Artificial Intelligence for Non-Orthogonal Multiple Access Networks

by

Yixuan Zou

Doctor of Philosophy

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

June 2022

TO MY FAMILY

Statement of originality

I, Yixuan Zou, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Yixuan Zou

Date: 14 June 2022

Details of collaboration and publications:

- Journal Papers

  1. Y. Xu, T. Zhang, **Y. Zou**, Y. Liu, "Reconfigurable Intelligence Surface Aided UAV-MEC Systems With NOMA", *IEEE Communication Letter*, 2022; (accepted for publication).

  2. **Y. Zou**, Y. Liu, X. Liu, X. Mu, X. Zhang, C. Yuen, "Comparisons between DL and DRL on the Optimization of RIS-assisted NOMA Networks", *IEEE Transactions on Wireless Communications*; (under revision).

  3. J. Chen, Z. Ma, **Y. Zou**, Y. Liu, J. Jia, X. Wang, "Joint Active and Passive Beamforming for EE Optimization in STAR-RIS assisted CoMP Systems", *IEEE Transactions on Communications*; (under revision)

  4. L. Guo, J. Jia, **Y. Zou**, Y. Liu, J. Chen, X. Wang, "Resource Allocation for Multiple RISs Assisted NOMA Empowered D2D Communication: A MAMP-DQN Approach", *IEEE Transactions on Vehicular Technology*; (under review).

  5. **Y. Zou**, W. Yi, Y. Liu, "Meta-Reinforcement Learning for Adaptive NGMA Networks with Energy Limitations", *IEEE Transactions on Wireless Communications*; (to be submitted).

- Conference Papers

  1. **Y. Zou**, Y. Liu, K. Han, X. Liu, and K. K. Chai, "Meta-learning for RIS-assisted Non-Orthogonal Multiple Access Networks", in *Proc. IEEE Global Communications Conf. (GLOBECOM'21)*, Madrid, Spain, December 2021.

  2. **Y. Zou**, Z. Qin and Y. Liu, "Joint User Activity and Data Detection in Grant-Free NOMA using Generative Neural Networks," in *Proc. IEEE Int.*

Communications Conf. (ICC'21), Montreal, Canada, June 2021.

3. **Y. Zou**, W. Yi, X. Xu, Y. Liu, "Adaptive NGMA Scheme for Energy-limited Networks: A Deep Reinforcement Learning Approach", in *Proc. IEEE Global Communications Conf. (GLOBECOM'22)*, Rio de Janeiro, Brazil, December 2022; (under review)

4. J. Chen, Z. Ma, **Y. Zou**, J. Jia, X. Wang, "DRL-based Energy Efficient Resource Allocation for STAR-RIS Assisted Coordinated Multi-cell Networks", in *Proc. IEEE Global Communications Conf. (GLOBECOM'22)*, Rio de Janeiro, Brazil, December 2022; (under review)

Acknowledgments

Foremost, I would like to express my deepest gratitude to my supervisors, Dr. Yuanwei Liu, Dr. Micheal Chai, and Prof. Ioannis Patras, for their kind supports in my Ph.D research. Special thanks to Dr. Yuanwei Liu, for his continuous and valuable guidance throughout my study. His levels of patience, knowledge, and ingenuity are something I will always aspire to and will have a profound effect on my future career. Without him, I would not be able to complete my Ph.D research. It is my distinct honour and privilege to have such a gracious supervisor. The past three years have been an extraordinary experience.

I am extremely thankful to my previous supervisor, Dr. Zhijin Qin, who patiently guided and supported me through the first year of my Ph.D. When I was a naive and clueless fresher, it was Dr. Qin who patiently taught me the necessary skills and the fundamental knowledge that a Ph.D student demands. She is the most amazing female scientist I have ever met and always will be.

I would like to thank all my collaborators: Dr. Wenqiang Yi, Dr. Xidong Mu, Dr. Xiao Liu, Dr. Yuen Chau, Dr. Kaifeng Han, Prof. Xiaodong Xu for their helpful suggestions and comments on my research.

I would also like to thank Yanling Hao, Chao Zhang, Ruikang Zhong, Xiaoxia Xu, Zhaolin Wang, Dr. Zhong Yang, Dr. Tianwei Hou, Zhixiong Chen, Xinyu Gao, Jiaqi Xu, Zelin Ji, Huiqiang Xie, Yiyu Guo, and all my colleges and friends at Queen Mary University of London, for their constant encouragement and kind help. I truly had extraordinary memories of my Ph.D life and study. Also, I would like to extend my thanks to my precious friends, Wenhua Zhou and Poro, who encouraged, comforted, and motivated me whenever I needed them.

Last and most importantly, I would like to express my endless gratitude to my beloved family, especially my parents, who always unconditionally support me and encourage me at any time.

Abstract

Massive connectivity, ultra-low latency, and high data rate are some of the fundamental requirements of the upcoming sixth-generation (6G) wireless networks. In this regard, non-orthogonal multiple access (NOMA) has been widely envisioned as a promising candidate for 6G due to its potential of achieving high spectral efficiency. By multiplexing the signals in the power or the code domain, NOMA allows multiple users to be served with the same orthogonal resources, such as frequency and time, hence outperforming the conventional orthogonal multiple access (OMA) systems with a significant spectral efficiency gain. The promising advantages of NOMA cannot be realized without proper optimization designs, of which the complexity escalates in the context of massive connectivity. As a remedy, artificial intelligence (AI) is capable of performing high-dimensional optimization at a lower computational complexity compared to the conventional iterative approaches. Hence, this thesis attempts to utilize AI technologies, including deep learning (DL) and deep reinforcement learning (DRL), to design systematic treatments for NOMA, from the uplink active user detection to the resource allocation in adaptive next-generation multiple access (NGMA) networks, to its combination with reconfigurable intelligent surfaces (RISs), as well as its application in multi-RIS aided device-to-device (D2D) networks.

First, this thesis investigates the application of generative neural networks for joint user activity and data detection in uplink NOMA networks. A generative neural network-enabled multi-user detection (MUD) framework is proposed, which outputs signal reconstructions in a fixed and small number of steps with low error rates, based on a low-complexity neural network. Moreover, a non-iterative sparsity estimator is provided to realize sparsity-blind MUD and is compatible with most existing MUD algorithms.

Second, to maximize the long-term sum data rate of NGMA networks with energy limi-

tations, DRL is employed to jointly design beamformers, power allocations, and user clustering strategies. To transform the non-trivial mixed-integer problem, a spatial correlation-based user clustering approach is proposed, which achieves higher sum rates compared to the existing channel condition-based clustering approach. To solve the formulated problem, the trust region policy optimization (TRPO) algorithm is employed, which demonstrates robust convergence under large learning rates and realizes a fast and stable training process.

Third, the integration of NOMA and RIS is examined, where the sum rate maximization performance of DL and DRL are investigated and compared from both short-term and long-term prospects. By utilizing model-agnostic-meta-learning (MAML), the DL method benefits from a low complexity network and a fast convergence rate. The DRL method, on the other hand, demonstrates superior sum rate performance, especially in the long term.

Fourth, this thesis addresses the sum rate maximization problem in multi-RIS assisted NOMA empowered D2D networks. The long-term dynamic optimization problem is reformulated into a Markov game (MG) and a multi-agent deep reinforcement learning (MADRL)-based framework is proposed to jointly learn sub-channel assignments, power allocations, and phase shifts, in a centralized training and decentralized execution (CTDE) manner. Furthermore, the mixed-integer action space is directly addressed by adopting multi-pass deep Q networks (MP-DQNs).

Table of Contents

Acknowledgments **iv**

Abstract **vi**

Table of Contents **viii**

List of Figures **xiv**

List of Tables **xvii**

List of Abbreviations **xviii**

1 Introduction **1**

1.1 Background 1

1.1.1 On the road to 6G 1

1.1.2 Evolution of Multiple Access Techniques 2

1.1.3 Artificial Intelligence for 6G Multiple Access 5

1.2 Motivation and Contributions 7

1.2.1 Joint User Activity and Data Detection in Grant-Free NOMA using Generative Neural Networks 8

1.2.2 Adaptive NGMA Scheme for Energy-limited Networks: A Deep Reinforcement Learning Approach 9

1.2.3 Comparisons between DL and DRL on the Optimization of RIS-assisted NOMA systems . . . 11
1.2.4 Multi-Agent Resource Allocation in NOMA-Enhanced Multi-RIS Aided D2D Networks . . . 12
1.3 Related Works . . . 13
1.3.1 NOMA Systems . . . 13
1.3.2 AI-empowered Networks . . . 15
1.3.3 AI-empower NOMA systems . . . 16
1.4 Dissertation Organization . . . 18

2 Fundamental Concepts **19**
2.1 Fundamental principles of NOMA . . . 19
2.1.1 Key Technologies of NOMA . . . 20
2.1.2 Mathematical Formulation of NOMA . . . 21
2.1.3 MIMO-NOMA . . . 23
2.2 Artificial Intelligence . . . 24
2.2.1 Deep Learning . . . 25
2.2.2 Deep Reinforcement Learning . . . 26
2.2.3 Meta-learning . . . 28
2.3 Related Technologies . . . 29
2.3.1 Compressed Sensing . . . 29
2.3.2 Spatial Division Multiple Access . . . 31
2.3.3 Reconfigurable Intelligent Surface . . . 31

3 Joint User Activity and Data Detection in Grant-Free NOMA using Generative Neural Networks **34**
3.1 Introduction . . . 34
3.2 System Model . . . 36
3.2.1 Frame-Wise Joint Sparsity Model . . . 37
3.2.2 Problem Formulation . . . 37

3.2.3 Performance Metrics . . . 38
3.3 Generative Networks for Multi-User Detection (GenMUD) . . . 39
3.3.1 Problem Reformulation . . . 40
3.3.2 GenMUD Framework . . . 43
3.3.3 Network Architecture . . . 43
3.3.4 Complexity Analysis . . . 45
3.3.5 Sparsity Estimator . . . 46
3.4 Simulations . . . 47
3.4.1 Benchmark Methods . . . 48
3.4.2 MSE Performance . . . 48
3.4.3 SER Performance . . . 49
3.4.4 Positive Detection Rate (P_d) Performance . . . 50
3.4.5 False Alarm Rate (P_{fa}) Performance . . . 50
3.4.6 Sparsity Estimation Performance . . . 52
3.5 Summary . . . 54

4 Adaptive NGMA Scheme for Energy-limited Networks: A Deep Reinforcement Learning Approach 55

4.1 Introduction . . . 55
4.2 Network Model . . . 56
4.2.1 Spatial Model . . . 56
4.2.2 Channel Model . . . 57
4.2.3 Adaptive NGMA . . . 58
4.2.4 Signal Model . . . 60
4.2.5 Problem Formulation . . . 62
4.3 DRL-based Resource Allocation for Adaptive NGMA . . . 64
4.3.1 Problem Reformulation . . . 64
4.3.2 Markov Decision Process (MDP) . . . 66
4.3.3 TRPO Learning Algorithm . . . 68

4.3.4 Complexity Analysis . . . 71
4.4 Numerical Results . . . 71
4.4.1 Simulation Settings . . . 71
4.4.2 Baseline Methods . . . 72
4.4.3 Algorithm Convergence . . . 72
4.4.4 Impact of Number of Antenna . . . 75
4.5 Summary . . . 76

5 Comparisons between DL and DRL on the Optimization of RIS-assisted NOMA Systems 77

5.1 Introduction . . . 77
5.2 Network Model . . . 79
5.2.1 System Model . . . 79
5.2.2 Channel Model . . . 80
5.2.3 PD-NOMA Signal Model . . . 80
5.3 Short-term Optimization Problem . . . 84
5.3.1 Short-term Problem Formulation . . . 84
5.3.2 DL-based Resource Allocation Scheme . . . 85
5.3.3 MAML-based Training Algorithm . . . 86
5.3.4 Convergence Analysis . . . 89
5.3.5 Complexity Analysis . . . 90
5.4 Long-term Optimization Problem . . . 91
5.4.1 Long-term Problem Formulation . . . 91
5.4.2 DRL-based Resource Allocation Scheme . . . 93
5.4.3 DDPG-based Training Algorithm . . . 94
5.4.4 Convergence Analysis . . . 97
5.4.5 Complexity Analysis . . . 98
5.5 Numerical Results . . . 99
5.5.1 Short-term Optimization with DL . . . 99

5.5.2 Long-term Optimization with DRL . . . 102
5.5.3 DL versus DRL . . . 104
5.6 Summary . . . 109

6 Multi-Agent Resource Allocation in NOMA-Enhanced Multi-RIS Aided D2D Networks 110

6.1 Introduction . . . 110
6.2 Network Model . . . 112
6.2.1 System Model . . . 112
6.2.2 Channel Model . . . 113
6.2.3 Signal Model . . . 114
6.2.4 Problem Formulation . . . 117
6.3 MAHA-DRL Algorithm for Resource Allocation . . . 119
6.3.1 Multi-Agent DRL . . . 119
6.3.2 MAHA-DRL Algorithm . . . 122
6.3.3 Complexity Analysis . . . 128
6.4 Numerical Results . . . 128
6.4.1 Algorithm Convergence . . . 129
6.4.2 Sum Rate versus Number of D2D Groups . . . 131
6.4.3 Sum Rate versus Number of REs . . . 132
6.4.4 Sum Rate versus Number of RISs . . . 133
6.5 Summary . . . 134

7 Conclusions and Future Works 135

7.1 Contributions and Insights . . . 135
7.2 Future Works . . . 138
7.2.1 Extensions of Current Works . . . 138
7.2.2 Promising Future Directions on AI-aided NOMA Systems . . . 139

Appendix A Proof in Chapter 5 141

A.1 Proof of Proposition 1 141
References 143

List of Figures

1.1 Visions for 6G wireless networks. 2

1.2 Evolution of wireless communication. 3

2.1 Block diagram of PD-NOMA. 22

2.2 Illustration of RIS-enabled wireless communications. 32

3.1 Illustration of a typical frame-based uplink grant-free CD-NOMA system. 35

3.2 Architecture of the developed neural network. 44

3.3 MSE versus SNR based on OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD. 48

3.4 SER versus SNR based on OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD under $S = 40$ active users and $M = 100$ sub-carriers. . . 49

3.5 Positive detection rate (P_d) versus SNR based on OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD. 50

3.6 False alarm probability (P_{fa}) versus SNR based on OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD. 51

3.7 SER comparison of OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD versus different number of active users under 6 dB SNR. 51

3.8 Normalized error (E_n) versus number of time slots of the proposed sparsity estimator under different numbers of sub-carriers and SNRs. 52

3.9 Detection probability (P_d) and false alarm probability (P_{fa}) versus SNR of the proposed GenMUD with known sparsity and estimated sparsity. . . 53

4.1 Illustration of the proposed adaptive NGMA-MISO downlink network. Users are grouped into clusters, where the users in the multi-user clusters employ SIC for decoding. 57
4.2 Block diagrams of the transmitter and the receiver in different transmission schemes, where s_k is the intended signal, p_k is the power allocation, and $\overline{w}_k$ is the normalized beamforming vector of user k: a) A SDMA scheme; b) A PD-NOMA scheme; c) The proposed NGMA scheme. 59
4.3 Flow diagram of the proposed TRPO-based resource allocation scheme for apative NGMA systems. 70
4.4 Episodic reward versus the number of episodes under different network architecture and batch sizes. 73
4.5 Episodic reward versus the number of episodes in SDMA, PD-NOMA, and NGMA systems under $K = 4$ users and $N = 4$ antennas. 74
4.6 Sum rate versus the number of antennas under SDMA, PD-NOMA, NGMA* and NGMA schemes with 4 users. 75

5.1 Illustration of the downlink RIS-assisted MISO-PDNOMA system. 79
5.2 Illustration of the MAML-based training framework. 87
5.3 Illustration of the DDPG framework. 94
5.4 Training loss versus the number of training iterations for different batch sizes and learning rates. 100
5.5 Sum rate versus the number of reflecting elements N for PD-NOMA and OMA cases, given 20 dBm BS transmit power. 101
5.6 Sum rate versus total transmit power at BS for PD-NOMA and OMA cases, given $N = 16$ reflecting elements. 101
5.7 Episode rewards of the DRL algorithm versus the number of training episodes, under different batch sizes and critic learning rates. 103
5.8 sum rate versus the number of RIS elements, under $P_{max} = 10$ dBm BS power and $M = 4$ BS antennas. 104

5.9 Sum rate of DL and DRL versus BS transmit power in PD-NOMA and OMA systems, for the short-term optimization problem. 105
5.10 Sum rate of DL and DRL versus the number of users under the QoS-based or channel-based clustering schemes, when solving the short-term optimization problem. 106
5.11 Sum rate of DL and DRL versus total transmit power when solving the long-term optimization problem in PD-NOMA systems. 107

6.1 Illustration of the PD-NOMA-enhanced multi-RIS aided D2D network underlaying cellular networks. 112
6.2 Flow diagram of MAHA-DRL algorithm based on CTDE training scheme. 123
6.3 Episodic reward versus the number of episodes under PD-NOMA transmission scheme. 130
6.4 Episodic reward versus the number of episodes under PD-NOMA-based D2D networks and OMA-based D2D networks, with or without the assistance of RISs. 130
6.5 Sum rate versus the number of D2D groups under PD-NOMA-based D2D networks and OMA-based D2D networks, with or without the assistance of RISs. 131
6.6 Sum rate versus the number of REs under PD-NOMA-based D2D networks and OMA-based D2D networks, with 2 RISs or no RIS. Each RIS consists of $Q = 2$ or $Q = 4$ sub-surfaces. 132
6.7 Sum rate versus the number of RISs under PD-NOMA-based D2D networks and OMA-based D2D networks, based on $D = 2$ or $D = 3$ D2D groups. 133

List of Tables

3-A List of main notations. 36
3-B Network and algorithm configurations. 47

4-A List of main notations. 56
4-B Network and algorithm configurations. 71

5-A List of main notations. 79
5-B DL simulation configurations. 98
5-C DDPG simulation configurations. 102

6-A List of main notations. 112
6-B Network and algorithm configurations. 129

List of Abbreviations

ADMM	Alternative Direction Method of Multipliers
AI	Artificial Intelligence
AR	Augmented Reality
AWGN	Additive White Gaussian Noise
BS	Base Station
CDMA	Code Division Multiple Access
CD-NOMA	Code-domain NOMA
CS	Compressive Sensing
CSI	Channel State Information
CoMP	Coordinated Multipoint
CUE	Cellular User
D2D	Device-to-Device
DDPG	Deep Deterministic Policy Gradient
DL	Deep Learning
DR	D2D Receiver
DRL	Deep Reinforcement Learning
DT	D2D Transmitter
DPC	Dirty-Paper Coding
DPG	Deterministic Policy Gradient
DQN	Deep Q-Learning

FDMA	Frequency Division Multiple Access
GAE	Generalized Advantage Estimator
HDTV	High-Definition Television
i.i.d.	Independent and Identically Distributed
Kbps	Kilobits per Second
KL	Kullback–Leibler
LoS	Line-of-Sight
LSTM	Long Short-Term Memory
MA	Multiple Access
MADRL	Multi-Agent Deep Reinforcement Learning
MAML	Model Agnostic Meta Learning
Mbps	Megabits per Second
MDP	Markov Decision Process
MG	Markov Game
MMS	Multimedia Messaging Service
mmWave	Millimeter Wave
MIMO	Multi-Input and Multi-Output
MISO	Multiple-Input Single-Output
MU	Mobile User
MUD	Multi-User Detection
NGMA	Next Generation Multiple Access
NOMA	Non-Orthogonal Multiple Access
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple access
OMA	Orthogonal Multiple Access
OMP	Orthogonal Matching Pursuit
PD-NOMA	Power-domain NOMA
QoS	Quality of Service

ReLU	Rectified Linear Unit
RIP	Restricted Isometry Property
RIS	Reconfigurable Intelligence Surface
RSMA	Rate-Splitting Multiple Access
SC	Superposition Coding
SCMA	Sparse Code Multiple Access
SDMA	Spatial Division Multiple Access
SIC	Successive Interference Cancellation
SISO	Single Input Single Output
SINR	Signal-to-Interference-Plus-Noise Ratio
SNR	Signal-to-Noise Ratio
TDMA	Time Division Multiple Access
THz	Terahertz
TRPO	Trust Region Policy Optimization
TS	Time Slot
UAV	Unmanned Aerial Vehicle
VR	Virtual Reality
ZF	Zero-Forcing
1D	One-Dimensional
1G	First-Generation
2G	Second-Generation
3G	Third-Generation
4G	Fourth-Generation
5G	Fifth-Generation
6G	Sixth-Generation

Chapter 1

Introduction

In this chapter, an overview of the sixth-generation (6G) wireless networks is presented, followed by the motivations for studying artificial intelligence (AI)-empowered non-orthogonal multiple access (NOMA) networks. Then, the main contributions of this thesis are outlined and the related works of this thesis are discussed. Finally, the organization of this thesis is presented.

1.1 Background

1.1.1 On the road to 6G

Since 2020, the fifth-generation (5G) wireless communication networks are being standardized and on their way to being deployed worldwide. As the next decades unfold, extremely rich multimedia applications (e.g., augmented reality (AR)/virtual reality (VR)), tactile/haptic-based communications, autonomous vehicles, super-smart city, and Internet of Everything are envisioned, whereas the requirements are yet to be fulfilled by 5G. In order to satisfy the future demands of the emerging applications, researchers in both academia and industry have been shifting their attentions to sixth generation (6G) wireless networks. 6G networks are expected to extend the capabilities of 5G networks to a brand new level, such as 10 times larger connectivity, 100 times higher

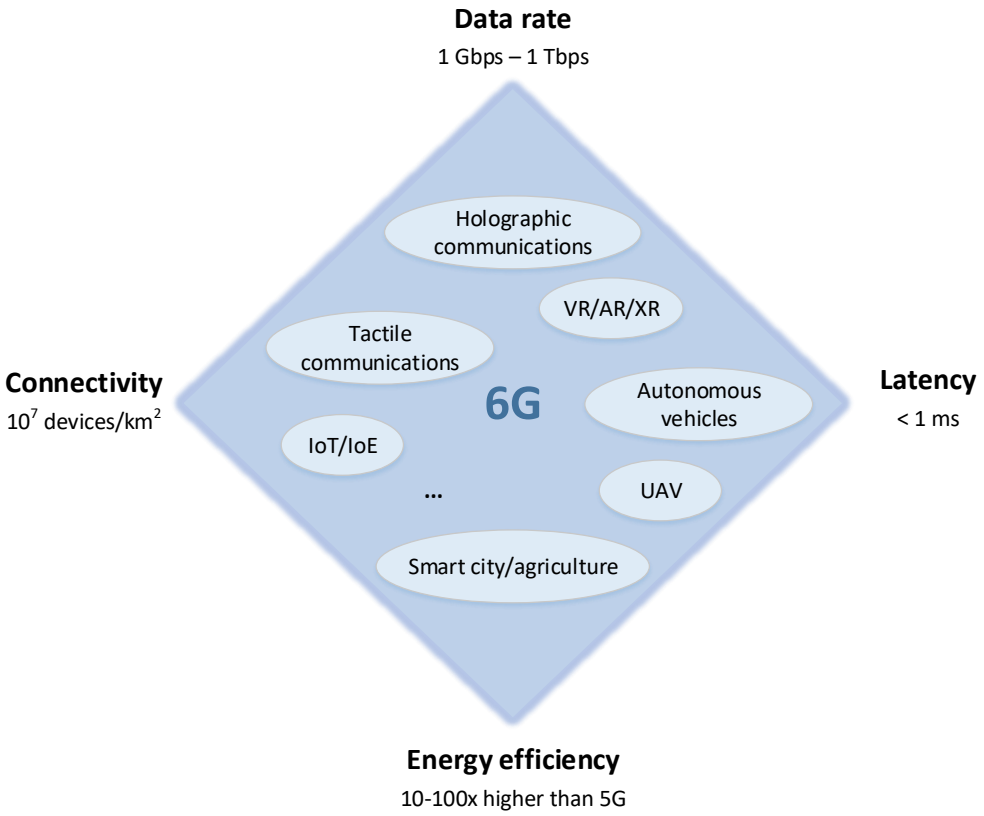


Figure 1.1: Visions for 6G wireless networks.

spectral efficiency, 10 - 100 times higher energy efficiency, terabit data rate, and sub-millisecond latency [1–3]. Fig. 1.1 depicts the visions for 6G wireless networks. Towards this direction, several key technologies such as massive multiple-input multiple-output (MIMO), reconfigurable intelligent surface (RIS), terahertz (THz) communications, coordinated multipoint (CoMP), unmanned aerial vehicle (UAV), compressive sensing (CS), AI, blockchain, and integrated sensing and communication (ISaC) have been envisioned as potential 6G technological enablers.

1.1.2 Evolution of Multiple Access Techniques

In addition to the aforementioned technologies, the multiple access (MA) technique is recognised as one of the most fundamental components in the physical layer, which continues to evolve with each generation of wireless networks and has a significant impact on the definition of technical features. A summary of the major milestones on the

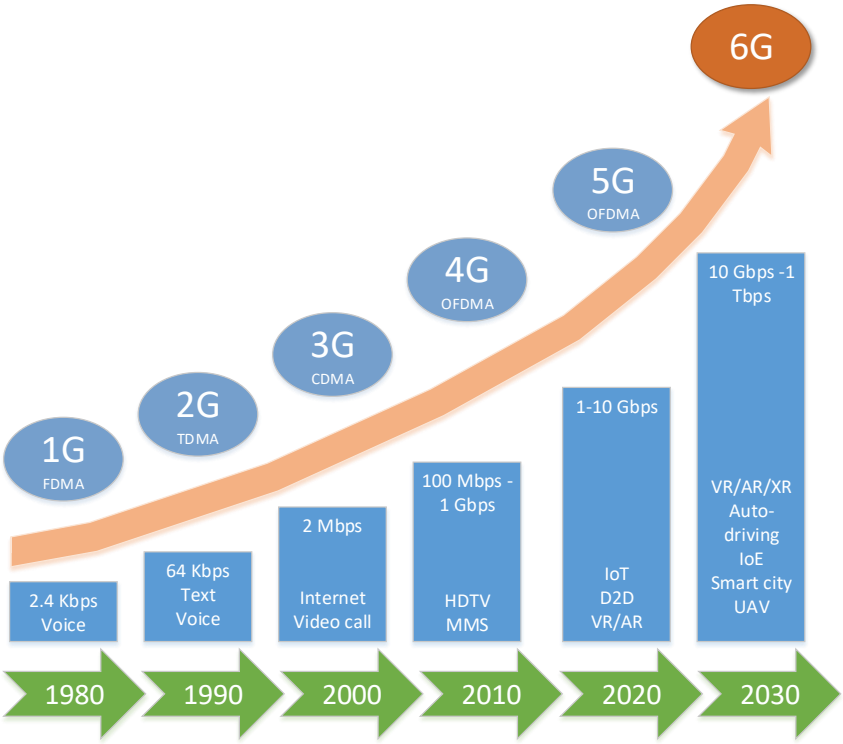


Figure 1.2: Evolution of wireless communication.

evolution of wireless communication is illustrated in Fig. 1.2.

Tracing back to the first-generation (1G) networks, which was introduced in the 1980s, an analog frequency modulation based technology known as frequency division multiple access (FDMA) was utilized to support voice services with a maximum data rate of 2.4 kilobits per second (Kbps). By dividing the available bandwidth into several non-overlapping sub-channels, FDMA can serve as many users as the number of sub-channels. However, since the channels are assigned permanently, spectrum is wasted when stations were idle. Moreover, to prevent interference, guard bands have to be inserted, leading to low spectrum efficiency.

In the 1990s, the second-generation (2G) networks shifted to digital modulation and adopted time division multiple access (TDMA) to achieve a peak data rate of 64 Kbps. In TDMA, multiple users share the same channel under different time slots. Signals are divided on a time basis and users transmit their signals in rapid succession, by utilizing

the allocated time slots. Hence, TDMA benefits from a higher spectrum efficiency and lower operational costs compared to FDMA. Moreover, services with different data rate requirements, such as voice and text messages, can be successfully achieved through time allocation. However, TDMA suffers from high synchronization overhead and time allocation complexity. Moreover, when switching between cells, the call might be disconnected due to fully occupied time slots in the next cell.

Around the year 2000, code division multiple access (CDMA) became the dominant MA standard for the third-generation (3G) communication networks [4]. By employing spread spectrum technology and orthogonal spreading codes, multiple users can share the whole bandwidth simultaneously without interference. Specifically, the 3G networks based on CDMA provided higher data transmission speeds of at least 2 Mbps, while supporting internet connections in addition to voice and text services. CDMA also benefit from the soft handoff feature, since the user can remain connected to both base stations (BS) when switching between cells. However, the near-far effect exists in CDMA systems and sophisticated power control schemes were required to overcome this problem.

In 2009, the fourth-generation (4G) networks were launched. In this generation, orthogonal frequency division multiple access (OFDMA) was dominantly adopted to support even higher data rates of at least 100 megabits per second (Mbps) [5]. The applications of 4G networks include video chat, Multimedia Messaging Service (MMS), and high-definition television (HDTV). OFDMA was developed based on orthogonal frequency division multiplexing (OFDM), which divides the whole bandwidth into orthogonal smaller bandwidths, known as sub-carriers, to eliminate mutual interference. In OFDMA, the sub-carriers are grouped into sub-channels, which are assigned to different users. The main limitations of this technique are that all sub-carriers have to be orthogonal to each other and accurate frequency synchronization between the transmitter and the receiver is required.

In the ongoing 5G networks, OFDMA is still the dominant MA technique. Meanwhile, a variety of potential MA techniques are also considered, including sparse code multiple

access (SCMA) [6], rate-splitting multiple access (RSMA) [7], as well as NOMA [8]. SCMA is a combination of OFDMA and CDMA with an additional restriction on sparse spreading codes for achieving low-complexity signal detection. RSMA split the messages into sub-massages, which are then combined and encoded into streams. Each stream may contain the message of one or more users for interference management purposes. Last but not least, NOMA utilizes superposition coding (SC) in the code or power domain at the transmitter to realize non-orthogonal signal multiplexing and to achieve significant capacity improvement. Both RSMA and power-domain NOMA (PD-NOMA) utilize SIC for signal decoding, where the main difference between PD-NOMA and RSMA can be explained as follows. In PD-NOMA, the involved users should be priorly ordered. Based on the predefined order, the signal of each user is successively decoded at the receiver employing SIC until its intended signal is decoded. However, for RSMA, there is no order predefined among users and the information of each user should be split into common and private parts at the transmitter. At the receiver, the common message is decoded before the private message employing SIC. As a result, user's signal should be first split and then combined in RSMA, while there is no such an operation in PD-NOMA.

Although NOMA has been widely investigated in 5G networks, its full potential has yet to be explored. To support 6G's massive connectivity, NOMA is becoming quality of service (QoS)-centric, focusing on not only data rates, but also latency and reliability. Moreover, as a highly compatible spectrum enhancement technology, NOMA can be flexibly integrated with numerous emerging 6G technologies and application scenarios [9], including RIS-empowered NOMA, THz-NOMA, AI-empowered NOMA, ISaC NOMA, NOMA in autonomous robotics networks, and NOMA in VR/AR multi-layer video transmission.

1.1.3 Artificial Intelligence for 6G Multiple Access

Since scenarios towards 6G networks are heterogeneous, dynamic, and complex, it demands advanced optimization solutions to tackle more challenging problems [10]. Specifically,

a heterogeneous network architecture with various QoS demands will require an intelligent resource allocation scheme that adapts to different network dynamics. Meanwhile, resource efficiency, reliability, and robustness are becoming stringent QoS requirements to 6G networks. To meet these demands, AI-based resource allocation schemes are promising tools. In contrast to conventional approaches that rely on strict preconditions and strong assumptions, the opportunities that arise from learning environmental knowledge under diverse wireless channels render AI technology an adaptive and general solution to provide 6G MA schemes with more optimized and adaptive data-driven decisions.

Among the broad variety of AI algorithms, deep learning (DL) and reinforcement learning (RL) are two major research directions [11, 12]. DL techniques utilize neural networks to extract the sophisticated relationships among variables. By offloading the complexity to the exhaustive offline training process, DL algorithms benefit from a lower complexity in application compared to conventional techniques, such as convex optimization methods [13] and greedy algorithms [14]. Promising applications of DL in communications include channel estimation, user detection, user localization, and resource allocation.

RL techniques, on the other hand, excel at maximizing long-term gains by modeling the problem as a Markov decision process (MDP). The training data of RL algorithms are collected through a trial-and-error process and the performance of the model is evaluated through a long-term expected reward. Hence, RL is particularly advantageous for scheduling problems, such as UAV, network caching, data offloading, energy harvesting, and long-term resource allocation. Moreover, the integration of DL and RL, namely, deep reinforcement learning (DRL), embraces the feature learning skills of DL to improve the learning speed and the performance of RL, and has drawn extensive research interests as a promising optimization tool for supporting future MA schemes [15].

1.2 Motivation and Contributions

As a promising candidate for future 6G systems, AI-empowered NOMA schemes offer the following main advantages.

- **High spectrum efficiency**: Through power-domain or code-domain multiplexing, NOMA enables multiple users to share one orthogonal resource block, resulting in an enhanced spectrum efficiency. With the aid of AI technologies, the intelligent resource allocation schemes of NOMA can adapt to diverse channel variations and long-term goals, allowing for further spectrum efficiency gains.

- **Massive connectivity**: The future 6G systems are envisioned to support 10 times higher connectivity than 5G systems. The existence of NOMA offers a promising solution to this non-trivial task by fully exploiting the non-orthogonal characteristic, while the AI technologies act as powerful high-dimensional optimization tools for supporting the massive connectivity networks.

- **High compatibility**: NOMA can be designed as an "add-on" implementation to any existing OMA techniques, such as TDMA, FDMA, CDMA, OFDMA, and spatial division multiple access (SDMA). More importantly, NOMA is compatible with numerous 6G technologies as a spectrum enhancement tool, while the AI technologies can realize a joint optimization of the strongly coupled variables of the associated technologies.

- **Low computational complexity**: Conventional optimization algorithms rely on iterative updates that consist of high-complexity calculations. By offloading the complexity to the training process, AI algorithms can accomplish highly-complex optimization tasks at a lower computational complexity compared to conventional solutions.

- **Adaptive resource management**: The heterogeneous and dynamic 6G wireless environments make resource management a principal concern in NOMA sys-

tems. While conventional solutions require strict preconditions, such as convexity, AI algorithms can be deployed to provide a general data-driven solution and are applicable to various network dynamics, optimization objectives, and QoS metrics.

Motivated by the aforementioned advantages and the recent advancements in the fields of NOMA and AI, this thesis spans the system design and the performance enhancement of AI-empowered NOMA systems. More specifically, the research of this thesis first investigates the optimization problems of NOMA systems, including user detection, power allocation, user clustering, and beamforming, by utilizing AI-based solutions. The thesis then focuses on the integration of NOMA with emerging technologies, including RIS and device-to-device (D2D) communications, with the assistance of AI algorithms. The specific motivations and contributions of this dissertation are summarized as follows.

1.2.1 Joint User Activity and Data Detection in Grant-Free NOMA using Generative Neural Networks

In the context of grant-free NOMA, the number of potential users can grow far beyond the number of orthogonal resources and data packets are transmitted immediately in the next available time slot without waiting for a grant. Without any prior scheduling, the BSs must perform multi-user detection (MUD) to identify the group of active users in addition to their transmitted data. One key characteristic of future 6G communication is sporadic traffic where a small proportion of users enter the system simultaneously while the majority of the users remains silent [16]. By exploiting sparsity for signal reconstruction, compressed sensing (CS) has become a promising solution to MUD problems [17]. In particular, the signals received at the BS can be viewed as a set of underdetermined equations of the sparse signals. Thus, CS theory guarantees full reconstructions of the signals with high probability.

In Chapter 3, the MUD problem in uplink NOMA is examined. Since users stay active during consecutive time slots to complete their transmissions, the received signals often exhibit strong temporal correlations. Hence, the frame-wise joint sparsity model is

considered, which assumes that each user is either active or inactive throughout a fixed number of time slots and signal recovery is jointly performed for all signals received over the whole time frame. To realize joint user activity and data detection, a generative neural network-based MUD (GenMUD) framework is proposed. By identifying the independent user behaviors, the network architecture is designed with a small number of 1x1 convolutional layers to greatly reduce computational complexity. Moreover, with the aid of meta-learning, signal recovery can be performed in as few as five iterations using a single model regardless of the number of available orthogonal resources in the system. Nonetheless, in practical scenarios, user sparsity is usually unknown at the BS whereas CS-based algorithms often rely on prior knowledge of user sparsity. To replace the exhaustive sparsity approximation procedures in most MUD algorithms, a closed-form low-complexity user sparsity estimator is obtained and examined. The estimator only requires the information of the received signals and the noise level, both of which can be easily retrieved in practice. Hence, it can be applied along side any MUD algorithms.

The novelty of this work is supported by the following publication

- **Y. Zou**, Z. Qin and Y. Liu, "Joint User Activity and Data Detection in Grant-Free NOMA using Generative Neural Networks," in *Proc. IEEE Int. Communications Conf. (ICC'21)*, Montreal, Canada, June 2021.

1.2.2 Adaptive NGMA Scheme for Energy-limited Networks: A Deep Reinforcement Learning Approach

Despite the superior spectral efficiency of NOMA, its performance gain over conventional OMA techniques can be limited in certain scenarios. For instance, the quasi-degradation condition [18] was proposed as a sufficient and necessary condition for NOMA to approach the optimal dirty paper coding (DPC) rate region in the MIMO context. Moreover, in QoS constrained cases, the performance of NOMA is shown to decrease when the channel gain difference is small [19]. Hence, in the diverse 6G environments, there does not exist one optimal MA technique that suits all network scenarios.

An adaptive MA scheme that can intelligently adjust its MA policy according to the varying environment is becoming a valuable research direction for supporting future 6G applications.

In Chapter 4, an adaptive next generation multiple access (NGMA) scheme is designed, where users can be adaptively allocated to SDMA or NOMA clusters and are served with the same orthogonal frequency and time resource. The long-term power-constrained sum rate maximization problem is investigated, where beamforming, power allocation, and user clustering are jointly optimized. The optimization problem is a non-convex mixed-integer problem. In Chapter 4, a spatial correlation-based user clustering algorithm is proposed to transform the problem, where user grouping can be performed by selecting a clustering threshold, which is a continuous-valued parameter that can be jointly optimized with the other continuous variables. As one of the critical performance targets of NGMA systems, low energy consumption is achieved by enforcing a long-term total power constraint. In this case, the BS can coordinate the power consumption among the time slots to enhance the long-term total sum rate, however, the problem is non-trivial for conventional iterative algorithms. Hence, a DRL-based resource allocation framework is proposed to address this dynamic optimization problem. To achieve a fast and stable training process, the trust region policy optimization (TRPO) learning algorithm is employed, which imposes a limitation on the maximum distributional distance between successive policies.

The results of this work are to be submitted for publication in

- **Y. Zou**, W. Yi, X. Xu, Y. Liu, "Adaptive NGMA Scheme for Energy-limited Networks: A Deep Reinforcement Learning Approach", in *Proc. IEEE Int. Communications Conf. (ICC'23)*, Rome, Italy, June 2023; (to be submitted)

1.2.3 Comparisons between DL and DRL on the Optimization of RIS-assisted NOMA systems

As two emerging 6G technologies, NOMA can be integrated into RIS-aided networks to facilitate a *win-win* transmission framework [20]. Specifically, NOMA can effectively enhance the spectral efficiency of RIS-aided networks, while the RIS can configure the channel conditions, allowing for increased flexibility in power allocation and decoding order designs and enabling better system performance and better QoS guarantees. Despite the integration of NOMA and RIS being promising, it also leads to challenging issues preventing the full benefits from being reaped. Firstly, the objective functions are often non-convex due to the range constraints on the absolute values of the RIS phase shifts. Secondly, the high-dimensional phase shift variables are strongly coupled with the resource allocation strategies of NOMA, leading to challenging optimization problems. As promising optimization tools, various AI technologies, such as DL and DRL, demonstrated outstanding performance when implemented in RIS-aided NOMA systems. However, the performance comparisons between them remain understudied.

In Chapter 5, a novel RIS-aided downlink NOMA system is proposed, in which a QoS-based NOMA clustering method is designed to enhance the resource efficiency under the zero-forcing (ZF) precoding scheme. Then, the joint design of the RIS phase shift and the BS power allocation is examined, subject to the sum rate maximization objective. To conduct a thorough comparison between DL and DRL, the optimization problem is formulated from both short-term and long-term perspectives. Specifically, the instantaneous power consumption of each time slot is fixed in the short-term formulation, whereas the long-term formulation allows the BS to coordinate the transmit power among the time slots subject to a long-term total power constraint. The DL algorithm transforms the joint optimization problem into a two-step optimization problem and utilizes meta-learning to improve the convergence rate. The DRL algorithm employs the state-of-the-art DDPG training algorithm, where the reward function is carefully designed to enforce both the short-term QoS constraints and the long-term transmit power constraints.

The novelty of this work is reinforced by the following works

- **Y. Zou**, Y. Liu, K. Han, X. Liu, and K. K. Chai, "Meta-learning for RIS-assisted Non-Orthogonal Multiple Access Networks", in *Proc. IEEE Global Communications Conf. (GLOBECOM'21)*, Madrid, Spain, December 2021.

- **Y. Zou**, Y. Liu, X. Liu, X. Mu, X. Zhang, C. Yuen, "Comparisons between DL and DRL on the Optimization of RIS-assisted NOMA Networks", *IEEE Transactions on Wireless Communications*; (under revision).

1.2.4 Multi-Agent Resource Allocation in NOMA-Enhanced Multi-RIS Aided D2D Networks

D2D communication has been recognized as a promising technique for enhancing system capacity and reducing traffic congestion in wireless networks. Nevertheless, co-channel interference is a growing concern due to the ever-increasing number of D2D equipment and applications. With the aid of successive interference cancellation (SIC) in NOMA, the severe interference of co-channel users can be effectively eliminated. Moreover, the integration with RIS can further promote the spectral efficiency gain of NOMA with high deployment flexibility and low deployment cost.

In Chapter 6, a RISs-assisted NOMA-empowered D2D communication underlay cellular network is considered. All cellular users (CUEs) and D2D users are assumed to roam continuously, where the direct links among the D2D users and between the CUEs and the BS are blocked by obstacles. To enhance the channel quality, multiple RISs are deployed to establish line-of-sight (LoS) links towards the D2D users, the BS, and the CUEs. Meanwhile, NOMA is employed by the D2D transmitters (DTs) to communicate with multiple D2D receivers (DRs) through the same orthogonal resource block simultaneously. Thus, D2D groups are formed instead of the conventional D2D pairs to enhance the spectral efficiency. The long-term sum rate maximization problem is investigated, where the sub-channel assignments of D2D groups, the power allocation at the DTs, and the RIS phase shifts are jointly optimized. To jointly optimize the

strongly coupled parameters of various technologies, a multi-agent resource allocation framework is designed, where the DTs and the RIS controllers are modelled as agents, who are trained in a centralized training and decentralized execution (CTDE) manner. Furthermore, all agents adopt the multi-pass deep Q-network (MP-DQN) to address the mixed-integer problem without any relaxations of the action space or modifications to the network architecture.

The novelty of this work is supported by the following work

- L. Guo, J. Jia, **Y. Zou**, Y. Liu, J. Chen, X. Wang, "Resource Allocation for Multiple RISs Assisted NOMA Empowered D2D Communication: A MAMP-DQN Approach", *IEEE Transactions on Vehicular Technology*; (under review).

1.3 Related Works

In this section, the related works of this thesis are discussed from three aspects: NOMA systems, AI-empowered systems, and AI-empowered NOMA systems.

1.3.1 NOMA Systems

The existing NOMA systems can be divided into two main categories, namely code-domain NOMA (CD-NOMA) systems and PD-NOMA systems. The concept of CD-NOMA is inspired by CDMA, in which multiple users share the same orthogonal time and frequency resources through unique spreading sequences. In contrast to CDMA, the spreading sequences in CD-NOMA are further restricted to sparse non-orthogonal sequences [21, 22]. Specifically, sparse spreading sequences have the advantage of interference reduction since each user only spreads its data over a small number of chips, while non-orthogonal sequences are capable of supporting much more users than the number of chips. Accordingly, sophisticated MUD algorithms are required at the receivers to accurately decode the superimposed signals. Hence, numerous research contributions have been established on the designs and the optimization of CD-NOMA systems [23–28].

Based on trellis-coded modulation techniques, the authors of [23] proposed a joint codebook and MUD design for CD-NOMA systems. To compare the performance of dense and sparse codebooks, the authors of [24] conducted theoretical analysis on the diversity order of these two types of CD-NOMA systems and concluded that dense codebooks can achieve a lower error rate with comparable complexity compared to sparse codebooks. In [25], an orthogonal matching pursuit (OMP)-based MUD algorithm was developed for uplink CD-NOMA systems. By exploiting the user activity sparsity, the proposed algorithm employed CS theory and adopted a Toeplitz matrix as the spreading code sequence for satisfying the necessary restricted isometry property (RIP) of CS. In [26], an iterative thresholding technique, namely approximate message passing, was employed in conjunction with the expectation maximization algorithm to solve the MUD problem in uplink CD-NOMA systems, where the spreading sequences are pseudo-random noise sequences. A block CS-based subspace pursuit algorithm was presented in [27], which utilized block compressed sensing to improve signal detection accuracy and designed an algorithm stopping criterion based on noise levels to enhance the sparsity estimation accuracy. Moreover, an alternative direction method of multipliers (ADMM)-based MUD solution was proposed in [28], which utilized the signal and the support detection of the previous time slot as prior knowledge to enhance MUD performance.

As another major category of NOMA, PD-NOMA exploits the near-far effect in wireless environments and superimposes the signals in the power domain. Owing to the superior spectral enhancement capability, many research efforts have been devoted to investigating the designs and the resource allocation problems of PD-NOMA systems [29–34]. Based on IoT scenarios, a joint design of user scheduling and power allocation was proposed in [29] with the objective of transmit power minimization in PD-NOMA systems. In [30], a MA selection framework was proposed, which adaptively switches between OMA and PD-NOMA for maximizing the sum rate. Specifically, the proposed utility function reflects both the rate and the complexity costs of the MA schemes, which captures the tradeoff between OMA and NOMA. In [31], the sum rate

maximization problem in millimeter wave (mmWave)-NOMA systems was investigated, where user clustering, beamforming, power allocation, and power splitting are optimized, respectively. The authors of [32] studied the resource allocation problems in uplink multi-UAV PD-NOMA systems, in which the sum rate maximization problem was formulated by optimizing sub-channel allocations, power allocations, and the UAVs' attitudes. The authors of [33] investigated the designs of user clustering, power allocation, and hybrid beamforming in mmWave-NOMA systems and the proposed approach outperformed conventional mmWave-OMA systems in terms of both sum rates and energy efficiency. To enhance the physical layer security of NOMA systems, the authors of [34] proposed a joint beamforming and power allocation design to maximize the secrecy sum rate, supported by an asymptotic analysis on the optimality of the proposed power allocation scheme.

1.3.2 AI-empowered Networks

In recent years, AI has emerged as a tremendous technology to address the problems of exploding data volume, non-convex optimization, and computational complexity [11, 12, 15]. The research directions of AI in communications can be divided into two categories, namely DL-empowered wireless networks [35–39] and DRL-empowered wireless networks [40–43]. In particular, DL techniques utilize extensive datasets to learn the unknown relationships among the variables, while offloading the optimization complexity to the training phase. Moreover, DL models often demonstrate strong generalization capability to unseen datasets and robustness to environment variations such as noise and imperfect CSI. For instance, the authors in [35] designed a deep learning-based MUD model for massive machine-type communications, which demonstrated improved detection accuracy while achieving a ten-fold decrease in computing time compared to the conventional algorithms. In [36], a model-driven deep learning-based joint channel and signal estimation framework was proposed, which exhibited strong adaptability to varying channel conditions. In [37], deep transfer learning was employed to solve the beamforming optimization problem in RIS-assisted networks, where the proposed algo-

rithm required only a small amount of training data. The problem was further extended into discrete phase shift cases to address hardware limitations. Moreover, the authors of [38] and [39] utilized neural networks to learn the interactions between the receiver locations and the optimal RIS phase shift to achieve maximal sum rate in RIS-assisted networks.

An alternative type of AI algorithm, namely DRL, collects training samples by interacting with the environment through trial and error. Hence, in contrast to the offline training mechanism of DL, DRL is often referred to as a type of online learning algorithm. Moreover, the performance of DRL algorithms is evaluated by their long-term expected returns, which enables them to maximize future rewards rather than only exploiting instantaneous benefits. Motivated by these advantages, many research contributions [40–43] have been devoted to investigating the implementations of DRL in communication networks. In [40], DRL was employed to optimize the long-term energy efficiency of RIS-aided multi-input single-output (MISO) networks, where the RIS was implemented with energy harvesting technologies. The authors in [41] proposed a double-DQN algorithm for solving the caching problem in mobile edge computing platforms. The proposed resource allocation scheme exploited vehicular mobility to reduce the cost of energy consumption, latency, and communication. In [42], the authors studied the resource management problem in network slicing and designed a DRL algorithm with a long short-term memory (LSTM) network by exploiting user mobility. Moreover, the authors of [43] investigated the implementation of DRL in MA protocol designs based on heterogeneous networks. The proposed algorithm demonstrated near-optimal performance subject to various objectives, including sum rate and user fairness.

1.3.3 AI-empower NOMA systems

With the evolution of NOMA technology towards 6G standards, conventional optimization techniques are struggling to cope with the escalating network complexity and the diverse application scenarios. Therefore, many researchers have shifted their atten-

tions to the implementations of AI technologies in NOMA, leading to DL-empowered NOMA systems [44–48] and DRL-empowered NOMA systems [49–54]. Among the DL implementations, the authors of [44] proposed an end-to-end transceiver for NOMA-based massive machine-type communications via both data-driven and model-driven DL designs. The authors in [45] utilized fully-connected layers and residual connections to improve the MUD detection accuracy in grant-free NOMA systems, without the knowledge of sparsity. In [46], a deep neural network was utilized to optimize the subcarrier assignment of OFDMA and the user clustering of NOMA in downlink video communications. In [47], the authors investigated the energy efficiency maximization problem in mmWave-NOMA systems, subject to QoS, interference, and power limitations. In this work, semi-supervised learning was employed to train a deep neural network for sub-channel assignment and power allocation. In [48], a DL-based outage probability and sum rate prediction framework was proposed for cognitive NOMA systems.

In terms of the implementations of DRL in NOMA systems, a prototype of transmit power pool was developed in [49] for grant-free NOMA, in which a multi-agent DQN network was designed to optimize the transmit power levels. The authors in [50] studied the long-term sum rate maximization problem in RIS-aided NOMA systems, by optimizing the stochastic phase shift with a deep deterministic policy gradient (DDPG) algorithm. In [51], two asynchronous DRL algorithms were proposed for joint relay selection and power allocation in hybrid NOMA/OMA systems. The authors of [52] investigated the long-term sum rate maximization problem in uplink grant-free NOMA systems by formulating the problem as a partially observable MDP. In [53], the power allocation of cache-aided NOMA systems was studied, where an optimal power allocation policy was derived in closed-form and a dual-network driven DRL solution was proposed. The DRL-based method outperformed the closed-form solution at the cost of extensive model training. In [54], a DRL-based power allocation and channel assignment algorithm was designed for NOMA systems, which demonstrated near-optimal performance.

1.4 Dissertation Organization

The remainder of this thesis is organized as follows. Chapter 2 introduces the fundamental concepts such as the basic principles of NOMA, DL, DRL, meta-learning, and RIS. Chapter 3 proposes a DL-based MUD framework for uplink grant-free NOMA systems. Chapter 4 designs an adaptive NGMA framework and investigates the application of DRL for resource allocation in the proposed network. Chapter 5 studies the performance comparisons between DL and DRL when applied to RIS-assisted NOMA systems. Chapter 6 examines the application of multi-agent DRL for resource allocation in NOMA-enhanced D2D networks with multiple RISs. Chapter 7 presents the conclusions of this thesis and discusses promising future research directions.

Chapter 2

Fundamental Concepts

This chapter provides the technical background knowledge that supports this thesis. First, the fundamental principles of NOMA are introduced, including CD-NOMA, PD-NOMA, and MIMO-NOMA, which lays a comprehensive foundation for the technical works. Second, the background knowledge of several AI technologies is discussed, including DL, DRL, and meta-learning, which provides the fundamental guidelines for the optimization frameworks. Finally, the concepts of several related technologies, namely CS, SDMA, and RIS, are outlined to offer a thorough understanding of the network designs.

2.1 Fundamental principles of NOMA

This section aims to provide a detailed introduction to NOMA principles, from the key technologies of NOMA, such as superposition coding (SC), SIC, and MUD, to the general mathematical formulations of both CD-NOMA and power domain NOMA, followed by an overview of MIMO-NOMA systems, including beamformer-based MIMO-NOMA and cluster-based MIMO-NOMA.

2.1.1 Key Technologies of NOMA

The core principles of NOMA consist of two concepts, namely signal multiplexing and signal decoding. In typical NOMA systems, SC is employed to multiplex the signals in the power or code domain. Then, to accurately decode the superimposed signal, the receivers need to carry out SIC or MUD, depending on the multiplexing technique.

2.1.1.1 Superposition Coding

How to effectively communicate with multiple transmitters or multiple receivers has been a challenging problem in communications. Conventionally, the solution is to set up orthogonal channels through time or frequency multiplexing. These orthogonal approaches have the advantage of ensuring zero interference between the channels, however, they often fail to achieve the optimal transmission rate for a given packet error rate [55]. As a non-orthogonal multiplexing technique, SC has been theoretically proved to achieve the capacities of scalar Gaussian broadcast channels [56], leading to extensive research interests in various channels, such as MA channels [57], interference channels [58], relay channels [59], and wiretap channels [60].

The fundamental concept of SC is to superimpose the intended signals before transmission to exploit the combined degrees of freedom available to these signals in orthogonal schemes. By recognizing the channel differences among the users, the SC method can be carefully designed with the decoding algorithm to ensure the successful separation of the superimposed signal at the receiver side.

2.1.1.2 Successive Interference Cancellation

In terms of power-domain multiplexing, the most dominant decoding technique is SIC, which has been demonstrated to achieve the capacity region in both additive white Gaussian noise (AWGN) channels and fading channels [61]. The main concept of SIC is to enable the receiver with a stronger channel to first decode the signal of the receiver with a weaker channel. Then, the decoded signal is subtracted from the superimposed

signal as interference. Therefore, the signal of the stronger transmitter can be decoded in an interference-free manner. To ensure successful and reliable transmission, the receiver with a weaker channel is often allocated with more transmit power and the optimal decoding order starts from decoding the signal of the user with the weakest channel gain to the user with the strongest channel gain [62].

2.1.1.3 Multi-User Detection

With code-domain multiplexing, MUD algorithms are often utilized to perform user detection and signal detection. Existing MUD algorithms are mostly designed based on CS theory, which relies on two conditions, namely RIP and sparsity. The necessary RIP condition of CS can be satisfied by carefully designing the spreading code sequence and the sparsity condition can be ensured by assuming sporadic traffic patterns [63]. Then, the signals at the receiver can be viewed as a set of underdetermined equations of the sparse signals and CS theory guarantees full reconstructions of the signals with high probability.

2.1.2 Mathematical Formulation of NOMA

2.1.2.1 PD-NOMA

Consider a downlink NOMA system of a single-antenna BS and K single-antenna users. The signal intended for user k is denoted by s_k. To employ power-domain multiplexing, the signal intended for each user k is allocated with transmit power p_k. Hence, the resulting superimposed signal x is formulated as $x = \sum_{k=1}^{K} \sqrt{p_k} s_k$. By denoting the channel between the BS and each user k as h_k, the signal received by user k is given by

$$y_k = h_k \sum_{k=1}^{K} \sqrt{p_k} s_k + n_k \tag{2.1}$$

$$= \underbrace{h_k \sqrt{p_k} s_k}_{\text{Desired signal}} + \underbrace{h_k \sum_{l=1,\ldots,K, l \neq k} \sqrt{p_l} s_l}_{\text{SIC signal}} + \underbrace{n_k}_{\text{noise}}. \tag{2.2}$$

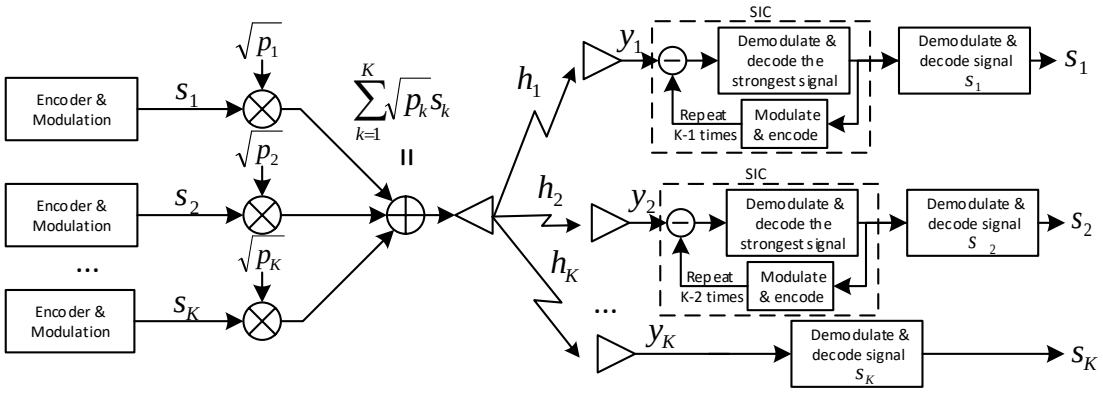


Figure 2.1: Block diagram of PD-NOMA.

According to the principles of SIC, each user performs a decode-then-subtract procedure following a pre-defined decoding order until the intended signal is obtained. To simplify the demonstration, the channel gains are assumed to follow $|h_1|^2 \geq |h_2|^2 \geq \cdots \geq |h_K|^2$, hence the achievable rate of user k is formulated by

$$R_k = \log_2 \left(1 + \frac{p_k |h_k|^2}{\sum_{i=1}^{k-1} p_i |h_k|^2 + \sigma_k^2} \right), \tag{2.3}$$

where σ_k^2 denotes the variance of the AWGN noise. A typical PD-NOMA transceiver system with the optimal decoding order is illustrated in Fig. 2.1.

Similarly, in an uplink PD-NOMA system, the BS is required to send controlling signals to the users for power allocation. Then, the users send their intended signals through the same orthogonal resource block. With the aid of the SIC technique, the BS decodes all the signals following the pre-defined decoding order. In this thesis, PD-NOMA systems are considered in Ch. 4–Ch. 6.

2.1.2.2 CD-NOMA

For the CD-NOMA, consider an uplink network with K single-antenna users and one single-antenna BS, supported by M subcarriers. The signal to be transmitted by user k is denoted by s_k. The concept of CD-NOMA is to assign sparse sequences or non-

orthogonal low-correlation sequences to each user. By denoting the spreading sequence assigned to user k as $\boldsymbol{\lambda}_k = [\lambda_{1k}, \ldots, \lambda_{Mk}]^T \in \mathbb{C}^{N\times 1}$, the signal received by the BS on subcarrier n is formulated by

$$y_m = \sum_{k=1}^{K} h_{mk}\lambda_{mk}s_k + n_m, \tag{2.4}$$

where h_{mk} denotes the channel of user k over subcarrier m and n_m denotes the AWGN noise on subcarrier m. Hence, the signal vector $\mathbf{y} = [y_1, \ldots, y_M]^T$ received by the BS is given by

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{2.5}$$

where $\mathbf{s} = [s_1, \ldots, s_K]^T$, $\mathbf{H}$ denotes the $M \times K$ channel matrix, whose entries are $[\mathbf{H}]_{mk} = h_{mk}\lambda_{mk}$, and $\mathbf{n} = [n_1, \ldots, n_M]^T$ denotes the noise vector.

Based on the received signal vector $\mathbf{y}$, the BS needs to extract the transmitted signal vector $\mathbf{s}$ by solving the following MUD problem:

$$\min_{\mathbf{s}} ||\mathbf{y} - \mathbf{H}\mathbf{s}||_2^2. \tag{2.6}$$

In 6G massive connectivity networks, the number of users can easily exceed the number of subcarriers, resulting in an overloaded system, i.e., $N >> M$. In this case, the MUD problem in (2.6) becomes a non-trivial under-determined system. Existing MUD methods find approximated solutions to (2.6) through two types of relaxation techniques, namely convex relaxation and greedy algorithms. In Ch. 3, the MUD problem in an uplink grant-free CD-NOMA system is investigated.

2.1.3 MIMO-NOMA

As an indispensable component of 6G, multi-antenna techniques introduce additional degrees of freedom in the spatial domain for performance enhancement. By carefully designing the beamformer at the transmitter, the signal power and the interference power

can be effectively adjusted, which consequently impacts the SINR of each user. Existing MIMO-NOMA systems can be divided into two categories, namely beamforming-based MIMO-NOMA systems [64] and cluster-based MIMO-NOMA systems [65]. In beamforming-based NOMA, a dedicated beamformer is allocated for each user, which is jointly optimized with the beamformer of all users. When there is a sufficient spatial degree of freedom, the multi-user interference can be effectively eliminated to achieve high spectral efficiency. However, the beamforming design needs to take into account the MUD algorithm, such as the SIC decoding order, to ensure successful signal decoding, which leads to high optimization complexity. Moreover, designing a dedicated beamformer for each user induces exponential complexity, which prevents the implementation of beamforming-based MIMO-NOMA in large scale networks.

In contrast to the beamforming-based NOMA, cluster-based NOMA allocates users into multiple clusters and assigns the same beamformer to all users in the same cluster. The SIC decoding order is designed within each cluster, while the signals of the users from other clusters are treated as inter-cluster interference. This technique exploits the spatial correlation features among the users to reduce or even eliminate inter-cluster interference. Moreover, the number of beamformers can be significantly less than the number of users, which is of vital importance in large scale networks.

In Ch. 4, an adaptive MA framework is designed based on MIMO networks, which serves OMA and NOMA users with the same time and frequency resources. In Ch. 5, the resource allocation problem in RIS-enhanced MIMO-NOMA systems is investigated.

2.2 Artificial Intelligence

This section presents the fundamental principles of the AI technologies used in this thesis, including DL, DRL, and meta-learning.

2.2.1 Deep Learning

The core concept of DL is to design intelligent algorithms, known as neural networks, that mimic the structure of the neurons in the human brain. A typical neural network has three main components, namely the input layer, the hidden layer, and the output layer. Each layer consists of a group of neurons, each of which performs mathematical calculations based on the values of the previous layer and the result is forwarded to the next layer. The input layer is often a linear layer of the same dimension as the input, where each neuron represents each input value. The hidden layer and the output layer can take various forms. The neurons in the hidden layer and the output layer are denoted by $f_1(\cdot), \ldots, f_H(\cdot)$ and $g_1(\cdot), \ldots, g_M(\cdot)$, respectively. Thus, the computation performed by the three-layer neural network is formulated as $y_m = g_m(\sum_{h=1}^{H} f_h(\mathbf{x})), \forall m = 1 \ldots, M$, where $\mathbf{x} = [x_1, \ldots, x_N]^T$ and $\mathbf{y} = [y_1, \ldots, y_M]^T$ denote the network input and output, respectively. Common choices of layers are introduced as follows:

- *Fully connected layer*: The fully connected layer is the most basic hidden layer, which can be represented by $f_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + b_i$, where $\mathbf{x}$ denotes the output of the previous layer, $\mathbf{w}_i$ is known as the weights, and b_i is the bias.

- *Activation layer*: Activation layers are indispensable components in modern neural network design, due to their vital impacts on the capability and performance of the neural network. Common choices of activation functions include rectified linear unit (ReLU), sigmoid, and hyperbolic tangent (tanh). Nonetheless, any customized functions can be employed as the activation function, as long as it is differentiable.

- *Convolutional layer*: A convolutional layer uses kernels to slide across the input, performing a convolution operation between each input region and the kernel. Each kernel has a window size, usually 3×3, which significantly reduces the computational complexity and often demonstrates outstanding learning performance in image processing tasks due to the spatial learning characteristic [66].

Recent developments in DL have also introduced more advanced network architectures,

such as generative adversarial networks (GANs) [67], autoencoders [68], and LSTM networks [69].

Apart from the network architecture, another crucial component of DL is the loss function, which evaluates how poorly the network performs and serves as the learning objective. Similar to the activation function, any differentiable function can be utilized as the loss function. The two most widely used loss functions are the mean squared error loss for regression problems and the cross-entropy loss for classification problems.

Given the network architecture and the loss function, the training procedure can be initiated. Neural networks are usually trained via gradient-based techniques through the following update equation

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \gamma \frac{1}{B} \sum_{b=1}^{B} \nabla_{\mathbf{w}} \mathcal{J}^b(\mathbf{w}_t), \tag{2.7}$$

where $\mathbf{w}_t$ indicates the network weights at iteration t, γ denotes the learning rate, B denotes the batch size, and $\mathcal{J}^b(\cdot)$ denotes the loss of the network evaluated using the data batch b.

In Ch. 3, a generative neural network with convolutional layers is designed to solve the MUD problem in grant-free NOMA systems. In Ch. 4, a neural network is trained to perform resource allocation in RIS-aided NOMA systems.

2.2.2 Deep Reinforcement Learning

In contrast to DL, where the training data is prepared beforehand, the training data of DRL is collected during training by interacting with the environment. The goal of DRL is to determine the optimal action that maximizes the future return based on the current observations. To be specific, the optimization problems of DRL must follow the MDP framework, which is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ are the state space and the action space, respectively. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the Markov transition probability, which specifies the probability of transitioning to a particular future state

given the current state. $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. γ is the discount factor, which determines how much the agent should care about rewards in the distant future relative to those in the immediate future. The agent is represented as a policy function $\pi : \mathcal{S} \to \mathcal{A}$, which specifies a mapping from the state space to the action space. In terms of the formulated MDP, the goal of DRL is to find the optimal policy π^* that maximizes the expected return $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)]$.

There are three commonly used metrics for evaluating the policy π, namely the action-value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the state-value function $V^\pi : \mathcal{S} \to \mathbb{R}$, and the advantage function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. First, the action-value function, also known as the Q-function is defined as

$$Q^\pi(\mathbf{s}, \mathbf{a}) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^n r(\mathbf{s}_t, \mathbf{a}_t) | \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_0 = \mathbf{a}, \mathbf{a}_t \sim \pi(\mathbf{s}_t), \mathbf{s}_{t+1} \sim \mathcal{P}(\cdot|\mathbf{s}_t, \mathbf{a}_t)], \tag{2.8}$$

which evaluates the goodness of taking action $\mathbf{a}$ at state $\mathbf{s}$. The state-value function is defined as

$$V^\pi(\mathbf{s}) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^n r(\mathbf{s}_t, \mathbf{a}_t) | \mathbf{s}_0 = \mathbf{s}, \mathbf{a}_t \sim \pi(\mathbf{s}_t), \mathbf{s}_{t+1} \sim \mathcal{P}(\cdot|\mathbf{s}_t, \mathbf{a}_t)], \tag{2.9}$$

which evaluates the goodness of being in state $\mathbf{s}$. Finally, the advantage function is computed by

$$A^\pi(\mathbf{s}, \mathbf{a}) = Q^\pi(\mathbf{s}, \mathbf{a}) - V^\pi(\mathbf{s}), \tag{2.10}$$

which describes the advantage of taking action $\mathbf{a}$ in state $\mathbf{s}$ among all other possible actions and states.

The training procedure of existing DRL algorithms can be generally classified into two categories, namely policy-based methods and value-based methods. In policy-based approaches, the policy function π is directly optimized, whereas, in value-based algorithms, the value function (action-value function/station-value function) is optimized and the optimal policy is implicitly derived from the optimal value function. Trust

region policy optimization (TRPO) and DQN are examples of policy-based and value-based DRL algorithms, respectively.

In Ch. 4, the DDPG algorithm is employed to perform resource allocation in RIS-aided NOMA systems. In Ch. 5, the TRPO algorithm is employed to solve the resource allocation problem in NGMA systems.

2.2.3 Meta-learning

Meta-learning, also known as learning to learn, refers to the process of improving a learning algorithm over multiple learning episodes. In conventional AI algorithms, the training data is assumed to follow a certain distribution, denoted as $\mathcal{T}_1$. With sufficient data and training time, the neural network is expected to perform well on any data following $\mathcal{T}_1$. However, poor performance is expected if the network is directly applied to data with a different distribution, such as $\mathcal{T}_2$. A common solution is to train a new neural network, which results in expensive computational and time costs.

Model agnostic meta-learning (MAML) [70] is a commonly used meta-learning algorithm that is developed to reduce the excessive training time on previously unseen tasks, i.e., $\mathcal{T}_2$ in the previous example. In particular, MAML assumes that these tasks share a common distribution $\mathcal{T}$, known as the task distribution, i.e., $\mathcal{T}_1, \mathcal{T}_2 \sim \mathcal{T}$. Each training iteration k of MAML consists of two steps: the task learning step and the task adaption step, which is described in the following.

- *Task learning step*: $N_{\mathcal{T}}$ tasks are sampled from the task distribution $\mathcal{T}$. Based on each task i, a provisional update is performed by

$$\mathbf{w}_{k+1}^{i} \leftarrow \mathbf{w}_k - \gamma_1 \nabla_{\mathbf{w}_k} J^i(\mathbf{w}_k), \tag{2.11}$$

  where $\mathbf{w}_k$ denotes the network weights at iteration k, γ_1 denotes the inner learning rate, $J^i(\cdot)$ denotes the loss function evaluated using the data of task $\mathcal{T}_i$. By the end of this step, $N_{\mathcal{T}}$ sets of updated weights are obtained as $\mathbf{w}_{k+1}^{1}, \ldots, \mathbf{w}_{k+1}^{N_{\mathcal{T}}}$.

- *Task adaption step*: The updated performance of the $N_{\mathcal{T}}$ networks are evaluated and the initial network weights $\mathbf{w}_k$ are optimized through a second order differentiation, given by

$$\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \gamma_2 \nabla_{\mathbf{w}_k} \left[\frac{1}{N_{\mathcal{T}}} \sum_{i=1}^{N_{\mathcal{T}}} J^i \left(\mathbf{w}_k + \gamma_1 \nabla_{\mathbf{w}_k} J^i(\mathbf{w}_k) \right) \right], \tag{2.12}$$

  where γ_2 denotes the outer learning rate. In (2.12), the true updated weights $\mathbf{w}_{k+1}$ is obtained, which becomes the initial weights of the next training iteration.

In Ch. 3 and Ch. 4, meta-learning is employed to improve the convergence rate of the proposed DL algorithms.

2.3 Related Technologies

This section introduces the fundamental principles of the related technologies used in this thesis, including CS, SDMA, and RIS.

2.3.1 Compressed Sensing

CS is a signal reconstruction approach, which guarantees an exact or an approximate signal recovery when the number of samples is below the minimum Nyquist rate [63]. The fundamental principles of CS rely on two concepts, namely sparsity and the RIP condition. The sparsity of a signal is defined as the number of non-zero elements under a certain domain, such as the spatial or the temporal domain, and the RIP condition states that the measurement matrix must preserve the distance between two signals to a large extent.

Let $\mathbf{s} \in \mathbb{C}^{N \times 1}$ defines the signal of interest and $\mathbf{\Phi} \in \mathbb{C}^{N \times N}$ defines the sparsifying basis, the sparse representation of $\mathbf{s}$ is computed by

$$\mathbf{x} = \mathbf{\Phi s}, \tag{2.13}$$

where the sparsity of $\mathbf{x}$ is denoted by $K << N$. Based on this sparsity, CS theory claims that only $M < N$ measurements of $\mathbf{x}$ is required to obtained a full reconstructed. The compressed signal is formulated by

$$\mathbf{y} = \mathbf{\Psi x} = \mathbf{\Psi \Phi s} = \mathbf{\Theta s}, \tag{2.14}$$

where $\mathbf{y}$ denotes the compressed signal, $\mathbf{\Psi} \in \mathbb{C}^{M \times N}$ denotes the measurements matrix, and $\mathbf{\Theta} \in \mathbb{C}^{M \times N}$ denotes the sensing matrix. The formulation in $\mathbf{y}$ is a set of underdetermined linear systems, hence solving $\mathbf{s}$ based on $\mathbf{y}$ is generally impossible. Fortunately, CS theory guarantees that a full reconstruction of $\mathbf{s}$ is possible as long as $\mathbf{s}$ is sparse and $\mathbf{\Psi}$ satisfies RIP.

Definition 1. *Restricted Isometry Property (RIP): The matrix* $\mathbf{\Psi}$ *is said to satisfy RIP of order* K *if there exists* $\epsilon > 0$ *such that, for any vector* $\mathbf{x}$ *of sparsity* K*,*

$$1 - \epsilon \leq \frac{||\mathbf{\Psi x}||_2^2}{||\mathbf{x}||_2^2} \leq 1 + \epsilon. \tag{2.15}$$

However, the RIP condition is computationally expensive to verify, but it can be achieved with a high probability simply by selecting $\mathbf{\Psi}$ as a random matrix, such as Gaussian matrices, Bernoulli matrices, and any matrices with independent and identically distributed (i.i.d.) entries.

Given the compressed signal $\mathbf{y}$ and the measurement matrix $\mathbf{\Psi}$ that satisfies RIP, the signal recovery of the sparse signal $\mathbf{x}$ is achieved by solving the following optimization problem:

$$\underset{\mathbf{x}}{\text{argmin}} \; ||\mathbf{y} - \mathbf{\Psi x}||_2^2, \qquad \text{subject to } ||\mathbf{x}||_0 = K. \tag{2.16}$$

In Ch. 3, CS is utilized to design the MUD algorithm for grant-free NOMA systems, where the sparsity constraint is satisfied by assuming sporadic user activity and the RIP condition is satisfied by enforcing it as a loss function of the neural network.

2.3.2 Spatial Division Multiple Access

SDMA is a MA technique designed on the basis of MIMO networks. In contrast to traditional cellular networks, where the BS radiates power in all directions, SDMA utilizes the spatial location of each user to design the beamformers. To be specific, by equipping the BS with multiple antennas, an additional spatial degree of freedom is introduced and can be exploited to reduce or even eliminate the multi-user interference.

The optimal SDMA scheme, known as DPC, was proposed in [71]. However, due to the excessive encoding and decoding, DPC exhibits expensive computational complexity, which motivated various suboptimal designs [72–74]. For instance, the opportunistic SDMA was proposed in [72] which studied the user scheduling problem under random beamforming and partial CSI and the proposed scheme demonstrated asymptotic optimality. The proposed designs in [73] and [74] utilized ZF beamforming with a semi-orthogonal user selection (SUS) algorithm, which also achieved asymptotic optimality.

Recently, the integration of NOMA and SDMA has also been investigated. For instance, the closed-form expression of the outage probability of NOMA-SDMA systems was derived and analyzed in [75]. In [76], the author studied the design of NOMA beamforming in SDMA systems and proposed two strategies for demonstrating the trade-off between the system performance and the complexity. In Ch. 4, an adaptive NGMA system is designed, where SDMA and NOMA users are simultaneously served with the same time and frequency resources.

2.3.3 Reconfigurable Intelligent Surface

A RIS composes of a large number of low-cost reflecting elements that can proactively reconfigure the propagation of incident signals. This technique can be especially beneficial in dense urban areas, where LoS links are often blocked by various obstacles such as trees and buildings. As shown in Fig. 2.2, consider a multi-user network, where the direct links between the users and the BS are obstructed. By employing a RIS on the facade of a building, LoS links can be successfully established through signal reflection

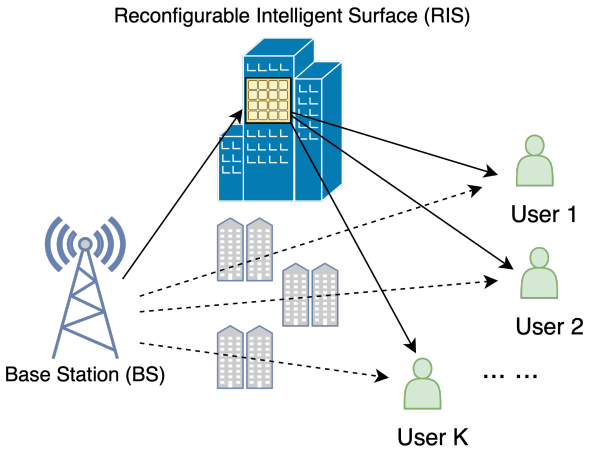


Figure 2.2: Illustration of RIS-enabled wireless communications.

from the BS to the blocked users.

To formulate the received signal in a RIS-aided downlink network, let K, M, and N denote the number of users, the number of antennas at the BS, and the number reflecting elements of the RIS, respectively. The RIS-user link and the BS-user link of user k are denoted by $\mathbf{h}_{R,k}^H \in \mathbb{C}^{1\times N}$, and $\mathbf{h}_{B,k}^H \in \mathbb{C}^{1\times M}$, respectively. The BS-RIS link is denoted by by $\mathbf{H}_{BR} \in \mathbb{C}^{N\times M}$. The phase shift of the RIS is denoted by $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n, \ldots, \theta_N]$, where $\theta_n \in [0, 2\pi)$. Thus, the diagonal phase-shifting matrix is expressed as $\boldsymbol{\Theta} = \text{diag}(\beta_1 e^{j\theta_1}, \ldots, \beta_n e^{j\theta_n}, \ldots, \beta_N e^{j\theta_N})$, where $\beta_n \in [0, 1]$ denotes the amplitude reflection coefficient.

Hence, the signal received by user k is derived as

$$y_k = \left(\mathbf{h}_{B,k}^H + \mathbf{h}_{R,k}^H \boldsymbol{\Theta} \mathbf{H}_{BR}\right) \sum_{k=1}^{K} \mathbf{w}_k x_k + n_k, \tag{2.17}$$

where $\mathbf{w}_k$ denotes the beamforming vector of user k and n_k denotes the AWGN.

Hence, by intelligently adjusting the phase shift of each RIS element, the communication channels can be beneficially manipulated to enhance various performance targets [77], including spectral efficiency, energy efficiency, sum rate, and user fairness. Benefiting from the low-cost meta-materials [78], RIS-enhanced networks can achieve

lower hardware cost and power consumption compared to MIMO networks, which rely on a large number of RF chains.

In Ch. 5, the sum rate maximization problem in RIS-aided NOMA systems is investigated. In Ch. 6, the sum rate maximization problem in multi-RIS enhanced NOMA-D2D networks is studied.

Chapter 3

Joint User Activity and Data Detection in Grant-Free NOMA using Generative Neural Networks

3.1 Introduction

In this chapter, the MUD problem in uplink grant-free CD-NOMA networks is investigated, where both user activity and the transmitted signals need to be detected and extracted based on the superimposed signal. Owing to the recent advancement in AI-based signal recovery techniques [79, 80], the proposed AI-enabled MUD framework demonstrates superior detection accuracy compared to conventional MUD approaches. The main contributions are as follows:

- The MUD problem in grant-free CD-NOMA systems is addressed based on generative neural networks such that signal recovery can be performed regardless of the number of available orthogonal resources in the system.

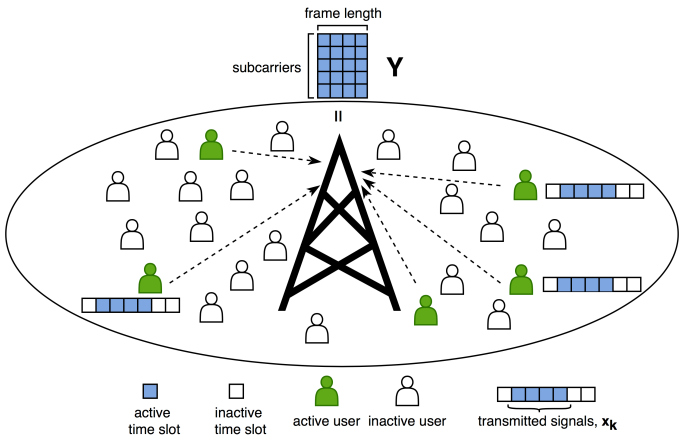


Figure 3.1: Illustration of a typical frame-based uplink grant-free CD-NOMA system.

- Exploiting the uncorrelated user behaviors in MUD data, a low-complexity generative network architecture is designed, which consists of a small number of 1x1 convolutional layers. By removing the fully-connected layers, the input latent dimension can be increased to achieve more accurate signal recovery with lower additional computational cost.

- To replace the exhaustive sparsity approximation procedures in most MUD algorithms, a closed-form user sparsity estimator is formulated. The estimator can be applied as an add-on technique to MUD algorithms since it only requires information about the received signals and the noise level, both of which are easy to obtain in practice.

- The extensive simulation results show that the proposed GenMUD is able to improve detection accuracy compared to conventional methods and the proposed sparsity estimator demonstrates high accuracy under various channel conditions and has neglectable impact on the support detection accuracy.

Table 3-A: List of main notations.

Notation	Description	Notation	Description
K	Number of users	λ_{mk}	Spreading code of user k on sub-carrier m
M	Number of sub-carriers	x_k	Intended signal of user k
T	Number of time slots (TSs)	g_{mk}	Channel gain of user k over sub-carrier m
S	User sparsity	$\mathbf{z}$	Latent vector

3.2 System Model

An uplink grant-free CD-NOMA system with K users and one BS is considered. Without loss of generality, all users and the BS are assumed to be equipped with a single antenna. The main notations are listed in Table 3-A.

By performing channel coding and modulation, the transmitted symbol x_k by user k is spread onto M orthogonal sub-carriers by an unique spreading sequence $\boldsymbol{\lambda_k} = (\lambda_{1k}, \lambda_{2k}, ..., \lambda_{Mk})^T \in \mathbb{C}^M$. Particularly, the overloaded CD-NOMA system is considered, i.e., $M < K$, where only a small proportion of users are actively transmitting signals at a given TS. For inactive users, their transmitted symbol is treated as zero. Hence, the signals received at the BS over sub-carrier m can be expressed individually as

$$y_m = \sum_{k=1}^{K} g_{mk}\lambda_{mk}x_k + n_m, \quad m = 1, 2, ..., M \ , \tag{3.1}$$

where g_{mk} denotes the channel gain of user k transmitted over sub-carrier m and $n_m \sim \mathcal{CN}(0, \sigma^2)$ denotes the Gaussian noise with noise power σ^2. The Rayleigh fading channel model is adopted whose channel gains are independent and identically distributed (i.i.d.) complex Gaussian random variables, i.e., $g_{mk} \overset{i.i.d.}{\sim} \mathcal{CN}(0, 1), \forall m, k$. To simplify the expression, the received signal vector $\mathbf{y} = (y_1, ..., y_M)^T$ can be expressed as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{3.2}$$

where $\mathbf{x} = (x_1, ..., x_K)^T$ is the transmitted signal vector, $\mathbf{H}$ denotes the $M \times K$ channel matrix whose entries are $h_{mk} = g_{mk}\lambda_{mk}$, and $\mathbf{n} = (n_1, ..., n_M)^T$ is the Gaussian noise vector.

3.2.1 Frame-Wise Joint Sparsity Model

Generally, users transmit data in consecutive TSs and remain active or inactive throughout the time frame [27] [81], as shown in Fig. 3.1. Motivated by the temporal correlations in user activity, the system is extended from a single transmission to a multiple transmission model, known as the *frame-wise joint sparsity* model. Given a time frame of length T, i.e. T consecutive TSs, the common sparsity support $\mathcal{S}$ is defined as

$$supp(\mathbf{x}^{(1)}) = supp(\mathbf{x}^{(2)}) = \cdots = supp(\mathbf{x}^{(T)}) \triangleq \mathcal{S}, \tag{3.3}$$

where $supp(\mathbf{x}^{(t)}) = \{k \mid x_k^{(t)} \neq 0, k \in \{1, ..., K\}\}$. The number of active users during each transmission is defined as $S = |\mathcal{S}|$, which is referred to as the sparsity of the system. The length of the time frame is restricted to be shorter than the channel coherence time, so that the channel matrix $\mathbf{H}$ remains unchanged throughout the entire time frame. Thus, the formulation of the signals received over the T consecutive TS is given by

$$\mathbf{Y} = \mathbf{HX} + \mathbf{N}, \tag{3.4}$$

where $\mathbf{Y} = [\mathbf{y}^{(1)}, ..., \mathbf{y}^{(T)}] \in \mathbb{C}^{M \times T}$, $\mathbf{X} = [\mathbf{x}^{(1)}, ..., \mathbf{x}^{(T)}] \in \mathbb{C}^{K \times T}$ and $\mathbf{N} = [\mathbf{n}^{(1)}, ..., \mathbf{n}^{(T)}] \in \mathbb{C}^{M \times T}$. The evaluation of channel estimation methods is outside the scope of this chapter and our results can serve as a theoretical system performance benchmark. Hence, perfect CSI is assumed to be available at the BS.

3.2.2 Problem Formulation

Based on (3.4), the MUD problem becomes a 2-dimensional CS problem, where the goal is to estimate the signal matrix $\mathbf{X}$ given the channel matrix $\mathbf{H}$ and the received signal matrix $\mathbf{Y}$. The optimization problem is formulated as

$$(\mathbf{P1}): \quad \min_{\mathbf{X}} ||\mathbf{Y} - \mathbf{HX}||_2^2, \quad \text{subject to } ||\mathbf{X}||_0 = ST. \tag{3.5}$$

Existing signal recovery methods consist of two major approaches, namely convex optimization and greedy algorithms. Convex optimization methods find suboptimal solutions to (3.5) by replacing the ℓ_0 regularization with a ℓ_1 constraint [82]. This type of technique has the advantages of high recovery accuracy and theoretical performance guarantees at the cost of heavy computational complexity and high sensitivity to noise. Greedy algorithms identify the sparse supports and perform signal detection iteratively until the termination criteria are met [83]. They have lower complexity than convex optimization methods, but usually require a large number of measurements for exact recovery and are sensitive to noise.

It has been shown that deep learning-based recovery algorithms provide even lower computational complexity than greedy algorithms while achieving higher recovery accuracy [35]. In particular, generative networks have been recently studied in solving CS-based image recovery problems, which demonstrated significant performance gain compared to conventional techniques. Moreover, unlike typical neural networks which use the received signals as the inputs, generative networks use arbitrary noise as the inputs, hence the same network architecture can be used for different input signal dimensions. In terms of the considered communication system, the same neural network can be employed regardless of the number of sub-carriers in the system, which introduces more flexibility in implementation. Hence, this chapter aims to investigate the designs of generative neural networks for solving MUD problems.

3.2.3 Performance Metrics

To achieve a thorough performance evaluation, four distinct performance metrics are adopted, namely mean squared error (MSE), symbol error rate (SER), positive detection rate (P_d), and false alarm probability (P_{fa}). Let $\hat{x}_k^{(t)}$ denotes the recovered signal of user k in TS t and $\tilde{x}_k^{(t)}$ denotes the recovered symbol based on $\hat{x}_k^{(t)}$, the performance metrics are formulated as follows:

- MSE: MSE is defined as the average squared error loss between the recovered signal

and the true signal of all users, which is computed as

$$\text{MSE} = \frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T}|x_k^{(t)} - \hat{x}_k^{(t)}|^2. \tag{3.6}$$

- SER: SER is defined as the ratio of the incorrectly recovered symbols to all symbols transmitted by the active users, which is given by

$$\text{SER} = \frac{1}{ST}\sum_{k\in\mathcal{S}}\sum_{t=1}^{T}\mathbb{1}_{\{x_k^{(t)}\neq\ \tilde{x}_k^{(t)}\}}. \tag{3.7}$$

- P_d: P_d is defined as the ratio of the number of correctly detected active users to the total number of active users, which can be given by

$$P_d = \frac{1}{ST}\sum_{k\in\mathcal{S}}\sum_{t=1}^{T}\mathbb{1}_{\{\tilde{x}_k^{(t)}\neq\ 0\}}. \tag{3.8}$$

- P_{fa}: P_{fa} is defined as the ratio of the number of inactive users who is detected as active to that of the total inactive users, given by

$$P_{fa} = \frac{1}{(K-S)T}\sum_{k\notin\mathcal{S}}\sum_{t=1}^{T}\mathbb{1}_{\{\tilde{x}_k^{(t)}\neq\ 0\}}. \tag{3.9}$$

3.3 Generative Networks for Multi-User Detection (GenMUD)

In this section, the offline model-agnostic meta-learning (MAML)-based training procedure of generative networks [79] is introduced and the network architecture is outlined, followed by the proposed GenMUD framework and the designed sparsity estimator.

3.3.1 Problem Reformulation

In existing deep learning-based MUD algorithms, the signal received at the BS is used as the input of the neural network. Hence, the dimension of the received signal has to be identical in training and in application due to the fixed network architecture. For instance, in the considered network, the dimension of the received signal is $(M \times T)$, where M is the number of sub-carriers and T is the length of the time frame. Therefore, if the number of sub-carriers varies during application, separate neural networks need to be trained and stored at the BS, which leads to excessive computational and memory cost.

In contrast to conventional neural networks, the generative neural network learns a mapping from the latent space to the space of all possible transmitted signals, which is formulated as

$$\mathbf{X} = G_\theta(\mathbf{z}), \tag{3.10}$$

where $\mathbf{z}$ represents a point in the latent space of arbitrary dimensions and G_θ represents a generative neural network, also known as a generator. Since the dimension of $\mathbf{z}$ can be arbitrary, the same neural network can be employed in systems with a varying number of sub-carriers, hence motivating the design of generative neural network-based MUD frameworks.

Hence, the generative neural network solves (3.5) by searching for the particular point $\mathbf{z}$ such that the neural network G_θ maps $\hat{\mathbf{z}}$ to the optimal solution $\mathbf{X}$ in (3.5). It leads to the following reformulated MUD problem:

$$(\mathbf{P2}): \quad \min_{\mathbf{z}} ||\mathbf{Y} - \mathbf{H}G_\theta(\mathbf{z})||_2^2, \quad \text{subject to } ||G_\theta(\mathbf{z})||_0 = ST. \tag{3.11}$$

To achieve the optimal solution $\hat{\mathbf{z}}$ in (3.11), gradient descent is initiated starting from

a randomly sampled point $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ through the following update equation:

$$\hat{\mathbf{z}} \leftarrow \hat{\mathbf{z}} - \alpha \frac{\partial ||\mathbf{Y} - \mathbf{H}G_\theta(\mathbf{z})||_2^2}{\partial \mathbf{z}}, \tag{3.12}$$

where α indicates the learning rate. However, it usually requires hundreds or even thousands of gradient descent steps until (3.12) converges. Fortunately, model-agnostic meta-learning (MAML) [70] can be employed in the training phase to further optimize this optimization procedure. As demonstrated in the simulations, the number of gradient descent steps can be reduced to as few as 20 iterations.

To formulate MAML in the context of MUD, the distribution of tasks is denoted by $p_{\text{task}}(\mathcal{T})$, in which each task $\mathcal{T}_i$ describes the optimization problem of finding the optimal $\hat{\mathbf{z}}_i$ that approximates the target signal matrix $\mathbf{X}_i$. Thus, MAML is employed by training the network weights $\boldsymbol{\theta}$ against the measurement error over all tasks, denoted by $\mathcal{L}_G$, through a second order differentiation, which is formulated as

$$\min_{\theta} \mathcal{L}_G, \quad \text{for } \mathcal{L}_G = \mathbb{E}_{\mathcal{T}_i \sim p_{\text{task}}(\mathcal{T})} \left[||\mathbf{Y}_i - \mathbf{H}_i G_\theta(\hat{\mathbf{z}}_i)||_2^2 \right], \tag{3.13}$$

where $\hat{\mathbf{z}}_i$ is the output of task $\mathcal{T}_i$ after iteratively performing the gradient descent steps in (3.12). It is worth pointing out that, the second order differentiation in (3.13) is performed by back-propagating through the gradient descent steps of all $\hat{\mathbf{z}}_i$. Additionally, to further increase the convergence rate, the same optimization procedure is applied to the learning rate α.

However, the network may exploit (3.13) and quickly approach small loss by mapping all $G_\theta(\hat{\mathbf{z}})$ into the null space of $\mathbf{H}$, which leads to divergence. To address this problem, an additional loss is designed to enforce the RIP condition in the CS theory. The RIP loss is given by

$$\mathcal{L}_H = \mathbb{E}_{\mathbf{X}^\star, \mathbf{X}'} \left[\left(\left|\left| \mathbf{H}\mathbf{X}^\star - \mathbf{H}\mathbf{X}' \right|\right|_2 - \left|\left| \mathbf{X}^\star - \mathbf{X}' \right|\right|_2 \right)^2 \right], \tag{3.14}$$

where $\mathbf{X}^{\star}$ and $\mathbf{X}^{'}$ are the samples of the signal in different stages of the optimization process. Specifically, the RIP loss is computed as an average over three pairs of signals, namely the true signal and the initial random signal before executing (3.12), the true signal and the optimized signal after executing (3.12), the initial random signal and the optimized signal of (3.12).

The training algorithm of the generative network, as illustrated in Algorithm 1, is described as follows: In each iteration, the received signal $\mathbf{Y}_i$ of each task $\mathcal{T}_i$ is first measured based on the true signal $\mathbf{X}_i$, the channel matrix $\mathbf{H}_i$, and the random noise $\mathbf{n}_i$; Then, for each task $\mathcal{T}_i$, an initial latent variable is sampled according to the Gaussian distribution and J gradient steps are performed using (3.12) to obtain the optimized latent variable $\hat{\mathbf{z}}_i$ and the corresponding signal recovery $G_\theta(\hat{\mathbf{z}}_i)$; By calculating the measurement loss $\mathcal{L}_G$ and the RIP loss $\mathcal{L}_H$ as an average over all tasks, the network weights θ and the learning rate α are updated through a second order differentiation over the J gradient steps; The previous procedures are repeated until convergence.

Algorithm 1 Generative Network Training Algorithm

Input: Number of training tasks N_d, training data $\{\mathbf{X}_i\}_{i=1}^{N_d}$, channel matrix $\{\mathbf{H}_i\}_{i=1}^{N_d}$, generator G_θ, number of latent update steps J

Output: Trained generator $G_{\hat{\theta}}$, optimized learning rate $\hat{\alpha}$

Initialize θ, α

repeat
  for $i = 1$ to N_d **do**
    Measure the signal $\mathbf{Y}_i \leftarrow \mathbf{H}_i\mathbf{X}_i + \mathbf{n}_i$
    Sample $\hat{\mathbf{z}}_i \sim \mathcal{N}(0, \mathbf{I})$
    for $j = 1$ to J **do**
      $\hat{\mathbf{z}}_i \leftarrow \hat{\mathbf{z}}_i - \alpha \frac{\partial}{\partial \hat{\mathbf{z}}_i} ||\mathbf{Y}_i - \mathbf{H}G_\theta(\hat{\mathbf{z}}_i)||_2^2$
    end for
  end for
  $\mathcal{L}_G = \frac{1}{N_d} \sum_{i=1}^{N_d} ||\mathbf{Y}_i - \mathbf{H}G_\theta(\hat{\mathbf{z}}_i)||_2^2$
  Compute $\mathcal{L}_H$ using (3.14)
  Update $\theta \leftarrow \theta - \frac{\partial}{\partial \theta}(\mathcal{L}_G + \mathcal{L}_H)$
  Update $\alpha \leftarrow \alpha - \frac{\partial}{\partial \alpha}(\mathcal{L}_G + \mathcal{L}_H)$
until reaches the maximum training steps
Return $G_{\hat{\theta}}$, $\hat{\alpha}$

Algorithm 2 GenMUD Algorithm

Input: Received signal $\mathbf{Y}$, channel matrix $\mathbf{H}$, pre-trained generator $G_{\hat{\theta}}$, number of latent update steps T, optimized learning rate $\hat{\alpha}$

Output: Reconstructed symbols $\tilde{\mathbf{X}}$

1: Sample $\hat{\mathbf{z}} \sim \mathcal{N}(0, \mathbf{I})$
2: **for** $t = 1$ to T **do**
3: $\quad \hat{\mathbf{z}} \leftarrow \hat{\mathbf{z}} - \hat{\alpha} \frac{\partial}{\partial \hat{\mathbf{z}}} ||\mathbf{Y} - \mathbf{H} G_{\hat{\theta}}(\hat{\mathbf{z}})||_2^2$
4: **end for**
5: Final signal reconstruction $\hat{\mathbf{X}} = G_{\hat{\theta}}(\hat{\mathbf{z}})$
6: Initialize symbol reconstruction $\tilde{\mathbf{X}} = \mathbf{0}_{K \times T}$
7: **for** $t = 1$ to T **do**
8: $\quad$ Order the users in terms of the magnitude of the recovered signal, i.e., $\hat{x}_{(1)}^{(t)} \geq \hat{x}_{(2)}^{(t)} \geq \cdots \geq \hat{x}_{(K)}^{(t)}$
9: $\quad$ Update $\tilde{\mathbf{X}}$ by mapping the signals of the first S users in the ordered list to the nearest symbol, i.e., $\tilde{x}_{(s)}^{(t)} = \{\hat{x}_{(s)}^{(t)}$ mapped to the nearest symbol$\}, s = 1, \ldots, S$
10: **end for**
11: **Return** $\tilde{\mathbf{X}}$

3.3.2 GenMUD Framework

Algorithm 2 illustrates the proposed GenMUD framework. Having obtained the trained generator $G_{\hat{\theta}}$ and the optimized learning rate $\hat{\alpha}$ from Algorithm 1, MUD is achieved by executing (3.12) for J iterations starting from a random latent variable to obtain the optimized latent variable and the corresponding signal recovery, which corresponds to line 1-5 of Algorithm 2. Since neural networks output continuous numbers, the recovered signals need to be further mapped to valid modulation symbols, represented by lines 6-10 of Algorithm 2. To be specific, in each TS, the signals are sorted in descending order of magnitude and the S signals with greater magnitudes are each mapped to the nearest constellation symbol. The rest $(K - S)$ signals are treated as $(0 + 0j)$, which represents the inactive users.

3.3.3 Network Architecture

Standard generative neural networks [80] are mainly designed for image data to extract spatial correlations and hidden structures. To be applied to MUD data which lack representative features, larger input dimensions are required to achieve greater learn-

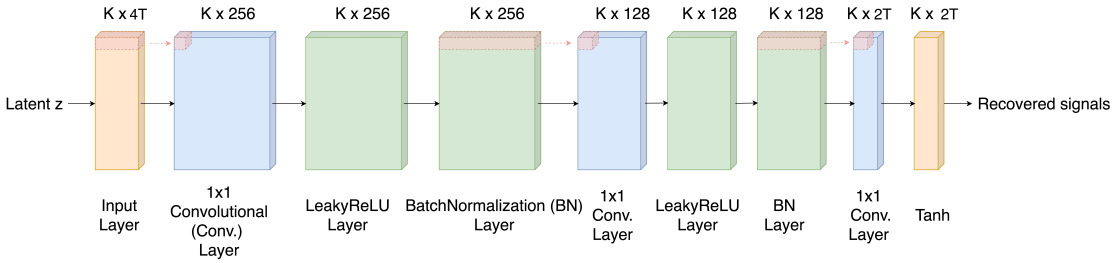


Figure 3.2: Architecture of the developed neural network.

ing capability. However, the use of fully-connected layers in conventional generative networks causes the network sizes to increase dramatically with the input dimensions, leading to high computational complexity. Moreover, since the considered model assumes independent user behaviours, i.e., the active status of one user does not influence the active/inactive state of any other user. Hence, the large kernel sizes used in convolutional neural network (CNN)-based networks [80] introduce redundant network parameters when applied to MUD data. Therefore, 1x1 convolutions are employed to construct a low-complexity generative network, which is tailored specifically for the considered MUD problems.

The detailed network architecture is described as follows. Based on the $K \times 2T$ output dimension, which corresponds to $[\Re(\mathbf{X}), \Im(\mathbf{X})]$, the input dimension is chosen as $K \times 4T$ to improve the network learning capability. Three one-dimensional (1D) convolutional layers are employed, each using a kernel size of 1×1. LeakyReLU activation and batch-normalization are employed after each convolutional layer except the last one, which employs the tanh activation function to scale the outputs to [-1, 1]. The architecture of the proposed network is illustrated in Fig. 3.2.

The proposed architecture has three main differences compared to standard generative networks [79, 80]:

- **Large input dimensions**: For generative neural networks, larger input dimensions are proved to improve the learning capability of the network at the cost of computational complexity [84]. Small latent dimensions are preferred in traditional

image CS problems because image datasets usually exhibit strong inter-correlations and can be compressed to smaller latent dimensions without much information loss. However, in the considered system model, users are expected to enter the communication system independently, causing zero correlation among the row of the signal matrix $\mathbf{X}$. Moreover, the signals transmitted by the active users are assumed to be independent in each TS. The lack of predictable features causes the generative networks to require a larger latent dimension. Given the considered model, the latent dimension is chosen to be twice the size of the output, since the network also needs to be adaptable to a different number of sub-carriers.

- **No fully-connected layers**: Fully-connected layers scale badly with input sizes, hence are computationally expensive for solving the considered MUD problems. In the proposed network, the input layer is followed directly by a convolutional layer to reduce the additional complexity cost due to the increase in the input dimension.

- **1x1 convolutional layers**: Considering the independent user behaviours, any spatial convolutions should be avoided among the users. In particular, all the 2D convolutional layers are replaced with 1x1 1D convolutional layers, such that the output of the resulting network does not exhibit any inter-user correlation.

3.3.4 Complexity Analysis

In both the training algorithm and the GenMUD algorithm, the majority of the computational complexity arises from the latent updates step, corresponding to lines 5-7 in Algorithm 1 and lines 2-4 in Algorithm 2. In each latent update step, the computational complexity mainly consists of the multiplication complexity of $\left(\mathbf{H} \times G_{\hat{\theta}}(\hat{\mathbf{z}})\right)$ and the forward propagation complexity of $G_{\hat{\theta}}(\hat{\mathbf{z}})$. The multiplication has a complexity of $\mathcal{O}(MKT)$. Then, for the 1×1 convolutional network with an input dimension of $4KT$ and an output dimension of $2TK$, the complexity is derived as $\mathcal{O}(KTL+n_H KL^2)$, where n_H and L denotes the number of hidden layers and the number of 1x1 kernels in each layer, respectively. Without loss of generality, it is assumed that all convolutional layers

have the same number of kernels. Hence, the total complexity of the J latent update steps is derived as $\mathcal{O}(J(MKT + KTL + n_H KL^2))$, which constitutes the computational complexity of the training algorithm and the GenMUD algorithm.

To demonstrate the complexity benefits of the designed network architecture, the deep convolutional generative adversarial network (DCGAN) [80] is considered as a baseline model. In terms of the MUD data, consider a DCGAN that compose of a fully connected layer and n_H 1D convolutional layers. Each convolutional layer consists of L kernels of size $s_k \times 1$ and has a stride of 2. Therefore, the computational complexity of the considered DCGAN is $\mathcal{O}\left(4K^2TL + n_H L^2 s_k(\lfloor \frac{K-s_k}{2} \rfloor + 1) + 2KTLs_k(\lfloor \frac{K-s_k}{2} \rfloor + 1)\right) = \mathcal{O}\left(K^2TL + n_H L^2 s_k \lfloor \frac{K-s_k}{2} \rfloor + KTLs_k \lfloor \frac{K-s_k}{2} \rfloor\right)$, which is significantly larger than that of the designed network.

3.3.5 Sparsity Estimator

The proposed GenMUD framework, along with many existing MUD algorithms, relies on the knowledge of sparsity prior to user detection. However, the exact sparsity is difficult to retrieve in practical systems, which makes sparsity estimation a challenging topic in MUD problems. Recently, some MUD algorithms are proposed with sparsity-blind strategies, including approximation algorithms for sparsity built inside the MUD algorithms [27] and stopping criteria of greedy algorithms with no prior knowledge of user sparsity [81]. Unfortunately, these sparsity approximation methods do not apply to other MUD frameworks.

By inspecting the equation (3.4) of the system model, a straightforward observation is that the power of each received signal is directly proportional to sparsity and inversely proportional to SNR. It raises the question of whether a closed-form relationship between the received signal, the sparsity, and the SNR can be formulated. Unfortunately, the randomness in noise and channel gain makes a direct formulation impossible. However, since the distribution of noise and channel gain is known, it is reasonable to suspect that the power of the received signals can also be modelled as a random variable with mean

Table 3-B: Network and algorithm configurations.

Parameter	Value	Parameter	Value
Total number of users, K	200	SNR	20 dB
Number of sub-carriers, M	100	Initial latent learning rate, α	0.01
Number of active users, S	40	Network learning rate	0.0001
Number of time slots, T	7	Task batch size	32

and variance expressed as a function of related quantities, namely sparsity. Based on extensive experiment results and data analysis, a sparsity estimator is formulated as

$$\hat{S} = \mathbb{E}\left[\frac{\tau}{2(\tau+1)}||\mathbf{y}||_2^2\right], \tag{3.15}$$

where τ is SNR in its linear scale and $\mathbf{y}$ is the $M \times 1$ signal vector received in a single TS. In terms of a frame-wise model of length T, the estimator is formulated as

$$\hat{S} = \frac{1}{T}\sum_{t=1}^{T}\frac{\tau}{2(\tau+1)}||\mathbf{y}^{(t)}||_2^2. \tag{3.16}$$

The estimation accuracy is evaluated by the normalized error (E_n), given by

$$E_n = \frac{|S-\hat{S}|}{S}. \tag{3.17}$$

3.4 Simulations

Unless otherwise stated, the configurations of the communication system and the network training parameters are listed in 3-B. It is worth pointing out that the SNR value is fixed to 20 dB and the sparsity is fixed to $S = 40$ in all training sessions, where the simulation results are produced under various SNRs and sparsity levels.

The transmitted signals are modulated by Quadrature Phase Shift Keying (QPSK). The codebook is designed as a random Toeplitz matrix [85] such that RIP is satisfied and CS can be implemented to ensure full signal reconstruction with a high probability.

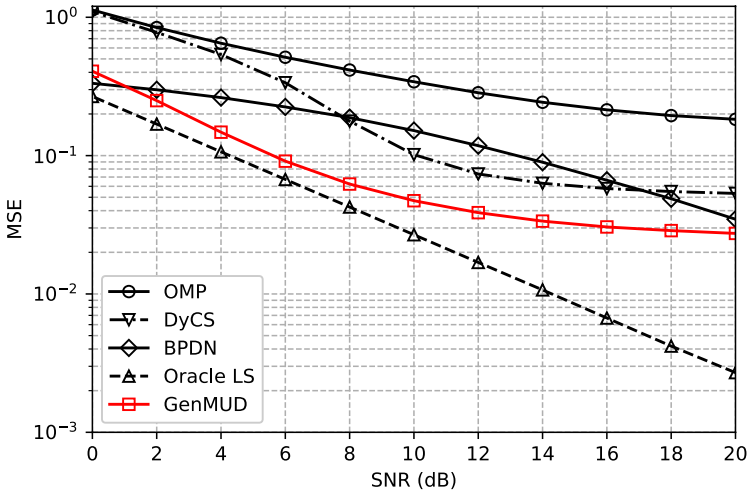


Figure 3.3: MSE versus SNR based on OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD.

3.4.1 Benchmark Methods

In this section, the MUD performance of the proposed GenMUD algorithm is investigated and compared to several existing solutions: 1) Oracle least squares (LS) algorithm, the widely used MUD performance upper bound, is chosen as the optimal solution since it performs LS estimation based on perfect knowledge of the sparse support.; 2) Orthogonal matching pursuit (OMP) [83], a greedy algorithm; 3) Basis pursuit de-noising (BPDN) [82], a convex optimization technique; 4) DyCS [25], a state-of-the-art MUD method based on dynamic CS.

3.4.2 MSE Performance

Fig. 3.3 compares the MSE performance of all considered methods, OMP, DyCS, BPDN, Oracle LS, and the proposed GenMUD, under different SNRs. It can be observed that the proposed GenMUD outperforms OMP and DyCS under all SNR values, which indicates a consistent and accurate signal detection performance for all considered values of SNR. Compared to BPDN, the proposed GenMUD demonstrates increasing performance gain as SNR increases from 0 dB to 10 dB. Moreover, GenMUD achieves almost the same MSE as Oracle LS between 0 dB and 8 dB, indicating the highly accurate support detection

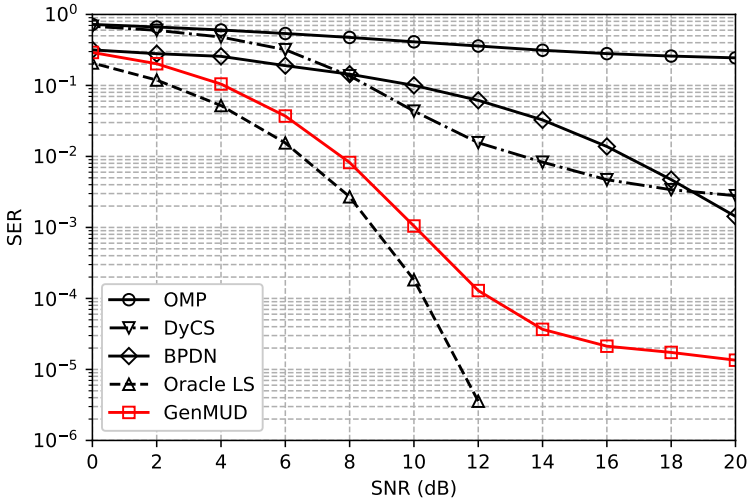


Figure 3.4: SER versus SNR based on OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD under $S = 40$ active users and $M = 100$ sub-carriers.

of GenMUD.

3.4.3 SER Performance

To investigate the symbol detection accuracy, Fig. 3.4 illustrates the SER performance of the considered algorithms under different SNRs. In the simulations. It can be noticed that the proposed GenMUD algorithm outperforms OMP and DyCS significantly in terms of SER over the whole range of SNR. In comparison to the BPDN approach, GenMUD shows increasing performance gain as SNR increases from 0 dB to 14 dB. Moreover, GenMUD demonstrates near oracle performance when SNR is lower than 8 dB and is the only method other than LS that achieves an SER lower than 1×10^{-3}.

By comparing the MSE and the SER performance of the considered methods in Fig. 3.3 and Fig. 3.4, an interesting insight is that the SER performance gain of GenMUD compared to OMP, DyCS, and BPDN is more significant than its MSE performance gain over these methods. This observation implies that, although the signals recovered by GenMUD are not precisely accurate, they are generally very close to the true signals and can be mapped to true symbols with higher accuracy than the other methods. Moreover,

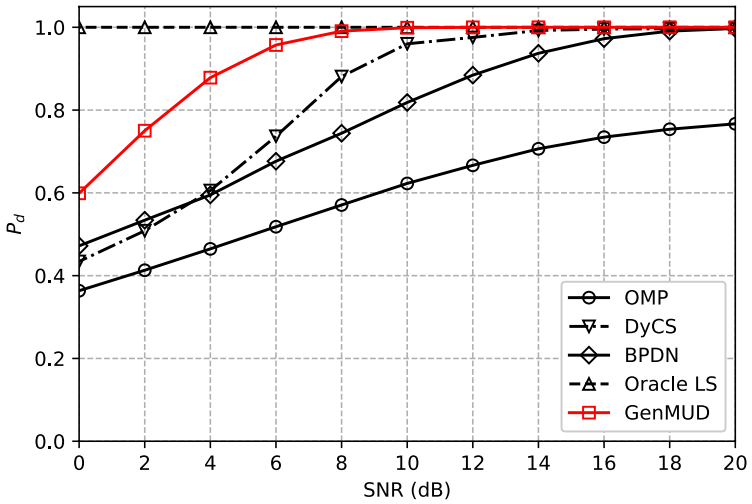


Figure 3.5: Positive detection rate (P_d) versus SNR based on OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD.

the neural network of GenMUD is trained at a fixed SNR of 20 dB but still demonstrates substantial performance gain compared to the conventional methods over other SNR values, indicating a strong generalization capability to tolerant noise variations.

3.4.4 Positive Detection Rate (P_d) Performance

Fig. 3.5 illustrates the positive detection rate, P_d, versus SNR. The number of active users is $S = 40$ and the number of sub-carriers is $M = 100$. Results show that the proposed GenMUD achieves the highest P_d among all methods, except for oracle LS which assumes perfect knowledge of the sparsity support. DyCS and BPDN achieve near 100% accuracy after SNR reaches 14 dB and 18 dB, respectively, whereas the proposed GenMUD achieves nearly 100% detection accuracy immediately after SNR achieves 10 dB. In other words, as SNR decreases from 20 dB to 10 dB, GenMUD demonstrates a consistent near-optimal detection accuracy, whereas other methods all suffer from observable growth in detection errors.

3.4.5 False Alarm Rate (P_{fa}) Performance

Fig. 3.6 compares the performance of the false alarm probability, P_{fa}, against SNR among the considered methods. It can be noticed that the proposed GenMUD achieves

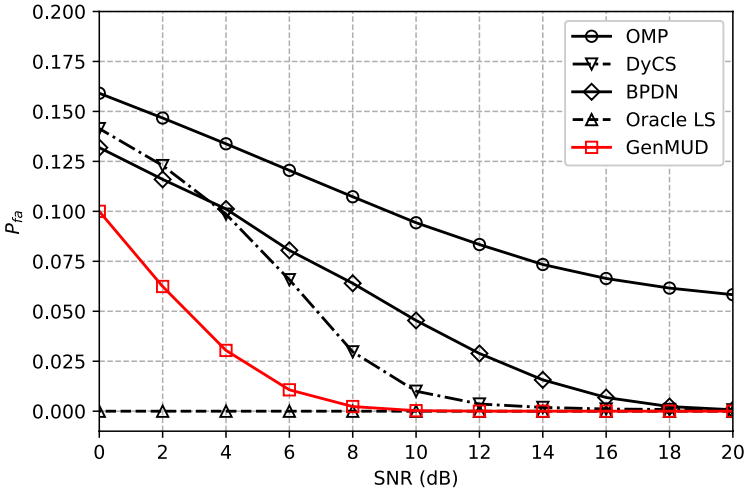


Figure 3.6: False alarm probability (P_{fa}) versus SNR based on OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD.

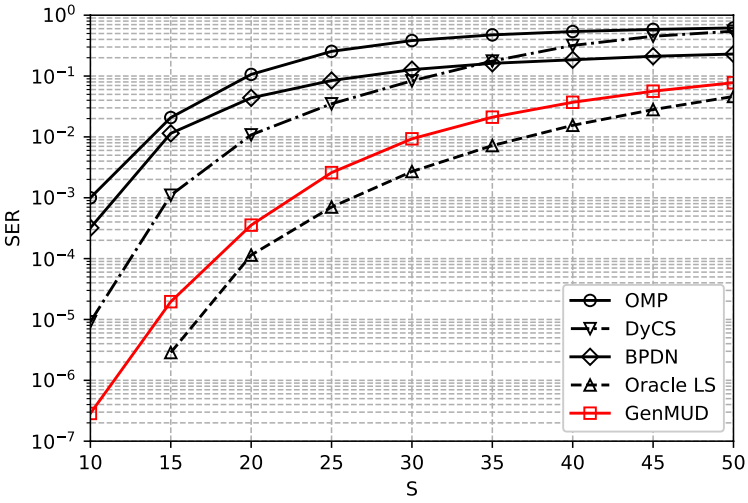


Figure 3.7: SER comparison of OMP, DyCS, BPDN, Oracle LS and the proposed GenMUD versus different number of active users under 6 dB SNR.

the lowest P_{fa} compared to OMP, BPDN and DyCS, and approaches zero P_{fa} as SNR increases beyond 10 dB. The consistent results in Fig. 3.5 and Fig. 3.6 validate that the proposed GenMUD is capable of accurately identifying active users from inactive users for all SNRs.

Fig. 3.7 illustrates the influence of user sparsity on SER performance under SNR =

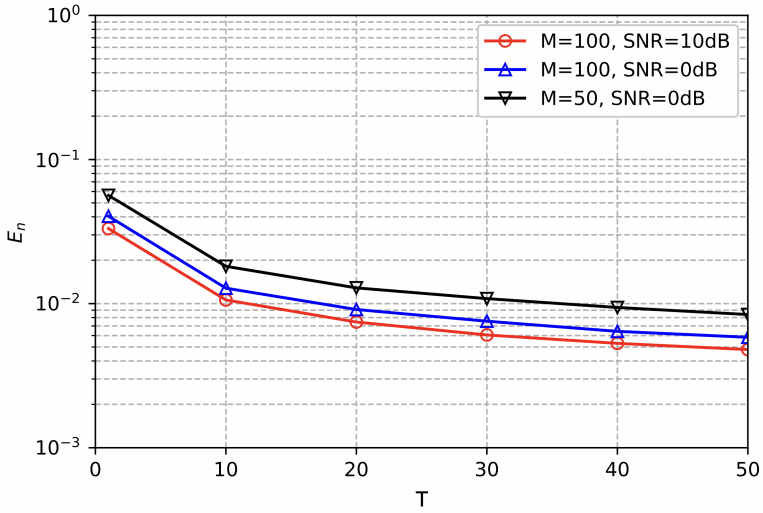


Figure 3.8: Normalized error (E_n) versus number of time slots of the proposed sparsity estimator under different numbers of sub-carriers and SNRs.

6 dB. For all the methods, the SER performance degrades as the number of active users increases, since the detection difficulty increases. However, the proposed GenMUD exhibits consistently lower SERs than OMP, DyCS and BPDN throughout the range of SNR. Given that the proposed neural network is trained under $S = 40$ active users, the consistent performance gains of the proposed method compared to its counterparts imply that the network has precisely captured the underlying relationships between user activity and the received signals.

3.4.6 Sparsity Estimation Performance

Fig. 3.8 depicts the normalized error, E_n, performance against frame length T of the proposed sparsity estimator in (3.16). Results are plotted for a different number of sub-carriers M and SNRs. Among all system settings, the sparsity estimation demonstrates a maximum normalized error of 0.05, which indicates a generally low estimation error. As the frame length increases from 1 to 50, E_n of the proposed estimator decreases gradually from around 0.05 to below 0.01, indicating an inverse relationship between estimation error and the number of TSs T, which is a valuable insight for practical applications. To be specific, in practical scenarios, the BS can perform an online update of the estimated

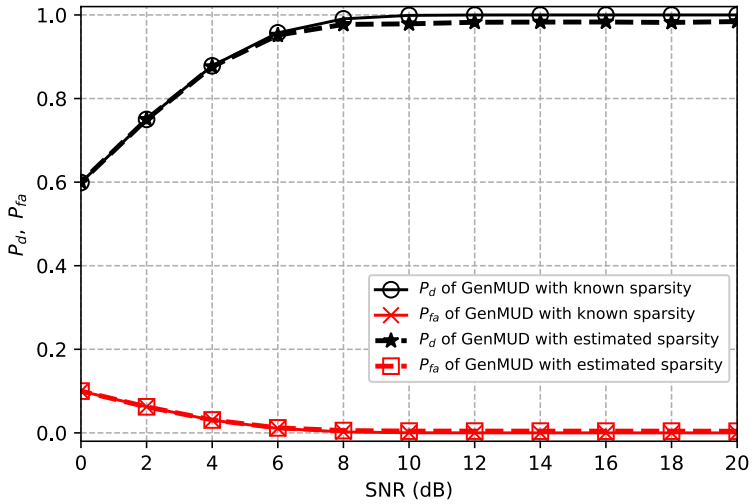


Figure 3.9: Detection probability (P_d) and false alarm probability (P_{fa}) versus SNR of the proposed GenMUD with known sparsity and estimated sparsity.

sparsity as new signals are received to preserve the low estimation error. It is also observed that there is little increase in E_n as the communication environment degrades, i.e., as M decreases from 100 to 50 and SNR decreases from 10 dB to 0 dB, which demonstrates the robustness of the proposed estimator to the varying environment.

To further investigate the influence of the sparsity estimator on MUD, Fig. 3.9 illustrates the P_d and P_{fa} performance of the proposed GenMUD with known sparsity and the estimated sparsity. It can be observed that the proposed GenMUD with the sparsity estimator achieves almost identical P_{fa} performance as when sparsity is known for all values of SNR. If the SNR is smaller than 6 dB, the sparsity estimator can be employed with no impact on P_d performance. For SNR values greater than 6 dB, the P_d performance difference between known and estimated sparsity is unnoticeable. Hence, it can be concluded that the proposed sparsity estimator provides an accurate approximation and has a neglectable influence on support detection performance.

3.5 Summary

In this chapter, a generative neural network-based MUD (GenMUD) framework was proposed. By identifying the uncorrelated user activity relationships, the proposed generative network was designed based on only 1x1 convolutional layers and no fully connected layers to achieve higher learning capability at a low additional network complexity cost. Moreover, a sparsity estimator was designed based on the received signal and the SNR level, both of which are easy to obtain in practical systems. This estimator can be employed as an add-on utility to existing MUD algorithms for realizing sparsity blind MUD. Simulation results showed that the proposed GenMUD method provided better detection performance compared to conventional MUD approaches in terms of SER, detection probability and false alarm probability. Experiments on the sparsity estimator proved the low estimation error and demonstrated the negligible impact of the estimator on MUD performance under various communication settings. In the next chapter, the critical resource allocation problem in downlink NOMA systems will be investigated and an adaptive NGMA scheme will be developed to jointly serve SDMA and PD-NOMA users.

Chapter 4

Adaptive NGMA Scheme for Energy-limited Networks: A Deep Reinforcement Learning Approach

4.1 Introduction

In this chapter, an adaptive NGMA scheme is proposed, which serves OMA and PD-NOMA users with the same orthogonal time and frequency resource. Based on this scheme, the long-term power-constrained sum rate maximization problem is formulated, where the beamforming, the power allocation, and the user clustering are jointly optimized. In particular, a spatial correlation-based user clustering algorithm is proposed to transform the mixed-integer optimization problem into a continuous one. Then, a TRPO-based resource allocation algorithm is designed to solve the formulated long-term optimization problem, which demonstrates fast and stable training performance. The main contributions are outlined as follows:

Table 4-A: List of main notations.

Notation	Description	Notation	Description
K	Number of users	p_k	Transmit power of user k
N	Number of antennas	s_k	Intended signal of user k
T	Number of time slots (TSs)	σ_ϕ	Angular standard deviation
M	Number of clusters	$\mathcal{G}_m$	Set of users in cluster m
$\bar{\mathbf{w}}_k$	Normalized beamformer of user k	$\alpha_{k,l}^m$	Decoding order between user k,l
$P^{max,t}$	Maximum power at TS t	$\lambda_{k,m}$	Cluster allocation of user k
P^{max}	Maximum power over T TSs	θ_k	Nominal angle of user k

- To embrace the complementary benefits of the conventional SDMA and PD-NOMA schemes, an adaptive NGMA scheme is proposed, in which users are adaptively allocated to SDMA or PD-NOMA clusters to share the same orthogonal resources.

- The long-term sum rate maximization problem is investigated, where the power allocation, the beamforming, and the user clustering are jointly optimized. To transform the mixed-integer problem, a spatial correlation-based clustering algorithm is proposed based on the spatially correlated channels in the practical systems.

- A DRL-based resource allocation scheme is designed, where the TRPO learning algorithm is employed to ensure a fast and stable training process.

- Simulation results show that the proposed clustering method outperforms the channel condition-based method in terms of sum rate. Results also demonstrate the sum rate gain of the proposed NGMA scheme against the conventional SDMA and PD-NOMA schemes.

4.2 Network Model

4.2.1 Spatial Model

As illustrated in Fig. 4.1, a downlink MISO multi-user network is considered, where a N-antenna BS serves K single-antenna users. The locations of the BS and the users are represented in a 2-dimensional Cartesian coordinate system, where the BS is located at

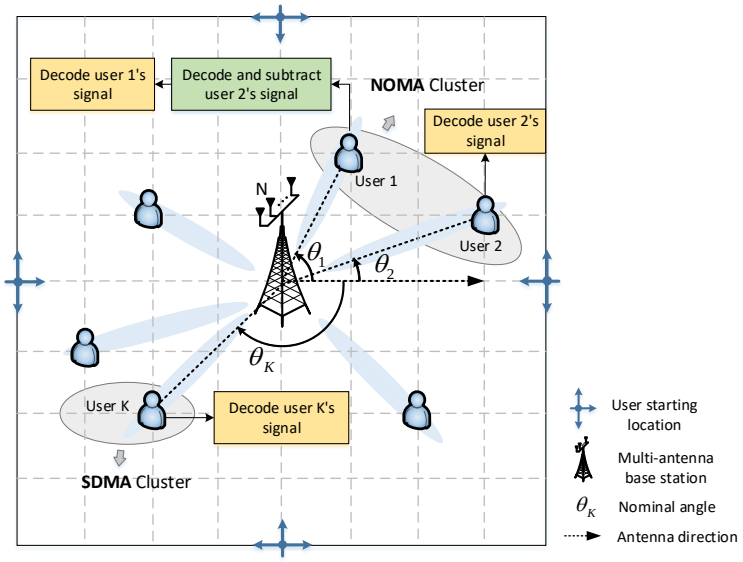


Figure 4.1: Illustration of the proposed adaptive NGMA-MISO downlink network. Users are grouped into clusters, where the users in the multi-user clusters employ SIC for decoding.

the origin with coordinates $[0, 0]$ and the coordinates of user k are denoted by $[c_{k,x}, c_{k,y}]$. A coverage area of D^2 meters for the BS is considered. Users are assumed to enter the area by travelling through the main roads, which is represented by the four starting locations at $[\frac{D}{2}, 0]$, $[-\frac{D}{2}, 0]$, $[0, \frac{D}{2}]$, and $[0, -\frac{D}{2}]$, respectively. After entering the coverage area, the mobility of each user is modelled as a random walk with fixed step size and an equal probability of travelling to one of the adjacent coordinates inside the coverage area. A nominal angle θ_k is defined for each user k, which is computed as the anti-clockwise angle deviation from the antenna direction to the LoS path between user k and the BS. The antenna direction is considered to be along the positive x-axis. Hence, the nominal angle θ_k of user k is formulated as $\theta_k = \tan(\frac{c_{k,y}}{c_{k,x}}) \in (-\pi, \pi)$. The main notations are listed in Table 4-A.

4.2.2 Channel Model

The downlink channel vector between the BS and user k is denoted by $\mathbf{h}_k^H \in \mathbb{C}^{1\times N}$. To simplify the analysis, it is assumed that $|\mathbf{h}_1^H| \geq |\mathbf{h}_2^H| \geq \cdots \geq |\mathbf{h}_K^H|$. The path loss between the BS and user k is formulated as $L_k = C_0 {d_k}^{-\alpha_{pl}}$, where d_k denotes the LoS

distance between user k and the BS, C_0 denotes the path loss intercept, and α_{pl} denotes the path loss exponent. Since the practical wireless communication environment has a finite number of scattering clusters, transmission channels are often spatially correlated, such that some spatial directions carry stronger signals than other directions. Therefore, a practical system is considered in this chapter, such that the received signal at the BS is the superposition of a large number of multipath components, where each multipath component reaches the BS from a particular angle similar to the nominal angle of the user. More specifically, this practical channel model follows the spatially correlated Rayleigh fading model, which is distributed as follows:

$$\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_k), \tag{4.1}$$

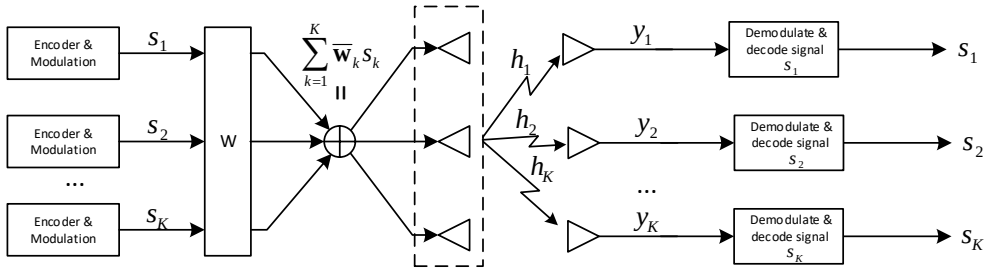
where the covariance matrix $\mathbf{R}_k$ advocates the local scattering model with Gaussian angular distribution [86]. The entries of $\mathbf{R}_k$ are given by

$$[\mathbf{R}_k]_{i,j} = \frac{L_k}{\sqrt{2\pi}\sigma_\phi} \int_{-\infty}^{+\infty} e^{j2\pi d_H (i-j)\sin(\theta_k+\delta)} e^{-\frac{\delta^2}{2\sigma_\phi^2}} d\delta, \tag{4.2}$$

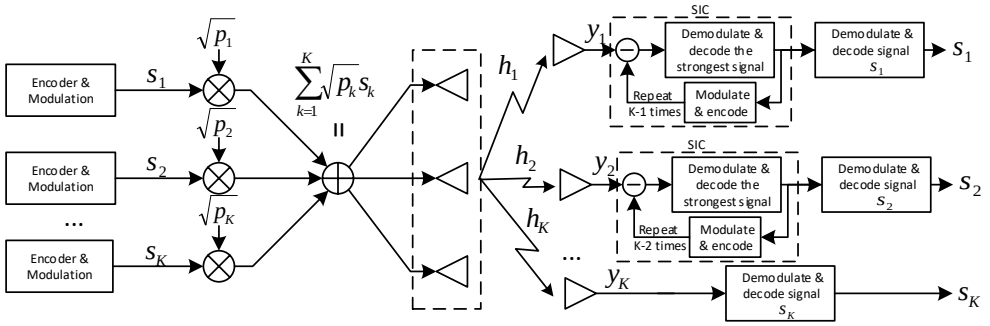
where d_H represents the antenna spacing, $\delta \sim \mathcal{N}(0, \sigma_\phi)$ describes the Gaussian distributed random deviation from the nominal angle, and σ_ϕ denotes the angular standard deviation.

4.2.3 Adaptive NGMA

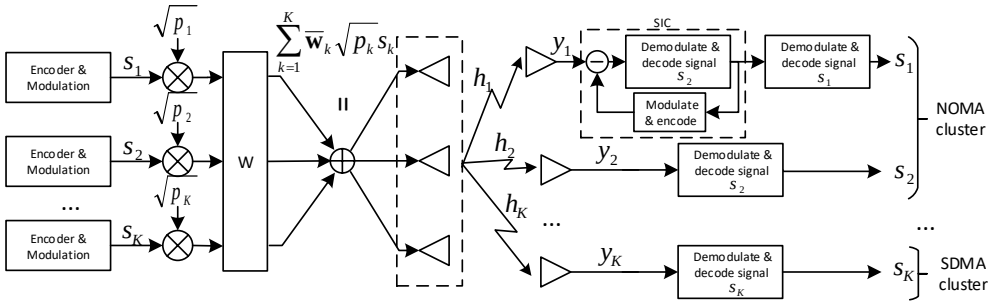
As illustrated in Fig. 4.2, conventional SDMA schemes rely on multi-user beamforming at the transmitters to eliminate the multi-user interference at the receivers, while the conventional PD-NOMA schemes utilize power domain multiplexing at the transmitters and SIC decoding at the receivers to improve the spectrum efficiency. Note that both schemes utilize the amplitude (power) and phase difference. This work aims to integrate them into one scheme. As a unified model, the proposed adaptive NGMA scheme employs both multi-user beamforming and power domain multiplexing at the transmitters. According to the user clustering outcome, the receivers either directly decode the intended signal



(a) SDMA.



(b) PD-NOMA.



(c) NGMA.

Figure 4.2: Block diagrams of the transmitter and the receiver in different transmission schemes, where s_k is the intended signal, p_k is the power allocation, and $\overline{w}_k$ is the normalized beamforming vector of user k: a) A SDMA scheme; b) A PD-NOMA scheme; c) The proposed NGMA scheme.

or employ SIC for decoding. Hence, the conventional SDMA and PD-NOMA schemes can be viewed as special cases of the proposed NGMA scheme. Moreover, in the proposed NGMA scheme, users can be adaptively allocated to PD-NOMA or SDMA clusters based on the knowledge of the CSI to embrace the complementary benefits of the conventional SDMA and PD-NOMA schemes, which can be used in any scales of networks,

e.g., $K < N, K \approx N, K > N$, etc. The parameters for the user clustering and SIC decoding in the proposed NGMA scheme are defined as follows:

4.2.3.1 User clustering

The transmission scheme allocated to each user is represented by the user clustering outcome, where the users in the single-user clusters employ SDMA and the users in the multi-user clusters employ PD-NOMA with a pre-defined SIC decoding order among the cluster members. The total number of clusters is denoted as M, constrained by $1 \leq M \leq N$. The set of users allocated to cluster $m \in \mathcal{M} \triangleq \{1, 2, \ldots, M\}$ is denoted by $\mathcal{G}_m$, where $\mathcal{G}_m \cap \mathcal{G}_n = \emptyset, \forall m \neq n \in \mathcal{M}$ and $\mathcal{G}_1 \cup \mathcal{G}_2 \cup \cdots \cup \mathcal{G}_K = \mathcal{K}$ are the clustering constraints that ensures no empty clusters and each user is allocated to only one of the clusters. Note that SIC is an interference-limited technique such that assigning a large number of users to one cluster can lead to severe decoding error propagation [87]. Therefore, a practical system of a maximum of two users per cluster is considered, i.e., $|\mathcal{G}_m| \leq 2, \forall m \in \mathcal{M}$.

4.2.3.2 Decoding order

The decoding order between user k and user i in cluster m is denoted by a binary variable $\alpha_{k,l}^m \in \{0, 1\}, \forall l \neq k \in \mathcal{G}_m, m \in \mathcal{M}$. In particular, $\alpha_{k,l}^m = 0$ indicates that, in cluster m, user k will employ SIC to decode user l's signal and remove it from the received signal, and $\alpha_{k,l}^m = 0$ indicates otherwise. Given the two-user cluster setup, the decoding order of each cluster is in the ascending order of channel gain, i.e., $\alpha_{k,l}^m = 1$ if $k > l$ and $\alpha_{k,l}^m = 0$ if $k < l$.

4.2.4 Signal Model

The signal intended for user $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$ is denoted by s_k and the beamforming vector for transmitting s_k is denoted by $\mathbf{w}_k \in \mathbb{C}^{N \times 1} = \sqrt{p_k}\ \overline{\mathbf{w}}_k$, where p_k and $\overline{\mathbf{w}}_k$ denote the allocated transmit power and the normalized beamforming vector, respectively. Hence, the signal vector $\mathbf{x} \in \mathbb{C}^{N \times 1}$ transmitted by the BS is formulated as

$\mathbf{x} = \sum_{k=1}^{K} \mathbf{w}_k s_k$.

On the receiver side, the SDMA users directly decode their intended signals by treating all other users' signals as noise, hence the signal received by a SDMA user k is given by

$$y_k^{\text{SDMA}} = \underbrace{\mathbf{h}_k^H \mathbf{w}_k s_k}_{\text{Desired signal}} + \underbrace{\sum_{j \in \mathcal{K}/\{k\}} \mathbf{h}_k^H \mathbf{w}_j s_j}_{\text{Inter-beam interference}} + \underbrace{\mathbf{n}}_{\text{noise}}. \tag{4.3}$$

The achievable rate of the SDMA user k is derived as

$$R_k^{\text{SDMA}} = \log_2 \left(1 + \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum\limits_{j \in \mathcal{K}/\{k\}} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_k^2} \right), \tag{4.4}$$

where σ_k^2 denotes the AWGN variance.

Since the PD-NOMA users decode the received signals based on a pre-defined SIC decoding order, the signal received by a PD-NOMA user k in cluster m, where $\mathcal{G}_m = \{k, l\}$, is formulated as

$$y_k^{\text{PD-NOMA}} = \underbrace{\mathbf{h}_k^H \mathbf{w}_k s_k}_{\text{Desired signal}} + \underbrace{\mathbf{h}_k^H \mathbf{w}_l s_l}_{\text{SIC signal}} + \underbrace{\sum_{j \in \mathcal{K}/\{k,l\}} \mathbf{h}_k^H \mathbf{w}_j s_j}_{\text{Inter-cluster interference}} + \underbrace{\mathbf{n}}_{\text{noise}}. \tag{4.5}$$

The achievable rate of the PD-NOMA user k in cluster m is derived as

$$R_k^{\text{PD-NOMA}} = \begin{cases} \log_2 \left(1 + \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum\limits_{j \in \mathcal{K}/\{k,l\}} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_k^2} \right), & \text{if } k > l \\ \log_2 \left(1 + \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{|\mathbf{h}_k^H \mathbf{w}_l|^2 + \sum\limits_{j \in \mathcal{K}/\{k,l\}} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_k^2} \right), & \text{if } k < l. \end{cases} \tag{4.6}$$

By utilizing the decoding order coefficient $\alpha_{k,l} \in \{0, 1\}$, (4.7) can be simplified into

$$R_k^{\text{PD-NOMA}} = \log_2 \left(1 + \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\alpha_{k,l}^m |\mathbf{h}_k^H \mathbf{w}_l|^2 + \sum\limits_{j \in \mathcal{K}/\{k,l\}} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_k^2} \right). \tag{4.7}$$

To simplify the expressions, by combining (4.3) and (4.5), the received signal of SDMA/PD-NOMA user k in cluster m can be formulated as

$$y_k = \underbrace{\mathbf{h}_k^H \mathbf{w}_k s_k}_{\text{Desired signal}} + \underbrace{\sum_{l \in \mathcal{G}_m/\{k\}} \mathbf{h}_k^H \mathbf{w}_l s_l}_{\text{SIC signal}} + \underbrace{\sum_{j \in \mathcal{K}/\mathcal{G}_m} \mathbf{h}_k^H \mathbf{w}_j s_j}_{\text{Interference}} + \underbrace{\mathbf{n}}_{\text{noise}}, \tag{4.8}$$

where the SIC signal in (4.8) becomes zero if user k is a SDMA user, since $\mathcal{G}_m/\{k\} = \{\}$.

Similarly, by combining (4.4) and (4.7), the achievable rate of user k in cluster m is given by

$$R_k = \log_2 \left(1 + \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum\limits_{i \neq k, i \in \mathcal{G}_m} \alpha_{k,i}^m |\mathbf{h}_k^H \mathbf{w}_i|^2 + \sum\limits_{j \in \mathcal{K}/\mathcal{G}_m} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_k^2} \right). \tag{4.9}$$

4.2.5 Problem Formulation

The objective is to maximize the total transmission sum rate over T transmission TSs by jointly optimizing the beamforming matrix $\mathbf{W}(t) = [\mathbf{w}_1(t), \ldots, \mathbf{w}_K(t)] \in \mathbb{C}^{N \times K}$ and the user clustering strategy $\mathcal{G}_m(t), \forall m \in \mathcal{M}$ of each TS $t = 1, \ldots, T$. To optimize the clustering strategy, it is reformulated into a binary user allocation matrix $\mathbf{\Lambda}(t) = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K]^T \in \mathbb{Z}^{K \times M}$, where $\lambda_{k,m}(t) = 1$ indicates that user k is allocated to cluster m in TS t and $\lambda_{k,m}(t) = 0$ indicates otherwise. Since each user can only be allocated to one of the clusters, the user allocation matrix is constrained by $\sum_{m \in \mathcal{M}} \lambda_{k,m}(t) = 1, \forall k \in \mathcal{K}, t = 1, \ldots, T$. As one of the critical performance targets of NGMA systems, low energy consumption can be achieved by enforcing a long-term average power constraint P_{max} [88, 89]. Hence, the BS is allowed to coordinate the power consumption among the

TSs to enhance the long-term total sum rate subject to the QoS constraints. Finally, the sum rate maximization problem is formulated as follows:

$$(\mathbf{P1}): \quad \max_{\{\mathbf{W}(t),\mathbf{\Lambda}(t)\}} \sum\nolimits_{t=1}^{T} \sum\nolimits_{k\in\mathcal{K}} R_k(t), \tag{4.10a}$$

$$\text{s.t.} \quad R_k(t) \geq R^{min}, \forall k, \tag{4.10b}$$

$$\sum\nolimits_{k\in\mathcal{K}} |\mathbf{w}_k(t)|^2 \leq P^{max,t}, \tag{4.10c}$$

$$\sum\nolimits_{t=1}^{T} \sum\nolimits_{k\in\mathcal{K}} |\mathbf{w}_k(t)|^2 \leq P^{max}, \tag{4.10d}$$

$$\sum\nolimits_{m\in\mathcal{M}(t)} \lambda_{k,m}(t) = 1, \forall k, t, \tag{4.10e}$$

$$\sum\nolimits_{k\in\mathcal{K}} \lambda_{k,m}(t) \leq 2, \forall m, t, \tag{4.10f}$$

where (4.10a) indicates the optimization objective, i.e., the transmission sum rate; (4.10b) denotes the minimum QoS constraint; (4.10c) represents the instantaneous transmit power constraint; (4.10d) represents the long-term total transmit power constraint; (4.10e) indicates that each user is allocated to only one of the clusters, and (4.10f) describes the constraint on the size of each cluster. Due to the integer-valued parameter $\mathbf{\Lambda}(t)$ and the total transmit power constraint in (4.10d), problem (**P1**) is indeed a long-term mixed-integer programming problem, which is non-trivial to be directly solved by standard convex optimization algorithms.

To tackle this challenging problem, the proposed optimization algorithm first transforms the binary user clustering variables $\mathbf{\Lambda}(t)$ into a continuous-valued clustering threshold, by identifying the characteristics of the spatially correlated channels. Then, a DRL-based resource allocation algorithm is employed to jointly optimize the instantaneous beamforming, the power allocation, and the clustering threshold, with respect to the long-term transmit power constraint.

4.3 DRL-based Resource Allocation for Adaptive NGMA

In this section, a DRL-based learning algorithm is designed to solve the long-term sum rate maximization problem in (**P1**). Firstly, it is demonstrated that the binary clustering variable can be represented by a single nominal angle threshold, which is continuous-valued. Then, the proposed DRL-based resource allocation scheme based on the TRPO learning algorithm is introduced.

4.3.1 Problem Reformulation

DRL is an efficient machine learning algorithm for maximizing the long-term rewards of time-varying environments. However, (**P1**) cannot be directly tackled by DRL algorithms, due to the mixed-integer variables. To address this problem, the binary clustering variables $\mathbf{\Lambda}(t)$ are transformed into continuous clustering variables, by exploiting the spatially correlated channel model and the superiority of PD-NOMA under correlated channels [18, 90, 91].

In the proposed system model, the spatially correlated channel model is employed, where the channel covariance matrix in (4.2) is formulated based on the nominal angle parameter $\theta_k \in [-\pi, \pi]$ of each user k, which is defined as the LoS angle between user k and the antenna direction. Therefore, users with similar nominal angles are likely to have strongly correlated channel vectors. Moreover, if the nominal angles of user k and user m satisfies $\theta_k + \theta_m = 2\pi$, their channel covariance matrices are the same, i.e., $\mathbf{R}_k = \mathbf{R}_m$. This is because the nominal angle is the input of a sine function. Hence, a regularized nominal angle $\theta_k^* = \tan(\frac{c_{k,y}}{|c_{k,x}|}) \in [-\pi/2, \pi/2]$ is defined for each user k, such that $\mathbf{R}_k = \mathbf{R}_m$ if and only if $\theta_k^* = \theta_k^*$ for any user k and m. To this end, instead of formulating a complicated channel correlation metric, the channel correlations between any two users can be evaluated based on the direct difference between their regularized nominal angles. Here, a threshold $\beta(t) \in [0, \pi]$ is proposed to control the nominal angle differences within all clusters at TS t, i.e., $|\theta_k^*(t) - \theta_l^*(t)| \leq \beta(t), \forall k, l \in \mathcal{G}_m(t), \forall m \in \mathcal{M}(t)$.

Algorithm 3 User Clustering Algorithm Based on Nominal Angle Threshold

Input: Regularized nominal angle $\boldsymbol{\theta}^* = [\theta_1^*, \ldots, \theta_K^*]$, nominal angle threshold β

Initialize $\boldsymbol{\Lambda} = \mathbf{0}_{K,N}, M = 1$,
for $k \in \mathcal{K}$ **do**
  if $\sum_{m=1}^{N} [\boldsymbol{\Lambda}]_{k,m} = 0$ **then**
    Allocate user k, i.e., $[\boldsymbol{\Lambda}]_{k,M} = 1$
    Find the candidates $\mathcal{C}_k := \{i \mid |\theta_i^* - \theta_k^*| \leq \beta, \sum_{m=1}^{N} [\boldsymbol{\Lambda}]_{i,m} = 0, i \in \mathcal{K}/\{k\}\}$
    Denote $i^* := \max(\mathcal{C}_k)$
    Pair user k with user i^*, i.e., $[\boldsymbol{\Lambda}]_{i^*,M} = 1$
  end if
  if $M < N$ **then**
    Update $M \leftarrow M + 1$
  else
    break
  end if
end for

The clustering algorithm based on $\beta(t)$ is described as follows: 1) Initialize the user allocation matrix as a zero matrix, i.e., $\boldsymbol{\Lambda}(t) = \mathbf{0}_{K,N}$; 2) Allocate user 1 to cluster 1, i.e., $\lambda_{1,1}(t) = 1$; 3) Find the set of candidates $\mathcal{C}_1 = \{i \mid |\theta_i^*(t) - \theta_1^*(t)| \leq \beta(t), \sum_{m=1}^{N} [\boldsymbol{\Lambda}]_{i,m} = 0, i \in \mathcal{K}/\{1\}\}$ to be paired with user 1, where $\mathcal{C}_1 = \{\}$ indicates a SDMA cluster; 4) Since each cluster can contain at most 2 users[1] and PD-NOMA is advantageous under large path loss differences, user 1 is paired with the user of the weakest channel gain in $\mathcal{C}_1$, i.e., $\lambda_{\max(\mathcal{C}_1),1} = 1$; 5) Repeat steps 2-4 for users $2, \ldots, K$ if they have not yet been allocated to a cluster.

The pseudocode of the proposed clustering algorithm is illustrated in Algorithm 3. The algorithm terminates when the number of non-empty clusters exceeds the number of antennas. Hence, if any user has not been allocated to a cluster after the algorithm terminated, the sum rate of the corresponding user will be regarded as zero. This situation is likely to happen in overloaded systems, i.e., $K > N$, with a poorly chosen β. Moreover, after the algorithm terminates, the zero columns in $\boldsymbol{\Lambda}(t)$ are discarded to reduce its dimension to $(K \times M)$, where M corresponds to the number of non-empty clusters.

[1]The proposed clustering scheme can be extended to systems with unbounded cluster size by allocating all candidates to the corresponding cluster.

By representing the user allocation matrix $\mathbf{\Lambda}(t)$ with the nominal angle threshold $\beta(t)$, (**P1**) can be reformulated into

$$(\mathbf{P2}): \quad \max_{\{\mathbf{W}(t),\beta(t)\}} \sum\nolimits_{t=1}^{T} \sum\nolimits_{k\in\mathcal{K}} R_k(t), \tag{4.11a}$$

$$\text{s.t.} \quad R_k(t) \geq R^{min}, \forall k \in \mathcal{K}, \tag{4.11b}$$

$$\sum\nolimits_{k\in\mathcal{K}} |\mathbf{w}_k(t)|^2 \leq P^{max,t}, \tag{4.11c}$$

$$\sum\nolimits_{t=1}^{T} \sum\nolimits_{k\in\mathcal{K}} |\mathbf{w}_k(t)|^2 \leq P^{max}. \tag{4.11d}$$

Now that both $\mathbf{W}(t)$ and $\beta(t)$ are continuous-valued variables, we can employ on-policy DRL algorithms to solve the long-term joint optimization problem in (**P2**).

4.3.2 Markov Decision Process (MDP)

To solve (**P2**) with DRL, it needs to be transformed into a MDP, which is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ are the state space and the action space, respectively. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the Markov transition probability, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and γ is the discount factor. Given a state $\mathbf{s}_t$, the agent chooses an action $\mathbf{a}_t$ according to a policy $\pi : \mathcal{S} \to \mathcal{A}$ and the objective of the agent is to find the optimal policy π^* that maximizes the expected return $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)]$.

In the proposed NGMA model, the BS acts as the agent, who makes decisions on beamforming and user clustering, and the adaptive NGMA system acts as the environment, which contains information about channel information and user locations. The aim of the BS is to maximize the total sum rate over the T TSs while ensuring the constraints in (4.11b), (4.11c), and (4.11d). The key elements of MDP in terms of the proposed NGMA model are described as follows:

4.3.2.1 State

The state at TS t is defined as $\mathbf{s}_t = [\mathbf{c}(t), \mathbf{h}^H(t), P(t), t]$. It consists of the user coordinates $\mathbf{c}(t) = [\mathbf{c}_{x,1}(t), \mathbf{c}_{y,1}(t), \ldots, \mathbf{c}_{x,K}(t), \mathbf{c}_{y,K}(t)]$, the channel vectors $\mathbf{h}^H(t) = [\mathbf{h}_1^H(t), \ldots, \mathbf{h}_K^H(t)]$,

the remaining power $P(t)$ at the BS, and the timestamp t. The state space has a dimension of $(2K + 2KN + 1)$.

4.3.2.2 Action

The action at TS t is defined as $\mathbf{a}_t = [\mathbf{W}(t), \beta(t)]$, which consists of the beamforming vectors $\mathbf{W}(t) = [\mathbf{w}_1(t), \ldots, \mathbf{w}_K(t)]$ and the nominal angle threshold $\beta(t)$. The action space has a dimension of $(2KN+1)$. Before executing the action, the agent first scales the beamforming vector, subject to $\sum_{k\in\mathcal{K}} |\mathbf{w}_k(t)|^2 \leq \min(P(t), P^{max,t})$. This ensures that the requested transmit power does not exceed the remaining power and the instantaneous power constraint in (4.11c) is enforced.

4.3.2.3 Transition probability

The probability of transitioning into a future state $\mathbf{s}_{t+1}$ depends on the action $\mathbf{a}_t$ and the state $\mathbf{s}_t$ of the current TS. Specifically, the state transition of the user coordinates $\mathbf{c}(t+1)$ is modelled as a 2-dimensional random walk based on the current coordinates $\mathbf{c}(t)$, where each user has an equal probability of moving to one of the adjacent coordinates within the coverage area. The state transition of the remaining transmit power $P(t + 1)$ is controlled by the beamforming vector $\mathbf{W}(t)$ through a deterministic power consumption formula, i.e., $P(t + 1) = P(t) - \sum_{k\in\mathcal{K}} |\mathbf{w}_k(t)|^2$, where $P(0) = P^{max}$.

4.3.2.4 Reward

At TS t, the instantaneous reward is computed as the total sum rate if all QoS requirements are satisfied. If the sum rate of any user is below the minimum QoS, the instantaneous reward is computed as the sum of QoS deficiency of all users. Hence, the reward function at TS t is formulated as

$$r_t(\mathbf{s}_t, \mathbf{a}_t) = \begin{cases} \sum_{k\in\mathcal{K}} R_k(t), & \text{if (4.11b) is satisfied,} \\ \sum_{k\in\mathcal{K}} \min(R_k(t) - R^{min}, 0), & \text{otherwise.} \end{cases} \tag{4.12}$$

Note that, if the remaining power at the BS is zero, i.e., $P(t) = 0$, the total sum rate of the system is regarded as zero. Hence, the instantaneous reward is equal to the QoS deficiency, i.e., $\sum_{k\in\mathcal{K}} \min(R_k(t) - R^{min}, 0)$, which indicates that the power consumption constraint in (4.11d) is enforced.

4.3.3 TRPO Learning Algorithm

Since the conventional on-policy DRL algorithms, such as the DDPG algorithm, utilizes gradient descent to improve the policy network π, their performance is extremely sensitive to the choice of step size. A large step size can lead to divergence and a small step size may cause the algorithm to stuck in local optima. Moreover, as the neural network gets deeper and wider, a small change in the policy parameters may result in a large difference in the learning outcome, which causes severe instability in the training process. To address this issue, the TRPO algorithm proposed in [92] specified a trust region, measured by the Kullback–Leibler (KL) divergence, around the current policy, which indicates the maximum distributional distance between successive policies.

Let $\boldsymbol{\omega}$ denotes the parameters of the policy network π, i.e., $\pi_{\boldsymbol{\omega}} \equiv \pi$, the theoretical TRPO update equation at training episode k is given by

$$\boldsymbol{\omega}_{k+1} \leftarrow \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \underset{\mathbf{s},\mathbf{a}}{\mathbb{E}} \left[\frac{\pi_{\boldsymbol{\omega}}(\mathbf{a}|\mathbf{s})}{\pi_{\boldsymbol{\omega}_k}(\mathbf{a}|\mathbf{s})} \cdot A_k^{\pi}(\mathbf{s},\mathbf{a}) \right], \qquad \text{s.t. } \overline{\mathrm{KL}}(\pi_{\boldsymbol{\omega}}(\mathbf{s})||\pi_{\boldsymbol{\omega}_k}(\mathbf{s})) \leq \delta, \tag{4.13}$$

where $A_k^{\pi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the advantage function, $\overline{\mathrm{KL}}$ denotes the average KL-divergence and δ denotes the radius of the trust region, which can be interpreted as the learning rate. In particular, the advantage function is estimated through the generalized advantage estimator (GAE), which is given by

$$\hat{A}_t^{\mathrm{GAE}} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \left(r_t + \gamma V(s_{t+1}) - V(s_t) \right), \tag{4.14}$$

where γ is the discount factor of the MDP, $\lambda \in [0, 1]$ is the exponential weight discount, and $V^{\pi} : \mathcal{S} \to \mathbb{R}$ denotes the value function. In the proposed algorithm, the value

function is approximated through a linear network with a time-varying feature vector as described in [93].

However, since it is difficult to directly compute (4.13), the TRPO algorithm utilizes a few approximations. First, Taylor expansion is employed to approximated (4.13) by

$$\boldsymbol{\omega}_{k+1} \leftarrow \underset{\boldsymbol{\omega}}{\operatorname{argmax}}\ \mathbf{g}^T(\boldsymbol{\omega}-\boldsymbol{\omega}_k), \qquad \text{s.t. } \frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\omega}_k)^T\mathbf{H}_{\mathrm{KL}}(\boldsymbol{\omega}-\boldsymbol{\omega}_k)) \leq \delta, \tag{4.15}$$

where $\boldsymbol{g}$ denotes the policy gradient, which is computed based on the estimated advantage $\hat{A}^{\pi}(\mathbf{s},\mathbf{a})$ and $\mathbf{H}_{\mathrm{KL}} = \frac{\partial^2}{\partial^2 \boldsymbol{\omega}}\overline{\mathrm{KL}}(\boldsymbol{\omega}||\boldsymbol{\omega}_k)|_{\boldsymbol{\omega}_k}$. Then, the problem in (4.15) can be analytically solved by the methods of Lagrangian duality, yielding the following solution:

$$\boldsymbol{\omega}_{k+1} \leftarrow \boldsymbol{\omega}_k + \eta^j \sqrt{\frac{2\delta}{\mathbf{g}^T\mathbf{H}_{\mathrm{KL}}^{-1}\mathbf{g}}}\mathbf{H}_{\mathrm{KL}}^{-1}\mathbf{g}, \tag{4.16}$$

where $\eta \in (0,1)$ is the back-tracking coefficient and $j \in \mathbb{Z}^+$ is the smallest possible value that satisfies the KL-divergence constraint. Moreover, to reduce the complexity in evaluating the matrix inverse, i.e., $\mathbf{H}_{\mathrm{KL}}^{-1}$, the conjugate gradient algorithm is employed to find $\hat{\mathbf{z}}$ that solves $\mathbf{H}_{\mathrm{KL}}\mathbf{z} = \mathbf{g}$ for $\mathbf{z} = \mathbf{H}_{\mathrm{KL}}^{-1}\mathbf{g}$. Hence, by substituting $\hat{\mathbf{z}}$ into (4.16), the closed-form TRPO update equation is finally derived as

$$\boldsymbol{\omega}_{k+1} \leftarrow \boldsymbol{\omega}_k + \eta^j \sqrt{\frac{2\delta}{\hat{\mathbf{z}}^T\mathbf{H}_{\mathrm{KL}}\hat{\mathbf{z}}}}\hat{\mathbf{z}}. \tag{4.17}$$

Fig. 4.3 demonstrates the proposed resource allocation algorithm for adaptive NGMA systems and Algorithm 4 illustrates the pseudocode of the TRPO training algorithm. The policy network consists of the input layer, a batch-normalization layer, five fully-connected hidden layers with ReLU activation functions, and the output layer with the linear activation function.

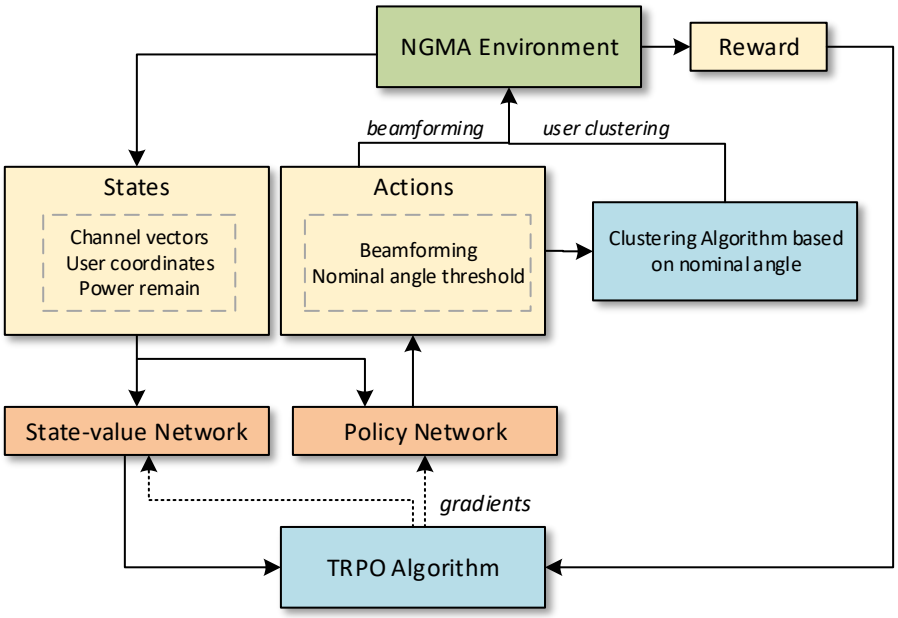


Figure 4.3: Flow diagram of the proposed TRPO-based resource allocation scheme for apative NGMA systems.

Algorithm 4 TRPO Algorithm for Resource Allocation in Adaptive NGMA

Input: initial policy parameters $\boldsymbol{\omega}$, length of time frame T, KL-divergence constraint δ, maximum number of line search steps N_{line}

for $k = 1, \ldots,$ max episode **do**
  Initialize the NGMA environment
  for $t = 1, \ldots, T$ **do**
    Observe states $\mathbf{s}_t$ and sample actions $\mathbf{a}_t \sim \pi_{\boldsymbol{\omega}_k}(\mathbf{s}_t)$
    Execute $\mathbf{a}_t$ and compute reward r_t using (4.12)
    Transition to next state $\mathbf{s}_{t+1}$ and store $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_t)$ in the trajectory $\mathcal{D}$
  end for
  Based on the collected trajectory $\mathcal{D}$, compute the estimated state-value $V_{\boldsymbol{\phi}_k}(\mathbf{s})$ and advantage $\hat{A}_k(\mathbf{s}, \mathbf{a})$
  for $j = 1, \ldots, N_{\text{line}}$ **do**
    Based on $V_{\boldsymbol{\phi}_k}(\mathbf{s})$ and $\hat{A}_k(\mathbf{s}, \mathbf{a})$, compute proposed update $\boldsymbol{\omega}_{k+1}$ using (4.17)
    if $\overline{\text{KL}}(\boldsymbol{\omega} || \boldsymbol{\omega}_k) \leq \delta$ **then**
      break
    end if
  end for
end for

Table 4-B: Network and algorithm configurations.

System parameters	Values	Algorithm parameters	Values
Bandwidth	1 MHz	Batch size	64
Number of TSs	$T = 10$	KL-divergence constraint	$\delta = 0.02$
Maximum power	$P_{max} = 10$ dBm	Neurons per layer	400
Noise spectral density	-120 dBm/Hz	NN layers	5
Minimum QoS	0.1 Mbps/Hz	Maximum line search steps	$N_{\text{line}} = 20$

4.3.4 Complexity Analysis

In the prediction stage, the algorithm performs one forward propagation of the policy network $\pi_{\boldsymbol{\omega}}$ to obtain the actions based on the environment. Let $m_{\pi,i}$ denotes the number of neurons in layer i of the policy network and L_π denotes the total number of layers. The complexity of one forward of the policy network is derived as $\mathcal{O}(\sum_{i=1}^{L_\pi-1} m_{\pi,i}\ m_{\pi,i+1})$, where $m_{\pi,1} = 2KN + 2K + 1$ and $m_{\pi,L_\pi} = 2KN + 1$ are the input and the output dimensions of the policy network, respectively.

In the training stage, the algorithm performs both forward and backward propagation of the policy π and one forward propagation of the state-value network $V_{\boldsymbol{\phi}}$. Since the complexity of forward and backward propagation are the same, the total complexity induced by the policy network is $\mathcal{O}(2\sum_{i=1}^{L_\pi-1} m_{\pi,i}\ m_{\pi,i+1})$. The state-value network $V_{\boldsymbol{\phi}}$ is a linear network, which maps the states to a single number, hence the complexity of $V_{\boldsymbol{\phi}}$ is $\mathcal{O}(2KN + 2K + 1)$. Finally, the total algorithm complexity during the training stage is derived as $\mathcal{O}(4KN + 4K + 2 + 2\sum_{i=1}^{L_\pi-1} m_{\pi,i}\ m_{\pi,i+1})$, where $m_{\pi,1} = 2KN + 2K + 1$ and $m_{\pi,L_\pi} = 2KN + 1$.

4.4 Numerical Results

4.4.1 Simulation Settings

A 40 m^2 outdoor space is considered, where the BS is located at the centre of the square area, defined as the origin of the plane, i.e., $(0, 0)$. Each trajectory begins when the users enter the area from one of the four starting points: $\{(-20, 0), (20, 0), (0, -20), (0, 20)\}$, which can be interpreted as the main roads. After entering the area, the mobility of each

user is modelled as a random walk with a step size of 5 m between consecutive TSs to one of the adjacent coordinates in the coverage area and the minimum distance between the user and the BS is 5 m. Each trajectory consists of $T = 10$ TSs. Unless otherwise stated, the system parameters and the algorithm parameters are provided in Table 4-B.

4.4.2 Baseline Methods

Three baseline methods are considered, namely SDMA, PD-NOMA, and NGMA with semi-orthogonal clustering (NGMA*), which are described as follows:

- *SDMA*: In SDMA, each cluster consists of only one user, i.e., $M = K$, and each user decodes their intended signals by treating all other users' signals as interference.
- *PD-NOMA*: In PD-NOMA, all users are allocated to the same cluster, i.e. $M = 1$ and $\mathcal{G}_1 = \mathcal{K}$, with the same beamforming vector. The SIC order is determined in the ascending order of channel gains, where the signal of the user with the weakest channel gain is decoded first and the signal of the user with the strongest channel gain is decoded last.
- *NGMA with semi-orthogonal clustering (NGMA*)*: In NGMA*, the clustering algorithm proposed in [94] is employed, where the channel orthogonality threshold is learned as a continuous variable to select the cluster heads. The rest of the users are each paired with a cluster head to achieve the lowest channel orthogonality within clusters.

4.4.3 Algorithm Convergence

Fig. 4.4 illustrates the convergence of the TRPO algorithm as the network architecture, the batch size, and the learning rate varies. For instance, the legend ($\text{nn} = 50 \times 4, \text{bs} = 64, \text{lr} = 0.1$) describes a neural network that consists of 50 neurons per hidden layer with 4 hidden layers, where the batch size and learning rate are 64 and 0.1, respectively. It can be observed that the deeper network with 50 neurons per layer and 4 hidden layers outperforms the wider network with 100 neurons per layer and 2 hidden layers throughout

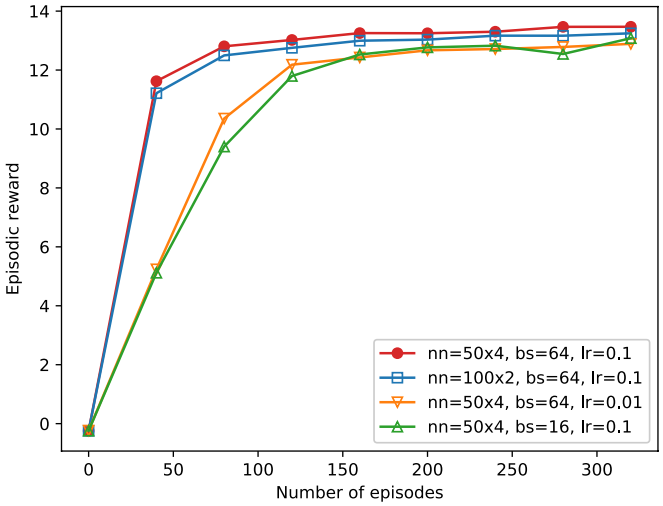


Figure 4.4: Episodic reward versus the number of episodes under different network architecture and batch sizes.

the whole training stage. In other words, under similar network complexity, a deeper network demonstrates better learning performance compared to a wider network since it represents a more complex and non-linear function. In terms of the batch size, a smaller batch size, such as 16, not only leads to slower training due to the small amount of data the network is trained with but also results in a suboptimal learning performance to that of a larger batch size of 64. Finally, learning performance under different learning rates is compared. Owing to the trust region technique, increasing the learning rate from 0.01 to 0.1 can greatly increase the convergence rate without suffering from strong fluctuations in the learning process. In practical scenarios, a stable and fast learning performance can greatly reduce the computational and time cost of model training.

Fig. 4.5 illustrates the convergence of the TRPO algorithm under different transmission schemes, namely, the NGMA scheme, the NGMA* scheme, the SDMA scheme, and the PD-NOMA scheme. It can be noticed that the TRPO algorithm demonstrates stable convergence under all four transmission schemes. The sum rate of the PD-NOMA scheme has the fastest convergence rate since the algorithm only needs to optimize one beamforming vector for all users. However, the PD-NOMA scheme achieves the lowest

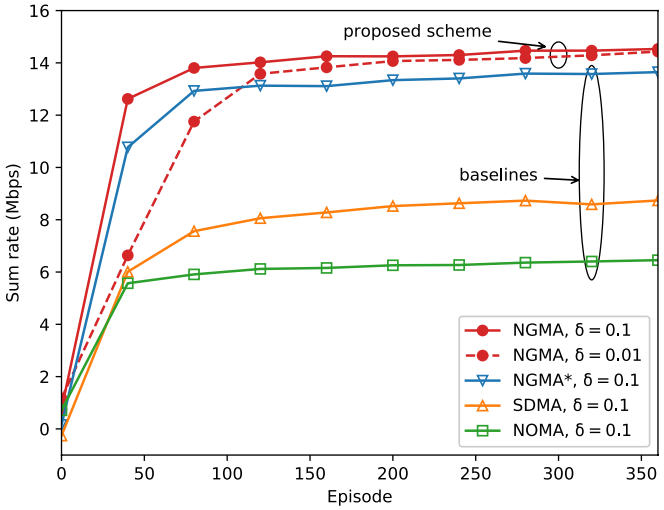


Figure 4.5: Episodic reward versus the number of episodes in SDMA, PD-NOMA, and NGMA systems under $K = 4$ users and $N = 4$ antennas.

sum rate among all schemes, because the user channels are not always correlated and they will suffer from strong inter-cluster interference due to the uncorrelated channels. By jointly designing the beamforming vectors of each user, the SDMA scheme can effectively reduce or even eliminate the inter-user interference for users with uncorrelated channels, hence achieving a higher sum rate compared to the PD-NOMA scheme. However, a large performance gap can be noticed between the proposed NGMA schemes and the SDMA scheme, since it is difficult to design beamforming vectors that eliminate the inter-user interference under a limited number of antennas and strongly correlated channels. By adaptively employing PD-NOMA and SDMA for different users in the same system, the NGMA scheme can exploit the complementary advantages of the two conventional schemes without suffering from their individual drawbacks, hence yielding significant performance gain. Finally, by comparing the results of the proposed NGMA to the NGMA* scheme, where the former scheme performs spatial correlation-based clustering and the later scheme performs channel correlation-based clustering, an observable sum rate gain is achieved by the proposed NGMA scheme, which indicates that the channel correlations can be effectively interpreted by spatial correlations in the considered

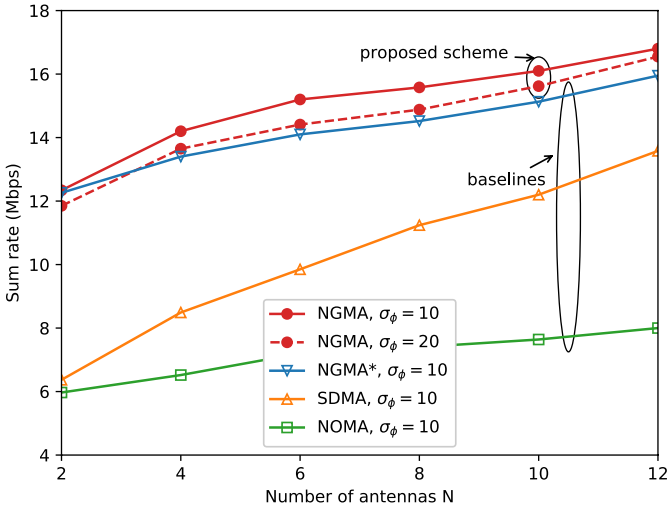


Figure 4.6: Sum rate versus the number of antennas under SDMA, PD-NOMA, NGMA* and NGMA schemes with 4 users.

network and, by additionally considering the channel gain differences during clustering, the proposed clustering method can further enhance the system sum rate.

4.4.4 Impact of Number of Antenna

Fig. 4.6 illustrates the sum rate performance of the adaptive NGMA scheme compared to the baseline methods in a 4-user network, where the number of antennas ranges from 2 to 12. First, it is observed that the PD-NOMA and the SDMA systems demonstrate similar sum rate performance when there are 2 antennas at the BS. Then, as the number of antennas increases, the SDMA system starts to outperform the PD-NOMA system because it can better utilize the increasing spatial degree of freedom to design dedicated beamforming that reduces the multi-user interference. However, the SDMA system is outperformed by the NGMA system for all values of N, which indicates that PD-NOMA can still achieve higher sum rates than SDMA in certain cases, e.g., when the 4 users have strongly correlated channels. Among two NGMA systems, namely the proposed NGMA with spatial correlation-based clustering and the NGMA* system with channel correlation-based clustering, the proposed NGMA system exhibits a generally higher sum

rate performance than NGMA*, which confirms the performance gain of the proposed clustering scheme compared to the benchmark technique. To further investigate the influence of the channel model on the proposed spatial correlation-based method, results are provided for different values of angular standard deviation, where a larger angular standard deviation leads to a lower channel correlation. When $\sigma_\phi = 20$ and $N = 2$, the proposed NGMA system experiences performance loss compared to NGMA*, which is caused by the weak channel correlations due to the limited number of antennas and a large angular standard deviation. At $\sigma_\phi = 10$, which is a reasonable value of an urban cellular network, the proposed NGMA system achieves an enhanced sum rate performance than NGMA*, where the performance gap further increases as the number of antennas increases, due to the increase in the channel correlations.

4.5 Summary

An adaptive NGMA scheme was proposed in this chapter to exploit the complementary benefits of OMA and PD-NOMA under diverse channel conditions. To solve the sum rate maximization problem, a spatial correlation-based clustering algorithm was proposed and a TRPO-based resource allocation algorithm was designed to jointly optimize beamforming, power allocation, and user clustering, subject to a long-term power constraint. Simulation results verified the sum rate gain of the NGMA scheme against the SDMA and the PD-NOMA baselines. Moreover, the proposed clustering method achieved comparable sum rate performance to the channel correlation-based baseline with increasing performance gain as the spatial correlation in the channel model increases. Having studied the critical problems in the uplink and downlink NOMA systems, the next chapter will investigate the integration of NOMA with emerging 6G technologies, such as RIS, to further enhance spectral efficiency.

Chapter 5

Comparisons between DL and DRL on the Optimization of RIS-assisted NOMA Systems

5.1 Introduction

In this chapter, the design of RIS-aided downlink MISO-PDNOMA systems is investigated. In conventional ZF precoding-based PD-NOMA, the strong/weak users are defined as the users with stronger/weaker channel conditions. Since weak users often suffer from strong multi-user interference, they usually demand a significant amount of transmit power to meet the minimum QoS requirements. Hence, in systems with diverse QoS requirements, users with weak channel conditions and high QoS requirements may suffer from outages. To address this issue, a QoS-based PD-NOMA clustering method is proposed to enhance resource efficiency under the ZF precoding scheme. By defining the weak users as the low QoS users, less transmit power is required to ensure their QoS requirements and more transmit power can be allocated to the strong users to enhance the overall transmission sum rate. To investigate the resource allocation problem in

the proposed system, the sum rate maximization problem is formulated by jointly optimizing the RIS phase shift and the BS power allocation. The problem is formulated from both short-term and long-term aspects, where DL and DRL are employed to solve both problems and their performance is compared through simulation results. The main contributions of this chapter can be summarized as follows:

- An RIS-enhanced PD-NOMA downlink framework is proposed, where a QoS-based PD-NOMA clustering scheme is employed to improve the resource efficiency by maximizing the QoS deviation within the clusters. A sum rate maximization problem is formulated by jointly optimizing the RIS phase shift and the BS power allocation from both short-term and long-term prospects.

- A meta-learning based DL algorithm is utilized to achieve a fast convergence rate at a low algorithm complexity. In particular, the proposed neural network is trained to output the optimized power allocation for any RIS phase shifts. Then the phase shift is optimized in an online manner, where MAML is employed to reduce the number of iterations required for the phase shift optimization.

- A DDPG-based optimization algorithm is invoked to learn the continuous phase shift and power allocation under the time-varying environments. A QoS-aware reward function is formulated to maximize the long-term transmission sum rate while ensuring the instantaneous QoS requirements, subject to the long-term transmit power constraint. In particular, a transmit power-based penalty term is designed to regulate the power consumption, by deducting the total reward when the long-term transmit power constraint is violated.

- Simulation results show that the implementation of RIS can induce approximately 5% to 25% sum rate gain as the number of RIS elements increases from 8 to 64, in both PD-NOMA and OMA systems. Moreover, it also shows that the performance difference between DL and DRL is negligible for the short-term optimization, while for the long-term optimization, DRL achieves a higher sum rate than DL, especially

Table 5-A: List of main notations.

Notation	Description	Notation	Description
K	Number of users	$p_{l,s}, p_{l,w}$	Transmit power
M	Number of antennas	$s_{l,s}, s_{l,w}$	Intended signals
T	Number of time slots (TSs)	θ_n	Phase shift n
N	Number of RIS reflecting elements	α	Path loss exponent
$\mathbf{w}_l$	Beamformer of cluster l		

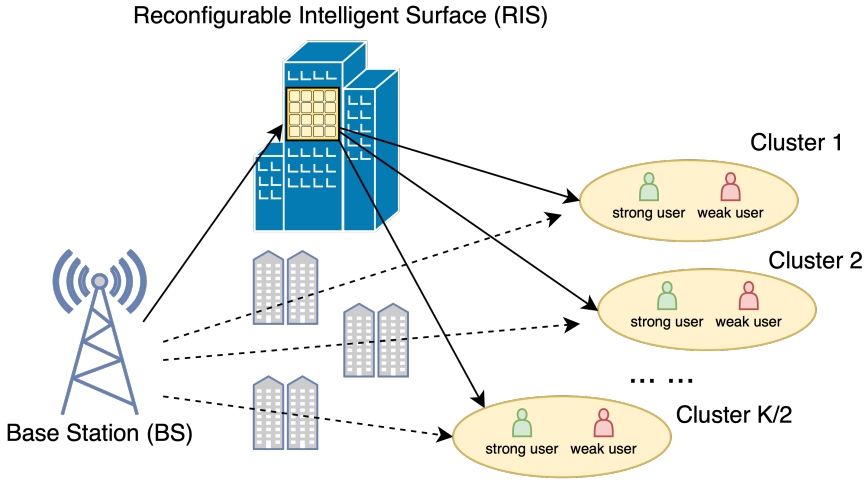


Figure 5.1: Illustration of the downlink RIS-assisted MISO-PDNOMA system.

in power-constrained scenarios.

5.2 Network Model

5.2.1 System Model

As illustrated in Fig. 5.1, we consider a downlink MISO system with one BS and K users. The BS is equipped with M antennas and each user is equipped with a single antenna. We consider a dense urban area, where no line-of-sight (LoS) link exists between the BS and the users. To enhance the wireless services, a RIS is deployed on the facade of a particular building where the LoS link exists between the RIS and the BS, as well as between the RIS and the users. The RIS consists of N reflecting elements, whose phase shift can be adjusted by a controller. The main notations are listed in Table 5-A.

5.2.2 Channel Model

The channels between the BS and the users compose of the non-LoS BS-user links and the reflecting LoS links from the BS to the RIS, denoted as BS-RIS links, and from the RIS to the users, denoted as RIS-user links. The BS-user links are modelled as Rayleigh fading channels and the LoS links are modelled as Rician fading channels. Then, the path loss of user k is modelled as $\mathrm{PL}_k = d_k^{-\alpha}$ where d_k denotes the distance, calculated in meters, between user k and the BS, and α denotes the path loss exponent.

5.2.3 PD-NOMA Signal Model

In this subsection, we formulate the signal model based on PD-NOMA and introduce the proposed QoS-based clustering method.

5.2.3.1 Signal model

To implement PD-NOMA, the BS groups the users into several clusters and utilizes power-domain multiplexing to superimpose the signals of all users in the same cluster [8]. To ensure the SIC decoding accuracy, it is assumed that each cluster is formed by two users, denoted as the strong user and the weak user. In contrast to the conventional user clustering algorithms, the proposed system defines the strong user as the user with a higher QoS requirement and the weak user as the user with a lower QoS requirement. Details of the proposed clustering method will be discussed in a later section. Here, the signal model is first formulated.

According to the principles of PD-NOMA, the signals of the users in the same cluster are multiplexed in the power-domain before transmission. Let $p_{l,s}$ and $p_{l,w}$ denote the transmit power of the strong user and the weak user in the l-th cluster, the BS transmits the following superimposed signal to users in the l-th cluster:

$$x_l = \sqrt{p_{l,s}} s_{l,s} + \sqrt{p_{l,w}} s_{l,w}, \tag{5.1}$$

where $s_{l,s}$ and $s_{l,w}$ denote the signals intended for the strong user and the weak user,

respectively.

The signal received at each user is a composition of the signals from the direct BS-user link and the reflecting BS-RIS-user link. In particular, for a user in the l-th cluster, the RIS-user link and the BS-user link are denoted by $\mathbf{h}_{R,l,i}^H \in \mathbb{C}^{1\times N}$, and $\mathbf{h}_{B,l,i}^H \in \mathbb{C}^{1\times M}$, respectively. The BS-RIS link is denoted by by $\mathbf{H}_{BR} \in \mathbb{C}^{N\times M}$. The phase shift of the RIS is denoted by $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n, \ldots, \theta_N]$, where $\theta_n \in [0, 2\pi)$. Thus, the diagonal phase-shifting matrix is expressed as $\boldsymbol{\Theta} = \text{diag}(\beta_1 e^{j\theta_1}, \ldots, \beta_n e^{j\theta_n}, \ldots, \beta_N e^{j\theta_N})$, where $\beta_n \in [0, 1]$ denotes the amplitude reflection coefficient. For simplicity, unit amplitude coefficients are assumed, i.e., $\beta_n = 1, \forall n$.

Hence, the signal received by a user in the l-th cluster can be expressed as

$$y_{l,i} = \left(\mathbf{h}_{B,l,i}^H + \mathbf{h}_{R,l,i}^H \boldsymbol{\Theta} \mathbf{H}_{BR}\right) \sum_{l=1}^{K/2} \mathbf{w}_l x_l + n_{l,i}, \tag{5.2}$$

where $\mathbf{w}_l$ denotes the beamforming vector of the l-th cluster and $n_{l,i}$ denotes the AWGN, modelled as $n_{l,i} \sim \mathcal{CN}\left(0, \sigma^2\right)$. For instance, the signal received by the strong user in the l-th cluster is expressed as

$$\begin{aligned} y_{l,s} &= \left(\mathbf{h}_{B,l,s}^H + \mathbf{h}_{R,l,s}^H \boldsymbol{\Theta} \mathbf{H}_{BR}\right) \mathbf{w}_l \left(\sqrt{p_{l,s}} s_{l,s} + \sqrt{p_{l,w}} s_{l,w}\right) \\ &+ \left(\mathbf{h}_{B,l,s}^H + \mathbf{h}_{R,l,s}^H \boldsymbol{\Theta} \mathbf{H}_{BR}\right) \sum_{j=1, j\neq l}^{K/2} \mathbf{w}_j x_j + n_{l,s}, \end{aligned} \tag{5.3}$$

where $\left(\mathbf{h}_{B,l,s}^H + \mathbf{h}_{R,l,s}^H \boldsymbol{\Theta} \mathbf{H}_{BR}\right) \sum\limits_{j=1, j\neq l}^{K/2} \mathbf{w}_j x_j$ is the inter-cluster interference induced by the transmission to other clusters and $\left(\mathbf{h}_{B,l,s}^H + \mathbf{h}_{R,l,s}^H \boldsymbol{\Theta} \mathbf{H}_{BR}\right) \mathbf{w}_l \sqrt{p_{l,w}} s_{l,w}$ is the intra-cluster interference imposed by the transmission to the weak user of the same cluster.

The low-complexity ZF precoding technique is employed to eliminate the inter-cluster interference. For simplicity, the combined channel of a user in the l-th cluster is denoted by $\mathbf{h}_{l,i}^H = \mathbf{h}_{B,l,i}^H + \mathbf{h}_{R,l,i}^H \boldsymbol{\Theta} \mathbf{H}_{BR}$, where $i = \{s, w\}$. Moreover, the combined channel matrix of all strong users is denoted as $\mathbf{H}_s^H = [\mathbf{h}_{1,s}^H; \mathbf{h}_{2,s}^H; \ldots; \mathbf{h}_{K/2,s}^H] \in \mathbb{C}^{K/2\times M}$. Hence, the

normalized ZF precoding matrix, denoted by $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_l, \ldots, \mathbf{w}_{K/2}] \in \mathbb{C}^{M \times K/2}$, is formulated as

$$\mathbf{W} = \mathbf{H}_s \left(\mathbf{H}_s^H \mathbf{H}_s\right)^{-1} \mathbf{\Lambda}, \tag{5.4}$$

where $\mathbf{\Lambda} = \text{diag}(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_{K/2}})$ is a diagonal matrix, introduced for column power normalization such that $|\mathbf{w}_l|^2 = 1, \forall\ l = 1, \ldots, K/2$. The corresponding ZF precoding constraints are expressed as follows:

$$\begin{cases} \mathbf{h}_{j,s}^H \mathbf{w}_l = 0, & \forall j \neq l, j = 1, \ldots, K/2, \\ \mathbf{h}_{j,s}^H \mathbf{w}_l = \frac{1}{\lambda_l}, & j = l. \end{cases} \tag{5.5}$$

To decode the intended signal from the multiplexed signal, each strong user employs SIC by first decoding the signal of the weak user. The decoded signal of the weak user is then subtracted from the received signal, so that the signals of the strong users can be decoded without any intra-cluster interference. The weak users, however, directly decode the signals without SIC. Therefore, the received SINR of the strong user in the l-th cluster is given by

$$\gamma_{l,s} = \frac{\left|\mathbf{h}_{l,s}^H \mathbf{w}_l \sqrt{p_{l,s}} s_{l,s}\right|^2}{\sigma^2} = \frac{p_{l,s}}{\lambda_l \sigma^2}. \tag{5.6}$$

Since the weak user decode the intended signal under both inter-cluster interference and intra-cluster interference, the received SINR of the weak user in the l-th cluster is derived as

$$\gamma_{l,w} = \frac{|\mathbf{h}_{l,w} \mathbf{w}_l|^2 p_{l,w}}{|\mathbf{h}_{l,w} \mathbf{w}_l|^2 p_{l,s} + \left|\mathbf{h}_{l,w} \sum\limits_{j=1, j\neq l}^{K/2} \mathbf{w}_j x_j\right|^2 + \sigma^2}. \tag{5.7}$$

5.2.3.2 QoS-based clustering scheme

As shown in (5.7), when both the ZF precoding and the SIC decoding techniques are employed in PD-NOMA, the weak users suffer from both the inter-cluster and the intra-cluster interference, thus resulting in low SINR and low achievable rate compared to the strong users, who are served in an interference-free manner. Conventional clustering methods allocate users by exploiting the difference between their channel conditions. However, when users have different QoS requirements, a user with a weak channel condition may acquire a high QoS, which is challenging to achieve due to the multiuser interference. Moreover, due to the low SINR, a great amount of transmit power has to be allocated to the weak user to fulfil the QoS. Hence, it is more sensible to assign users with lower QoS requirements as the weak users to improve resource efficiency and enhance the system sum rate. Therefore, a QoS-based clustering scheme is proposed, which assigns the users with higher or lower QoS requirements as the strong or weak users, respectively.

The objective of the QoS-based clustering method is to maximize the minimum QoS deviation among all clusters. The clustering problem can be formulated as

$$\max \min_{l=1,\ldots,K/2} (R_{QoS}^{l,s} - R_{QoS}^{l,w}), \tag{5.8}$$

where $R_{QoS}^{l,s}$ and $R_{QoS}^{l,w}$ denote the QoS requirements of the strong and the weak users in the l-th cluster, respectively. To achieve the maximal QoS deviation, a simple but optimal clustering method is proposed as follows:

Proposition 1. *Assuming that K is an even number and all users are ordered in terms of their QoS requirements, i.e., the k-th user has the k-th highest QoS requirement, the optimal solution to* (5.8) *is achieved by assigning the k-th user and the $(k+K/2)$-th user into the same cluster, for all $k \leq K/2$.*

Proof. See Appendix A.1. □

5.3 Short-term Optimization Problem

In this section, the short-term optimization problem is formulated and a meta learning-enabled DL algorithm is proposed that jointly optimizes both power allocation and RIS phase shifts, subject to the QoS requirements.

5.3.1 Short-term Problem Formulation

The short-term optimization goal is to maximize the total transmission sum rate for a single TS, subject to the maximum instantaneous transmit power constraint, which is denoted by P_{max}. The optimization variables are the phase shifts $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n, \ldots, \theta_N]$ of the RIS, as well as the power allocation vector $\mathbf{P} = [\mathbf{P}_s, \mathbf{P}_w]$ of the BS, where $\mathbf{P}_s = [p_{1,s}, \ldots, p_{K/2,s}]$ and $\mathbf{P}_w = [p_{1,w}, \ldots, p_{K/2,w}]$. The short-term optimization problem is formulated as follows:

$$\max_{\boldsymbol{\theta},\mathbf{P}} \; R = \sum\nolimits_{l=1}^{K/2} (R_{l,s} + R_{l,w}) \tag{5.9a}$$

$$\text{s.t.} \quad R_{l,i} \geq R_{\text{QoS}}^{l,i}, \forall l, \forall i \in \{s, w\} \tag{5.9b}$$

$$\left| e^{j\theta_n} \right| = 1, \forall n \tag{5.9c}$$

$$\sum\nolimits_{l=1}^{K/2} (p_{l,s} + p_{l,w}) \leq P_{max} \tag{5.9d}$$

where $R_{l,i} = B_l \log_2(1+\gamma_{l,i})$ denotes the sum rate achieved by the user in the l-th cluster and $R_{\text{QoS}}^{l,i}$ denotes the minimal QoS requirement of the user. Here, (5.9b) represents the minimum transmit rate constraint. (5.9c) denotes the unit modulus constraint of each RIS element and (5.9d) qualifies the instantaneous transmit power constraint of the BS. Due to the non-convex constraint (5.9c) and the non-concave objective function, the optimization problem cannot be directly solved by conventional approaches. To address this issue, DL is employed to tackle this non-convex joint optimization problem by employing DL.

5.3.2 DL-based Resource Allocation Scheme

The main idea of DL is to extensively train a neural network such that, given any inputs that follow the same distribution as the training data, the outputs of the network achieve minimal loss. A conventional design is to construct a neural network that outputs all optimization variables, i.e., the RIS phase shift and the power allocation. However, this design will result in a large input space that consists of all channel information and QoS information. In particular, all channel matrices contribute an input dimension of $(2K \times N + 2K \times M + 2N \times M)$, leading to exceedingly expensive computational costs. Moreover, the phase shift and the power allocation have vastly different value ranges and distributions, which greatly increase the training difficulty.

Remark 1. *Optimizing the phase shift requires the knowledge of all channels among the BS, the RIS, and the users. However, the optimization of the power allocation only requires the information of the combined channel.*

Since the combined channel of all users provides sufficient channel information for optimizing $\mathbf{P}$, the neural network can be designed to input the combined channel, denoted by $\mathbf{H} = [\mathbf{h}_{1,s}, \mathbf{h}_{1,w}, \ldots, \mathbf{h}_{K/2,s}, \mathbf{h}_{K/2,w}]$ and output the optimized power allocation $\mathbf{P}$. The real and the imaginary parts of the combined channel $\mathbf{H}$ contribute $(2K \times M)$ to the input dimension, which is significantly smaller than the input dimension of the intuitive design. Hence, the neural network G_η is formulated as follows:

$$\mathbf{P} = G_\eta(\mathbf{H}(\boldsymbol{\theta}), \mathbf{R}_{QoS}, \mathbf{L}_{\text{path}}), \tag{5.10}$$

where $\mathbf{H}(\boldsymbol{\theta})$ denotes the combined channel calculated using the phase shift $\boldsymbol{\theta}$, $\mathbf{R}_{QoS} \in \mathbb{R}^K$ denotes the QoS requirement vector, and $\mathbf{L}_{\text{path}} \in \mathbb{R}^K$ is the path loss vector. The phase shift $\boldsymbol{\theta}$ is optimized separately using a gradient descent algorithm. Two optimization algorithms are connected in an alternating structure, by using the output of the other algorithm as the input.

In contrast to the conventional alternating optimization approach, the neural network

G_η is trained to output the optimized power allocation for any given RIS phase shift. Hence, for a given trained G_η, the optimized pair of $\boldsymbol{\theta}$ and $\mathbf{P}$ can be obtained by solely performing the optimization on $\boldsymbol{\theta}$. To improve the convergence rate of the phase shift optimization algorithm, MAML, a meta-learning technique, is employed in the network training phase.

5.3.3 MAML-based Training Algorithm

Meta-learning, also known as learning-to-learn, refers to a ML technique that aims to improve the convergence rate of learning algorithms, by feeding them with experience over multiple training episodes [95]. MAML is able to optimize the model parameters such that a few gradient steps will produce a maximally effective performance on a new task [70].

As demonstrated in [79], MAML can be employed to reduce the number of gradient descent steps required to optimize the inputs of the neural network. In the proposed model, the network inputs consist of the combined channel matrix, which is directly affected by the phase shift $\boldsymbol{\theta}$. Thus, optimizing the phase shift is equivalent to optimizing the input of the network. Moreover, the gradient descent steps on $\boldsymbol{\theta}$ are performed by back-propagating through the weights of G_η. Hence, MAML can be employed to train G_η such that a small number of gradient steps is sufficient to optimize the phase shift.

Then, the loss function needs to be formulated. Since the optimization targets of $\boldsymbol{\theta}$ and $\mathbf{P}$ are the same, they share the same loss function, denoted by $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is the weights of the neural network G_η. The loss function consists of two parts, the total sum rate and a penalty term for enforcing the QoS requirements, which is formulated as follows:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = -w_1 \sum_{l=1}^{K/2} \sum_{i=s,w} R_{l,i}(\boldsymbol{\theta}, \boldsymbol{\eta}) + w_2 \sum_{l=1}^{K/2} \sum_{i=s,w} \max\Big(R_{l,i}(\boldsymbol{\theta}, \boldsymbol{\eta}) - R_{QoS}^{l,i}, 0 \Big), \quad (5.11)$$

where $R_{l,i}(\boldsymbol{\theta}, \boldsymbol{\eta})$ denotes the sum rate calculated using $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, and $\max\Big(R_{l,i}(\boldsymbol{\theta}, \boldsymbol{\eta}) -$

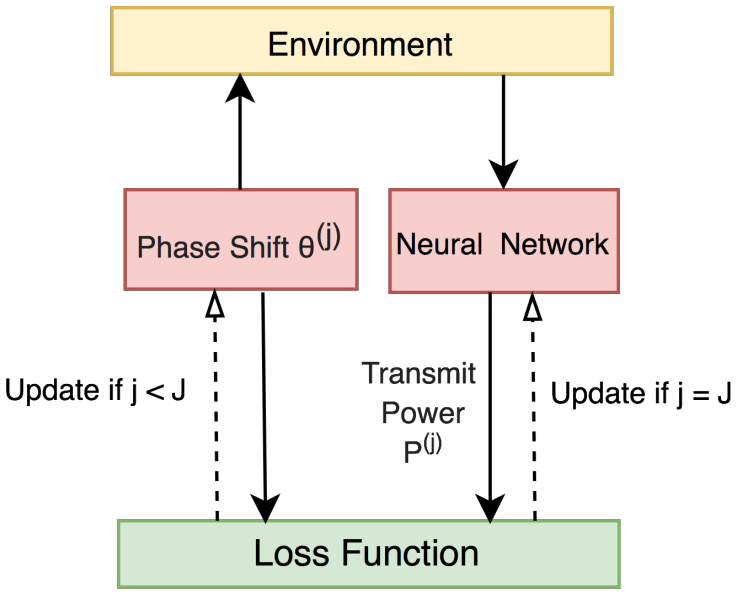


Figure 5.2: Illustration of the MAML-based training framework.

$R_{QoS}^{l,i}, 0\Big)$ indicates the QoS deficiency of user i in the l-th cluster. The weights w_1 and w_2 are tuned during training.

Suppose $\boldsymbol{\theta}$ needs to be optimized in J gradient steps, the gradient descent update in the j-th gradient step of the p-th training episode can be formulated as

$$\boldsymbol{\theta}^{(j)} \leftarrow \boldsymbol{\theta}^{(j-1)} - \gamma_\theta \frac{\partial}{\partial \boldsymbol{\theta}^{(j-1)}} \mathcal{L}(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\eta}^{(p-1)}), \tag{5.12}$$

where γ_θ denotes the phase shift learning rate, $\boldsymbol{\eta}^{(p-1)}$ denotes the neural network weights obtained in the previous training episode. In order to satisfy the phase shift constraint in (5.9c), the entries of $\boldsymbol{\theta}$ are clipped to $[0, 2\pi)$ after each update. Based on (5.12), the loss function after completing the J gradient steps is therefore $\mathcal{L}(\boldsymbol{\theta}^{(J)}, \boldsymbol{\eta}^{(p-1)})$, where $\boldsymbol{\theta}^{(J)}$ is the optimized phase shift. Then, the neural network G_η is optimized to minimize $\mathcal{L}(\boldsymbol{\theta}^{(J)}, \boldsymbol{\eta}^{(p-1)})$ through the following update formula

$$\boldsymbol{\eta}^{(p)} \leftarrow \boldsymbol{\eta}^{(p-1)} - \gamma_\eta \frac{\partial}{\partial \boldsymbol{\eta}^{(p-1)}} \mathcal{L}(\boldsymbol{\theta}^{(J)}, \boldsymbol{\eta}^{(p-1)}), \tag{5.13}$$

where γ_η denotes the network learning rate.

To employ MAML, (5.13) is calculated by implicitly performing the second order differentiation with respect to the loss function $\mathcal{L}(\boldsymbol{\theta}^{(J)}, \boldsymbol{\eta}^{(p-1)})$ and back-propagating through the J phase shift optimization steps in (5.12). Moreover, MAML is performed on the phase shift learning rate γ_θ against the loss function $\mathcal{L}(\boldsymbol{\theta}^{(J)}, \boldsymbol{\eta}^{(p-1)})$ to reduce the need for further hyper-parameter tuning. The update equation of γ_θ can be derived in the same way as in (5.13).

As shown in Fig. 5.2, each training epoch can be divided into two stages, corresponding to the inner and the outer MAML steps:

1. *Phase shift optimization (inner step)*: The initial phase shift is sampled according to a random uniform distribution, i.e. $\boldsymbol{\theta}^{(0)} \sim \mathcal{U}(0, 2\pi)$. In the j-th gradient loop, the corresponding power allocation $\mathbf{P}^{(j)}$ based on $\boldsymbol{\theta}^{(j)}$ is obtained using (5.10). Then, $\boldsymbol{\theta}^{(j)}$ is optimized with respect to the loss function $\mathcal{L}(\boldsymbol{\theta}^{(j)}, \boldsymbol{\eta})$, as in (5.12). After repeating (5.12) for J iterations, the final optimized phase shift is denoted as $\boldsymbol{\theta}^{(J)}$.

2. *Power allocation optimization (outer step)*: After completing J gradient descent loops, the current optimal power allocation $\mathbf{P}^{(J)}$ can be computed using $\boldsymbol{\theta}^{(J)}$ and (5.10). Then, the network weights $\boldsymbol{\eta}$ are updated according to (5.13), by backpropagating through all J gradient descent iterations.

The pseudocode of the training algorithm is presented in Algorithm 5, where lines 2-8 correspond to the phase shift optimization and lines 9-11 correspond to the power allocation optimization. To apply the trained network to new datasets, the phase shift optimization procedure only needs to be executed for J times, after which the optimized phase shift $\boldsymbol{\theta}^{(J)}$ and the corresponding power allocation $\mathbf{P}^{(J)}$, are the solutions to the joint optimization problem in (5.9).

Algorithm 5 Meta-learning Based Training Algorithm

Input: Channel matrix $\mathbf{H}$, QoS vector $\mathbf{R}_{QoS}$, user locations, neural network G_η, number of phase shift update steps J

Output: Trained neural network $G_{\hat{\eta}}$

Initialize $\boldsymbol{\eta}$

repeat
  for each episode **do**
    Initialize phase shift $\theta_1, ..., \theta_N \overset{\text{iid}}{\sim} \mathcal{U}(0, 2\pi)$
    Calculate path loss vector $\mathbf{L}_{\text{path}}$
    for $j = 0$ to $J - 1$ **do**
      Obtain power allocation $\mathbf{P}^{(j)} = G_\eta(\mathbf{H}(\boldsymbol{\theta}^{(j)}), \mathbf{R}_{QoS}, \mathbf{L}_{\text{path}})$
      Calculate loss function $\mathcal{L}(\boldsymbol{\theta}^{(j)}; \boldsymbol{\eta})$
      Update phase shift using (5.12)
    end for
    Given the optimized phase shift $\boldsymbol{\theta}^{(J)}$, calculate the optimized power allocation $\mathbf{P}^{(J)} = G_\eta(\mathbf{H}(\boldsymbol{\theta}^{(J)}), \mathbf{R}_{QoS}, \mathbf{L}_{\text{path}})$
    Calculate loss function $\mathcal{L}(\boldsymbol{\theta}^{(J)}, \boldsymbol{\eta})$ using the optimized phase shift
    Update network weights using (5.13)
  end for
until reaches the maximum training steps
Return $G_{\hat{\eta}}$

5.3.4 Convergence Analysis

According to [96], a MAML algorithm finds an ϵ-first-order stationary point in $\mathcal{O}(1/\epsilon)$ iterations for any $\epsilon > 0$, given a sufficient number of learning samples. The convergence theorem was then extended to the multi-step MAML algorithms in [97]. Both theorems were built on several assumptions on the loss function, one of which requires the loss function to be smooth. In the proposed model, the loss function $\mathcal{L}(\boldsymbol{\theta}^{(J)}, \boldsymbol{\eta})$ contains a non-smooth penalty term which represents the QoS requirements of individual users. To be specific, the penalty term $\max\left(R_{l,i}(\boldsymbol{\theta}, \boldsymbol{\eta}) - R_{QoS}^{l,i}, 0\right)$ is non-differentiable at point $(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$, where $R_{l,i}(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) = R_{QoS}^{l,i}$, which causes the loss function to be non-smooth at this particular point. More importantly, the adopted MAML algorithm carries out inner gradient descent steps on the network input space, rather than on the model weights, which is significantly different to conventional MAML techniques. This MAML variant was proposed recently in [79] and its theoretical convergence has not been explored yet.

Here, two main challenges of the convergence analysis are discussed. First, the min-

imum QoS requirements must be enforced by adding a non-smooth penalty term to the loss function. Although the neural network can be successfully trained using sub-gradients, the conventional smoothness-based convergence theorems are violated. Secondly, the inner and the outer loops in the nested optimization path have different optimization variables. This type of MAML algorithm has been implemented in several research contributions [79, 98], however, no analytical results were provided to support the convergence. Due to the difficulty in proving the algorithm convergence, extensive experiments are provided in Sec. 5.5.1 to verify the convergence of the proposed algorithm under various training configurations.

5.3.5 Complexity Analysis

In this section, the asymptotic computational complexity of the MAML-based DL algorithm is discussed. Since the online optimization complexity is at most the complexity of one offline training episode, the analysis will be focused on deriving the offline complexity. In particular, asymptotic complexity is derived in terms of three important model variables: the number of users (K), the number of reflecting elements (N), and the number of BS antennas (M).

The complexity of the proposed DL algorithm is dominated by three parts: 1) the forward and backward propagation of the neural network; 2) the calculation of the loss function; 3) the gradient descent of the phase shift. The computational complexity of the forward and backward propagation of the same neural network is identical, hence, without loss of generality, the complexity of the forward propagation is derived. In the proposed DL algorithm, the neural network inputs the combined complex-valued channel matrix, the QoS requirement vector and the path loss vector, then outputs the power allocations of all users. Thus, the input dimension is $(2KM + 2K)$ and the output dimension is K. It is assumed that the number of neurons in each hidden layer is independent of the variables of interest. Moreover, the first and the last hidden layers are assumed to be the conventional fully-connected layers, whose operations are 2-

dimensional matrix multiplications. Therefore, by denoting α_0 as the number of neurons in the first hidden layer, the complexity associated with the input layer can be derived as $\mathcal{O}(\alpha_0(2KM + 2K)) = \mathcal{O}(KM)$. Similarly, the complexity associated with the output layer can be derived as $\mathcal{O}(K)$. Hence, the complexity of one forward propagation or one backward propagation is $\mathcal{O}(KM) + \mathcal{O}(K) = \mathcal{O}(KM)$. In each training episode of the proposed DL algorithm, the neural network undergoes $J+1$ forward and $J+1$ backward propagation operations, hence, the complexity associated with the neural network in a single training episode is $\mathcal{O}(JKM)$.

Then, additional complexity is induced when calculating the loss function, i.e. (5.11). Trivially, the complexity of (5.11) is dominated by the calculation of the individual user's combined channel vector $\mathbf{h}_{l,i}^H$, of which the complexity is $\mathcal{O}(NM)$. Therefore, since the combined channel is computed for each user, the total complexity induced by calculating the loss function is $\mathcal{O}(NKM)$. Finally, the complexity of the gradient descent algorithm for optimizing the phase shift is equal to $\mathcal{O}(JN)$.

To sum up, the offline training complexity of the DL algorithm is derived as $\mathcal{O}(N_{\text{ep}}(JKM + NKM + JN))$, where N_{ep} denotes the total number of training episodes before reaching convergence. Hence, the online application complexity is $\mathcal{O}((JKM + NKM + JN))$.

5.4 Long-term Optimization Problem

In this section, the long-term optimization problem is formulated and the proposed DDPG-based DRL algorithm is introduced.

5.4.1 Long-term Problem Formulation

The short-term problem, formulated in the previous section, is subject to an instantaneous transmit power constraint, which assumes that all transmit power at the BS should be consumed for the considered one-time-slot transmission. However, in practice, the transmission generally involves multiple TSs and requires a long-term strategy. To

investigate this problem, an average transmit power constraint over multiple transmission TSs is considered. To satisfy the instantaneous QoS requirements, the BS has to coordinate the transmit power of different TSs. Accordingly, the long-term problem is formulated as follows:

$$\max_{\boldsymbol{\theta},\mathbf{P}} \ \sum\nolimits_{t=1}^{T} R(t) = \sum\nolimits_{t=1}^{T} \sum\nolimits_{l=1}^{K/2} (R_{l,s}(t) + R_{l,w}(t)) \tag{5.14a}$$

$$\text{s.t.} \quad R_{l,i}(t) \geq R_{\text{QoS}}^{l,i}(t), \forall l, \forall i \in \{s, w\}, \forall t \tag{5.14b}$$

$$\left| e^{j\theta_n(t)} \right| = 1, \forall n, \forall t \tag{5.14c}$$

$$\frac{1}{T} \sum\nolimits_{t=1}^{T} \sum\nolimits_{l=1}^{K/2} (p_{l,s}(t) + p_{l,w}(t)) \leq P_{max} \tag{5.14d}$$

where $R_{l,i}(t) = B_l \log_2(1 + \gamma_{l,i}(t))$ denotes the sum rate achieved by the users in the l-th cluster at the t-th TS and $R_{\text{QoS}}^{l,i}(t)$ denotes the minimal QoS requirement of the users in the l-th cluster at the t-th TS. Then, (5.14b) defines the QoS requirement constraints, (5.14c) defines the phase shift constraint of the RIS, and (5.14d) specifies the average transmit power constraint of the BS, which indicates that farsighted network evolution has to be considered instead of only striking the current benefits.

In addition to the non-convex constraint (5.14c), the formulated problem has a long-term power constraint (5.14d), which introduces correlations among the solutions of different TSs and cannot be directly solved by either traditional convex-based techniques or conventional DL algorithms. As an algorithm that is designed to maximize long-term rewards, DRL is employed to tackle the formulated problem. In particular, the DDPG algorithm is employed since it is capable of learning continuous states, i.e., the channel information and the QoS requirements, and actions, i.e., the RIS phase shift and the power allocation.

5.4.2 DRL-based Resource Allocation Scheme

In the proposed reinforcement learning algorithm, the BS acts as the agent, who controls the power allocation of each user, as well as the phase shift adjustments of the RIS. In each TS, the BS first observes the states, denoted by $\mathbf{s}_t$, which include the channel information, the QoS requirements, the path loss, and the remaining transmit power, denoted by $P_{max,t}$. In particular, $P_{max,t}$ is formulated as follows:

$$P_{max,t} = TP_{max} - \sum_{j=1}^{t-1} \sum_{l=1}^{K/2} (p_{l,s}(j) + p_{l,w}(j)) \tag{5.15}$$

$$= P_{max,t-1} - (p_{l,s}(t-1) + p_{l,w}(t-1)), \tag{5.16}$$

where P_{max} is the average transmit power constraint and $P_{max,0} = TP_{max}$. Then, the BS carries out a set of actions, denoted by $\mathbf{a}_t$, to adjust the RIS phase shift and to assign the transmit power for each user. The state space is denoted as $\mathcal{S}$ and the action space is denoted as $\mathcal{A}$. Then, the relationship between the states and the actions can be represented by a policy function $\mu_\phi : \mathcal{S} \rightarrow \mathcal{A}$, parameterized by ϕ. The actions are assessed through a reward function $r(\mathbf{s}_t, \mathbf{a}_t)$. Finally, to determine the optimal policy μ_ϕ^* that maximizes the long-term reward, an objective function is formulated as follows:

$$\mu_\phi^* = \underset{\phi}{\text{argmax}} \sum_t \kappa_t \; r(\mathbf{s}_t, \mu_\phi(\mathbf{s}_t)), \tag{5.17}$$

where $\kappa_t \in [0, 1]$ denotes the discount factor, which is utilized to prevent an infinite sum of rewards. For simplicity, the long-term reward can be represented by an action-value function, denoted by $Q(\mathbf{s}_t, \mathbf{a}_t)$, where $Q(\mathbf{s}_t, \mathbf{a}_t) = \sum_t \kappa_t \; r(\mathbf{s}_t, \mathbf{a}_t)$. Hence, the equation in (5.17) can be simplified into

$$\mu_\phi^* = \underset{\phi}{\text{argmax}} \; Q(\mathbf{s}_t, \mathbf{a}_t), \tag{5.18}$$

where $\mathbf{a}_t = \mu_\phi(\mathbf{s}_t)$.

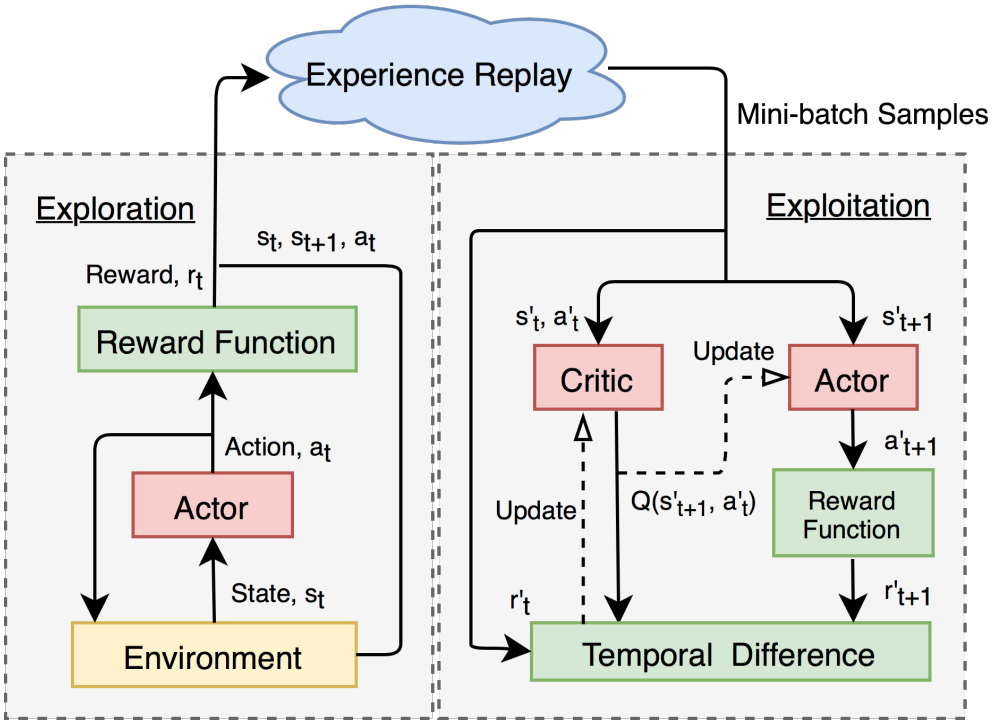


Figure 5.3: Illustration of the DDPG framework.

5.4.3 DDPG-based Training Algorithm

The DDPG algorithm [99] is deployed as the training algorithm since both the state space and the action space is high-dimensional and continuous-valued. In the DDPG algorithm, two neural networks, known as the actor and the critic networks, are trained to approximate the optimal policy function μ^*_ϕ and the action-value function $Q(s, a)$, respectively.

As illustrated in Fig. 5.3, the actor network is updated based on the outputs of the critic network, following (5.18), and the critic network is updated using the temporal difference loss. To prevent the data from being highly correlated, the DDPG algorithm utilizes the experience replay technique. To be specific, in the exploration stage, new training data is generated by interacting with the environment and the data samples are stored in an experience buffer. Then, in the exploitation stage, random samples are obtained from the buffer to train the actor and the critic networks. Algorithm 6 illustrates the pseudocode of the adopted DDPG training algorithm. The state space,

Algorithm 6 DDPG-based Phase Shift and Power Allocation Optimization Algorithm

Input: Critic network Q_ψ, actor network μ_ϕ, length of time frame T, empty replay buffer $\mathcal{D}$

Construct target parameters $\phi_{\text{targ}} \leftarrow \phi$ and $\psi_{\text{targ}} \leftarrow \psi$

repeat

  for $t = 1, \ldots, T$ **do**

    Observe states $\mathbf{s}_t$ and select actions $\mathbf{a}_t = \text{clip}(\mu_\phi(\mathbf{s}_t) + \epsilon, -1, 1)$, where $\epsilon \sim \mathcal{N}$

    Execute $\mathbf{a}_t$ and obtain reward r_t

    Observe next state $\mathbf{s}_{t+1}$ and store $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_t)$ in replay buffer $\mathcal{D}$

    Randomly sample a batch $B^* = \{(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_t)\}$ from $\mathcal{D}$

    Compute $y_t = r_t + \kappa_t Q_\psi(\mathbf{s}_{t+1}, \mu_\phi(\mathbf{s}_{t+1}))$

    Update critic network with gradient

$$\Delta_\psi \frac{1}{|B^*|} \sum_{(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, r_t) \in B^*} (Q_\psi(\mathbf{s}_t, \mathbf{a}_t) - y_t)^2$$

    Update actor network with gradient

$$\Delta_\phi \frac{1}{|B^*|} \sum_{\mathbf{s}_t \in B^*} Q_\psi(\mathbf{s}_t, \mu_\phi(\mathbf{s}_t))$$

    Perform soft updates on target parameters by

$$\phi_{\text{targ}} \leftarrow \rho \phi_{\text{targ}} + (1 - \rho)\phi$$

$$\psi_{\text{targ}} \leftarrow \rho \psi_{\text{targ}} + (1 - \rho)\psi$$

  end for

until reaches the maximum training steps

Return ϕ_{targ} and ψ_{targ}

action space, reward function, and the neural networks are defined as follows:

5.4.3.1 State space

The state space consists of all environment parameters that may affect the agent's decisions. In terms of our model, the state space consists of all channel information in the current TS, the transmit power available at the beginning of the TS, the current QoS requirements and the path loss of all users. At the t-th TS, the state vector is defined as

$$\mathbf{s}_t = [P_{max,t}, \mathbf{h}_{1,s}(t), \mathbf{h}_{1,w}(t), \ldots, \mathbf{h}_{K/2,s}(t), \mathbf{h}_{K/2,w}(t), \mathbf{R}_{QoS}, \mathbf{L}_{\text{path}}]^T. \tag{5.19}$$

5.4.3.2 Action space

The action space consists of two parts: the phase shift $\theta_n(t)$ of each RIS element and the power allocation $p_{l,i}(t)$ of each user. Note that the value of $p_{l,i}(t)$ may varies significantly, which can cause unstable learning phase. To address this issue, we further represent the power allocation as $p_{l,i}(t) = \alpha_{l,i}(t)P_{max}/T$, where $\alpha_{l,i}(t) \in [0, 1]$ and P_{max}/T can be interpreted respectively as the power allocation factor and the average maximum power of individual TSs. Hence, the action vector of the t-th TS is defined as

$$\mathbf{a}_t = [\boldsymbol{\theta}(t), \alpha(t)]. \tag{5.20}$$

5.4.3.3 Reward function

The reward function consists of the system sum rate, a QoS constraint term, and a transmit power constraint term. The QoS constraint term follows the formulation in the DL loss function, i.e., (5.11). The transmit power constraint is calculated as the absolute value of the difference between the average power consumption and the average power constraint. The use of absolute value helps to prevent excess power consumption, while encouraging the agent to allocate all transmit power available. The reward function is formulated as

$$\begin{aligned} r_t = w_1 \sum_{l=1}^{K/2} \sum_{i=s,w} R_{l,i} + w_2 \sum_{l=1}^{K/2} \sum_{i=s,w} \max(R_{l,i} - R_{QoS}^{l,i}, 0) \\ + w_3 \mathbb{1}_T(t) \left| P_{max} - \frac{1}{T} \sum\nolimits_{j=1}^{T} \sum\nolimits_{l=1}^{K/2} (p_{l,s}(j) + p_{l,w}(j)) \right|, \end{aligned} \tag{5.21}$$

where $\mathbb{1}_T(t)$ is an indicator function, which returns one if $t = T$ and zero otherwise. w_1, w_2, and w_3 are tuning parameters to be adjusted in training.

5.4.3.4 Neural Networks

The architectures of the critic and the actor networks are described as follows. The actor network is a fully-connected network, which consists of two layers. Note that the counting

index starts from the first hidden layer up to the output layer. The first layer in the actor network is a dense layer of 1024 neurons, followed by a ReLU activation function. The second layer is another dense layer of 1024 neurons with the tanh activation function that outputs the actions in the range [-1, 1]. Then, the outputs are scaled to their desired ranges.

The critic network is a three-layer network, where the actions and the states are processed separately through two dense layers, each of 1024 neurons, followed by the ReLU layers. The outputs are concatenated before feeding to a dense layer of 1024 neurons with the ReLU activation. Then, the final layer outputs a real number as the approximation for the action-value function. It is worth pointing out that, before the first dense layers in both actor and critic networks, a batch-normalization layer is inserted to increase the convergence rate and to ensure the stability of the training process.

5.4.4 Convergence Analysis

Inspired by the success of DQN [100, 101], DDPG is developed by extending the deterministic policy gradient (DPG) algorithm [102] with non-linear neural networks to better approximate the policy functions and the action-value functions, which can be especially beneficial in high-dimensional problems. The DPG algorithm is guaranteed to converge since the gradient of the cumulative reward with respect to the policy parameters ϕ has been proved to exist. However, this convergence is only compatible with a function approximator of $Q(s, a)$ that is linear of the policy features.

Remark 2. *The use of non-linear neural networks implies that the convergence of DDPG is not guaranteed.*

In general, the convergence of DRL algorithms depends heavily on the accuracy of the function approximator, because a function can be approximated to greater or lesser degrees by using more or less complex polynomials for approximation. Hence, the convergence of the proposed DRL algorithm is verified based on extensive simulation results, as shown in Sec. 5.5, where the results demonstrate that the DRL algorithm is

Table 5-B: DL simulation configurations.

Parameter	Value
Number of MUs, K	4
Number of BS antennas, M	16
Path loss exponent	3
Noise power spectral density	-169 dBm/Hz
Bandwidth, B_l	4 MHz
Training batch size	128
Testing batch size	1,000
Network learning rate, γ_η	0.0001
Phase shift initial learning rate, γ_θ	0.3
Phase shift update steps, J	5
Maximum training episode	10,000

able to converge under various hyper-parameter values.

5.4.5 Complexity Analysis

The DRL algorithm does not have an offline training phase, hence, the online learning complexity of the DRL algorithm is derived in terms of two factors: the neural network complexity and the loss function complexity.

The DRL algorithm utilizes four neural networks, namely, the actor network, the critic network, the target actor network and the target critic network. In particular, two types of neural network architectures are adopted, corresponding to the actor and critic networks. The actor network inputs the states, which consist of three channel matrices, a QoS requirement vector, the path loss vector, and the remaining transmit power. Hence, the input dimension of the actor network is $(2KM + 2KN + 2MN + 2K + 1)$. The actor network then outputs the actions, which consist of both the phase shift and the power allocation vectors, leading to an output dimension of $(N + K)$. Therefore, the asymptotic complexity of the actor network is $\mathcal{O}(2KM+2KN+2MN+2K+N+K+1) = \mathcal{O}(KM + KN + MN)$.

The complexity of the critic network can be derived similarly. To be specific, the critic network inputs both the states and the actions, then outputs a real number, resulting in a total complexity of $\mathcal{O}(2KM + 2KN + 2MN + K + N + K + 2) = \mathcal{O}(KM + KN +$

MN). Combing the complexity of the actor and the critic network, the overall network complexity is derived as $\mathcal{O}(KM + KN + MN)$, after omitting the simpler terms.

The complexity of the loss function can be obtained following the results of the DL algorithm, that is, $\mathcal{O}(NKM)$. Hence, given N_{ep} number of training episodes, the asymptotic complexity of the DRL algorithm is derived as $\mathcal{O}(N_{\text{ep}}NKM)$.

5.5 Numerical Results

In this section, simulation results are provided to evaluate the sum rate maximization performance of the DL and the DRL algorithms in the proposed RIS-assisted PD-NOMA downlink systems. Users are assumed to travel in a 10x10 square area, where the BS is located at a corner of the area and a RIS is randomly located within the area. As the baseline model, OMA deploys the ZF precoding based on the channels of each user and the BS transmits signals without superposition coding. After receiving the signal, each user directly decodes the intended signal by considering all other users' signals as interference.

5.5.1 Short-term Optimization with DL

This subsection presents the simulation results of the proposed MAML-based DL algorithm for solving the short-term optimization problem, as formulated in (5.9). The simulation configurations are presented in Table 5-B unless otherwise stated.

5.5.1.1 Convergence of DL

The convergence of the DL algorithms mainly depends on the values of the training parameters. To evaluate the convergence of the proposed DL algorithm, we illustrate the loss function values as the number of training iterations increases, using different batch sizes and different learning rates, as shown in Fig. 5.4. It can be observed that, as the batch size increases from 32 to 128, the loss function values decrease at a faster rate and converge more steadily to a lower value. Moreover, when the learning rate is 0.0001,

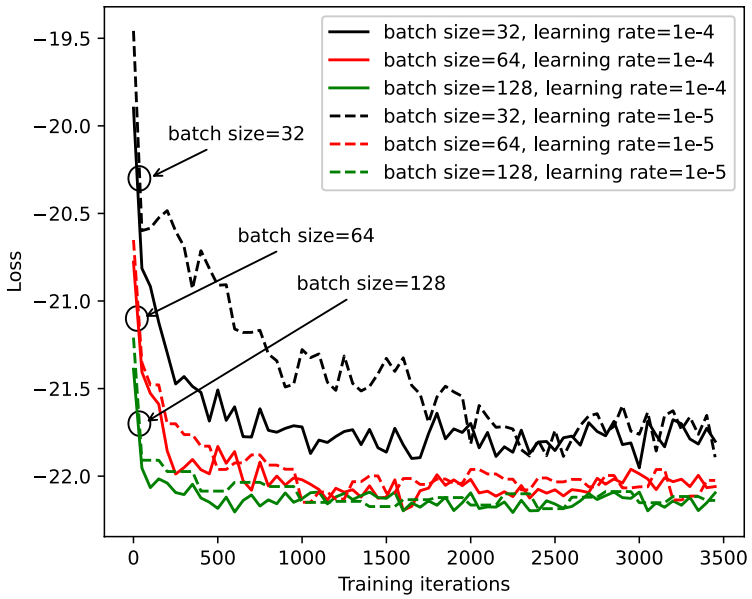


Figure 5.4: Training loss versus the number of training iterations for different batch sizes and learning rates.

we notice that fewer iterations are required for the loss function values to converge, compared to the case when the learning rate is 0.00001. Hence, in consideration of the stability and the computational efficiency, we adopt a batch size of 128 and a learning rate of 0.0001 in the subsequent simulations.

5.5.1.2 Sum rate versus the number of RIS elements

In Fig. 5.5, it can be observed that the PD-NOMA system with the proposed QoS-based clustering scheme outperforms the conventional OMA system by around 4 dBm/Hz of sum rate, without the enhancement of RIS. In both PD-NOMA and OMA systems, the deployment of RIS induces approximately 5% to 25% sum rate gain as the number of reflecting elements ranges from $N = 8$ to $N = 64$. Higher performance gain can be attained by increasing the number of RIS elements, however, the cost of optimization complexity and deployment increases as well.

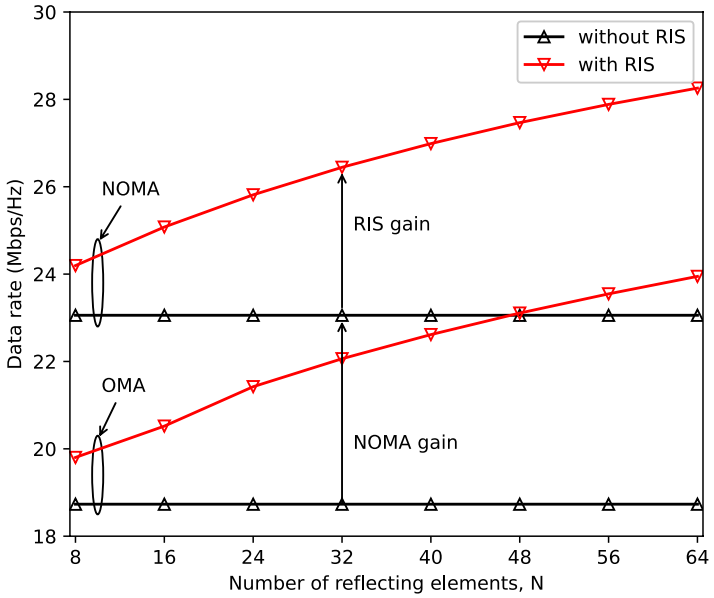


Figure 5.5: Sum rate versus the number of reflecting elements N for PD-NOMA and OMA cases, given 20 dBm BS transmit power.

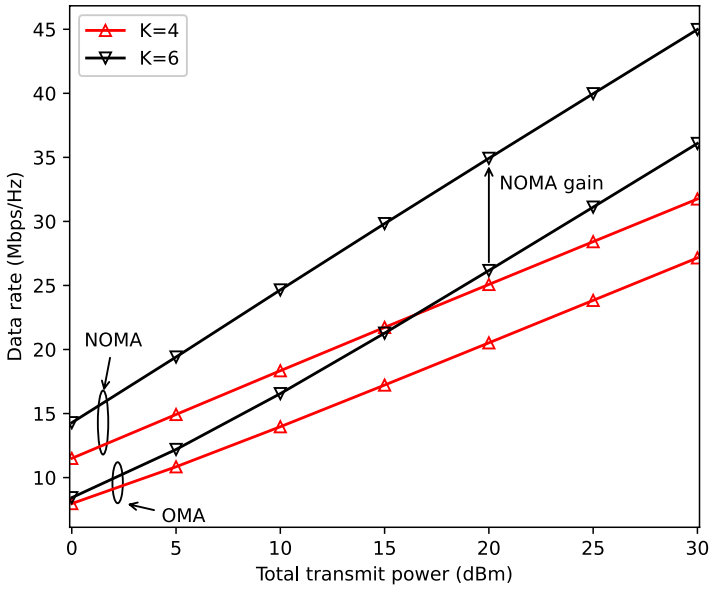


Figure 5.6: Sum rate versus total transmit power at BS for PD-NOMA and OMA cases, given $N = 16$ reflecting elements.

5.5.1.3 Sum rate versus BS total transmission power

Fig. 5.6 illustrates the sum rate performance between OMA and PD-NOMA systems as the BS power varies between 0 dBm and 30 dBm. It can be observed that the PD-NOMA

Table 5-C: DDPG simulation configurations.

Parameter	Value
Number of MUs, K	4
Number of BS antennas, M	4
Path loss exponent	3
Noise power spectral density	-169 dBm/Hz
Bandwidth, B_l	1 MHz
Batch size	128
Actor learning rate	0.0001
Critic learning rate	0.001
Discount factor, κ_t	1.0
Exploration noise standard deviation	0.05
Number of steps per episode, T	10
Number of maximum episodes	10,000

system outperforms the OMA system for all values of BS transmit power, given the same number of users. It can also be noticed that the PD-NOMA system with 4 users achieves a higher sum rate than the OMA system with 6 users when the BS power is less than 15 dBm. Moreover, as the number of users increases, the sum rate of the PD-NOMA systems increases by a larger amount compared to the sum rate of the OMA systems.

5.5.2 Long-term Optimization with DRL

This subsection demonstrates the simulation results of the DDPG-based DRL algorithm when applied to the long-term optimization problem, as formulated in (5.14). The DRL configurations are presented in Table 5-C unless otherwise stated. In particular, the discount factor is set to 1.0 because the length of each episode is finite, hence there is no risk of infinite rewards.

5.5.2.1 Convergence of DRL

To investigate the convergence of the adopted DDPG algorithm, experiments are conducted based on different batch sizes and critic learning rates, where the results are illustrated in Fig. 5.7. Since the performance of the DDPG algorithms is strongly determined by the approximation accuracy of the critic function, the discussions are focused on the impacts of the critic learning rates. From Fig. 5.7, it can be noticed that the rewards increase and converge as the number of training episodes increases, demonstrat-

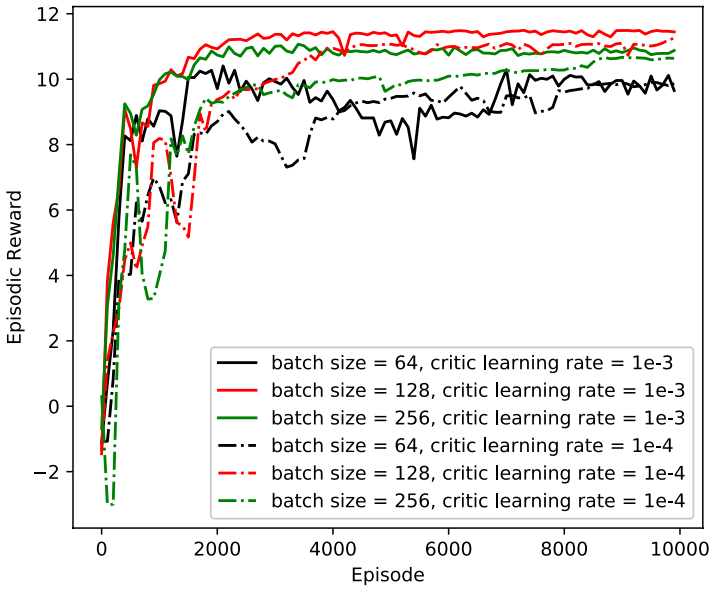


Figure 5.7: Episode rewards of the DRL algorithm versus the number of training episodes, under different batch sizes and critic learning rates.

ing that all agents have successfully learned to improve their policies for achieving higher rewards. In particular, as the batch size increases from 64 to 256, the rewards increase more steadily to a higher value. Moreover, as the learning rate increases, the rewards increase at a faster rate. However, different learning rates do not have a noticeable impact on the reward after convergence. Hence, a batch size of 128 and a critic learning rate of 0.001 is adopted in the subsequent experiments to achieve higher rewards at a faster convergence speed.

5.5.2.2 Sum rate versus the number of RIS elements

Fig. 5.8 illustrates the sum rate performance of the DRL algorithm as the number of RIS elements increases from $N = 4$ to $N = 16$. It can be observed that, under both PD-NOMA and OMA models, the sum rate is improved by a significant amount after the implementation of RIS. Moreover, as the number of RIS elements increases, the sum rate continues to rise, demonstrating the benefits of employing RIS for sum rate improvement. Results also show that the proposed PD-NOMA systems achieve a higher sum rate compared to the traditional OMA systems. More importantly, without the aid

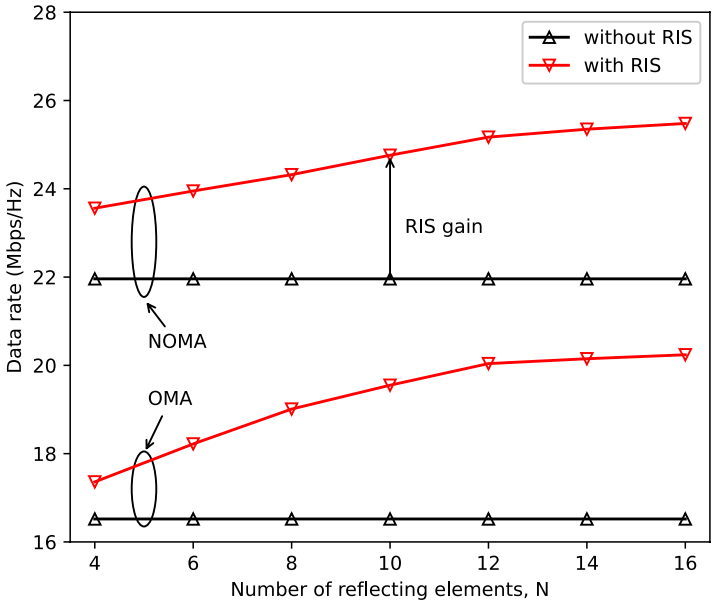


Figure 5.8: sum rate versus the number of RIS elements, under $P_{max} = 10$ dBm BS power and $M = 4$ BS antennas.

of RIS, the proposed PD-NOMA system can outperform RIS-assisted OMA systems by a substantial amount of sum rate.

5.5.3 DL versus DRL

This subsection investigates the sum rate difference between DL and DRL when both are applied to the short-term and the long-term problems, respectively. The computational complexity of the two algorithms is compared to further illustrate the performance difference. To solve the short-term problem with DRL, the algorithm considers it as a long-term problem that consists of only one TS, i.e., $T = 1$. To solve the long-term problem with DL, the algorithm decomposes it into several individual short-term problems and solves them separately, where the total transmit power of each TS is the average transmit power in the long-term problem. Then, the long-term sum rate of DL is calculated as the total sum rate of all the obtained short-term solutions.

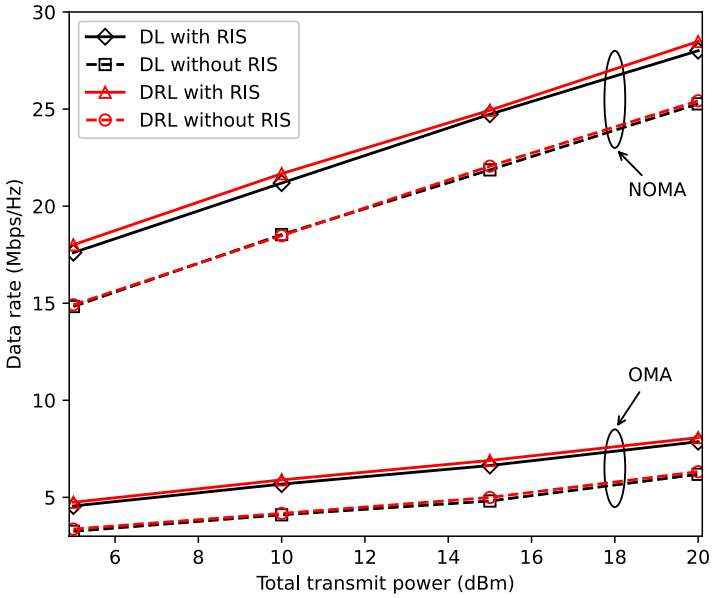


Figure 5.9: Sum rate of DL and DRL versus BS transmit power in PD-NOMA and OMA systems, for the short-term optimization problem.

5.5.3.1 Short-term optimization performance

The training configurations of the DL algorithm follow Table 5-B. The configurations of the DRL algorithm follow Table 5-C, except for the bandwidth, which is set to 4 MHz. The number of RIS elements is $N = 4$ for both methods. Fig. 5.9 illustrates the sum rate performance of both methods as the BS transmit power varies from 5 dBm to 20 dBm. It can be observed that DL and DRL achieve similar transmission sum rates in all scenarios, where the data rate of both approaches grows linearly with the BS transmit power. It can also be observed that PD-NOMA provides significant performance gain against OMA and the deployment of RIS has enhanced the system sum rate by a substantial amount in both PD-NOMA and OMA systems.

Fig. 5.10 compares the system sum rate given different clustering methods, namely, the proposed QoS-based clustering scheme and the conventional channel condition-based clustering scheme. Simulations are performed with $M = 24$ antennas at the BS and $P_{\max} = 20$ dBm maximum transmit power. It can be noticed that the proposed QoS-based method outperforms the conventional channel condition-based approach and the

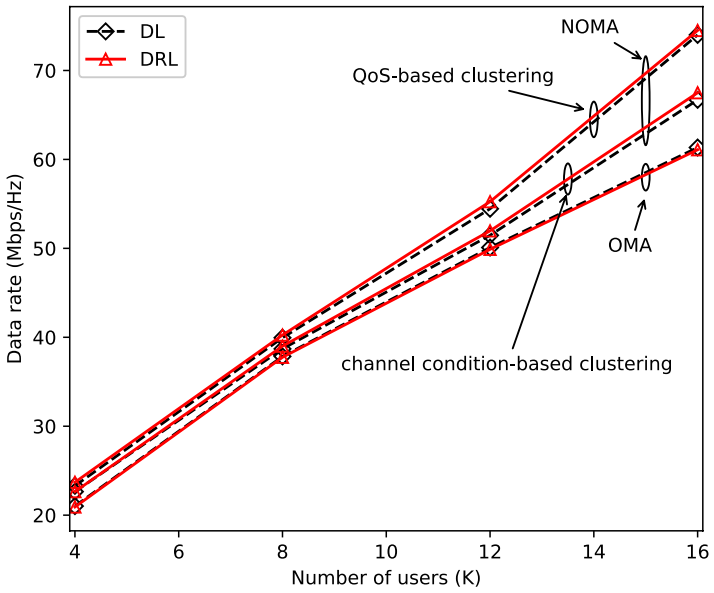


Figure 5.10: Sum rate of DL and DRL versus the number of users under the QoS-based or channel-based clustering schemes, when solving the short-term optimization problem.

performance gain of the QoS-based method further increases as the number of users increases. It implies that the QoS-based clustering scheme allows for a more resource-efficient power allocation strategy, where a higher system sum rate can be achieved compared to the baseline method given the same amount of transmit power.

Combining the observations from Fig. 5.9 and Fig. 5.10, it can be noticed that there is a negligible sum rate difference between DL and DRL for the short-term optimization in PD-NOMA and OMA systems, with or without the RIS. Hence, to further illustrate the performance difference between DL and DRL, we compare the computational complexity of the two approaches later in this subsection.

5.5.3.2 Long-term optimization performance

In Fig. 5.11, the long-term optimization performance of DL and DRL is studied by comparing their total sum rates versus total transmit power over 10 consecutive TSs, where $M = 4$ and $K = 4$. To illustrate the long-term performance difference between DL and DRL, a power-constrained scenario, is considered, where the total transmit power is

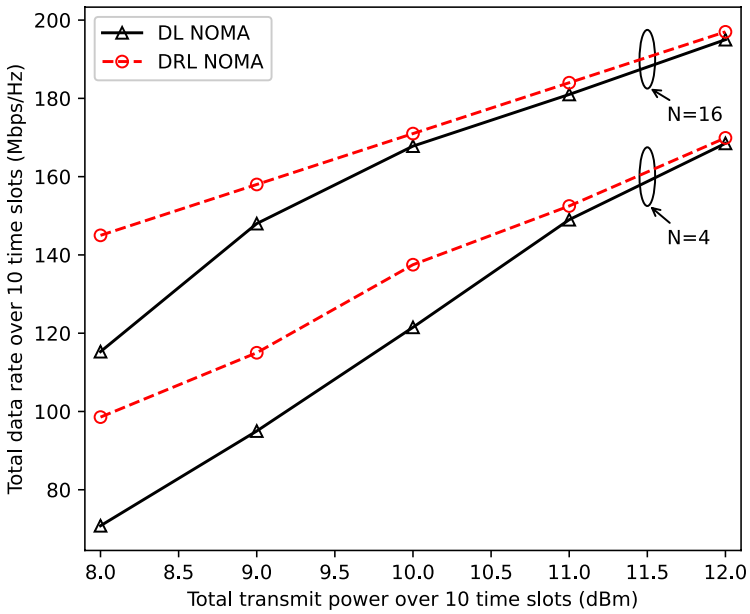


Figure 5.11: Sum rate of DL and DRL versus total transmit power when solving the long-term optimization problem in PD-NOMA systems.

insufficient to fulfil all QoS requirements in certain TSs. Note that, if at least one QoS requirement is unsatisfied, the sum rate of the corresponding TS is deducted to 0 dBm.

From Fig. 5.11, a noticeable performance gap between DL and DRL solutions is observed, which becomes more significant as the transmit power decreases. This is because, to solve the long-term problem, the DL method needs to decouple the problem into individual TSs, where each of them is treated as a short-term problem. Hence, due to the limited transmit power, the DL solutions are more likely to fail the QoS requirements compared to the DRL solutions, which are obtained by optimizing the power consumption of each TS. Moreover, as the number of RIS elements increases from $N = 4$ to $N = 16$, the performance difference between the two methods is reduced because the transmission sum rate is improved and more users' QoS requirements can be fulfilled given the same total transmit power.

5.5.3.3 Computational Complexity Comparison

Comparing the computational complexity between DL and DRL is not straightforward, because DL offloads the majority of the complexity to the offline training stage, whereas DRL is based on the online training. However, in practical scenarios, the available computational resources are more sufficient for offline training than those for online training. Therefore we compare the complexity of DL and DRL with a particular focus on the online training stage.

For both DL and DRL, the computational complexity of the long-term optimization is T times the computational complexity of the short-term optimization, where T denotes the number of TSs. Therefore, the analysis is focused on comparing the computational complexity of the short-term optimization between DL and DRL. As derived in the Sec. 5.3.5 and 5.4.5, the online computational complexity of DL and DRL for the short-term optimization are $\mathcal{O}(JKM + NKM + JN)$ and $\mathcal{O}(N_{ep}NKM)$, respectively, where J denotes the number of phase shift optimization iterations in the DL algorithm and N_{ep} denotes the number of training episodes before the DRL algorithm converges. It is worth noting that the practical values of N_{ep}, are significantly larger than all the other variables. For instance, when $K = N = M = 4$, the DL algorithm converges for $J = 5$, whereas the DRL algorithm converges for $N_{ep} \approx 10,000$. The sufficiently large values of N_{ep} cause DRL to have much higher computational complexity than DL during the online training stage. As a result, although DL and DRL demonstrated a similar sum rate for the short-term optimization, DL is superior due to the relatively low computational complexity.

5.5.3.4 Overall comparisons between DL and DRL

Having compared the short-term sum rate performance, the long-term sum rate performance, and the computational complexity between DL and DRL, it can be concluded that, compared to DRL, DL achieves a slightly lower sum rate but yields significantly lower online training complexity. Hence, DL is preferred when the computational

resources are limited. By contrast, when there are sufficient computational resources, DRL is preferred because it achieves a higher transmission sum rate, especially when solving the long-term optimization problems due to the advantage of coordinating the transmit power among the TSs.

5.6 Summary

In this article, a QoS-based clustering method is proposed to improve resource efficiency in RIS-assisted PD-NOMA systems. The sum rate maximization problem was formulated by jointly optimizing the RIS phase shift and the BS power allocation. The optimization problem was formulated from both short-term and long-term prospects. Both DL and DRL techniques were employed to solve the formulated problems. The DL algorithm adopted a low-complexity network architecture and utilized MAML to improve the convergence rate in the application. The DRL algorithm utilized a transmit power-based penalty term in the reward function to regulate the power consumption, such that the average transmit power constraint can be satisfied. Simulation results demonstrated that the proposed QoS-based PD-NOMA scheme achieved a higher sum rate compared to the conventional channel condition-based PD-NOMA and OMA schemes. It also revealed that the implementations of RISs significantly improved the achieved system sum rate. Moreover, we noticed that DRL achieved a higher transmission sum rate than DL, especially for the long-term problems. In terms of the algorithm complexity, DL was superior as a low-complexity solution compared to DRL. Overall, DL is preferred when the computational resources are scarce and DRL is preferred to achieve a higher transmission sum rate given sufficient computational resources. The work of this chapter demonstrated the spectral enhancement of employing a RIS in PD-NOMA systems. To further investigate the integration of RIS and PD-NOMA, the next chapter will study the implementation of multiple RISs in PD-NOMA systems for enhancing D2D communications.

Chapter 6

Multi-Agent Resource Allocation in NOMA-Enhanced Multi-RIS Aided D2D Networks

6.1 Introduction

By providing proximity communication for paired mobile users, D2D communication has been recognized as one of the promising technologies to enhance network capacity and alleviate traffic burdens in wireless networks. However, as the number of D2D equipment increases, the severe co-channel interference greatly limits the spectral efficiency, which can be addressed by the SIC technique of PD-NOMA. Moreover, as a low-cost spectral enhancement technique, RISs can be easily deployed in the environment to establish LoS links for enhancing signal strength and configure the wireless signals for interference reduction. Hence, in this chapter, a PD-NOMA-enhanced multi-RIS aided D2D communication network is investigated, where the sum rate maximization problem is formulated by jointly optimizing RIS phase shifts, PD-NOMA power allocations, and sub-channel assignments for D2D receivers. Based on the time-varying channels, a multi-agent hybrid

action DRL (MAHA-DRL) algorithm is designed for resource allocation, where each D2D transmitter (DT) and each RIS controller act as an agent and the challenging hybrid action space is addressed by utilizing MP-DQNs. The main contributions are outlined as follows:

- A PD-NOMA-enhanced multi-RIS aided D2D underlaying cellular network is proposed, where each DT can communicate with multiple D2D receivers (DRs) simultaneously through PD-NOMA transmission. To maximize the long-term sum rates of DRs, the resource allocation problem is formulated by jointly optimizing D2D channel assignments, PD-NOMA power allocations, and RIS phase shifts, subject to a time-varying channel model.

- The high-dimensional long-term optimization problem is addressed by constructing a multi-agent DRL-based resource allocation algorithm. In particular, each DR and each RIS controller act as an agent, who are trained in a CTDE structure.

- MP-DQN is employed for each agent to directly incorporate the hybrid discrete-continuous action space while exploiting the correlations between the discrete actions and the continuous actions.

- Simulation results verify the stable convergence of the proposed MAHA-DRL algorithm under various network scenarios. Results also illustrate the sum rate enhancement capability of PD-NOMA compared to OMA when applied to D2D networks. Moreover, the implementations of multiple RISs also introduce significant sum rate improvements in both PD-NOMA-based and OMA-based networks, compared to those without RIS.

Table 6-A: List of main notations.

Notation	Description	Notation	Description
I	Number of cellular users k	$p^k_{d,i}, p^k_i$	Transmit power allocation
N	Number of RISs	$x_{d,i}, x_i$	Intended signals
D	Number of D2D groups	v^d_k, v^i_k	Sub-channel assignments
T	Number of time slots (TSs)	$H^k_{b,i}$, $H^k_{d,j}$	Combined channel gain
K	Number of sub-channels	$\theta_{n,m}$	Phase shift m of RIS n
M	Reflecting elements per RIS	$\pi^k_{d,1}, \pi^k_{d,2}$	Decoding order of D2D receivers
Q	Number of sub-surfaces per RIS		

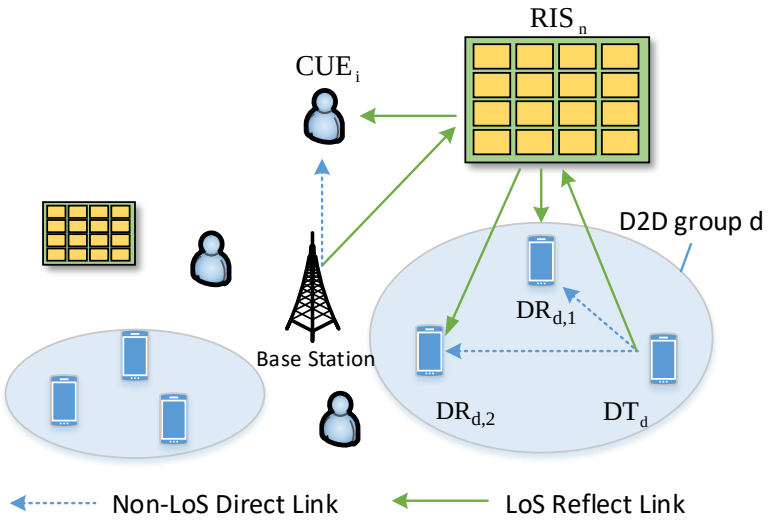


Figure 6.1: Illustration of the PD-NOMA-enhanced multi-RIS aided D2D network underlaying cellular networks.

6.2 Network Model

6.2.1 System Model

Consider a downlink D2D communication underlaying cellular network, which consists of D D2D groups, one BS, and I cellular users (CUEs), as illustrated in Fig. 6.1. Without loss of generality, the DTs, the DRs, the BS, and the CUEs are all assumed to be equipped with a single antenna. To avoid severe co-channel multi-user interference, each D2D group d consists of one DT and DRs, denoted by DT_d, $DR_{d,1}$, and $DR_{d,2}$, respectively. Hence, the set of D2D transmitters is denoted as $\mathcal{D}_T = \{DT_1, \ldots, DT_D\}$ and the set of D2D receivers is denoted as $\mathcal{D}_R = \{DR_{1,1}, DR_{1,2}, \ldots, DR_{D,1}, DR_{D,2}\}$. Moreover, the set of CUEs is denoted as $\mathcal{I} = \{1, \ldots, I\}$. The total bandwidth, denoted

by W, is divided into K orthogonal sub-channels, denoted by $\mathcal{K} = \{1, \ldots, K\}$, where each sub-channel occupies a bandwidth of ΔB Hz. Furthermore, the DTs, DRs, and CUEs are considered to be mobile equipment, travelling randomly during T time slots (TSs). The main notations are listed in Table 6-A.

As shown in Fig. 6.1, the direct links between the DTs and the DRs, and between the BS and the CUEs are blocked by obstacles such as buildings or trees. To address this issue, N RISs are deployed to assist the transmission, where LoS links can be established between the RIS and all other equipment, including the BS, the CUEs, the DRs, and the DTs. Each RIS composes of M reflecting elements (REs), which are divided into Q sub-surfaces. Within each sub-surface, the phase shifts of all reflecting elements are the same.

6.2.2 Channel Model

To avoid co-channel interference, each CUE is considered to occupy one of the K orthogonal sub-channels and each channel can be assigned to only one of the CUEs. Hence, the combined channel from the BS to CUE i over sub-channel k at TS t is formulated as

$$H_{b,i}^{k}(t) = \sum_{n \in \mathcal{N}} \mathbf{g}_{n,k,i}^{H}(t) \mathbf{\Theta}_n(t) \mathbf{f}_{b,k,n}(t) + h_{b,k,i}(t), \tag{6.1}$$

where $\mathbf{g}_{n,k,i}^{H}(t) \in \mathbb{C}^{1 \times M}$ denotes the channel between RIS n and CUE i over sub-channel k at TS t, $\mathbf{\Theta}_n(t) = \text{diag}(\beta e^{j\theta_{n,1}}, \ldots, \beta e^{j\theta_{n,M}}) \in \mathbb{C}^{M \times M}$ denotes the phase shifting matrix of RIS n at TS t, $\mathbf{f}_{b,k,n}(t) \in \mathbb{C}^{M \times 1}$ denotes the channel between the BS and RIS n over sub-channel k at TS t, and $h_{b,k,i}(t) \in \mathbb{C}$ denotes the channel between the BS and CUE i over sub-channel k at TS t. All RISs are assumed to have unit amplitude coefficients and continuous phase shifts, i.e., $\beta = 1$, $\theta_{n,m} \in (0, 2\pi]$, $\forall m = 1, \ldots, M$. The path loss is modelled as $L^{-\alpha}$ for all transmissions, where L denotes the transmission distance and α denotes the path loss exponent.

Similarly, in D2D group d, the combined channel between transmitter DT_d and

receiver $DR_{d,j}$ is formulated as

$$H_{d,j}^{k}(t) = \sum_{n \in \mathcal{N}} \mathbf{g}_{n,k,d,j}^{H}(t) \mathbf{\Theta}_{n}(t) \mathbf{f}_{d,k,n}(t) + h_{d,k,d,j}(t), \tag{6.2}$$

where $\mathbf{g}_{n,k,d,j}^{H}(t) \in \mathbb{C}^{1 \times M}$ denotes the channel between RIS n and receiver $DR_{d,j}$ over sub-channel k at TS t, $\mathbf{f}_{d,k,n}(t) \in \mathbb{C}^{M \times 1}$ denotes the channel between transmitter DT_d and RIS n over sub-channel k at TS t, and $h_{d,k,d,j}(t)$ denotes the channel between transmitter DT_d and receiver $DR_{d,j}$ over sub-channel k at TS t. The transmission channels between the BS and DTs/DRs, and between the DTs and the CUEs can be derived in a similar manner. Hence, the detailed formulation will not be discussed here. In terms of the notations, $H_{b,d,j}^{k}(t)$ denotes the combined channel between the BS and receiver $DR_{d,j}$ over sub-channel k in TS t, $H_{b,d}^{k}(t)$ denotes the combined channel between the BS and transmitter DR_d over sub-channel k in TS t, and $H_{d,i}^{k}(t)$ denotes the combined channel between transmitter DR_d and CUE i over sub-channel k in TS t.

Moreover, the practical time-varying channels are considered [103], where the first-order Gaussian Markov channel model is adopted to formulate the small-scale fading components. Let $h(t)$ denote a small scale fading channel at TS t, the channel at TS $(t+1)$ is formulated as

$$h(t+1) = \epsilon \cdot h(t) + \sqrt{1-\epsilon^2} \cdot u(t+1), \tag{6.3}$$

where $u(t+1)$ is a random sample following the distribution of $h(t)$. Hence, $\epsilon = 1$ represents the stationary scenario and $\epsilon < 1$ represents the mobility scenario. In the considered network, all non-LoS direct links are modelled as Rayleigh fading channels and all LoS reflect links are modelled as Rician fading channels.

6.2.3 Signal Model

In each D2D group, PD-NOMA is employed so serve the DRs with the same orthogonal resource. In particular, let $x_{d,1}(t)$ and $x_{d,2}(t)$ denote the signals intended for receiver

$DR_{d,1}$ and receiver $DR_{d,2}$ in TS t, the superimposed signal transmitted by transmitter DT_d over sub-channel k is formulated as $\sqrt{p_{d,1}^k(t)}x_{d,1}(t) + \sqrt{p_{d,2}^k(t)}x_{d,2}(t)$, where $p_{d,1}^k(t)$ and $p_{d,2}^k(t)$ indicate the transmit power allocated to receiver $DR_{d,1}$ and receiver $DR_{d,2}$, respectively. Meanwhile, the signal intended for CUE i is denoted by $x_i(t)$ and $v_k^i(t) \in \{0,1\}$ denotes the sub-channel assignment variable of CUE i on sub-channel k at TS t. In particular, $v_k^i(t) = 1$ indicates that sub-channel k is assigned to CUE i at TS t and $v_k^i(t) = 0$ indicates otherwise. Similarly, $v_k^d(t)$ denotes the sub-channel assignment variable of D2D group d over sub-channel k. Hence, the signal received by CUE i over sub-channel k is formulated as

$$\begin{aligned} y_i^k(t) = & \underbrace{v_i^k(t)H_{b,i}^k(t)\sqrt{p_i^k(t)}x_i(t)}_{\text{Desired signal}} \\ & + \underbrace{\sum_{d\in\mathcal{D}} v_d^k(t)H_{d,i}^k(t)\left(\sqrt{p_{d,1}^k(t)}x_{d,1}(t) + \sqrt{p_{d,2}^k(t)}x_{d,2}(t)\right)}_{\text{D2D interference}} + \underbrace{\zeta_k}_{\text{Noise}}. \end{aligned} \tag{6.4}$$

For D2D group d, the signal received by receiver $DR_{d,1}$ over sub-channel k in TS t is formulated as

$$\begin{aligned} y_{d,1}^k(t) = & \underbrace{v_d^k(t)H_{d,1}^k(t)\sqrt{p_{d,1}^k(t)}x_{d,1}(t)}_{\text{Desired signal}} + \underbrace{v_d^k(t)H_{d,1}^k(t)\sqrt{p_{d,2}^k(t)}x_{d,2}(t)}_{\text{SIC signal}} \\ & + \underbrace{\sum_{i\in\mathcal{I}} v_i^k(t)H_{b,d,1}^k(t)\sqrt{p_i^k(t)}x_i(t)}_{\text{CUE interference}} \\ & + \underbrace{\sum_{\check{d}\neq d} v_{\check{d}}^k(t)H_{\check{d},d,1}^k(t)\left(\sqrt{p_{\check{d},1}^k(t)}x_{\check{d},1}(t) + \sqrt{p_{\check{d},2}^k(t)}x_{\check{d},2}(t)\right)}_{\text{D2D interference}} + \underbrace{\zeta_k}_{\text{Noise}}, \end{aligned} \tag{6.5}$$

where ζ_k indicates the AWGN over sub-channel k with variance σ^2.

To decode the superimposed signal, SIC needs to be employed by the DRs, where $\pi_{d,1}^k(t)$ and $\pi_{d,2}^k(t)$ denote the decoding order of receiver $DR_{d,1}$ and receiver $DR_{d,2}$ on sub-channel k at TS t, respectively. Without loss of generality, it is assumed that receiver

$DR_{d,2}$ employs SIC to decode and subtract the signal of receiver $DR_{d,1}$ from the received signal, i.e., $\pi^k_{d,1}(t) < \pi^k_{d,2}(t)$, where receiver $DR_{d,1}$ directly decodes its intended signal by treating the signal of $DR_{d,2}$ as interference. To ensure the accuracy of SIC, the following decoding constraint must be preserved:

$$\frac{|H^k_{d,2}(t)|^2 p^k_{d,1}(t)}{I^{2,in}_{d,2}(t) + I^{out}_{d,2}(t) + I^c_{d,2}(t) + \sigma^2} \geq \frac{|H^k_{d,1}(t)|^2 p^k_{d,1}(t)}{I^{2,in}_{d,1}(t) + I^{out}_{d,1}(t) + I^c_{d,1}(t) + \sigma^2}, \tag{6.6}$$

where $I^{2,in}_{d,2}(t) = |H^k_{d,2}(t)|^2 p^k_{d,2}(t)$ and $I^{2,in}_{d,1}(t) = |H^k_{d,1}(t)|^2 p^k_{d,2}(t)$ denote the intra-group interference at receiver $DR_{d,2}$ and receiver $DR_{d,1}$, respectively. Moreover, $I^{out}_{d,2}(t) = \sum_{\breve{d}\neq d} v^k_{\breve{d}}(t)|H^k_{\breve{d},d,2}(t)|^2(p^k_{\breve{d},1}(t)+p^k_{\breve{d},2}(t))$ and $I^{out}_{d,1}(t) = \sum_{\breve{d}\neq d} v^k_{\breve{d}}(t)|H^k_{\breve{d},d,1}(t)|^2(p^k_{\breve{d},1}(t)+p^k_{\breve{d},2}(t))$ denote the inter-group interference at receiver $DR_{d,2}$ and receiver $DR_{d,1}$, respectively. Nonetheless, $I^c_{d,2}(t) = \sum_{i\in\mathcal{I}} v^k_i(t)|H^k_{b,d,2}(t)|^2 p^k_i(t)$ and $I^c_{d,1}(t) = \sum_{i\in\mathcal{I}} v^k_i(t)|H^k_{b,d,1}(t)|^2 p^k_i(t)$ denote the CUE interference at receiver $DR_{d,2}$ and receiver $DR_{d,1}$, respectively. To simplify the expression in (6.6), let $F^k_d(2,1)(t)$ denotes the SIC condition of D2D group d on sub-channel k at TS t, which is given by

$$F^k_d(2,1)(t) = |H^k_{d,2}(t)|^2(I^{out}_{d,1}(t)+I^c_{d,1}(t)+\sigma^2) - |H^k_{d,1}(t)|^2(I^{out}_{d,2}(t)+I^c_{d,2}(t)+\sigma^2) \geq 0. \tag{6.7}$$

Hence, $F^k_d(2,1)(t) \geq 0$ indicates that the SIC condition is satisfied and $F^k_d(2,1)(t) < 0$ indicates otherwise. Finally, the SINR of receiver $DR_{d,1}$ over sub-channel k is given by

$$\gamma^k_{d,1}(t) = \frac{|H^k_{d,1}(t)|^2 p^k_{d,1}(t)}{I^{2,in}_{d,1}(t) + I^{out}_{d,1}(t) + I^c_{d,1}(t) + \sigma^2}. \tag{6.8}$$

Since receiver $DR_{d,2}$ employs SIC to decode and subtract the signal of $DR_{d,1}$ before decoding its own signal, the SINR of receiver $DR_{d,2}$ is computed as

$$\gamma^k_{d,2}(t) = \frac{|H^k_{d,2}(t)|^2 p^k_{d,2}(t)}{I^{out}_{d,2}(t) + I^c_{d,2}(t) + \sigma^2}. \tag{6.9}$$

For the cellular network, the SINR of CUE i over sub-channel k is formulated as

$$\gamma_i^k(t) = \frac{|H_{b,i}^k(t)|^2 v_i^k(t) p_i^k(t)}{I_{\mathcal{D},i}^k(t) + \sigma^2}, \tag{6.10}$$

where $I_{\mathcal{D},i}^k(t) = \sum_{d\in\mathcal{D}} |H_{d,i}^k(t)|^2 v_d^k(t)(p_{d,1}^k(t)+p_{d,2}^k(t))$ denotes the aggregated interference at CUE i imposed by the transmissions of all D2D groups on sub-channel k at TS t.

Then, the data rates of receiver $DR_{d,1}$ and $DR_{d,2}$ are given by

$$R_{d,1}(t) = \Delta B \log_2(1 + \gamma_{d,1}^k(t)), \tag{6.11}$$

and

$$R_{d,2}(t) = \Delta B \log_2(1 + \gamma_{d,2}^k(t)), \tag{6.12}$$

respectively. Similarly, the data rate of CUE i is computed as

$$R_i(t) = \Delta B \log_2(1 + \gamma_i^k(t)). \tag{6.13}$$

6.2.4 Problem Formulation

The aim is to maximize the sum rate of all DRs, by jointly optimizing the channel assignments of D2D groups, i.e., $\mathbf{v} = [v_1^1(1), \ldots, v_D^K(T)]^T$, the power allocations of DRs, i.e., $\mathbf{p} = [p_{1,1}^1(1), p_{1,2}^1(1), \ldots, p_{D,1}^K(T), p_{D,2}^K(T)]^T$, and the phase shifts, i.e., $\mathbf{\Theta} = \{\mathbf{\Theta}_1(1), \ldots, \mathbf{\Theta}_N(T)\}$. Since the optimization problem is focused on the resource allocation of D2D communications, random sub-channel assignments of CUEs are considered and the power allocation of each CUE is considered as the minimum transmit power that satisfies the QoS requirement. The sum rate maximization problem in the considered PD-NOMA-enhanced multi-RIS aided D2D network is formulated as follows:

$$(\mathbf{P1}): \quad \max_{\mathbf{v},\mathbf{p},\boldsymbol{\Theta}} \ \frac{1}{T}\sum\nolimits_{t=1}^{T}\sum_{d\in\mathcal{D}}(R_{d,1}(t)+R_{d,2}(t)) \tag{6.14a}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} v_d^k(p_{d,1}^k(t)+p_{d,2}^k(t)) \le p_d^{kmax}, \forall d, \forall t, \tag{6.14b}$$

$$R_i(t) \ge R_c^{min}, \forall i, \forall t, \tag{6.14c}$$

$$R_{d,j}(t) \ge R_d^{min}, \forall d, \forall t, \forall j \in \{1,2\}, \tag{6.14d}$$

$$\sum_{k\in\mathcal{K}} v_d^k(t) = 1, \forall d, \forall t, \tag{6.14e}$$

$$\sum_{d\in\mathcal{D}} v_d^k(t) \le K_d, \forall d, \forall t, \tag{6.14f}$$

$$F_d(u,v)(t) \ge 0, \text{if } \pi_{d,u}(t) > \pi_{d,v}(t), \forall u,v \in \{1,2\}, \forall d, \forall t, \tag{6.14g}$$

where p_d^{max} denotes the instantaneous maximum transmit power at transmitter DT_d in each TS, R_c^{min} denotes the minimum QoS requirement of each CUE, and R_d^{min} denotes the minimum QoS requirement of receiver DR_d. Hence, (6.14b) indicates the transmit power constraints; (6.14c) and (6.14d) specify the minimum QoS constraints of all CUEs and D2D groups, respectively; (6.14e) and (6.14f) denote the sub-channel assignment constraints, which indicate that each D2D group is assigned to one of the sub-channels and each sub-channel can be occupied by at most K_d D2D groups; (6.14g) specifies the SIC constraint for ensuring successful SIC decoding outcomes.

Unfortunately, (**P1**) is NP-hard and is non-trivial to be directly solved by conventional approaches. Moreover, due to the mobile DTs and DRs, and the time-varying channels, the communication environment is highly dynamic, which may not be fully exploited by conventional techniques. To address this NP-hard time series problem, a DRL-based resource allocation framework is proposed.

6.3 MAHA-DRL Algorithm for Resource Allocation

6.3.1 Multi-Agent DRL

As a major research interest of ML, RL learns an action selection policy, known as an agent, in a trial-and-error manner through continuous interactions with the unknown dynamic environment. The goal of the agent is to learn an optimal policy π^* which can maximize a long-term discounted reward. However, in complex networks involving a great variety of optimization parameters, a single agent may not be powerful enough to learn the joint optimal actions. By jointly training multiple agents, each of which is assigned a dedicated optimization task, the learning complexity can be offloaded and distributed to individual agents, such that the overall performance can be enhanced. Hence, the idea of multi-agent RL (MARL) is inspired [104]. MARL can be modeled as a Markov Game (MG), which is defined as a tuple $(\mathcal{C}, \mathcal{S}, (\mathcal{A}_i)_{i\in\mathcal{C}}, \mathcal{P}, \mathcal{R}, (\pi_i)_{i\in\mathcal{C}})$, where $\mathcal{C}$ denotes the set of agents and $C = |\mathcal{C}| > 1$. The main components of MG are described as follows:

- $\underline{\mathcal{A}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_C$: The joint action space which is composed of the action space of all agents. In particular, the joint action of all agents in TS t is denoted as $\underline{\mathbf{a}}(t)$.
- $\mathcal{S}$: The state space that consists of the observed environmental status, where $\mathbf{s}(t)$ denotes the state in TS t.
- $\underline{\pi}(\mathcal{S}) \rightarrow \mathcal{A}$: $\underline{\pi}$ denotes the joint policy of all agents. In particular, $\underline{\pi}(\underline{\mathbf{a}}|\mathbf{s}) = \prod_{i\in\mathcal{C}} \pi_i(\mathbf{a}_i|\mathbf{s})$, where $\pi_i(\mathbf{a}_i|\mathbf{s})$ represents the probability of agent i choosing action $\mathbf{a}_i$ after observing state $\mathbf{s}$, such that $\sum_{\mathbf{a}_i\in\mathcal{A}_i} \pi_i(\mathbf{a}_i|\mathbf{s}) = 1, \forall \mathbf{s} \in \mathcal{S}$.
- $\mathcal{P}(\mathcal{S} \times \underline{\mathcal{A}} \times \mathcal{S}) \rightarrow \mathbb{R}$: $\mathcal{P}$ denotes the transition probability function, where $\mathcal{P}(\mathbf{s}(t+1), \underline{\mathbf{a}}(t), \mathbf{s}(t))$ is interpreted as the probability of transitioning into state $\mathbf{s}(t+1)$ after executing the joint action $\underline{\mathbf{a}}(t)$ in state $\mathbf{s}(t)$.
- $\mathcal{R}(\mathcal{S} \times \underline{\mathcal{A}}) \rightarrow \mathbb{R}$: $\mathcal{R}$ indicates the reward function, where $\mathcal{R}(\underline{\mathbf{a}}(t), \mathbf{s}(t))$ is interpreted as the immediate reward obtained by executing the joint action $\underline{\mathbf{a}}(t)$ in state $\mathbf{s}(t)$.

For solving the sum rate maximization problem in the considered multi-RIS aided D2D network, a total of $(N+D)$ agents are deployed, which consist of N RIS agents for optimizing the phase shifts and D D2D agents for optimizing the sub-channel assignments and the power allocations. The key components of the multi-agent framework for solving (**P1**) are described as follows:

- *State space*: Both D2D and RIS agents require the information of all channel links and the previous actions, which is expressed as

$$\mathbf{s}_n(t) = \{\underline{\mathbf{h}}_{b,i}(t), \underline{\mathbf{g}}_{n,i}(t), \underline{\mathbf{f}}_{b,n}(t), \underline{\mathbf{h}}_{d,\check{d}}(t), \underline{\mathbf{g}}^D_{n,j}(t), \underline{\mathbf{f}}_{d,n}(t), \underline{\mathbf{h}}_{b,d,j}(t), \underline{\mathbf{h}}^D_{b,d}(t), \underline{\mathbf{h}}_{d,i}(t), \mathbf{v}(t-1), \mathbf{p}(t-1), \theta_{n,1}(t-1), \ldots, \theta_{N,Q}(t-1)\}, \quad (6.15)$$

  where $\underline{\mathbf{h}}_{b,i}(t) = \{h_{b,1,1}(t), \ldots, h_{b,K,I}(t)\}$ is the set of channels between the BS and all CUEs, $\underline{\mathbf{g}}_{n,i}(t) = \{\mathbf{g}_{1,1,1}(t), \ldots, \mathbf{g}_{N,K,I}(t)\}$ is the channels between the RISs and the CUEs, $\underline{\mathbf{f}}_{b,n}(t) = \{\mathbf{f}_{b,1,1}(t), \ldots, \mathbf{f}_{b,K,N}(t)\}$ is the channels between the BS and the RISs, $\underline{\mathbf{h}}_{d,\check{d}}(t) = \{h_{1,1,1,j}(t), \ldots, h_{D,K,D-1,j}(t)\}$ is the channels among the D2D groups, $\underline{\mathbf{g}}^D_{n,d}(t) = \{\mathbf{g}^D_{1,1,1}(t), \ldots, \mathbf{g}^D_{N,K,D}(t)\}$ is the channels between the RIS and the DRs, $\underline{\mathbf{f}}_{d,n}(t) = \{\mathbf{f}_{1,1,1}(t), \ldots, \mathbf{f}_{D,K,N}(t)\}$ is the channels between the DTs and the RISs, $\underline{\mathbf{h}}_{b,d,j}(t) = \{h_{b,1,1,j}(t), \ldots, h_{b,K,D,j}(t)\}$ is the channels between the BS and the DRs, $\underline{\mathbf{h}}^D_{b,d}(t) = \{h^D_{b,1,1}(t), \ldots, h^D_{b,K,D}(t)\}$ is the channels between the BS and the DTs, and $\underline{\mathbf{h}}_{d,i}(t) = \{h_{1,1,1}(t), \ldots, h_{D,K,I}(t)\}$ is the channels between the DTs and the CUEs, where $j \in \{1, 2\}$ indicates the index of the DRs in each D2D group. Moreover, $\mathbf{v}(t-1)$ denotes the channel assignments at TS $(t-1)$, $\mathbf{p}(t-1)$ denotes the power allocation in D2D group d at TS $(t-1)$, and $\theta_{n,q}(t-1)$ denotes the phase shift of sub-surface q on RIS n at TS $(t-1)$.

- *Action Space*: The D2D agents and the RIS agents have distinct action space. In particular, the action space of each D2D agent consists of the sub-channel assignment and the power allocations, which constitute a hybrid discrete-continuous

action space, given by

$$\mathcal{A}^{\text{D2D}} = \left\{ \left(k, p_{d,1}^{k}(t), p_{d,2}^{k}(t)\right) \quad \text{s.t. } p_{d,j}^{k} \in \mathcal{A}^{\text{D2D, cts}}, k \in \mathcal{K} \right\}, \tag{6.16}$$

where $\mathcal{A}^{\text{D2D, cts}}$ denotes the action space of power allocations. Then, the joint action of all D2D agents is expressed as $\underline{\mathbf{a}}^{D}(t) = \{\mathbf{a}_1^D(t), \ldots, \mathbf{a}_D^D(t)\}$, where $\mathbf{a}_d^D(t) \in \mathcal{A}^{\text{D2D}}$ indicates the hybrid action taken by D2D agent d. The action space of each RIS agent is also hybrid. In particular, a discrete variable $w_{n,q}(t) \in \{-1, 1\}$ is defined, known as phase shift adjustment direction, which specifies whether the phase shift of sub-surface q on RIS n should increase or decrease compared to the phase shift in the last TS. Then, a continuous phase shift variable $\breve{\theta}_{n,q}(t)$ is determined, which specifies the amount of phase shift adjustment compared to last TS. Therefore, the phase shift at TS t is computed based on the following equation:

$$\theta_{n,q}(t) = \theta_{n,q}(t-1) + w_{n,q}(t)\breve{\theta}_{n,q}(t). \tag{6.17}$$

Thus, the hybrid action space of RIS agent n is given by

$$\mathcal{A}^{\text{RIS}} = \left\{ \left(w_{n,q}(t), \breve{\theta}_{n,q}(t)\right) \quad \text{s.t. } w_{n,q}(t) \in \{-1, 1\}, \breve{\theta}_{n,q}(t) \in [0, \pi] \right\}. \tag{6.18}$$

Finally, the joint action of all RIS agents at TS t can be expressed as $\underline{\mathbf{a}}^{R}(t) = \{\mathbf{a}_1^R(t), \ldots, \mathbf{a}_{NQ}^R(t)\}$ and the joint action of all agents at TS t is given by

$$\underline{\mathbf{a}}(t) = \{\underline{\mathbf{a}}^{D}(t), \underline{\mathbf{a}}^{R}(t)\}. \tag{6.19}$$

- *Reward*: Since the aim is to maximization the sum rate of DRs, subject to the

constraints in (**P1**), the instantaneous reward of D2D agents at TS t is defined as

$$r^D(t) = \begin{cases} \sum\limits_{d\in\mathcal{D}} \sum\limits_{j\in\{1,2\}} (R_{d,j}(t) - R_d^{min}), & \text{if all constraints are satisfied} \\ U \sum\limits_{d\in\mathcal{D}} \sum\limits_{j\in\{1,2\}} |(R_{d,j}(t) - R_d^{min})|, & \text{otherwise,} \end{cases} \tag{6.20}$$

where U denotes the number of D2D groups that fail to meet the constraints. For the RIS agents, the instantaneous reward is computed as the reward gain compared to the network without RISs. Let $r^{D'}(t)$ denotes the reward achieved without the assistance of RIS, the reward of RIS agents is formulated as

$$r^R(t) = r^D(t) - r^{D'}(t). \tag{6.21}$$

Remark 3. *The carefully designed reward function for D2D agents is defined as the difference of the sum data rate of the D2D groups compared to the minimum data rate requirements, while that for the RIS agents is defined as the difference in sum rate between multi-RIS aided networks and conventional networks without RISs. Thus, as stated in [105], the proposed reward functions can provide useful guidance for the agents to improve their policy, thus achieving a higher convergence rate.*

6.3.2 MAHA-DRL Algorithm

The proposed MAHA-DRL algorithm integrates the multi-agent framework with the MP-DQN network to accommodate hybrid action space in multi-agent scenarios. To be specific, each agent consists of two networks, a policy network that outputs the associated continuous actions of each possible discrete action and a MP-DQN that outputs the Q values of all pairs of discrete and continuous actions, where the Q value can be interpreted as the goodness of choosing the given pair of actions in the given state. The multi-agent training algorithm adopts a CTDE structure, as illustrated in Fig. 6.2. During the training process, each agent observes the current state and computes all pairs of discrete

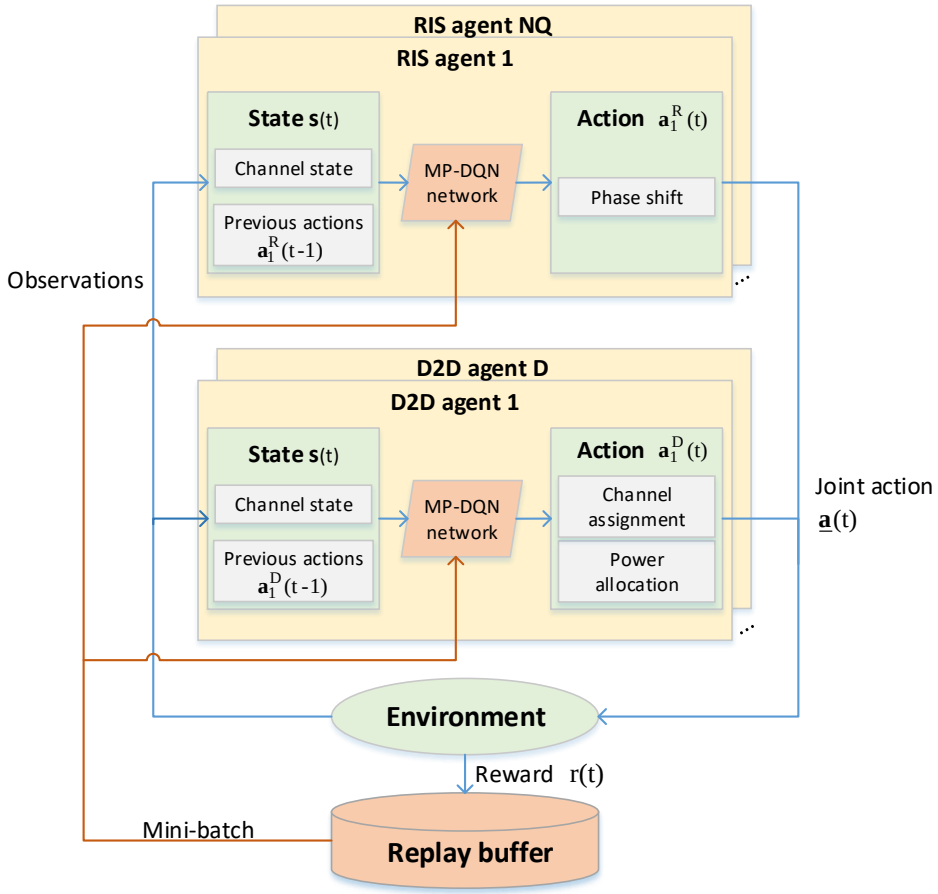


Figure 6.2: Flow diagram of MAHA-DRL algorithm based on CTDE training scheme.

and continuous actions using the policy network. Then, the actions are fed to Q-network to obtain all Q values. The optimal action pair that achieves the highest Q value is selected and uploaded to the central unit. After receiving the actions from all agents, the joint action is constructed and is executed in the environment, where a reward and the next state are obtained. The current state, the joint action, the reward, and the next state are stored in an experience replay buffer at the central unit. Then, the neural networks are trained by sampling mini-batches of data from the experience replay buffer. The main principles of the proposed MAHA-DRL algorithm are discussed as follows.

6.3.2.1 MP-DQN

The main challenge in the considered framework is the hybrid action space. In conventional DRL algorithms, the action space is either discrete, such as DQN [101] or continuous, such as TRPO [92]. The implementation of DRL with hybrid action space has recently emerged as a new research direction, where agents can exhibit more flexible and complex behaviors [106–108]. The existing solutions consist of two main approaches: transform the hybrid action space into a continuous one [108], or optimize the discrete and the continuous actions in an alternating manner [107]. However, both approaches do not fully exploit the correlations between discrete actions and continuous actions. For instance, continuous actions are often computed based on the decisions of the discrete action. To address this issue, the P-DQN framework [109] was introduced, in which the Q network inputs all continuous actions to calculate the Q value of each discrete action. This technique, however, introduces unnecessary correlations between the Q values of different pairs of actions. Hence, the MP-DQN network was recently proposed [110], which calculates the Q value of each individual pair of discrete and continuous actions separately as a mini-batch of data, such that the Q value gradient of a particular pair of actions is independent of that of the other pairs of actions.

In contrast to the conventional DQN framework, where the Q-network inputs the state and outputs the Q values of all actions, the Q-network in the MP-DQN framework inputs the state as well as $|\mathcal{A}^{\text{dis}}|$ pairs of discrete and continuous actions, and outputs the Q values of each pair of actions. To achieve this without introducing excess correlations among the actions, all $|\mathcal{A}^{\text{dis}}|$ continuous actions are concatenated as a vector, denoted by $\tilde{\mathbf{a}}^{\text{cts}} \in \mathbb{R}^{|\mathcal{A}^{\text{dis}}|}$. Then, define a total of $|\mathcal{A}^{\text{dis}}|$ basis vectors, where basis vector k is given by $\boldsymbol{e}_k = (0, \ldots, 1, \ldots, 0)$, such that $[\boldsymbol{e}_k]_i = 0, \forall i \neq k$. The continuous action vector $\tilde{\mathbf{a}}^{\text{cts}}$ is multiplied with each of the $|\mathcal{A}^{\text{dis}}|$ basis vectors to form $|\mathcal{A}^{\text{dis}}|$ product vectors, each of which represents a continuous action. Together with the state, a mini-batch of size $|\mathcal{A}^{\text{dis}}|$ is constructed as the inputs to the Q-network. For each input, also referred to as

a pass, in the mini-batch, the Q-network outputs $|\mathcal{A}^{\text{dis}}|$ Q values, which is formulated as

$$\begin{pmatrix} Q(s,\cdot,\tilde{\mathbf{a}}^{\text{cts}}\boldsymbol{e}_1;w) \\ \vdots \\ Q(s,\cdot,\tilde{\mathbf{a}}^{\text{cts}}\boldsymbol{e}_{|\mathcal{A}^{\text{dis}}|};w) \end{pmatrix} = \begin{pmatrix} Q_{1,1} & Q_{1,2} & \dots & Q_{1,|\mathcal{A}^{\text{dis}}|} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{|\mathcal{A}^{\text{dis}}|,1} & Q_{|\mathcal{A}^{\text{dis}}|,2} & \dots & Q_{|\mathcal{A}^{\text{dis}}|,|\mathcal{A}^{\text{dis}}|} \end{pmatrix}, \tag{6.22}$$

where $Q(\cdot)$ denotes the Q-network, w denotes the network weights, and $Q_{i,j}$ denotes the Q value computed for action pair j in pass i. Only the diagonal entries of the output contain useful information and are used in the action selection process. In terms of the proposed framework, a sub-channel assignment and the power allocation over the selected sub-channel form a pair of discrete and continuous actions of a D2D agent, i.e., a total of K passes are required. Similarly, a phase shift adjustment direction and the adjustment amount constitute the action pair of a RIS agent, i.e. a total of 2 passes are required. Generally, in the MAHA-DRL framework, the Q network of agent c, regardless if it is a D2D agent or a RIS agent, is denoted as $Q_c(\mathbf{s}(t), a_c^{\text{dis}}(t), \tilde{\mathbf{a}}_c^{\text{cts}}(t)\boldsymbol{e}_k; w_c)$, parameterized by w_c, and the deterministic policy network that outputs the continuous actions is denoted as $\pi_c(\mathbf{s}(t); w_{\pi,c})$, parameterized by $w_{\pi,c}$.

6.3.2.2 Exploration and Exploitation

Due to the hybrid action space, different exploration strategies need to be designed for the discrete and the continuous actions, respectively. All discrete actions are selected according to the ϵ-greedy strategy. The ϵ-greedy based action of agent c is expressed as

$$\tilde{a}_c^{\text{dis}}(t) = \begin{cases} \underset{a_c^{\text{dis}} \in \mathcal{A}^{\text{dis}}}{\text{argmax}}\ Q_c(\mathbf{s}(t), a_c^{\text{dis}}(t), \tilde{\mathbf{a}}_c^{\text{cts}}(t)\boldsymbol{e}_k; w_c), & \text{with probability } 1-\epsilon, \\ \text{sample } a_c^{\text{dis}}(t) \sim \mathcal{A}^{\text{dis}}, & \text{with probability } \epsilon, \end{cases} \tag{6.23}$$

where $\epsilon \in (0,1)$. Based on the selected discrete action, the exploration on the continuous action is conducted by adding a stochastic noise, generated by the Ornstein-Uhlenbeck

(OU) process, through the following equation:

$$\tilde{a}_c^{\text{cts}}(t) = \left[a_c^{\text{cts}}(t) + \mathcal{N}(0,1)\right]_0^{\overline{a}^{\text{cts}}}, \tag{6.24}$$

where $\overline{a}^{\text{cts}}$ denotes the maximum value of the continuous action.

After all agents have selected their actions, the joint action $\underline{\mathbf{a}}(t)$ is constructed. Then, by performing the joint action in the environment, the instantaneous reward $r(t)$ and the next state $\mathbf{s}(t+1)$ are received as the feedback. The tuple $(\mathbf{s}(t), \underline{\mathbf{a}}(t), r(t), \mathbf{s}(t+1))$ is stored in the experience replay buffer as potential training samples.

To update the neural networks, a mini-batch of B tuples are sampled from the replay buffer. For each agent c, the objective of the policy network π_c is to minimize the Q value, hence the policy loss function is formulated as

$$\mathcal{L}(w_{\pi,c}) = -\sum_{a_c^{\text{dis}} \in \mathcal{A}^{\text{dis}}} Q_c\Big(\mathbf{s}(t), a_c^{\text{dis}}(t), \pi_c(\mathbf{s}(t); w_{\pi,c}); w_c\Big). \tag{6.25}$$

For the Q-network, the objective is to minimize the difference between the estimated Q value and the target Q value. In particular, the target Q value is given by

$$y_c(t) = r(t) + \tau \max_{a_c^{\text{dis}} \in \mathcal{A}^{\text{dis}}} Q\Big(\mathbf{s}(t+1), a_c^{\text{dis}}(t+1), \pi_c(\mathbf{s}(t+1); w_{\pi,c}); w_c\Big). \tag{6.26}$$

Hence, the Q-network loss function of agent c is formulated as

$$\mathcal{L}(w_c) = \frac{1}{2}\left(y_c(t) - Q_c\Big(\mathbf{s}(t), \tilde{a}_c^{\text{dis}}(t), \pi(\mathbf{s}(t); w_{\pi,c}); w_c\Big)\right)^2. \tag{6.27}$$

Furthermore, to achieve a stable training process, the soft update technique is employed. To be specific, two target networks are constructed, namely the target Q-network and the target policy network. These two target networks are parameterized by w_c^- and $w_{\pi,c}^-$ and share the same structures as the Q-network and the policy network, respectively. In

Algorithm 7 MAHA-DRL Training Algorithm

Input: Maximum episodes E, maximum steps in each episode T, discounted factor τ
Output: The joint policy $\underline{a}$

1: Initialize network weights: w_c, $w_{\pi,c}$
2: Initialize target network weights:$w_c^- \leftarrow w_c$, $w_{\pi,c}^- \leftarrow w_{\pi,c}$
3: **for** $e = 1, 2, \ldots, E$ **do**
4: Initialize/reset the environment
5: **for** $t = 1, 2, \ldots, T$ **do**
6: **for** D2D group $d = 1, \ldots, D$ **do**
7: Obtain the joint sub-channel assignment k and power allocation $(p_{d,1}^k, p_{d,2}^k)$ using (6.23) and (6.24)
8: **end for**
9: **for** RIS $n = 1, \ldots, N$ **do**
10: **for** Sub-surface $q = 1, \ldots, Q$ **do**
11: Obtain the phase shift adjustment direction $w_{n,q}$ and adjustment amount $\breve{\theta}_{n,q}$ using (6.23) and (6.24)
12: **end for**
13: **end for**
14: Construct joint actions $\underline{\mathbf{a}}(t)$, obtain reward $r(t)$ using (6.21), and observe next state $\mathbf{s}(t+1)$
15: Store tuple $(\mathbf{s}(t), \underline{\mathbf{a}}(t), r(t), \mathbf{s}(t+1))$ in the experience replay buffer
16: Sample B tuples $(\mathbf{s}_i, \underline{\mathbf{a}}_i, r_i, \mathbf{s}_{i+1})$ from the experience replay buffer
17: For each agent, update weights w_c and $w_{\pi,c}$ by minimizing the loss functions (6.27) and (6.25), respectively
18: Update weights w_c^- and $w_{\pi,c}^-$ according to (6.28)
19: **end for**
20: **end for**

each training episode, after the networks are updated, the target networks are adapted through the following equations

$$\begin{aligned} w_c^- &\leftarrow \iota_Q w + (1 - \iota_Q) w_c^-, \\ w_{\pi,c}^- &\leftarrow \iota_\pi w_{\pi,c} + (1 - \iota_\pi) w_{\pi,c}^-, \end{aligned} \tag{6.28}$$

where $0 < \iota_Q \ll 1$ and $0 < \iota_\pi \ll 1$ denote the soft update coefficients. The pseudocode is presented in Algorithm 7. In particular, lines 4-13 present the decentralized execution process and lines 14-17 present the centralized learning process.

6.3.3 Complexity Analysis

The computational complexity of the MAHA-DRL algorithm during training mainly comprises of two parts, namely the policy optimization process and the action selection process. Let $|\mathcal{S}|$ denotes the dimension of the state space in (6.15) and $|\mathcal{A}|$ denotes the dimension of the action space in (6.16) and (6.18). Since $|\mathcal{A}| = K$ for D2D agents and $|\mathcal{A}| = 2$ for RIS agents, $|\mathcal{A}| = \mathcal{O}(K)$ is considered. Without loss of generality, it is assumed that each of the Q-networks and the policy networks consists of one fully-connected hidden layer of μ neurons. Hence, the complexity of the policy network is $\mathcal{O}(\mu|\mathcal{S}| + \mu|\mathcal{A}|)$. Similarly, the complexity of a MP-DQN is $\mathcal{O}(|\mathcal{A}| \times (\mu|\mathcal{S}| + 2\mu|\mathcal{A}|))$. Thus, the total complexity of the action selection procedure of one agent can be derived as $\mathcal{O}\left(\mu|\mathcal{S}||\mathcal{A}| + 2\mu|\mathcal{A}|^2 + \mu|\mathcal{S}| + \mu|\mathcal{A}|\right)$. The complexity due to activation functions are neglected and the complexity of forward and backward propagation are regarded as equivalent. Considering a total number of $(D+NQ)$ agents, the complexity of each time step is derived as $\mathcal{O}\left((D + NQ)(\mu|\mathcal{S}||\mathcal{A}| + 2\mu|\mathcal{A}|^2 + \mu|\mathcal{S}| + \mu|\mathcal{A}|)\right)$. Moreover, assuming that the total number training episodes is E, the number of steps in each episode is T, and the batch size is B, the total training complexity of the proposed MAHA-DRL algorithm can be derived as $\mathcal{O}\left(ETB(D + NQ)(\mu|\mathcal{S}||\mathcal{A}| + 2\mu|\mathcal{A}|^2 + \mu|\mathcal{S}| + \mu|\mathcal{A}|)\right)$.

6.4 Numerical Results

This section presents the simulation results of the MAHA-DRL algorithm when solving the sum rate maximization problem in PD-NOMA-enhanced multi-RIS aided D2D networks. All CUEs, DRs, and DTs are considered to be randomly roaming within a specific area of size 100 meters^2, where the BS is located at the centre of the area with coordinates $(0, 0)$. At most 4 RISs are considered in the experiments and their coordinates are $(0, 40)$, $(0, -40)$, $(40, 0)$, and $(-40, 0)$, respectively. The movement of each D2D group is restricted to one of the quadrants. The directional random model [111] is employed to model the movements of all CUEs, DRs, and DTs. For instance, the movement direction and speed of CUE i is modelled as $(\theta_i + \mathcal{U}(0, 2\pi))$ and $(\frac{4}{5} + \mathcal{U}(0, \frac{1}{5}))$ meters per second,

Table 6-B: Network and algorithm configurations.

Parameters	Values	Parameters	Values
Maximum power at DRs	$p_d^{max} = 20$ dBm	Batch size	64
Total bandwidth	10 MHz	Discount factor	$\tau = 0.9$
Number of CUEs	$I = 2$	Replay memory size	10000
Number of D2D groups	$D = 2$	Q-network learning rate	$\alpha_Q = 0.1$
Number of RISs	$N = 2$	Policy network learning rate	$\alpha_\pi = 0.001$
Number of sub-channels	$K = 2$	Number of sub-surfaces per RIS	$Q = 4$
Noise spectral density	-174 dBm/Hz	Number of REs per RIS	$M = 16$

respectively, where θ_i indicates the angle with respect to the positive x-axis.

All neural networks consist of two fully-connected hidden layers of 512 and 128 neurons, respectively. Each hidden layer is followed by the leaky ReLU activation function. The Adam optimizer is employed to minimize the loss functions. The path loss exponents of the BS-RIS, BS-CUE, RIS-CUE, RIS-DR, DT-RIS, and DT-DR links are set to 2.2, 3.5, 2.8, 2.6, 2.2 and 3.5, respectively. Unless otherwise stated, the network parameters and the algorithm configurations are listed in Table 6-B.

6.4.1 Algorithm Convergence

Fig. 6.3 verifies the convergence behaviour of the proposed MAML-DQN algorithm for maximizing the sum rate of PD-NOMA-enhanced multi-RIS aided D2D networks. The convergence analysis is focused on the learning rates of the Q-networks, denoted by α_Q, and the learning rates of the policy networks, denoted by α_π. By comparing the results under $\alpha_Q = 0.1$, it can be observed that, as α_π decreases from 0.1 to 0.001, the learning process becomes more stable and converges to a higher sum rate. This result indicates that the learning rate of the policy network should be smaller than that of the Q-network to ensure a stable learning process and improved performance. Moreover, as α_Q decreases from 0.1 to 0.001, the achieved sum rate decreases significantly. Hence, in the rest of the experiments, the learning rates are set to $\alpha_Q = 0.1$ and $\alpha_\pi = 0.001$, respectively.

Fig. 6.4 illustrates the convergence of the proposed MAHA-DRL algorithm under various D2D network scenarios, namely PD-NOMA with RIS, PD-NOMA without RIS,

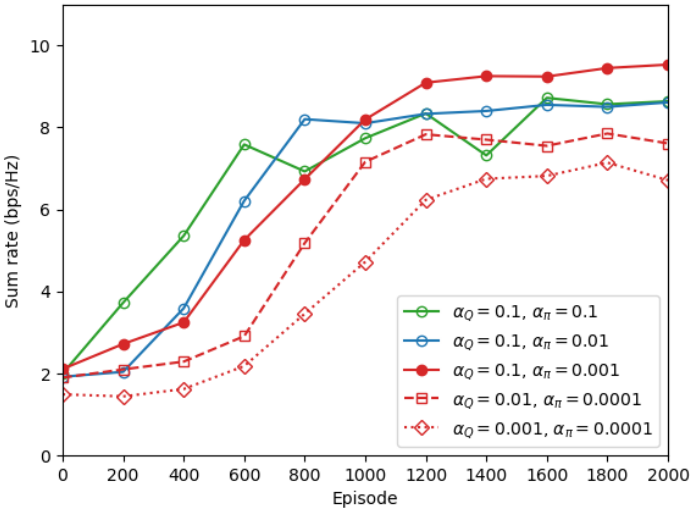


Figure 6.3: Episodic reward versus the number of episodes under PD-NOMA transmission scheme.

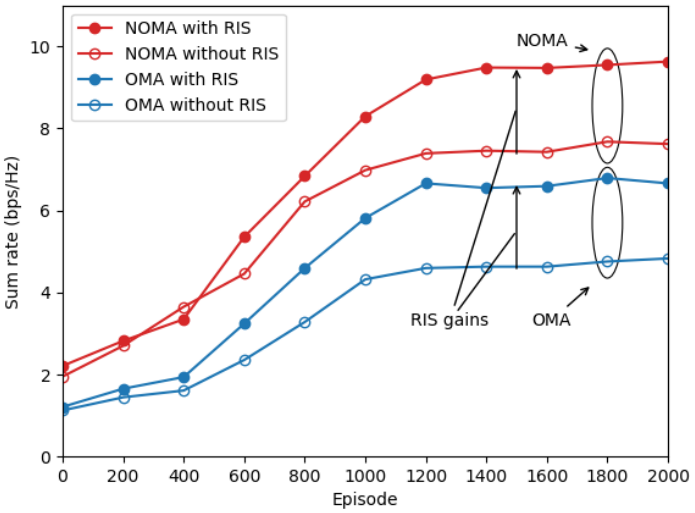


Figure 6.4: Episodic reward versus the number of episodes under PD-NOMA-based D2D networks and OMA-based D2D networks, with or without the assistance of RISs.

OMA with RIS and OMA without RIS. All RIS-aided networks consist of 2 RISs. The proposed algorithm demonstrates stable convergence in all considered networks, where the sum rate of the PD-NOMA-based networks is significantly higher than that of the

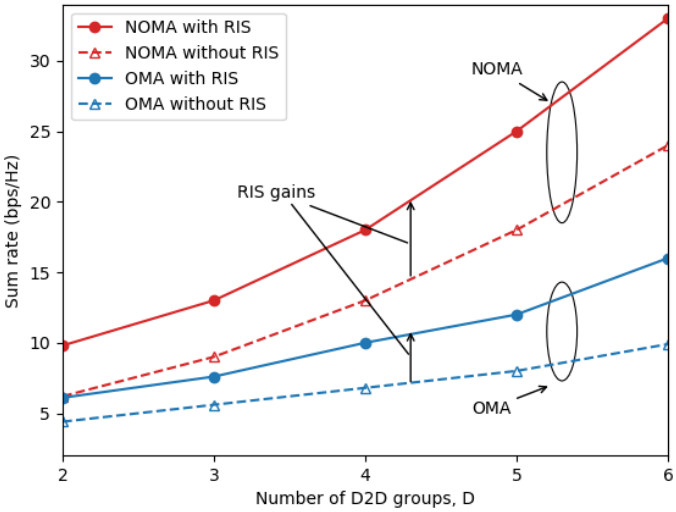


Figure 6.5: Sum rate versus the number of D2D groups under PD-NOMA-based D2D networks and OMA-based D2D networks, with or without the assistance of RISs.

OMA-based networks. During the early training stage, the sum rates of the RIS-aided networks are similar to that of the conventional networks without RIS. Then, with extensive learning and optimization, RIS-aided networks demonstrate increasing sum rate gain compared to the non-RIS networks, which indicates the benefits of RIS in sum rate enhancement in both PD-NOMA-based and OMA-based networks. Moreover, the sum rates of all considered networks have similar convergence rates, which implies that the training of additional RIS agents and phase shift variables has a neglectable impact on the overall convergence rate.

6.4.2 Sum Rate versus Number of D2D Groups

Fig. 6.5 evaluates the sum rate performance of the PD-NOMA-enhanced multi-RIS aided D2D networks. The number of sub-channels is $K = 2$ in both PD-NOMA-based and OMA-based networks. It can be observed that, as the number of D2D groups increases from $D = 2$ to $D = 6$, the PD-NOMA-based networks demonstrate increasing sum rate gains compared to the OMA-based networks, by serving both DRs in each D2D group with the same sub-channel to improve resource efficiency. Moreover, by intelligently con-

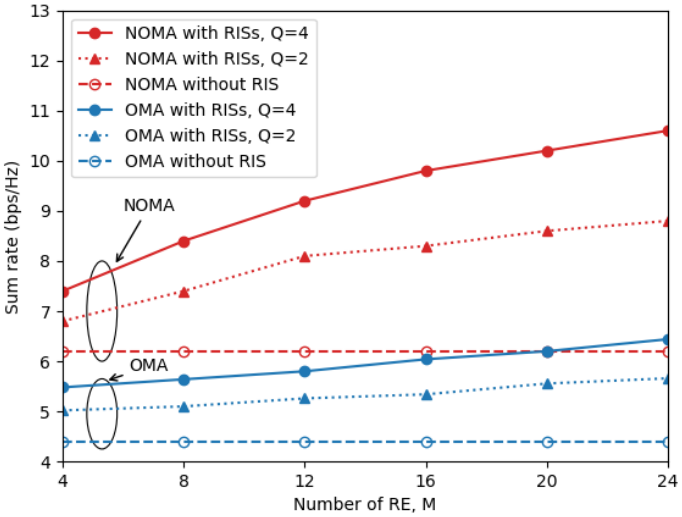


Figure 6.6: Sum rate versus the number of REs under PD-NOMA-based D2D networks and OMA-based D2D networks, with 2 RISs or no RIS. Each RIS consists of $Q = 2$ or $Q = 4$ sub-surfaces.

figuring the incident signals to reduce the multi-user interference, while enhancing the signal strength, the implementations of RISs introduce substantial sum rate improvements in both PD-NOMA-based and OMA-based networks. Nonetheless, by comparing the sum rates of PD-NOMA-based networks without RIS and that of the OMA-based networks with RIS, it can be concluded that PD-NOMA demonstrates a more significant sum rate enhancement than RISs when employed in the considered D2D networks.

6.4.3 Sum Rate versus Number of REs

Fig. 6.6 depicts the influence of the number of REs on the transmission sum rate in both PD-NOMA-based and OMA-based networks, with or without the assistance of RISs. 2 RISs are employed in all RIS-aided networks, where the number of sub-surfaces on each RIS varies from $Q = 2$ to $Q = 4$. It can be observed that the sum rate improvements of RISs are more significant in PD-NOMA-based networks than in OMA-based networks, which verifies the benefits of integrating PD-NOMA with RISs. However, due to the limited number of sub-surfaces on each RIS, the performance increment gradually decreases as the number of REs increases. Hence, by increasing the number of sub-surfaces from

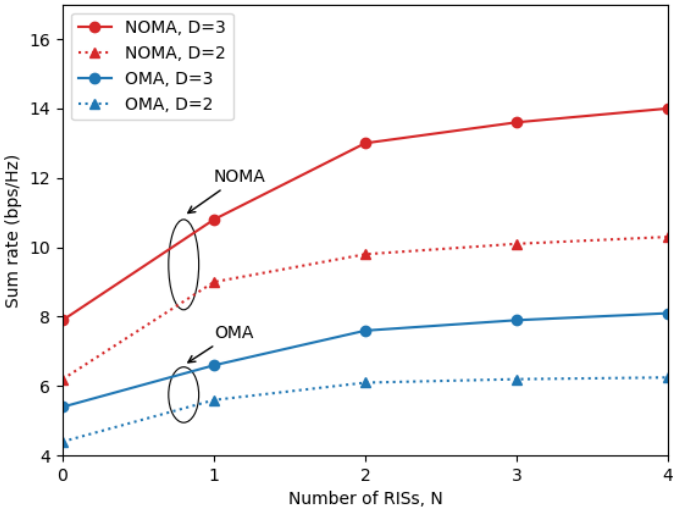


Figure 6.7: Sum rate versus the number of RISs under PD-NOMA-based D2D networks and OMA-based D2D networks, based on $D = 2$ or $D = 3$ D2D groups.

2 to 4, more degrees of freedom are introduced to the phase shift optimization, leading to significant sum rate improvements in all networks.

6.4.4 Sum Rate versus Number of RISs

Fig. 6.7 illustrates the impact of the number of RISs on sum rates based on a different number of D2D groups in PD-NOMA-based and OMA-based D2D networks. Specifically, $N = 0$ indicates a network without RIS. When the network consists of 2 D2D groups, the increase in sum rate becomes negligible after the number of RISs reaches 2. However, when the network consists of 3 D2D groups, the sum rate continues to increase as the number of RISs exceeds 2. Hence, it can be concluded that the deployment of multiple RISs is beneficial for enhancing the transmission sum rate, especially in large-scale D2D networks.

6.5 Summary

In this chapter, the resource allocation problem of PD-NOMA-enhanced multi-RIS aided D2D communication underlaying cellular networks is investigated. In particular, the sum rate maximization problem is formulated by jointly optimizing phase shifts, power allocations, and sub-channel assignments, under a time-varying channel model. To address this high-dimensional time-varying problem, a MAHA-DRL resource allocation framework is designed, following a CTDE structure. Moreover, MP-DQN networks are employed to directly manage the hybrid actions without any relaxation of the action space. Simulation results verify the convergence of the proposed algorithm under different learning rates and different network scenarios. Results also demonstrate the sum rate enhancement of PD-NOMA compared to OMA in D2D networks, as well as the performance gains of employing multiple RISs. In the next chapter, the summary and conclusion to the thesis and possible future work will be provided.

Chapter 7

Conclusions and Future Works

7.1 Contributions and Insights

This thesis concentrates on the designs of NOMA-enhanced wireless networks, with a particular focus on the implementation of AI-enabled optimization algorithms. The following three aspects are presented in this thesis: 1) The fundamental NOMA principles are studied, including the basic uplink and downlink NOMA transmission protocols and the mathematical formulations of NOMA systems; 2) Two key optimization problems of NOMA, namely user detection and resource allocation, are investigated with the assistance of AI technologies; 3) The potential integration of NOMA with the emerging technologies such as MIMO, RIS, and D2D communications. The main contributions and insights are summarized and outlined as follows.

In Chapter 3, the MUD problem in uplink grant-free NOMA systems was investigated. By exploiting the sporadic user activity patterns, CS theory was utilized to design the MUD algorithm in overloaded systems. A generative neural network was employed to learn the underlying relationships between the received signals and the transmitted signals. By assuming independent user behaviours, a carefully designed low-complexity neural network was proposed and meta-learning was employed in the training process

to improve the convergence rate in the application. Moreover, a sparsity estimator was provided to realize sparsity-blind MUD, which can be employed as an add-on technique to existing MUD algorithms. Simulation results verified the outstanding signal recovery accuracy and activity detection accuracy of the proposed MUD algorithm compared to several conventional iterative MUD techniques, under different noise levels and user sparsity levels. Furthermore, the proposed sparsity estimator demonstrated robust activity detection accuracy under different noise levels and showed a neglectable impact on MUD accuracy.

In Chapter 4, by exploiting the performance advantages of NOMA and OMA in different network scenarios, an adaptive NGMA framework was proposed. In particular, users were allocated to NOMA or OMA clusters by considering all users' channel information. The NOMA power allocation, user clustering, and beamforming were jointly optimized for maximizing the transmission sum rate, subject to a long-term total power constraint. To transform the mixed-integer problem, a spatial correlation-based clustering algorithm was proposed, where the user clustering can be determined based on a continuous-valued nominal angle threshold. Then, a DRL-based resource allocation scheme was designed, where the TRPO algorithm was employed to ensure training stability. As shown in the simulation results, the proposed TRPO algorithm demonstrated stable convergence under large learning rates, which indicates a fast and stable training process. Compared to NOMA and OMA systems, the proposed adaptive NGMA achieved significant sum rate gains under a different number of antennas, which indicates the limitations of pure NOMA or OMA systems in diverse wireless environments.

In Chapter 5, the integration of NOMA and RIS in multi-antenna networks was studied, where the optimization performance of DL and DRL was investigated and compared. The RIS was deployed to establish LoS links between the BS and the blocked users and NOMA is employed to serve multiple users with the same orthogonal resources simultaneously. To enhance resource efficiency, a QoS-based NOMA clustering strategy was proposed, where users with higher/lower QoS are defined as the strong/weak users.

The resource allocation problem in the considered network was investigated by designing a sum rate maximization problem, where RIS phase shifts, NOMA power allocations, and BS beamforming were jointly designed. To strike a thorough comparison between DL and DRL, the sum rate maximization problem was formulated with instantaneous transmit power constraints and long-term total power constraints, respectively. A meta-learning enabled DL algorithm and a DDPG-based DRL algorithm were proposed to solve the resource allocation problems. The extensive simulation results illustrated the sum rate improvement of NOMA systems compared to OMA systems, as well as the performance gains of RIS in both NOMA and OMA systems. Moreover, by comparing the results of DL and DRL, a similar sum rate performance was observed in the short-term problem, whereas DRL achieved a higher sum rate in the long-term due to the capability of maximizing long-term rewards. However, based on the complexity analysis, DL demonstrated lower algorithm complexity compared to DRL, which indicates that DL is more advantages in solving instantaneous problems and DRL is preferred for solving long-term problems at the cost of higher training complexity.

In Chapter 6, the integration of NOMA, RIS, and D2D communication underlaying cellular networks were investigated. In contrast to the conventional D2D pairs, the D2D transmitters in the proposed network can communicate with multiple D2D receivers through the same orthogonal resource by utilizing NOMA transmission. To further enhance the signal strength, multiple RISs were deployed to assist the transmissions of both D2D groups and the cellular networks. The sum rate maximization problem was designed by jointly optimizing RIS phase shifts, NOMA power allocations, and D2D sub-channel assignments. To address the high-dimensional mix-integer action space, the MP-DQN technique was integrated with the MADRL framework, where the DRs and the RIS controllers served as the agents. The training algorithm utilized a CTDE structure. Simulation results verified the convergence of the proposed MAMP-DQN algorithm under various learning rates and in different network scenarios. Results also illustrated that NOMA-enhanced D2D networks outperformed OMA-based D2D networks in terms of

sum rate. Also, the implementation of multiple RISs introduced significant sum rate improvement, especially in the NOMA-based networks, which indicates the merits of integrating NOMA with RIS in D2D networks.

7.2 Future Works

7.2.1 Extensions of Current Works

In this subsection, the potential extensions of the current works in this thesis are described in the following.

7.2.1.1 Imperfect CSI

The current works on NOMA in this thesis have relied on the perfect CSI assumption, which is difficult to realize in practical communication systems. Additional pilot signals are required to achieve an accurate channel estimation, which may reduce the spectral efficiency. In terms of MUD, the imperfect CSI may significantly degrade detection accuracy. Moreover, many NOMA protocols demand perfect CSI at the transmitters to determine the optimal resource allocation strategies, which may cause severe signalling overhead. However, the requirement of channel feedback can be relaxed in power-domain NOMA, since a few bits of feedback is sufficient for resource allocation tasks such as power allocation. Hence, it is important and necessary to study the impact of imperfect CSI on the performance of NOMA systems, as well as the designs of advanced NOMA systems with strong robustness to imperfect CSI.

7.2.1.2 Simultaneously Transmitting and Reflecting (STAR) RISs

In a RIS-aided network, both the transmitter and the receiver have to be on the same side of the RIS. This results in a half-space coverage of RIS and limits the flexibility of RIS deployment. To address this issue, the concept of STAR-RIS was proposed in [112]. STAR-RISs can realize a full-space coverage by simultaneously reflecting the signal on the same side of the RIS and transmitting the signal into the other side of the RIS. The

full space coverage is especially beneficial in NOMA-enhanced systems because users will have more diverse channel conditions. The differences in channel conditions can be exploited by NOMA to further enhance the spectral efficiency. Hence, the integration of STAR-RIS and NOMA is a valuable research extension to my existing works.

7.2.2 Promising Future Directions on AI-aided NOMA Systems

In this subsection, some promising future directions in terms of AI-aided NOMA systems are discussed.

7.2.2.1 End-to-End Channel Estimation and MUD Framework in NOMA Systems via DL

User activity detection and channel estimation for active users are often coupled, especially in grant-free NOMA systems. Hence, the research of NOMA-MUD is shifting towards a more general framework that jointly designs channel estimation and MUD [113, 114], where the user activity is carried out during channel estimation, followed by signal detection. By utilizing multiple neural networks, the joint channel estimation, user activity detection, and signal detection can be designed in an end-to-end structure. Owing to the superior performance of the DL-based channel estimation designs and the DL-based MUD designs, the end-to-end framework is expected to further improve the MUD accuracy as well as the channel estimation accuracy, hence is a valuable research topic.

7.2.2.2 DRL-enabled Age of Information (AoI) Optimization in NOMA Systems

The research contributions of NOMA in terms of resource allocation mostly aim to maximize the system's spectral efficiency or energy efficiency. Recently, a new metric, namely AoI, has received extensive research interest for characterising information timeliness and freshness in ubiquitous wireless networks. Moreover, as a spectral efficiency enhancement technique, NOMA has been envisioned as a promising technique to reduce

AoI [115]. Nonetheless, the optimization AoI is often modelled as a scheduling problem, in which DRL has been considered a favourable solution. Motivated by this, the implementation of DRL for AoI optimization in NOMA systems is another promising research direction.

Appendix A

Proof in Chapter 5

A.1 Proof of Proposition 1

For simplicity, the QoS subscript is removed and the ordered QoS requirements of K MUs is denoted as $R_1 \leq R_2 \leq \cdots \leq R_K$. For $K = 4$, the QoS requirements satisfies

$$\min((R_3 - R_1), (R_4 - R_2)) \geq \min((R_3 - R_2), (R_4 - R_1)), \tag{A.1.1}$$

and

$$\min((R_3 - R_1), (R_4 - R_2)) \geq \min((R_2 - R_1), (R_4 - R_3)). \tag{A.1.2}$$

Hence, intuitively, Proposition 1 is the optimal solution to (5.8) when $K = 4$. Then, proof by induction is employed to prove Proposition 1 for all even values of K.

It is further assumed that $\exists K' > 0$ where Proposition 1 is the optimal solution to (5.8) for $K = K'$. Then, if Proposition 1 can be proved to be optimal for $K = K' + 2$, it can be concluded, based on the principles of mathematical induction, that this is the optimal solution for all even and positive values of K. The ordered QoS requirements of

the $K'+2$ MUs is denoted as

$$R_1 \leq \cdots \leq R_{K'/2} \leq R_{n_1} \leq \cdots \leq R_{K'} \leq R_{n_2}. \tag{A.1.3}$$

The minimum QoS difference achieved by Proposition 1 is

$$D_{n1,n2} = \min\Big((R_{K'/2+1} - R_1), \cdots, (R_{K'} - R_{K'/2}), (R_{n_2} - R_{n_1})\Big). \tag{A.1.4}$$

Without loss of generality, an alternative clustering method that pairs MU n_2 with MU m, $m < n_1$, is considered. Based on the assumption made earlier, the optimal solution to cluster the rest K' MUs follows Proposition 1. Therefore, the minimum QoS achieved by this clustering method is

$$\begin{aligned} D_{m,n2} = \min\Big(&(R_{K'/2+1} - R_1), \cdots, (R_{K'/2+m-1} - R_{m-1}), \\ &(R_{K'/2+m} - R_{m+1}), \cdots, (R_{K'-1} - R_{K'/2}), (R_{K'} - R_{n_1}), (R_{n_2} - R_m)\Big). \end{aligned} \tag{A.1.5}$$

Two clustering schemes can be compared by computing the difference between $D_{n1,n2}$ and $D_{m,n2}$, given by

$$\begin{aligned} &D_{n1,n2} - D_{m,n2} = \\ &\quad \min\Big((R_{K'/2+m} - R_m), \cdots, (R_{K'-1} - R_{K'/2-1}), (R_{K'} - R_{K'/2}), (R_{n_2} - R_{n_1})\Big) \\ &\quad - \min\Big((R_{K'/2+m} - R_{m+1}), \cdots, (R_{K'-1} - R_{K'/2}), (R_{K'} - R_{n_1}), (R_{n_2} - R_m)\Big). \end{aligned} \tag{A.1.6}$$

Based on (A.1.3), $(R_{K'/2+m'} - R_{m'}) \geq (R_{K'/2+m'} - R_{m'+1})$ for all m'. Similarly, since $R_{K'/2} \leq R_{n_1}$ and $R_{n_2} > R_{K'}$, it can be derived that $D_{n1,n2} - D_{m,n2} > 0$. Hence, this alternative clustering method is suboptimal to Proposition 1. The similar proof

can be derived for all $m > n_1$. Therefore, Proposition 1 is the optimal solution when $K = M + 2$, and thus for all even and positive values of K.

References

[1] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020.

[2] P. Yang, Y. Xiao, M. Xiao, and S. Li, "6G wireless communications: Vision and potential techniques," *IEEE Network*, vol. 33, no. 4, pp. 70–75, Jul. 2019.

[3] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020.

[4] D. M. Novakovic and M. L. Dukic, "Evolution of the power control techniques for DS-CDMA toward 3G wireless communication systems," *IEEE Commun. Surv. Tut.*, vol. 3, no. 4, pp. 2–15, Oct. 2000.

[5] J. Li, X. Wu, and R. Laroia, *OFDMA mobile broadband communications: A systems approach.* Cambridge University Press, 2013.

[6] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sept. 2013, pp. 332–336.

[7] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, "Rate-splitting multiple access: A new frontier for the PHY layer of 6G," in *Proc. IEEE Veh. Technol. Conf. (VTC 2020)*, Nov. 2020, pp. 1–7.

[8] 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, Dec. 2017.

[9] Y. Liu, S. Zhang, X. Mu, Z. Ding, R. Schober, N. Al-Dhahir, E. Hossain, and X. Shen, "Evolution of NOMA toward next generation multiple access (NGMA) for 6G," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1037–1071, Apr. 2022.

[10] 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.

[11] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surv. Tut.*, vol. 21, no. 3, pp. 2224–2287, Mar. 2019.

[12] Y. Qian, J. Wu, R. Wang, F. Zhu, and W. Zhang, "Survey on reinforcement learning applications in communication networks," *J. Commun. Information Netw.*, vol. 4, no. 2, pp. 30–39, Jun. 2019.

[13] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, Mar. 2004.

[14] E. Crespo Marques, N. Maciel, L. Naviner, H. Cai, and J. Yang, "A review of sparse recovery algorithms," *IEEE Access*, vol. 7, pp. 1300–1322, Dec. 2018.

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

[16] S. Zeb, M. A. Rathore, A. Mahmood, S. A. Hassan, J. Kim, and M. Gidlund, "Edge intelligence in softwarized 6G: Deep learning-enabled network traffic predictions," in *IEEE Global Commun. Conf. (GLOBECOM) Workshops 2021)*, Dec. 2021, pp. 1–6.

[17] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li, "Sparse representation for wireless communications: A compressive sensing approach," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 40–58, May 2018.

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

[19] H. V. Cheng, E. Björnson, and E. G. Larsson, "Performance analysis of NOMA in training-based multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 372–385, Jan. 2018.

[20] Y. Liu, X. Mu, X. Liu, M. Di Renzo, Z. Ding, and R. Schober, "Reconfigurable intelligent surface (RIS) aided multi-user networks: Interplay between NOMA and RIS," *arXiv:2011.13336*, Nov. 2020.

[21] Q. Luo, Z. Liu, G. Chen, Y. Ma, and P. Xiao, "A novel multitask learning empowered codebook design for downlink SCMA networks," *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1268–1272, Mar. 2022.

[22] Q. Luo, H. Wen, G. Chen, Z. Liu, P. Xiao, Y. Ma, and A. Maaref, "A novel non-coherent SCMA with massive MIMO," *IEEE Wireless Commun. Lett.*, pp. 1–1, Aug. 2022.

[23] B. Di, L. Song, Y. Li, and G. Y. Li, "TCM-NOMA: Joint multi-user codeword design and detection in trellis-coded modulation-based NOMA for beyond 5G," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 766–780, Jun. 2019.

[24] Z. Liu and L.-L. Yang, "Sparse or dense: A comparative study of code-domain NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 4768–4780, Aug. 2021.

[25] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, p. 2320–2323, Nov. 2016.

[26] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640–643, Mar. 2017.

[27] Y. Du, C. Cheng, B. Dong, Z. Chen, X. Wang, J. Fang, and S. Li, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894–7909, Oct. 2018.

[28] A. C. Cirik, N. Mysore Balasubramanya, and L. Lampe, "Multi-user detection using ADMM-based compressive sensing for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 46–49, Sept. 2018.

[29] 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, Jun. 2018.

[30] M. Baghani, S. Parsaeefard, M. Derakhshani, and W. Saad, "Dynamic non-orthogonal multiple access and orthogonal multiple access in 5G wireless networks," *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 6360–6373, Sept. 2019.

[31] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.

[32] R. Duan, J. Wang, C. Jiang, H. Yao, Y. Ren, and Y. Qian, "Resource allocation for multi-UAV aided IoT NOMA uplink transmission systems," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7025–7037, Apr. 2019.

[33] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, Nov. 2019.

[34] Y. Feng, S. Yan, Z. Yang, N. Yang, and J. Yuan, "Beamforming design and power allocation for secure transmission with NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2639–2651, May 2019.

[35] Y. Bai, B. Ai, and W. Chen, "Deep learning based fast multiuser detection for massive machine-type communication," in *Proc. IEEE Veh. Technol. Conf. (VTC 2019)*, Nov. 2019, pp. 1–5.

[36] H. He, C. Wen, S. Jin, and G. Y. Li, "Model-driven deep learning for MIMO detection," *IEEE Trans. on Signal Process.*, vol. 68, pp. 1702–1715, Feb. 2020.

[37] Y. Ge and J. Fan, "Beamforming optimization for intelligent reflecting surface assisted MISO: A deep transfer learning approach," *IEEE Trans. Veh. Technol.*, pp. 1–1, Mar. 2021.

[38] C. Huang, G. C. Alexandropoulos, C. Yuen, and M. Debbah, "Indoor signal focusing with deep learning designed reconfigurable intelligent surfaces," in *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[39] B. Sheen, J. Yang, X. Feng, and M. M. U. Chowdhury, "A deep learning based modeling of reconfigurable intelligent surface assisted wireless communications for phase shift configuration," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 262–272, Jan. 2021.

[40] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces," in *IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[41] L. T. Tan, R. Q. Hu, and L. Hanzo, "Twin-timescale artificial intelligence aided mobility-aware edge caching and computing in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3086–3099, Apr. 2019.

[42] R. Li, C. Wang, Z. Zhao, R. Guo, and H. Zhang, "The LSTM-based advantage actor-critic learning for resource management in network slicing with user mobility," *IEEE Commun. Lett.*, vol. 24, no. 9, pp. 2005–2009, Sept. 2020.

[43] Y. Yu, T. Wang, and S. C. Liew, "Deep-reinforcement learning multiple access for heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1277–1290, Jun. 2019.

[44] N. Ye, J. An, and J. Yu, "Deep-learning-enhanced NOMA transceiver design for massive MTC: Challenges, state of the art, and future directions," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 66–73, Aug. 2021.

[45] W. Kim, Y. Ahn, and B. Shim, "Deep neural network-based active user detection for grant-free NOMA systems," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2143–2155, Apr. 2020.

[46] S.-M. Tseng, Y.-F. Chen, C.-S. Tsai, and W.-D. Tsai, "Deep-learning-aided cross-layer resource allocation of OFDMA/NOMA video communication systems," *IEEE Access*, vol. 7, pp. 157 730–157 740, Oct. 2019.

[47] H. Zhang, H. Zhang, K. Long, and G. K. Karagiannidis, "Deep learning based radio resource management in NOMA networks: User association, subchannel and power allocation," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2406–2415, Oct. 2020.

[48] T.-H. Vu, T.-V. Nguyen, D. B. da Costa, and S. Kim, "Performance analysis and deep learning design of underlay cognitive NOMA-based CDRT networks with imperfect SIC and co-channel interference," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8159–8174, Dec. 2021.

[49] M. Fayaz, W. Yi, Y. Liu, and A. Nallanathan, "Transmit power pool design for

grant-free NOMA-IoT networks via deep reinforcement learning," *IEEE Trans Wireless Commun.*, vol. 20, no. 11, pp. 7626–7641, Nov. 2021.

[50] Z. Yang, Y. Liu, Y. Chen, and J. T. Zhou, "Deep reinforcement learning for RIS-aided non-orthogonal multiple access downlink networks," in *IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[51] C. Huang, G. Chen, Y. Gong, P. Xu, Z. Han, and J. A. Chambers, "Buffer-aided relay selection for cooperative hybrid NOMA/OMA networks with asynchronous deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2514–2525, Aug. 2021.

[52] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, "Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6369–6379, Jul. 2020.

[53] K. N. Doan, M. Vaezi, W. Shin, H. V. Poor, H. Shin, and T. Q. S. Quek, "Power allocation in cache-aided NOMA systems: Optimization and deep reinforcement learning approaches," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 630–644, Jan. 2020.

[54] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for noma with deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019.

[55] S. Vanka, S. Srinivasa, Z. Gong, P. Vizi, K. Stamatiou, and M. Haenggi, "Superposition coding strategies: Design and experimental evaluation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 7, pp. 2628–2639, Jul. 2012.

[56] P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inf. Theory*, vol. 19, no. 2, pp. 197–207, Mar. 1973.

[57] A. J. Grant, B. Rimoldi, R. L. Urbanke, and P. A. Whiting, "Rate-splitting multiple access for discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 873–890, Mar. 2001.

[58] A. Carleial, "Interference channels," *IEEE Trans. Inf. Theory*, vol. 24, no. 1, pp. 60–70, Jan 1978.

[59] T. Cover and A. E. Gamal, "Capacity theorems for the relay channel," *IEEE*

Trans. Inf. Theory, vol. 25, no. 5, pp. 572–584, Sep. 1979.

[60] I. Csiszar and J. Korner, "Broadcast channels with confidential messages," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 339–348, May 1978.

[61] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge Univ. Press, 2005.

[62] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tut.*, vol. 20, no. 3, pp. 2294–2323, May 2018.

[63] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li, "Sparse representation for wireless communications: A compressive sensing approach," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 40–58, May 2018.

[64] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

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

[66] L. Jiao and J. Zhao, "A survey on the new generation of deep learning in image processing," *IEEE Access*, vol. 7, pp. 172 231–172 263, Nov. 2019.

[67] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[68] J. Zhai, S. Zhang, J. Chen, and Q. He, "Autoencoder and its various variants," in *Proc. IEEE Int. Conf. Systems, Man, and Cybern. (SMC)*, Oct. 2018, pp. 415–419.

[69] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019.

[70] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *Proc. Int. Conf. Mach. Learn. (ICML 2017)*, pp.

1126–1135, Jul. 2017.

[71] M. Costa, "Writing on dirty paper (corresp.)," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[72] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

[73] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[74] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1478–1491, Sept. 2007.

[75] Q. Yang, H.-M. Wang, D. W. K. Ng, and M. H. Lee, "NOMA in downlink SDMA with limited feedback: Performance analysis and optimization," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2281–2294, Oct. 2017.

[76] Z. Ding, "NOMA beamforming in SDMA networks: Riding on existing beams or forming new ones?" *IEEE Commun. Lett.*, vol. 26, no. 4, pp. 868–871, Apr. 2022.

[77] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tut.*, vol. 23, no. 3, pp. 1546–1577, May 2021.

[78] 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, Jul. 2020.

[79] Y. Wu, M. Rosca, and T. Lillicrap, "Deep compressed sensing," *Proc. Int. Conf. Mach. Learn. (ICML 2019)*, vol. 97, pp. 6850–6860, May 2019.

[80] Y. Yu, Z. Gong, P. Zhong, and J. Shan, "Unsupervised representation learning with deep convolutional neural network for remote sensing images," in *Proc. Int. Conf. Image Graph.* Springer, Sept. 2017, pp. 97–108.

[81] N. Y. Yu, "Multiuser activity and data detection via sparsity-blind greedy recovery

for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 2082–2085, Aug. 2019.

[82] S. Chen and D. Donoho, "Basis pursuit," in *Proc. Asilomar Conf. Signals Syst. Comput.*, vol. 1, Oct. 1994, pp. 41–44.

[83] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, p. 1094–1121, Feb. 2012.

[84] B. Bailey and M. Telgarsky, "Size-noise tradeoffs in generative networks," in *Proc. Int. Conf. NeurIPS*, Dec. 2018, pp. 6490–6500.

[85] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, Apr. 2009.

[86] Ö. Özdogan, E. Björnson, and E. G. Larsson, "Massive MIMO with spatially correlated rician fading channels," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3234–3250, May 2019.

[87] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-s. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tut.*, vol. 19, no. 2, pp. 721–742, Oct. 2016.

[88] J. Wang, B. Liang, M. Dong, G. Boudreau, and H. Abou-Zeid, "Online distributed coordinated precoding for virtualized MIMO networks with delayed CSI," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 1012–1016, May 2022.

[89] X. Deng, J. Li, L. Shi, Z. Wang, J. H. Wang, and T. Wang, "On dynamic resource allocation for blockchain assisted federated learning over wireless channels," *Proc. 14th IEEE Int. Conf. Internet Things*, pp. 306–313, Dec. 2021.

[90] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6174–6189, Aug. 2016.

[91] Z. Sun and Y. Jing, "Transmission and clustering designs for multi-antenna NOMA based on average transmit power," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3412–3427, Apr. 2021.

[92] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy

optimization," *Proc. Int. Conf. Mach. Learn. (ICML 2015)*, pp. 1889–1897, Jun. 2015.

[93] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," *Int. Conf. Mach. Learn. (ICML)*, pp. 1329–1338, 2016.

[94] Y. Fu, M. Zhang, L. Salaün, C. W. Sung, and C. S. Chen, "Zero-forcing oriented power minimization for multi-cell MISO-NOMA systems: A joint user grouping, beamforming, and power control perspective," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1925–1940, Aug. 2020.

[95] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *arXiv:2004.05439*, Apr. 2020.

[96] A. Fallah, A. Mokhtari, and A. Ozdaglar, "On the convergence theory of gradient-based model-agnostic meta-learning algorithms," in *Proc. Int. Conf. Artif. Intell. Statist.* PMLR, Jun. 2020, pp. 1082–1092.

[97] K. Ji, J. Yang, and Y. Liang, "Theoretical convergence of multi-step model-agnostic meta-learning," *arXiv:2002.07836*, Feb. 2020.

[98] Y. Wu, J. Donahue, D. Balduzzi, K. Simonyan, and T. Lillicrap, "LOGAN: Latent optimisation for generative adversarial networks," *arXiv:1912.00953*, Dec. 2019.

[99] 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:1509.02971*, Sept. 2015.

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

[101] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[102] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn. (ICML*

2014). PMLR, Jul. 2014, pp. 387–395.

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

[104] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook Reinforcement Learn. Control*, pp. 321–384, Jun. 2021.

[105] 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, Aug. 2019.

[106] A. Hussein, E. Elyan, and C. Jayne, "Deep imitation learning with memory for robocup soccer simulation," in *Proc. Int. Conf. Eng. Appl. Neural Netw.* Springer, Jul. 2018, pp. 31–43.

[107] M. Khamassi, G. Velentzas, T. Tsitsimis, and C. Tzafestas, "Active exploration and parameterized reinforcement learning applied to a simulated human-robot interaction task," in *Proc. IEEE Int. Conf. Robot. Comput. (IRC)*, Apr. 2017, pp. 28–35.

[108] M. Hausknecht and P. Stone, "Deep reinforcement learning in parameterized action space," *arXiv:1511.04143*, Nov. 2015.

[109] J. Xiong, Q. Wang, Z. Yang, P. Sun, L. Han, Y. Zheng, H. Fu, T. Zhang, J. Liu, and H. Liu, "Parametrized deep Q-networks learning: Reinforcement learning with discrete-continuous hybrid action space," *arXiv:1810.06394*, Oct. 2018.

[110] C. J. Bester, S. D. James, and G. D. Konidaris, "Multi-pass Q-networks for deep reinforcement learning with parameterised action spaces," *arXiv:1905.04388*, May 2019.

[111] R. Zhong, X. Liu, Y. Liu, and Y. Chen, "Multi-agent reinforcement learning in NOMA-aided UAV networks for cellular offloading," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1498–1512, Mar. 2022.

[112] Y. Liu, X. Mu, J. Xu, R. Schober, Y. Hao, H. V. Poor, and L. Hanzo, "STAR: Simultaneous transmission and reflection for 360^o coverage by intelligent surfaces,"

IEEE Wireless Commun., vol. 28, no. 6, pp. 102–109, Dec. 2021.

[113] Y. Zhang, Q. Guo, Z. Wang, J. Xi, and N. Wu, "Block sparse bayesian learning based joint user activity detection and channel estimation for grant-free NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9631–9640, Oct. 2018.

[114] Y. Zhang, Z. Yuan, Q. Guo, Z. Wang, J. Xi, and Y. Li, "Bayesian receiver design for grant-free NOMA with message passing based structured signal estimation," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8643–8656, Aug. 2020.

[115] Q. Wang, H. Chen, C. Zhao, Y. Li, P. Popovski, and B. Vucetic, "Optimizing information freshness via multiuser scheduling with adaptive NOMA/OMA," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1766–1778, Mar. 2022.