine

which approach is preferred in certain communication scenarios. In this thesis, not only several DL and RL algorithms are developed for the optimisation of the RISs, but also the performance between DL and RL algorithms is compared in the same scenario to reveal the superiority and inferiority of the algorithms.

1.2.4 FL for Flexible Deployments of RISs

Meanwhile, since multiple mobile RISs can be deployed in different cells, federated learning (FL) is employed to strengthen their training efficiency and effectiveness for the proposed multi-cell multi-agent scenario [21]. FL arouses the interest of researchers as a distributed learning framework since it can effectively utilize computational resources [22] with a protection of user privacy [23]. Especially for the DRL algorithm, FL can improve training efficiency and learning effect, since agents can explore the environments simultaneously and their knowledge can be transferred to each other through a global neural networks model. Therefore, a DRL algorithm with a framework of FL is proposed, namely the FL enhanced deep deterministic policy gradient (FL-DDPG) algorithm to jointly optimise the passive beamforming, dynamic deployment of RISs, and the power allocation for NOMA users.

1.3 Research Contributions

The main contributions of this thesis can be summarised as following

1. A RIS-assisted multi-user downlink communication system over fading channels is investigated. In particular, the time overhead for configuring the RIS reflective elements at the beginning of each fading channel is considered. Two ML algorithms are proposed, including a deep learning algorithm named environment trained deep learning (ETDL) and a reinforcement learning algorithm named exploration attenuated deep deterministic policy gradient (EA-DDPG), to solve the

resulting overhead-dependent joint optimisation problems. The ETDL algorithm trains the DL agent through the communication environment, which eliminates the requirement of the training data set. The EA-DDPG algorithm has the ability to achieve continuous and deterministic control of phase shifts. The proposed solutions are capable of balancing the trade-off between configuration overhead and configuration accuracy to improve the throughput of the RIS-assisted networks.

2. A novel framework of RISs-enhanced indoor wireless networks is proposed, where a RIS mounted on a robot is invoked to enable the mobility of the RIS and enhance the service quality for mobile users. Meanwhile, non-orthogonal multiple access (NOMA) techniques are adopted to further increase the spectrum efficiency since RISs are capable of providing NOMA with artificially controlled channel conditions, which can be seen as a beneficial operation condition to obtain NOMA gains. To optimise the sum rate of all users, a deep deterministic policy gradient (DDPG) algorithm is invoked to optimise the deployment and phase shifts of the mobile RIS as well as the power allocation policy. In a multi-cell scenario, where each cell employs a robot, in order to improve the efficiency and effectiveness of agent training for the DDPG agents, a federated learning (FL) training framework is adopted to enable multiple agents to simultaneously explore similar environments and exchange experiences. The proposed FL-enhanced DDPG algorithm can control RISs to provide additional data rate gain compared with the fixed RIS deployment.

3. A STAR-RIS assisted multi-user downlink multiple-input single-output (MISO) communication system is investigated. In contrast to the existing ideal STAR-RIS model assuming an independent transmission and reflection phase-shift control, a practical coupled phase-shift model is considered. Then, a joint active and passive beamforming optimisation problem is formulated for minimising the long-term transmission power consumption, subject to the coupled phase-shift constraint and the minimum data rate constraint. Despite the coupled nature of the

phase-shift model, the formulated problem is solved by invoking a hybrid continuous and discrete phase-shift control policy. Inspired by this observation, a pair of hybrid RL algorithms, namely the hybrid deep deterministic policy gradient (hybrid DDPG) algorithm and the joint DDPG & deep-Q network (DDPG-DQN) based algorithm are proposed. The hybrid DDPG algorithm controls the associated high-dimensional continuous and discrete actions by relying on hybrid action mapping. By contrast, the joint DDPG-DQN algorithm constructs two MDPs relying on an inner and an outer environment, thereby amalgamating the two agents to accomplish a joint hybrid control. The proposed hybrid beamforming algorithm outperformed the conventional continuous beamforming approach in terms of energy efficiency.

4. A STAR-RISs aided downlink NOMA communication framework is proposed, where two STAR-RIS protocols are investigated, namely the energy splitting (ES) and the mode switching (MS). However, since the STAR-RIS has a massive number of reconfigurable elements, the passive beamforming problem has enormous action dimensions and extremely high complexity, resulting in an increased training time and performance degradation for the artificial intelligent agent. To resolve this predicament, a partitioning approach is proposed to divide the STAR-RIS into several tiles. A distributed learning approach is conceived for the partitioning and the corresponding tile-based passive beamforming of each STAR-RIS, as well as the power allocation for users to maximise the average throughput achieved by multiple base stations. In particular, the deep reinforcement learning (DRL) agent is employed at each BS, and an access-free federated learning (AFFL) model, which gathers the features of FL and transfer learning (TL) is proposed to accelerate or even exempt the training process for agents. When STAR-RIS has a massive number of STAR elements, the proposed joint partitioning and tile-based beamforming outperforms the conventional element-based beamforming with significantly lower complexity.

1.4 Outline of the Thesis

Chapter 1 introduces an overview including research motivations, methodologies, and main contributions of this thesis.

Chapter 2 summarizes background and a state-of-the-art for the RIS-assisted networks and the related machine learning algorithms.

Chapter 3 presents a RIS-assisted multi-user downlink communication system over fading channels, where the time overhead for configuring the RIS reflective elements at the beginning of each fading channel is considered. A DL approach and an RL approach are proposed to solve the beamforming problem of the RIS and their performance is investigated.

Chapter 4 describes a novel framework of mobile RISs-enhanced indoor wireless networks, where a RIS mounted on a robot is invoked to enable mobility of the RIS and enhance the service quality for mobile users. To optimise the sum rate of all users, an FL enhanced DDPG algorithm is invoked to optimise the deployment and phase shifts of the mobile RIS as well as the power allocation policy.

Chapter 5 proposes a novel STAR-RIS model. A practical coupled phase-shift model of the transmitting and reselecting signals is investigated. A pair of hybrid DRL algorithms are proposed for jointly optimising the transmitting and reselecting beamforming.

Chapter 6 conceives a low-complexity beamforming approach that employs the DRL method to intelligently divide STAR elements into multiple tiles, solving the beamforming matrix by regarding tiles as units.

Chapter 7 presents the conclusion including contributions, publications, and a plan for future work.

1.5 Author's Publications

Journal papers

[1] **R. Zhong**, Y. Liu, X. Mu, Y. Chen and L. Song, "AI Empowered RIS-Assisted NOMA Networks: Deep Learning or Reinforcement Learning?" in *J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 182-196, Jan. 2022.

[2] **R. Zhong**, X. Liu, Y. Liu and Y. Chen, "Multi-Agent Reinforcement Learning in NOMA-aided UAV Networks for Cellular Offloading," in *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1498-1512, March 2022.

[3] **R. Zhong**, X. Liu, Y. Liu, Y. Chen and X. Wang, "Path Design and Resource Management for NOMA enhanced Indoor Intelligent Robots", in *IIEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8007-8021, Oct. 2022.

[4] **R. Zhong**, Y. Liu, X. Mu, Y. Chen, X. Wang and L. Hanzo, "Hybrid Reinforcement Learning for STAR-RISs: A Coupled Phase-Shift Model Based Beamformer", in *J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2556-2569, Sept. 2022.

[5] **R. Zhong**, X. Liu, Y. Liu, Y. Chen and Z. Han, "Mobile Reconfigurable Intelligent Surfaces for NOMA Networks: Federated Learning Approaches", in *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 10020-10034, Nov. 2022.

[6] **R. Zhong**, X. Mu, Y. Liu, Y. Chen, J. Zhang and P. Zhang, "STAR-RISs Assisted NOMA Networks: A Distributed Learning Approach", in *IEEE J. Sel. Top. Signal Process.*, vol. 17, no. 1, pp. 264-278, Jan. 2023.

[7] **R. Zhong**, X. Mu, Y. Chen and Y. Liu "Semantic Multiple Access", submitted to *IEEE Communi. Lett.*.

[8] M. Zhang, **R. Zhong**, X. Mu and Y. Liu "Heterogeneous Semantic and Bit Communication for Extended Reality: OMA and NOMA", submitted to *IEEE J. Sel. Top.*

Signal Process..

Conference papers

[1] **R. Zhong**, X. Liu, Y. Liu and Y. Chen, "NOMA in UAV-aided cellular offloading: A machine learning approach", *IEEE Globecom Workshops*, 2020, pp. 1-6.

[2] **R. Zhong**, X. Liu, Y. Liu, D. Zhang and Y. Chen, "Path Design for NOMA-Enhanced Robots: A Machine Learning Approach with Radio Map", *IEEE ICC Workshops*, 2021, pp. 1-6.

[3] **R. Zhong**, X. Liu, Y. Liu, Y. Chen and Z. Han, "Federated Learning Empowered Mobile RISs for NOMA Networks", *IEEE ICC*, 2022, pp. 4956-4961.

[4] **R. Zhong**, X. Mu, X. Xu, Y. Chen and Y. Liu, "STAR-RISs Assisted NOMA Networks: A Tile-based Passive Beamforming Approach", *International Symposium on Wireless Communication Systems (ISWCS)*, 2022, pp. 1-6.

[5] Y. Liu, **R. Zhong** and M. Jaber, "A Reinforcement Learning Approach for Energy Efficient Beamforming in NOMA Systems", *IEEE Globecom*, 2022, 3827-3832.

[6] M. Zhang, **R. Zhong**, X. Mu, Y. Chen and Y. Liu, "Resource Management for Heterogeneous Semantic and Bit Communication Systems", accepted by *IEEE ICC*, 2023.

Chapter 2

Background and the State-of-the-Art

This chapter summarizes the background and a state-of-the-art on related fields of this thesis. This chapter first briefly reviews the literature on RIS assisted networks, as well as STAR-RIS assisted wireless networks. Then the recent development of NOMA and RIS-aided NOMA networks is introduced. Finally, a number of state-of-the-art ML approaches for optimising communication systems are reviewed in the last section.

2.1 Background of RIS

The RIS is regarded as a promising technique in recent years since it offers an approach to artificially revise the fading channel to improve the channel gain and eliminate interference among users. In addition, RISs can provide noticeable gains since they can provide additional line-of-sight (LoS) paths for users who do not originally have an LoS path [24]. Compared with other candidate technologies for coverage extension, such as active relays (e.g. UAV relays, WIFI extenders), RIS has advantages in terms of economical cost [25] and energy efficiency [3, 26] due to its portable shape, simple structure, and

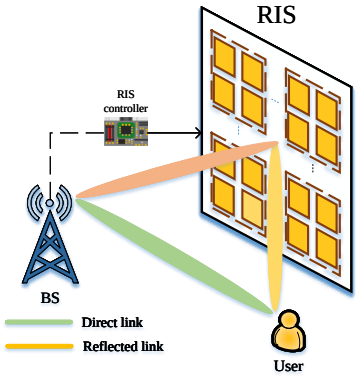


Figure 2.1: Model of reflecting-RIS assisted wireless networks

passive features, respectively. Furthermore, RISs do not need to take up a considerable space, these thin panels can be installed or attached to the wall, furniture or even clothes [27]. Therefore, RISs have been considered as a potent candidate technique for future wireless networks.

2.1.1 Reflecting RIS

RIS is a type of smart surface that can manipulate electromagnetic waves to control and optimise wireless communication systems. RIS is composed of a large number of small elements or units, which can be electronically adjusted to alter the reflection, refraction, and absorption of electromagnetic waves as shown in Fig. 2.1. These surfaces are capable of dynamically modifying the propagation characteristics of radio waves in their vicinity, enabling them to adapt to changing wireless communication environments and improve system performance. The basic principle of RIS is to use passive elements with electronically reconfigurable properties to modify the behavior of wireless signals. RIS can be made up of different types of materials, such as conductive or dielectric materials, and each unit or element is equipped with an electronic control circuit that can adjust its properties. These control circuits can be programmed to manipulate the phase, amplitude, and polarization of the incident waves, allowing the RIS to selectively

reflect, refract, or absorb wireless signals in a controlled manner.

RISs can be deployed in various scenarios, including indoor and outdoor environments, and can be integrated into different types of communication systems, such as cellular networks, Wi-Fi, and Internet of Things (IoT) networks. They have the potential to significantly improve wireless communication performance by mitigating signal interference, extending coverage range, increasing spectral efficiency, and enhancing security and privacy [28].

2.1.2 From Reflecting-RIS to STAR-RIS

The performance of RISs and wireless relays has been comprehensively compared due to their similar functions and roles [26, 29]. Despite the appealingly low complexity and noise figure [26], an undeniable fact is that the reflecting-only RISs are likely to have a coverage disadvantage compared to omnidirectional relays due to their 180° half-plane reflection limitation. However, in practical scenarios, users roam at both sides of the RIS, and the reflecting-only RIS cannot provide signal enhancements for users located at the back of the RIS. This deficiency of reflecting-only RISs inspires the design of double-sided RISs to achieve 360° coverage boundary.

Hence, a simultaneous transmitting and reflecting reconfigurable intelligent surface (STAR-RIS) model emerged as an ameliorated version of the reflecting-only RIS [30]. As an advanced derivative, an additional transmission function is facilitated by the simultaneous transmitting and reflecting (STAR) elements. The STAR-RIS accommodates a number of STAR elements on a surface to separate the incident electromagnetic waves into transmitted and reflected signals. Consequently, the transmitted signals and reflected signals can form a pair of sectors on both sides of the surface simultaneously, thereby solving the limitations of RISs in terms of their coverage region. In contrast to the reflecting-only RIS, which has to be in the vicinity of either the base station (BS) or the users, STAR-RIS may be positioned flexibly. Furthermore, if the BS or users

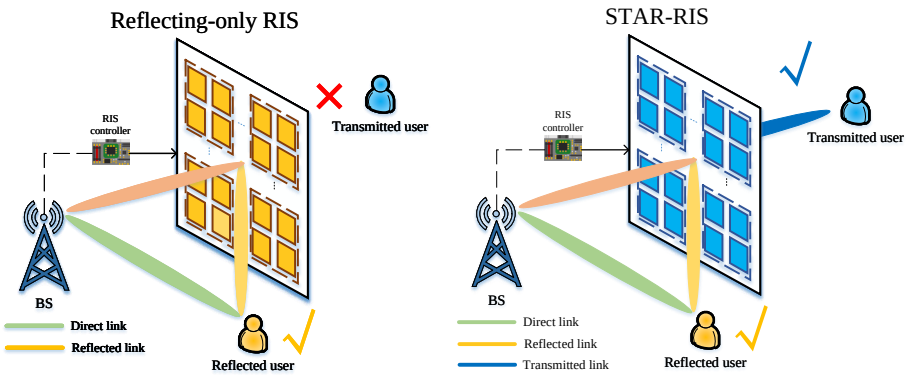


Figure 2.2: Model of STAR-RIS assisted wireless networks

are out of the optimal orientation of the RIS, the gain provided by the reflecting-only RIS deteriorates distinctly [31]. Fortunately, the emergence of STAR-RISs can relax the orientation requirement of (STAR) RISs[1], since the transmission and reflection sectors have the opposite orientation.

2.1.3 STAR-RIS

STAR-RIS has the capability of controlling and manipulating electromagnetic waves by simultaneously transmitting and reflecting them. This allows RIS to perform both functions of transmitting and reflecting the incident signals, providing coverage of signals at both sides of the STAR-RIS as shown in Fig. 2.2. The basic concept behind STAR-RIS is that each STAR element in the RIS array can be configured to dynamically adjust its properties to transmit and reflect the incident electromagnetic waves. This is achieved by electronically controlling the phase, amplitude, and polarization of the incident waves at each RIS element. By carefully designing the configuration and operation of the RIS, it is possible to achieve desired beamforming and beam steering effects for both the transmitted and reflected waves.

[1]The abbreviation '(STAR) RISs' refers to both reflecting-only RISs and STAR-RISs.

2.1.4 Multiple Access for RIS Engaged Communications

The multiple access (MA) schemes of the RIS engaged future wireless communication network is a critical problem since the number of mobile terminals and Internet-of-Things (IoT) devices maintains a continuously increasing trend. On one hand, various orthogonal multiple access (OMA) technologies are widely adopted by existing cellular networks that can be inherited to RIS-assisted networks. On the other hand, another promising candidate, the non-orthogonal multiple access (NOMA) technique has potential advantages in RIS-assisted communication systems [32, 33]. To tackle the explosively increasing number of user equipment in next-generation wireless networks, with the advantages of high spectral efficiency and high flexibility, NOMA has been envisioned as an effective solution [34, 35]. By serving users via the same time/frequency/code resource block, NOMA can significantly outperform conventional OMA based counterparts in terms of average sum-rate and outage probability [36, 37].

There are potential compatibility and affinity between the RIS and NOMA [38]. As pointed out by the authors of [38], affinities between RISs and the NOMA scheme include that RISs can provide additional signal diversity, desired channel condition, and undemanding multi-antenna constrain for NOMA systems. The artificial intervention channels and additional degrees-of-freedom (DoF) provided by RIS offer an opportunity to further improve the NOMA gain, and it also enables the communication quality of NOMA users to be determined by quality of service (QoS) requirements instead of being restricted by the natural channel gain [38]. In a conventional NOMA enhanced wireless network, the decoding order of successive interference cancelation (SIC) is determined by the natural channel state information (CSI) of users, which is not likely to be fully consistent with the users' data rate demand. However, RISs can artificially modify the CSI for each user and thereby provide desired propagation condition for superposed signals. Therefore, motivated by the aforementioned reasons NOMA techniques are invoked in the mobile RIS model to obtain further capacity and data rate gains.

2.2 RIS-assisted Communications

2.2.1 Passive Beamforming in Reflecting RISs

The ascendant of RIS has spawned a number of related academic contributions in recent years, especially in the passive beamforming design [39]. The authors of [40] studied the joint optimisation of the reflection coefficients and the resource block allocation for a RIS-assisted orthogonal frequency division multiple access (OFDMA) network to maximise the minimum rate among all users. In [10], the authors proposed a robust beamforming for the transmitter to against the imperfect cascaded BS-IRS-user channels. The authors of [41] proposed a framework of self-sustainable RIS by adopting the energy harvest technique, and a computationally-efficient iterative algorithm was proposed to maximise the sum data rate as well as maintaining the self-sustainability of the RIS. The authors of [13] proposed multiple RISs collaboration paradigms and theoretically analysed that the double-RIS with cooperation has the ability to capture superior signal to the SINR, compared to the single-RIS paradigm. To counteract the interference caused by RISs, the authors of [42] employed a dynamic rate splitting method to improve the energy efficiency for the system. In addition to theoretical research, some entities have also been constructed. For example, in [43] the authors constructed a RIS prototype with 1100 controllable elements, and their experimental results could prove the potential and effectiveness of RISs.

2.2.2 STAR-RIS-assisted Communications

The STAR-RIS is regarded as an emerging and promising technique, and a number of research contributions on STAR-RISs have been published due to its coverage advantage compared with reflecting RISs [8, 44] and refracting RISs [45]. The authors of [46] proposed a design of STAR elements that can split the incident signal into the reflected signal and the transmission signal, which inspired the proposal of the energy splitting

(ES) mode of STAR-RISs. Another simpler scheme to achieve STAR is to switch the mode of elements through binary switches that arrange a part of elements working in reflection mode and others in transmission mode [30, 47]. A novel bilayer structure was proposed for the STAR-RIS in [48] to achieve almost independently reflection and refraction beamforming, and the spectral efficiency was maximised by a weighted mean square error minimisation approach. An indoor communication network model was proposed in [49], where a STAR-RIS embedded in the wall was invoked to enhance the signals from two independent access points (APs) to suppress inter-cell interference. The joint transmission and reflection beamforming is one of the main research directions of the STAR-RIS, since beamforming and interference suppression of the STAR-RIS are more complicated than unidirectional RISs. Several researchers focused on the correlated phase-shift optimisation for the STAR-RIS, an efficient element-wise alternating optimisation algorithm was proposed for NOMA and OMA networks [50], and a convex optimisation algorithm was proposed in [12] for the joint beamforming of the STAR-RIS and the multi-antenna small base station (SBS) to extend the coverage of the SBS. Moreover, most recently, three transmission and reflection coefficients (TRC) configuration strategies, namely the primary-secondary phase-shift configuration, the diversity preserving phase-shift configuration, and the T/R-group phase-shift configuration strategies were proposed in [51] for the STAR-RIS assisted NOMA networks.

2.3 RIS for NOMA Networks

2.3.1 NOMA Enhanced Wireless Communications

Given the trend that the number of users in cellular networks continues increasing, NOMA techniques are considered to be suitable for future communications since the huge number of devices requires enormous capacity [52, 53]. Furthermore, NOMA techniques are regarded as one of the candidate technologies for the next generation wireless communication since it has superior spectral efficiency and user fairness as well [54].

The author of [55] developed the analytical frameworks of up-link and down-link NOMA in a dense wireless network and proved the NOMA gain in terms of achievable rate and outage probability. Due to the high sensitivity on power, the power control policy of the NOMA enhanced system is a kernel of the optimisation. Therefore, a series of related studies were proposed to optimise the delay [56], outage probability [57], and energy efficiency [58] of the NOMA enhanced communication system.

2.3.2 RIS-assisted NOMA Networks

As the combination of RISs and NOMA techniques is considered promising, a series of related research contributions have been proposed in the past years. To combine the advantages of RISs and NOMA, authors of [59] proposed a new RIS-aided downlink NOMA system to improve the reliability of the wireless network, and they derived the analytical expression of the bit error rate (BER) performance of RIS enhanced NOMA systems. The author of [60] investigated the physical layer security of a RIS enhanced NOMA system. A NOMA based model of RIS-unmanned aerial vehicles (UAVs) communication was proposed in [61], where the authors deployed RISs on the outer surface of a skyscraper to assist the wireless link of UAVs. UAVs' trajectories, passive beamforming of the RISs, and power allocation are treated as optimisation variables to minimise the energy cost of the UAVs. A partitioned RIS was employed in [62] to enhance the spectrum efficiency by improving the ergodic rate of all users, and the physical resource distribution was optimised by three efficient search algorithms.

The authors of [63] optimised user clustering, passive beamforming, and power allocation for a downlink NOMA system with RISs by iteratively optimising three subproblems. A RISs enhanced NOMA cellular network with the joint transmission coordinated multipoint was proposed in [64] to improve the data rate of edge users, while the network spectral efficiency was evaluated and validated through Monte-Carlo simulations. Meanwhile, in [65] and [11], joint optimisations of the base station beamforming and the passive beaming at the RIS were proposed with the aim of minimising the total

transmit power of the base station.

The authors of [66] proposed an algorithm to jointly optimise the active beamforming of the base station and the passive beamforming for a RIS-assisted downlink NOMA network to maximise the data rate and ensure users' fairness. The authors of [67] employed RIS to enhance the channel for cell-edge users in a two-cell NOMA network and proposed an algorithm to minimise the total transmit power under the constraints of users' SINR demand. To solve the deployment problem of RISs, the authors of [68] proposed a monotonic-based optimisation method, where the frequency division multiple access (FDMA), and time division multiple access (TDMA), and NOMA schemes were considered.

2.4 Background of Machine learning

2.4.1 Deep Learning

DL is an important subfield of ML that focuses on neural networks. It involves training DNNs, which are composed of multiple layers of interconnected nodes, to perform tasks such as fitting, recognition, regression, etc.

The basic concepts of DL include:

Neural Networks: Deep learning models are built using DNNs, which are composed of interconnected neural nodes organized into form of layers. Each neural node receives input from its connected neural nodes. The input is processed into an output, which is passed on to the next layer.

Activation Functions: Activation functions introduce non-linearity into DNNs, enabling them to learn complex non-linear patterns from data. Common activation functions used in deep learning include ReLU (Rectified Linear Unit), sigmoid, tanh, and etc.

Backpropagation: Backpropagation is the main approach for training DNNs. It involves calculating the gradient of the loss function with respect to the parameters of the network and then using this gradient to update the weights of each node in order to minimise the loss. This process is repeated iteratively until the network converges to a expected set of weights.

Training: DL models typically require large amounts of data for training. More data allows the network to learn more accurate representations of the underlying patterns in the data, leading to superior performance.

Transfer Learning: Transfer learning is a technique in DL where a pre-trained DNN, trained on one task or domain, is used as an initialized model for training a new model on a different but related task or domain. TL can help in cases where labeled data for the target task or domain is changed.

DL has achieved state-of-the-art results in a number of fields, including computer vision, speech recognition, natural language processing, and robotics, and it is excepted as a powerful tool to resolve problems in wireless communications.

2.4.2 Reinforcement Learning

Reinforcement learning is another type of ML approach that focuses on training agents to make decisions in an environment to maximise a cumulative reward. It is commonly used in scenarios where an agent needs to interact with an environment. The continuous learning process through trial and error to take actions can lead to desirable outcomes.

The basic concepts of RL include:

Agent and Environment: In RL, there is an agent that takes actions in an environment. The agent interacts with the environment by observing its current state, taking actions, and receiving feedback in the form of rewards or penalties based on the actions it takes.

State, Action, and Reward: The state represents the current situation of the environment, which would be observed by the agent. The action represents the decision made by the agent to perform in the environment. The agent then receives a reward or penalty from the environment based on the action taken, which serves as feedback to guide the agent's learning.

Step and Episode: Each state-action-reward circulation is called a step. The RL agent continues to cycle until the desired state is reached or the maximum number of steps expires. This sequence of steps is called an episode. At the start of each episode, the environment is set to the initial state and the agent's reward is reset to zero.

MDP: Long-term problems are often modeled as Markov Decision Processes (MDPs), which formalize the interaction between an agent and an environment. MDPs are often characterized by states, actions, rewards, and transition probabilities, which can be used to model a wide range of sequential decision-making problems.

Policy: A policy is a mapping from states to actions, and it determines the behavior of the agent. The agent's goal is to learn an optimal policy that maximises the cumulative reward over time. Policies can be deterministic, where the agent always selects the same action for a given state, or stochastic, where the agent selects actions with a certain probability.

Value Function: The value function represents the expected cumulative reward that an agent can obtain from a given state or state-action pair, while following a certain policy. It is used to evaluate the desirability of different states or actions, and can be used to guide the agent's decision-making. Commonly used value functions include the state-value function ($V(s)$) and the action-value function ($Q(s, a)$).

Exploration and Exploitation: RL agents need to strike a balance between exploration and exploitation. Exploration refers to trying out different actions to discover their effects on the environment and learn about the environment, while exploitation refers to selecting actions that are expected to yield high rewards based on the agent's current

knowledge. Finding the balance between exploration and exploitation is crucial for the agent to learn an optimal policy.

RL has been applied to a wide range of applications, including robotics, game playing, recommendation systems, and autonomous vehicles, etc. It is preferred in wireless communication systems since it does not need to collect huge data samples for training but it can be training by interreact with environment to resolve the practical problems.

2.4.3 Federated Learning

Federated learning is a machine learning framework that enables models training across multiple decentralized devices or servers while keeping the data local. In federated learning, the training process takes place on the edge devices without requiring the data to be sent to a central server. The FL framework has a number of common advantages for all ML algorithms, for example, it can save more hardware resources, improve training speed, and with the protection of user privacy [69]. The basic working flow of FL includes:

Initialization: A central server or authority distributes an initial model to the participating devices. This model is usually pre-trained on a large dataset and serves as a starting point.

Local Updates: Each device performs model training using its local data without sharing the data itself. This training step is typically performed through multiple iterations or epochs, using techniques like stochastic gradient descent.

Global Aggregation: After local training, the devices send only the updated model parameters (not the raw data) back to the central server. The server aggregates these model updates using techniques like averaging or weighted averaging, creating a global model update.

Local Model Update: The central server incorporates the aggregated model update into the global model. The updated model is then redistributed to the devices for the

next round of local training.

Iteration: The process of local training, model aggregation, and model update is repeated for multiple iterations, allowing the model to improve over time.

2.5 Machine learning for Wireless Network Optimisations

2.5.1 DL for RIS-assisted Networks

DL has been successfully applied in RIS-assisted networks [5, 70, 71]. DL is primarily used for channel estimation in RIS-assisted networks, in order to extract the channel state information (CSI) in support of RIS-based passive beamforming. Gao *et al.* [72] employed a synthetic deep neural network (DNN) for sequentially estimating the BS-RIS channel and RIS-user channel, while reducing the pilot overhead and guaranteeing the estimation accuracy. Obtaining the channel state information (CSI) with the assistance of ML algorithms aroused the interest of researchers, since the channel estimation is challenging for the conventional minimum mean square error (MMSE) estimator in RIS-assisted networks [73, 74]. Thus, DL was widely applied for the channel estimation for RIS-assisted networks [75, 76]. Furthermore, a deep transfer learning (DTL) algorithm was proposed in [77] and the DTL method is able to achieve a comparable bit error rate (BER) performance with the optimal detection method with perfect CSI. DL is used to circumventing CSI instead of inferring, which can reduce the transceiver's dependence on CSI [78] and even work without instantaneous CSI feedback [79].

In addition to using DL for the estimation, using DL to optimize the performance of communication systems is also a major research direction [80]. In [81], a DL approach was proposed for the resource management problem of vehicle networks. Moreover, a communication deep neural network (CDNN) approach was proposed in [82] for the energy efficiency optimization in a MIMO-NOMA network. To be more specific, DL was also invoked for the phase-shift control of RISs. For example, an unsupervised

DL algorithm was proposed for joint active and passive beamforming optimisation in [83]. The passive beamforming design, as the main research problem of RIS, has also induced some deep learning solutions. The authors of [84] proposed a DL solution for the joint active and passive beamforming design for a RIS-assisted downlink multiple-input single-output (MISO) network. Moreover, a DL-empowered predictive beamforming method was proposed in [85] to maximise the average sum-rate. In [86], a well-trained convolutional neural network (CNN) was employed to help RISs infer the interfering signals from the incident signals and thereby maximise the received SINR.

2.5.2 DRL for RIS-assisted Networks

Compared with DL algorithms, RL and deep reinforcement learning (DRL) algorithms were employed to optimise diverse optimisation variables, including the deployment of the aerial RIS [87], the secrecy rate of communication systems [88], and the user scheduling [89]. In [90], three machine learning algorithms were jointly invoked to maximise the throughput of a RIS-assisted NOMA network, where a long short-term memory (LSTM) algorithm was adopted to predict the position of users, a K-means based Gaussian mixture model was used to determine users' clustering and a deep Q-network (DQN) was in charge of optimising the phase shift matrix and power allocation policy. In [61], the authors envisaged deploying RISs at high altitudes to establish a RIS-NOMA network for UAVs, and a DRL algorithm was proposed to jointly optimise UAVs' trajectories, the RIS's passive beamforming, and the NOMA resource allocation. Similar with [61], an RISs assisted UAV communication system was invoked in [91] that a UAV and multiple RISs were paired to serve a number of ground users and two DRL approaches were adopted to maximise the overall weighted data rate and geographical fairness of by optimising the UAV's trajectory and phase shifts of RISs.

2.5.3 DRL for Beamforming Problem

DRL has demonstrated commendable performance in various wireless network systems [92]. By invoking a DRL approach, the authors of [93] investigated the joint design of transmit beamforming at the base station and the phase shifts at the RIS. By adopting a deep deterministic policy gradient (DDPG) algorithm, Huang *et al.* [5] conceived the joint optimisation of the transmit beamforming matrix of the BS and the phase-shift matrix of the RIS. Yang *et al.* [94] proposed a deep Q network (DQN) for secure beamforming guarding against eavesdroppers in dynamic environments. Their simulation results verified that the DQN approach is capable of enhancing the secrecy rate and the satisfaction probability of users. Furthermore, a DRL scheme was invoked in [95] for optimising a RIS-assisted NOMA network, where a long short-term memory (LSTM) based echo state network (ESN) algorithm collaborated with a decaying double deep Q-network (D3QN) for intelligently controlling the RIS according to the users' data demand. In [17], a proximal policy optimisation (PPO) algorithm was developed for minimising the expected Age-of-Information (AoI) for an aerial RIS. Specifically for the STAR-RIS scenario, a federated learning algorithm was proposed in [96] for maximising the achievable data rate of a STAR-RIS assisted heterogeneous NOMA network. The author of [97] proposed a hill-climbing algorithm to optimise the power allocation at the base station and reflecting beamforming to achieve an anti-jamming communication. The author of [98] optimised the reflecting beamforming of a RIS by an RL approach and their results suggest that the participation of the RL significantly improves the secrecy and quality of service (QoS) satisfaction probability of the system. In [99], a dueling double deep Q-network algorithm was proposed to optimise the user association and resource allocation to maximise the long-term performance of the cellular network. The authors of [100] employed the DDPG algorithm to resolve the energy efficiency maximisation problem of a STAR-RIS assisted NOMA downlink network by jointly optimising the active beamforming at the BS and TRCs at the STAR-RIS.

2.5.4 DRL for Deployment Planning

Invoking RL algorithms for deployment planning has achieved several remarkable successful cases [101, 102]. In [103], the authors applied Q-learning and State-Action-Reward-State-Action (SARSA) algorithms to plan motions for a swarm of robots so that they can be deployed to a user-defined target distribution timely. The author of [104] adopted a deep reinforcement learning (DRL) algorithm with an actor-critic structure to optimise a complete coverage path for the tiling robot with minimal energy cost. In addition to robots, RL algorithms have been used to design the trajectory for other kinds of craft, such as vehicles [105, 106] and UAVs [107–110]. From a perspective of the communication, since RL algorithms have a prominent ability to solve non-convex problems, it manifests extraordinary potential in deployment optimisation to ensuring the communication quality [111, 112].

2.5.5 Deployment of RISs

The authors of [113] employed statistical geometric tools to investigate the impact of large-scale deployment of RIS on cellular network performance. It is theoretically proved that the deployment of RIS greatly improves the coverage performance of BS. A deployment optimisation of the RIS assisted relay system was proposed in [114], where the optimal strategy is to deploy the RIS near the relay. The authors of [95] optimised the deployment of the RIS by DRL algorithm. Although the 3D position of the RIS is calculated, how to arrange the RIS to the optimal position with height is not considered. Another work focused on RIS 3D deployment is presented in [115], theoretically revealing the statistical optimal position of RIS in single-input single-output systems. Obtaining more LoS coverage by optimizing RIS deployment is studied in [116], which has a similar scope to Chapter 4, but the location and behaviors of users are not considered in [116].

2.5.6 FL in Wireless Networks

FL was also considered to be invoked in the wireless network due to its advantages [117], and some research contributions on FL in wireless networks have been proposed [118]. The author of [119] proposed an FL approach to estimate channels for a RISs-assisted massive Multi-input Multi-output (MIMO) system. To optimise the data rate of RIS-assisted networks, the authors of [120] proposed an FL-based beam reflection optimisation algorithm to achieve high-speed communication with the sparse CSI. In addition, since the FL process needs to exchange data between agents via wireless networks, research on how to use wireless communications to support federated learning is also challenging [121–123]. The authors of [122] formulated the joint learning and communication problem, and proposed an iterative algorithm to minimise the total energy consumption for an FL-based system. For example, FL was adopted for channel estimation in a RIS-assisted massive multi-input multi-output (MIMO) network, and the simulation results indicate the FL framework provides significantly lower transmission overhead than the centralized learning framework.

2.5.7 Distributed DRL for Wireless Networks

DRL algorithms are considered potential candidates for solving optimisation and control problems in future communication systems [124], and the distributed DRL is preferred in the wireless network due to the superior convergence [125]. Distributed proximal policy optimisation (DPPO) [126], which is broadly recognized for its stable and fast convergence, was adopted in wireless networks to optimise spectral efficiency in [127]. However, the disadvantage of this distributed framework is that the agents and the chief need to exchange policy gradients (PG) frequently, which is likely to cause significant communication overhead and synchronization problems. A DRL-based distributed dynamic downlink-beamforming coordination approach was introduced in [128] for the beamforming problem in a downlink MISO network.

2.6 Knowledge Gap and Distinctions of This Thesis

Although there are a number of existing research contributions on RIS assisted wireless networks, there are still many important scientific questions that have not been addressed. Existing research that has already been done on RIS beamforming has the following deficiencies, including **Limitation 1.** Most existing research does not consider imperfect problems in practice (e.g. configuration overhead and constrained reflection coefficient values); **Limitation 2.** Apply statistical models (e.g. random user distributions) rather than any specific case; **Limitation 3.** In most existing research contributions, RISs are assumed to be fixed on a wall or other bearing, and the rigid deployment may prevent RIS from obtaining LoS paths and optimal channel enhancement, especially in an environment with obstructions; **Limitation 4.** The coverage of the reflection beam from RIS is limited; **Limitation 5.** The complexity of existing beamforming schemes is generally high. If a RIS has a massive number of reconfigurable elements, the complexity of the beamforming algorithm will be unacceptable.

Since the existing research still has deficiencies in the above problems, this thesis aims to investigate the configuration overhead problem, mobile deployment, two-directional beamforming, and low-complexity beamforming for RISs in the following four chapters. Various practical concerns, such as signaling and configuration overhead, non-ideal responsiveness of programmable elements, etc., are taken into account in Chapter 3 and Chapter 5, respectively, against **Limitation 1**. Second, a mobile RIS model is considered with modelling of specific surroundings in Chapter 4, instead of using fixed RIS and statistical models in **Limitation 2** and **Limitation 3**. In Chapter 5, a STAR-RIS model is proposed for coverage extension against **Limitation 4**, and the joint transmitting and reflecting beamforming of STAR-RIS has to be resolved in order for the beam to effectively cover both sides of the RIS panel. In Chapter 6, how to reduce the complexity is considered against **Limitation 5**.

Chapter 3

Passive Beamforming with configuration overhead

This chapter introduces a RIS-assisted multi-user downlink communication system over fading channels, where the time overhead for configuring the RIS reflective elements at the beginning of each fading channel is considered. A DL approach and an RL approach are proposed to solve the beamforming problem of the RIS and their performance is compared.

3.1 Configuration Overhead Problem of the RIS

Employing the RIS in wireless networks, the existing literature mainly focused on the optimisation for one specific channel coherence block (i.e., static channel scenario) but ignored the time overhead for configuring the reflection coefficients of the RIS. However, given the fact that RISs are equipped with a large number of reflective elements, it leads to a non-negligible time overhead for configurations, thus degrading the performance of data transmission. Sources of the configuring overhead include but are not limited to programmable controllers [14] or shift registers. This is also confirmed by

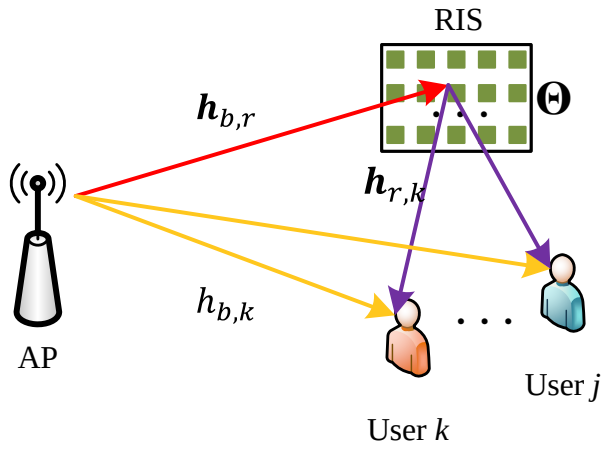


Figure 3.1: Schematic of the RIS-assisted multi-user downlink communication system.

the prototype experiment conducted by the research team of Massachusetts Institute of Technology [129]. By taking the RIS configuration overhead into consideration, a fundamental tradeoff problem is emerged. Specifically, although totally reconfiguring the reflection coefficients of the RIS can well match the current channel condition and achieve a high instantaneous data rate. However, the precise configuration reduces the time duration for data transmission due to the time overhead caused by the reconfiguration of each RIS element. The situation can be even worse for practical data transmission over fading channels. Since if the reflection coefficient of one element is selected not to be reconfigured at the current channel fading, it will reserve the previous value. Therefore, a sophisticated transmission policy has to be developed to balance the configurations between different channel fadings and to achieve a long-term performance gain.

3.2 System Model and Problem Formulation

3.2.1 System Model

As shown in Fig. 3.1, a RIS-assisted multi-user downlink network is considered, where a single-antenna access point (AP) equipped with a single antenna provides service to K single-antenna users and a RIS with M passive reflecting elements are employed to enhance the channel. The passive beamforming of the RIS is controlled by an attached smart controller, which generates the configuration instruction for the reflective elements. The locations of the AP, RIS, and users are denoted by $(x_{ap}, y_{ap}, z_{ap})^T$, $(x_r, y_r, z_r)^T$, and $(x_k, y_k, z_k)^T$, respectively, in a three-dimensional (3D) Cartesian coordinate system. In particular, the locations of the RIS and the AP are fixed, and users are randomly moving over different time slots and are assumed to be static for each time slot. As a consequence of the location variation, all channels are assumed to follow the quasi-static block fading channel model, and the channel coefficient remains approximately constant in each channel coherence block and varies independently from block to block. Specifically, one particular transmission period $\mathcal{T}$ is considered, which consists of N channel blocks with the duration of T, i.e., $\mathcal{T} = NT$. Thus, in order to fit the fading channel blocks, the phase shifts of the RIS has to be reconfigured at the beginning of each channel block[1].

The RIS's diagonal reflection matrix for the n-th block is denoted by

$$\mathbf{\Theta}[n] = \mathrm{diag}\left(e^{j\theta_1[n]}, e^{j\theta_2[n]}, \cdots, e^{j\theta_M[n]}\right), \tag{3.1}$$

where $\theta_m[n] \in [0, 2\pi)$ denotes the reflection phase shift of the m-th element. In this chapter, the values of $\theta_m[n]$ are assumed to be continuous in the interval $[0, 2\pi)$. In practice, this value may have to come from a fixed discrete set, such as $\{0, \pi\}$. The

[1]Since the scope of this research is not channel estimation, CSI knowledge is assumed to be known. The channel estimation method can refer to the approaches mentioned in Chapter 2. In practice, the CSI need to be measured at the beginning of each fading block.

proposed scheme only needs to add a simple decision threshold to meet the requirements. If less configuration precision is applied, less configuration overhead is required, but at the expense of configuration precision, which will lead to a trade-off problem.

3.2.2 Channel Model

Let $h_{b,k}[n] \in \mathbb{C}^{1\times 1}$ and $\mathbf{h}_{b,r}[n] \in \mathbb{C}^{M\times 1}$ denote the corresponding AP-user k and AP-RIS channels for the n-th block. Correspondingly, the channel between the RIS and the user k is denoted by $\mathbf{h}_{r,k}[n] \in \mathbb{C}^{M\times 1}$. Since the positions of AP and RIS are selected, and the distance between RIS and users is relatively close, we assume that the channels from AP to RIS and from RIS to users following Rician fading. On the contrary, since the AP can be far away from the user and the user roams randomly, there is a high possibility of occlusion between them, and the channel from the AP to the user is assumed to be a Rayleigh channel. Thus, assuming that direct channel follows Rayleigh fading and the RIS-assisted channels are modelled as Rician fading [130], the channels can be modeled as

$$h_{b,k}[n] = \sqrt{PL\left(d_{AU,k}\right)} h_{b,k}^{\mathrm{NLoS}}[n], \tag{3.2a}$$

$$\mathbf{h}_{b,r}[n] = \sqrt{PL\left(d_{AR}\right)}\left(\sqrt{\frac{K_{AR}}{K_{AR}+1}}\mathbf{h}_{b,r}{}^{\mathrm{LoS}} + \sqrt{\frac{1}{K_{AR}+1}}\mathbf{h}_{b,r}{}^{\mathrm{NLoS}}[n]\right), \tag{3.2b}$$

$$\mathbf{h}_{r,k}[n] = \sqrt{PL\left(d_{RU,k}\right)}\left(\sqrt{\frac{K_{RU}}{K_{RU}+1}}\mathbf{h}_{r,k}^{\mathrm{LoS}} + \sqrt{\frac{1}{K_{RU}+1}}\mathbf{h}_{r,k}^{\mathrm{NLoS}}[n]\right), \tag{3.2c}$$

where the distances for the direct AP-user link, the AP-RIS link, and the RIS-user link are denoted by $d_{AU,k}$, d_{AR}, and $d_{RU,k}$, respectively. The propagation path loss for all channels is modeled as $PL(d) = \rho_0\left(\frac{d}{d_0}\right)^{-\alpha}$ [131], where ρ_0 represent the path loss at the reference distance, d denotes the distance between the transmitter and the receiver, α represents the path loss exponent, and K_{AR}, K_{RU} denote the Rician factors [132]. $\mathbf{h}_{b,r}{}^{\mathrm{LoS}}$, and $\mathbf{h}_{r,k}^{\mathrm{LoS}}$ are the deterministic line-of-sight (LoS) components, which can be

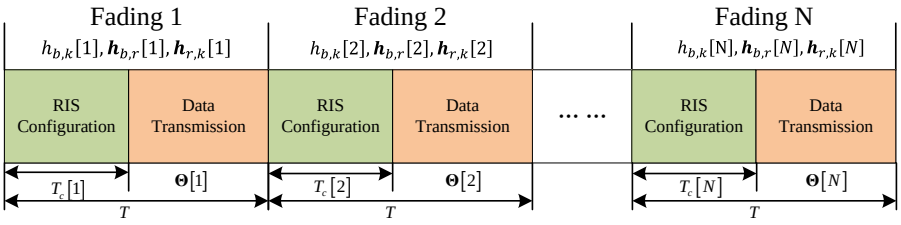


Figure 3.2: Illustration of the transmission over N fadings.

calculated as

$$\mathbf{h}_{b,r}{}^{\mathrm{LoS}} = \left[1, \ldots, e^{j(m-1)\pi \sin(a_{AR})} \ldots, e^{j(M-1)\pi \sin(a_{AR})}\right]^T,$$

$$\mathbf{h}_{r,k}^{\mathrm{LoS}} = \left[1, \ldots, e^{j(m-1)\pi \sin\left(a_{RU,k}\right)}, \ldots, e^{j(M-1)\pi \sin\left(a_{RU,k}\right)}\right]^T,$$

where $\sin(a_{AR}) = \frac{y_r - y_{ap}}{\sqrt{(x_r - x_{ap})^2 + (y_r - y_{ap})^2}}$ and $\sin(a_{RU,k}) = \frac{y_k - y_r}{\sqrt{(x_k - x_r)^2 + (y_k - y_r)^2}}$. $h_{b,k}$, $\mathbf{h}_{b,r}{}^{\mathrm{NLoS}}$, and $\mathbf{h}_{r,k}^{\mathrm{NLoS}}$ are the non-line-of-sight (NLoS) components modeled as Rayleigh fading. Each element of $\mathbf{h}_{b,r}{}^{\mathrm{NLoS}}$ or $\mathbf{h}_{r,k}^{\mathrm{NLoS}}$ follows $\mathcal{CN}(0, 1)$.

Therefore, the combined channel power gain from the AP to the user k with the aid of the RIS for the n-th block is given by

$$h_k[n] = \left| h_{b,k}[n] + \mathbf{h}_{r,k}^H[n] \boldsymbol{\Theta}[n] \mathbf{h}_{b,r}[n] \right|^2. \tag{3.3}$$

3.2.3 RIS Configuration Model

The communication model with a consideration of the time overhead for configuring the RIS is investigated. As illustrated in Fig. 3.2, at the beginning of each fading block, the reflection matrix of the RIS is configured by the controller (intelligent agent). Once the elements are configured, the AP can start the transmission to users with the assistance of the configured RIS reflection matrix. Let $T_c[n]$ denote the time duration consumed by configuring the RIS in the n-th block. In general, $T_c[n]$ depends on the number of

RIS elements that need to be adjusted since it consumes time to send instructions for each adjusted element [129] . Let T_0 and $0 \leq M_c[n] \leq M$ denote the time duration for adjusting one RIS element and the number of RIS elements to be configured in the n-th block. Then, $T_c[n] = M_c[n] T_0$. Therefore, the AP can choose to only configure a part of RIS elements to reduce the configuration overhead, and keep the remaining RIS elements staying in the previous state. Let $\tau_m[n] \in \{0, 1\}$ denote whether the m-th RIS element is configured at the n-th block or not. Then, it can be given by

$$\sum_{m=1}^{M} \tau_m[n] = M_c[n], \forall n, \tag{3.4}$$

$$\theta_m[n] = \theta_m[n-1], \text{if } \tau_m[n] = 0, \ \forall m, n. \tag{3.5}$$

Since RISs may contain an enormous number of reflective elements, following equation (3.4), precise phase shift for each individual element can lead to excessive configuration overhead [2]. Therefore, in order to reduce configuration overhead, several reflective elements can be grouped into a element group, where the elements in one group are spatially adjacent and have a high channel correlation [133]. For ease of presentation, we refer to each element group as a tile. As a result, the controller can no longer design the phase shift for each reflective element, but instead, design a common phase shift for each tile. Assuming that each tile has u reflective elements, the overhead can be reduced to

$$T_c[n] = M_c[n] T_0 / u. \tag{3.6}$$

In this chapter, the reflective element number per each element tile is not considered as an optimisation variable and it would be further discussed in Chapter 6. It is also worth

[2]Parallel control circuits can also be employed to reduce the transmission time of configuration commands, but as long as the parallel control scheme cannot achieve the independent transmission of each reflective element, the control information queues still exist, then the proposed solutions can be invoked to further reduce the configuration overhead. Compared with the parallel control circuit scheme, the proposed solution has advantages in terms of hardware, complexity, and economical cost.

noting that RIS also has other overheads in each fading block, such as pilot overhead, signaling overhead, etc. The pilot overhead depends on the channel estimation method, and the signalling between the smart controller and RIS depends on the bandwidth and transmission lag of the cable. These overheads may also result in reduced transmission times. In order to focus on the configuration overhead issue, the complexity and time overheads from other source is ignored in this chapter, and they can be discussed further in future work.

3.2.4 Signal Model

The achievable throughput for both NOMA and OMA scenarios is investigated. In the NOMA scenario, a resource block is shared by the users in a cluster, while multiple users are considered to utilise an orthogonal resource block in the OMA scenario.

3.2.4.1 OMA

For the OMA scheme, the AP communicates with K users over orthogonal time/ frequency resource of equal size. Thus, the transmitted signal from the AP to user k can be denoted by

$$x^O[n] = \sum_{k=1}^{K} \sqrt{P_k[n]} s_k[n], \tag{3.7}$$

where $s_k[n]$ denotes the symbol sequence and $P_k[n]$ represents the allocated power for user k at block n. With an ideal orthogonal model, user k will not be interfered by other users but only suffer from the noise $n_0[n]$, the received signal at user k can be expressed as

$$y_k^O[n] = \left(h_{b,k}[n] + \mathbf{h}_{r,k}^H[n] \boldsymbol{\Theta}[n] \mathbf{h}_{b,r}[n]\right) \sqrt{P_k[n]} s_k[n] + n_0. \tag{3.8}$$

Thus, the achievable communication rate of user k for the OMA case at the n-th block is given by

$$R_k^O[n] = \frac{1}{K}\log_2\left(1 + \frac{p_k[n]\, h_k[n]}{\frac{1}{K}\sigma_k^2}\right), \tag{3.9}$$

where $p_k[n]$ denotes the allocated transmit power to user k at the n-th block and σ_k^2 denotes the received noise power of user k. Considering the time overhead for configuring the RIS, the effective communication rate of user k at the n-th block is given by

$$\overline{R}^O[n] = \left(1 - \frac{T_c[n]}{T}\right)\sum_{k=1}^{K} R_k^O[n]. \tag{3.10}$$

3.2.4.2 NOMA

For the NOMA scenario, assuming there is a NOMA cluster contains K users associated with the AP, where the users utilise the same time/frequency resources. Therefore, the superimposed signal can be denoted as

$$x[n] = \sum_{k=1}^{K} \sqrt{P_k[n]} s_k[n], \tag{3.11}$$

Therefore, after transmitting through the RIS modified integrated channel, the received signal at user k is given by

$$y_k = (h_{b,k}[n] + \mathbf{h}_{r,k}^H[n]\,\mathbf{\Theta}[n]\,\mathbf{h}_{b,r}[n])x_k[n] + \sum_{j\neq k}^{K}(h_{b,k}[n] + \mathbf{h}_{r,k}^H[n]\,\mathbf{\Theta}[n]\,\mathbf{h}_{b,r}[n])x_j[n] + n_0, \tag{3.12}$$

where the firt term $(h_{b,k}[n] + \mathbf{h}_{r,k}^H[n]\,\mathbf{\Theta}[n]\,\mathbf{h}_{b,r}[n])x_k[n]$ represents the received desired signal for user k, and $x_k[n]$ represents the transmitting signal for user k, and the second term $\sum_{j\neq k}^{K}(h_{b,k}[n] + \mathbf{h}_{r,k}^H[n]\,\mathbf{\Theta}[n]\,\mathbf{h}_{b,r}[n])x_j[n]$ represents the inter user interference.

However, not all undesired signals cause interference, since SIC allows user k to

eliminate some of the signals of other users [3]. Specifically, for each fading block, the user with a higher channel power gain first decodes the signal of users with weaker channel power gains, before decoding its own signal. Let π_k and π_j indicates the decoding order among any users k and j. If $h_k[n] \geq h_j[n]$, it can be obtained $\pi_k > \pi_j$, which indicate the signal for user j has to be decoded earlier than the signal for user k. Then, the interfering signals from users with weaker channel conditions are removed and the achievable communication rate of user k for NOMA at the n-th block is given by

$$R_k^N[n] = \log_2\left(1 + \frac{p_k[n]\, h_k[n]}{\sum_{\pi_j > \pi_k} p_j[n]\, h_k[n] + \sigma_k^2}\right). \tag{3.13}$$

Accordingly, the effective communication rate for the NOMA case at the n-th block is given by

$$\overline{R}^N[n] = \left(1 - \frac{T_c[n]}{T}\right) \sum_{k=1}^{K} R_k^N[n]. \tag{3.14}$$

3.2.5 Problem Formulation

This chapter aims to maximise the effective throughput of the entire duration, by jointly optimising the configuration policy of the RIS $\{\boldsymbol{\Theta}[n], n \in \mathcal{N}\}$, and the power allocation policy of the AP, $\{p_k[n], n \in \mathcal{N}, k \in \mathcal{K}\}$. Thus, the long-term joint optimisation problem can be formulated as

$$\max_{\{\boldsymbol{\Theta}[n], p_k[n], \tau_m[n]\}} \quad \sum_{n=1}^{N} \sum_{k=1}^{K} \overline{R}_k^X[n] T \tag{3.15a}$$

$$\text{s.t.} \quad p_k^O[n] \leq p_{\max}^O / K, \forall n, \tag{3.15b}$$

$$p_k^N[n] \leq p_{k\max}^N, p_{k\max}^N \leq p_{j\max}^N, \forall n, \tag{3.15c}$$

$$\theta_m[n] \in [0, 2\pi), \forall m, n, \tag{3.15d}$$

[3] Perfect SIC is assumed in this thesis since the research scope of this thesis is not on the SIC decoder. In practice, according to [134], the imperfect may lead to a possible BER increase depending on the modulation scheme and power allocation policy (2dB in [134]). Some robust SIC decoding approaches have also been developed for NOMA receivers [135]

$$\sum_{m=1}^{M} \tau_m [n] = M_c [n], \forall n, \tag{3.15e}$$

$$\theta_m [n] = \theta_m [n-1], \text{if } \tau_m [n] = 0, \ \forall m, n, \tag{3.15f}$$

$$\pi_k > \pi_j, \ \text{if} \ \ h_k [n] \geq h_j [n], \forall k \neq j, \tag{3.15g}$$

where $X \in \{N, O\}$ indicates NOMA and OMA schemes. Constraint (3.15b) and constraint (3.15c) indicate the maximum allowed transmit power for each user, and considering user fairness, each user has a maximum power limitation. In the OMA case, it can be obtained that $p_k^O [n] \leq p_{\max}^O / K$, where $p_{\max}^O$ represents the total transmission power. For NOMA scenarios, it can be obtained that $p_k^N [n] \leq p_{k\max}^N$, and $p_{k\max}^N \leq p_{j\max}^N$ according to the SIC principle. Constraint (3.15d) represents the phase shift constraint for each element of the RIS. Constraints (3.15e) and (3.15f) are the constraints for configuring the RIS elements, where (3.15e) indicates the time cost of configuring reflecting elements and (3.15f) represents the initial phase shift of elements for a time block. Constraints (3.15g) is NOMA decoding order constraints to ensure the decoding order for each user is correct, which are only valid when $X = N$. Due to the existence of configuration overhead, it is necessary to determine whether each reflecting unit needs to be reconfigured at every moment, which is $\tau_m [n] \in \{0, 1\}$. This binomial decision constraint makes it difficult for phase shifts optimisation with configuration overhead to be solved by conventional methods such as convex optimisation. Therefore, two ML algorithms are proposed to solve this challenging problem.

3.3 Deep learning Solution

An ETDL algorithm is proposed in this section, where a DNN is employed to map the phase shift as well as the power allocation policy with the corresponding channel state. The motivation for proposing ETDL is that the conventional DL algorithms require rich and diverse training data sets, which are onerous to obtain for communication scenarios.

On the contrary, different from the conventional offline pre-training mode, the DL agent is trained by the interaction with the environment and converges the phase shift and power allocation to an optimal solution. Another advantage of the proposed ETDL algorithm is that it has low complexity and is capable of obtaining fast convergence.

3.3.1 ETDL Algorithm

Algorithm 1 ETDL algorithm for problem (3.15)

Initialize the environment, the neural network size, and the training epoch number e
Initialize the DNN ω with random parameters
for each episode **do**
  Reset the environment and initial state
  for each step $0 \leq n \leq N$ **do**
    Input the current phase shifts $\boldsymbol{\Theta}[n-1]$ and the CSI $\{\mathbf{h}_{r,k}[n], \mathbf{h}_{b,r}[n], h_{b,k}[n]\}$ to the DNN
    DNN predict and output the phase shifts and power allocation according to (3.16)
    Calculate the datarate and feed into loss function
    Train DNN ω_n with the loss (3.18) for e epoches
    $[n] \leftarrow [n+1]$
  end for
end for

3.3.1.1 Environment based training process

There is a predicament when DL-based algorithms are applied in wireless networks, which is the acquisition of training data sets. Therefore, in order to solve the training problem of the DL, the ETDL paradigm is proposed. In the ETDL mode, the training data set used by the agent no longer needs to be prepared in advance, but allows the agent to continuously interact with the environment to obtain training data. By adopting this online training mode, the ETDL agent can be trained by the environment and has the ability to adapt to the environment expeditiously. Furthermore, in order to compare the throughput and convergence rate of RL and DL algorithms over a period of time. The concept of episodes is also introduced in the ETDL training process.

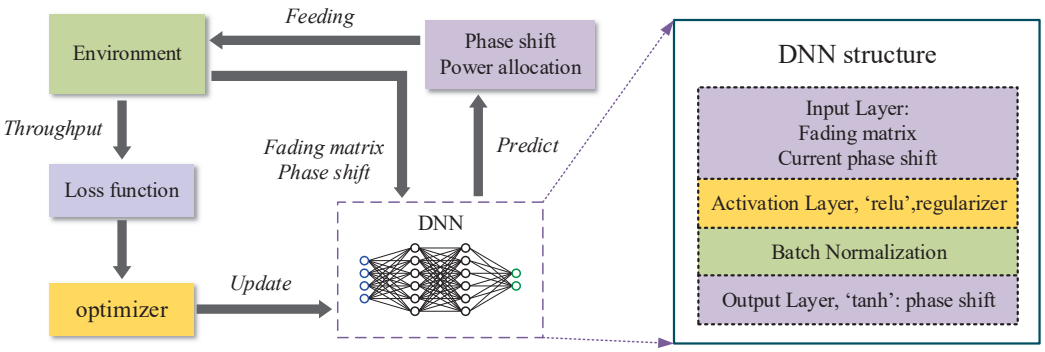


Figure 3.3: Flow diagram of the proposed ETDL algorithm

As described in **Algorithm 1**, at each channel block n, the DL agent needs to collect the previous phase-shift matrix, $\boldsymbol{\Theta}[n-1]$, and the CSI, $\{\mathbf{h}_{r,k}[n], \mathbf{h}_{b,r}[n], h_{b,k}[n]\}$, and input them into the DNN. In case that the configuration overhead is not considered, the previous phase-shift matrix does not have to be included as the input value for the DNN, and the DNN only needs to match the optimal phase shift based on the CSI. However, this scheme is not adaptable for the RIS with configuration overhead since in this design DNN has no access to learn the data rate impact caused by the configuration overhead. Then, DNN predicts the current phase-shift matrix, $\boldsymbol{\Theta}[n]$, and the power allocation, $\{p_k[n], \forall k \in K\}$ as follows

$$\{\boldsymbol{\Theta}[n], p_k[n]\} = \omega(\boldsymbol{\Theta}[n-1], \mathbf{h}_{r,k}[n], \mathbf{h}_{b,r}[n], h_{b,k}[n]|\omega_n), \tag{3.16}$$

where ω represents the function fitted by the DNN and the ω_n represents the parameters of the DNN for block n. Once the action is executed in the environment, the achieved data rate can be obtained and feed into the loss function to train the DNN. After proper training procedures, the DNN is able to select a profitable action based on the existing CSI and phase shifts.

3.3.1.2 Loss Function

In order to maximise the throughput for both NOMA and OMA schemes, the following loss function $L[n] = -\overline{R}^{X}[n]$ is invoked.

In fact, for the beamforming in OMA networks, solving the optimisation problem can be equivalently seen as solving:

$$L[n] = -\sum_{k=1}^{K}(\left|h_{b,k}[n] + \mathbf{h}_{r,k}^{H}[n]\,\mathbf{\Theta}[n]\,\mathbf{h}_{b,r}[n]\right|)^2. \tag{3.17}$$

However, this channel is not necessarily favored by NOMA, since the NOMA gain increase with the discrepancy of users' channel strength [33]. Moreover, since NOMA users suffer from intra-cluster interference, the power allocation for each user has to be jointly considered with the channel gain. Thus, the RIS phase shifts optimisation problem in the NOMA scenario can be more complicated. In addition, equation (3.17) cannot be invoked to jointly optimise phase shifts and power allocation. Therefore, since DNN has the ability to fit complex or even non-analytic function relationships, the throughput obtained from the environment is taken as the input value of the loss function.

3.3.1.3 Overfitting Problem

Compared with the conventional DL algorithms, ETDL exempts the burden of preparing a large number of training data sets. However, due to the limited data samples that can be obtained in each episode, the overfitting problem becomes the cost of the shortcut. Overfitting problem is a kind of high incidence of DNN training, which can cause serious optimisation performance degradation. In order to resolve the overfitting problem, a series of methods have been proposed, such as reducing the number of hidden layers (neurons), early stopping [136], and cross-validation. Unfortunately, these solutions are not suitable for the proposed ETDL algorithm or the communication scenario. The

cross-validation method requires additional data sets for verification and it is unlikely to be obtained in the considered scenario. There are also attempts to reduce hidden layers. However, even if a DNN with only one hidden layer is employed, the degree of overfitting is still unacceptable. Early stopping is indeed an effective solution to solve the overfitting problem, but the practical problem is how to determine the time slot that the optimal neural network parameters appear and stop the training process at the optimal moment accurately.

Therefore, for the proposed ETDL algorithm , L_2 regularisation is adopted to remittence the overfitting problem [137]. By adding a penalty term, the regularised loss function is given by

$$L[n] = -\overline{R}^X[n] + \lambda \|\omega\|_2^2, \tag{3.18}$$

where λ is the regularization weight that balances the loss minimisation and regularization, and $\|\omega\|_2^2$ is the square root of the sum of the squares of the model parameters ω.

3.3.2 DNN Structure

The proposed neural network consists of an input layer, an activation layer, a batch normalization (BN) layer, and an output layer as shown in Fig. 3.3. In the proposed RIS model, the size of the DNN input layer is determined by $\{\mathbf{\Theta}, \mathbf{h}_{r,k}[n], \mathbf{h}_{b,r}[n], h_{b,k}[n]\}$, and it is indirectly determined by the number of RIS elements and the number of served users. One or more activation layers can be used to fit the functional relationship between input and output values, and 'relu' is selected as the activation function for activation layers. The output layer is responsible for outputting all the action parameters. Since the phase shift is periodic, a 'tanh' function is employed in the output layer to ensure that the output is normalized. Even if the allocated power and phase shift have discrepant magnitudes, the normalized output can be converted according to the corresponding

demand. Finally, a BN layer is inserted before the output layer, since the BN layer needs to ensure that the value passed into the output layer is within the effective range of the 'tanh' function.

In addition to the hierarchical structure of the neural network, the size of the neural network is another issue that needs to be determined. Since the size of the input layer is constantly changing, the size of the activation layer of the fully-linked network also has to be adjusted accordingly to achieve the optimal fitting effect. The empirical formula that can be given is that the reasonable number of neurons in the activation layer is $4 * \|\boldsymbol{\Theta}, \mathbf{h}_{r,k}[n], \mathbf{h}_{b,r}[n], h_{b,k}[n]\|$, where $\|\|$ represents the number of elements in the collection. In practical scenarios, the devices may need to store multiple DNN models in order to deal with different situations. Please note that processing these models are also limited by the computational and storage performance of the processor, which is not considered in this thesis.

Remark 1. *The size of the input layer, hidden layers, and the output layer of the DNN are all proportional to the number of reflective elements. Therefore, the reflective element tile not only reduces the configuration overhead, but also reduce the scale and complexity of the DNN.*

3.3.3 Complexity Analysis

The main complexity of DL algorithms is from predicting and training, which can be discussed separately. Assume that the employed DNN has I layers, and each layer has ω_i neural nodes. Furthermore, since different types of layers are employed, the floating point calculation complexity consumed by the corresponding neurons is also different. Therefore, the sum nodes of BN layers, 'relu' layers, and the 'tanh' layer are assumed to be ω_b, ω_r, and ω_t, respectively.

- Predicting complexity: As aforementioned, for any block n, the DNN needs to predict the phase shifts and power allocation policy, and the complexity required

to complete a single prediction is given by $\mathcal{O}(\sum_{i=0}^{I} \omega_i \cdot \omega_{i+1})$ as indicated in [138].

- Training complexity: For the back propagation training, different types of nodes require different computations. A single 'BN' node, a 'relu' node and a 'tanh' node require 5, 1, and 6 times floating point operations, respectively [139]. Thus, a single back propagation training step for the proposed DNN structure will contribute the complexity of $\mathcal{O}(5\omega_b + \omega_r + 6 \cdot \omega_t + \sum_{i=0}^{I} \omega_i \cdot \omega_{i+1})$.

Therefore, for the transmission period $\mathcal{T}$, the overall complexity of the DL algorithm can be calculated as

$$\mathcal{O}(N \cdot e \cdot (5\omega_b + \omega_r + 6 \cdot \omega_t + \sum_{i=0}^{I} 2\omega_i \cdot \omega_{i+1})), \tag{3.19}$$

where e represent the number of epoch for one training step.

3.4 Reinforcement Learning Solution

An RL algorithm, namely EA-DDPG is proposed in this section to solve the exactly same problem as the proposed ETDL algorithm, which is the joint optimisation for phase shifts of the elements and the power allocation policy for each user. Different from DL algorithms, RL algorithms can formulate the phase shift control as a Markov decision process [140] and predict the rewards in the future, which suggests that RL is more capable of maximizing the long-term throughput for the wireless network. The phase shifts, power allocation, and corresponding data rate changes caused by each decision of the agent can be abstracted into a Markov transition, composed of four components, namely state ($\mathbf{s}[n] \in \mathbf{S}$), action ($\mathbf{a}[n] \in \mathbf{A}$), reward ($r[n]$), and next state ($\mathbf{s}[n+1] \in \mathbf{S}$). RL is committed to finding the optimal action $\mathbf{a}[n]$ for each state $\mathbf{s}[n]$, and the RL agent introduces action value (Q-value) to quantitatively evaluate the long-term value of each action. The long-term value of action $\mathbf{a}[n]$ is defined as the following the Bellman

equation

$$Q(\mathbf{s}[n],\mathbf{a}[n]) = r[n](\mathbf{s}[n],\mathbf{a}[n]) + \beta \max Q(\mathbf{s}[n+1],\mathbf{a}[n+1]), \tag{3.20}$$

where β represents the discount factor in a range of [0,1], and the optimal action is defined as

$$\mathbf{a}^*[n] = \underset{\mathbf{a}[n]\in\mathbf{A}}{\arg\max} \quad Q(\mathbf{s}[n],\mathbf{a}[n]). \tag{3.21}$$

3.4.1 Algorithm flow

In order to achieve a firm phase shifts control to obtain the configuration-overhead-free experiences, the EA-DDPG algorithm is proposed, as the algorithm flow presented in Fig. 3.4 and **Algorithm 2**. Contrary to the original deep deterministic policy gradient (DDPG) algorithm [141], EA-DDPG adopts attenuated action noise to obtain deterministic actions to avoid additional configuration overhead. Two pairs of neural networks are employed in the RL agent, namely the actor network μ, the critic network Q, and their corresponding target networks μ' and Q'. The RL agent can obtain the current phase shift of the elements and the CSI can be obtained by channel estimation to form the state $\mathbf{s}[n]$. Then the actor network is able to give the action as

$$\mathbf{a}[n] = \mu(\mathbf{s}[n]|\omega_n^\mu), \tag{3.22}$$

After action $\mathbf{a}[n]$ is taken in the communication environment, the phase shifts are changed to $\mathbf{s}[n+1]$ and the reward $r[n]$ can also be calculated according to the data rate. After that, the transition $\{\mathbf{s}[n],\mathbf{a}[n],r[n],\mathbf{s}[n+1]\}$ has to be recorded into the memory buffer. According to the memory replay approach, for each training step the agent can be trained by a number of random samples from the memory bank. The actor

network is trained by the policy gradient calculated by the critic network

$$\nabla_{\omega^\mu} J = \frac{1}{e} \sum_{e} \nabla_{\mathbf{a}} Q(\mathbf{s}_e[n], \mathbf{a}_e[n]|\omega^Q) \nabla^{\mu}_{\omega} \mu(\mathbf{s}_e[n]|\omega^\mu). \tag{3.23}$$

where e represents the number of samples. With the actor-critic structure, the function of the critic network is to evaluate the Q-value and the critic network is updated by minimising the loss function

$$L[n] = \frac{1}{e} \sum_{e} (y_e[n] - Q(\mathbf{s}_e[n], \mathbf{a}_e[n]|\omega^Q_n))^2, \tag{3.24}$$

where

$$y_e[n] = \ r_e[n](\mathbf{s}_e[n], \mathbf{a}_e[n]) \ + \beta Q'(\mathbf{s}_e[n+1], \mu'(\mathbf{s}_e[n+1]|\omega^{\mu'}_n)|\omega^{Q'}_n). \tag{3.25}$$

Algorithm 2 EA-DDPG algorithm for problem (3.15)

Initialize the environment and the neural network size
Initialize the actor network ω^{μ}_u, critic network ω^{Q}_u, target actor network $\omega^{\mu'}_u$, target critic network $\omega^{Q'}_u$ with random parameters
for each episode **do**
  Reset the environment and initial state
  Action noise attenuation
  for each step in $t_0 \leq t \leq t_{\max}$ **do**
    Observe $\mathbf{s}[n]$
    Choose $\mathbf{a}[n]$ according to (3.22)
    Action a is execute in the environment
    $r[n]$ is calculated based on the environment and the next state $s[n+1]$ is predicted
    Record $e\{\mathbf{s}[n], \mathbf{a}[n], r[n], \mathbf{s}[n+1]\}$ in memory buffer
    Random sample a batch of transection e from memory buffer
    Calculate target according to (3.25)
    Train critic network $Q(\mathbf{s}[n], \omega^Q_n)$ with a gradient descent step (3.24)
    Train actor network $\mu(\mathbf{s}[n], \omega^{\mu}_n)$ with (3.23)
    Update the target networks $\omega^{\mu'}_n \leftarrow (1-\tau)\omega^{\mu'}_n + \tau\omega^{\mu}_n$, $\omega^{Q'}_n \leftarrow (1-\tau)\omega^{Q'}_n + \tau\omega^{Q}_n$
    $\mathbf{s}[n] \leftarrow \mathbf{s}[n+1]$
  end for
  Each agent save the network models ω^{μ}_n, ω^{Q}_n, $\omega^{\mu'}_n$, $\omega^{Q'}_n$
end for

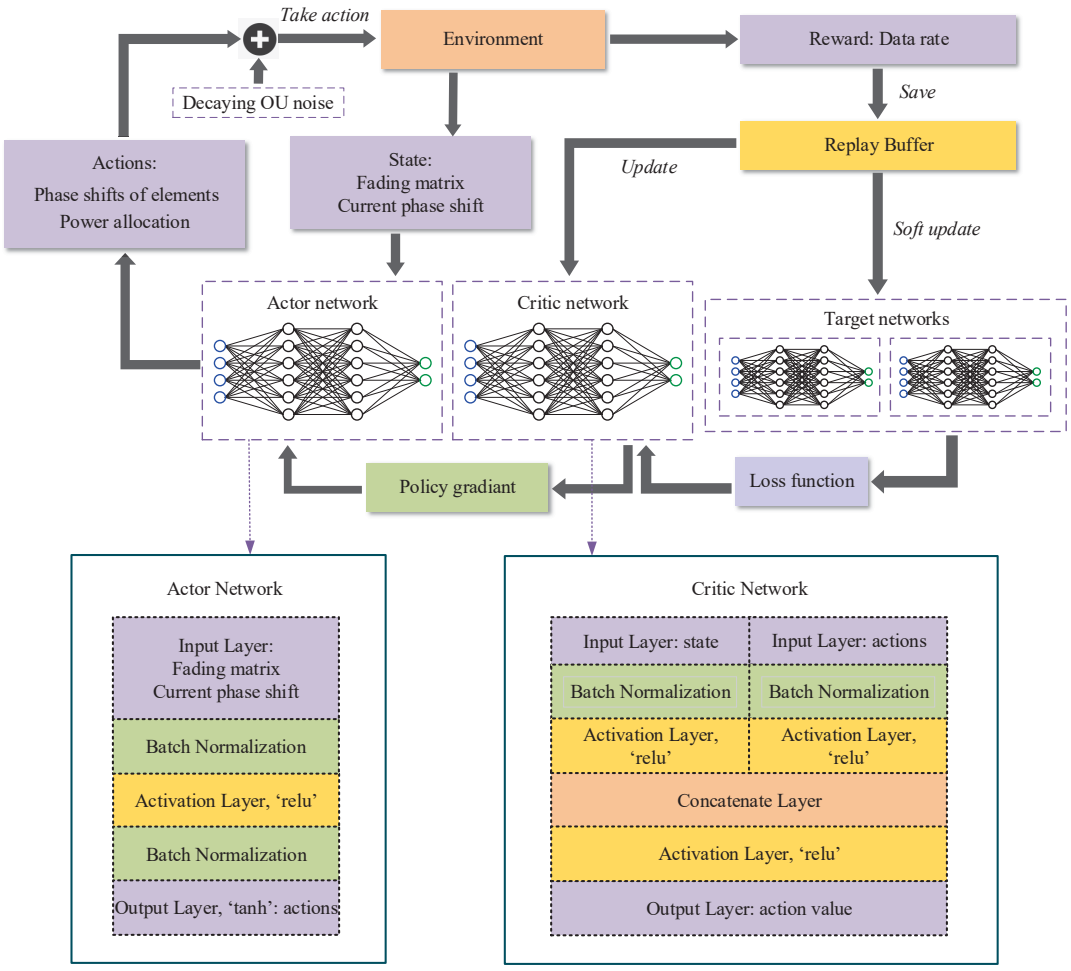


Figure 3.4: Flow diagram of the proposed EA-DDPG algorithm

3.4.2 State, Action and Reward Function

In order to employ DRL for the optimisation problem (3.15), a fundamental step is to design the state space, action space, and reward function. Meanwhile, in order to ensure both the efficiency of the exploration and convergence, the EA-DDPG algorithm is armed with an action policy having decaying action noise.

3.4.2.1 State Space

Similar to the input obtained by the DNN in the DL algorithm, the fading matrix is also provided to the EA-DDPG algorithm to adjust the phase shifts of elements. Not only that, in order for the RL to determine whether the elements need to be adjusted, the phase shifts of the reflective elements in block $n-1$ are also provided to the RL agent as a reference. Therefore, the state for each fading block n is formed by $\{\mathbf{\Theta}[n-1], \mathbf{h}_{r,k}[n], \mathbf{h}_{b,r}[n], h_{b,k}[n], \forall k\}$. Since the data type input to the neural network needs to be unified [4], the state array can be expressed as

$$\begin{aligned}\mathbf{s}[n] =&\{\text{real}\{\mathbf{\Theta}[n-1], \mathbf{h}_{r,k}[n], \mathbf{h}_{b,r}[n], h_{b,k}[n]\},\\ &\text{imag}\{\mathbf{\Theta}[n-1], \mathbf{h}_{r,k}[n], \mathbf{h}_{b,r}[n], h_{b,k}[n]\}\}, \forall k \in K. \end{aligned} \tag{3.26}$$

3.4.2.2 Action Space

The action space is composed of two optimisation variables, phase shifts of the elements and power allocation for users. It is worth mentioning that the output action can also be selected as the changed phase $\triangle\mathbf{\Theta}[n]$, but in order to control the difference as much as possible to compare the performance of RL and DL algorithms, the RL agent also output $\mathbf{\Theta}[n]$.

- Phase shifts: The actor network outputs the optimal $\mathbf{\Theta}[n]$ for block n in a radian. Then the output phase-shift matrix $\mathbf{\Theta}[n]$ will be judged whether there is a difference with the current phase $\mathbf{\Theta}[n-1]$. It is worth noting that since that the action noise invoked by the DRL algorithm may cause tiny phase shifts but the impact of these actions is negligible. Thus, in order to avoid inconsequential configuration, insignificant phase shifts raised by the agent will be ignored, and substantial phase changes will be carried out in the environment [5].

[4] At this stage, deep neural network implementations do not support complex input. This problem may be resolved with the improve of AI programming language.

[5] In the simulation, the phase shift less than 2 degrees is ignored.

- Power allocation policy: The allocated power $\{p_k[n], \forall n\}$ for each user k will get an independent output value from the actor network.

Thus, the overall action space can be noted as

$$\mathbf{a}[n] = \{\mathbf{\Theta}[n], p_k[n], \forall k \in K\}. \tag{3.27}$$

3.4.2.3 Exploration and Action Policy

The RL algorithm family generally requires action noise as an engine for exploring diversity, and the action policy in original DDPG algorithm can be denote as

$$\mathbf{a}[n] = \mu(\mathbf{s}[n]|\omega_n^\mu) + \mathcal{N}_{\mathrm{OU}}(0, 0.2), \tag{3.28}$$

where $\mu(s[n]|\omega_n^\mu)$ is the decision made by the actor network, ω_n^μ denotes the parameters of the actor network, $\mathcal{N}_{\mathrm{OU}}(0, 0.2)$ represents the zero mean Ornstein Uhlenbeck (OU) noise that following $\mathcal{N}_{\mathrm{OU}} \sim OU(0, \xi)$, and ξ is the volatility of the OU noise [142]. OU noise is selected as the action noise since the exploration efficiency of the OU noise is higher than the conventional Gaussian noise, which is proved by [143].

However, the accession of the action noise makes the agent unable to maintain the phase shifts. For example, assuming $\theta_m[n]$ was adopted at time block n, once the agent decide to maintain the $\theta_m[n]$ at time block $n+1$, it is supposed to have $\mathbf{a}_m[n+1] = \mu(\mathbf{s}[n+1]|\omega_{[n+1]}^\mu) = \theta_m[n]$, and in this manner the element m do not need to be reconfigured. Unfortunately, the action noise destroys this deterministic relationship and it will cause $\mathbf{a}_m[n+1] = \mu(\mathbf{s}[n+1]|\omega_{[n+1]}^\mu) + \mathcal{N}_{\mathrm{OU}}(0, 0.2)$, which suggested that the reflective element m needs to be reconfigured, and configuration overhead is incurred.

Therefore, in order to alleviate the contradiction between configuration overhead and exploration, a decaying OU noise is selected as the action noise. The strong initial noise is able to offer a variety of random actions, which can explore the environment efficiently

at the early stage of the training. With the decaying of the noise power, the agent finally obtains firm actions and stable convergence at the late training stage. Cooperating with the aforementioned phase shift threshold, in the late stage of training, the agent can maintain a fixed phase shift if necessary. Thus, the action with decaying OU noise can be expressed as

$$\mathbf{a}[n] = \mu(\mathbf{s}[n]|\omega_n^{\mu}) + \mathcal{N}_{\mathrm{OU}}(0, \xi), \xi_n = \xi_0 \to 0, \tag{3.29}$$

where $\xi = \xi_0$ at the beginning of the training stage and decay with the number of training episodes.

Remark 2. *Action noise is generally regarded as a boost for the RL exploration. However, for scenarios that need to maintain the state, the action noise is likely to enforce the agent to change the state, which makes the agent lack the experience of maintaining the state. Therefore, for the proposed scenario with configuration overhead, the decaying action noise is employed and the agent is expected to obtain transition of keeping the state in the late stage of training.*

3.4.2.4 Reward Function

Since the optimisation goal is to maximise the throughput, similar to the loss function employed in DL, the reward function for RL is set to the effective sum data rate of each block n as

$$r[n] = \sum_{k=1}^{K} \overline{R}_k^X[n]T. \tag{3.30}$$

Correspondingly, since no additional rewards are provided for the increase in total throughput, the discount factor have to be set to 1 to ensure the agent has enough foresight to maximise long-term benefits.

3.4.3 Neural Network Structure

Since the actor and critic networks have different inputs, they also have different structures, and the target network has exactly the same shape as the main network. The actor network has a similar structure with the DNN employed in the proposed DL algorithm. The actor network consists of the input layer, BN layer, activation layer(s), another BN layer, and output layer in turn. As mentioned above, the size of the activation layer needs to be coordinated with the number of RIS elements. An insufficient number of neurons in the activation layer will result in a decrease in the fitting effect. On the contrary, according to empirical conclusions, extra activation layers will not only increase the computational complexity, but also lead to a worse fitting effect. On the other hand, the critic network has a larger scale, trainable parameters, and complexity, since the critic network has to take both state $s[n]$ and action $a[n]$ as input to evaluate the action value. Therefore, two BN layers connected to the input layer are still necessary, since the value of the state and the action may have different sizes.

3.4.4 Complexity Analysis

Since the RL algorithm employs 4 neural networks, it has a higher complexity than the DL algorithm. Inheriting the semiotics of DL complexity analysis, ω^{μ} represents the action networks and ω^{Q} denotes the critic networks since the target networks have the same size as the main networks.

- Predicting complexity: For each single action prediction, the complexity only caused by the actor, which can be calculated as $\mathcal{O}(\sum_{i=0}^{I} \omega_i^{\mu} \cdot \omega_{i+1}^{\mu})$ following the same approach as the DL.

- Training complexity: Since both the actor network and the critic network need to be trained, the most intuitive complexity is caused by the back propagation, which can be given by $\mathcal{O}(5\omega_b^{Q} + \omega_r^{Q} + 6 \cdot \omega_t^{Q} + \sum_{i=0}^{I} \omega_i^{Q} \cdot \omega_{i+1}^{Q})$ and $\mathcal{O}(5\omega_b^{\mu} + \omega_r^{\mu} +$

$6 \cdot \omega_t^{\mu} + \sum_{i=0}^{I} \omega_i^{\mu} \cdot \omega_{i+1}^{\mu})$. It is also worth to notice that the training process needs the prediction results from target networks, which can be calculated as $\mathcal{O}(\sum_{i=0}^{I} \omega_i \cdot \omega_{i+1}^{Q}) + \mathcal{O}(\sum_{i=0}^{I} \omega_i^{\mu} \cdot \omega_{i+1}^{\mu})$.

Therefore, the total complexity of the RL algorithm is given by

$$\begin{aligned} \mathcal{O}(N \cdot e \cdot ((5\omega_b^Q + \omega_r^Q + 6 \cdot \omega_t^Q + \sum_{i=0}^{I} 2\omega_i^Q \cdot \omega_{i+1}^Q) \\ +(5\omega_b^{\mu} + \omega_r^{\mu} + 6 \cdot \omega_t^{\mu} + \sum_{i=0}^{I} 2\omega_i^{\mu} \cdot \omega_{i+1}^{\mu}))) \end{aligned} \tag{3.31}$$

where e represents the batch size of the RL algorithm. Comparing equation (3.19) and equation (3.31), when the DNN employed in the DL approach has the same size as the action network in the RL approach, the RL algorithm has a significantly higher complexity.

3.5 Numerical Results and Analysis

This section presents the numerical results obtained by the developed RL and DL algorithms to optimise the RIS's phase shifts with configuration overhead, aiming to compare and analyze their advantages, disadvantages, and strategic features.

3.5.1 Parameters Settings

To ensure a fair comparison of the convergence behavior, the 'Adam' optimiser is employed for both algorithms to train parameters for neural networks since it has high-efficiency gradient descent strategy [144]. However, two algorithms are not likely to have the same optimal parameter setting, thus, in order to ensure the performance of the algorithms, different empirical optimal neural network sizes and learning rates are set in the RL and DL algorithms. For the DL algorithm, the learning rate is set to be exponentially

Table 3-A: Simulation Parameters

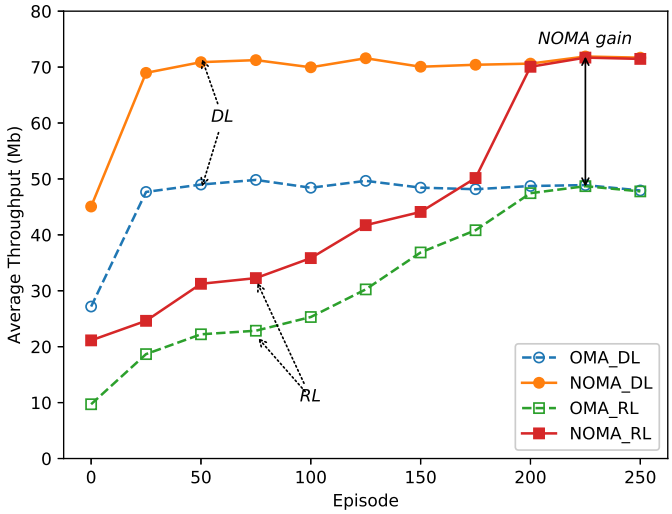
Parameter	Description	Value	Parameter	Description	Value
f_c	carrier frequency	2GHz	K	number of users	4
T_{n}	Fading block length	1ms	T_0	Time overhead per element	0.03ms
B_k^u	bandwidth	1 MHz	$P_{\max}$	maximum transmitting power	26 dBm
$d_{a,r}$	AP to RIS distance	300 m	σ	noise power density	-55 dBm/MHz
ρ_0	path loss at the reference distance	-30	d_0	reference distance	1 m
α_{AU}	path loss exponent, AP-users	3.5	α_{AR}	path loss exponent, AP-RIS	2.2
α_{RU}	path loss exponent, RIS-users	2.2	K_{AR}, K_{RU}	Rician factors	3dB
α	learning rate	3×10^{-4}	γ	discount factor	1
e	batch size	32 samples	τ	target update rate	0.002
u	tile size	4 elements	λ	regularization weight	0.001

decayed starting from 0.001, and for the RL algorithm, both the actor and the critic learning rate are assigned as 0.0003. On the other hand, neural networks follow the aforementioned stature and each neural network contains one activation layer, where the scale of the activation layer is in a range of 64-512, depending on the element number of the RIS and the user number. Although training neural networks of this scale is completely affordable for a communication system, the parameter number of neural networks required by the RL algorithm is larger, which can lead to further complexity. The value of T_0 is determined by the experiment in [129], where a reasonable range is in $[0.06, 0.015]$(ms), and two representative values 0.1 and 0.03 is used in this simulation. The rest of the default parameters in this simulation are listed in Table 3-A. [6]

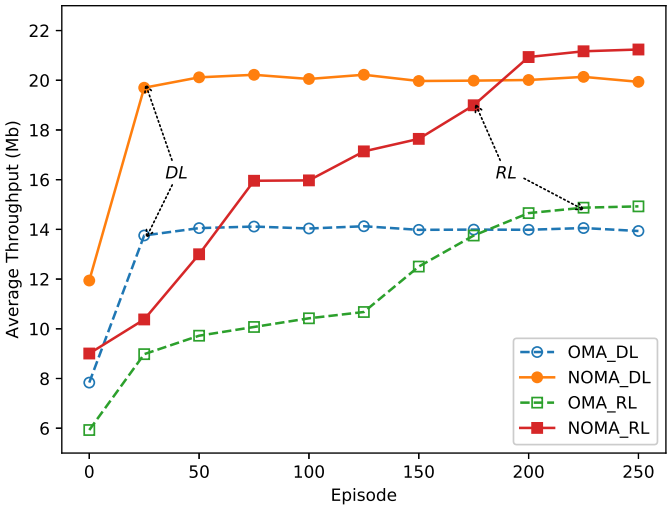
3.5.2 Convergence and Optimality

Fig. 3.5 demonstrates the convergence of the proposed RL and DL algorithms. Following the aforementioned system model, the AP-RIS link and RIS-users links are assumed to have a Ricain fading channel and AP-users links follow the Rayleigh channel. At the

[6] In this thesis, the parameters of the simulation are generally selected according to the 3GPP standard, common values found in other references, or our empirical optimum. Specifically, the parameters of communication systems, such as the path loss, maximum transmit power, etc. are derived from the standards [145] mentioned in the channel model; other parameters, such as center frequency, bandwidth, number of users, Rice factor, etc., are selected according to the commonly used values in reference materials; finally, machine learning The parameters of the agent, such as learning rate, bench size, exploration rate, etc., are selected based on the empirical optimum figured out in other works (e.g. [146]) or the simulation.



(a) RL/DL optimised throughput under Rician channels for AP-RIS and RIS-user links.



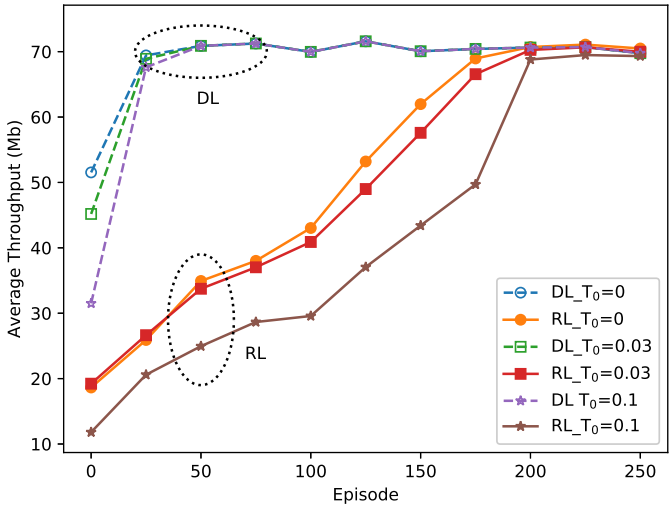
(b) RL/DL optimised throughput under Rayleigh channels for AP-RIS and RIS-user links.

Figure 3.5: RL/DL performance under different channel conditions.

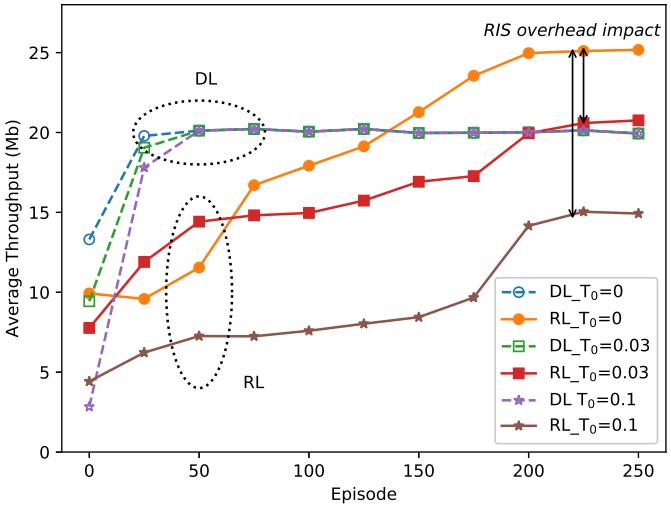
early episodes of the training, the DL approach earns a larger throughput compared with the RL algorithm, since the large action noise in early training compels the RL agent to configure elements to random phase shifts. Two key phenomena can be captured by observing the curves. Firstly, although both algorithms can converge with proper parameter settings, DL has a notable convergence advantage, i.e., DL can converge within 30

episodes while RL requires a considerable long period to converge since the RL algorithm requires a repetitive exploration process and large action noise. Meanwhile, in the later episodes, both algorithms achieve comparable performance. Moreover, it can be seen that the average NOMA gain reached about 42% compared to the OMA scheme, which illustrates that the NOMA scheme has compatibility with the RIS-assisted networks. In contrast, since there is a possibility that the LoS paths of the AP-RIS and RIS-user links are blocked, and then the RIS may lose the Rician channel. Fig. 3.5(b) depicts the obtained performance by the DL and RL algorithms for Rayleigh channel models. As illustrated, the DL approach still maintains a conspicuous convergence advantage for both NOMA and OMA schemes.

The optimality of the DL algorithm mainly depends on the fitting accuracy of the DNN and the distribution of the training data. On the other hand, the optimality of DRL is also plagued by local optimisation and DNNs' fitting performance [147]. The fitting performance of DNN is closely related to the size and structure of the DNNs. Therefore, for the proposed DL and RL algorithms, their empirically optimal neural network layout is adopted in the simulation. Assuming that all DNNs have perfect fitting functions and unlimited training time, DRL theoretically has the possibility of obtaining the global optimal solution [148]. However, in a finite training process, the optimality of DRL is mainly limited by the action policy and local optimisation. Therefore, the optimality of the proposed DL and RL algorithms is evaluated by comparing their achieved sum rate under different channel conditions. In Fig. 3.5(b), the achieved long-term throughput of DL is slightly inferior to the RL approach. The reason behind this can be explained as follows. The phase of the signal over each channel block becomes poignantly random due to the loss of the LoS path. Therefore, the RL algorithm with exploration examples is likely to recognize more different states, and then to determine whether worth adjusting the phase shifts of the elements and what the optimal phase shift to be configured. Therefore, a moderate optimality advantage of the RL algorithm can be detected in the Rician channel.



(a) Performances with different configuration overhead under Rician channels for AP-RIS and RIS-user links (NOMA).



(b) Performances with different configuration overhead under Rayleigh channels for AP-RIS and RIS-user links (NOMA).

Figure 3.6: RL/DL performance under different channel conditions.

3.5.3 Impact of the Overhead and Elements Number

Fig. 3.6 presents the achieved throughput performance of DL and RL with different time consumption T_0 for configuring one element in NOMA scenario. In particular, T_0 is set to $0, 0.03T$, and $0.1T$. For the case of $T_0 = 0$, it means that the time overhead

for configuration is not considered. In the case of Rician channels (i.e., Fig. 3.6(a)), the increase of the time overhead only causes a slight drop in the achieved throughput. Under the Rician channel hypothesis, the optimal RIS phase shifts are which conforms to the maximization of the main propagation path, thus, the change of overhead only has an impact on the performance during the convergence process[7]. However, the DL strategy is not preferable for the Rayleigh channel (or Rician channel with a small Rician factor) since the randomness of the channel is grim, so that the strategy of sticking to a fixed phase shifts for the RIS can no longer well-match the channel blocks. Therefore, the DL algorithm achieves deficient performance in the Rayleigh channel. Another reality that can be revealed is that the configuration overhead in a channel with larger randomness can lead to a graver impact on the throughput. In the inferior channel, the agent can only choose between two problems, suffer from the loss of transmission time caused by the configuration overhead or give up matching the channel. Finally, when the Rayleigh channel is invoked, the RL agent may choose to unwisely adjust the elements, even if it bears the throughput degeneration caused by the configuration overhead, and the reason for this phenomenon is revealed in **Remark 2**. Even if the RL agent takes firm action, the inherent disadvantage of exploration makes it inferior to DL in terms of averting configuration overhead. In fact, the performance achieved by the RL agent for $T_0 = 0.1$ is a local optimum, since logically maintaining certain phase shifts is likely to be the optimal solution when the configuration overhead is significant. Undeniably, assuming the training process is infinite, which means that all actions and states are traversed, the RL agent has the ability to obtain solutions that are equivalent to the DL agent. However, in practice, system cannot afford unlimited training time for intelligent agents, and thus the advantage of the DL algorithm in the Rayleigh channel exists.

Fig. 3.7 shows the achieved throughput versus the maximum allowed transmit power for different numbers of RIS elements, where a Rician channel and the NOMA scheme are invoked. It can be observed that for both DL and RL algorithms, the obtained

[7]For ease of the simulation, any phase change less than 2 degrees is ignored, which suggested elements will only be adjusted when the shift range is greater than 2 degrees.

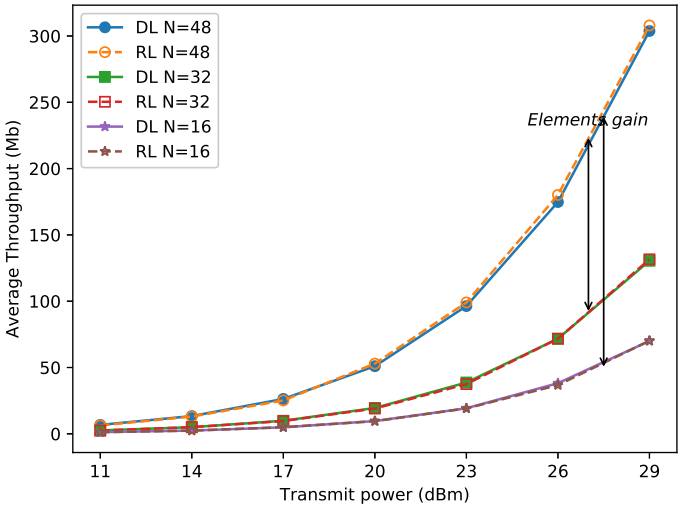


Figure 3.7: RL/DL performance over transmitting power with different elements number in Rician channel (NOMA).

throughput increases with the increase of N and $p_{\max}$ due to a higher reflecting array gain and a higher transmitting signal strength. From an algorithm perspective, the results illustrate that when the Rician channel is obtained, the RL and DL methods can achieve equal throughput without being affected by the number of the RIS elements.

3.5.4 Configuration Strategies

In order to further demonstrate the configuration of RIS elements and explore the strategies heterogeneity between the agents, the number of the precision configured elements number per episode is displayed in Fig. 3.8. It can be observed that even though the two algorithms have achieved similar throughput, they adopted different strategies. Regardless of the value of T_0, DL has a consistent strategy to converge to the optimal phase shifts and maintain them. The configuration tendency of the elements is completely consistent with the tendency of the throughput displayed in Fig. 3.6(a). During the early training process (before convergence), as depicted in Fig. 3.8, the DL agent continuously changes phase shifts over each fading block, and in this period it suffers from significant configu-

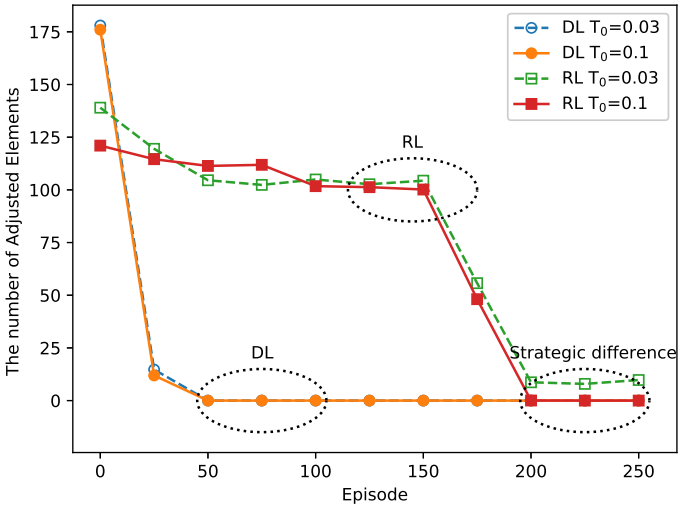


Figure 3.8: Configuration stratagem of DL/RL algorithm.

ration overhead. After the DL algorithm converges, the phase shift would converge and stick to the optimum, and the agent does not need to suffer from the configuration overhead. Since the optimal phase shifts are obtained by the agent trained by all fading block samples, in the Rician channel with less randomness, the optimal phase obtained by the DL algorithm has an excellent performance. It is interesting and worth noting that the RL agent has shown greater flexibility and it adopts inconsistent strategies in the case of different values of the configuration overhead. In the case of $T_0 = 0.03$, which could be a tolerable overhead, the well-trained RL agent chooses to bear the overhead and adjust a part of elements, though it is not frequently. However, when the configuration overheads are eloquent, for example $T_0 = 0.1$, the RL agent presents the same strategy as the DL agent. Therefore, the RL approach has preponderance in terms of strategic diversity. The RL algorithm has this feature since it formulates the states during a transmission period (episode) into the Markov decision process model, and can predict the change of the state by each action. However, the proposed DL algorithm does not have the capability of a similar long-term planning, prediction, and memory functions. However, it is worth considering that 'flexible strategy' does not necessarily result in better overall performance as demonstrated in Fig. 3.6.

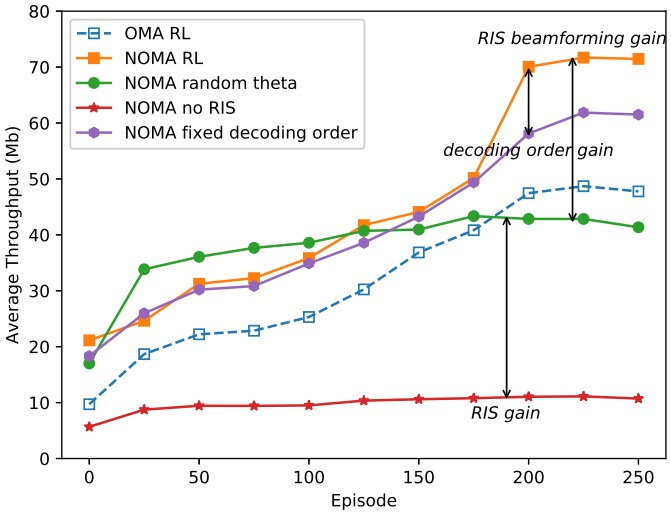


Figure 3.9: Throughput gain of the RIS, NOMA, decoding order, and phase shift optimisation.

3.5.5 Source of the Gains

Fig. 3.9 plots the achieved throughput of the cases with incomplete procedures versus episodes to identify the gain of each technological process. First of all, with the RIS and the corresponding Rician channel, the case with RIS significantly outperforms the case without RIS. On the other hand, the phase shifts optimised by machine learning algorithms can achieve a gain of 95.2%, compared to the case of random configured RIS phase shifts. It is worth to mention that since the configuration overhead is considered, the achieved gain not only comes from the optimised phase shifts, but also from the intelligent decision which diminishes the configuration overhead. Although employing RIS and ML lead to an increase in the computational complexity and hardware cost of RIS, the throughput gain is substantial. Another process worth noting is the determination of the decoding order. Once the decoding order is not in line with the CSI, the NOMA gain will be severely weakened. The curve achieved by the case with a fixed decoding order is plotted as a benchmark. About 40% of the NOMA gain is lost, in the case of incorrect decoding order.

Table 3-B: Performance Summary

Algorithm	Throughput(Rician/Rayleigh)	Convergence	Complexity	Strategic diversity	Overhead Resistance
DL	Equivalence/Inferior	Ascendant	Ascendant	Inferior	Ascendant
RL	Equivalence/Ascendant	Inferior	Inferior	Ascendant	Inferior

3.5.6 Performance Summary

The aforementioned analysis and simulation compared the pros and cons of the DL and RL methods from multiple aspects. In summary, the DL algorithm can avoid the detriment caused by the configuration overhead by virtue of the optimal phase shifts obtained by training. However, this stubborn strategy can result in performance degradation in the Rayleigh channel. On the contrary, the RL algorithm has a flexible strategy, which can be customised according to the extend of the configuration overhead, but the price is the significantly higher complexity and the lengthy training process. Therefore, once the RIS has the opportunity to obtain the Rician channel during the service period, the DL algorithm can be a preferable choice.

3.6 Summary

In this chapter, a RIS-assisted NOMA network model was studied, in which the time overhead of configuring the reflective elements of the RIS is taken into account. To solve the tradeoff problem caused by the configuration overhead, an ETDL algorithm (DL) and an EA-DDPG algorithm (RL) were proposed for the formulated throughput maximization problem and their performances were comprehensively investigated and compared in the simulation. The simulation results first revealed that the NOMA scheme achieved considerable gains compared to the OMA scheme in a variety of cases. From the perspective of algorithm performance, DL and RL algorithms achieved approximately equivalent overall performances but with their own decision-making characteristics. The DL algorithm can fast converge to a proper phase shift and remain fixed, thus, its performance was less affected by the RIS configuration overhead. On the contrary, RL

is able to intelligently choose countermeasures based on the severity of the configuration overhead. Thanks to the flexible strategy, the RL algorithm had the opportunity to adapt to the Rayleigh channel and achieve higher gain. In terms of complexity, RL has significantly higher complexity and storage space, since extra DNNs are employed. As a summary, the RL algorithm is more adaptable for complex and fickle communication scenarios. Note that only single-antenna transmission is considered in this chapter, the investigation of multiple-antenna transmission, i.e., the overhead-dependent joint beamforming optimisation, is an interesting but challenging research topic in the future.

Chapter 4

Flexible Deployment of Reconfigurable Intelligent Surfaces

This chapter describes a novel framework of mobile RISs-enhanced indoor wireless networks, where a RIS mounted on a robot is invoked to enable mobility of the RIS and enhance the service quality for mobile users. It is worth noting that RIS is an attached component on the robot, and these carrier robots can have other functions. It is also possible to add RIS panels to existing service robots, such as sweeping robots, robot guider in shopping malls, etc. This implementation does not cause considerable cost on resources, which only need to instal a light panel on the robot. To optimise the sum rate of all users, an FL enhanced DDPG algorithm is proposed to optimise the deployment and phase shifts of the mobile RIS as well as the power allocation policy.

4.1 Flexible Deployments of RISs

One of the main factors that RISs can provide noticeable gain is that they can provide further possible line-of-sight (LoS) paths for users who do not originally have an LoS path [24]. However, in most existing research contributions, RISs are assumed to be fixed on a wall or other bearing, and the rigid deployment may prevent RIS from obtaining LoS paths and optimal channel enhancement, especially in an environment with obstructions. In an effort to complement this defect, a mobile RIS scheme in which RISs are mounted on intelligent robots is proposed to achieve flexible deployment.

To maximise the data rate gain of empowering mobility to RISs, how to plan proper dynamic deployments for mobile RISs is a problem worth exploring. Since users are considered as moving as well, the optimisation problem is highly dynamic, and the joint optimisation problem of movements and phase shifts of RISs is an emerging problem worth exploring. In addition, since obstacles that hinder the movement of RISs and shield LoS paths are likely to have irregular and non-analytic shapes, this also raises challenges for conventional optimisation approaches. In contrast to convex optimisation, DRL is considered to be a more competent methodology for dynamic optimisation problems since DRL is able to recognize the current state of the environment [18, 19]. Meanwhile, since multiple mobile RISs can be deployed in different cells, FL is employed to strengthen their training efficiency and effectiveness for the proposed multi-cell multi-agent scenario [21]. FL arouses the interest of researchers as a distributed learning framework since it can effectively utilize computational resources [22] with a protection of user privacy [23]. Especially for the DRL algorithm, FL can improve training efficiency and learning effect, since agents can explore the environments simultaneously and their knowledge can be transferred to each other through a global neural networks model. Therefore, a DRL algorithm with a framework of FL is proposed, namely the FL-DDPG algorithm to jointly optimise the passive beamforming, dynamic deployment of RISs, and the power allocation for NOMA users.

4.2 System Model

This section first describe assumptions and system model of the proposed mobile RISs enhanced wireless networks in subsection 4.2.1. The layout modeling method of the indoor environment and the propagation model are illustrated in subsection 4.2.2 and subsection 4.2.3, respectively. The signal models of both OMA and NOMA scheme are illustrated in subsection 4.2.4. At last, the optimisation problem is formulated in subsection 4.2.5.

4.2.1 System Description and Assumption

An indoor downlink multiple-input and single-output (MISO) scenario is considered, where RISs are employed and each RIS is carried by a robot to enhance indoor propagation for a wireless access point (AP) to serve users in the room as illustrated in Fig. 4.1. The served building is assumed having multiple floors or rooms, denoting each of them as a cell, and each cell is configured with an AP. In order to relieve the interference between each floor, a spectrum strategy that similar to what is adapted in the cellular networks is employed to diminish adjacent cell interference. The frequency band of the system is divided into at least two, and then adjacent floors can use different frequency bands. For example, if the frequency band is divided into two, odd-numbered floors can occupy the same frequency band, and even-numbered floors have to apply the other frequency band. Since the floors using the same frequency band are guaranteed to have a sufficient spatial distance and the signal is obstructed by ceilings, so that the interference between APs can be reduced to a negligible level.

Each floor of the building is assumed to have a similar architectural structure and layout, which is common in office buildings or flats. For each cell, the AP is equipped with M antenna, while each user only has a single antenna. The RIS is armed with N reflecting elements, which can provide concatenated LoS propagation for the transmitter and receivers by reflecting and reconfiguring the signals. Multiple users are considered

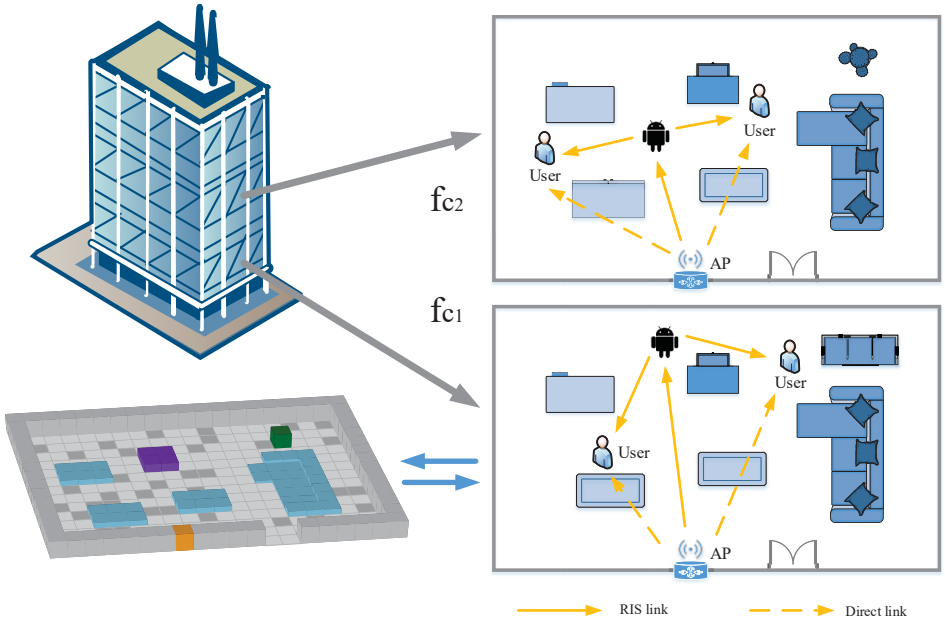


Figure 4.1: System model of NOMA enhanced mobile RIS

in the room and they follow independent random movements [149]. Since users are constantly roaming, in order to maximise the channel gain, the robot carried RISs have to be deployed opportunely according to the real-time user distribution. The robot operates on the floor and the RIS is set at a fixed height on the robot, as a result the altitude of RIS is considered as a constant. In order to ensure safe operations, the robot cannot cross or collide with any obstacles, it also has to be guaranteed that the RIS will not collide with people.

Remark 3. *The fixed-position RIS is likely to encounter blind spots when it is employed in indoor scenarios since furniture and room structures form a complex sheltered environment. Whether the RIS is mounted on the wall or ceiling, the LoS blind zone may be caused by girders, pillars, or chandeliers, and users in the blind zone can only get the NLoS channel. On the contrary, the RIS mounted on the robot can be deployed timely according to the user's location, which can improve the probability of LoS propagation for users.*

In this model, the set of APs is denoted as $u \in \mathbb{U} = \{1, 2, 3...U\}$, and the set of users associated with AP u is denoted as $k_u \in \mathbb{K}_u = \{1, 2, 3...K_u\}$. Users have to be associated with the AP on the same floor and the RISs employed are denoted as $r \in \mathbb{R} = \{1, 2, 3...R\}$. For a clear expression, the AP, RIS, and the agent employed in the same cell are defaulted to have a corresponding order, for instance, if the AP order is $u = 1$, the RIS working with u is $r = 1$. To express the concatenated propagation caused by RIS, the channel matrix of the link between AP and RIS is denoted as $\boldsymbol{h}_{u,r} \in \mathbb{C}^{M \times N}$, and the links of the r-th RIS to users are denoted as $\boldsymbol{h}_{r,k} \in \mathbb{C}^{N \times 1}$. On the other hand, users can also receive the signal via the direct link (AP to user link). Thus, the channel between AP u and user k can be denoted as $\boldsymbol{h} \in \mathbb{C}^{M \times 1}$.

The passive beamforming of RISs is considered as one of the main optimisation variables. The phase shift matrix of the RIS is denoted as $\boldsymbol{\Theta}_r = \text{diag}[\beta_1 e^{j\theta 1}...\beta_n e^{j\theta_n}...\beta_N e^{j\theta_N}]$, where β_n represents the amplitude of complex reflection coefficient and $\theta_n \in [0, 2\pi)$ represents the phase shift. On the contrary, since the main research scope of this chapter is the joint optimisation of the deployment and passive beamforming of RIS, the active beamforming at the AP side is solved by a conventional zero-forcing beamforming [150].

4.2.2 Interior Layout Modeling

Prior to discussing indoor propagation and RISs' deployments, it is necessary to establish an interior layout model. With the assistance of the layout model, it would be able to determine whether there is LoS path between any two points in the indoor environment, which is one of the key knowledge for RISs to obtain significant channel gains.

In order to accurately represent the outline of the furniture, a number of fictitious bricks are engaged to construct the layout model instead of simple columns. For example, a digitized layout model for the office is shown in the lower left corner of Fig. 4.1. Please note that theoretically this modeling method can describe any shape or object, but it will lead to a rise of computational complexity since each virtual brick has to be traversed

to determine whether it occludes the LoS path.

4.2.3 Propagation Model

In order to simulate the indoor propagation, the statistical prorogation model is not preferred since there is only a close range for indoor transmission distance (in metres) and a deterministic propagation model is more conducive to precise planning the path of the carrier robot. Thus, the aforementioned interior layout model and the indoor propagation model proposed by ITU recommendation [145] are adopted to obtain a deterministic indoor propagation model.

A propagation model including path loss and small-scale fading is considered, which can be express as

$$\mathcal{L}_{k_u}^{u}(d) = L_{k_u}^{u}(d) - 10\log_{10} h_{k_u}^{u}, \tag{4.1}$$

where $h_{k_u}^{u}$ denotes the Rician fading and $L_{k_u}^{u}(d)$ represents the pass loss described in [145, 151]. With the aid of interior layout model and intersection detection [152], the LoS state of a link can be calculated. Then we can obtain deterministic pass loss

$$L_{k_u}^{u}(d) = \begin{cases} L_{\mathrm{LoS}}(d), & \text{if LoS,} \\ L_{\mathrm{NLoS}}(d,n), & \text{if NLoS.} \end{cases} \tag{4.2}$$

For the NLoS link, the path loss can be calculated as

$$L_{\mathrm{NLoS}}(d,n) = L_0 + \mathcal{N}\log_{10} d + L_f(n), \tag{4.3}$$

where variable d represents the separation distance between the transmitter and the receiver and n represents the the number of completely blocked obstacles, determined by the number of the walls or floors. $\mathcal{N}$ denotes the distance power loss coefficient, as

suggested in [145], $\mathcal{N} = 25.5$ is chosen for the proposed office scenario. The parameter f represents the carrier frequency in MHz. Please note that although discrepant frequency bands are invoked on adjacent floors, these center frequencies have to be adjacent to avoid the heterogeneity in transmission characteristics.

The term L_0 represents the basic transmission loss that can be calculated as

$$L_0 = 20 \log_{10} f - 28, \tag{4.4}$$

and

$$L_f(n) = 15 + 4(n-1). \tag{4.5}$$

The path loss for the LoS link can be calculated as

$$L_{\text{LoS}}(d) = 16.9 \log_{10} d - 27.2 + 20 \log_{10} f. \tag{4.6}$$

4.2.4 Signal Model

4.2.4.1 OMA Scheme

In each cell, an FDMA scheme is adopted and some users utilize the same frequency band by invoking ZF beamforming to eliminate interference. The pre-coded transmitting signal from AP u can be expressed as

$$x^u(t) = \sum_{k_u=1}^{K_u} \sqrt{P^u_{k_u}(t)} \boldsymbol{g}^u_{k_u}(t) s^u_{k_u}(t), \tag{4.7}$$

where $s^u_{k_u}(t)$ represents the data symbol sequence from AP u to user k_u and $P^u_{k_u}(t)$ is the allocated power for user k_u. $\boldsymbol{g}^u_{k_u} \in \mathbb{C}^{M \times 1}$ represents the active beamforming vector.

Thus, the received signal at user k can be calculated as

$$y_{k_u} = (\boldsymbol{h}_{u,k_u} + \boldsymbol{h}_{r,k_u}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\sum_{k_u=1}^{K_u}\sqrt{P_{k_u}^u}\boldsymbol{g}_{k_u}^u s_{k_u}^u + n_0, \tag{4.8}$$

where n_0 denotes the additive white Gaussian noise (AWGN)[1]. As aforementioned, the active beamforming matrix $\boldsymbol{g}_{j_u}^u$ is derived by a ZF approach to mitigate the interferences. For a given user k_u and interference user j_u the normalized pre-coding matrix of can be calculated as

$$\begin{cases} (\boldsymbol{h}_{u,k_u} + \boldsymbol{h}_{r,k_u}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_{k_u}^u = 1, \\ (\boldsymbol{h}_{u,j_u} + \boldsymbol{h}_{r,j_u}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_{k_u}^u = 0, \quad j_u \neq k_u. \end{cases} \tag{4.9}$$

We denote the ZF pre-coding matrix of AP u as $\boldsymbol{G_u} = [\boldsymbol{g_1^u}, ...\boldsymbol{g_{k_u}^u}, ...\boldsymbol{g_{K_u}^u}]$. If we denote the channel response as matrix, where $\boldsymbol{H}_{u,k_u} = [\boldsymbol{h}_{u,1}, ..., \boldsymbol{h}_{u,K_u}]$, and $\boldsymbol{H}_{r,k_u} = [\boldsymbol{h}_{r,1}, ..., \boldsymbol{h}_{r,K_u}]$. As a result, the direct channel and the concatenated channel can be regarded as an overall channel response as $\boldsymbol{H_u} = \boldsymbol{H}_{u,k_u} + \boldsymbol{H}_{r,k_u}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r}$. Thus, the pre-coding matrix $\boldsymbol{G^u}$ can be calculated as the pseudo-inverse of overall channel response following $\boldsymbol{G_u} = \boldsymbol{H_u}(\boldsymbol{H_u}^H\boldsymbol{H_u})^{-1}$.

Therefore, based on (4.8) the SINR for user k can be calculated as

$$\gamma_{k_u} = \frac{|\,(\boldsymbol{h}_{u,k_u} + \boldsymbol{h}_{r,k_u}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_{k_u}^u\sqrt{P_{k_u}^u}\,|^2}{|\,(\boldsymbol{h}_{u,k_u} + \boldsymbol{h}_{r,k_u}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\sum\limits_{j_u\neq k_u}\boldsymbol{g}_{j_u}^u\sqrt{P_{j_u}^u}\,|^2 + \sigma^2}, \tag{4.10}$$

where σ^2 is the average power of the AWGN [2]. Consequently, the data rate of user k_u at time t can be calculated as

$$\mathcal{R}_{k_u}^u = B_{k_u}\log 2\,(1+\gamma_{k_u})\,, \tag{4.11}$$

[1]For brevity of the express, the path loss $\mathcal{L}_{k_u}^u(t)$ is implicitly included in $\boldsymbol{h}_{u,k_u}$ and the time symbol (t) is omitted in the subsequent equation to achieve a concise expression.

[2]If the multiple access approach is assumed to be ideally orthogonal, the inter-user interference can be considered as zero.

where B_{k_u} represents the bandwidth allocated to user k served by AP u.

4.2.4.2 NOMA Scheme

Contrary to the OMA scheme, the NOMA technique allows multiple users to form a cluster and utilize the same frequency band simultaneously. Hence, for each user cluster $v \in \mathbb{V} = (1, 2...V)$, and we denotes the users in cluster v as k_v. We also assume that the maximum callable power of each cluster is the same, and the transmitted signal for cluster v can be expressed as

$$x^v = \sum_{k_v=1}^{K_v} \sqrt{P_{k_v}^u(t)} s_{k_v}^u(t), \tag{4.12}$$

and then the transmitting signal of AP can be expressed as

$$x^u = \sum_{v=1}^{V} \boldsymbol{g}_v^u \sum_{k_v=1}^{K_v} \sqrt{P_{k_v}^u(t)} s_{k_v}^u(t), \tag{4.13}$$

where $\boldsymbol{g}_v^u$ represents the ZF pre-coding matrix. Therefore, the received signal of user k in the NOMA cluster v served by AP u can be expressed as

$$\begin{aligned} y_{k_v} =& (\boldsymbol{h}_{u,k_v} + \boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u x_{k_v}^u + \underbrace{(\boldsymbol{h}_{u,k_v} + \boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u \sum_{j_v>k_v}^{K_v} x_{j_v}^v}_{\text{intra-cluster interference}} + \\ & \underbrace{(\boldsymbol{h}_{u,k_v} + \boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r}) \sum_{\mathfrak{v}=1,\mathfrak{v}\neq v}^{V} \boldsymbol{g}_{\mathfrak{v}}^u x^{\mathfrak{v}}}_{\text{inter-cluster interference}} + n_0, \end{aligned} \tag{4.14}$$

which include the desired signal $(\boldsymbol{h}_{u,k_v} + \boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{h}_{u,r})\boldsymbol{g}_v^u x_{k_v}^u$ of user k_v, the intra-cluster interference, and the inter-cluster interference received by user k_v.

In order to obtain comparable results, the same ZF beamforming is also invoked at the NOMA AP to partially eliminate inter-cluster interference. Similar with the OMA

case, the pre-coding for NOMA can be expressed as

$$\begin{cases} (\boldsymbol{h}_{u,k_v} + \boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u = 1, \\ (\boldsymbol{h}_{u,k_\mathfrak{v}} + \boldsymbol{h}_{r,\mathfrak{v}}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u = 0, \quad v \neq \mathfrak{v}. \end{cases} \tag{4.15}$$

The derivation process of the pre coding matrix for NOMA is the same as the OMA scheme, thus we can also obtain it as

$$\boldsymbol{G}_{\boldsymbol{v}}^{\boldsymbol{u}} = \boldsymbol{H}_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{H}_{\boldsymbol{u},\boldsymbol{v}}{}^H\boldsymbol{H}_{\boldsymbol{u},\boldsymbol{v}})^{-1}. \tag{4.16}$$

where $\boldsymbol{H}_{\boldsymbol{u},\boldsymbol{v}}$ represents the overall fading matrix for NOMA clusters. It can be observed in (4.13) that instead of design beamforming for each individual user in the OMA scheme, a beam in the NOMA system is designed for a NOMA cluster. However, users in the same NOMA cluster also have a probability to own different CSI, and ZF beamforming cannot perfectly eliminate inter-cluster interference for all users. For example, assuming users j_v and i_v are in the same cluster v with channel $(\boldsymbol{h}_{u,j_v} + \boldsymbol{h}_{r,j_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r}) \neq (\boldsymbol{h}_{u,i_v} + \boldsymbol{h}_{r,i_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})$. According to (4.15) we have $(\boldsymbol{h}_{u,j_v} + \boldsymbol{h}_{r,j_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_{k_\mathfrak{v}}^u = 0$ and it is easy to figure out $(\boldsymbol{h}_{u,i_v} + \boldsymbol{h}_{r,i_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_{k_\mathfrak{v}}^u \neq 0$, which suggests the inter-cluster interference cannot be removed completely at user j_v. Given the background that ZF beamforming can only remove inter cluster interference for a part of user, the user with the highest equivalent channel gain[3] in each cluster is selected as representative j_v to determine the ZF beamforming vector for the cluster. Therefore, the inter-cluster interference can be eliminated at the strongest user in each cluster but weaker users still have to suffer.

A portion of intra-cluster interference can be eliminated by SIC and the intra-cluster interference for each user can be calculated with a given decoding order. Since users and RISs keep moving, the channel quality will be changing, so a dynamic decoding order has to be determined in each time slot. For the convenience of presentation, we assume

[3]For convenience of expression, we call the user with the highest equivalent channel gain in each cluster as the strongest user.

that users in NOMA cluster v have a consistent numbering order with channel quality at time t, where user K is the strongest user. Consider user j_v and k_v at time t have relationship described as

$$\frac{\mid (\boldsymbol{h}_{u,j_v} + \boldsymbol{h}_{r,j_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u \mid^2}{\sum\limits_{\mathfrak{v}=1,\mathfrak{v}\neq v}^{V} \mid (\boldsymbol{h}_{u,j_v} + \boldsymbol{h}_{r,j_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_{\mathfrak{v}}^u \mid^2 + \sigma^2} > \frac{\mid (\boldsymbol{h}_{u,k_v} + \boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u \mid^2}{\sum\limits_{\mathfrak{v}=1,\mathfrak{v}\neq v}^{V} \mid (\boldsymbol{h}_{u,k_v} + \boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_{\mathfrak{v}}^u \mid^2 + \sigma^2}. \tag{4.17}$$

According to the SIC principle, user j_v can adopt the SIC to remove the signal for user k_v in prior of decoding the signal for itself [153]. Thus, rewriting (4.17) in to a concise shape as $\Omega_{j_v}^u > \Omega_{k_v}^u$, where $\Omega_{k_v}^u = \frac{|(\boldsymbol{h}_{u,k_v}+\boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u|^2}{\sum\limits_{\mathfrak{v}=1,\mathfrak{v}\neq v}^{V}|(\boldsymbol{h}_{u,k_v}+\boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_{\mathfrak{v}}^u|^2+\sigma^2}$, $\Omega_{j_v}^u = \frac{|(\boldsymbol{h}_{u,j_v}+\boldsymbol{h}_{r,j_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u|^2}{\sum\limits_{\mathfrak{v}=1,\mathfrak{v}\neq v}^{V}|(\boldsymbol{h}_{u,j_v}+\boldsymbol{h}_{r,j_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_{\mathfrak{v}}^u|^2+\sigma^2}$, and then the decoding order can be denote as $\Omega_{j_v}^u > \Omega_{k_v}^u, j_v > k_v$.

Therefore, with a correct decoding order, the received SINR for user k_v in the NOMA network can be calculated by

$$\gamma_{k_v}^u = \frac{\mid (\boldsymbol{h}_{u,k_v} + \boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u\sqrt{P_{k_v}^u}s_{k_v}^u \mid^2}{\mid (\boldsymbol{h}_{u,k_v} + \boldsymbol{h}_{r,k_v}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r})\boldsymbol{g}_v^u \sum\limits_{j_v=k_v+1}^{K_v}\sqrt{P_{j_v}^u}s_{j_v}^u \mid^2 + \mid (\boldsymbol{h}_{u,k_u} + \boldsymbol{h}_{r,k_u}\boldsymbol{\Theta}_r\boldsymbol{H}_{u,r}) \sum\limits_{\mathfrak{v}=1,\mathfrak{v}\neq v}^{V}\boldsymbol{g}_{\mathfrak{v}}^u x^{\mathfrak{v}} \mid^2 + \sigma^2}. \tag{4.18}$$

At last, the data rate of user k_v served by AP u can be calculated as

$$\mathcal{R}_{k_v}^u = B_{k_v}^u \log 2\left(1 + \gamma_{k_v}^u\right). \tag{4.19}$$

4.2.5 Problem Formulation

This chapter aims to maximise the sum data rate of users by jointly optimising robot-mounted RISs' deployment $\mathcal{D}_r = \{D_r(1), D_r(2), ...D_r(t)...\}, r \in \mathbb{R}$, and the phase shift

for all reflecting element $\mathbf{\Theta}_r = \{\mathbf{\Theta}_r(1), \mathbf{\Theta}_r(2), ...\mathbf{\Theta}_r(t)...\}, r \in \mathbb{R}$ of the mobile RIS, where $D_r(t) = [x_r(t), y_r(t), z_r(t)]$ represents the position of mobile RIS r at time t. Meanwhile, since APs need to collaborate with RISs, the corresponding power allocation policy $\mathcal{P}_r = \{P^u(1), P^u(2), ...P^u(t)\}, u \in \mathbb{U}$. Thus, the optimisation problem can be formulated as

$$\max_{\mathcal{D}_r, \mathcal{P}_r, \mathbf{\Theta}_r} \sum_{k_u=1}^{K_u} \sum_{u=1}^{U} \mathcal{R}_{k_u}, \tag{4.20a}$$

$$\text{s.t.} \quad x_{\min} \leq x_r(t) \leq x_{\max}, \forall r, \forall t,$$

$$y_{\min} \leq y_r(t) \leq y_{\max}, \forall r, \forall t, \tag{4.20b}$$

$$\sum_{k_v \in \mathbb{K}_v} P^u_{k_v}(t) \leq P^u_{v\max}, \forall t, \forall v, \forall u, \tag{4.20c}$$

$$k^u_v(t) < j^u_v(t), \forall (k, j), \forall t, \forall k, \forall u, \tag{4.20d}$$

$$\mathcal{R}(t) \geq \mathcal{R}_{\text{QoS}}, \forall t, \forall k, \forall u, \tag{4.20e}$$

where constraint (4.20b) ensures that the mobile RISs have to be deployed in the appointed room, since once a mobile RIS is moved to other areas, it may cause unexpected interference especially when multiple RISs are deployed in the same room. Constraint (4.20c) is a power constraint that the total power allocated to users in a cluster cannot exceed the maximum power that the cluster is authorized to invoke while in the OMA scheme a signal user can be regarded as a cluster. It is worth noting that the power constraint is for the AP side instead of the RIS, and the energy consumption of the robot is not count in this constraint. Constraint (4.20d) is introduced to ensure that the user ordering and decoding order can be performed correctly in each NOMA cluster. Finally, taking into account the fairness of users, constraint (4.20e) represents the data rate of each user at any time t is guaranteed to meet the minimum rate of QoS requirement. As mentioned above, the predicament of the optimisation is that the formulated problem is dynamic, non-convex [154] and the obstructive environment is non-functional. It is worth mentioning that the phase shift optimisation in the NOMA scenario not only provides channel enhancement for users, but the channel modification has to be NOMA-friendly

as well to take care of user fairness. Therefore, a DRL algorithm is invoked to solve the formulated problem.

4.3 Federated Learning Model

Subsection 4.3.1 elaborates on the role and superiority of invoking FL to coordinate multiple agents and prove that there is a theoretical gain in FL-DRL framework. In subsection 4.3.2, an FL model with local training is proposed to serve multiple cell networks.

4.3.1 Enhancing DRL by FL

Federated learning is competent to be invoked for optimising the proposed communication model. As aforementioned, the proposed indoor network composed of APs has cellular characteristics to extend, and independent agents served in each cell have great common functions and attributes. For example, the pursuit of service quality, the equipment of RISs and the propagation characteristics of signals in each cell are equal, which constitutes the cornerstone of adopting the FL framework.

In addition to advantages mentioned in Chapter 2, DRL algorithms are specifically suitable to be applied in the FL framework. The learning process of RL comes from continuously interacting with the environment and exploring different states and actions. However, the exploration of the environment is not likely to exhaust all states though the action policy contains random actions or noise [155, 156], which leads to the global optimum may not being discovered. In particular, the proposed communication scenarios and indoor layouts have high complexity, impelling the efficient and sufficient exploration to be a problem. Therefore, the training effect of DRL is determined by whether the agent has sufficient exploration and experience, fortunately, the participation of FL is helpful to reveal more different states since multi-agents are investigating the environment, which

allows the environment to be explored more sufficiently.

Remark 4. *When the environments explored by the DRL agents have similarities and the state transitions have not been traversed by agents, the FL framework can provide potential gains than independent training scheme since it is likely to obtain more sufficient environmental knowledge.*

Therefore, the FL scheme has more expected gain than independent agents until all state transitions have been traversed by agents. Furthermore, the global model can also greatly enriches the experience diversity since each agent has different initialization and pseudorandom. In summary, by establishing a global model and exchanging neural network parameters, agents located on different floors or cells can learn from each other's experiences. The introduction of FL can improve the training efficiency and effect of DRL algorithms and the gain is also validated by the simulation results in Section 4.5.

4.3.2 FL Model for DRL

The proposed FL framework adopts decentralized training and uses federated averaging to generate a global model. The operation process can be divided into three parts: local training, updating global model, and downloading global model, which is illustrated in Fig. 4.2.

- Local training: Each local agent set up their local model ω_t^u and uses its own computing resources to train the local model, where t represents the time and u represents the agent number. Local neural network models have random initialization to increase the diversity of exploration at early training.

- Global model update: After a period of training interval F_G, the parameters of the global model ω_t^G can be upgraded by averaging the parameters of each local

model, which can be express as

$$\omega_t^G = \frac{1}{U}\sum_{u=0}^{U}\omega_t^u. \tag{4.21}$$

- Local model update: After the global model is updated, each agent downloads the global model and then updates the local model according to the global model.

$$\omega_t^u = \omega_t^G, \forall u. \tag{4.22}$$

After the updating is complete, the new model can be used for the next round of local training.

In user-based federated learning models, edge devices, such as mobile phones, play the role of agents. Consequently, federated learning needs to upload and download models over wireless channels, which incurs spectral overhead. In the proposed model, since there is a wired connection between the controller (agent) and the AP, no spectral overhead is required for model transmission.

4.4 FL-DDPG executed optimisation for Mobile RISs

With the aforementioned FL framework, this section details the FL enhanced DDPG algorithm to optimise the deployment, phase shifts of RISs, and the power allocation for users. The algorithm training and decision flow is explained in subsection 4.4.1. As a DRL approach, the specific state space and action space design for mobile RIS scenario is presented in subsection 4.4.2, and the adaptive neural network structure is introduced in subsection 4.4.3. Furthermore, subsection 4.4.4 analyses the convergence and complexity of the FL-DDPG algorithm.

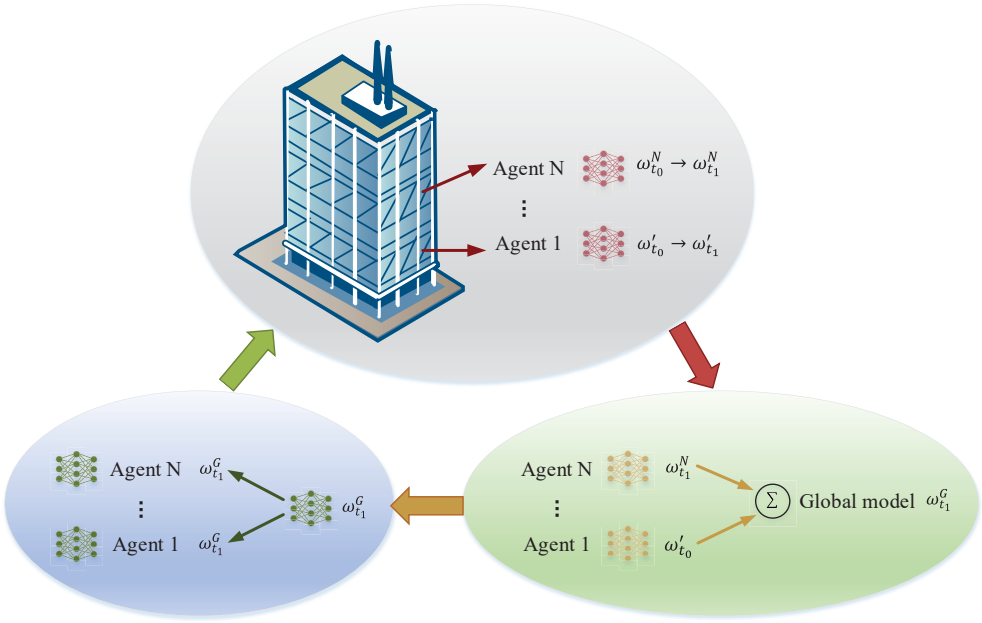


Figure 4.2: Federated learning enhanced indoor mobile RIS networks

4.4.1 FL-DDPG Algorithm and Training

An FL-DDPG algorithm is proposed to jointly optimise deployments, phase shifts of mobile RIS and the corresponding power allocation policy for users in each cell. Additionally, several improvements are implemented on the original DDPG algorithm [141], such as decaying OU noise and adaptive neural network structure to adapt the algorithm into the proposed communication scenarios. Each local agent is assumed to be deployed within the AP and it can control the actions of the RIS and the carrier robot via the control channel. Due to the actor-critic structure, four neural networks are used in the DDPG agent, namely the actor network μ, the critic network Q, the actor target network μ' and the critic target network Q'. Once observing the environment state $\mathbf{s}_t$, the actor network calculates the action $\mathbf{a}_t$ and then it will be executed. After the action is executed, the state will be changed to $\mathbf{s}_{t+1}$, and the reward r_t will be calculated according

to the data rate $\mathcal{R}_t$ and QoS requirement threshold. The detailed update flow of a single DDPG agent is presented in Fig. 4.3.

In order to train the agent efficiently, decaying OU noise is adopted in the training process

$$\mathbf{a}_t = \mu(\mathbf{s}_t|\omega_t^\mu) + N(0, \xi_t), \xi_t = \xi_0 \to 0, \xi_0 \in [1, 0), \tag{4.23}$$

where ω_t^μ represents the parameters of neural network μ and ξ_t denotes the scale of the OU noise. The OU action noise can drive the agent to explore further diversely compare to the Gaussian noise [143], and decreasing noise can improve exploration efficiency without loss of convergence. On the other hand, memory replay technology is adopted in the model. The agent record and store the transition $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ for each step into a replay memory buffer and randomly sample experiences at each step and train neural networks according to the samples. For single sample at each step, the actor network can be updated according to the policy gradient. Assuming the minibatch has e transition samples, the policy gradient can be calculated as

$$\nabla_{\omega^\mu} J = \frac{1}{e} \sum_e \nabla_{\mathbf{a}} Q(\mathbf{s}_{t=e}, \mathbf{a}_{t=e}|\omega^Q) \nabla_\omega^\mu \mu(\mathbf{s}_{t=e}|\omega^\mu). \tag{4.24}$$

The critic network is in charge of evaluating the action value (Q-value) of the action taken actions taken in a certain state, which is similar as the Q-learning and deep Q-network (DQN) algorithms. A Q-value with a concern of long-term reward is defined by the Bellman equation

$$Q(\mathbf{s}_t, \mathbf{a}_t) = r_t(\mathbf{s}_t, \mathbf{a}_t) + \beta \max Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}). \tag{4.25}$$

In order to accurately estimate Q-value, the critic network is updated by minimising

the loss function

$$L_e = \frac{1}{e}\sum_{e}(y_{t=e} - Q(\mathbf{s}_{t=e}, \mathbf{a}_{t=e}|\omega^Q))^2, \tag{4.26}$$

where

$$y_t = r_t(\mathbf{s}_t, \mathbf{a}_t) + \beta Q'(\mathbf{s}_{t+1}, \mu'(\mathbf{s}_{t+1}|\omega^{\mu'})|\omega^{Q'}). \tag{4.27}$$

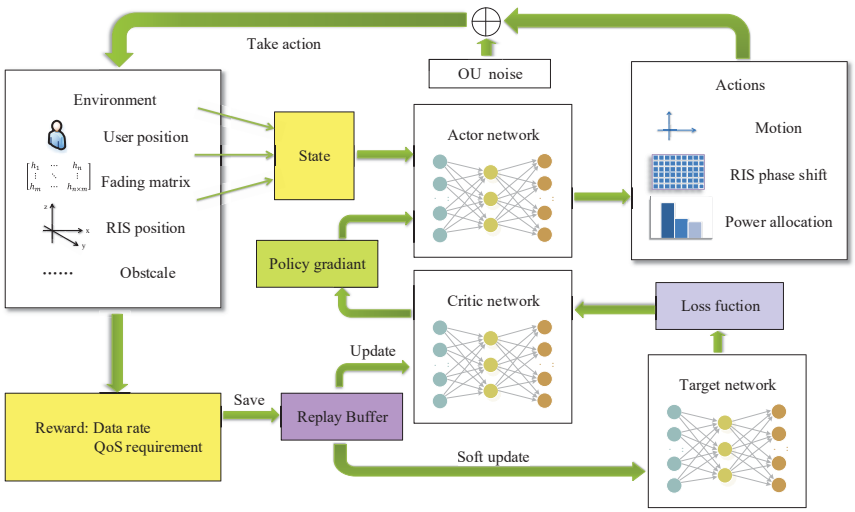


Figure 4.3: Flow diagram of the local training in the FL-DDPG algorithm

4.4.2 State, Action and Reward Function

The DDPG algorithm supports continuous state and action space. Therefore, regardless of movement, phase shifts and power allocation are designed to be continuous to obtain the accurate action, and the state, action space, and reward function are designed as

Algorithm 3 FL-DDPG algorithm for the sum rate optimisation

for each cell $u \in \mathbb{U}$ **do**
  Initialize the environment and determine the neural network specifications based on the number of RIS elements
  Initialize the actor network ω_u^{μ}, critic network ω_u^{Q}, target actor network $\omega_u^{\mu'}$, target critic network $\omega_u^{Q'}$ with random parameters
  for each episode $\mathcal{E}$ **do**
    if $\mathcal{E}$ % F_G =0 **then**
      Update global model $\omega_G^{\mu,\mu',Q,Q'} = \frac{1}{U}\sum_{u=0}^{U} \omega_u^{\mu,\mu',Q,Q'}$
      Update local models $\omega_u^{\mu,\mu',Q,Q'} = \omega_G^{\mu,\mu',Q,Q'}$.
    end if
    Reset the environment and initial state
    for each step in $t_0 \leq t \leq t_{\max}$ **do**
      Observe $\mathbf{s}_t$ according to the radio map
      Choose $\mathbf{a}$ according to action policy and $\mu(\mathbf{s}, \omega_t^{\mu})$
      IRs take action $\mathbf{a}$, observe r_t and $\mathbf{s}_{t+1}$
      Record $e\{\mathbf{s}_t, \mathbf{a}, r, \mathbf{s}_{t+1}\}$
      Random sample a batch of transection e from memory buffer
      Calculate target according to (4.27)
      Train critic network $Q(\mathbf{s}, \omega^{Q})$ with a gradient descent step (4.26)
      Train actor network $\mu(\mathbf{s}, \omega^{\mu})$ with (4.24)
      Update the target networks $\omega^{\mu'} \leftarrow (1-\tau)\omega^{\mu'} + \tau\omega^{\mu}$,$\omega^{Q'} \leftarrow (1-\tau)\omega^{Q'} + \tau\omega^{Q}$
      $\mathbf{s}_t \leftarrow \mathbf{s}_{t+1}$
    end for
    Each agent save the network models ω_u^{μ}, ω_u^{Q}, $\omega_u^{\mu'}$, $\omega_u^{Q'}$
  end for
end for

follows.

4.4.2.1 State Space

For a single agent, the state space $\mathbf{s}_t$ contains four components, the RIS location $D_r(t)$ in time slot t, user location $D_{k_u}(t), k_u \in \mathbb{K}_u$, pass loss for each user $L_{k_u}^{u}(t)$ and fading matrixes $\boldsymbol{h}_{u,k}, \boldsymbol{h}_{u,r}$ and $\boldsymbol{h}_{r,k}$. Thus, the state for time slot t can be noted as

$$\mathbf{s}_t = \{D_r(t), D_{k_u}(t), L_{k_u}^{u}(t), \text{real}\{\boldsymbol{h}_{u,k}, \boldsymbol{h}_{u,r}, \boldsymbol{h}_{r,k}\}, \text{imag}\{\boldsymbol{h}_{u,k}, \boldsymbol{h}_{u,r}, \boldsymbol{h}_{r,k}\}\}, k_u \in \mathbb{K}_u. \tag{4.28}$$

Since the elements in fading matrixes are complex numbers, the real and imaginary parts of each element can be split and input to different nodes. These selected input carry the necessary information for decision making, while the deployment plan requires location information, and the optimisation of power allocation and phase shifts is based on CSI.In addition, since the state is composed of different variable categories, and the values may be very different (e.g. Value of $D_r(t)$ may be 20 but the values in $\boldsymbol{h}_{u,k}$ may be $10e-10$). Thus, proper scaling is necessary to avoid some values being neglect by the DNN.

4.4.2.2 Action Space

The composition of the action space completely corresponds to the three optimisation parameters, including motion, phase shift and power allocation.

- Deployments: For the deployment, the agent does not calculate the optimal position but choosing the next move $\Delta D_r(t)$ for the robot at each time slot t. The proposed approach allows the agent to find the optimal movement at each moment, with a consideration of long-term reward. However, the method of directly finding an optimal position will cause the moving path of mobile RIS may not be optimal.

- Phase shifts: The agent calculates the optimal $\boldsymbol{\Theta}_r(t)$ at the current moment for each element express them in a radian system. The time for rotating the angle of reflecting elements is neglected.

- Power allocation policy: The agent allocate power $P^u_{k_u}(t)$ to each associated user k_u at each time slot, where the allocated power meets $P^u_{k_u}(t) < P^u_{\max k_u}(t)$. For the OMA scenario, $P^u_{\max k_u}(t) = P^u_{\max}/K_u$ but in NOMA cases users can have their own power upper bound while $\sum_{k_u} P^u_{\max k_u}(t) \leq P^u_{\max}/K_u$.

In summary, the action space can be noted as

$$\mathbf{a}_t = \{\Delta D_r(t), \boldsymbol{\Theta}, P_{k_u}^u(t)\}, k_u \in \mathbb{K}_u. \tag{4.29}$$

4.4.2.3 Reward Function

For each cell, in order to maximise the data rate of the system, the reward is set to be proportional to the sum rate of all users. As mentioned in (4.20e), in order to meet the user fairness constraint, once the data rate of any user does not meet the QoS requirements, a penalty has to be imposed. The agent will receive a discounted reward as in (4.30), where λ is the reduction factor

$$R_t = \begin{cases} \mathcal{R}^u(t), & \text{QoS requirement satisfied,} \\ \frac{\mathcal{R}^u(t)}{\lambda}, & \text{QoS requirement not satisfied.} \end{cases} \tag{4.30}$$

4.4.3 Neural Network Structure

The structures of the actor network and the critic network is presented in Fig. 4.4. Two batch normalization (BN) layers and an activation layer with relu function are employed in the actor network. The first BN layer is in charge of normalizing input data and the second BN layer ensures a valid input range for the tanh layer. Since all elements of fading matrices $\boldsymbol{h}_{u,k}, \boldsymbol{h}_{u,r}$ and $\boldsymbol{h}_{r,k}$ need to be input to the actor network as the basis for the phase shift optimisation, then the size of the input dimension have to be adaptive and determined by the number of users and the number of elements in RIS. The size of the hidden layer should also be adjusted accordingly to the communication system to achieve a proper fitting effect. The empirical number of the activation layer nodes is $\omega_{\text{relu}} = 4MN$, where the position input is not counted since it adds a negligible input dimension. A similar structure is adopted in the critic network. Since the critic network only needs to output a Q-value, its hidden layers can have a minor size, although the

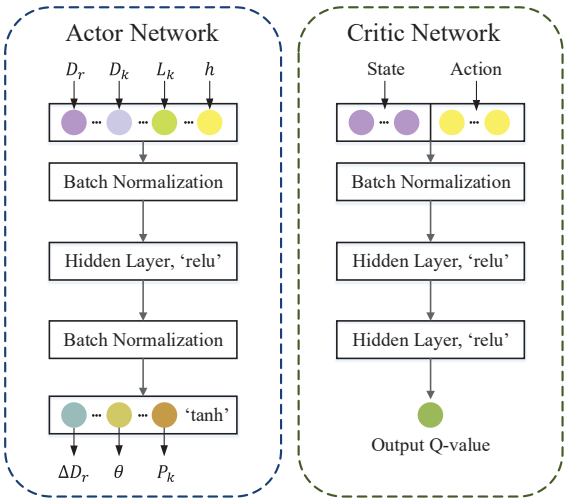


Figure 4.4: Neural network structure of the proposed FL-DDPG algorithm

critic network has a larger input dimension.

4.4.4 Convergence and Computational Complexity Analysis

The convergence of the basic Q-learning has been proved in a series of literatures, such as [157]. However, due to the introduction of neural networks, the convergence of the DDPG algorithm is no longer guaranteed [141]. In fact, DRL algorithms may fail to converge under the interference of improper parameters setting. Nevertheless, the proposed FL-DDPG algorithm is capable to converge when a few constraints are met. If the learning rate, target network update rate and action noise are properly set, FL-DDPG can converge stably, which can be proved by simulation results displayed in Section 4.5.

The complexity of the FL-DDPG algorithm is determined by the size of the neural network employed. Since the local training approach is adopted, each agent trains the neural network by itself, so the complexity of each agent can be denoted as ζ^u can be calculated independently and the total complexity of the multi-agent system is $\zeta =$

$\zeta^G + \sum_{u=0}^{U} \zeta^u$, where ζ^G represents the complexity caused by the updating and downloading parameters of the globe model. Note that we assume a perfect model exchange, the error and complexity caused by communication and positioning are not taken into account.

The action selection for each step is the responsibility of the actor network ω^μ, and the number of nodes in the actor network is denoted as ω_{n}^μ for normalized nodes, ω_{r}^μ for relu nodes and ω_{t}^μ for 'tanh' nodes. Thus, the calculations complexity caused by the node computation is $5 \cdot \omega_{\mathrm{n}}^\mu + \omega_{\mathrm{r}}^\mu + 6 \cdot \omega_{\mathrm{t}}^\mu$ as suggested in [139]. Further, assuming the actor network has I layers in total and each layer i has $\|\omega_i^\mu\|$ nodes, the complexity required to propagate values between neural nodes and adding bias can be calculated as $\sum_{i=0}^{I} \|\omega_i^\mu\| \cdot \|\omega_{i+1}^\mu\|$. Then the complexity of actor network for a single step is $\zeta_{\omega^\mu} = 5 \cdot \omega_{\mathrm{n}}^\mu + \omega_{\mathrm{r}}^\mu + 6 \cdot \omega_{\mathrm{t}}^\mu + \sum_{i=0}^{I} \|\omega_i^\mu| \cdot \|\omega_{i+1}^\mu\|$. If the same assumption is applied to the critic network Q, since the critic network has to train e samples at each step, with the same calculation method, the complexity of the critic network is $\zeta_{\omega^Q} = e \cdot (5 \cdot \omega_{\mathrm{n}}^Q + \omega_{\mathrm{r}}^Q + 6 \cdot \omega_{\mathrm{t}}^Q + \sum_{i=0}^{I} \|\omega_i^Q| \cdot \|\omega_{i+1}^Q\|)$. Then, for the proposed scenario, which has t steps per episode, for a single agent the total complexity can be calculate as $\zeta_u = \mathcal{E} \cdot t \cdot (\zeta_{\omega^\mu} + \zeta_{\omega^Q})$, where $\mathcal{E}$ represents the episode number. On the other hand, the complexity caused by the globe model is $2 \cdot \mathcal{E}/F_G \cdot \|\omega^\mu\| + \|\omega^Q\|$, where F_G represents number of episodes interval of global model update, which is negligible compared to the local model training. Therefore, the total complexity can be express as $\zeta = \sum_{u=0}^{U} \zeta_{\omega_\mu}^u + \zeta_{\omega_Q}^u$. Considering that the DRL algorithm needs a period of training time, even if FL is adopted, the model training will take several minutes or even hours. Thanks to the fact that the DDPG algorithm is an off-policy algorithm, off-line training can be used in practice to avoid delays on decision making.

4.5 Numerical Results and Analysis

This section aims to exhibit numerical results of the FL-DDPG optimised mobile RIS system. In the simulation, each cell is assumed to serve four users and these users are

Table 4-A: Simulation Parameters

Parameter	Description	Value	Parameter	Description	Value
f_c	carrier frequency	2GHz	K	number of users	4
B_k^u	bandwidth	1 MHz	$P_{\max k}^u$	maximum transmitting power	20 dBm
$V_{\max}$	maximum speed of RIS	0.5 m/s	λ	QoS penalty coefficient	2
$y_{\max}$	room length	20 m	$x_{\max}$	room width	15 m
R_{QoS}	QoS require	10 kb/s	σ	noise power density	-30 dBm/MHz
α	learning rate	3×10^{-4}	γ	discount factor	1
e	batch size	64 samples	τ	target update rate	0.002

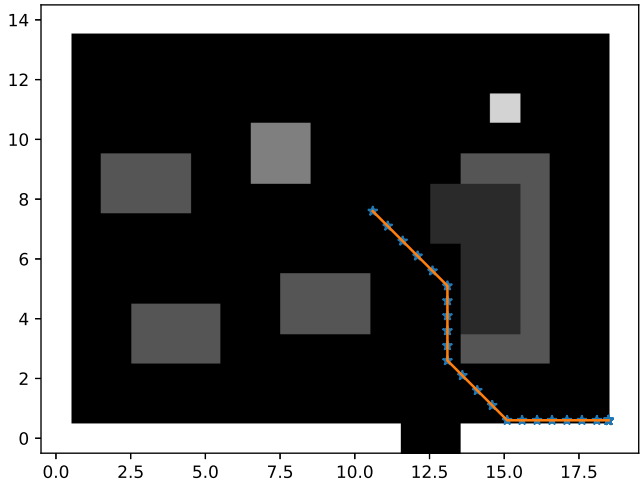


Figure 4.5: Optimised path for the mobile RIS

partitioned into two clusters. Each user makes a random movement on the horizontal plane at every time slot, the moving distance conforms to the Rayleigh distribution and the direction following the uniform distribution. The building structure of each cell is assumed to be the same, and the global model update frequency for the FL is 20 episodes. As for the agent, Adam optimisers are employed for the neural network training and the proper learning rate range is 5×10^{-4} to 10^{-5} according to simulation scenarios. The initial action noise scale is set as 0.4. The rest of the default parameters have been given in Table 4-A.

Fig. 4.5 exhibits a trajectory example of the mobile RIS derived from the proposed DDPG algorithm. In this figure, the orange curve records the trajectory of the mobile

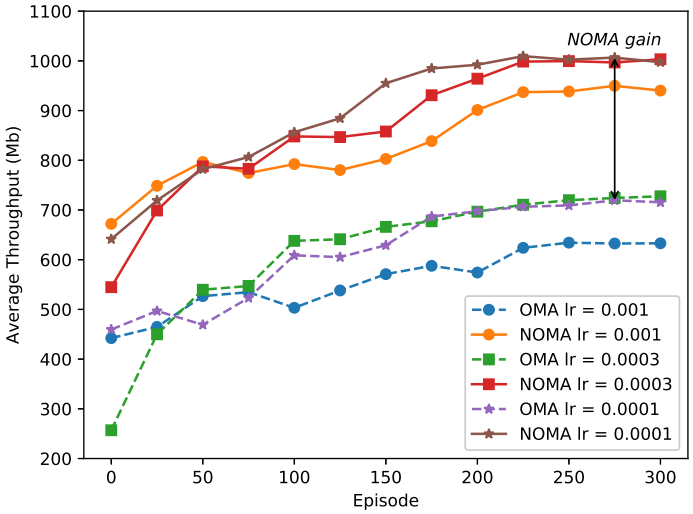


Figure 4.6: Mobile RIS performance with different learning rate

RIS and the blue stars represent the position where the robot stops at each discrete time slot, which is also the RIS position that is input into the neural network as a part of the state information. The mobile RIS is initially placed in the middle area of the office, and it moves to a corner gradually so that provides LoS cover for the large area blocked by the sofa. The gray and white blocks correspond to the furniture and walls of different heights. It can be observed that the derived path avoids obstacles and the data rate gain for the flexible deployment will be discussed later.

Fig. 4.6 demonstrates the training performance of the DDPG algorithm in a single cell. It can be observed that the average throughput of the system increases steadily over the training episodes and gradually flattens out in the late stage of training, which proves that the algorithm has stable convergence within a proper learning rate range. The throughput of well trained agents indicates that an inappropriately large learning rate can result in a debuff in optimisation performance. For example, when the learning rate is 0.001, the throughput suffers a decrease of approximately 7% compared to the other two learning rates. Moreover, a significant NOMA gain can be observed, which is around 42% compared to the OMA scheme under the same conditions.

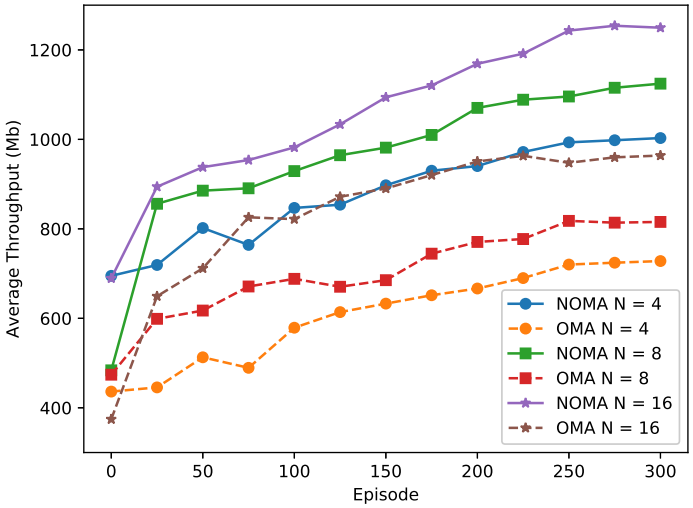


Figure 4.7: Mobile RIS performance with different number of reflection elements

The impact of the RIS reflecting elements number on system performance is investigated in Fig. 4.7. Logically, a larger amount of reflection elements can enhance the propagation to a superior extent and obtain further power gain. It can be observed that with the enhancement of 16 reflection elements, the OMA scheme obtain a data rate equivalent to the NOMA scheme with 4 reflection elements. Meanwhile, the stable convergence of results indicates although the different values of reflecting elements number N cause tremendous dimensional differences of the input state, by correspondingly adjusting the size of the neural network, the proposed algorithm can serve RIS with different specifications.

The throughput curves versus the transmit power are plotted in Fig. 4.8 and display both OMA and NOMA cases where the number of antennas M is 2 or 4. The data rate gain of the 4 antennas case is approximately 11.6% on average, compared to the case of double antennas. The NOMA gain is higher with the growth of the transmission power, the reason is when the transmission power is low, the weaker users are not likely to meet the QoS requirements and need to be allocated more power to ensure the fairness. Although this fairness-dominated power allocation scheme results in a reduction in data

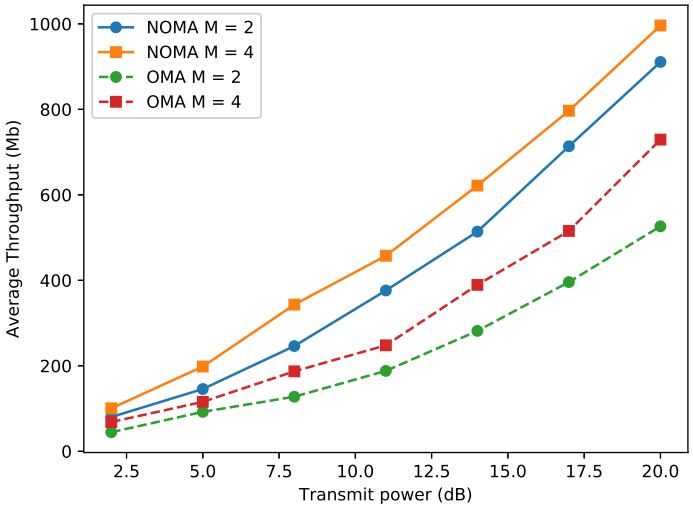


Figure 4.8: Achievable sum rate versus AP transmit power

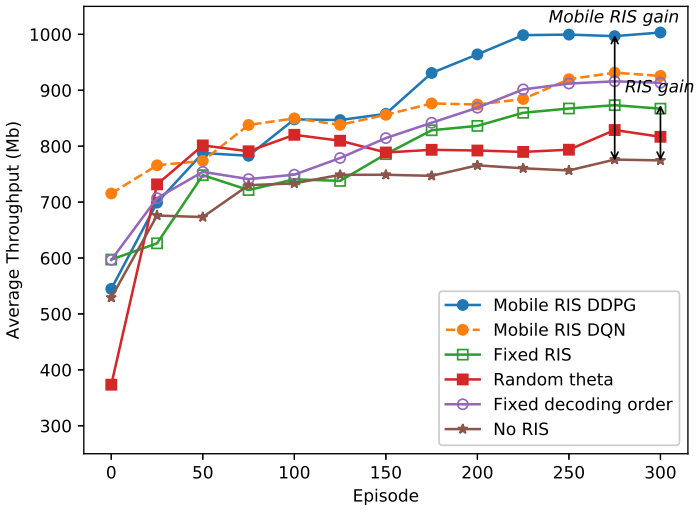


Figure 4.9: Date rate gain of each component in mobile RIS enhanced networks

rate gain, the NOMA scheme still achieves a noticeable gain in the case of small transmit power.

In order to determine the gain of the maneuver deployment and each other component in the mobile RIS model, Fig. 4.9 is plotted to show the throughput of the proposed

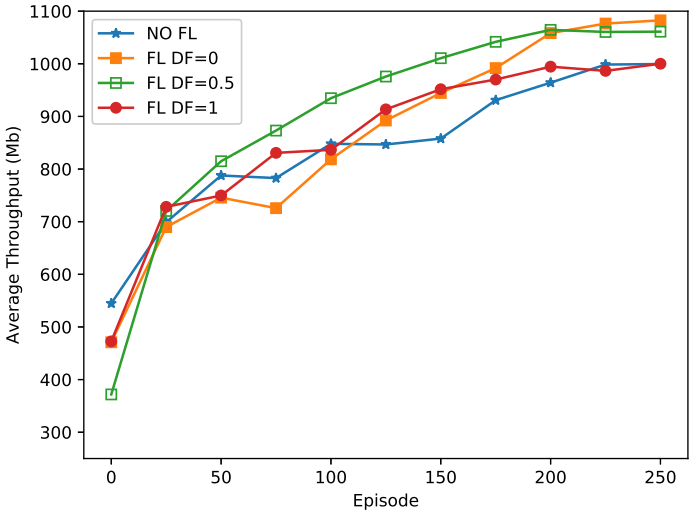


Figure 4.10: The performance of federated learning

model and benchmarks. First of all, the dynamic decoding order achieves a gain of 10.2% compared to the pre-settled static decoding order. By observing the blue and yellow curves, it can be found that the flexibly deployed RIS obtains an additional 15.1% data rate gain compared to the fixed RIS scheme, where the fixed RIS is settled at the start position in Fig. 4.5. It is worth noting that the performance improvement provided by the mobile RIS even exceeds the performance difference between the fixed RIS model and no RIS engaged network, which indicates the superiority of the mobile RIS framework is substantial. In addition, in contrast to the fixed RIS model, the mobile RIS has compelling compatibility for various user distributions. In order to investigate the effect of the phase shift optimisation, a RIS with random phase shifts is employed as another benchmark. It is undeniable that the RIS with random phase shifts also leads to a diminutive gain compared to the no RIS mode, but it is far inferior to the DRL optimised case. Meanwhile, the curve behaves unevenness even in the final episodes since the phase is not controlled by the agent.

Fig. 4.10 shows the impact of environmental differences on the performance of federated learning, where the difference factor (DF) represents the correlation of the fading

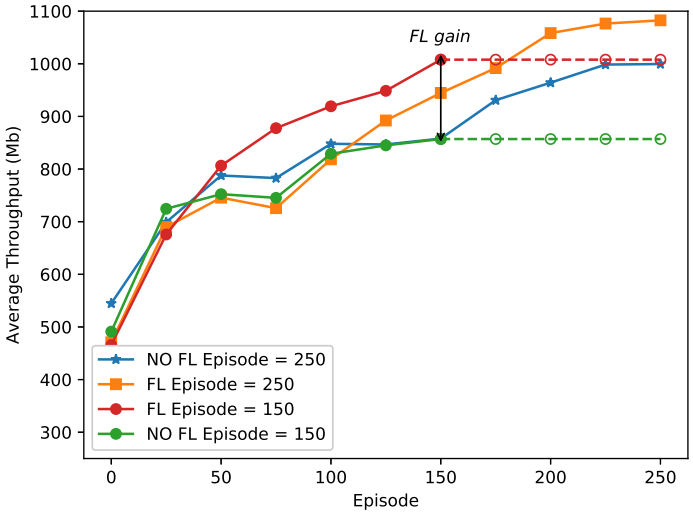


Figure 4.11: Training effect with/without federated learning

characteristic in different cells. DF is 0 means that the cells have the same channel characteristics. Obviously, FL achieved the optative performance in this case, since agents are in the same environments so that the model update has the highest efficiency. It is worth noting that even though the rooms have similar architectural structures, they have different fading characteristics due to the difference in decoration and surface materials. Therefore, the cases that cells with propagation differences are investigated, and DF = 1 suggests that the propagation characteristics of each cell are completely independent. It can be observed that even in the case of DF=1, FL-enhanced DRL still has stable convergence, and is capable to achieve a matched average sum rate to the single-cell case.

Fig. 4.11 reveals the gain of FL at different training maturities. It can be found at first that the introduction of FL can effectively save training time. With the aid of FL, agents only spend 150 episodes of training to achieve an equal performance that the single-cell scheme needs 250 episodes, which supports the statement in **Remark 4**. Since the DRL approaches train agents by replaying the obtained experiences, more diverse and richer experience of transitions obtained by FL makes the agents' decision-making

wiser. It is worth clarifying that despite fluctuations in throughput during early training epochs, this is not necessarily evidence of faster or slowe convergence. The throughput is jointly determined by the optimal strategy and the action noise. Due to a large action noise in the early stage, the early throughput has a strong randomness. It is undeniable that spending infinite training episodes can enable all states to be explored, so that the agents can converge to the same optimal level. However, in practice, under the condition that the training time is limited, FL has a significant training advantage compared to the mode without FL.

4.6 Summary

This chapter presented a NOMA enhanced wireless network model with the aid of mobile RISs that can provide NOMA craved channel conditions and improve channel quality for users. In order to optimise deployments and phase shifts of RISs and the corresponding power allocation for users, an FL enhanced DDPG algorithm has been proposed, which has preponderant performance under the same training extend compared to the independent DLR scheme since the engagement of FL lead to more sufficient exploration and experience exchange for agents. Simulation results proved that 1) Compared to the scenario without RIS, mobile RISs are capable to provide around 30.1% data rate gain that significantly exceeds the gain of the fixed RISs paradigm, which is 12.4%; 2) The NOMA scheme, where the proposed dynamic decoding identification method is applied, outperforms the OMA scheme by obtaining approximately 42% gain in terms of the sum rate; 3) The FL enhanced DDPG algorithm has stable convergence while the parameters are within an appropriate range and the participation of the federated learning can considerably reduce the training time of the DDPG agents or improve the training effect under a limited equal training process.

Although mobile RIS exhibits significant gains, the cost is also obvious, including the hardware overhead of RIS, the power consumption of the robot, complex signaling,

etc. A critical thought is whether the gain is worth compared to the overhead. In the author's perspective, whether it is worthwhile or not depends on multiple factors in the specific scenario, such as the value of communication tasks, whether there are available robot resources, electricity prices, etc. The application cost of mobile RIS is relatively high, and some simpler solutions, such as RIS with guide rails, can be considered instead in budget-limited cases.

Chapter 5

STAR-RISs: A Coupled Phase-Shift Model Based Beamformer

Although STAR-RISs have the advantages of coverage, the effective design of the RC and TC of STAR-RISs has become a new challenge. Firstly, the STAR-RIS requires joint transmission and reflection beamforming, which is exceedingly more complex than reflection-only beamforming. What aggravates the situation further is that the STAR-RISs cannot independently adjust the TCs and RCs in practice, since the electric and magnetic impedances are unlikely to leave arbitrary values, but they depend on the electromagnetic properties of the STAR elements [46]. Furthermore, the coupling of the TCs and RCs requires a hybrid continuous and discrete control scheme for the phase-shift design. Given the above-mentioned adversities, it is a challenge to jointly solve the transmission and reflection beamforming problem for STAR-RISs, especially considering that the existing convex optimisation and machine learning solutions basically only support either continuous or discrete control. Although several hybrid algorithms have been proposed in the field of computer science [158–160], they are designed for minuscule action

dimensions. For example, only four discrete actions were assumed in [159] since gaming controllers usually have four buttons. However, the possible number of actions can be a^N for the STAR-RIS scenario, where a represents the possible number of actions for a single STAR element and N represents the total number of elements employed. Based on the current assumptions, prototypes of (STAR) RISs are likely to have a massive number of elements, which implies that the action dimension of STAR-RISs substantially exceeds the design in existing algorithms. Since there are lacking suitable hybrid algorithms for the optimisation of STAR-RISs, two hybrid reinforcement learning (RL) algorithms are proposed for joint active and passive beamforming design for the BS and the STAR-RIS.

5.1 System Model

5.1.1 Model of STAR-RISs

An energy splitting model is employed for supporting simultaneous transmission and reflection [30], where the STAR-RIS is capable of splitting the incident signal into the transmitted and reflected signals, partitioning the space into the transmission and reflection zones. Mobile users can be served by the transmitted or reflected signal, respectively, depending on which region they happen to be roaming in. In order to perform joint beamforming to covering both the transmission and reflection sectors, the TC and RC of each STAR element have to be appreciatively integrated, which are denoted as $\beta_{\mathcal{R},n}e^{j\theta_{\mathcal{R},n}}$ and $\beta_{\mathcal{T},n}e^{j\theta_{\mathcal{T},n}}, n = 1, 2, ..., N$.

It is worth noting that for any STAR element, the TCs an RCs are determined by its resistance and reactance. Therefore, it is non-trivial to independently adjust the coefficients. For a given RC of $\beta_{\mathcal{R},n}e^{j\theta_{\mathcal{R},n}}$, according to the conservation of energy[1], we

[1] According to the research in [46], the energy splitting does not require additional complexity or time consumption but possible energy loss due to the imperfect feature of the electromagnetic material. This chapter assumes that there is no energy loss in the process of transmission and reflection. The energy loss has an impact on the phase shift relationship between transmission and reflection [46], which would make the beamforming design more complex.

have $\beta_{\mathcal{T},n} = \sqrt{1-\beta_{\mathcal{R},n}^2}$. Then, simplifying RC $\beta_{\mathcal{R},n}$ as β_n, the TC can be calculated as $\sqrt{1-\beta_n^2}e^{j\theta_{\mathcal{T},n}}$. As pointed out in [46], for STAR elements, the coupling between the TC's phase-shift $\theta_{\mathcal{T},n}$, the RC's phase-shift $\theta_{\mathcal{R},n}$ and the amplitude β_n follows a relationship as

$$\beta_n\sqrt{1-\beta_n^2}\,\cos(\theta_{\mathcal{R},n}-\theta_{\mathcal{T},n}) = 0. \tag{5.1}$$

Thus, for a STAR-RIS having N elements, the transmission and reflection matrices have a diagonal structure given by

$$\boldsymbol{\Theta}_{\mathcal{R}} = \text{diag}\left(\beta_1 e^{j\theta_{\mathcal{R},1}}, \beta_2 e^{j\theta_{\mathcal{R},2}}, \cdots, \beta_N e^{j\theta_{\mathcal{R},N}}\right), \tag{5.2}$$

$$\boldsymbol{\Theta}_{\mathcal{T}} = \text{diag}\left(\sqrt{1-\beta_1^2}e^{j\theta_{\mathcal{T},1}}, \cdots, \sqrt{1-\beta_N^2}e^{j\theta_{\mathcal{T},N}}\right). \tag{5.3}$$

5.1.2 System Description

A downlink scenario described in Fig. 5.1 is considered, where the AP is equipped with M antennas, and the STAR-RIS has N STAR elements. There are K randomly roaming users and each having a single antenna. Each of the N STAR elements has the amplitude response $\beta_n, n = 1, 2, ..., N$. The locations of the BS, RIS, and users are denoted by $(x_b, y_b, z_b)^T$, $(x_r, y_r, z_r)^T$, and $(x_k, y_k, z_k)^T$, respectively. The STAR-RIS naturally partitions the users into two groups according to their locations. The users between the BS and the STAR-RIS receive direct signals from the BS and reflected signals from the STAR-RIS. This fraction of the users in the reflective region of the STAR-RIS are denoted by $\mathcal{R}_u$. For simplicity, this set of users is termed as $\mathcal{R}$ users in the rest of the text. Correspondingly, the users served by direct BS signals and transmitted STAR-RIS signals is denoted by $\mathcal{T}_i$. The number of users obeys $K = U + I$, where U and I are the

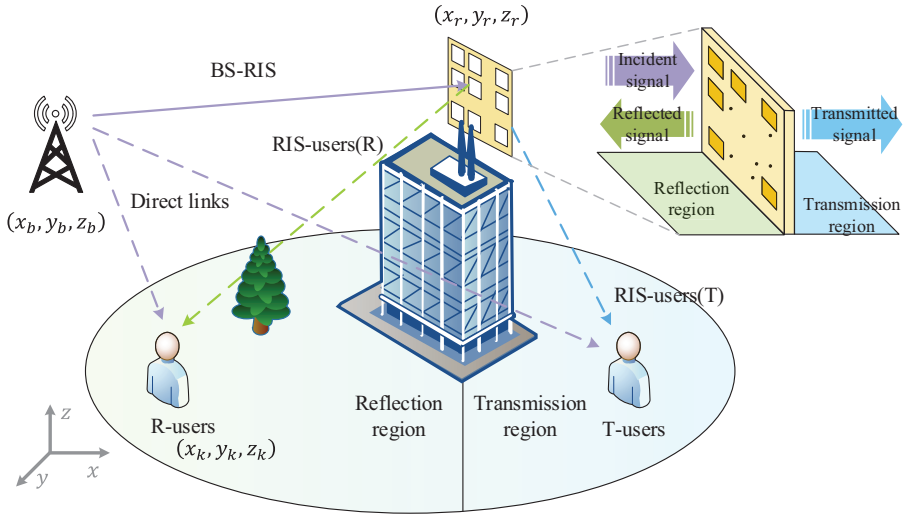


Figure 5.1: System model of STAR-RIS assisted wireless networks

numbers of $\mathcal{R}$ users and $\mathcal{T}$ users[2]. Although the mobile users are considered as moving, the users are assumed that do not cross the region. The reason for this assumption is that since a finite time frame is considered and the users are not able to cross the region due to the moving speed, which suggests that $\mathcal{R}$ users and not become $\mathcal{R}$ users, and vice versa.

5.1.3 Channel Model

In the STAR-RIS scenario, multiple channels have to be considered, including the BS to STAR-RIS channel $\boldsymbol{H}_{b,r} \in \mathbb{C}^{M\times N}$, the direct channel spanning the BS to $\mathcal{R}$ and $\mathcal{T}$ users $\boldsymbol{H}_{b,\mathcal{R}} \in \mathbb{C}^{M\times U}$, $\boldsymbol{H}_{b,\mathcal{T}} \in \mathbb{C}^{M\times I}$, and the channel impinging from the STAR-RIS to $\mathcal{R}$ and $\mathcal{T}$ users $\boldsymbol{H}_{r,\mathcal{R}} \in \mathbb{C}^{N\times U}$, $\boldsymbol{H}_{r,\mathcal{T}} \in \mathbb{C}^{N\times I}$. For each specific user $\mathcal{R}_u$ and $\mathcal{T}_i$, the direct channels, and the $\mathcal{T}$ & $\mathcal{R}$ channels can be denoted as $\boldsymbol{h}_{b,u} \in \mathbb{C}^{M\times 1}$, $\boldsymbol{h}_{b,i} \in \mathbb{C}^{M\times 1}$, $\boldsymbol{h}_{r,u} \in \mathbb{C}^{N\times 1}$ and $\boldsymbol{h}_{r,i} \in \mathbb{C}^{N\times 1}$, respectively.

All channels are assumed to follow the quasi-static block fading model, where the

[2]For simplicity of equations, the subscript $k = 1, 2, ..., K$ refers to any user, user $\mathcal{R}_u, u = 1, 2, ..., U$ or $\mathcal{T}_i, i = 1, 2, ..., I$ refers to a $\mathcal{R}$ user or $\mathcal{T}$ user.

fading coefficient remains constant in each time slot (TS) t. The channel $\boldsymbol{H}_{b,r}$ is assumed to have a line-of-sight (LoS) path and obeys to the Rician distribution, since the BS and STAR-RIS have a LoS component owing to their selected positions. Thus, upon considering the path loss and the small scale fading, the Rician channel can be formulated following the channel model given in Chapter 3. [3]

The pathloss $\mathcal{L}$ follows the urban propagation model presented in 3GPP specification TR 36.873 [161]. For the path with LoS, the path loss can be given by

$$\mathcal{L}_{\text{LoS}}\left(d, f_c\right) = 22.0 \log_{10} d + 28.0 + 20 \log_{10} f_c, \tag{5.4}$$

where d represents the 3D distance between the transmitter and the receiver, while f_c is the carrier frequency. For the NLoS propagation, the path loss is given by

$$\mathcal{L}_{\text{NLoS}} = \max[\mathcal{L}_{\text{LoS}}\left(d, f_c\right), \mathcal{L}_{\text{NLoS}}\left(d, f_c\right)], \tag{5.5}$$

$$\mathcal{L}_{\text{NLoS}}\left(d, f_c\right) = 36.7 \log 10d + 22.7 + 26 \log 10 f_c - 0.3(z_r - 1.5).$$

On the other hand, due to the random movements of users, LoS propagation may not necessarily be guaranteed, regardless whether the transmitting side is the BS or the STAR-RIS. Therefore, the channel to users, $\boldsymbol{H}_{b,\mathcal{R},t}, \boldsymbol{H}_{b,\mathcal{T},t}, \boldsymbol{H}_{r,\mathcal{R},t}$ and $\boldsymbol{H}_{r,\mathcal{T},t}$ are assumed to be NLoS channels and follow Rayleigh fading channel as described in Chapter 3.

[3] In order to focus on the hybrid beamforming problem of the STAR-RIS, the overhead model in Chapter 3 is not included in this Chapter.

5.1.4 Signal Model

The information sequence and the active beamforming vectors for user k at the BS are denoted by $s_{k,t}$ and $\mathbf{w}_{k,t}, \in \mathbb{C}^{M\times K}$. The signal transmitted at TS t can be expressed as

$$\mathbf{x}_{b,t} = \sum_{k=1}^{K} \mathbf{w}_{k,t} s_{k,t}. \tag{5.6}$$

Then, the incident signal at the STAR-RIS is given by

$$\mathbf{x}_{r,t} = \boldsymbol{H}_{b,r,t} \sum_{k=K}^{U} \mathbf{w}_{k,t} s_k + n_0, \tag{5.7}$$

where n_0 represents the Gaussian noise, and the received signal of user $\mathcal{R}^u$ is given by

$$y_{u,t} = [\boldsymbol{h}_{b,u,t} + \boldsymbol{h}_{r,u,t}\boldsymbol{\Theta}_{\mathcal{R},t}\boldsymbol{H}_{b,r,t}] \sum_{u=1}^{U} \mathbf{w}_{u,t} s_{u,t} + n_0. \tag{5.8}$$

Correspondingly, the received signal of user $\mathcal{T}^i$ can be represented in a similar form as

$$y_{i,t} = [\boldsymbol{h}_{b,i,t} + \boldsymbol{h}_{r,i,t}\boldsymbol{\Theta}_{\mathcal{T},t}\boldsymbol{H}_{b,r,t}] \sum_{i=1}^{I} \mathbf{w}_{i,t} s_{i,t} + n_0. \tag{5.9}$$

Given the received signal, the SINR of user $\mathcal{R}^u$ and $\mathcal{T}^i$ is given by

$$\gamma_{u,t} = \frac{\mid [\boldsymbol{h}_{b,u,t} + \boldsymbol{h}_{r,u,t}\boldsymbol{\Theta}_{\mathcal{R},t}\boldsymbol{H}_{b,r,t}]\mathbf{w}_{u,t} \mid^2}{\mid [\boldsymbol{h}_{b,u,t} + \boldsymbol{h}_{r,u,t}\boldsymbol{\Theta}_{\mathcal{R},t}\boldsymbol{H}_{b,r,t}] \sum\limits_{k\leq K, k\neq u} \mathbf{w}_{k,t} \mid^2 + \sigma^2}, \tag{5.10}$$

$$\gamma_{i,t} = \frac{\mid [\boldsymbol{h}_{b,i,t} + \boldsymbol{h}_{r,i,t}\boldsymbol{\Theta}_{\mathcal{T},t}\boldsymbol{H}_{b,r,t}]\mathbf{w}_{i,t} \mid^2}{\mid [\boldsymbol{h}_{b,i,t} + \boldsymbol{h}_{r,i,t}\boldsymbol{\Theta}_{\mathcal{T},t}\boldsymbol{H}_{b,r,t}] \sum\limits_{k\leq K, k\neq i} \mathbf{w}_{k,t} \mid^2 + \sigma^2}, \tag{5.11}$$

where σ^2 represents the noise power. Therefore, given a bandwidth B, the achievable

data rate of each user is given by

$$R_{k,t} = B \log_2 (1 + \gamma_{k,t}) . \tag{5.12}$$

5.1.5 Problem Formulation

This chapter aims for minimising the power consumption of the BS by jointly optimising the beamforming vector $\mathbf{w}_k$ for the BS and the TCs as well as RCs of the STAR-RIS. Again, the TCs and RCs of the STAR-RIS can be represented by the phase-shifts $\boldsymbol{\Theta}_{\mathcal{T}}, \boldsymbol{\Theta}_{\mathcal{R}}$, and amplitude coefficients $\boldsymbol{\beta}$. Therefore, the optimisation problem can be formulated as

$$\min_{\mathbf{w}, \boldsymbol{\Theta}_{\mathcal{T}}, \boldsymbol{\Theta}_{\mathcal{R}}, \boldsymbol{\beta}} \sum_{t=1}^{T} \sum_{k=1}^{K} \parallel \mathbf{w}_{k,t}^2 \parallel, \tag{5.13a}$$

$$\text{s.t.} \quad -\pi \leq \theta_{\mathcal{T},n,t} \leq \pi, \forall n, \forall t, \tag{5.13b}$$

$$-\pi \leq \theta_{\mathcal{R},n,t} \leq \pi, \forall n, \forall t, \tag{5.13c}$$

$$R_{k,t} \geq R_{\text{QoS}}, \forall k, \forall t, \tag{5.13d}$$

$$0 < \beta_{n,t} \leq 1, \forall n, \forall t, \tag{5.13e}$$

$$\beta_{n,t} \sqrt{1 - \beta_{n,t}^2} \cos(\theta_{\mathcal{R},n,t} - \theta_{\mathcal{T},n,t}) = 0, \tag{5.13f}$$

$$P_{b,t} \leq P_{\max}, \tag{5.13g}$$

where constraint (5.13b) and (5.13c) represent the legitimate range of the TC and RC phase-shifts. Constraint (5.13d) is a QoS constraint specifically the minimum data rate. Since (STAR) RISs are passive devices, their amplitude response is limited by the conservation of energy, as shown in (5.13e). Constraint (5.13f) characterizes the phase and amplitude relationship of the TCs and RCs. Finally, (5.13g) is the maximum power constraint for the BS.

The challenge of solving the formulated problem is not only owing to the joint consideration of TCs and RCs, but also due to the constraint (5.13e). Given the coupling

between $\theta_{\mathcal{T},n}$ and $\theta_{\mathcal{R},n}$, the STAR elements n cannot have independent arbitrary TCs and RCs. Furthermore, although the STAR-RIS aspire continuous phase-shift control for both transmission and reflection, one party of them can only have dualistic options. For example, assuming $\beta_n \neq 0$[4], once the TC is determined as $\theta_{\mathcal{T},n}$, the RC can only select the phase-shift from $\{\theta_{\mathcal{T},n}+\frac{\pi}{2}, \theta_{\mathcal{T},n}-\frac{\pi}{2}\}$. Therefore, the coupled phase-shift model of STAR-RISs requires hybrid continuous and discrete control for the transmission and reflection, which needs hybrid DRL algorithms for solving this challenge.

5.2 The Hybrid DDPG Algorithm

The DDPG algorithm was shown to constitute an efficient solution for continuous control problems [141]. For applying DRL approaches to solve the optimisation problem of (STAR) RIS , the transmission period has to obey a Markov decision process (MDP) [162]. In TS $t \in T$, by checking the CSI of the current channels, the agent determines the current state $\mathbf{s}_t \in \mathbf{S}$ and decides to carry out the action $\mathbf{a}_t \in \mathbf{A}$, where $\mathbf{S}$ and $\mathbf{A}$ represent the state space and action space. The action refers to a vector storing the active and passive beamforming coefficients at the BS and STAR-RIS.

Since passive and active beamforming are jointly considered, the MDP state for each TS t includes the CSI of the BS to STAR-RIS, BS to users (both $\mathcal{R}$ and $\mathcal{T}$ users), and the STAR-RIS to users channels, as shown in Fig. 4. Thus, $\mathbf{s}_t$ is given by

$$\mathbf{s}_t = \{\boldsymbol{H}_{b,r,t}, \boldsymbol{H}_{b,\mathcal{R},t}, \boldsymbol{H}_{b,\mathcal{T},t}, \boldsymbol{H}_{r,\mathcal{R},t}, \boldsymbol{H}_{r,\mathcal{T},t}\}. \tag{5.14}$$

Once action $\mathbf{a}_t$ is executed, the agent has to determine the reward r_t according to the data rate and power consumption of the transceiver, and then the state would be constrained to $\mathbf{s}_{t+1} \in \mathbf{S}$. Once the above steps are completed, $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ would be saved as a Markov transition in the replay buffer for the agent's training.

[4]If $\beta_n = 0$ or $\beta_n = 1$, the STAR-RIS operates in either the full transmission or full reflection mode, which is not a preferred mode.

5.2.1 DDPG Training

The objective of the DRL based agent training is to find the specific action $\mathbf{a}_t$ for each state $\mathbf{s}_t$, which maximizes the expected accumulated reward $\mathbb{E}[\sum_{i=t}^{T}\gamma r_{t+1}]$, where γ represents the discount factor $\gamma \in [0,1]$. For a DDPG agent, the Q value of action $\mathbf{a}_t$ can be quantified by the Bellman equation of [163]

$$Q^{\mu}(\mathbf{s}_t,\mathbf{a}_t) = \mathbb{E}\Big[r(\mathbf{s}_t,\mathbf{a}_t) + \gamma Q^{\mu}(\mathbf{s}_{t+1},\mathbf{a}_{t+1})\Big], \tag{5.15}$$

where μ represents the action policy function $\mathbf{S} \leftarrow \mathbf{A}$. The training of DRL agents aims for ascertaining the optimal action yielding the maximum Q value, as given by

$$Q^{*}(\mathbf{s}_t,\mathbf{a}_t) = \mathbb{E}\Big[r(\mathbf{s}_t,\mathbf{a}_t) + \max_{\mathbf{a}\in\mathbf{A}}\gamma Q^{*}(\mathbf{s}_{t+1},\mathbf{a}_{t+1})\Big]. \tag{5.16}$$

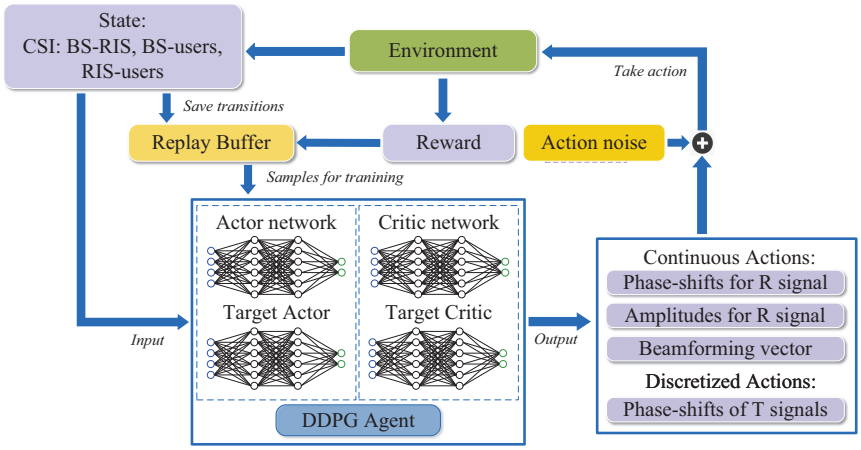
A parameterized actor function $\mu(\mathbf{s}|\boldsymbol{\omega}^{\mu})$ is considered and a function approximator associated with a batch of parameters is denoted by $\boldsymbol{\omega}^{Q}$. By sampling the aforementioned transition experiences in the memory of the replay buffer, the DRL agent can be trained by minimising the loss function

$$L(\boldsymbol{\omega}^{Q}) = \frac{1}{e}\sum_{e}[y_t - Q(\mathbf{s}_t,\mathbf{a}_t|\boldsymbol{\omega}_t^{Q})]^2, \tag{5.17}$$

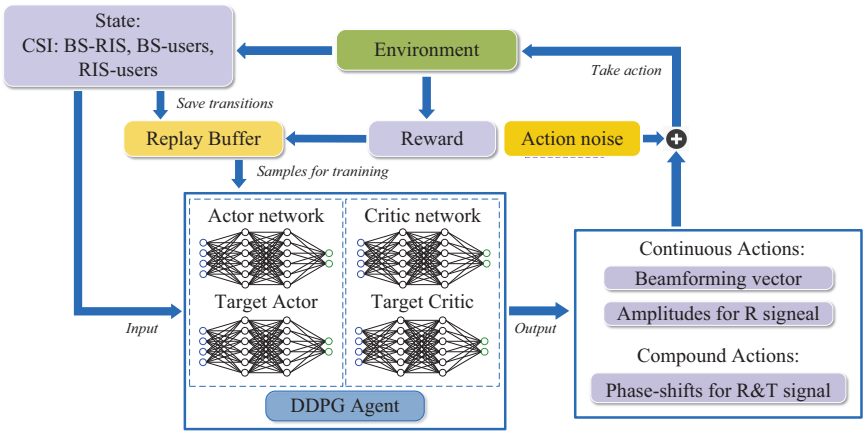
where e is the size of the sampled transitions. However, to avoid oscillations or divergence during the training process, y_t has to be provided by the target network, which has the same structure as the training network, but associated with a deferred parameter update. Upon denoting the parameters of the target networks by $\boldsymbol{\omega}^{Q'}$, y_t can be expressed as

$$y_t = r_t(\mathbf{s}_t,\mathbf{a}_t) + \gamma Q'[\mathbf{s}_t,\mu'(\mathbf{s}_t|\boldsymbol{\omega}_t^{\mu'})|\boldsymbol{\omega}_t^{Q'}]. \tag{5.18}$$

According to the principle of the original DDPG algorithm [141], the actor network



(a) Flow diagram of the conventional DDPG algorithm with output discretization.



(b) Flow diagram of the proposed hybrid DDPG algorithm.

Figure 5.2: Flow diagrams of the DDPG/hybrid DDPG algorithms.

is trained by the policy gradient calculated by the critic network as

$$\nabla_{\boldsymbol{\omega}^{\mu}} J = \frac{1}{e}\sum_{e} \nabla_{\mathbf{a}} Q(\mathbf{s}_e, \mathbf{a}_e|\boldsymbol{\omega}^{Q})|_{\mathbf{s}_e=\mathbf{s}_t, \mathbf{a}_e=\mu(\mathbf{s}_t)} \nabla_{\boldsymbol{\omega}^{\mu}} \mu(\mathbf{s}_e|\boldsymbol{\omega}^{\mu}|_{\mathbf{s}_e=\mathbf{s}_t}). \tag{5.19}$$

Algorithm 4 DDPG/Hybrid DDPG algorithm

Initialize the environment and the agent with the actor network $\boldsymbol{\omega}^{\mu}$, critic network $\boldsymbol{\omega}^{Q}$, target actor network $\boldsymbol{\omega}^{\mu'}$, target critic network $\boldsymbol{\omega}^{Q'}$
for each episode **do**
  Reinitialize the environment to $\mathbf{s}_{t=0}$
  for each step in $t_0 \leq t \leq T$ **do**
    Observe $\mathbf{s}_t$
    Choose $\mathbf{a}_t$ according to (5.21)
    if DDPG algorithm **then**
      Discretize a part of $\mathbf{a}_t$ into $\mathbf{a}_{t,d}$
    end if
    if Hybrid DDPG algorithm **then**
      Map $\mathbf{a}_t$ with $\boldsymbol{\Theta}_{\mathcal{R},t}$ and $\boldsymbol{\Theta}_{\mathcal{T},t}$
    end if
    Execute $\mathbf{a}_t$ in the environment
    Calculated the reward r_t and observe the next state $\mathbf{s}_{t+1}$
    Record $e\{\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}\}$ in memory buffer
    Random sample a batch of transection e from memory buffer
    Calculate target according to (5.18)
    Train critic network $Q(\mathbf{s}_t, \boldsymbol{\omega}_t^{Q})$ with a gradient descent step (5.17)
    Train actor network $\mu(\mathbf{s}_t, \boldsymbol{\omega}_t^{\mu})$ with (5.19)
    Update the target networks $\boldsymbol{\omega}_t^{\mu'} \leftarrow (1-\tau)\boldsymbol{\omega}_t^{\mu'} + \tau\boldsymbol{\omega}_t^{\mu}$, $\boldsymbol{\omega}_t^{Q'} \leftarrow (1-\tau)\boldsymbol{\omega}_t^{Q'} + \tau\boldsymbol{\omega}_t^{Q}$
    $\mathbf{s}_t \leftarrow \mathbf{s}_{t+1}$
  end for
end for

5.2.2 Continuous-discrete Actions and Hybrid DDPG

The actions designed for the DRL agent have to contain all optimised variables, resulting in the action space of $\mathbf{a}_t = \{\mathbf{w}, \boldsymbol{\Theta}_{\mathcal{T}}, \boldsymbol{\Theta}_{\mathcal{R}}, \boldsymbol{\beta}\}$. Among these optimisation variables, $\mathbf{w}$ and $\boldsymbol{\beta}$ can be handled by a continuous control scheme to achieve precise control. Therefore, this subsection focuses on the discussion of $\boldsymbol{\Theta}_{\mathcal{T}}$ and $\boldsymbol{\Theta}_{\mathcal{R}}$. As described above, due to the existence of the constraint (5.13f), the transmission and reflection phase-shifts cannot be adjusted independently by the STAR-RIS. For element n, assume $\beta_n \neq 0$ and $\theta_{\mathcal{R},n}$ is determined, in order to satisfy the constraint (5.13f). In this context, it is no hard to discover that $\theta_{\mathcal{T},n} = \theta_{\mathcal{R},n} \pm \frac{\pi}{2}$. Therefore, regardless of whether the phase-shift $\theta_{\mathcal{R},n}$ or $\theta_{\mathcal{T},n}$ is selected to be continuously controlled, the phase control of the other one is no longer a continuous control problem, but a binary selection problem. Therefore, (5.13a) requires continuous control for $\mathbf{w}, \boldsymbol{\Theta}_{\mathcal{R}}, \boldsymbol{\beta}$ and discrete control for $\boldsymbol{\Theta}_{\mathcal{T}}$,

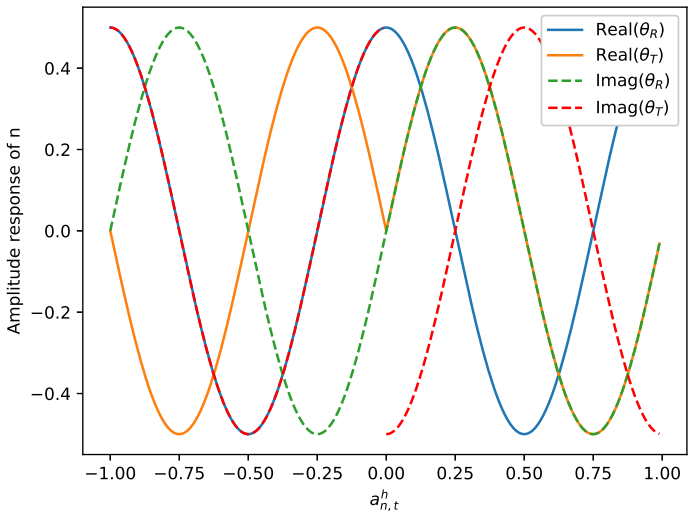


Figure 5.3: Amplitude response over normalized action output of hybrid DDPG algorithm for STAR-RIS ($\beta = 0.5$).

which lead to a hybrid action space associated with continuous and discrete actions. The continuous sub-space and the discrete sub-space are denoted by $\mathbf{a}^c = (a_1^c, a_2^c ... a_n^c ...)$ and $\mathbf{a}^d = (a_1^d, a_2^d ... a_n^d ...)$, respectively. For the convenience of presentation, there are assumptions $\mathbf{\Theta}_{\mathcal{R}} \subset a^c$ and $\mathbf{\Theta}_{\mathcal{T}} \subset a^d$ without loss of generality.

5.2.2.1 Action Space for the DDPG Algorithm

If the DDPG algorithm is applied for solving the optimisation problem associated with the hybrid action space, a direct and intuitive approach is to discretize a part of the continuous output of the DDPG algorithm as presented in Fig. 5.2(a) and **Algorithm 4**. Including the action noise, a classic action policy is given for the DDPG algorithm by

$$\mathbf{a}_t = \mu(\mathbf{s}_t|\boldsymbol{\omega}_t^{\mu}) + \mathcal{N}_{\mathrm{OU}}(0, \xi), \tag{5.20}$$

where $\mathcal{N}_{\mathrm{OU}}(0, \xi)$ represents the zero mean OU noise [164] that follows $\mathcal{N}_{\mathrm{OU}} \sim OU(0, \xi)$, and ξ is the volatility of the OU noise.

Thus, the continuous action space can be formulated as

$$\mathbf{a}_t^c = \{\mathbf{a}_t^{\mathbf{w}}, \mathbf{a}_t^{\boldsymbol{\Theta}_\mathcal{R}}, \mathbf{a}_t^{\boldsymbol{\beta}}\}. \tag{5.21}$$

Then, the normalized output of the actor network may be decoded into executable actions for the communication environment by following the actions of

$$\mathbf{w} \leftarrow \mathbf{a}_t^{\mathbf{w}}, \tag{5.22}$$

$$\boldsymbol{\Theta}_\mathcal{R} \leftarrow \mathbf{a}_t^{\boldsymbol{\Theta}_\mathcal{R}}, \tag{5.23}$$

$$\boldsymbol{\beta} \leftarrow \mathbf{a}_t^{\boldsymbol{\beta}}. \tag{5.24}$$

Since the mapping of $\mathbf{a}_t^c$ and the continuous actions are trivial, it is not necessary to discuss them in detail. On the other hand, for a specific STAR element n, $\theta_{\mathcal{T},n}$ can be obtained by the binary discretized $a_{n,t}^d$

$$\theta_{\mathcal{T},n,t} = \begin{cases} \theta_{\mathcal{R},n,t} + \frac{\pi}{2}, & a_{n,t}^d > 0, \\ \theta_{\mathcal{R},n,t} - \frac{\pi}{2}, & a_{n,t}^d \leq 0. \end{cases} \tag{5.25}$$

5.2.2.2 Action Space for Hybrid DDPG Algorithm

In order to deal with the high-dimensional continuous and discrete action components, specifically for the STAR-RIS scenario, a hybrid DDPG algorithm is proposed. By exploiting $\boldsymbol{\Theta}_\mathcal{R}$ and $\boldsymbol{\Theta}_\mathcal{T}$ having the same dimension for the STAR-RIS, the codebook between the actor outputs and the phase-shifts is designed to achieve hybrid control. The hybrid action policy for the STAR-RIS at time t can be expressed by

$$\begin{aligned} \mu(\mathbf{s}_t|\boldsymbol{\omega}_t^\mu) &= \mu^c(\mathbf{s}_t|\boldsymbol{\omega}_t^\mu)\mu^d(\mathbf{s}_t|\boldsymbol{\omega}_t^\mu) \\ &= \prod_{n=1}^{N} \mu^c(s_{n,t}|\boldsymbol{\omega}_t^\mu) \prod_{n=1}^{N} \mu^d(s_{n,t}|\boldsymbol{\omega}_t^\mu). \end{aligned} \tag{5.26}$$

Specifically, for the STAR element n, the values of $\theta_{\mathcal{R},n,t}$ and $\theta_{\mathcal{T},n,t}$ are compounded and given in a normalized output as

$$\mu(s_{n,t}|\boldsymbol{\omega}_t^{\mu}) = \mu^c(s_{n,t}|\boldsymbol{\omega}_t^{\mu})\mu^d(s_{n,t}|\boldsymbol{\omega}_t^{\mu}). \tag{5.27}$$

In the face of the action noise similar to (5.20), the hybrid action $a_{n,t}^h$ can be obtained from (5.27). Then, the normalized action $a_{n,t}^h$ has to be mapped to $\theta_{\mathcal{R},n,t}$ and $\theta_{\mathcal{R},n,t}$ as

$$\theta_{n,\mathcal{R},t} = 2\pi a_{n,t}^h, \tag{5.28}$$

$$\theta_{n,\mathcal{T},t} = \begin{cases} \theta_{n,\mathcal{R},t} + \frac{\pi}{2}, & a_{n,t}^h > 0, \\ \theta_{n,\mathcal{R},t} - \frac{\pi}{2}, & a_{n,t}^h \leq 0. \end{cases} \tag{5.29}$$

Since the normalized action $a_{n,t}^h$ is in the interval $[-1, 1]$, $\theta_{n,\mathcal{R},t}$ and $\theta_{\mathcal{T}}$ have to be mapped with $a_{n,t}^h$. Furthermore, $\theta_{n,\mathcal{R},t}$ vs $a_{n,t}^h$ is modelled by a periodic linear function and $\theta_{\mathcal{T}}$ vs $a_{n,t}^h$ by a piecewise linear function. The amplitude response $a_{n,t}^h$ of a s single STAR element is plotted in Fig. 5.3 for $\beta = 0.5$. Since STAR-RIS is generally equipped with a large number of STAR elements, in this case, the hybrid DDPG algorithm has a significantly smaller action dimension than the conventional DDPG algorithm, since $|\mathbf{a}_t^h| = \frac{|\mathbf{a}_t^c|+|\mathbf{a}_t^d|}{2}$.

Remark 5. *In some existing hybrid DRL schemes, the agent can only output a single action for the discrete action space, which is a disadvantage for problems associated with high-dimensional discrete action spaces of multi-antenna or multi-element-RIS scenarios. By contrast, the proposed hybrid scheme is eminently suitable for the high-dimensional hybrid action spaces, where the discrete action dimension is no larger than the continuous action dimension.*

5.2.3 Reward Function

The plain DDPG algorithm and the hybrid DDPG algorithm have identical reward functions. In order to ensure that the agent meets the users' QoS constraint, each satisfied $\mathcal{T}$ and $\mathcal{R}$ user contributes a distinguishable positive reward. Meanwhile, to minimise the power consumption, the sum power cost associated with $\mathbf{w}$ results in a negative reward. Thus, the reward function at TS t can be formulated as

$$r_t = \sum_{k=1, R_{k,t}>R_c}^{K} \dot{r} - \sum_{t=1}^{T}\sum_{k=1}^{K} \hat{r} \parallel \mathbf{w}_{k,t}^2 \parallel, \tag{5.30}$$

where the constant coefficient $\dot{r}$ represents the reward gleaned for satisfying the data rate requirement per user, and the coefficient $\hat{r}$ cost (negative reward) for power consumption. It is worth noting that, to guarantee the primary goal of the agent is satisfying the QoS requirement of each and every user rather than saving energy, the coefficients have to satisfy $\dot{r} > \sum_{k=1}^{K} \hat{r} P_{\max}$.

5.2.4 Neural Network Structure

In order to ensure accurate fitting, the structure and scale of the (target) actor network and the (target) critic network have to be selected appropriately. The actor networks include the input layer, batch normalization (BN) layer, and activation layer(s) with a 'relu' function in turn. Moreover, 'tanh' is assigned as the activation function of the output layer. To ensure a valid range for the input values of the output layer, another BN layer has to be employed above the output layer. The critic networks consist of an input layer, BN layer, concatenate layer, and activation layers. The size of the hidden layers has to be determined according to the state and action dimension, which is dependent both on the number of antennas M and the number of STAR elements N. It is also worth noting that due to the difference in output dimensions between the DDPG scheme and the hybrid DDPG scheme, the requireed size of the hidden layer is smaller in the

hybrid DDPG algorithm due to its mapping function, especially when a large number of STAR elements is considered.

5.3 Joint DDPG-DQN Algorithm

The hybrid DDPG algorithm aims for covering the action space $\mathbf{a}^c$ and $\mathbf{a}^d$ with the aid of a compound action mapping. By contrast, this section explores another promising option, namely that of employing two agents for covering the action spaces $\mathbf{a}^c$ and $\mathbf{a}^d$, respectively. Thus, a DDPG algorithm is harnessed for optimising $\mathbf{a}^c = \{\mathbf{a}_t^{\mathbf{w}}, \mathbf{a}_t^{\boldsymbol{\Theta}_{\mathcal{R}}}, \mathbf{a}_t^{\boldsymbol{\beta}}\}$, and a DQN algorithm is responsible for giving $\mathbf{a}^d = \mathbf{a}_t^{\boldsymbol{\Theta}_{\mathcal{T}}}$. DDPG and DQN algorithms can be regarded either as a joint agent, or as a pair of cooperative agents. For the convenience of the presentation, the DDPG agent and the DQN agent of the joint DDPG-DQN algorithm are regard as a pair of components, but they have to be installed on the same device in practice.

5.3.1 MDP for Joint DDPG-DQN

The DDPG and DQN algorithms are capable of optimising independent continuous and discrete problems [165]. Nonetheless, how to harness them in the interest of joint optimisation results is an open conundrum. If two agents adopt a parallel relationship at $\mathbf{s}_t$ to output $\mathbf{a}_t^c$ and $\mathbf{a}_t^d$, then according to the classic MDP model, the execution of $\mathbf{a}_t$ will result in reward r_t. However, both agents need corresponding rewards for estimating the action value of $\mathbf{a}_t^c$ and $\mathbf{a}_t^d$ independently. The dilemma is that in the scenario considered, r_t is the result of the combined action of $\mathbf{a}_t^c, \mathbf{a}_t^d$ and cannot be split into r_t^c, r_t^d. In other words, from the perspective of the wireless network, the user-side SINR of the STAR-RIS network is jointly determined by active beamforming, reflection beamforming, and transmission beamforming. As a consequence, any gain or loss cannot be solely and unambiguously attributed. If r_t is regarded as a common rewards for both two agents, the agents will authenticate the MDP transitions as $\{\mathbf{s}_t, \mathbf{a}_t^c, r_t, \mathbf{s}_{t+1}\}$ and $\{\mathbf{s}_t, \mathbf{a}_t^d, r_t, \mathbf{s}_{t+1}\}$.

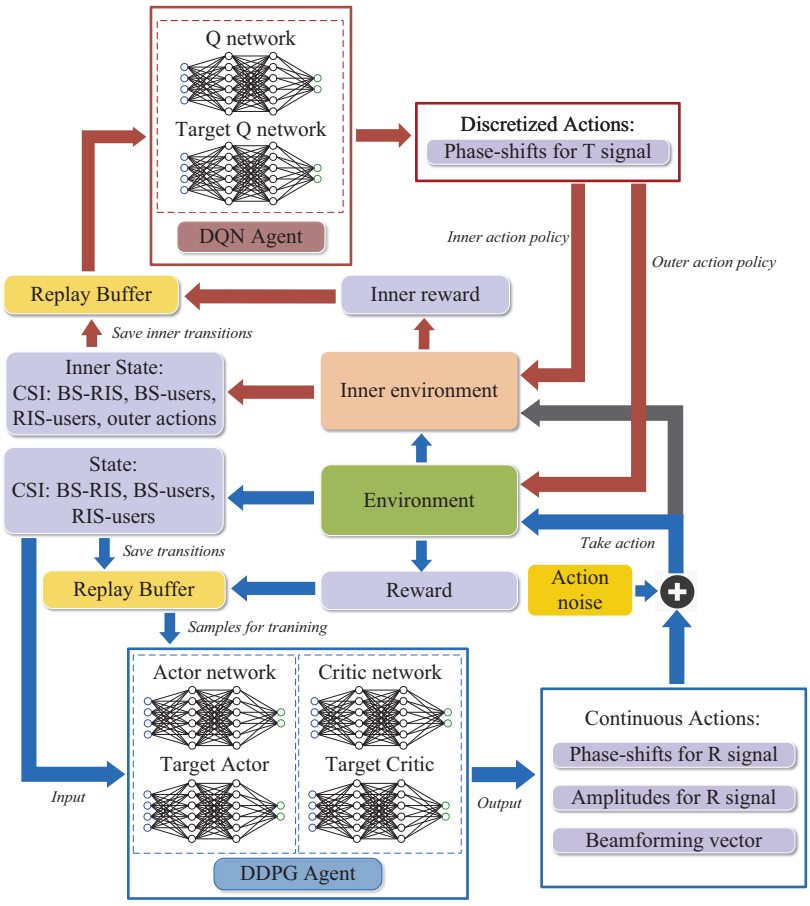


Figure 5.4: Flow diagram of the proposed DDPG-DQN algorithm.

Unfortunately, these transitions are not correct, as in fact the correct transitions have to be formulated as $\{\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}\}$ or $\{\mathbf{s}_t, \mathbf{a}_t^c, r_t^c, \mathbf{s}_{t+1}^c\}, \{\mathbf{s}_t, \mathbf{a}_t^d, r_t^d, \mathbf{s}_{t+1}^d\}$. In order to resolve this dilemma, inspired by [162], an inner environment is artificially added to formulate the MDPs for a pair of agents.

As shown in Fig. 5.4, an extra environment is derived from the original environment, and the environments refer to the 'inner environment' and 'outer environment' to distinguish them. Thus, the DDPG agent interacts with the outer environment and the DQN agent interacts with the inner environment. The outer environment represents the STAR-RIS assisted wireless network in the reality, the inner environment is only a fictitious environment that contains the knowledge of the outer environment for the agent's training. From the perspective of the DQN agent, the inner environment can

be regarded as a collection of the outer environment and the DDPG agent. Based on this framework, for the DQN agent and the inner environment, the MDP can be formulated as $\{\mathbf{s}_{i,t}, \mathbf{a}_{i,t}^d, r_{i,t}, \mathbf{s}_{i,t+1}\}$ and for the outer environment as $\{\mathbf{s}_{o,t}, \mathbf{a}_{o,t}^c, r_{o,t}, \mathbf{s}_{o,t+1}\}$. The remainder of this section discusses the details of the DQN and DDPG agents, respectively.

Algorithm 5 Joint DDPG-DQN algorithm

Initialize the outer environment and the DDPG agent with the actor network $\boldsymbol{\omega}^{\mu}$, critic network $\boldsymbol{\omega}^{Q}$, target actor network $\boldsymbol{\omega}^{\mu'}$, target critic network $\boldsymbol{\omega}^{Q'}$
Initialize the DQN agent with the deep Q network $\boldsymbol{\omega}^{\mathcal{Q}}$ and target Q network $\boldsymbol{\omega}^{\mathcal{Q}'}$
for each episode **do**
  Reinitialize the outer environment to $\mathbf{s}_{t=0}$
  for each step in $t_0 \leq t \leq T$ **do**
    Observing $\mathbf{s}_{o,t}$ and DDPG agent choose $\mathbf{a}_{o,t}^c$ according to (5.21)
    DQN agent choose inner action $\mathbf{a}_{i,t}^d$ with (5.35)
    Execute action $\mathbf{a}_{i,t}^d$ in the inner environment
    Calculated the reward $r_{i,t}$ and observe the next state $\mathbf{s}_{i,t+1}$
    Record $e\{\mathbf{s}_{i,t}, \mathbf{a}_{i,t}^d, r_{i,t}, \mathbf{s}_{i,t+1}\}$
    Sample random transitions of e from DQN memory
    Train $\omega^{\mathcal{Q}}$ with (5.33)
    DQN agent choose outer action $\mathbf{a}_{o,t}^d$ with (5.36)
    Execute $\mathbf{a}_{o,t}^c$ and $\mathbf{a}_{o,t}^d$ in the outer environment
    Calculated the reward $r_{o,t}$ and observe the next state $\mathbf{s}_{r,t+1}$
    Record $e\{\mathbf{s}_{o,t}, \mathbf{a}_{o,t}, r_{o,t}, \mathbf{s}_{o,t+1}\}$ in DDPG memory buffer
    Random sample a batch of transection e from memory buffer
    Calculate target according to (5.18)
    Train critic network $Q(\mathbf{s}_t, \boldsymbol{\omega}_t^Q)$ with a gradient descent step (5.17)
    Train actor network $\mu(\mathbf{s}_t, \boldsymbol{\omega}_t^{\mu})$ with (5.19)
    Update the target networks: $\boldsymbol{\omega}_t^{\mu'} \leftarrow (1-\tau)\boldsymbol{\omega}_t^{\mu'} + \tau\boldsymbol{\omega}_t^{\mu}$, $\boldsymbol{\omega}_t^{Q'} \leftarrow (1-\tau)\boldsymbol{\omega}_t^{Q'} + \tau\boldsymbol{\omega}_t^{Q}$, $\boldsymbol{\omega}_t^{\mathcal{Q}'} \leftarrow (1-\tau)\boldsymbol{\omega}_t^{\mathcal{Q}'} + \tau\boldsymbol{\omega}_t^{\mathcal{Q}}$
    $\mathbf{s}_{o,t} \leftarrow \mathbf{s}_{o,t+1}$
  end for
end for

5.3.2 Inner Environment and the DQN agent

5.3.2.1 DQN Training

As a value-based RL algorithm, the DQN algorithm identifies the action values as (5.15), since the DQN agent also aims for the maximum long-term reward. In the inner envi-

ronment, the Q function is given by

$$Q(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}^d) = \mathbb{E}\Big[r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}^d) + \gamma Q(\mathbf{s}_{i,t+1}, \mathbf{a}_{i,t+1})\Big]. \tag{5.31}$$

In the training process, the Q value for the actions has to be updated in each step by following

$$\begin{aligned} Q_{t+1}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}^d) \leftarrow (1-\alpha) Q_t(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}^d) \\ + \alpha[r_{i,t} + \gamma \max Q_t(\mathbf{s}_{i,t+1}, \mathbf{a}_{i,t+1}^d)], \end{aligned} \tag{5.32}$$

where α represents the learning rate ($0 < \alpha \leq 1$). In the DQN algorithm, the Q function is approximated by a DNN having the parameter vector $\boldsymbol{\omega}^{\mathcal{Q}}$, and $Q(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}^d) \approx Q(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}^d | \omega^{\mathcal{Q}})$. In order to accurately fit the Q function, the DNN has to be appreciatively trained. Similar to the DDPG algorithm, the transitions of the DQN agent are stored in the DQN memory buffer and the memory replay technique is adopted for training the DNN by the following loss function

$$\begin{aligned} L(\omega^{\mathcal{Q}}) = \frac{1}{e} \sum_{e} \Bigg[r_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}^d) + \\ \gamma \max_{\mathbf{a}_{i,t+1} \in \mathbf{A}^d} Q'(\mathbf{s}_{i,t+1}, \mathbf{a}_{i,t+1} | \omega_t^{\mathcal{Q}'}) - Q(\mathbf{s}_{i,t}, \mathbf{a}_{i,t} | \omega_t^{\mathcal{Q}}) \Bigg]^2. \end{aligned} \tag{5.33}$$

5.3.2.2 Inner State and Action Space

As designed above, the DQN agent regards the DDPG agent as a part of the inner environment. Therefore, the state of the inner environment consists of two parts, including the state inherited from the outer environment and the actions $\mathbf{a}_t^c$ of the DDPG agent. Thus the inner state can be formulated as

$$\mathbf{s}_{i,t} = \{\boldsymbol{H}_{b,r,t}, \boldsymbol{H}_{b,\mathcal{R},t}, \boldsymbol{H}_{b,\mathcal{T},t}, \boldsymbol{H}_{r,\mathcal{R},t}, \boldsymbol{H}_{r,\mathcal{T},t}, \mathbf{a}_{o,t}^c\}. \tag{5.34}$$

The action space of the DQN agent is $\mathbf{a}_{i,t}^{d} = \boldsymbol{\Theta}_{\mathcal{T}}$, since the DQN agent is only responsible for the discrete actions. The state space and action indicate a pair of main facts. Primarily, the decision of the DDPG agent has to be known to the DQN agent, when it is interacting with the inner environment. Additionally, it also suggests that the DQN agent has to 'interrupt' the interactions of the DDPG agent. Specifically, as shown in **Algorithm 5**, the timing of activating the inner environment and the DQN agent is after the action selection of the DDPG agent, but before the outer environment's execution of the action, since the transition of the outer environment relies on $\mathbf{a}_{o,t} = \{\mathbf{a}_{o,t}^{c}, \mathbf{a}_{o,t}^{d}\}$.

5.3.2.3 Action Policy

For conventional DQN agents, the $\epsilon-$greedy action policy constitutes an efficient technique of carefully balancing exploration and exploitation. To elaborate, the $\epsilon-$greedy policy authorizes the DQN agent to choose a random action with the probability of ϵ, and the optimal action with a probability of $1-\epsilon$. However, it is worth noting that in the joint DDPG-DQN algorithm, the DQN agent has to output a pair of actions one for the inner and one for the outer environment. Action $\mathbf{a}_{i,t}^{d}$ for the inner environment is adopted to form the inner transitions for the training process of the DQN agent. On the other hand, the action $\mathbf{a}_{o,t}^{d}$ produced by the DQN for the outer environment has to assist the DDPG agent. Therefore, once the $\epsilon-$greedy action policy is adopted in the outer environment, it may impose interference on $r_{o,t}$ and lead to an estimation error on the Q value of $\mathbf{a}_{o,t}^{c}$. Thus, the action policy of the DQN agent can be formulated as

$$\mathbf{a}_{i,t}^{d} = \begin{cases} \text{random action}, & \epsilon, \\ \underset{\mathbf{a}_{i,t} \in \mathbf{A}^{d}}{\arg\max}\, Q(\mathbf{s}_{i,t+1}, \mathbf{a}_{i,t} | \boldsymbol{\omega}_{t}^{Q}), & 1-\epsilon, \end{cases} \tag{5.35}$$

$$\mathbf{a}_{o,t}^{d} = \underset{\mathbf{a}_{i,t} \in \mathbf{A}^{d}}{\arg\max}\, Q(\mathbf{s}_{o,t}, \mathbf{a}_{o,t} | \boldsymbol{\omega}_{t}^{Q}). \tag{5.36}$$

Remark 6. *In the joint DDPG-DQN algorithm, the DQN is trained in the inner environment. The DQN agent produces actions for the outer environment, which may be viewed as a service instead of training for the DQN. Therefore, the $\epsilon-$ greedy action policy is employed in the inner environment to train the DQN agent and the optimal action policy in the outer environment to obtain optimal actions.*

5.3.2.4 Reward Function

Since the reward function depends on the optimisation goal and on the constraints, similarly to (5.30), the reward function of the DQN agent is also given by

$$r_{i,t} = \sum_{k=1, R_{k,t} > R_c}^{K} \dot{r} - \sum_{t=1}^{T} \sum_{k=1}^{K} \hat{r} \parallel \mathbf{w}_{k,t}^{2} \parallel, \tag{5.37}$$

where the constant settings have to be consistent with the corresponding discussion of Sub-section 5.2.4.

5.3.3 Outer Environment and the DDPG agent

The principle and algorithm's flow of the outer environment and the DDPG agent are the same as in Section 5.2, thus the principle and training process are not repeated here but only the difference is highlighted. For the DDPG agent, state $\mathbf{s}_{o,t}$ is the same as (5.14), and the reward is $r_{o,t} = \sum_{k=1, R_{k,t} > R_c}^{K} \dot{r} - \sum_{t=1}^{T} \sum_{k=1}^{K} \hat{r} \parallel \mathbf{w}_{k}^{2} \parallel$, but the difference is that the action space is reduced to $\mathbf{a}_{o,t}^{c} = \{\mathbf{a}_{t}^{\mathbf{w}}, \mathbf{a}_{t}^{\boldsymbol{\Theta}_{\mathcal{R}}}, \mathbf{a}_{t}^{\boldsymbol{\beta}}\}$. Hence, for any TS t, the optimised value $\{\mathbf{w}_t, \boldsymbol{\Theta}_{\mathcal{R},t}, \boldsymbol{\beta}_t\}$ can be completely covered by $\mathbf{a}_{o,t}^{c}$ using the mapping approach described in (5.22) (5.23) and (5.24).

In the joint DDPG-DQN algorithm, since both agents are employed and each has its

own DNNs, their optimal DNN scales are likely to be smaller than these of the hybrid DDPG algorithm. For the DDPG agent in the joint algorithm, the same structure is adopted as for the DNNs of the hybrid DDPG algorithm, not only as its optimal experimental structure, but also to ensure the fairness of the performance comparison.

5.3.4 Discussions

Although facing an identical formulated problem, Solution I and Solution II revealed two disparate measures. The hybrid DDPG approach aimed for achieving hybrid control via carefully tailored mapping function without any additional DNN or other processes. The advantage of this scheme is that its complexity is not increased. By contrast, since it has a lower output dimension than the DDPG agent, i.e. $|\mathbf{a}_n^h| < |\mathbf{a}_n^c| + |\mathbf{a}_n^d|$, it can employ a DNN with fewer trainable parameters, thereby reducing the complexity of the training process. The joint DDPG-DQN algorithm employs a pair of agents having different capabilities for jointly solving the problem having a hybrid action space. Since an extra DQN agent has to be harnessed, given the complexity of the DDPG and the DQN algorithm [139], the total complexity becomes significantly higher than that of the hybrid DDPG scheme.

The complexity of the DDPG algorithm depends on the specification of the employed DNN. Assuming that an actor network having I layers is employed, while each layer contains $\boldsymbol{\omega}_i^\mu$ nodes, the complexity of propagation is given by $\sum_{i=0}^{I} \boldsymbol{\omega}_i^\mu \boldsymbol{\omega}_{i+1}^\mu$. Given the number of nodes in the actor network as $\boldsymbol{\omega}_{\mathrm{b}}^\mu$ for BN layers, $\boldsymbol{\omega}_{\mathrm{r}}^\mu$ for 'relu' layers and $\boldsymbol{\omega}_{\mathrm{t}}^\mu$ for 'tanh' layers, according to [139], the required number of floating-point operations is given by $5\boldsymbol{\omega}_{\mathrm{b}}^\mu + \boldsymbol{\omega}_{\mathrm{r}}^\mu + 6\boldsymbol{\omega}_{\mathrm{t}}^\mu$. Applying the same theory for the critic networks, the complexity of a single prediction and training step can be formulated by $\mathcal{O}(\sum_{i=0}^{I} \boldsymbol{\omega}_i^\mu \boldsymbol{\omega}_{i+1}^\mu + \sum_{i=0}^{I} \boldsymbol{\omega}_i^Q \boldsymbol{\omega}_{i+1}^Q + 5\boldsymbol{\omega}_{\mathrm{b}}^\mu + \boldsymbol{\omega}_{\mathrm{r}}^\mu + 6\boldsymbol{\omega}_{\mathrm{t}}^\mu + 5\boldsymbol{\omega}_{\mathrm{b}}^Q + \boldsymbol{\omega}_{\mathrm{r}}^Q + 6\boldsymbol{\omega}_{\mathrm{t}}^Q)$. Since $\boldsymbol{\omega}_i \gg 5$, the computational complexity can be approximated by $\mathcal{O}(\sum_{i=0}^{I} \boldsymbol{\omega}_i^\mu \boldsymbol{\omega}_{i+1}^\mu + \sum_{i=0}^{I} \boldsymbol{\omega}_i^Q \boldsymbol{\omega}_{i+1}^Q)$. As mentioned, since the hybrid DDPG algorithm has smaller output and DNN scale, $\boldsymbol{\omega}_i^{\mu,h} < \boldsymbol{\omega}_i^\mu$. Thus, the

hybrid DDPG algorithm has lower complexity than the conventional DDPG algorithm.

The complexity of the joint DDPG-DQN algorithm is partially due to the DDPG agent and partially from the DQN agent. The complexity of the DQN agent is given by $\mathcal{O}(3\sum_{i=0}^{I}\boldsymbol{\omega}_i^{\mathcal{Q}}\boldsymbol{\omega}_{i+1}^{\mathcal{Q}})$ as suggested in [166]. The overall complexity of the joint DDPG-DQN algorithm is $\mathcal{O}(\sum_{i=0}^{I}\boldsymbol{\omega}_i^{\mu}\boldsymbol{\omega}_{i+1}^{\mu}+\sum_{i=0}^{I}\boldsymbol{\omega}_i^{Q}\boldsymbol{\omega}_{i+1}^{Q}+3\sum_{i=0}^{I}\boldsymbol{\omega}_i^{\mathcal{Q}}\boldsymbol{\omega}_{i+1}^{\mathcal{Q}})$. Therefore, the joint DDPG-DQN algorithm only has an increased complexity compared to the hybrid DDPG algorithm if the DQN agent has a significantly smaller $\boldsymbol{\omega}_i^{\mathcal{Q}}$.

5.4 Numerical Results and Analysis

The performance of the STAR-RIS is compared to that of the reflecting-only RIS and double spliced RIS. The reflecting-only RIS can only serve $\mathcal{R}$ users and fails to provide signal enhancements for the $\mathcal{T}$ users. As for the other benchmark, the double spliced RIS is formed by splicing a pair of RISs facing in opposite directions, where the total number of elements is set to N. This also ensures the fairness of comparisons. The double spliced RIS can also be equivalently regarded as a STAR-RIS employing the 'Mode Switching' scheme [30], but the proportion of transmission elements and reflection elements is fixed to 1. In terms of optimisation algorithms, the DDPG algorithm having a partial discrete action space is employed as the baseline. The performance of the hybrid DDPG algorithm and DDPG-DQN algorithm is compared.

As for the simulation parameters, the (STAR) RIS located 4km away from the BS, where the $\mathcal{T}$ users and $\mathcal{R}$ users are randomly distributed on both sides of the (STAR) RIS. The wireless channel model was introduced in Section 5.1. The simulated transmission process is 30 seconds, and the block fading envelope of the channel is assumed to vary once per second. For the intelligent agent, the 'Adam' optimiser is employed for both the hybrid DDPG and for the DDPG-DQN algorithms. The default learning rates for all agents, including the learning rate for the DQN, the actor/critic learning rates for the

Table 5-A: Default Parameters

Parameter	Description	Value	Parameter	Description	Value
M	antenna number	4	N	element number	12
f_c	carrier frequency	5GHz	B	bandwidth	1 MHz
K	number of users	4	$P_{\max}$	maximum power	29 dBm
K_{AR}, K_{RU}	Rician factors	3dB	σ	noise power density	-95.2 dBm/MHz
r	replay buffer size	10000	γ	discount factor	1
e_{DDPG}	batch size	32	e_{DQN}	batch size	32
τ_{DDPG}	target update rate	0.002	τ_{DQN}	target update rate	0.003

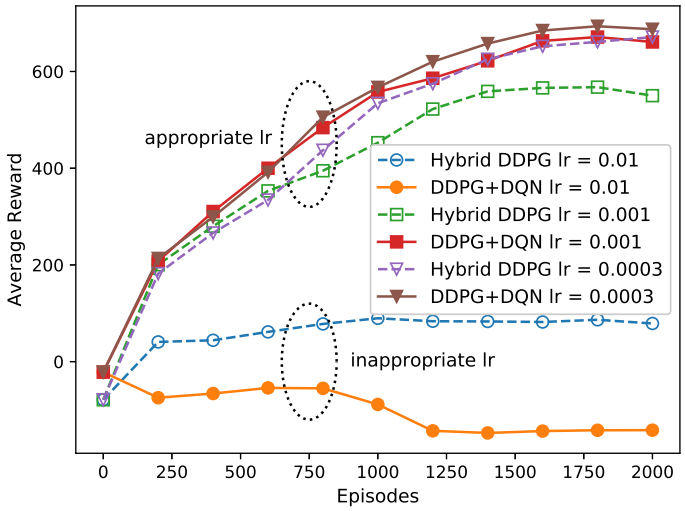


Figure 5.5: Reward of hybrid DDPG algorithm and DDPG-DQN algorithm with different learning rate

DDPG agent are set to be 3×10^{-4} [167]. For the DDPG agent, the actor network has a single activation layer, and the critic network has two activation layers. As for the DQN agent, the deep Q network has 1-2 activation layer(s). The activation function of the hidden layers is 'relu' in both algorithms. As mentioned, the size of the activation layers has to be determined by the complexity of the communication system. Empirically, in the majority of realizations, the actor and critic network are equipped with hidden layer(s) containing 256-512 neuron nodes, and the DQN agent employs hidden layer(s) with 200-300 neuron nodes. The default parameters used in the simulations are listed in Table 5-A.

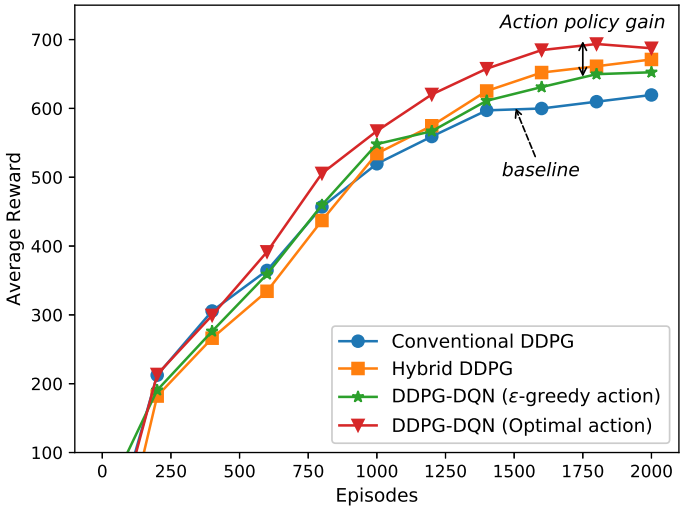


Figure 5.6: Performances comparison of different algorithms for STAR-RIS

Fig. 5.5 presents the reward obtained by the proposed algorithms having different learning rates. Based on the rewards exhibited by the different learning rates, two main conclusions can be obtained. On the one hand, both the hybrid DDPG and the DDPG-DQN schemes converge if the appropriate learning rate is selected. It has to be clarified that, in general, a higher learning rate leads to faster convergence. However, the curves $\mathrm{lr} = 0.001$ and $\mathrm{lr} = 0.0003$ of Fig. 5.5 exhibit similar convergence rates, since decaying action noise and exploration rate are invoked in the (hybrid) DDPG and DQN agents, which affects the reward obtained. On the flip side, the joint DDPG-DQN approach has an excellent capability of obtaining rewards. According to (5.30), the rewards obtained by the agents represent the overall performance of the wireless links, which translate into superior user satisfaction or reduced energy consumption.

The rewards obtained by the different algorithms are plotted in Fig. 5.6. The reward indicates the quality of the decisions, and the physical mean of the reward in this case is jointly determined by the number of satisfied users and the power spent on it. After a training process, the conventional DDPG algorithm used as the baseline has achieved an average reward of about 590, which is inferior to the proposed algorithms. The DDPG-DQN algorithm having optimal action output achieved a slight advantage of about $3-6\%$

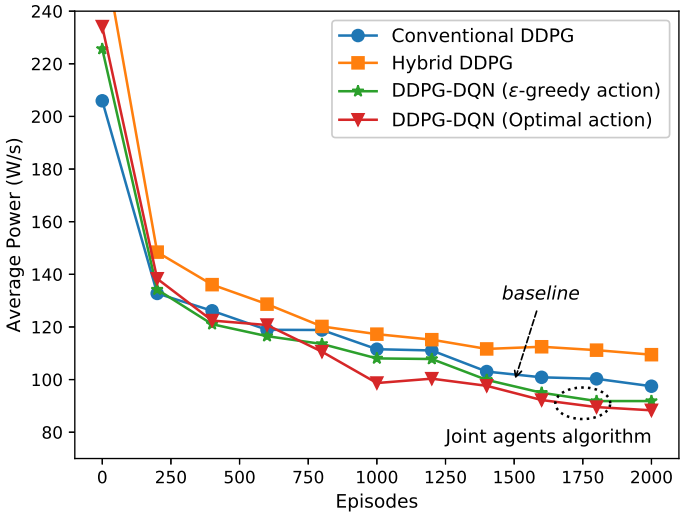


Figure 5.7: Power consumption of different algorithms for STAR-RIS

compared to that of the hybrid DDPG algorithm, but the price of this performance gain is that the DDQP-DQN algorithm has higher complexity, since two agents are employed. Although the hybrid DDPG scheme is slightly inferior in terms of optimality, it provides a low-complexity rapidly converging solution. Observing the two branches of the DDPG-DQN scheme, it can be observed that compared to the scheme in [162], the separated action policy of inner and outer environments has achieved superior performance, which supports the arguments in **Remark 8**.

Fig. 5.7 shows the performance of the proposed algorithm from the perspective of power minimisation. In the early stage of training, the agent executes fairly random actions, which can be considered as a chaotic system and its transmit power is relatively high. By contrast, for well-trained scenarios, the joint DDPG-DQN scheme has achieved superior power minimisation effects regardless of the action policy, which is in line with their trends in the reward analyses. The hybrid DDPG algorithm consumes more power than the DDPG-DQN algorithm, which is why it obtains less rewards. What is worth noting is that the baseline algorithm requires lower consumption than the hybrid DDPG algorithm. Recalling reward function (5.30) reveals that the solution provided by the DDPG agent does not satisfy (5.13d) even in some feasible conditions. Therefore, it can

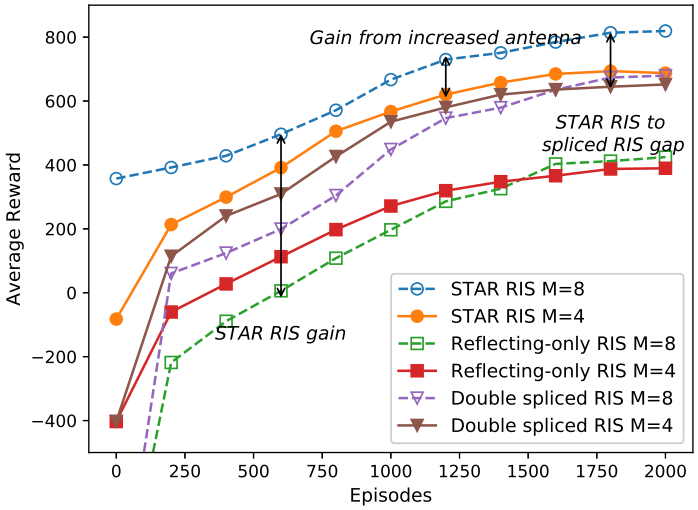


Figure 5.8: Performance comparison between STAR-RIS, reflecting-only RIS, and double spliced RIS

be found that the DDPG algorithm relying on discretized actions can result in completely different actions on both sides of the threshold, which may lead to some action errors and may result in violations of the QoS constraint.

The comparison between STAR-RIS, double spliced RIS (two reflecting-only RIS), and single reflecting-only RIS is presented in Fig. 5.8. STAR-RIS has achieved the best overall performance in all training stages, and the double spliced RIS has about 7% performance disadvantages over the STAR-RIS. This phenomenon confirms the conclusion that even if the transmitted and reflected signals of the STAR-RIS are constrained by each other, the STAR-RIS can provide further signal enhancement for users than double spliced RISs, since the STAR-RIS has higher multipath gain. The reflecting-only RIS has achieved a lower reward than the scheme having double-sides coverage, since the reflecting-only RIS is not capable of serving $\mathcal{T}$ users. In terms of the number of antennas, the simulation results prove that the proposed model and algorithm have general applicability. Furthermore, the multi-antenna gain of the STAR-RIS is higher than that of the reflecting-only RIS.

The supporting evidence for Fig. 5.8 is provided by Fig. 5.9, which reveals the

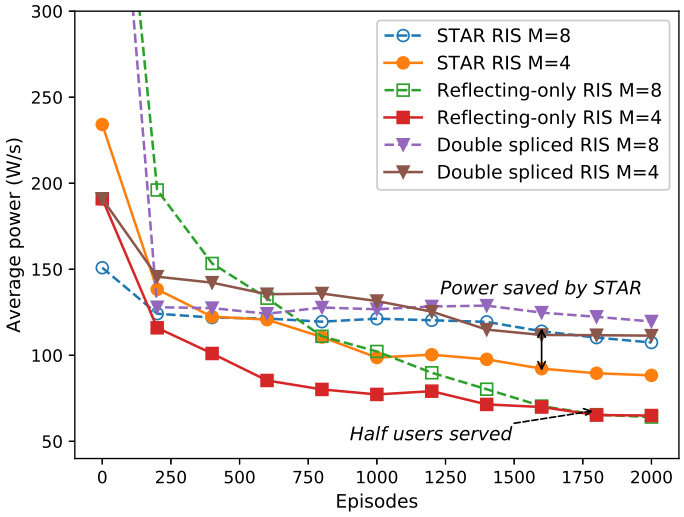


Figure 5.9: Power consumption of STAR-RIS, reflecting-only RIS, and double spliced RIS

transmission power consumption of different types of RISs. The energy consumption characteristics shown by the STAR-RIS and double spliced RIS follow their reward trend. It can be observed that since the STAR-RIS can have a higher number of STAR elements to serve users, it obtained higher gains than the double spliced RIS. Thus, the BS dissipates less transmission power to meet the data rate requirements of users. It is worth paying attention to the fact that the reflecting-only RIS has the lowest power consumption in Fig. 5.9. However, this does not suggest that it has an overall favorable energy efficiency, since it only serves about half the users. Based on similar logic, in the case of $M = 8$, the power consumption seen in Fig. 5.9 is higher than for $M = 4$, since the feasibility of the constraint (5.13d) has changed due to the increased number of antennas.

Fig. 5.10 plots the reward against the number of STAR or reflection elements of (STAR) RISs. Upon increasing the number of elements, the rewards have also been improved to varying degrees, which is in line with the theoretical expectations of the diversity gain. The result indicates that the performance of the proposed algorithm is not significantly affected by the action dimension (element number). Finally, it can be

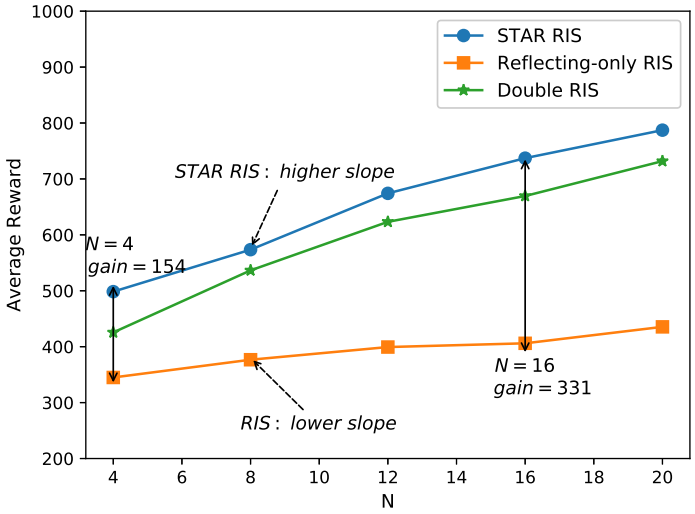


Figure 5.10: Reward against the number of STAR/reflecting elements

observed in Fig. 5.10 that the gains obtained by the double-sided coverage of RISs are more significant than these of the single-sided RISs.

5.5 Summary

A STAR-RIS assisted downlink network model was proposed in this chapter and the effects of coupled transmission and reflection phase-shift model were considered for the STAR-RIS. Although the STAR-RIS expanded the service range of the reflecting-only RIS, optimising the beamforming of the coupled transmission and reflection became a challenging problem, which required both continuous-valued and discrete-valued control. Thus, a hybrid DDPG algorithm and a joint DDPG-DQN algorithm are designed for jointly optimising the active and passive beamforming to minimise the energy consumption. The analysis and simulation results indicated that 1) STAR-RIS exhibited superiority over the double spliced and the reflecting-only RISs in terms of consuming transmission energy; 2) The proposed hybrid DDPG algorithm and the DDPG-DQN algorithm have outperformed the conventional DDPG algorithm; 3) The DDPG-DQN algorithm achieved superior performance compared to the hybrid DDPG algorithm albeit

at an increased complexity. As a new member of RIS, the electromagnetic model of STAR-RIS still needs further discussion. In practice, an imperfect model may result in energy loss or tighter bounds on phase shifts, leading to performance degradation. However, the proposed machine learning algorithm has a generality and is still likely to solve the beamforming problem for non-ideal STAR-RISs.

Chapter 6

Tile-based Beamforming for STAR-RIS

6.1 The Complexity for STAR-RIS Beamforming

Although the STAR-RIS assisted NOMA network has its superiorities, the control of STAR-RISs presents another main challenge due to its large number of STAR elements and bidirectional beams. The DRL algorithms are considered capable of making decisions according to specific scenarios and channel state information (CSI) to achieve long-term intelligent control of the RIS and adaptive resources allocation [168]. However, there are still several challenges to applying distributed DRL for optimising the beamforming of STAR-RISs, and the author commits to providing some insights and solutions for these challenges. The first challenge is the complexity of the passive beamforming problem for a single agent. Furthermore, from the perspective of the distributed framework, although a number of research and cases applying single-agent DRL to optimise wireless networks have been reviewed, there are still challenges in how to deploy distributed DRL agents in large cellular communication systems. Therefore, the following challenges are discussed and addressed in this chapter.

- ***Challenge 1:*** Since the reconfigurable elements do not have any power amplification function and the incident signal has experienced fading before arriving at the STAR-RIS, the transmitted or reflected signal from each single element has limited energy [169]. Hence, in order to ensure sufficient strength of the reflected signal, a large STAR-RIS with a massive number of elements is necessary. Once an enormous number of elements are employed in the STAR-RIS, the optimisation complexity of their TRCs will increase significantly, which leads to challenges for invoking the DRL algorithm. The resultant damages include but are not limited to the increase of the action dimension, the extension of the training time, the difficulty on hyperparameter setting, and finally the decline in optimality.

- ***Challenge 2:*** One practical problem is that ML algorithms require high-performance processing equipment and considerable training time [170]. The distributed deployment scheme is considered to have compatibility with cellular communication networks, which enables agents to contribute their training experience to condense the training load for each agent. Although some studies such as [126] demonstrated that distributed DRL schemes are feasible, they often require frequent exchange of parameters, resulting in a huge communication overhead and synchronization problems. Then, with a number of agents and various scales of deep neural network (DNN) models, how to manage the exchange of the models in a succinct manner is still a challenge.

- ***Challenge 3:*** Since the DRL algorithm is sensitive to the state dimension, the action dimension, and the scale of the DNN, it suggests that the agent needs to dynamically switch the running models according to the current user status in the cellular. Unfortunately, the number of users in a cellular tends to change frequently, if the agents train a new model based on the current state during each model switching, the system performance will be damaged and even outage due to the untrained model. Therefore, how to warm start or switch the model to avoid performance degradation becomes a challenge.

6.1.1 Contributions

A distributed DRL approach is developed to empower the intelligent STAR-RIS assisted NOMA network as a candidate to provide accessibility and coverage for IoT devices in the 6G wireless network. Distinguishing from the existing research contributions [9] and the previous work on STAR-RIS [38], this chapter focus on resolving ***Challenge 1-3*** for the STAR-RIS aided NOMA networks. The contributions and novelties of this chapter can be summarized as follows.

- A STAR-RIS assisted NOMA network model is conceived, where a sum rate maximisation problem is formulated for the proposed model. Two STAR-RIS operating protocols, ES and mode switching (MS) are investigated and their performances are compared.

- A novel tile-based passive beamforming approach is proposed for the STAR-RIS, where the STAR-RIS is intelligently partitioned into several tiles. A proximal policy optimisation (PPO) based algorithm is invoked to optimise the partitioning of STAR elements, as well as the tile-based transmission and reflection beamforming, which is the first research on the partition problem of the STAR-RIS.

- An AFFL paradigm is proposed for deploying artificial intelligent agents to cellular networks, where each BS plays a role as an agent. This novel paradigm utilizes the feature of both FL and TL, authorizing agents to freely join or quit federations according to the network status. The new member of a federation can inherit well-trained global models when joining in the federation, which does not require additional time for retraining agents to ensure uninterrupted high-quality communication service.

- The analysis and simulation results indicate that the tile-based beamforming approach is a stable and low-complexity solution for resolving ***Challenge 1***. It is also proved that the AFFL paradigm not only increases training speed and optimality, but also

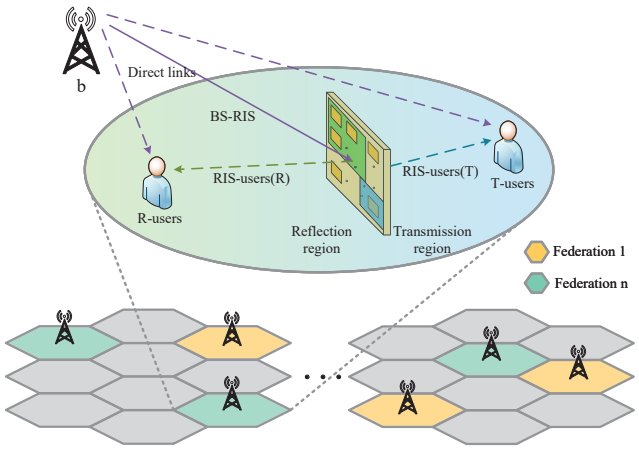


Figure 6.1: System model of STAR-RIS assisted NOMA networks

enables training-free model switching for the agent who changes its federation attribution, providing efficient solutions for resolving ***Challenge 2*** and ***Challenge 3***.

6.1.2 Organizations

The rest of this chapter is organized as follows: Section 6.2 introduces two implementations of the STAR-RIS, the system model of STAR-RIS assisted wireless networks, and the problem formulation for the tile-based passive beamforming of the STAR-RIS. Section 6.3 introduces the proposed solution from two major aspects, including the DRL algorithm for the single STAR-RIS control and the AFFL framework for federating DRL agents. Section 6.4 demonstrates the simulation results, analyzing the results in terms of the STAR-RIS partitioning and the distributed learning framework. Finally, Section 6.5 is the summary of this chapter.

6.2 System Model

6.2.1 System Description

A STAR-RISs assisted downlink NOMA network is considered, where each cellular comprises a BS $b, b \leq B$, K_b users, and a STAR-RIS $r, r \leq R$ equipped with N STAR

elements as illustrated in Fig. 6.1. The STAR-RIS is equipped with plenty of STAR elements, which consequent in a challenge for the TRC control [171], while the BS and users are assumed to have single antenna. The locations of the BS, RIS, and users are denoted by $(x_b, y_b, z_b)^T$, $(x_r, y_r, z_r)^T$, and $(x_{k_b}, y_{k_b}, z_{k_b})^T$, respectively. The STAR-RIS is assumed to be capable of achieving 360^o coverage by the transmitted (T) and reflected (R) signal to serve users around the STAR-RIS. Each user k_b receives T or R signal depending on its location, where users having $x_{k_b} < x_r$ are served by the R signal and users having $x_{k_b} > x_r$ are covered by the T signal. Without loss of generality, different BSs are assumed to be likely to be associated with a different number of users K_b, but STAR-RISs have the same value of N and M as the communication equipment is standardized. In each cell, the intelligent agent is deployed on a controller, which is connected with both the BS and the STAR-RIS by wired links to achieve a joint control. Since BSs and controllers have wired links, agents can transfer commands and parameters without spectrum overhead. Multiple BSs in the considered scenario do not necessarily have similar or adjacent locations. The BSs are observed and managed according to the number of active users they have, thus, the cells considered can be located in different locations, cities, or even countries. Hence, due to the separate spatial location, note that the scope of this chapter does not include suppression of inter-cell interference between adjacent cells, and average inter-cell interference is considered among the cellular instead.

6.2.2 Proposed Tile-based STAR-RIS Operation

Several physical models were proposed in [30] for splitting the incident signal at the STAR-RIS. Two practical operating protocols are considered for STAR-RISs, namely the MS mode and the ES mode. In order to efficiently resolve the passive beamforming problem for the STAR-RIS, a partitioning-based beamforming scheme is proposed, which partitions the STAR elements into M subsets (tiles), and elements in the same tile have a uniform TRC. This subsection will briefly introduce the MS protocol and the ES protocol and discuss their corresponding tile partition schemes, respectively.

6.2.2.1 STAR-RIS Partitioning for MS Protocol

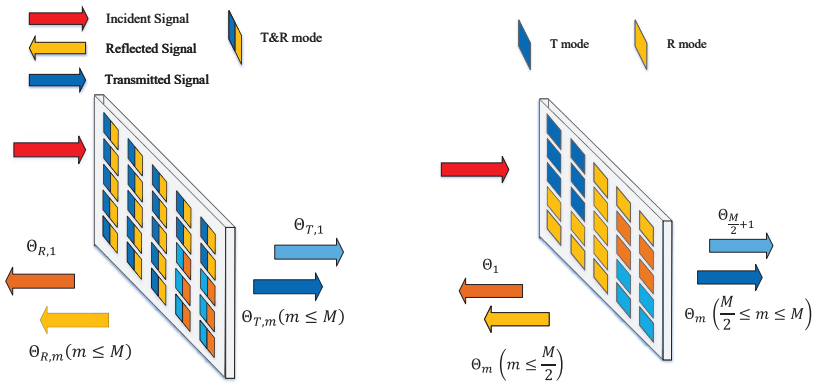
In the MS mode, the STAR-IRS consists of two groups of elements, where one group is operating in reflection function and the other one is responsible for the transmissions. Each MS element employed in the STAR-RIS can be either working in reflection mode or transmission mode as shown in Fig. 6.2(a), and the resulting 'on-off' working mode switching makes MS protocol easy to implement in practice. Assuming there are $N_{\mathcal{T}}$ elements working in the transmission mode and $N_{\mathcal{R}}$ elements in the reflection mode, the number of elements are $N_{\mathcal{T}} + N_{\mathcal{R}} = N$. Then, the TRC matrices can be given by

$$\mathbf{\Theta}_{\mathcal{T}} = \text{diag}\left(\beta_{\mathcal{T},1}e^{j\theta_{\mathcal{T},1}}, \beta_{\mathcal{T},2}e^{j\theta_{\mathcal{T},2}}, \cdots, \beta_{\mathcal{T},n}e^{j\theta_{\mathcal{T},N}}\right), \tag{6.1}$$

$$\mathbf{\Theta}_{\mathcal{R}} = \text{diag}\left(\beta_{\mathcal{R},1}e^{j\theta_{\mathcal{R},1}}, \beta_{\mathcal{R},2}e^{j\theta_{\mathcal{R},2}}, \cdots, \beta_{\mathcal{R},N}e^{j\theta_{\mathcal{R},N}}\right), \tag{6.2}$$

where $\beta_{\mathcal{T},n}, \beta_{\mathcal{R},n}$ represents the amplitude response and θ_n represents the phase-shift of element n. According to the principle of the MS elements and the law of conservation of energy, it can be obtained that $\beta_{\mathcal{T},n}, \beta_{\mathcal{R},n} \in \{0, 1\}$ and $\beta^2_{\mathcal{T},n} + \beta^2_{\mathcal{R},n} = 1$. In terms of the phase shift, the MS element is assumed to be able to provide arbitrary $\theta_{\mathcal{T},n}, \theta_{\mathcal{T},n} \in [0, 2\pi)$.

For the element partitioning problem in the case of MS, $N_{\mathcal{T}}$ elements have to be assigned into $M/2$ tiles for transmission and $N_{\mathcal{R}}$ elements into another $M/2$ tiles for reflection. Please note that the number of transmission tiles and reflection tiles can be set arbitrarily according to the need. In this chapter, they are assumed to be $M/2$ without loss of generality. The element partition matrix $\mathbf{\Lambda}^m_{\mathcal{T}} \in \mathbb{C}^{N\times N}, m \leq M/2$ and $\mathbf{\Lambda}^m_{\mathcal{R}} \in \mathbb{C}^{N\times N}, M/2 < m \leq M$ are defined for the reflection and the transmission, respectively. $\mathbf{\Lambda}^m_{\mathcal{T}}, \mathbf{\Lambda}^m_{\mathcal{R}}$ are diagonal matrixes, while $\mathbf{\Lambda}^m_{\mathcal{R}}[n, n] = 1$ indicates element n is in tile m, and $\mathbf{\Lambda}^m_{\mathcal{R}}[n, n] = 0$ represents element n is not in tile m. In each tile, the elements persist in



(a) Tile model of the ES based STAR-RIS.

(b) Tile model of the MS based STAR-RIS.

Figure 6.2: Tile model of the STAR-RIS.

the same TRC $\tilde{\boldsymbol{\Theta}}_{\mathcal{T}}^{m}$ or $\tilde{\boldsymbol{\Theta}}_{\mathcal{R}}^{m}$, and the overall TRC matrices are given by

$$\tilde{\boldsymbol{\Theta}}_{\mathcal{T}} = \sum_{m=1}^{M/2} e^{j\tilde{\theta}_{\mathcal{T}}^{m}} \boldsymbol{\Lambda}_{\mathcal{T}}^{m}, \tag{6.3}$$

$$\tilde{\boldsymbol{\Theta}}_{\mathcal{R}} = \sum_{m=M/2+1}^{M} e^{j\tilde{\theta}_{\mathcal{R}}^{m}} \boldsymbol{\Lambda}_{\mathcal{R}}^{m}. \tag{6.4}$$

6.2.2.2 STAR-RIS Partitioning for ES Protocol

The ES mode employs ES elements that split the incident signal into the transmitted and reflected signals. Since each ES element simultaneously provides transmitted and reflected signals, the TRCs of each STAR element can be denoted as $\beta_{\mathcal{R},n}e^{j\theta_{\mathcal{R},n}}$ and $\beta_{\mathcal{T},n}e^{j\theta_{\mathcal{T},n}}$, respectively. Therefore, the ES based STAR-RIS can have M multi-path to both T users and R users which indicates the ES model can obtain more degrees of freedom and multi-path gain compared with the MS mode.

However, the price of the favorable multi-path gain is that the ES elements are not able to configure TRCs arbitrarily and independently. The TRCs of the ES elements

are determined by their resistance and reactance. As pointed out in [50], the coupling between the TRCs for a given ES element n is given by

$$\beta_{\mathcal{T},n}\beta_{\mathcal{R},n}\cos(\theta_{\mathcal{R},n}-\theta_{\mathcal{T},n})=0. \tag{6.5}$$

which suggest that while the transmission signal and the reflection signal exists, their phase-shift always have to be orthogonal in each other, which follows $\Delta\theta_n = \theta_{\mathcal{R},n} - \theta_{\mathcal{T},n}, \Delta\theta_n \in \{-\pi/2, \pi/2\}$. Meanwhile, according to the conservation of energy, the relationship between $\beta_{\mathcal{T},n}$ and $\beta_{\mathcal{R},n}$ follows $\beta_{\mathcal{R},n} = \sqrt{1-\beta^2_{\mathcal{T},n}}$. Therefore, for the ES based STAR-RIS model, the TRCs matrices is given by

$$\mathbf{\Theta}_{\mathcal{T}} = \text{diag}\left(\beta_{\mathcal{T},1}e^{j\theta_{\mathcal{T},1}}, \beta_{\mathcal{T},2}e^{j\theta_{\mathcal{T},2}}, \cdots, \beta_{\mathcal{T},N}e^{j\theta_{\mathcal{T},N}}\right), \tag{6.6}$$

$$\mathbf{\Theta}_{\mathcal{R}} = \text{diag}\left(\sqrt{1-\beta^2_{\mathcal{T},1}}e^{j\theta_{\mathcal{T},1}+\Delta\theta_1}, \cdots, \sqrt{1-\beta^2_{\mathcal{T},N}}e^{j\theta_{\mathcal{T},N}+\Delta\theta_N}\right). \tag{6.7}$$

The partitioning problem in the ES mode is different from the MS mode, where the tiles in the MS mode are responsible for transmission or reflection. However, in the ES mode, each tile has both transmission and reflection functions. Thus, all N elements have to be assigned into M tiles, and the elements in each tile have the same TRCs. Similar to the case in MS mode, the element partition matrix in ES mode is denoted by $\mathbf{\Lambda}^m \in \mathbb{C}^{N\times N}, m \leq M$. In the tile having TRCs $\tilde{\mathbf{\Theta}}^m_{\mathcal{T}} = \tilde{\beta}^m_{\mathcal{T}}e^{j\tilde{\theta}^m_{\mathcal{T}}}$ and $\tilde{\mathbf{\Theta}}^m_{\mathcal{R}} = \sqrt{1-\tilde{\beta}^{m2}_{\mathcal{T}}}e^{j\tilde{\theta}^m_{\mathcal{T}}+\Delta\tilde{\theta}^m}$, and the overall TRC matrices are given by

$$\tilde{\mathbf{\Theta}}_{\mathcal{T}} = \sum_{m=1}^{M}\tilde{\mathbf{\Theta}}^m_{\mathcal{T}}\mathbf{\Lambda}^m, \tag{6.8}$$

$$\tilde{\mathbf{\Theta}}_{\mathcal{R}} = \sum_{m=1}^{M}\tilde{\mathbf{\Theta}}^m_{\mathcal{R}}\mathbf{\Lambda}^m. \tag{6.9}$$

Remark 7. *The disadvantage of the ES mode is that TRC cannot be adjusted independently. Assuming that users $k_{\mathcal{T}}$ and $k_{\mathcal{R}}$ are served by T and R signals, respectively, when ES-based STAR-RIS provides a perfect beam for user k that $h_{k_{\mathcal{T}}} = 1$, $h_{k_{\mathcal{T}}}$ is likely to be a random value. However, this defect is tolerated in NOMA networks, since the NOMA gain increases as the channel strength difference between users increases. Therefore, the ES mode has a higher affinity to NOMA networks than the MS mode.*

6.2.3 Channel Model

The quasi-static block fading channel is invoked to characterize the channels in the STAR-RIS scenario, where the fading blocks are assumed to be a finite set. The channel spanning BS b to the STAR-RIS r, which is denoted by $\boldsymbol{h}_{b,r} \in \mathbb{C}^{1\times N}$ is assumed to follow Rician fading since the LoS path is likely to be obtained. The direct channel between the BS and user k are denoted by $h_{b,k}$[1]. The channel linking the STAR-RIS to user k is denoted by $\boldsymbol{h}_{r,k} \in \mathbb{C}^{N\times 1}$. Due to the random movement of the users, the LoS path is not guaranteed at the user equipment. Therefore, the channel $h_{b,k}, \boldsymbol{h}_{r,k}$ are assumed to obey Rayleigh fading model. Meanwhile, the distance-based path loss for all channels follows the classic model that $\mathcal{L}(d) = \rho_0 \left(\frac{d}{d_0}\right)^{-\alpha}$ as suggested in [172], where ρ_0 represents the path loss at the reference distance d_0, d is the distance between the transceivers, and α is the path loss exponent. Following the aforementioned channel models, the BS-user channel $h_{b,k}$ and the RIS-user channel $\boldsymbol{h}_{r,k}$ is assumed to be Rayleigh fading channel, and the channel $\boldsymbol{h}_{b,r}$ is assumed to be Rician fading channel described in Chapter 3.

Remark 8. *Since two distant elements are likely to receive signals with different AOAs and require different AODs, it is not wise to assign these elements to the same tile. Therefore, the size of a tile and its element allocation need to take this factor into account.*

[1] Please note that since the user k is assumed associated with BS b, for concision, the superscript b for user k_b is omitted.

6.2.4 Communication Model

Without loss of generality, the T users and R users are assumed in one NOMA cluster. According to the principle of the superposition coding (SC), the transmitted signal at the BS b can be given by

$$x_b = \sum_{k=1}^{K} \sqrt{P_{b,k}} x_{b,k}, \tag{6.10}$$

where P_k represents the allocated power for user k. Please note that all the signals are time-varying during the transmission period, and the superscript t is omitted for a concise presentation. Since the users are assumed to receive the signal from both the direct link and the cascaded STAR-RIS link, the composite channel for user k can be given by

$$\hat{h}_{b,k} = \begin{cases} [h_{b,k} + \boldsymbol{h}_{r,k} \tilde{\boldsymbol{\Theta}}_{\mathcal{T},t} \boldsymbol{h}_{b,r}], \forall k \in \mathbb{K}_{\mathcal{T}}, \\ [h_{b,k} + \boldsymbol{h}_{r,k} \tilde{\boldsymbol{\Theta}}_{\mathcal{R},t} \boldsymbol{h}_{b,r}], \forall k \in \mathbb{K}_{\mathcal{R}}, \end{cases} \tag{6.11}$$

where $\mathbb{K}_{\mathcal{T}}$ and $\mathbb{K}_{\mathcal{R}}$ represent the set of T and R users, respectively. According to the principle of the successive interference cancelation (SIC), NOMA users in the same cluster can remove a part of the intra-cluster interference. To ensure a successful SIC for each user, the decoding order has to be correctly determined. As suggested in [173], the decoding order has to be determined by the channel gain order. Assuming there are two users k, j in the same NOMA cluster, the decoding order is given by $\Omega_j > \Omega_k, \forall |h_j|^2 > |h_k|^2$. Then, the received signal at the user k is given by

$$y_{b,k} = \sqrt{P_{b,k}} \hat{h}_{b,k} x_{b,k} + \sum_{j, \Omega_j > \Omega_k,} \sqrt{P_{b,j}} \hat{h}_{b,j} x_{b,j} + n_0, \tag{6.12}$$

where the first term is the desired signal, the second term is intra-cluster interference, and n_0 represents the sum of the average inter-cluster interference and Gaussian noise.

Given the received signal, the signal-to-interference-plus-noise ratio (SINR) of user k can be calculated by

$$\gamma_{b,k} = \frac{P_{b,k} \mid \hat{h}_{b,k} \mid^2}{\sum\limits_{j,\Omega_j > \Omega_k} P_{b,j} \mid \hat{h}_{b,k} \mid^2 + \sigma^2}, \tag{6.13}$$

where σ represents the inter-cluster interference and noise power. Therefore, the data rate of each user is given by

$$\mathcal{R}_{b,k} = W \log_2 \left(1 + \gamma_{b,k}\right). \tag{6.14}$$

where W represents the frequency bandwidth.

6.2.5 Problem Formulation

This chapter aims for the average throughput maximisation of the BSs by jointly optimising the power allocation P_k and the corresponding tile-based beamforming design $\tilde{\mathbf{\Theta}}_{\mathcal{T}}^{m}$, $\tilde{\mathbf{\Theta}}_{\mathcal{R}}^{m}$. The optimisation problem can be formulated as

$$\max_{\mathbf{\Lambda}_m, \tilde{\mathbf{\Theta}}_{\mathcal{T}}^{m}, \tilde{\mathbf{\Theta}}_{\mathcal{R}}^{m}, P_{b,k}} \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{t=1}^{T} \mathcal{R}_{b,k}, \tag{6.15a}$$

$$\text{s.t.} \quad 0 \leq \tilde{\theta}_{\mathcal{T},n} < 2\pi, \forall n, \tag{6.15b}$$

$$0 \leq \tilde{\theta}_{\mathcal{R},n} < 2\pi, \forall n, \tag{6.15c}$$

$$\mathcal{R}_{b,k} \geq \mathcal{R}_{\text{QoS}}, \forall k, \tag{6.15d}$$

$$\beta_{\mathcal{T},n}^2 + \beta_{\mathcal{R},n}^2 = 1, \forall n, \tag{6.15e}$$

$$\sum_{m=1}^{M} \mathbf{\Lambda}_m = \mathbf{I}, \tag{6.15f}$$

$$\sum_{k=1}^{K} P_{b,k} \leq P_{\max}, \tag{6.15g}$$

where constraint (6.15b) and (6.15c) represent the phase-shift range of the reflection coefficient and transmission coefficient. Constraint (6.15d) is a quality of service (QoS) con-

straint that stipulated the minimum data rate for users. Since STAR-RISs are regarded as passive devices that do not equip power amplifiers, according to the law of conservation of energy, the amplitude response constraint is given by (6.15e). Constraint (6.15f) indicates that each element has to be and only be assigned to a tile. Finally, (6.15g) is the maximum power constraint for the BS. Neglecting the complexity contributed by the optimisation of the power allocation for NOMA users and the TRCs of each tile, the choice of $\mathbf{\Lambda}_m$ is an integer choice problem for each. If each element selects its tile in turn, it will result in M^N choices. Considering the massive number of elements in the STAR-RIS, that will be a complex and challenging problem. Furthermore, since multiple BSs, random moving users, and a period of transmission time are considered in the scenario, the optimisation problem is a dynamic long-term optimisation problem. Therefore, a DRL-based distributed ML scheme is proposed to solve the formulated problem.

6.3 Proposed Distributed Learning Solution

This section describes the proposed algorithm in detail to solve the formulated optimisation problem, where a PPO algorithm is invoked for the STAR-RIS partitioning and corresponding beamforming, and an AFFL multi-agent distributed learning framework is proposed for the experience exchange among agents. This section will first briefly introduce the principles of the PPO algorithm and describe how to deploy the PPO agent in a single cell, including the design of states, actions, and reward functions. Then, the scope will be expanded to the multi-agent case. This section will explain why the AFFL framework is desired by communication systems, and it will also illustrate the principle of the AFFL framework.

6.3.1 PPO-based STAR-RIS Partition and Beamforming

The PPO algorithm [174] is invoked to provide solutions for the STAR-RIS in each cellular. After the advent of the PPO algorithm, it achieved impressive results in a variety of training environments, becoming one of the current mainstream DRL algorithms

[175]. As a novel PG algorithm, the PPO algorithm overcomes a common challenge in PG algorithms that they are sensitive to the stepsize of the policy update, and the oversized update can result in devastating performance degradation. By employing the clipped surrogate objective or the KL-penalized objective, the PPO algorithm successfully constrains the size of a policy update, which ensures its stable convergence and leads to a lower complexity compared to other conventional PG algorithms. Due to the aforementioned superiority, it is employed for the considered optimisation problem.

Following the basic principle of the DRL algorithm, the PPO agent has to be trained by interacting with the communication environment. During the transmission period $0 \leq t \leq T$, the agent have to recognize the current $\mathbf{s}_t \in \mathbf{S}$ and choose an action $\mathbf{a}_t \in \mathbf{A}$ according to the policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$. As a consequence of taking action $\mathbf{a}_t$, the agent receives a reward that indicates the quality of action $\mathbf{a}_t$ and the state will transfer to $\mathbf{s}_{t+1} \in \mathbf{S}$. Thus, a crucial factor in successfully using the PPO algorithm to resolve the STAR-RIS optimisation problem is to determine an appropriate state space $\mathbf{S}$, action space $\mathbf{A}$ and reward function to determine r_t.

6.3.1.1 State Space

State space has to contain the relevant information required for the STAR-RIS partitioning, TRC optimisation, and power allocation. The PPO agent has the identical state space whether ES or MS are considered, where the CSI at fading block t is included in the state $\mathbf{s}_t$ as given in (6.16).

$$\mathbf{s}_t = c_0 \times \{\boldsymbol{h}_{b,r,t}, \boldsymbol{h}_{r,k}, h_{b,k,t}\}. \tag{6.16}$$

where c_0 is a scaling coefficient. Scaling coefficient is needed since the path loss in the wireless network can be severe, the value of the elements in $\mathbf{s}_t$ is sometimes close to 0. It can cause the actor and critic network to receive a state vector with insignificant difference or even invalid data type, which is not conducive to the DNN fitting. Hence,

empirically an adaptive scaling coefficient c_0 or a normalization needs to be considered in contracting $\mathbf{s}_t$.

6.3.1.2 Action Space

The PPO algorithm is capable of handling both discrete and continuous action spaces, and PPO implementations with continuous action spaces is employed. Although the partitioning problem of tiles prefers discrete actions as the number of STAR elements has to be an integer, a continuous action space is invoked for obtaining accurate TRCs and power distribution. Once discrete actions are required, decision thresholds can be incorporated to discretize the sampled continuous actions succinctly. For the conventional element-based RIS beamforming problem, a straightforward action space design as shown in (6.17) is widely acknowledged

$$\mathbf{a}_t = \{\mathbf{a}_t^{\mathbf{\Theta}}, \mathbf{a}_t^p\}, \tag{6.17}$$

where $\mathbf{a}_t^{\mathbf{\Theta}}$ and $\mathbf{a}_t^p$ represent the output action of the agent for the phase shift and the power allocation, respectively. The reflection coefficient $\mathbf{\Theta}$ and the power P_k for each user can be obtained by applying simple linear mapping functions. However, this method leads to an enormous demission of the action space.

For the partitioning based approach, the element partition matrix $\mathbf{\Lambda}_{\mathcal{T}}^m, \mathbf{\Lambda}_{\mathcal{R}}^m$ cannot be directly given by the agent. Indeed, the agent can assign each element to the tile that it belongs to. Taking the MS mode as an example, the action space can be designed as

$$\mathbf{a}_t = \{\mathbf{a}_t^{\mathbf{\Lambda}_{n,\mathcal{R}}^M}, \mathbf{a}_t^{\mathbf{\Lambda}_{n,\mathcal{T}}^M}, \mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{T}}^M}, \mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{R}}^M}, \mathbf{a}_t^p\}, \tag{6.18}$$

where $\mathbf{a}_t^{\mathbf{\Lambda}_{n,\mathcal{R}}^M}, \mathbf{a}_t^{\mathbf{\Lambda}_{n,\mathcal{R}}^M}$ represents the element partition matrix and $\mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{T}}^M}, \mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{R}}^M}$ includes the information of the TRCs for each tile. Unfortunately, this scheme still has a large action dimension since $\mathbf{a}_t^{\mathbf{\Lambda}_{n,\mathcal{R}}^M}, \mathbf{a}_t^{\mathbf{\Lambda}_{n,\mathcal{R}}^M}$ have a dimension N in total, which contrary to the

original intention of reducing complexity and action space dimension. Moreover, since the adjacent element $[n, n]$ and $[n+1, n+1]$ in the diagonal matrixes $\mathbf{\Lambda}^{M_{\mathcal{R}}}, \mathbf{\Lambda}^{M_{\mathcal{T}}}$ is not guaranteed to be adjacent in the 2D STAR-RIS panel, there is a possibility that multiple scattered elements are assigned to one tile. Recalling the discussion in **Remark 8**, such an allocation is not in line with the principles of the propagation.

In order to achieve a low-complexity partition for the STAR-RIS, an indirect method is proposed to determine the tile. It is not necessary to directly seek the attribution of each element, but assign the agent to output the horizontal and vertical dividing boundary of the tile. For example, if a STAR-RIS plane needs to be divided into $M = M_x \times M_y$ tiles, where M_x and M_y represent the number of horizontal and vertical tiles, respectively, the agents outputs the $N_{m,x}$ and $N_{m,y}$. Thus $\mathbf{\Lambda}^m[n_x, n_y] = 1, N_{m,x} \leq n_x \leq N_{m+1,x}, N_{m,y} \leq n_y \leq N_{m+1,y}$ can be obtained. According to the above method, for the ES mode, the action spaces can be designed as

$$\mathbf{a}_t = \{\mathbf{a}_t^{\mathbf{n}_x^M}, \mathbf{a}_t^{\mathbf{n}_y^M}, \mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{R}}^M}, \mathbf{a}_t^{\Delta\tilde{\mathbf{\Theta}}^M}, \mathbf{a}_t^p\}, \tag{6.19}$$

where $\mathbf{a}_t^{\mathbf{n}_x^M}, \mathbf{a}_t^{\mathbf{n}_y^M}$ represents the vector contains the horizontal and vertical boundary for tiles, and $\mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{T}}^M}$ represents the transmission coefficient for tiles, and $\mathbf{a}_t^{\Delta\tilde{\mathbf{\Theta}}^M}$ contains the phase shift difference between the reflection coefficient and the transmission coefficient. After harvesting the above actions, $\mathbf{\Lambda}^M$, $\tilde{\mathbf{\Theta}}_{\mathcal{R}}^M$, and $\tilde{\mathbf{\Theta}}_{\mathcal{T}}^M$ can be calculated. Please note that for the ES mode, $\mathbf{\Lambda}_{\mathcal{R}}^M = \mathbf{\Lambda}_{\mathcal{T}}^M = \mathbf{\Lambda}^M$ Then, following (6.8), (6.9), and (6.19), TRC can be calculated for all elements.

Following the similar partitioning logic, the action space of the MS mode can be given by

$$\mathbf{a}_t = \{\mathbf{a}_t^{\mathbf{n}_x^M}, \mathbf{a}_t^{\mathbf{n}_y^M}, \mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{R}}^M}, \mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{T}}^M}, \mathbf{a}_t^p\}. \tag{6.20}$$

where $\mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{T}}^M}$ and $\mathbf{a}_t^{\tilde{\mathbf{\Theta}}_{\mathcal{R}}^M}$ represent the transmission coefficient and reflection coefficient for tiles. Pleas note that due to the different feature between the ES and MS, in MS mode,

it is necessary to distinguish the $\mathbf{\Lambda}_{\mathcal{T}}^{M}, \mathbf{\Lambda}_{\mathcal{R}}^{M}$, and it need independent outputs for the $\tilde{\mathbf{\Theta}}_{\mathcal{T}}^{M}$ and $\tilde{\mathbf{\Theta}}_{\mathcal{R}}^{M}$.

Remark 9. *The action dimension of element-based beamforming is* $3N+K$*, but only be about* $4M+K$ *in tile-based beamforming. Considering the practical case that* $N > 100 >> M$*, the action dimension of element-based beamforming will be huge, which requires a large scale DNNs and are likely to lead to increasing complexity, slow convergence, and inferior optimality.*

6.3.1.3 Reward Function

The reward function determines the optimisation orientation for the intelligent agent, which will have a major impact on the optimisation effect. Therefore, the reward function has to be consistent with the objective function. The reward function is defined as

$$r_t = \begin{cases} \mathcal{R}_t, & \text{QoS requirement satisfied,} \\ \frac{\mathcal{R}_t}{c_p}, & \text{QoS requirement not satisfied,} \end{cases} \tag{6.21}$$

where c_p is a penalty for violation of QoS requirements to enforce the agent to meet the constraint (6.15d).

6.3.1.4 Training Process

PPO algorithm genesis form policy gradient methods which aims to train stochastic policy $\pi_\omega(\boldsymbol{a}_t|\boldsymbol{s}_t)$ to obtain the optimal parametric probability distribution for actions against $\boldsymbol{s}_t$. A conventional gradient estimator is given by

$$L_t^{\mathrm{PG}}(\omega) = \mathbb{E}\left[\log \pi_\omega(\boldsymbol{a}_t|\boldsymbol{s}_t)\, \mathcal{A}_t^s\right], \tag{6.22}$$

where $\mathcal{A}_t^s$ is an advantage function at state s_t. Then, for a given length of time frame $0 < t < T$, tt can be expressed as

$$\mathcal{A}_t^s = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}, \tag{6.23}$$

$$\delta_t^s = r_t + \gamma V_\omega\left(\boldsymbol{s}_{t+1}\right) - V_\omega\left(\boldsymbol{s}_t\right), \tag{6.24}$$

where γ is the discount factor, and $V_\omega(s)$ represents the state-value function [176]. In order to solve the defect that the transition samples of the policy gradient algorithm cannot be reutilized for training, different from the conventional policy gradient algorithm, the PPO algorithm adopts a surrogate objective, which enables the agent to train the current policy π by using samples generated by the previous policy π_{old}, where the surrogate objective is given by

$$L^{\text{SO}} = \mathbb{E}\left[u_t(\omega)\mathcal{A}_t^s\right], \tag{6.25}$$

where

$$u_t\left(\omega\right) = \frac{\pi_\theta\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)}{\pi_{\theta_{\text{old}}}\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)}. \tag{6.26}$$

However, maximising L^{SO} without any constraint is likely to lead to an excessively large update on policy. Therefore, following the design in [174], a loss function that combines the policy surrogate and a value function error term is designed, which is given by

$$L_t\left(\omega\right) = \mathbb{E}\left[L_t^{\text{CLIP}}\left(\omega\right) - c_1 L_t^{\text{VF}}\left(\omega\right) \ + \ c_2 S\left[\pi_\theta\right]\left(\boldsymbol{s}_t\right)\right], \tag{6.27}$$

where $c1$ and $c2$ are coefficients, L_t^{VF} is the squared value function error term that

$$L_t^{\mathrm{VF}}(\omega) = \mathbb{E}_t\left[\left(V_\omega(\boldsymbol{s}_t) - V_{\pi_{\omega_{\mathrm{old}}}}(\boldsymbol{s}_t)\right)^2\right], \tag{6.28}$$

and $c_2 S[\pi_\theta](\boldsymbol{s}_t)$ an entropy bonus for exploration. The most distinction contribution of the PPO algorithm is embodied in the first term, which is a clipped surrogate objective proposed in the PPO2 algorithm. It can be given by

$$L_t^{\mathrm{CLIP}}(\omega) = \mathbb{E}\left[\min\left(u_t^s(\omega)\mathcal{A}_t^s, \mathrm{clip}\left(u_t^s(\omega), 1-\epsilon, 1+\epsilon\right)\mathcal{A}_t^s\right)\right]. \tag{6.29}$$

where ϵ is a hyperparameter to limit the update on the policy.

With the calculated loss L_t, the DNNs are updated following

$$\omega \leftarrow \omega + \alpha \nabla_\omega L_t(\omega), \tag{6.30}$$

where α represents the learning rate.

6.3.2 AFFL Model

The AFFL model invokes both FL and TL, establishing multiple federations according to the network status of each BS and authorizing the agent to freely shuttle between each federation, thereby exempting the time overhead required by retraining the models. As reviewed in the state of the art, several distributed frameworks were proposed for the PPO or other DRL algorithms. These successful implementations proved that incorporating DRL agents into distributed frameworks can effectively reduce training time, distribute computing load to hardware devices, and lay a solid theoretical foundation for distributed reinforcement learning. However, these schemes may encounter practical difficulties when using them to arm communication systems. Since in a communication system, thousands of BSs having the same technical specifications work in parallel, this feature becomes the technical prerequisite for applying distributed learning among BSs.

However, the existing distributed frameworks are not specifically designed for the communication network, which would meet practical issues once they are employed. Taking the DPPO [126] as an example, if the DPPO paradigm is employed to coordinate BSs, then each BS needs to play the role of worker, while the central server will play the role of the chief. Then in each training slot, the chief needs to wait until it receives the policy gradient from all workers before starting training. This scheme not only requires frequent data exchanges between BSs and the central server, but data synchronization and long latency will be foreseeable issues due to the possible significant physical distance and transmission delays between BSs and the central server. Therefore, the distributed framework for the communication system still needs further discussion (***Challenge 2***).

Another practical issue mentioned in ***Challenge 3*** is that a pair of trained DNN models cannot handle the dynamic circumstances in the wireless networks. Observing (19), (22), and (23), it can be confirmed that the dimensions of the state and action spaces (the scale of the input and the output layers of the DNN model) are jointly determined by multiple parameters, such as the number of antennas, STAR-elements, and active users served by the STAR-RIS. Although the numbers of antennas and STAR elements are generally constants, the number of connected users is likely to be a time-varying variable. Therefore, when the number of active users in the cell changes, the old policy may no longer be applicable due to changes in the state and action space. In this case, as mentioned in ***Challenge 3***, it needs additional time and exploration to training a pair of new models. During the training process, since naive models are employed, mismatched estimations are likely to be produced, resulting in degraded beamforming and a consequential decline in the SINR. Therefore, in order to provide a comprehensive distributed solution for the BSs and to resolve ***Challenge 2*** and ***Challenge 3***, the AFFL framework is proposed.

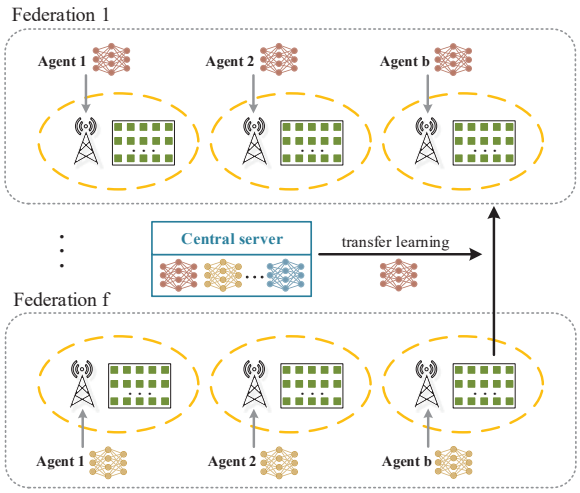


Figure 6.3: Structure of the AFFL Framework.

6.3.2.1 Generalized Framework

The generalized framework of the AFFL is shown in Fig. 6.3. BSs are allocated into different federations according to the number of active users associated with the BSs, which guarantees that the model $\omega_{b,t}^{\kappa}$ of the agents in federation κ have the same specifications and they can share the global model $\omega_{G,t}^{\kappa}$, where κ represents the federation. Since the BSs are likely to have diverse number of active users, thus, the agent in each cell needs different scales of input layer, hidden layer, and output layers, and it challenging to federate all agents with different scales of DNN models via a single federation. As shown in Fig. 6.3, agents have to form several federations, which are indexed by $1...f$, and the global model of each federation has to be uploaded and storied at the central server. The online training is applied for each agent, and the agents in the same federation use federated averaging to achieve parameter exchange. Since the BSs are linked by wire and fiber, this model does not incur huge spectrum overhead like applying federated learning among terminal devices. With the assistance of FL, the problem mentioned in ***Challenge 2*** is relieved, efficiently utilizing the extensive transition samples and hardware resources from different BSs to accelerate training.

To address ***Challenge 3***, TL is employed in the AFFL framework to enable a join and quit mechanism for federations. In classic FL frameworks, the federation is assumed to be unchanged during the learning process. As mentioned above, agents in the communication system need to change models and be sent to another federation once the number of users in the cell varies. Before the agent changes its federation, the central server has to issue the agent the global model of the target federation as a smart initialization, and then the agents can provide high-quality decisions for users immediately. Since the agent can flow freely among multiple federations, the framework is named as AFFL. The technical details of the federated aggregation and transfer learning will be explained, respectively.

Algorithm 6 AFFL-PPO algorithm for the sum rate optimisation

for each federation κ **do**
  Initialization: Determine the DNN specifications based on the number of active users and STAR-RIS elements
  Initialize the actor network $\omega^{\kappa}_{a,b,0}$, critic network $\omega^{\kappa}_{c,b,0}$, and global model $\omega^{\kappa}_{a,G,0}, \omega^{\kappa}_{c,G,0}$
  for each BS b^{κ} **do**
    for each iteration **do**
      if Globe update = True **then**
        Update globe model $\omega^{\kappa}_{a,G,t}, \omega^{\kappa}_{c,G,t}$ and local critic models $\omega^{\kappa}_{c,b,t}$ according to (6.31) (6.32) (6.33)
      end if
      if New agent b_0 joining in = True **then**
        Initialize models $\omega^{\kappa}_{a,b_0,t}$ and $\omega^{\kappa}_{c,b_0,t}$ according to (6.34) (6.35)
      end if
      for each time step t **do**
        Observe $\boldsymbol{s}_t$ and choose $\boldsymbol{a}_t$ according to the policy $\pi_{\omega_{\text{old}}}(\boldsymbol{a}_t|\boldsymbol{s}_t)$
        Execute action $\boldsymbol{a}_t$, observe r_t and $\boldsymbol{s}_{t+1}$
        Record transition $(\boldsymbol{s}_t, \boldsymbol{a}, r_t, \boldsymbol{s}_{t+1})$
        Calculate advantage estimates according to (6.23)
      end for
      for Each training epoch **do**
        Random sample and calculate loss according to (6.29)
        Calculate the policy gradient $\nabla_{\omega} L_t(\omega)$ and train actor/critic networks according to (6.30)
      end for
    end for
  end for
end for

6.3.2.2 Federated Aggregation

For PPO agents in the same federation, their training process is divided into four stages loop, which are local model update, local model upload, global model update, and globe model download.

- Local model update: Each agent update the actor $\omega^{\kappa}_{a,b}$ and critic $\omega^{\kappa}_{c,b}$ locally according to the principle of PPO2. The models are time varying during the training, which can also be denoted as $\omega^{\kappa}_{a,b,t}$ and critic $\omega^{\kappa}_{c,b,t}$.

- Local model upload: After a interval of training, each agent has to send its models $\omega^{\kappa}_{a,b,t}$ and critic $\omega^{\kappa}_{c,b,t}$ to the central server. Thus the agent only needs to save its own models but the central server has access to all models.

- Global model update: After receiving models from all agent belongs to federation κ, the globe models of κ can be generated by averaging the parameters of local models, which can be express as

$$\omega^{\kappa}_{a,G,t} = \frac{1}{B^{\kappa}} \sum_{b^{\kappa}=1}^{B^{\kappa}} \omega^{\kappa}_{a,b,t}, \tag{6.31}$$

$$\omega^{\kappa}_{c,G,t} = \frac{1}{B^{\kappa}} \sum_{b^{\kappa}=1}^{B^{\kappa}} \omega^{\kappa}_{c,b,t}, \tag{6.32}$$

  where B^{κ} represents the number of BSs in κ.

- Globe model download: Finally, each agent downloads the globe critic model to replace its local critic model

$$\omega^{\kappa}_{c,b,t} = \omega^{\kappa}_{c,G,t}, \forall b. \tag{6.33}$$

  Please note that the local actor model cannot be updated by the globe actor model since the CSI and user distribution is not identical in each cell.

This mode enables distributed learning and avoids frequent data transfers compared to the DPPO framework, since the frequency of federated aggregation can be adjusted.

6.3.2.3 Transfer Federation

When the agent b^{κ} transfer from federation $\kappa = 0$ to $\kappa = k$, where k is any non-zero integer, it no longer participates in federated aggregation of federation 0 in the upcoming time slot. Before the agent b^0 transferred to b^k, the central server needs to transfer models to b to replace the previous local models, which can be expressed as

$$\omega^{\kappa}_{a,b,t} = \omega^{k}_{c,G,t}, \tag{6.34}$$

$$\omega^{\kappa}_{c,b,t} = \omega^{k}_{c,G,t}. \tag{6.35}$$

Combining all the above steps, the pseudocode of the AFFL-PPO algorithm for federation κ is given in **Algorithm 6**. Note that the given pseudocode is for one federation for a clear presentation, in practice multiple federations are working in parallel.

Remark 10. *Thanks to the global model provided by transfer learning, on the one hand, the new joined agent owns a trained model, on the other hand, it does not need to have a large exploration rate as a fresh agent, thus it can keep a small action variance as a well-trained agent. Thus, there is only a negligible communication performance degradation for the agent who transferred federation.*

6.4 Numerical Results and Analysis

This section presents the simulation results to reveal the performance of the proposed STAR-RIS assisted NOMA network and the ML algorithm. Meanwhile, by analysing the numerical results, the provided solution and insights for ***Challenge 1-3*** are verified.

This section will first discuss the performance of the aforementioned two STAR-RIS models and the partition approaches, and then the performance of the proposed AFFL model is verified by comparing the performance of multi-agent AFFL training and the single-cell training.

6.4.1 Simulation Setup

In the simulation, the STAR-RIS located 4km away from the BS [2], where the users associated with the STAR-RIS are randomly distributed around the STAR-RIS, and the users are assumed to have random moving during each time step. There are 30 time steps in the transmission duration, and the CSI is assumed to vary per each time step. It is worth noting that the spacial relationship between the BS and the RIS is the same in different cellular, but the realization of the user distribution, user movements, as well as the channel fading blocks are independent in each cell. For the DRL agent, both the actor and the critic network have 4 layer structure, where the 'tanh' function is employed for the hidden layer. The scale of the DNN is adjusted according to the environment and the presented simulation results are all given by the agent with empirically optimal parameters. A decaying action noise is invoked to balance the exploration and the exploitation, starting with 0.6 at the beginning of the training, linearly decays and finally stays at 0.05. At the beginning of training, a larger exploration rate can better avoid local optimal problems, and at the end of training, a smaller exploration rate can ensure the training stability. The specific value selection is based on the empirical optimal value obtained in our simulation. The rest of the default parameters are listed in Table 6-A.

6.4.2 STAR-RIS and the Partitioning Problem

Fig. 6.4 demonstrates a pair of example partitions of the partitioned STAR-RIS in a fading block, where Fig. 6.4(a) is an example of the ES mode, and Fig. 6.4(b) is an

[2] A larger BS to STAR-RIS distance is adopted as users at the edge of the cell eager to signal enhancements, and users with closer distances can obtain a satisfactory data rate with a direct link.

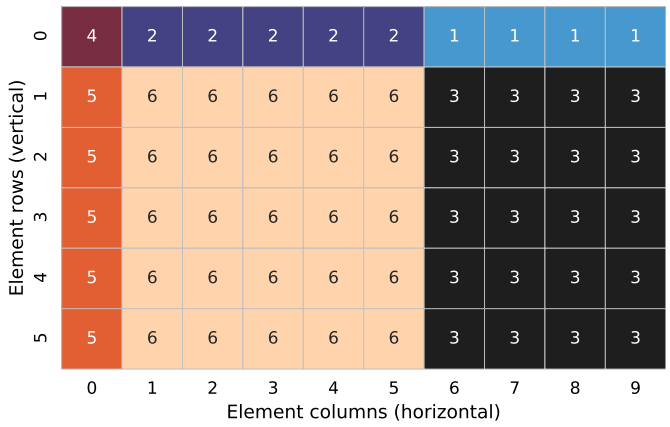
Table 6-A: Default Parameters

Parameter	Description	Value	Parameter	Description	Value
M	tile number	6	N	total element number	60
f_{c}	carrier frequency	5GHz	W	bandwidth	1 MHz
K	users per cell	4	$P_{\max}$	maximum power	29 dBm
K_{AR}, K_{RU}	Rician factors	3dB	σ	noise power density	-95.2 dBm/MHz
α	learning rate	3×10^{-4}	γ	discount factor	0.99
e	batch size	80 samples	N/A	Optimiser	Adam

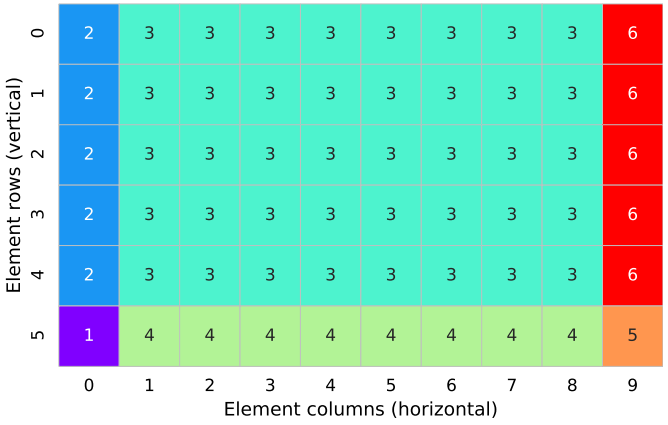
example of the MS mode. Each square represents a STAR element, and there are 60 elements on the STAR-RIS in the example. The elements with the same number and color label are allocated into the same tile. It is worth noting that, in the simulation of the MS mode, tiles 1-3 work in the R mode, while tiles 4-6 are employed to serve T users. It can be observed that the partition given by the agent has significant differences for ES and MS models under the same fading block. The partition given for the MS-based STAR-RIS can be explained by the classical greedy algorithm and the NOMA gain theory. However, the partition for the ES-based STAR-RIS produces two large tiles, tile 3 and tile 6. The reason leading to this phenomenon may be the QoS constraint. Since the ES model cannot adjust the phase arbitrarily, it is harder to meet the QoS requirements with small-scale tiles, such as tiles 1 and 5 in Fig. 6.4(b). Hence, the reason for having two larger tiles in Fig. 6.4(a) may not be that the two major tiles are needed, but to avoid the appearance of very small tiles. Therefore, although the interpretability of ML methods is relatively deficient, the partition results given by the DRL agent are interpretable with the help of theory in communication.

Fig. 6.5 plots the averaged reward[3] obtained during the simulated duration of 3 STAR-RIS assisted NOMA cells to reveal the performance of the STAR-RIS. The reward is positively related to the sum data rate of all users meeting the minimum rate requirement. Significant NOMA gains and STAR-RIS gains can be observed Fig. 6.5 aside from

[3]The average reward and average data rate refer to the average value of the reward and data rate obtained by the agent of multiple BSs and episodes, which is invoked to reduce the affect caused by the random action explorations and channel realizations.



(a) Example partitioning for the STAR-RIS(ES).



(b) Example partitioning for the STAR-RIS(MS).

Figure 6.4: Example partitioning for the STAR-RIS.

which type of STAR-RIS is employed. In the NOMA networks, the STAR-RISs obtain a reward of around 400 and outperform the conventional RIS investigated in [38], since the reflecting-only are not capable of severing all users. Furthermore, in both OMA and NOMA cases, the ES mode achieved superior reward compared to the MS mode. This result indicates that though ES elements cannot independently config the TRC, a larger

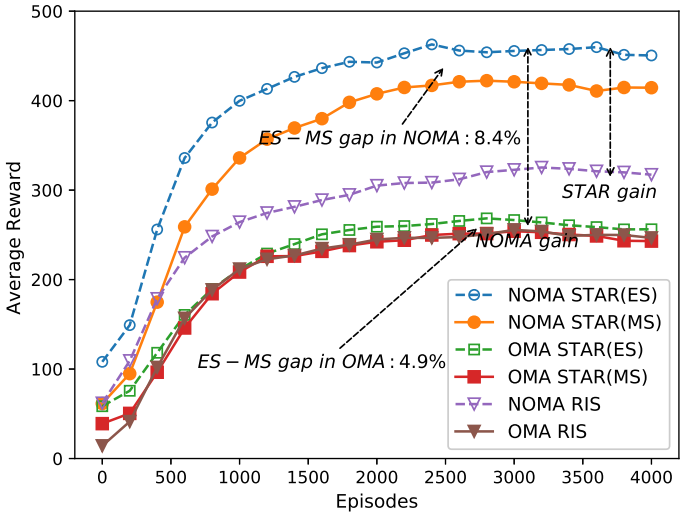
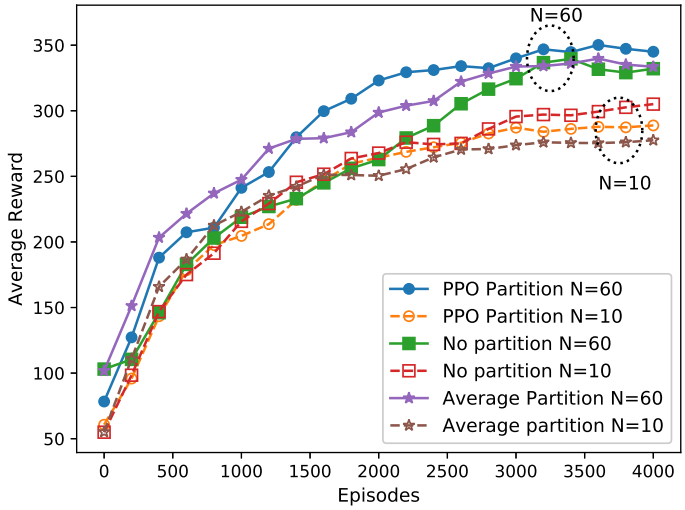


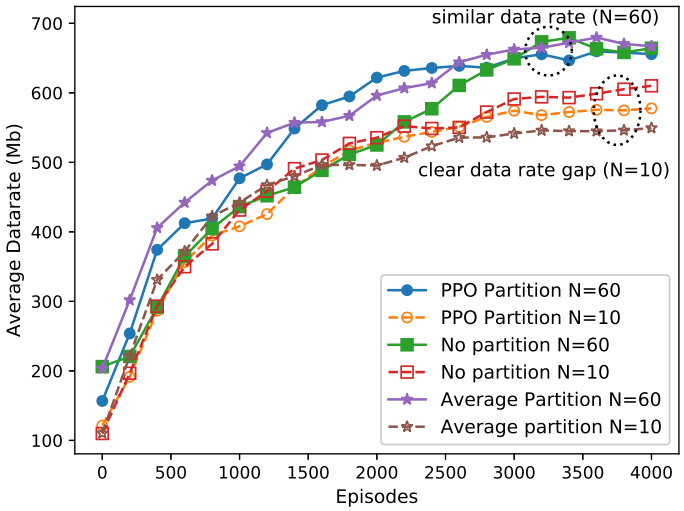
Figure 6.5: Performances of different types for STAR-RIS in OMA/NOMA networks

equivalent element number still dominates the performance. It is worth emphasizing that the ES mode has a stronger affinity to NOMA networks. In the NOMA network, the ES mode achieved a gain of about 8.4% compared to the MS mode, while in contrast to the OMA network, only a gain of 4.9% was achieved, which is in line with the discussion in **Remark 7**.

To explain and verify the necessity and advantage of the STAR-RIS partition, the system performance of three TRC controlling schemes are compared in Fig. 6.6, namely intelligent partitioning, average partitioning, and element-based intelligent control. Fig. 6.6(b) plots the obtained reward of the three approaches in cases of $N = 10$ and $N = 60$. While only $10(2 \times 5, M = 2)$ elements of STAR-RIS are employed, the element-based TRC control scheme achieves the highest reward, as the independent control can provide precise TRCs for each element. However, once a STAR-RIS having a large number of elements is invoked, the rank of the reward is changed, where the independent control scheme achieved an inferior reward to the partition-based control, which is caused by the increased element number and complexity that degrades the performance of the trained agent as analyzed in **Remark 9**. This tend is evidence to prove that when



(a) Reward obtained by different TRC controlling schemes.



(b) Throughput obtained by different TRC controlling schemes.

Figure 6.6: Performance comparison of different partitioning scheme for the STAR-RIS.

a larger scale of STAR-RIS is employed, unless accepting an extremely large training cost, the partition-based TRC control is likely to outperform the element-based control scheme, since the STAR-RIS partition can significantly simplify the problem to improve the performance of DRL agents. Meanwhile, the PPO-based partition outperformed the

average partition in both realizations, which indicates the advantage of the intelligent partition. Shifting the focus to Fig. 6.6(b), it can be observed that in the case of $N = 10$, the throughput has exactly the same ranking and trend as the reward, but at $N = 60$ the throughput of the three schemes became nearly coinciding. The results indicate that the intelligent partitioning scheme obtains additional rewards due to meeting the users' QoS requirements, suggesting that the DRL partitioning does not unilaterally pursue data rate maximisation like the water-filling algorithm, and coincides with the partitioning example shown in 6.4(a).

6.4.3 AFFL Distributed Model

This subsection refocus on the validation of the proposed AFFL framework. To evaluate the training process and performance of the agents, the performance indicators of the communication system are no longer in account, and the reward is considered as the sole performance indicator to evaluate the system performance.

Fig. 6.7 demonstrates the training process of a 3-cell federation started with random initialized models. In order to validate the performance of the DRL based beamforming and the AFFL framework, the algorithm in [100] is employed as a state-of-the-art (SOTA) benchmark. It can be observed that the proposed solutions outperform the benchmark algorithm in terms of convergence and optimality. Comparing two pairs of curves for the ES mode and the MS mode, firstly, it can be observed that a significant reward gain in the multi-agent case where AFFL is applied. Another significant advantage of AFFL is convergence. Although the large action noise at the early training stage affects the observation of convergence to some extent, taking the ES case as an example, the convergence shown by the blue dashed curve is better than the green one. Therefore, in a limited training period, the AFFL framework achieves better optimality and convergence by aggregating the critic model from multiple agents.

Fig. 6.8 plots the training process against different learning rates of a cell that

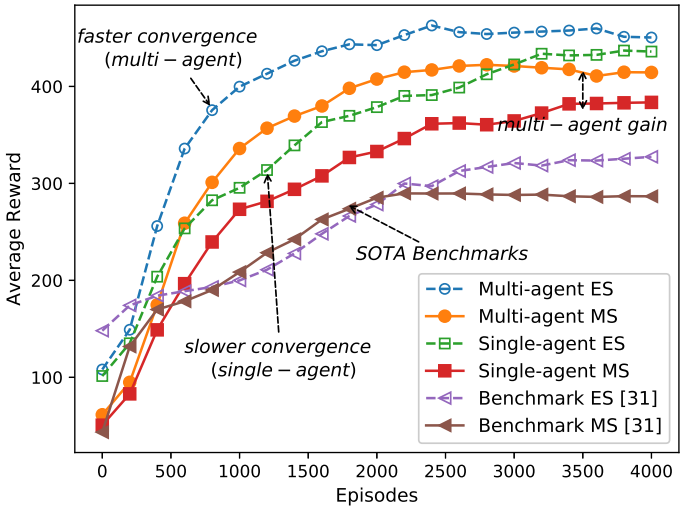


Figure 6.7: Performances of different learning frameworks for STAR-RIS

Table 6-B: Training time consumption of different partitioning approaches and training frameworks

Element number	no partition	average partition	individual PPO	AFFL-PPO (initialization)	AFFL-PPO (transfer)
10	41m55s	25m16s	27m02s	17m12s	0s
60	321m08s	34m26s	34m30s	19m39s	0s

switches federation needs to experience. In this scenario, assuming a cell becomes a new member of the federation mentioned in Fig. 6.7 due to the change in the number of users, and the reward of the new member agent is plotted. In case of the AFFL model is employed, since a fully trained model is distributed to the agent, it can obtain eligible rewards from episode 1, which indicates that users do not have to wait for the training of the agent and can be served immediately. Although the agent still obtains a small reward improvement with training, the initial service quality is still guaranteed as claimed in **Remark 10**. Conversely, users who did not under the AFFL model are not so lucky, during the first 1000 episodes of the training time, these users are likely to experience poor service due to the immature model. Even if optimality is sacrificed a bit and set a larger learning rate 0.001 to shorten the training process, a wait of about 500 epochs is still necessary. Therefore, this result illustrates that with the aid of transfer learning,

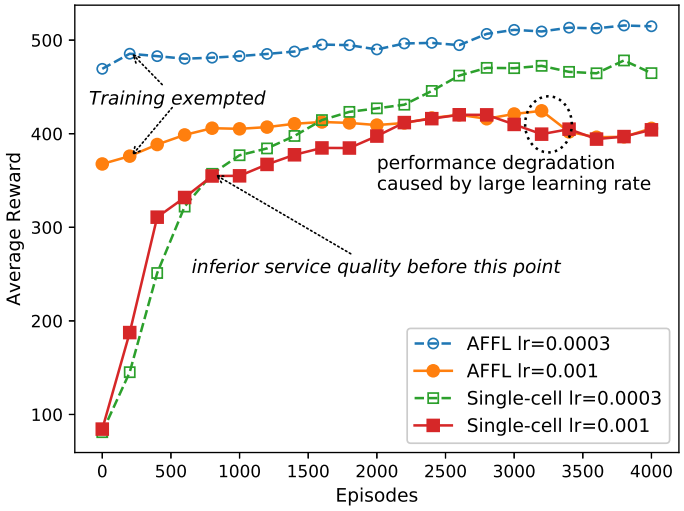


Figure 6.8: Performances of different learning frameworks for STAR-RIS

the AFFL model is capable of solving ***Challenge 3***.

Table 6-B presents the training time consumption per agent of different partitioning approaches and training frameworks. It can be observed that the training time consumed by the tile-based beamforming scheme is significantly diminished than the element-based beamforming, and the DRL-based beamforming does not produce discernible complexity compared to the averaged partitioning. In terms of the training framework, by adopting the model aggregation of multiple agents, the AFFL framework greatly reduces the training time of the model. Specifically, for the problem of dynamic network requirements mentioned in ***Challenge 3***, the AFFL model does not require any additional retraining but invokes model transfers to solve the problem. Although AFFL framework is able to exempt the training time by replacing models, the initial model still need to be trained (e.g. around 30 min in Table 6-B). Thus, the initial model have to be pre-trained before employing the agent in practice.

6.5 Summary

In this chapter, STAR-RISs assisted downlink NOMA networks were investigated, where both ES mode and MS mode were considered for the STAR-RIS. In order to achieve a low-complexity passive beamforming control for the STAR-RISs, a joint STAR-RIS partitioning and tile-based passive beamforming optimisation problem was formulated for maximising the average throughput of the BSs while satisfying the QoS requirements of the users. A PPO-based solution was proposed for the management of the STAR-RIS, and an AFFL distribute learning framework was proposed to accelerate or exempt the training process for the intelligent agents deployed at BSs. The analysis and the simulation results of this chapter provided the following insights and conclusion: 1) When STAR-RIS has a massive number of STAR elements, the joint partitioning and tile-based beamforming outperforms the conventional element-based beamforming. However, once the STAR-RIS does not have a massive number of elements, there are a trade-off problem between the beamforming complexity and the data rate gain; 2) ES-based STAR-RIS have compatibility with NOMA networks since the ES mode is likely to results in channel strength differences that NOMA prefers; 3) With the global critic network provided by AFFL, even if each environment has diverse CSI, the average reward has advantages compared to the independently trained DRL agent; 4) The AFFL model allows agents to freely switch between federations with a hot-start, which can protect the base station and STAR-RIS from short-term performance degradation due to un-trained DNN models.

Chapter 7

Conclusion

In this thesis, the prospects and applications of ML technologies in the RIS assisted wireless networks have been investigated. The models for RIS-assisted networks were presented and a number of optimisation issues were addressed, which include passive beamforming with configuration, mobile RIS deployments, hybrid beamforming, and tile-based beamforming for STAR-RISs. As answers to the optimisation problems, a number of ML approaches including ETDL, EA-DDPG, FL enhanced DDPG, Hybrid-DDPG, DDPG-DQN, and AFFL-PPO algorithms were developed. The simulation results confirmed that the improvement and contribution of RISs engaged in wireless networks. Computer simulation also comprehensively confirms and evaluates the performance of the proposed algorithm from multiple perspectives.

The proposed techniques have multiple application scenarios to enhance the communication quality and energy efficiency of terrestrial mobile users. Specifically, 1. the research in Chapter 3 can reduce the time overhead of RIS beamforming in practice; 2. The deterministic propagation model and mobile RIS proposed in Chapter 4 can well assist home WIFI networks; 3. The STAR-RIS proposed in Chapter 5 can be applied in the window, using two beams to serve indoors and outdoors, thereby enhancing the performance of the macro base station; 4. The low-complexity beamforming technique in

Chapter 6 provides a feasible beamforming scheme for large RIS with a massive number of reconfigurable elements.

7.1 Summary of Contributions

In this thesis, the author made the following main research contributions

- In Chapter 3, a RIS-assisted multi-user communication system over multiple fading channels was investigated, where the time overhead for configuring the reflection coefficients at the RIS is taken into consideration. A long-term effective throughput maximisation problem is formulated for joint optimisation of the passive beamforming of the RIS and the power allocation of the AP over each channel fading block, subject to the hardware constraints of RIS configurations. ETDL and EA-DDPG algorithms were proposed to solve the resulting overhead-dependent joint optimization problems. The ETDL algorithm trains the DL agent through the communication environment in anticipation of eliminating the requirement of the training data set. EA-DDPG algorithm has the ability to achieve continuous and deterministic control of phase shifts. The author compared and analyze the optimisation performances of the algorithms via simulation, which provides insights into the methodology for the optimisation of wireless communications. The simulation results reveal that the RL and DL algorithms are capable of achieving comparable overall performance but each has its own advantages. The training complexity required by the DL approach is significantly less than the RL approach. On the contrary, the RL algorithm can better identify the system state, and thereby choose more flexible strategies and decisions.

- In Chapter 4, a novel indoor communication model was proposed, which employs mobile RIS to enhance the channel quality for users. Compared to existing fixed RIS paradigms, the proposed framework is capable to cover indoor users who suffered from obstructed environments with the aid of flexible deployments of RISs,

thereby increasing the sum data rate. In order to further increase user capacity and increase spectrum efficiency, NOMA techniques are invoked. A corresponding dynamic decoding order scheme is adopted, since the channels intervened by mobile RISs are likely to significantly impact the user's CSI. The DDPG algorithm is invoked to jointly optimise the deployments, phase shifts of mobile RISs, and the power allocation policy for users. Furthermore, a federated learning enabled DRL framework was developed to reduce the training time of agents and theoretically prove that within limited training processes, the FL framework is capable to provide reward gains for DRL agents. An FL model with local training and periodic global model was invoked update to enable the agent in each cell to learn from others' experiences and thereby improve the efficiency of exploration and training. In addition, the impact of the different propagation characteristics of each cell on FL learning effects is also investigated.

- In Chapter 5, a STAR-RIS model for broadening the coverage of reflecting-only RISs was conceived. Specifically, the practical electromagnetic property of STAR elements is considered, resulting in a coupled phase-shift of the transmission and reflection. Based on the proposed model, a joint active and passive beamforming problem that requires hybrid control for the phase-shift and amplitude is formulated for minimizing the long-term power consumption. The author proposed a hybrid DDPG algorithm for solving the hybrid control problem caused by the energy splitting nature of STAR elements. The hybrid control is carried out by mapping each output node of the actor network to the transmission and reflection actions of each STAR element. The proposed hybrid DDPG solution provides high-dimensional continuous and discrete phase-shift optimisation for STAR-RISs. A joint DDPG-DQN algorithm was developed as a high performance-solution. The joint DDPG-DQN scheme can handle hybrid control by employing two collaborated agents, where a DDPG agent is in charge of the continuous control and a DQN agent is invoked for the discrete control. The performance of the STAR-RIS

relying on the proposed algorithms is evaluated by computer simulation revealing that it outperforms both the reflecting-only and the double spliced RISs. Furthermore, the hybrid DDPG algorithm outperforms its plain DDPG counterpart without increasing its complexity, while the joint DDPG-DQN algorithm attains optimality at an increased complexity.

- In Chapter 6, a STAR-RIS assisted NOMA network model was conceived, where a sum rate maximisation problem is formulated for the proposed model. Two STAR-RIS operating protocols, ES and MS are investigated and their performances are compared in this chapter. A novel tile-based passive beamforming approach was proposed for the STAR-RIS, where the STAR-RIS is intelligently partitioned into several tiles. A PPO based algorithm is invoked to optimise the partitioning of STAR elements, as well as the tile-based transmission and reflection beamforming, which is the first research on the partition problem of the STAR-RIS. An AFFL paradigm was proposed for deploying artificial intelligent agents to cellular networks, where each BS plays a role as an agent. This novel paradigm utilizes the feature of both FL and TL, authorizing agents to freely join or quit federations according to the network status. The new member of a federation can inherit well-trained global models when joining in the federation, which does not require additional time for retraining agents to ensure uninterrupted high-quality communication service. The analysis and simulation results indicate that the tile-based beamforming approach is a stable and low-complexity solution. It is also proved that the AFFL paradigm not only increases training speed and optimality, but also enables training-free model switching for the agent who changes its federation attribution, providing efficient solutions.

7.2 Limitation and Future Work

7.2.1 Limitations of This Thesis

Although this thesis provides some possible schemes and research contributions for RIS-assisted wireless communication, there are still some limitations. 1. Some assumptions are idealized, for example, perfect CSI is assumed to be available, and NOMA's SIC decoding is assumed to be perfect. These assumptions are unlikely to be implemented in practical communication systems, but we were unable to investigate these issues further due to the limitations of the research scope. 2. Some power consumption and overhead are not fully considered, such as pilot overhead and the energy overhead of robots in mobile RIS. These overheads cannot be avoided in the physical world, and more work on RIS overheads is worth carrying out to further promote the standardization and application of RIS. 3. This thesis does not introduce experiments to verify the proposed algorithms. Although some effort has been made to purchase a STAR-RIS prototype to carry out experiments, the author failed to get the opportunity to deploy AI algorithms on hardware due to product and time constraints.

7.2.2 Joint Active & Passive Beamforming with Practical Reflection Model and Imperfect CSI

As discussed in this thesis, the joint beamforming of the base station and RIS is one of the core issues of RIS research. The existing research generally assumed the perfectly ideal reflection coefficient for RIS. However, some RIS-related materials and electromagnetic studies have proved that the phase shift and amplitude of the reflection coefficient may have non-ideal features (e.g. phase-shift error). RIS modeling and optimisation in non-ideal scenarios is still a problem to be solved in this field, since the non-ideal characteristics of reflective elements bring irregular constraints to the optimisation.

Meanwhile, in practice, the acquisition of accurate CSI is another severe problem,

especially for mobile users moving at high velocity. Therefore, how to accomplish beamforming with limited or imperfect CSI is a practical challenge. It is well-known that the perfect CSI cannot be obtained in practice. Furthermore, due to consideration of power consumption and economic cost, whether RIS needs to have channel estimation capability has also been discussed in academia.

In order these two practical issues, future work will focus on (1) the robust RIS beamforming with a consideration of configuration error or elements' non-ideal characteristics; (2) the RIS beamforming with imperfect, partial, or non-CSI. Machine learning is a potential technology to solve this problem it can quickly adapt to limited observations on CSI and non-ideal properties by training.

7.2.3 The Comprehensive Configuration Overhead Investigation

Chapter 3 of this thesis focuses on the time overhead caused by configuring reconfigurable elements arrays. In addition to the signaling overhead for reconfigurable elements, there are other possible sources of overheads, which may lead to performance degradation of the RIS-assisted transmission. The control signaling generated by the RIS intelligent controller needs to be transmitted to the RIS, which may be another source of configuration time overhead [177]. Moreover, Although some CSI acquisition methods for RIS-assisted communication have been proposed [72], the time overhead of these methods has not been systematically and quantitatively investigated. Therefore, identifying the comprehensive overhead of multiple sources and investigating how to reduce these configuration overheads will be an important issue to improve the performance of RISs in practice.

7.2.4 Competition and Collaboration between Multi-RISs

The lack of research focus in the area of RISs is multi-RIS collaboration. The main challenge is the complexity brought about by the interaction among multiple RISs since RISs

can affect one another. Iterative optimisation will have a tough time-solving problem brought on by more than two RISs in particular. In practice, RIS cannot be established in isolation in practice, thus, it is promising to utilize a mix of game theory and multi-agent machine learning to address the collaboration issue between multiple RISs.

7.2.5 RIS for Near-field Communications

The increase in the number of RIS elements makes large-array RIS a trend in the future development of RIS. The classic far-field propagation model may have significant inaccuracies due to the growth in the number and size of RIS array elements. When the user is close to the RIS, the conventionally used plane wave model is no longer fit in this case, which would decrease the quality of communication. In order to enhance the communication quality of near-field communication, research will be conducted on the propagation model, communication protocol, and beamforming in the near-field communication of large-array RIS.

Appendix A

Proof of Remark 4 in Chapter 4

Assuming a Markov process has $\mathbf{S}$ states, each state $\mathbf{s} \in \mathbf{S}$ has action space $\mathbf{A}_N^{\mathbf{s}}$ and the corresponding reward set $\mathbf{R}_N^{(\mathbf{s},\mathbf{a})}$, denoting the explored action space of FL agents as $\mathbf{A}_F^{\mathbf{s}} \subseteq \mathbf{A}_N^{\mathbf{s}}$,$\mathbf{R}_F^{(\mathbf{s},\mathbf{a})} \subseteq \mathbf{R}_N^{(\mathbf{s},\mathbf{a})}$ and explored action space of the independent agent as $\mathbf{A}_I^{\mathbf{s}} \subseteq \mathbf{A}_N^{\mathbf{s}}$,$\mathbf{R}_I^{(\mathbf{s},\mathbf{a})} \subseteq \mathbf{R}_N^{(\mathbf{s},\mathbf{a})}$. Assuming that a repetitive tolerant random action policy is adopted during the exploration process, we can get $E[\mid \mathbf{A}_F^{\mathbf{s}} \mid] \geq E[\mid \mathbf{A}_I^{\mathbf{s}} \mid]$ and also for the reward set $E[\mid \mathbf{R}_F^{(\mathbf{s},\mathbf{a})} \mid] \geq E[\mid \mathbf{R}_I^{(\mathbf{s},\mathbf{a})} \mid]$. For the reward sets, the maximum known reward $\max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})}, \mathbf{R}_I^{(\mathbf{s},\mathbf{a})})$ always exists, though there may have $\max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})}, \mathbf{R}_I^{(\mathbf{s},\mathbf{a})}) < \max(\mathbf{R}_N^{(\mathbf{s},\mathbf{a})})$. Then the probability that the known maximum reward is found by FL agents and independent agents can be calculated as

$$P[\max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})}, \mathbf{R}_I^{(\mathbf{s},\mathbf{a})}) \in \mathbf{R}_I^{(\mathbf{s},\mathbf{a})}] = 1 - (1 - \frac{1}{\mid \mathbf{R}_N^{(\mathbf{s},\mathbf{a})} \mid_{\mathbf{R}_N^{(\mathbf{s},\mathbf{a})} \leq \max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})}, \mathbf{R}_I^{(\mathbf{s},\mathbf{a})})})^{\mathbb{E}[\mid \mathbf{A}_I^{\mathbf{s}} \mid]}, \quad \text{(A.1)}$$

$$P[\max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})}, \mathbf{R}_I^{(\mathbf{s},\mathbf{a})}) \in \mathbf{R}_F^{(\mathbf{s},\mathbf{a})}] = 1 - (1 - \frac{1}{\mid \mathbf{R}_N^{(\mathbf{s},\mathbf{a})} \mid_{\mathbf{R}_N^{(\mathbf{s},\mathbf{a})} \leq \max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})}, \mathbf{R}_I^{(\mathbf{s},\mathbf{a})})})^{\mathbb{E}[\mid \mathbf{A}_F^{\mathbf{s}} \mid]}. \quad \text{(A.2)}$$

Since $\mathbb{E}[|\ \mathbf{A}_F^{\mathbf{s}}\ |] \geq \mathbb{E}[|\ \mathbf{A}_I^{\mathbf{s}}\ |]$, then

$$P[\max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})}, \mathbf{R}_I^{(\mathbf{s},\mathbf{a})}) \in \mathbf{R}_I^{(\mathbf{s},\mathbf{a})}] \leq P[\max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})}, \mathbf{R}_I^{(\mathbf{s},\mathbf{a})}) \in \mathbf{R}_F^{(\mathbf{s},\mathbf{a})}], \forall \mathbf{s} \in \mathbf{S}, \tag{A.3}$$

and it can be obtained that

$$\mathbb{E}[\max(\mathbf{R}_I^{(\mathbf{s},\mathbf{a})})] \leq \mathbb{E}[\max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})})], \forall \mathbf{s} \in \mathbf{S}. \tag{A.4}$$

It is worth to point out that $\mathbb{E}[\max(\mathbf{R}_I^{(\mathbf{s},\mathbf{a})})] = \mathbb{E}[\max(\mathbf{R}_F^{(\mathbf{s},\mathbf{a})})]$ if $\mathbf{A}_F^{\mathbf{s}} = \mathbf{A}_I^S = \mathbf{A}_N^S$, which reveals that when the action space is traversed, FL will no longer provide gain. Then, the cumulative reward which described by value function

$$Q(\mathbf{s},\mathbf{a}) = R(\mathbf{s},\mathbf{a}) + \beta \sum_{\mathbf{s}' \in \mathbf{S}} \max_{\mathbf{a}'} Q(\mathbf{s}',\mathbf{a}'). \tag{A.5}$$

According to (A.4) and (A.5), we can obtain

$$\mathbb{E}[\max(Q_I(\mathbf{s},\mathbf{a}))] \leq \mathbb{E}[Q_F(\mathbf{s},\mathbf{a}))]. \tag{A.6}$$

For DQL algorithms, assuming the neuronal network fitting correctly, we have $Q^* \rightarrow Q$, then

$$\mathbb{E}[\max(Q_I^*(\mathbf{s},\mathbf{a}))] \leq \mathbb{E}[Q_F^*(\mathbf{s},\mathbf{a}))]. \tag{A.7}$$

Remark 4 is proven.

References

[1] G. Oliveri, D. H. Werner, and A. Massa, "Reconfigurable electromagnetics through metamaterialsa review," *Proceedings of the IEEE*, vol. 103, no. 7, pp. 1034–1056, 2015.

[2] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[3] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, 2021.

[4] J. Hu, H. Zhang, B. Di, L. Li, K. Bian, L. Song, Y. Li, Z. Han, and H. V. Poor, "Reconfigurable intelligent surface based RF sensing: Design, optimization, and implementation," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2700–2716, Jul. 2020.

[5] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[6] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2019.

[7] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May. 2020.

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

[9] T. Hou, Y. Liu, Z. Song, X. Sun, Y. Chen, and L. Hanzo, "Reconfigurable intelligent surface aided NOMA networks," *IEEE J. Sel. Areas Commun.*, vol. 38,

no. 11, pp. 2575–2588, Nov. 2020.

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

[11] Y. Li, M. Jiang, Q. Zhang, and J. Qin, "Joint beamforming design in multi-cluster MISO NOMA reconfigurable intelligent surface-aided downlink communication networks," *IEEE Trans. on Commun.*, vol. 69, no. 1, pp. 664–674, Jan. 2021.

[12] S. Zhang, H. Zhang, B. Di, Y. Tan, M. Di Renzo, Z. Han, H. Vincent Poor, and L. Song, "Intelligent omni-surfaces: Ubiquitous wireless transmission by reflective-refractive metasurfaces," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 1, pp. 219–233, 2022.

[13] B. Zheng, C. You, and R. Zhang, "Double-IRS assisted multi-user MIMO: Cooperative passive beamforming design," *IEEE Trans. Wirel. Commun.*, 2021, early access, doi:10.1109/TWC.2021.3059945.

[14] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[15] J. M. Rojas and G. Fraser, "Is search-based unit test generation research stuck in a local optimum?" in *IEEE/ACM SBST*, 2017, pp. 51–52.

[16] C. Huang, G. Chen, J. Tang, P. Xiao, and Z. Han, "Machine-learning-empowered passive beamforming and routing design for multi-RIS-assisted multihop networks," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 25 673–25 684, 2022.

[17] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Trans. Veh.*, vol. 70, no. 4, pp. 3978–3983, 2021.

[18] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 4, pp. 3039–3071, Jul. 2019.

[19] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. V. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Trans. Wirel. Commun.*, vol. 12, no. 7, pp. 3202–3212, July 2013.

[20] P. Papadimitroulas, L. Brocki, N. C. Chung, W. Marchadour, F. Vermet, L. Gaubert, V. Eleftheriadis, D. Plachouris, D. Visvikis, G. C. Kagadis *et al.*, "Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization," *Physica Medica*, vol. 83, pp. 108–121, 2021.

[21] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May. 2020.

[22] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wirel. Commun.*, vol. 27, no. 2, pp. 72–80, Apr. 2020.

[23] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S.-L. Kim, and M. Debbah, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *arXiv preprint arXiv:2008.02608*, 2020.

[24] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8238–8242, Aug. 2019.

[25] R. Liu, M. Li, H. Luo, Q. Liu, and A. L. Swindlehurst, "Integrated sensing and communication with reconfigurable intelligent surfaces: Opportunities, applications, and future directions," *IEEE Wirel. Commun.*, vol. 30, no. 1, pp. 50–57, 2023.

[26] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, and Y.-C. Liang, "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 4, pp. 2283–2314, 2020.

[27] C. Pan, H. Ren, K. Wang, J. F. Kolb, M. Elkashlan, M. Chen, M. Di Renzo, Y. Hao, J. Wang, A. L. Swindlehurst, X. You, and L. Hanzo, "Reconfigurable intelligent surfaces for 6g systems: Principles, applications, and research directions," *IEEE*

Commun. Mag., vol. 59, no. 6, pp. 14–20, 2021.

[28] C. Pan, G. Zhou, K. Zhi, S. Hong, T. Wu, Y. Pan, H. Ren, M. D. Renzo, A. Lee Swindlehurst, R. Zhang, and A. Y. Zhang, "An overview of signal processing techniques for RIS/IRS-aided wireless systems," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 5, pp. 883–917, 2022.

[29] S. Basharat, S. Ali Hassan, H. Pervaiz, A. Mahmood, Z. Ding, and M. Gidlund, "Reconfigurable intelligent surfaces: Potentials, applications, and challenges for 6G wireless networks," *IEEE Wirel. Commun.*, pp. 1–8, 2021.

[30] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 5, pp. 3083–3098, 2022.

[31] S. Zeng, H. Zhang, B. Di, Z. Han, and L. Song, "Reconfigurable intelligent surface (RIS) assisted wireless coverage extension: RIS orientation and location optimization," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 269–273, Jan. 2021.

[32] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "Multi-beam NOMA for hybrid mmwave systems," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1705–1719, Feb. 2019.

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

[34] Y. Cai, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann, "Modulation and multiple access for 5G networks," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 1, pp. 629–646, Oct. 2017.

[35] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of NOMA over OMA in uplink communication systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 536–568, Jan. 2020.

[36] M. Zeng, R. Du, V. Fodor, and C. Fischione, "Computation rate maximization for wireless powered mobile edge computing with NOMA," in *IEEE 20th International Symposium on WoWMoM*, Washington, DC, USA, Aug. 2019, pp. 1–9.

[37] V. Kumar, Z. Ding, and M. Flanagan, "On the performance of downlink NOMA

in underlay spectrum sharing," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4523 – 4540, March. 2021.

[38] Y. Liu, X. Mu, X. Liu, M. Di Renzo, Z. Ding, and R. Schober, "Reconfigurable intelligent surface-aided multi-user networks: Interplay between NOMA and RIS," *IEEE Wirel. Commun.*, vol. 29, no. 2, pp. 169–176, Apr. 2022.

[39] Q. Xu, C. Jiang, Y. Han, B. Wang, and K. J. R. Liu, "Waveforming: An overview with beamforming," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 1, pp. 132–149, Sep. 2017.

[40] Y. Yang, S. Zhang, and R. Zhang, "IRS-enhanced OFDMA: Joint resource allocation and passive beamforming optimization," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 6, pp. 760–764, Jun. 2020.

[41] S. Hu, Z. Wei, Y. Cai, C. Liu, D. W. K. Ng, and J. Yuan, "Robust and secure sum-rate maximization for multiuser MISO downlink systems with self-sustainable IRS," *arXiv preprint arXiv:2101.10549*, 2021.

[42] K. Weinberger, A. A. Ahmad, A. Sezgin, and A. Zappone, "Synergistic benefits in irs- and rs-enabled c-ran with energy-efficient clustering," *IEEE Trans. Wirel. Communi.*, vol. 21, no. 10, pp. 8459–8475, 2022.

[43] X. Pei, H. Yin, L. Tan, L. Cao, Z. Li, K. Wang, K. Zhang, and E. Björnson, "RIS-aided wireless communications: Prototyping, adaptive beamforming, and indoor/outdoor field trials," *arXiv preprint arXiv:2103.00534*, 2021.

[44] E. Basar, "Reconfigurable intelligent surface-based index modulation: A new beyond MIMO paradigm for 6G," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3187–3196, Feb. 2020.

[45] Z. Lin, H. Niu, K. An, Y. Wang, G. Zheng, S. Chatzinotas, and Y. Hu, "Refracting RIS aided hybrid satellite-terrestrial relay networks: Joint beamforming design and optimization," *IEEE Trans. Aerosp. Electron. Syst.*, p. Early Access, Mar. 2022.

[46] B. O. Zhu, K. Chen, N. Jia, L. Sun, J. Zhao, T. Jiang, and Y. Feng, "Dynamic control of electromagnetic wave propagation with the equivalent principle inspired tunable metasurface," *Sci. rep.*, vol. 4, no. 1, pp. 1–7, May. 2014.

[47] A. R. Ndjiongue, T. Ngatched, O. A. Dobre, and H. Haas, "Double-sided beam-

forming in OWC systems using omni-digital reconfigurable intelligent surfaces," *arXiv preprint arXiv:2203.03781*, Mar. 2022.

[48] Q. Wu, T. Lin, M. Liu, and Y. Zhu, "BIOS: An omni RIS for independent reflection and refraction beamforming," *IEEE Wirel. Commun. Lett.*, vol. 11, no. 5, pp. 1062–1066, May. 2022.

[49] Y. Zhang, B. Di, H. Zhang, Z. Han, H. Vincent Poor, and L. Song, "Meta-wall: Intelligent omni-surfaces aided multi-cell MIMO communications," *IEEE Trans. Wirel. Commun.*, p. Early Access, Mar. 2022.

[50] Y. Liu, X. Mu, R. Schober, and H. V. Poor, "Simultaneously transmitting and reflecting (STAR)-RISs: A coupled phase-shift model," in *ICC*, Aug. 2022, pp. 2840–2845.

[51] J. Xu, Y. Liu, X. Mu, R. Schober, and H. V. Poor, "STAR-RISs: A correlated T&R phase-shift model and practical phase-shift configuration strategies," *IEEE J. Sel. Top. Signal Process*, vol. 16, no. 5, pp. 1097–1111, May 2022.

[52] I. Budhiraja, S. Tyagi, S. Tanwar, N. Kumar, and J. J. Rodrigues, "Tactile internet for smart communities in 5G: An insight for NOMA-based solutions," *IEEE Trans. Ind. Inform.*, vol. 15, no. 5, pp. 3104–3112, 2019.

[53] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 55–61, Sept. 2017.

[54] Y. Liu, Z. Qin, Y. Cai, Y. Gao, G. Y. Li, and A. Nallanathan, "UAV communications based on non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 52–57, 2019.

[55] Z. Zhang, H. Sun, and R. Q. Hu, "Downlink and uplink non-orthogonal multiple access in a dense wireless network," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2771–2784, 2017.

[56] D. Zhai, R. Zhang, L. Cai, and F. R. Yu, "Delay minimization for massive Internet of Things with non-orthogonal multiple access," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 3, pp. 553–566, 2019.

[57] 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 Trans. Wirel. Commun.*, vol. 15, no. 11, pp. 7244–7257, Aug. 2016.

[58] D. Zhai, R. Zhang, L. Cai, B. Li, and Y. Jiang, "Energy-efficient user scheduling and power allocation for NOMA-based wireless networks with massive iot devices," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1857–1868, 2018.

[59] V. C. Thirumavalavan and T. S. Jayaraman, "BER analysis of reconfigurable intelligent surface assisted downlink power domain NOMA system," in *Proc. 2020 COMSNETS, Bengaluru, India*, Jan. 2020, pp. 519–522.

[60] L. Yang and Y. Yuan, "Secrecy outage probability analysis for ris-assisted NOMA systems," *Electron. Lett.*, vol. 56, no. 23, pp. 1254–1256, Nov. 2020.

[61] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE J. Sel. Areas Commun.*, Dec. 2020, doi: 10.1109/JSAC.2020.3041401.

[62] A. Khaleel and E. Basar, "A novel NOMA solution with RIS partitioning," *arXiv preprint arXiv:2011.10977*, 2020.

[63] M. Zhang, M. Chen, Z. Yang, H. Asgari, and M. Shikh-Bahaei, "Joint user clustering and passive beamforming for downlink NOMA system with reconfigurable intelligent surface," in *Proc. IEEE 31st PIMRC, London, UK*, Aug. 2020, doi:10.1109/PIMRC48278.20

[64] M. Elhattab, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Reconfigurable intelligent surface assisted coordinated multipoint in downlink NOMA networks," *IEEE Commun. Lett.*, Oct. 2020, doi:10.1109/LCOMM.2020.3029717.

[65] M. Fu, Y. Zhou, Y. Shi, and K. B. Letaief, "Reconfigurable intelligent surface empowered downlink non-orthogonal multiple access," *arXiv preprint arXiv:1910.07361*, 2019.

[66] G. Yang, X. Xu, Y.-C. Liang, and M. D. Renzo, "Reconfigurable intelligent surface-assisted non-orthogonal multiple access," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 5, pp. 3137–3151, Jan. 2021.

[67] H. Wang, C. Liu, Z. Shi, Y. Fu, and R. Song, "Power minimization for two-cell IRS-aided NOMA systems with joint detection," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1635–1639, Dec. 2021.

[68] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Joint deployment and multiple access design for intelligent reflecting surface assisted networks," *IEEE Trans. Wirel. Commun.*, 2021, early access, doi:10.1109/TWC.2021.3075885.

[69] Z. Qin, G. Y. Li, and H. Ye, "Federated learning and wireless communications," *arXiv preprint arXiv:2005.05265*, 2020.

[70] H. Ye, L. Liang, G. Y. Li, and B.-H. Juang, "Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 5, pp. 3133–3143, Feb. 2020.

[71] H. Yang, Z. Xiong, J. Zhao, D. Niyato, Q. Wu, H. V. Poor, and M. Tornatore, "Intelligent reflecting surface assisted anti-jamming communications: A fast reinforcement learning approach," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 3, pp. 1963–1974, Mar. 2021.

[72] S. Gao, P. Dong, Z. Pan, and G. Y. Li, "Deep multi-stage CSI acquisition for reconfigurable intelligent surface aided MIMO systems," *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 2024–2028, 2021.

[73] C. Liu, X. Liu, D. W. K. Ng, and J. Yuan, "Deep residual learning for channel estimation in intelligent reflecting surface-assisted multi-user communications," *IEEE Trans. Wirel. Commun.*, Aug. 2021, early access, doi:10.1109/TWC.2021.3100148 .

[74] X. Wei, D. Shen, and L. Dai, "Channel estimation for RIS assisted wireless communicationspart I: Fundamentals, solutions, and future opportunities," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1398–1402, May. 2021.

[75] T. Jiang, H. V. Cheng, and W. Yu, "Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation," *IEEE J. Sel. Areas Commun.*, 2021, early access, doi:10.1109/JSAC.2021.3078502.

[76] F. Jiang, L. Yang, D. B. da Costa, and Q. Wu, "Channel estimation via direct calculation and deep learning for RIS-aided mmwave systems," *arXiv preprint arXiv:2008.04704*, 2020.

[77] C. Liu, Z. Wei, D. W. K. Ng, J. Yuan, and Y.-C. Liang, "Deep transfer learning for signal detection in ambient backscatter communications," *IEEE Trans. Wirel.*

Commun., vol. 20, no. 3, pp. 1624–1638, Mar. 2021.

[78] H. Hojatian, V. N. Ha, J. Nadal, J.-F. Frigon, and F. Leduc-Primeau, "RSSI-based hybrid beamforming design with deep learning," in *IEEE International Conference on Communications*, 2020, pp. 1–6.

[79] A. M. Elbir, "A deep learning framework for hybrid beamforming without instantaneous CSI feedback," *IEEE Trans. Veh. Tech.*, vol. 69, no. 10, pp. 11 743–11 755, 2020.

[80] H. Huang, S. Guo, G. Gui, Z. Yang, J. Zhang, H. Sari, and F. Adachi, "Deep learning for physical-layer 5g wireless techniques: Opportunities, challenges and solutions," *IEEE Wirel. Commun.*, vol. 27, no. 1, pp. 214–222, 2020.

[81] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," *Proc. IEEE*, vol. 108, no. 2, pp. 341–356, 2020.

[82] H. Huang, Y. Yang, Z. Ding, H. Wang, H. Sari, and F. Adachi, "Deep learning-based sum data rate and energy efficiency optimization for MIMO-NOMA systems," *IEEE Trans. Wirel.Commun.*, vol. 19, no. 8, pp. 5373–5388, 2020.

[83] H. Song, M. Zhang, J. Gao, and C. Zhong, "Unsupervised learning-based joint active and passive beamforming design for reconfigurable intelligent surfaces aided wireless networks," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 892–896, Mar. 2021.

[84] ——, "Unsupervised learning-based joint active and passive beamforming design for reconfigurable intelligent surfaces aided wireless networks," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 892–896, Dec. 2021.

[85] C. Liu, X. Liu, Z. Wei, S. Hu, D. W. K. Ng, and J. Yuan, "Deep learning-empowered predictive beamforming for IRS-assisted multi-user communications," *arXiv preprint arXiv:2104.12309*, 2021.

[86] B. Yang, X. Cao, C. Huang, C. Yuen, L. Qian, and M. D. Renzo, "Intelligent spectrum learning for wireless networks with reconfigurable intelligent surfaces," *IEEE Trans. Veh. Tech.*, vol. 70, no. 4, pp. 3920–3925, March, 2021.

[87] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforce-

ment learning approach," *IEEE Trans. Veh. Tech.*, vol. 70, no. 4, pp. 3978–3983, March. 2021.

[88] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[89] A. Al-Hilo, M. Samir, M. Elhattab, C. Assi, and S. Sharafeddine, "Reconfigurable intelligent surface enabled vehicular communication: Joint user scheduling and passive beamforming," *arXiv preprint arXiv:2101.12247*, 2021.

[90] X. Gao, Y. Liu, X. Liu, and L. Song, "Machine learning empowered resource allocation in IRS aided MISO-NOMA networks," *arXiv preprint arXiv:2103.11791*, 2021.

[91] L. Wang, K. Wang, C. Pan, W. Xu, and N. Aslam, "Joint trajectory and passive beamforming design for intelligent reflecting surface-aided UAV communications: A deep reinforcement learning approach," *arXiv preprint arXiv:2007.08380*, 2020.

[92] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 3, pp. 1472–1514, Jan. 2020.

[93] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[94] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[95] X. Liu, Y. Liu, Y. Chen, and H. V. Poor, "RIS enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1057–1071, 2021.

[96] W. Ni, Y. Liu, Y. C. Eldar, Z. Yang, and H. Tian, "STAR-RIS enabled heterogeneous networks: Ubiquitous NOMA communication and pervasive federated learning," *arXiv preprint arXiv:2106.08592*, 2021.

[97] H. Yang, Z. Xiong, J. Zhao, D. Niyato, Q. Wu, H. V. Poor, and M. Torna-

tore, "Intelligent reflecting surface assisted anti-jamming communications: A fast reinforcement learning approach," *IEEE Trans. Wirel. Commun.*, Nov. 2020, doi:10.1109/TWC.2020.3037767.

[98] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wirel. Commun.*, 2020.

[99] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 11, pp. 5141–5152, 2019.

[100] Y. Guo, F. Fang, D. Cai, and Z. Ding, "Energy-efficient design for a NOMA assisted STAR-RIS network with deep reinforcement learning," *arXiv preprint arXiv:2111.15464*, Nov. 2021.

[101] K. Zhang, S. McLeod, M. Lee, and J. Xiao, "Continuous reinforcement learning to adapt multi-objective optimization online for robot motion," *Int J Adv Robot Syst.*, vol. 17, no. 2, 2020.

[102] K. Lobos-Tsunekawa, F. Leiva, and J. Ruiz-del Solar, "Visual navigation for biped humanoid robots using deep reinforcement learning," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3247–3254, 2018.

[103] Z. M. Kakish, K. Elamvazhuthi, and S. Berman, "Using reinforcement learning to herd a robotic swarm to a target distribution," *arXiv preprint arXiv:2006.15807*, 2020.

[104] A. K. Lakshmanan, R. E. Mohan, B. Ramalingam, A. V. Le, P. Veerajagadeshwar, K. Tiwari, and M. Ilyas, "Complete coverage path planning using reinforcement learning for tetromino based cleaning and maintenance robot," *Autom. Constr.*, vol. 112, p. 103078, 2020.

[105] J. James, W. Yu, and J. Gu, "Online vehicle routing with neural combinatorial optimization and deep reinforcement learning," *IEEE trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3806–3817, 2019.

[106] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Enhancing the fuel-economy of V2I-assisted autonomous driving: A reinforcement learning approach," *IEEE Trans.*

Veh. Technol., 2020.

[107] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun*, vol. 36, no. 9, pp. 2059–2070, 2018.

[108] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-UAV assisted wireless networks: A machine learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7957–7969, 2019.

[109] M. Chen, W. Saad, and C. Yin, "Liquid state machine learning for resource and cache management in lte-u unmanned aerial vehicle (uav) networks," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 3, pp. 1504–1517, 2019.

[110] Y. Wang, M. Chen, Z. Yang, T. Luo, and W. Saad, "Deep learning for optimal deployment of UAVs with visible light communications," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 11, pp. 7049–7063, 2020.

[111] Y. Qian, J. Wu, R. Wang, F. Zhu, and W. Zhang, "Survey on reinforcement learning applications in communication networks," *J. Commun. Info. Netw.*, 2019.

[112] J. Wang, S. Guan, C. Jiang, D. Alanis, Y. Ren, and L. Hanzo, "Network association in machine-learning aided cognitive radar and communication co-design," *IEEE IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2322–2336, 2019.

[113] M. A. Kishk and M.-S. Alouini, "Exploiting randomly located blockages for large-scale deployment of intelligent surfaces," *IEEE J. Sel. Areas Commun.s*, vol. 39, no. 4, pp. 1043–1056, 2021.

[114] Q. Bie, Y. Liu, Y. Wang, X. Zhao, and X. Y. Zhang, "Deployment optimization of reconfigurable intelligent surface for relay systems," *IEEE Trans. Green Commun.*, vol. 6, no. 1, pp. 221–233, 2022.

[115] B. Xu, T. Zhou, T. Xu, and Y. Wang, "Reconfigurable intelligent surface configuration and deployment in three-dimensional scenarios," in *IEEE ICC Workshops*, 2021, pp. 1–6.

[116] Z. Li, H. Hu, J. Zhang, and J. Zhang, "Enhancing indoor mmwave wireless coverage: Small-cell densification or reconfigurable intelligent surfaces deployment?" *IEEE Wirel. Commun. Lett.*, vol. 10, no. 11, pp. 2547–2551, 2021.

[117] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Commun. Magaz.*, vol. 58, no. 6, pp. 46–51, June. 2020.

[118] D. Chen, C. S. Hong, L. Wang, Y. Zha, Y. Zhang, X. Liu, and Z. Han, "Matching theory based low-latency scheme for multi-task federated learning in mec networks," *IEEE Internet Things J.*, Jan. 2021, doi:10.1109/JIOT.2021.3053283.

[119] A. M. Elbir and S. Coleri, "Federated learning for channel estimation in conventional and irs-assisted massive MIMO," *arXiv preprint arXiv:2008.10846*, 2020.

[120] D. Ma, L. Li, H. Ren, D. Wang, X. Li, and Z. Han, "Distributed rate optimization for intelligent reflecting surface with federated learning," in *Proc. IEEE ICC Workshops, Dublin, Ireland*, 2020, doi:10.1109/ICCWorkshops49005.2020.9145388.

[121] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. on Wirel. Commun.*, vol. 20, no. 1, pp. 269–283, Oct. 2021.

[122] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. on Wirel. Commun.*, Nov. 2020, doi:10.1109/TWC.2020.3037554.

[123] Z. Wang, J. Qiu, Y. Zhou, Y. Shi, L. Fu, W. Chen, and K. B. Lataief, "Federated learning via intelligent reflecting surface," *arXiv preprint arXiv:2011.05051*, 2020.

[124] Y. Huang, C. Xu, C. Zhang, M. Hua, and Z. Zhang, "An overview of intelligent wireless communications using deep reinforcement learning," *J. Commun. Info. Net.*, vol. 4, no. 2, pp. 15–29, June. 2019.

[125] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S.-L. Kim, and M. Debbah, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 796–819, May. 2021.

[126] N. Heess, D. TB, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. Eslami *et al.*, "Emergence of locomotion behaviours in rich environments," *arXiv preprint arXiv:1707.02286*, Jul. 2017.

[127] H. Zhang, N. Yang, W. Huangfu, K. Long, and V. C. M. Leung, "Power control

based on deep reinforcement learning for spectrum sharing," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 6, pp. 4209–4219, Mar. 2020.

[128] J. Ge, Y.-C. Liang, J. Joung, and S. Sun, "Deep reinforcement learning for distributed dynamic MISO downlink-beamforming coordination," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6070–6085, 2020.

[129] V. Arun and H. Balakrishnan, "RFocus: Beamforming using thousands of passive antennas," in *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, Santa Clara, CA, USA, Feb. 2020, pp. 1047–1061.

[130] M. Hua, L. Yang, Q. Wu, C. Pan, C. Li, and A. Lee Swindlehurst, "UAV-assisted intelligent reflecting surface symbiotic radio system," *IEEE Trans. Wirel. Commun.*, 2021, early access, doi:10.1109/TWC.2021.3070014.

[131] Z. Li, W. Chen, Q. Wu, H. Cao, K. Wang, and J. Li, "Robust beamforming design and time allocation for IRS-assisted wireless powered communication networks," *arXiv preprint arXiv:2105.06226*, 2021.

[132] S. Liu, M. Lei, and M.-J. Zhao, "Deep learning based channel estimation for intelligent reflecting surface aided MISO-OFDM systems," in *IEEE 92nd VTC2020-Fall*, Victoria, BC, Canada, Feb. 2021.

[133] X. Yu, V. Jamali, D. Xu, D. W. K. Ng, and R. Schober, "Smart and reconfigurable wireless communications: From IRS modeling to algorithm design," *arXiv preprint arXiv:2103.07046*, 2021.

[134] M. R. Usman, A. Khan, M. A. Usman, Y. S. Jang, and S. Y. Shin, "On the performance of perfect and imperfect SIC in downlink non orthogonal multiple access (NOMA)," in *IEEE ICSGTEIS*, 2016, pp. 102–106.

[135] J.-M. Kang, I.-M. Kim, and C.-J. Chun, "Deep learning-based mimo-noma with imperfect sic decoding," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3414–3417, 2020.

[136] Y. Song, M. R. Khandaker, F. Tariq, K.-K. Wong, and A. Toding, "Truly intelligent reflecting surface-aided secure communication using deep learning," *arXiv preprint arXiv:2004.03056*, 2020.

[137] E. Phaisangittisagul, "An analysis of the regularization between L2 and dropout

in single hidden layer neural network," in *2016 7th ISMS*, Bangkok, Thailand, Jan. 2016, pp. 174–179.

[138] O. Sidelnikov, A. Redyuk, and S. Sygletos, "Equalization performance and complexity analysis of dynamic deep neural networks in long haul transmission systems," *Opt. Express*, vol. 26, no. 25, pp. 32 765–32 776, 2018.

[139] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8577–8588, Oct. 2019.

[140] J. Lin, Y. Zout, X. Dong, S. Gong, D. T. Hoang, and D. Niyato, "Deep reinforcement learning for robust beamforming in IRS-assisted wireless communications," in *IEEE GLOBECOM*, Taipei, Taiwan, Dec. 2020, pp. 1–6.

[141] 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.

[142] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the brownian motion," *Phys. Rev.*, vol. 36, no. 5, p. 823, 1930.

[143] C. Colas, O. Sigaud, and P.-Y. Oudeyer, "Gep-pg: Decoupling exploration and exploitation in deep reinforcement learning algorithms," *arXiv preprint arXiv:1802.05054*, 2018.

[144] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, 2018, pp. 1–2.

[145] R. I.-R. P.1238-10, "Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 450 GHz," 2019.

[146] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[147] G. Matheron, N. Perrin, and O. Sigaud, "The problem with DDPG: understanding failures in deterministic environments with sparse rewards," *arXiv preprint arXiv:1911.11679*, 2019.

[148] R. Zhong, X. Liu, Y. Liu, Y. Chen, and Z. Han, "Mobile reconfigurable intelligent surfaces for NOMA networks: Federated learning approaches," *arXiv preprint arXiv:2105.09462*, 2021.

[149] N. Cao, Y. Chen, and Z. Yang, "Secrecy outage probability with randomly moving interferers in nakagami-m fading," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 76–79, Oct. 2018.

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

[151] R. I.-R. M.2135-1, "Guidelines for evaluation of radio interface technologies for IMT-advanced," 2009.

[152] A. Majercik, C. Crassin, P. Shirley, and M. McGuire, "A ray-box intersection algorithm and efficient dynamic voxel rendering," *Journal of Computer Graphics Techniques Vol*, vol. 7, no. 3, pp. 66–81, Jun. 2018.

[153] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1502–1517, Mar. 2018.

[154] Z. Na, J. Wang, C. Liu, M. Guan, and Z. Gao, "Join trajectory optimization and communication design for UAV-enabled OFDM networks," *Ad Hoc Networks*, vol. 98, pp. 1–10, Mar. 2020.

[155] T.-R. Lin, D. Penney, M. Pedram, and L. Chen, "A deep reinforcement learning framework for architectural exploration: A routerless NoC case study," in *Proc. IEEE HPCA, San Diego, CA, USA*, Feb. 2020, pp. 99–110.

[156] S. Zhang, H. Peng, S. Nageshrao, and E. Tseng, "Discretionary lane change decision making using reinforcement learning with model-based exploration," in *Proc.18th IEEE ICMLA, Boca Raton, FL, USA*, Dec. 2019, pp. 844–850.

[157] F. S. Melo, "Convergence of Q-learning: A simple proof," *Institute Of Systems and Robotics, Tech. Rep*, pp. 1–4, 2001.

[158] M. Neunert, A. Abdolmaleki, M. Wulfmeier, T. Lampe, T. Springenberg, R. Hafner, F. Romano, J. Buchli, N. Heess, and M. Riedmiller, "Continuous-discrete reinforce-

ment learning for hybrid control in robotics," in *Proceedings of the CoRL*, vol. 100, 30 Oct–01 Nov 2020, pp. 735–751.

[159] O. Delalleau, M. Peter, E. Alonso, and A. Logut, "Discrete and continuous action representation for practical rl in video games," *arXiv preprint arXiv:1912.11077*, 2019.

[160] B. Li, H. Tang, Y. Zheng, J. Hao, P. Li, Z. Wang, Z. Meng, and L. Wang, "Hyar: Addressing discrete-continuous action reinforcement learning via hybrid action representation," *arXiv preprint arXiv:2109.05490*, 2021.

[161] "Study on 3d channel model for lte (3gpp tr 36.873 release 12)," *3rd Generation Partnership Project*, Jan. 2018.

[162] Y. Li, A. H. Aghvami, and Y. Deng, "Joint resource block and beamforming optimization for cellular-connected UAV networks: A hybrid D3QN-DDPG approach," *arXiv preprint arXiv:2102.13222*, 2021.

[163] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 2, pp. 1226–1252, Mar. 2021.

[164] I. Abraham, A. Prabhakar, and T. D. Murphey, "An ergodic measure for active learning from equilibrium," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 917–931, July 2021.

[165] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

[166] C. Li, J. Xia, F. Liu, D. Li, L. Fan, G. K. Karagiannidis, and A. Nallanathan, "Dynamic offloading for multiuser muti-CAP MEC networks: A deep reinforcement learning approach," *IEEE Trans. Vehi. Technol.*, vol. 70, no. 3, pp. 2922–2927, Mar. 2021.

[167] M. Babaeizadeh, I. Frosio, S. Tyree, J. Clemons, and J. Kautz, "Reinforcement learning through asynchronous advantage actor-critic on a gpu," *arXiv preprint arXiv:1611.06256*, 2016.

[168] I. Chafaa, R. Negrel, E. Veronica Belmega, and M. Debbah, "Self-supervised deep

learning for mmwave beam steering exploiting sub-6 GHz channels," *IEEE Trans. Wirel. Commun.*, Early Access, May. 2022.

[169] M. Najafi, V. Jamali, R. Schober, and H. V. Poor, "Physics-based modeling and scalable optimization of large intelligent reflecting surfaces," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2673–2691, Apri 2021.

[170] A. Abels, D. Roijers, T. Lenaerts, A. Nowé, and D. Steckelmacher, "Dynamic weights in multi-objective deep reinforcement learning," in *International Conference on Machine Learning*. PMLR, Jun 2019, pp. 11–20.

[171] X. Yu, V. Jamali, D. Xu, D. W. K. Ng, and R. Schober, "Smart and reconfigurable wireless communications: From IRS modeling to algorithm design," *IEEE Wirel. Commun.*, vol. 28, no. 6, pp. 118–125, Dec. 2021.

[172] J. Zuo, Y. Liu, Z. Ding, L. Song, and H. Vincent Poor, "Joint design for simultaneously transmitting and reflecting (STAR) RIS assisted NOMA systems," *IEEE Trans. Wirel. Commun.*, pp. 1–1, 2022.

[173] Z. Yang, W. Xu, C. Pan, Y. Pan, and M. Chen, "On the optimality of power allocation for NOMA downlinks with individual QoS constraints," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1649–1652, July, 2017.

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

[175] N. Naughton, J. Sun, A. Tekinalp, T. Parthasarathy, G. Chowdhary, and M. Gazzola, "Elastica: A compliant mechanics environment for soft robotic control," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3389–3396, 2021.

[176] T. Schaul, D. Horgan, K. Gregor, and D. Silver, "Universal value function approximators," in *International conference on machine learning*. PMLR, 2015, pp. 1312–1320.

[177] A. Elzanaty, A. Guerra, F. Guidi, and M.-S. Alouini, "Reconfigurable intelligent surfaces for localization: Position and orientation error bounds," *IEEE Trans. Signal Process.*, vol. 69, pp. 5386–5402, 2021.