

**INTERFERENCE MITIGATION AND ENERGY EFFICIENT RESOURCE
ALLOCATION SCHEME FOR D2D COMMUNICATION
IN 5G AND BEYOND**

A Thesis submitted in fulfillment of the requirement for the award of the

degree of

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by:

Vineet Vishnoi

(Registration No. : E19SOE825)

Under the guidance of:

Dr. Ishan Budhiraja

(Associate Professor, SCSET)

Dr. Suneet Gupta

(Associate Professor, SCSET)



BENNETT UNIVERSITY, GREATER NOIDA – 201310

August 2023

CERTIFICATE

I, Vineet Vishnoi, Enrollment. No. E19SOE825, hereby declare that the thesis entitled **“Interference Mitigation and Energy Efficient Resource Allocation Scheme For D2D Communication in 5G and Beyond”** submitted to the School of Computer Science Engineering and Technology, Bennett University, Greater Noida, Uttar Pradesh, India is an authenticated record of my own work for the award of the degree of "Doctor of Philosophy". I pursued this research work under the guidance of Dr. Ishan Budhiraja (Assistant Professor) and Dr. Suneet Gupta (Associate Professor). This report has not been presented to any other institution for the purpose of obtaining another degree.



Vineet Vishnoi

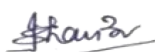
Place: Greater Noida, Uttar Pradesh (India)

Enrollment No. E19SOE825

Date: 21 August, 2023

This is to certify that the aforementioned declaration provided by the candidate is accurate to the utmost of our knowledge.

Verified by:



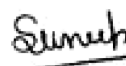
Dr. Ishan Budhiraja

(Supervisor)

Assistant Professor, SCSET

Bennett University

Greater Noida, Uttar Pradesh (India)



Dr. Suneet Gupta

(Supervisor)

Associate Professor, SCSET

Bennett University

Greater Noida, Uttar Pradesh (India)

ABSTRACT

The proliferation of mobile users, smart gadgets, and multimedia applications generates an unprecedented growth of data traffic in 5G and beyond networks. This massive increase in data traffic puts a lot of burden on efficient spectrum utilization in the years to come. To address this problem, researchers recommended various D2D communication (D2D-C) techniques. In D2D-C technology, two neighboring devices can share the data directly without the base station (BS). As a result, it enhances mobile users' quality of service by reducing the transmission delay. Also, in D2D-C, the D2D pairs (DDPs) reuse the same resources as used by cellular users (CUs) to boost the spectral efficiency. Despite these advantages, key challenges such as cross-channel (CR-CI) and co-channel (CO-CI), as well as ultra-massive connectivity (UMC), need to be investigated. To overcome these challenges, academic and industry researchers suggested the non-orthogonal multiple access (NOMA) approach.

NOMA is a scheme in which more than one user shares the same spectrum resources at any instant but with different power levels, resulting in considerable improvements in the spectral and UMC. Despite these benefits, NOMA brings additional challenges of intra-user interference, and fairness. The receiver implements the SIC technique to reduce the effects of intra-user interference. However, it is necessary to explore new techniques to improve fairness in dynamic environments. To address this challenge, researchers used the deep reinforcement learning (DRL).

DRL proves to be a highly proficient method for optimizing embedded systems, endowed with the remarkable ability to promptly respond in wireless communication networks (WCNs). In the realm of DRL approaches, it is customary to prepare neural networks (NNs) via offline training prior to their implementation on terminal devices or controllers. It employs the trained

model to estimate the most optimized transmission power management while keeping a low level of computational complexity with regards to resource allocation.

Energy represents yet another significant challenge in D2D -C given the restricted capacity for energy storage within a battery. Several researchers have proposed various schemes for increasing EE in the UAV-RIS integrated network. In order to create an optimal transmission setting for both the UAV and D2D users, we opted for the implementation of reconfigurable intelligent surfaces (RIS). This choice was made due to its ability to operate without the use of energy-intensive sources, resulting in a notable increase in energy efficiency.

In this study, we have put forth the two approaches to resolve the previously mentioned concerns.

Firstly, a DRL Scheme for Sum Rate and Fairness Maximization Among D2D Pairs Underlying Cellular Network With NOMA is proposed. The objective of this proposal is to enhance the overall network's sum rate and fairness among NOMA-enabled CUs and DDPs while considering the resource allocation, power allocation, and QoS demands of CUs and DDPs. To attain the objective, we have formulated the problem as MINLP, taking into account the resources and power constraints of both BS and DDPs. Our ultimate aim is to optimize the sum rate and fairness amongst the NOMA enabled CUs and DDPs. We first used the centralized deep deterministic policy gradient (DDPG) and arithmetic-geometric mean approximation (AGMA) technique to reduce cross-channel interference (CR-CI) and control the power. Then, to provide fairness to all the users, we transformed the proposed solution into a distributed deep deterministic policy gradient (D3PG). Also, the successive convex approximation technique is then integrated into the D3PG to mitigate the impact of CO-CI among DDPs.

The second approach is DRL Based Energy Consumption Minimization for Intelligent Reflecting Surfaces (RIS) Assisted D2D Users Underlying UAV Network. In this investigation, we delve into the communication system assisted by up-link RIS, while taking into account the presence of D2DPs. RIS is introduced into WCNs with the assistance of an unmanned aerial vehicle (UAV) in order to solve the issues of CO-CI and CR-CI. A method known as the centralized declining deep-Q network (C-DDQN) has been proposed to accurately estimate the UAV's trajectory and RIS's phase shift, while simultaneously meeting the data demands of

of DDUs and CUs. In the C-DDQN method, a central controller acts as an agent. The central controller governs both the trajectory of the UAV and the phase shift of the RIS with precision.

ACKNOWLEDGMENT

Before delving into my Ph.D. journey, I must express my deepest gratitude to the almighty God who bestowed upon me the fortitude and bravery to conquer all the hindrances and successfully accomplish this pursuit. The realization of my lifelong goal to earn the title of ‘Doctor’ has come to fruition as I have been granted admission to the Doctorate of Philosophy program at Bennett University in Greater Noida, Uttar Pradesh. Without recognizing the individuals who have provided me with unwavering support during this expedition, this undertaking would remain unfinished. I am aware that words alone are insufficient to convey the depth of my appreciation; nevertheless, I offer them as a token of my heartfelt regards.

First and foremost, I wish to extend my utmost gratitude to my parents. It is with their unwavering approval, encouragement, and inspiration that I have decided to embark on this monumental endeavor in my existence. Secondly, I express my profound gratitude to my supervisors, Dr. Ishan Budhiraja (Assistant professor, SCET) and Dr. Suneet Gupta (Associate Professor, SCET), for their unwavering support and guidance throughout my PhD journey. Their remarkable patience and wealth of knowledge have been invaluable to me, and I am deeply grateful for the space they have given me to work on my own terms. Throughout this journey, they not only provided me with exceptional supervision but also actively cooperated with me and constantly encouraged me.

I am immensely thankful to Prof. (Dr.) Abhay Bansal (Dean-SCET), Dr. Rama Komaragiri (Dean-R & C), Dr. Akansha Singh (Ph.D. Coordinator), and the esteemed members of my doctoral committee, Dr. Aditya Bhardwaj and Dr. Ajay Yadav, for their invaluable feedback and unwavering support in maintaining the perfect momentum of my work. I am immensely grateful to the Vice Chancellor, Dr. Prabhu Kumar Aggarwal, Pro Vice Chancellor Prof. (Dr.) Ajith Abraham, and the astute management of Bennett University. Their unwavering support and provision of all the essential resources and facilities were instrumental in enabling me to

successfully complete my PhD work.

The expression of my gratitude would certainly be inadequate if I neglected to acknowledge my entire family, including my esteemed father Shri. Rajendra Vishnoi, my esteemed mother Mrs. Sudha Vishnoi, my beloved wife Mrs. Shweta Vishnoi, and my adorable children Arpan Vishnoi and Aakriti Vishnoi, for their unwavering love, assistance, and motivation during every stage of my life. I would like to extend my sincere gratitude to my father-in-law, Shri Rajkumar Vishnoi, as well as my mother-in-law, Mrs. Sareeta Vishnoi, for their invaluable guidance and unwavering financial support throughout this journey.

I would like to sincerely extend my gratitude to Prof. Deepak Garg (Vice Chancellor) of SRU University, Warangal, Prof. Neeraj Kumar (Dean DCT) of Thapar University, Patiala, Dr. Rajat Chaudhary (Assistant Professor) and Dr. Praveen Malik of LPU, Ludhiana for their invaluable guidance and unwavering support throughout my journey. I would like to extend my heartfelt appreciation towards all my beloved relatives and cousins for their unwavering inspiration and encouragement. They provided me with words of encouragement that made this journey much more comfortable and aided me in completing my research work.

I would also like to express my gratitude towards my dear friends and esteemed colleagues who have accompanied me throughout this remarkable journey of conducting research. I would like to extend a heartfelt appreciation to my esteemed research group members, namely Mr. Prakhar Consul, Mr. Neeraj Joshi, Mr. Shivam Chaudhary, and Ms. Deepanshi. These individuals have truly made my research journey more remarkable and enjoyable. As it is not feasible to enumerate all the names of my well-wishers, friends, and loved ones, I would like to express my heartfelt gratitude to each and every one who extended their support, be it direct or indirect, during my pursuit of knowledge.

(Vineet Vishnoi)

List of Publications

Journal Publications (SCIE and Conference):

1. Vineet Vishnoi, Ishan Budhiraja, Suneet Gupta, "Sum Rate and Fairness Maximization Scheme for D2D Pairs Underlying Cellular Network with NOMA Using Deep Reinforcement Learning", *IEEE Transactions on Vehicular Technology*, pp. 1-17, May. 2023. (IF 6.8-Q1)
2. Vineet Vishnoi, Ishan Budhiraja, Suneet Gupta, "Deep Reinforcement Learning Based Energy Consumption Minimization for Intelligent Reflecting Surfaces Assisted D2D Users Underlying UAV Network", in *IEEE International Conference on Computer Communications (INFOCOM)*, New York, USA., April 2023 [Accepted, A* Conference, Indexed in IEEE-Xplore & SCOPUS].
3. Vineet Vishnoi, Ishan Budhiraja, Suneet Gupta, "Deep Reinforcement Learning based Energy Efficiency Maximization Scheme for Uplink NOMA enabled D2D Users", in *IEEE Global Communications Conference (Globcom)*, Kuala Lumpur, Malaysia, August 2023 [Accepted, Indexed in IEEE-Xplore & SCOPUS].
4. Vineet Vishnoi, Ishan Budhiraja, et al., "Energy Efficient Optimization Scheme for RIS-Assisted Communication Underlying UAV with NOMA", in *IEEE International Conference on Communication (ICC)*, pp. 1-6 Seoul, South Korea, August 2022.

Contents

Certificate	i
Abstract	ii
Acknowledgment	v
List of Publications	vii
List of Figures	xv
List of Tables	xvii
List of Abbreviations	xviii
1 Introduction	1
1.1 D2D Communication	2
1.1.1 Potential Benefits of D2D-C	2
1.1.2 Use Cases of D2D	3
1.1.2.1 Local Services	3
1.1.2.2 Emergency Communication:	3
1.1.2.3 Enhancing the Internet of Things (IoT)	3
1.1.2.4 D2D in Multiuser-Multiple Input Multiple Output (MU-MIMO):	4
1.2 Deep Reinforcement Learning (DRL)	4
1.3 Motivation	6
1.4 Thesis Organization	7

1.5	Summary	9
2	Literature Review	10
2.1	DRL OVERVIEW	13
2.1.1	DRL Fundamental Building Blocks	13
2.1.1.1	Fundamental of RL	13
2.1.1.2	Fundamental of DL	16
2.2	Classification of DRL Algorithm	19
2.2.1	Basic DRL Classifications	19
2.2.1.1	VB DRL Algorithms	20
2.2.2	PG Based DRL Algorithm	24
2.2.2.1	Stochastic Policy Gradients (SPG) Algorithm	26
2.2.2.2	Deterministic Policy Gradients (DPG) Algorithm	27
2.2.2.3	Monte Carlo Policy Gradient(MCPG) Algorithm	27
2.2.2.4	Algorithm Based on Natural Policy Gradient (NPG)	29
2.2.3	DRL Advanced Techniques	30
2.2.3.1	Actor-Critic Technique(ACT)	30
2.2.3.2	Hierarchical Reinforcement Learning (H-RL)	32
2.2.3.3	Multi-Agent DRL (MA-DRL) Algorithm	33
2.3	Deep Reinforcement Learning (DRL)	35
2.3.1	DRL with D2D-C	35
2.3.1.1	FC-MAQ (Fuzzy clustering multi-agent Q-learning)	35
2.3.1.2	S-DDPG (Sharing-deep deterministic policy gradient)	36
2.3.1.3	MAAC-NAAC DRL (Multi-agent actor critic, neighbor-agent actor critic-DRL)	36
2.3.1.4	ICWN-DRL (Information-centric wireless networking-DRL)	36
2.3.1.5	NC-DRL (Non cooperative-DRL)	37
2.3.1.6	ST-DRL (Social Trust-DRL)	37
2.3.1.7	C-MADRL-PER (Coordinated-multi-agent DRL-prioritized experience replay)	37

2.3.1.8	CSSCA-DRL (Constrained stochastic successive convex approximation-DRL)	38
2.3.1.9	TP-DQN (Two parallel-deep Q networks)	38
2.3.1.10	SRPR-DRL (Successive rounding and power refinement-DRL)	39
2.3.1.11	MSRA-DDPG (Mode selection and resource allocation DDPG)	39
2.3.1.12	D-MARL (Deep-multi-agent RL)	39
2.3.1.13	ADMM-DRL (Alternating direction method of multipliers based DRL)	40
2.3.1.14	C-DRL (Centralized-DRL)	40
2.3.1.15	SOTPSR-DRL (Self-organization of transmission power and power splitting-DRL)	40
2.3.1.16	BDRFL (Double-layer blockchain-based deep reinforcement federated learning)	41
2.3.1.17	D4SA (Double deep Q-network based D2D spectrum access)	41
2.3.1.18	AW-FDRL (Attention weighted federated -DRL)	42
2.3.1.19	CO-DDDPG (Conventional optimization scheme with DDPG)	42
2.3.1.20	SGMA-DRL (Stackelberg game guided multi-agent DRL) . .	42
2.3.1.21	MSPA-DRL (Mode selection integrated with power allocation-DRL)	43
2.3.1.22	D3QN-UARA (Double-dueling-deep Qnetwork based joint user association and resource allocation)	43
2.3.1.23	MAOD-DRL (Multi-agent online distributed DRL)	43
2.3.1.24	DSM-DRL (Dynamic spectrum matching-DRL)	44
2.3.1.25	PS-D3QN-DRL (Priority sampling based dueling double deep Q-network-DRL)	44
2.3.1.26	FD-DRA-DRL (Federated learning aided decentralized resource allocation-DRL)	45
2.3.1.27	DSA-DRL-ToT (Dynamic spectrum access- DRL-Internet of Things)	45

2.3.1.28	UAVs-SWIPT-MADQN-D2D	45
2.3.1.29	DTD3-D2D (Dinkelbach combined twin delayed deep deterministic policy gradient)	46
2.3.1.30	SACC-D2D (Social-aware cooperative caching)	46
2.3.2	DRL with RIS	49
2.3.2.1	DLR-DQN (Decaying learning rate deep Q-network)	49
2.3.2.2	WL-DDPG (Water filling-deep deterministic policy gradient)	49
2.3.2.3	EA-DDPG (Exploration attenuate-deep deterministic policy gradient)	50
2.3.2.4	PSD-DRL (Phase shift design-DRL)	50
2.3.2.5	MC-PD-NOMA-DRL (Multi carrier power domain non-orthogonal multiple access-DRL)	50
2.3.2.6	ERAP-DRL (Efficient resource allocation parallel-DRL)	51
2.3.2.7	UCA-HSE-DRL (UAV collision avoidance-high sample efficient-DRL)	51
2.3.2.8	DRPO-PGAC (Decomposition and relaxation-based precoding optimization-policy gradient Actor- Critic)	52
2.3.2.9	SAC-AO-RIS (Soft actor-critic-alternating optimization-RIS)	52
2.3.2.10	WPT-DDPG (Wireless power transfer-DDPG)	52
2.3.2.11	PSD-TD3-RIS (Primal-dual sub gradient descent-twin-delayed DDPG-RIS)	53
2.3.2.12	MAML-DDPG (Model-agnostic-meta-learning-DDPG)	53
2.3.2.13	EH-R-DRL (Energy harvesting-robust-DRL)	53
2.3.2.14	ARSO-DDPG (Active RIS subarray optimization scheme based on deep deterministic policy gradient)	55
2.3.2.15	IS-UAV-TN-MO-DDPG (Integrated satellite-unmanned aerial vehicle-terrestrial network-multi objective-DDPG)	56
2.3.2.16	MU-MISO-PW-DRL (Multi-user-multiple input single output-piecewise-DRL)	56

2.3.2.17	DDPG-TD3 (DDPG integrated twin delayed deep deterministic)	56
2.3.2.18	TTD3-DRL (Twin twin-delayed deep deterministic policy gradient based DRL)	57
2.3.2.19	SCA-AC-PPO (Successive convex approximation-actor critic-proximal policy optimization)	57
2.3.2.20	RAFD-6GV2X-LCPPO (RIS assisted full duplex-6G vehicle to everything-low complexity proximal policy optimization scheme)	57
2.4	Summary	58
2.5	Research Gaps	58
2.6	Objectives	61
2.7	Methodology	61
2.8	Methodology for Objective 1	61
2.9	Methodology for Objective 2	63
3	A Deep Reinforcement Learning Scheme for Sum Rate and Fairness Maximization Among D2D Pairs Underlying Cellular Network With NOMA	64
3.1	Network Model	65
3.1.1	Network Architecture	65
3.1.2	Channel Model	68
3.1.2.1	CU Channel Model	68
3.1.2.2	D2D Channel Model	70
3.1.3	Network Sum Rate Calculation	71
3.1.4	Fairness Utility Function	71
3.1.5	Total Fairness Utility Function Calculation	73
3.1.6	Problem Formulation	73
3.2	PRELIMINARIES OF DRL	74
3.2.1	Value Function (VF)	75
3.2.2	Policy Search (PS)	76
3.2.3	Actor Method	77

3.2.4	Critic Method	77
3.3	Proposed Scheme	77
3.3.1	MDP	78
3.3.1.1	Agent	78
3.3.1.2	Network Space	78
3.3.1.3	Network Action	78
3.3.1.4	Network Reward function	80
3.3.2	DDPG For Centralized Optimization	82
3.3.3	Arithmetic-Geometric Mean Approximation Scheme For CUs Power Allocation	83
3.4	Proposed Scheme	86
3.4.1	Architecture of Multi-Agent Power Control Scheme With DRL	86
3.4.2	Fair Resource Allocation Scheme Using CO-D3PG	89
3.4.3	Complexity Analysis	94
3.4.3.1	Algorithm 1	94
3.4.3.2	Algorithm 2	95
3.4.3.3	Algorithm 3	95
3.4.3.4	Algorithm 4	95
3.5	Performance Assessment	95
3.5.1	Simulation Parameters	96
3.5.2	Baseline Schemes	96
3.5.3	Experimental Graph and Discussion	97
3.5.3.1	Comparative Metric	97
3.5.3.2	QoS Probability Analysis	99
3.5.3.3	Jain Fairness Indexing	101
3.5.3.4	Convergence Behavior	102
3.6	Summary	102
4	Deep Reinforcement Learning Based Energy Consumption Minimization for Intelligent Reflecting Surfaces Assisted D2D Users Underlying UAV Network	104

4.1	Network Model and Problem Formulation	105
4.1.1	Network Model	105
4.1.2	Channel Model	106
4.1.2.1	MU-BS Channel Model	107
4.1.2.2	D2DT-D2DR Channel Model	108
4.1.3	Energy Efficiency Estimation	109
4.1.4	Problem Formulation	110
4.2	PROPOSED SCHEME	111
4.2.1	Markov Decision Process	111
4.2.2	The C-DDQN Algorithm for the Joint Optimization of Both TR-UAV and PS-RIS.	111
4.2.2.1	Agent	113
4.2.2.2	State Space	113
4.2.2.3	Action Space	113
4.2.2.4	Reward Function	114
4.3	Performance Assessment	115
4.3.1	Parameters for Simulation	115
4.3.2	Results and Discussion	116
4.4	Summary	118
5	Conclusion and Future Scope	119
	References	121

List of Figures

1.1	CISCO VNI Report	1
1.2	General scenario Of D2D	2
1.3	DQN training based DRL scheme	5
2.1	Description of RL Process.	14
2.2	Feed forward NN.	17
2.3	Recurrent NN.	18
2.4	Basic DRL Classifications.	20
2.5	Generalized Value-Based Approaches for DRL.	21
2.6	DQN with Target Network.	23
2.7	DDQN with Target Network.	25
2.8	DRL Techniques Based on PG Methods.	26
2.9	MCPG Method.	28
2.10	Actor-Critic Method.	31
2.11	An example of a H-RL approach, wherein sub-policy π_1 is chosen to generate the environmental action.	32
2.12	An exemplification of the POMDP model is provided, wherein an arrow is sketched from the reliant category to the corresponding dependence, signifying the perceivable information of the surroundings.	34
2.13	A system model to meet the objectives	62
3.1	Network Architecture	67
3.2	Flow Diagram of the Proposed Scheme.	79

3.3	Multi-Agent Power Control Scheme With DRL	87
3.4	Comparative Metric (a) Network Sum Rate vs. Network Size (b) Network Sum Rate vs. Number of DDPs (c) Network Sum Rate vs. Number of CUs (d) Network Sum Rate vs. Minimum SINR requirement Requirement.	98
3.5	Probability Ananalysis (a) QoS Probability vs. Total Number of DDPs (b) QoS Probability vs. Total Number of CUs (c) QoS Probability vs. Minimum SINR Requirement for CUs (d) QoS Probability vs. Interference Threshold.	98
3.6	Jain Fairness Indexing (a) Jain Fairness Indexing vs. Total Number of DDPs (b) Jain Fairness Indexing vs. Total Number of CUs	99
3.7	Convergence Behavior (a) QoS Probability vs. Episode (b) Network Sum Rate vs. Episode.	100
4.1	Network Architecture.	106
4.2	Flow Diagram of Proposed Scheme.	112
4.3	Comparative Analysis (a) e suggested C-DDQN algorithm's convergence rate. (b) The instant rate of transmission over a period of time (c) Mean usage of energy in relation to transmission power.	118

List of Tables

2.1	Comparative Analysis of DRL in D2D	47
2.2	Comparative Analysis of DRL in RIS	54
3.1	A List of Symbols Utilized.	66
3.2	Simulation Parameters	97
4.1	Simulation Parameters	117

List of Abbreviations

Abbreviations	Definitions
1G	First Generation
2G	Second Generation
3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
6GV2X	Sixth Generation Vehicle to Everithing
AC-N	Actor Network
ACT	Actor-Critic Technique
ADMM	Alternating Direction Method of Multipliers
AGMA	Arithmetic-Geometric Mean Approximation
ARSO	Active RIS Subarray Optimization
AWF	Attention Weighted Federated
BDRFL	Block chain-based Deep Reinforcement Federated Learning
BS	Base Station
CC	Computational Complexity
C-DDQN	Centralized Declining Deep-Q Network
C-MADRL	Coordinated-Multi-Agent DRL
CO	Conventional Optimization
CO-CI	Co-Channel Interference
CO-D3PG	Conventional Optimizing Integrated D3PG
CR-CI	Cross-Channel Interference
CR-N	Critic Network
CSSCA	Constrained Stochastic SCA
CUs	Cellular Users
D2D-C	Device to Device Communication
D2DPs	D2D Pairs
D2DR	D2D Receivers
D2DT	D2D Transmitters
D2D-U	D2D Users
D3PG	Distributed Deep Deterministic Policy Gradient
DDPG	Deep Deterministic Policy Gradient
DDPs	D2D Pairs
DDQN	Double Deep Q-Networks

DDRs	D2D Receivers
DDTs	D2D Transmitters
DL	Deep Learning
DLR	Decaying Learning Rate
DNNs	Deep Neural Networks
DPGA	Deterministic Policy Gradient Approach
DQN	Deep Q-Networks
DRL	Deep Reinforcement Learning
DRPO	Decomposition and Relaxation-based Precoding Optimization
DSM	Dynamic Spectrum Matching
DTD3	Dinkelbach Combined Twin Delayed DDPG
EA-DDPG	Exploration Attenuate-Deep Deterministic Policy Gradient
EE	Energy Efficiency
EH	Energy Harvesting
EHR	Energy Harvesting Robust
ERA	Efficient Resource Allocation
ERAP	Efficient Resource Allocation Parallel
FC	Fuzzy Clustering
FFNN	Feed-Forward Neural Network
FS	Fair Scheduling
FUF	Fairness Utility Function
GLSTM	Long Short-Term Memory Network
HLA	Homogeneous Linear Array
HRL	Hierarchical Reinforcement Learning
ICWN	Information-Centric Wireless Networking
IMT	International Mobile Telecommunications
IoT	Internet of Things
IS-UAV-TN	Integrated Satellite-Unmanned Aerial Vehicle-Terrestrial Network
LB	Lower Bound
LCPP0	Low Complexity Proximal Policy Optimization
MA	Multi-Agent Approaches
MAAC	Multi-Agent Actor Critic
MAC	Medium Access Control
MA-DRL	Multi-Agent Deep Reinforcement Learning
MAML	Model Agnostic Meta Learning
MAOD	Multi-Agent Online Distributed
MAQ	Multi-Agent Q-Learning
MARL	Multi-Agent Reinforcement Learning
MCPG	Monte Carlo Policy Gradient
MDP	Markov Decision Processes
MIMO	Multiple Input Multiple Output
MINLP	Mixed-Integer Non-Linear Programming
ML	Machine Learning
MSPA	Mode Selection Integrated with Power Allocation
MSRA	Mode Selection and Resource Allocation
MUs	Mobile Users
NAAC	Neighbor-Agent Actor Critic
NC	Non Cooperative
NOMA	Non-Orthogonal Multiple Access

NPG	Natural Policy Gradient
OMA	Orthogonal Multiple Access
P2P	Peer-to-Peer Services
PDCP	Packet Data Convergence Protocol
PER	Prioritized Experience Replay
PG	Policy Gradient
PGAC	Policy Gradient Actor- Critic
POMDP	Partially Observable Markov Decision Processes
PS	Priority Sampling
PSD	Phase Shift Design
PDSGD	Primal-Dual Sub Gradient Descent
PW	Piece-wise
QF	Q-Function
QoS	Quality of Service
RAFD	RIS Assisted Full Duplex
RBs	Resource Blocks
RIS	Reconfigurable Intelligent Surfaces
RL	Reinforcement Learning
RLC	Radio Link Control
RNN	Recurrent Neural Network
SACAO	Soft Actor-Critic-Alternating Optimization
SAC-AO	Soft Actor-Critic-Alternating Optimization
SACC	Social-Aware Cooperative Caching
SAMDP	Semi-Aggregated Markov Decision Process
SC	Sub Channel
SCA	Successive Convex Approximation
SE	Spectral Efficiency
SGMA	Stackelberg Game Guided Multi-Agent
SIC	Successive Interference Cancellation
SOTPSR	Self-Organization of Transmission Power and Power Splitting
SPGA	Stochastic Policy Gradient Approach
SRPR	Successive Rounding and Power Refinement
ST	Social Trust
TDE	Temporal-Difference Error
TNs	Target Networks
TP	Two Parallel
UARA	User Association and Resource Allocation
UAV	Unmanned aerial vehicle
UCA-HSE	UAV Collision Avoidance-High Sample Efficient
UEs	User Equipments
UMC	Ultra-Massive Connectivity
V2V	Vehicle-to-Vehicle Communication
VF	Value Function
WLAN	Wireless Local Area Network
WLAN	Wireless Local Area Network
WL-DDPG	Water Filling Deep Deterministic Policy Gradient
VF	Value Function
WLAN	Wireless Local Area Network
WLAN	Wireless Local Area Network
WL-DDPG	Water Filling Deep Deterministic Policy Gradient
WPT	Wireless Power Transfer ^{XX}

Chapter 1

Introduction

The proliferation of mobile users, smart gadgets, and multimedia applications generates an unprecedented growth of data traffic in 5G and beyond networks. As per the Cisco visual networking index report of 2023, 100 billion devices are going to generate 75 exabytes of data traffic per month and illustrated in Fig. 1.1 [1]. This massive increase in data traffic puts a lot of burden on efficient spectrum utilization in the years to come. To address this problem, researchers recommended various D2D communication (D2D-C) techniques.

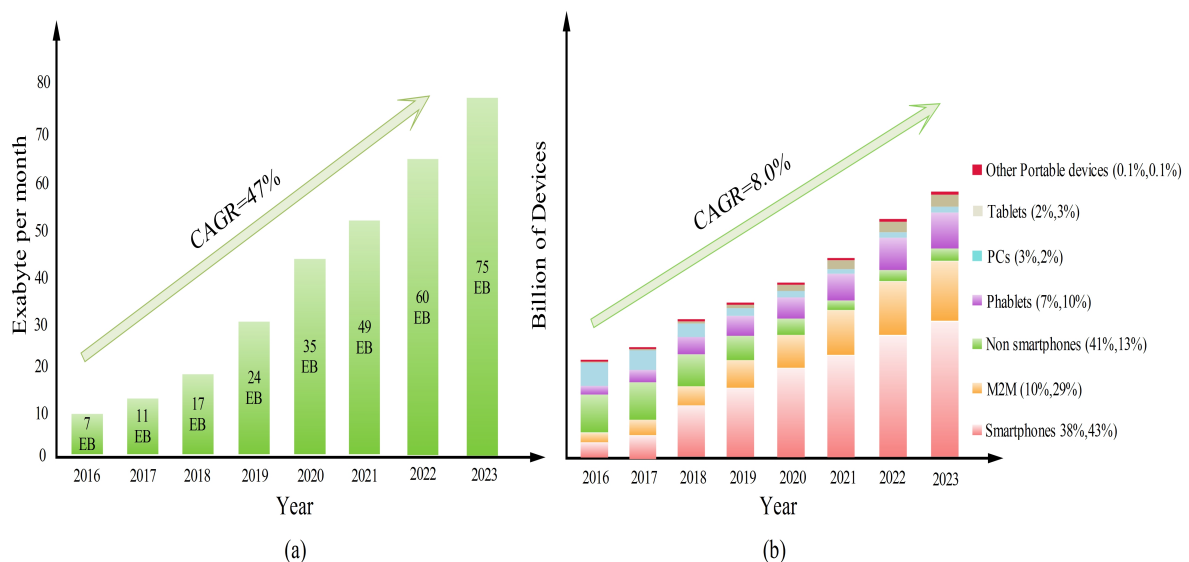


Figure 1.1: CISCO VNI Report

1.1 D2D Communication

In D2D-C technology, two neighboring devices can share the data directly without the base station (BS) as shown in Fig. 1.2 [2]. As a result, it enhances mobile users' quality of service by reducing the transmission delay. In D2D-C, the D2D pairs (DDPs) utilize the same RBs as the CUs to enhance SE and create a direct link. D2D can utilize either licensed (in-band) or unlicensed (out-band) bands with the CUs to achieve this goal [3]. Furthermore, when it comes to band-to-band direct communication, it is divided into two distinct categories: overlay and underlay. On the other hand, out-of-band communication is classified as either controlled or self-governing. In the context of D2D overlay, a precise allocation of RBs is designated solely to D2D users, while the remaining RBs are employed by CUs. On the contrary, within the context of D2D underlay, both D2D users and CUs utilize an identical set of RBs. In the case of out-band communication, when communication is controlled, the resources among D2D users are governed by the BS. However, in autonomous communication, it is the user equipment (UEs) that governs the resources.

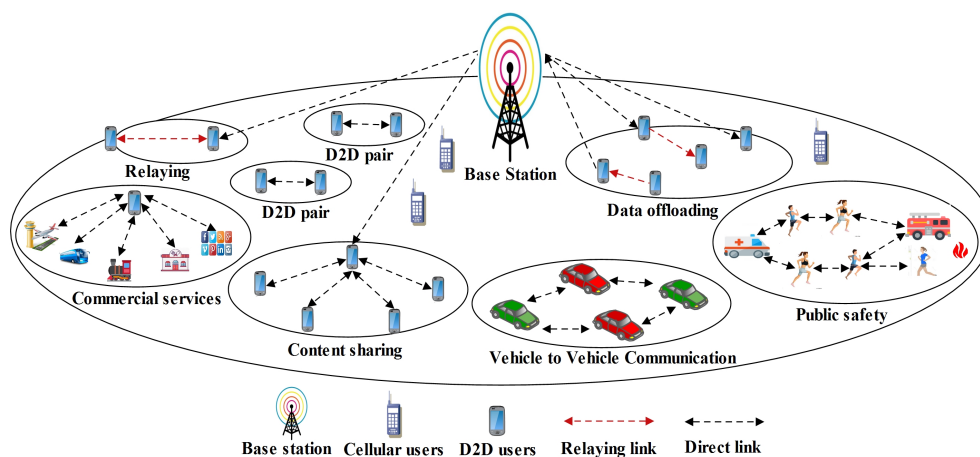


Figure 1.2: General scenario Of D2D

1.1.1 Potential Benefits of D2D-C

The advantages that can be gained from D2D communication [4, 5] include: (i) increasing energy efficiency through low power proximity services; (ii) improving SE by allowing the

sharing of RBs between DDU and CU; (iii) being able to provide assistance for a wide range of peer-to-peer (P2P) services; (iv) reducing the data traffic burden on base stations; and (v) enhancing QoS and lowering latency for cell edge users during short-distance communication. Despite these advantages, key challenges such as CR-CI and CO-CI, as well as ultra-massive connectivity (UMC), need to be investigated more for QoS provision to the end users [6].

1.1.2 Use Cases of D2D

1.1.2.1 Local Services

UEs can exchange data with one another directly without requiring the involvement of BS. This service is typically utilized for local information distribution, social applications, and offloading cellular data traffic. For instance, when it comes to transmitting local information, the information is transmitted directly between DDU and CU, taking into account their respective proximity. The implementation of D2D-C not only serves to enhance mobile applications, but it also effectively conserves spectrum resources, thereby creating new revenue streams for operators. Therefore, local advertising companies aim to maximize benefits by targeting individuals based on their proximity. For instance, a cinema located within a shopping mall entices individuals by providing them with show schedules and promotional show-tickets.

1.1.2.2 Emergency Communication:

In the case of natural disasters like earthquakes, floods, or tsunamis, the traditional communication infrastructure is often damaged, thereby disrupting the network and adversely affecting the rescue process. In such a scenario, multi-hop wireless connectivity can be established among D2D users to guarantee seamless communication.

1.1.2.3 Enhancing the Internet of Things (IoT)

The main purpose behind commencing cellular communication was to create an extensive system that enables smooth and uninterrupted communication between various User Equipment (UEs). This thought generate an idea for developing IoT using cellular communication system.

The report published by Cisco [1] predicts that by the year 2020, there will be a staggering 50 billion wireless devices in use worldwide. Out of those, around 30 billion devices may rely on IoT capabilities, specifically machine terminals. When D2D-C is combined with IoT, it results in the creation of a flawless interconnected wireless system. An exemplary instance of this would be vehicle-to-vehicle (V2V) communication.

1.1.2.4 D2D in Multiuser-Multiple Input Multiple Output (MU-MIMO):

In a typical MU-MIMO scenario, the pre-coding weights are determined through the assistance of BS in order to provide feedback to the UEs. This particular approach implemented at terminals generates nulls and effectively reduces interference among UEs. However, it is limited to single user MIMO. On the contrary, the implementation of D2D-C results in the creation of paired users who are able to exchange data via the same channel. As a consequence, it improves the performance of multi-user MIMO systems with regards to SE.

1.2 Deep Reinforcement Learning (DRL)

A considerable amount of resources has been dedicated towards enhancing the SE and EE of the network. This has involved a rigorous focus on optimization theory, with the aim of devising algorithms that are more efficient and can yield optimal or near-optimal outcomes. However, it is noteworthy that a considerable number of prior investigations have assumed a static network setting. In wireless networks that are advanced and have a huge number of devices, the surroundings are frequently subject to dynamic instability. Hence, it is highly preferable to empower network nodes with the capacity for independent decision-making. The optimization of network performance can be achieved by solely relying on local observations when making decisions, including power allocation, spectrum access, and interference management.

Reinforcement learning (RL) is a powerful technique in which an autonomous agent performs sequential decisions using various mathematical functions [7]. An agent in RL interacts with the dynamic environment through trial-and-error learning. After learning, the agent optimizes the performance of its previous action. Then, the agent initiates a new action, analyses

the outcomes of the interaction, and takes a decision in an interactive learning manner. In the past two decades, numerous widely-used applications of RL have surfaced, spanning across various fields such as robotics, natural language processing [8], and game solving [9]. Among these, AlphaGo [10], which was the first software program to outdo a world-class professional player in the game of Go, stands out as one of the most familiar applications of RL.

RL has emerged as a significant method for dealing with issues in modern networks in the field of wireless communications. Due to this, the convergence of existing learning schemes slows down as the agent must explore the entire state space frequently in order to discover the correct policy for the overall system. As a result, in complex domains, traditional RL schemes become ineffective and impracticable. Deep learning (DL) [11] approaches have recently introduced a new era of computer vision breakthroughs. DL is a subdivision of ML that focuses on the development of scalable and adaptable algorithms for the purpose of optimizing complex functions on massive datasets. In recent years, DL has been effectively utilized in a multitude of domains such as gaming [10], facial recognition [12]- [13], and cybersecurity [14]. Deep reinforcement learning (DRL) [15] is an advanced approach that integrates DL with RL.

DRL employs Deep Neural Networks (DNNs) to improve the process of training and learn-

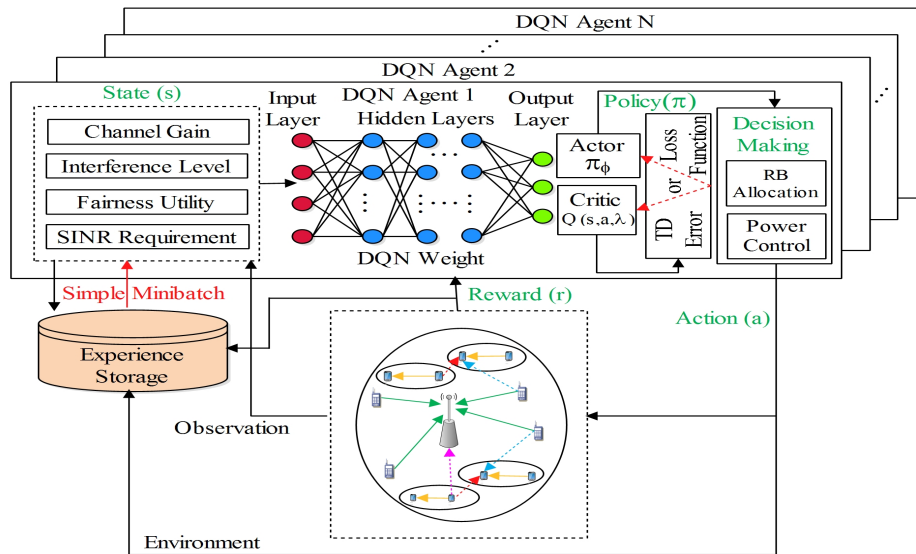


Figure 1.3: DQN training based DRL scheme

ing, resulting in faster learning and better performance of RL algorithms as shown in Fig. 1.3 [15]. The following are some of the main benefits of DRL: (i) DRL has the ability to

solve complex network optimizations. As a result, network controllers are able to solve non-convex and complex problems. It's also used to find the best options in the absence of full and reliable network data. (ii) DRL enables network system to learn about and improve their understanding of the communication and networking environment. (iii) DRL enables the capacity for autonomous and self-directed choices. Using DRL schemes, users in the network can observe each other and determine the optimum strategy, with minimal or no need for information exchange. Network entities may use DRL schemes to find the best strategy by making observation locally with less or no information sharing. DRL approaches accomplish this by reducing communication overhead while simultaneously enhancing network security and dependability. (iv) For large state and action spaces DL may increase the learning speed remarkably.

1.3 Motivation

In the existing proposal it has been observed that authors have not studied the impact of interference, ultra massive connectivity (UMC), and fairness using NOMA and DRL to their full potential. In this thesis, our target is to optimize the fairness utility function (FUF) of the whole network with respect to spectrum resources and power allocation. Non orthogonal multiple access (NOMA) is used at the BS to serve a set of CUs across each resource with respect to their different power levels. Also, D2D pairs (DDPs) reuse the same spectrum resource in an orthogonal manner while maintaining interference across different cellular users (CUs). Here, traditional DDPs are used instead of NOMA-based DDPs to reduce the computation on the resource-constrained DDPs, and hence they are more practically applicable. Also, unlike NOMA-based DDPs, where two or three D2D users may often be scheduled to maintain minimal SIC receiver complexity, the number of D2D users that can join the network is not tightly restricted if the interference is properly handled. Also, to improve fairness and reduce the cross channel interference (CR-CI) and co channel interference (CO-CI), the DRL scheme is proposed. Using the DRL, each agent is assigned a certain reward in order to meet the goal. To remove the instabilities of a multi-agent framework, every agent exchanges its historical status, activities, and objectives in a centralized manner. This contradicts our previous claims, as agent

always consider the previous record to take the next action, but in the proposed scheme, agents analyze the previous data to train the RL model. It results in decreasing the communication time imposed by the algorithm to train the model. Also, this technique reduces the hardware complexity and computational time on the resource-constrained D2D devices, which results in lower intra-user interference and faster convergence speed. Despite these advantages, fairness, CR-CI and CO-CI are still issues that need to be investigated. This motivates us to develop a solution for underlay DDPs by coordinating them with the NOMA-based CUs. Optimization of Energy efficiency (EE) in NOMA up-link networks is not yet investigated enough. In most of the existing proposals, authors conveyed matching game theory and NOMA as a cogent tool to address massive connectivity and spectral efficiency in D2D-C. To accommodate exigencies of 5G and 6G users, it is the need of the hour to eliminate various interferences by optimizing EE in the underlay D2D network. Therefore, in our thesis, we got motivation to use the DRL technique to effectively allocate SCs for optimization of EE by maximizing throughput.

1.4 Thesis Organization

The structure of this thesis is presented as such, accompanied by a succinct overview of every individual chapter.

Chapter 2: Literature Review

In this particular chapter,, we explore the various DRL variants that are applicable in the context of 5G and beyond. The literature review is partitioned into three distinct sections. The first section of the literature review thoroughly examines the overview and categorization of DRL. Firstly, a comprehensive discussion is presented on the basics of RL and DL. Then we present the categorization of DRL schemes, which is established upon the distinct characteristics of the policy functions, various approaches to policy evaluation,, and the parameter updates in various learning techniques. The second part of the chapter discussed the various DRL variants in the D2D-C. The third part of the literate review discussed the various DRL variants in the implementation of RIS technology.

Chapter 3: A Deep Reinforcement Learning Scheme for Sum Rate and Fairness Maximization Among D2D Pairs Underlying Cellular Network With NOMA

This chapter investigates the sum-rate and fairness maximization among NOMA-enabled CUs and DDPs while considering the resource and power constraints of BS and DDT. To achieve the target, firstly, the centralized DDPG is used to allocate the resources to CUs with NOMA. Afterwards, to mitigate the CR-CI, CO-CI and improve fairness among the CUs, the AGMA technique is integrated with DDPG. The AGMA optimize the power of the CUs and maintains the power distribution among the CUs based on their channel gain. However, it is found that the DDPs do not train simultaneously because they cannot access the instantaneous global CSI in real time. This problem results in an increase in CO-CI and a decrease in fairness. To address these problems, first of all, the D3PG is proposed to improve fairness. D3PG helps provide resources to DDPs by reusing the CUs' resources. Finally, to reduce CO-CI, CO-D3PG is proposed. CO-D3PG is the combination of CO and D3PG that controls the power of the DDPs. The experimental results reveal that the proposed scheme enhances the overall network's sum rate by maintaining fairness among the CUs and DDPs as compared to baseline schemes.

Chapter 4: Deep Reinforcement Learning Based Energy Consumption Minimization for Intelligent Reflecting Surfaces Assisted D2D Users Underlying UAV Network

In this chapter, we put forth a framework based on DRL for leveraging RIS in a UAV-assisted wireless network. The aim of this framework is to enhance the EE of the entire network. The UAV's trajectory and RIS's phase shift control issue have been formulated using the C-DDQN method. The C-DDQN method achieved a trade-off between an increasing learning rate and convergent local optimality. The C-DDQN approach not only circumvents oscillation, but also guarantees that the data necessities of each user are satisfied in each time slot. Furthermore, the findings illustrate that the proposed technique performs better than the cutting-edge schemes.

Chapter 5: Conclusion and Future Scope

This chapter provides a comprehensive overview of the contributions made through the uti-

lization of the proposed strategies, thus concluding the thesis. Moreover, this specific chapter provides comprehensive coverage of the forthcoming opportunities regarding the research areas of D2D-C, DRL, and RIS in 5G and future networks. In this study, we have put forth the two approaches to resolve the previously mentioned concerns.

1.5 Summary

This chapter delves deep into the essential idea of employing D2D-C and DRL, providing a comprehensive discourse on the topic. Furthermore, within this chapter, we elucidate how DRL schemes effectively address the challenges of co-channel and cross-channel interference. We have also deliberated on how implementing the DRL technique can improve the throughput, SE, and EE of the D2D underlay cellular network.

Chapter 2

Literature Review

The progression of wireless communication commenced with the inception of first generation (1G) and advanced to the fourth generation (4G). To fulfil the demand of users and to cope up with the continuous increase of data traffic, fifth generation (5G) will be introduced by 2020. 1G was limited to the audio communication with analog services only. Whereas, digital technology came in to existence from 2G with direct communication between users equipments (UEs) for example walkie-talkie. The direct communication between UEs became popular from 3G with the introduction of wireless local area network (WLAN) and wireless personal area network (WPAN) technologies. With the help of WLAN and WPAN communication can be done through unlicensed band due to which interference levels can not controllable. As a result, quality of service (QoS) cannot be guaranteed and consumes high power.

With the increase in number of subscribers and massive usage of smart devices produce large interference which becomes more severe. To overcome this problem a trend of sharing data shifts from unlicensed band to licensed band in 4G. The sharing of data via licensed band reduced the problem of interference as well as improves the networks SE. The exponential demand in usage of high data rate multimedia applications, video streaming, live conferences, 3D gaming provoked researchers towards 5G. The Direct communication between devices from 2G to 4G is network centric. In network centric wireless communication system, BSs act as a controlling entity to establish the communication. On the other hand, in 5G the deployment of small cells shift the technology from network to device centric where, in device centric devices

can directly communicate with or without BSs. In D2D-C, users actively participate to perform storage, relaying, safety and content distribution. Where QoS of the cell edge users can also be guaranteed due to cooperative sharing of data from cell center users to cell edge users.

Since wireless technology is constantly improving, the number of users is growing at an exponential rate. There is a need of better quality of service in dynamic environment. RL represents one of the most crucial research directions in the field of ML and plays a significant role in the advancement of AI [7]. RL represents a self-learning approach where an agent can regularly make choices, monitor the results, and automatically adapt its strategy to achieve the most efficient policy. It has been demonstrated that the learning process in RL does converge; however, it does require a significant amount of time to reach the optimal policy. The cause behind this is that RL necessitates extensive exploration and acquisition of knowledge pertaining to the entire system, thereby rendering it unsuitable and unsuited for deployment in ultra-dense networks. As a consequence, the implementations of RL in practical applications are exceedingly diminutive. DL [16] has been successfully implemented in a variety of fields in recent years. The integration of DL with RL through the use of DNNs, known as DRL, represents a modern approach to training the learning process. This approach leads to faster learning and improved performance of RL algorithms.

Markov Decision Processes (MDP) [17] provide a powerful tool for modeling decision-making issues that arise in uncertain and stochastic environments. Dynamic programming [18], [19] as well as other algorithms including value iteration and RL approaches can be applied to solve the MDP. Modern networks, on the other hand, are vast and complex, so computational complexity of the techniques quickly becomes incommensurate. Hence DRL has evolved into a viable alternative to overcoming the issue. The DRL methods have the following advantages in general [20]:

- DRL possesses the capability to effectively resolve complex network optimization issues. As a consequence of this, DRL empowers BS to tackle complex and nonlinear challenges. Furthermore, DRL also handles the association of UEs jointly, as well as computing and transmitting scheduling, even in scenarios where network information is incomplete or imprecise.

- DRL empowers its users to acquire knowledge and enhance their understanding of the WCN environment. By implementing DRL strategies, CUs have the ability to acquire optimal policies for actions such as selecting BSs, channels, and making handover decisions. Additionally, caching and offloading decisions can also be made without prior knowledge of system models or users mobility patterns.
- DRL provides the opportunity for autonomous decision-making. Through the implementation of DRL approaches, various network entities are able to individually observe and deduce the most optimal policy of action with minimal to zero exchange of knowledge or information. Network entities may use DRL schemes to find the best strategy by making observation locally with less or no information sharing. By doing this DRL approaches lowers the communication overhead but at the same time increases network security and reliability.
- DRL greatly enhances the pace of learning, particularly when faced with challenges involving vast state and action domains, as seen in IoT networks equipped with numerous devices. DRL allows for the monitoring of user association, spectrum access, and transmission power for a vast number of IoT terminals and CUs on a dynamic basis through IoT gateways.
- There are various problems in communication and networking. These issues consist of interference control, cyber-physical attacks, and data offloading. It is possible to conceptualize these issues as games, specifically non-cooperative games

The value functions (VF) as well as Q-functions (QF) use a huge amount of memory when used in the RL setting. However, in real world applications, the sets representing the state are often huge. Therefore, it is a challenging task to store the VF and QF in tabular form due to the huge memory and computational complexity requirements. The domain of DL emerges as a relevant contributor to the field as it enables the approximation of certain RL functions, including value/Q-functions and policy functions, through the utilization of a reduced number of parameters. The integration of DL into RL yields a more potential technique commonly referred to as DRL.

2.1 DRL OVERVIEW

The term DRL refers to the combination of reinforcement learning (RL) and deep learning (DL). The value functions (VF) as well as Q-functions (QF) use a huge amount of memory when used in the RL setting. However, in real world applications, the sets representing the state are often huge. Therefore, it is a challenging task to store the VF and QF in tabular form due to the huge memory and computational complexity requirements. The domain of DL emerges as a relevant contributor to the field as it enables the approximation of certain RL functions, including value/Q-functions and policy functions, through the utilization of a reduced number of parameters. The integration of DL into RL yields a more potential technique commonly referred to as DRL.

2.1.1 DRL Fundamental Building Blocks

2.1.1.1 Fundamental of RL

RL is a distinct form of algorithm that is employed in the realm of machine learning (ML). It possesses the ability to attain optimal control over a MDP. The RL framework comprises of primarily two units, as depicted in Fig. 2.1 [20], namely the agent and the environment. The environment is in a constant state of evolution, occurring in a random manner and capable of existing in any of the possible states inside a particular state space at any given moment. RL is a ML technique that relies on feedback mechanism. In this methodology, an agent acquires knowledge on how to act within a specific surrounding by carrying out actions and monitoring the consequences of said actions. The agent is consistently receiving positive feedback for every remarkable action taken, while any unfavorable action results in negative feedback or a penalty. Furthermore, the action that has been mentioned earlier will inevitably impact the subsequent state that the environment will shift towards. The state-action pair's stochastic evolution over time results in the formation of a MDP[4]. The MDP is described by five tuples $(S, A, P, \gamma, \Gamma)$, where S defined as a set of an agent's finite states, A defined as a set of finite states taken by agents, P defined as a probability matrix that expresses the probability of transition from one state to another state, γ defined as a set of rewards based on the current situation

and the action taken, and Γ defined as a discount factor.

A policy dictates the manner in which an agent chooses its actions in varying states, and

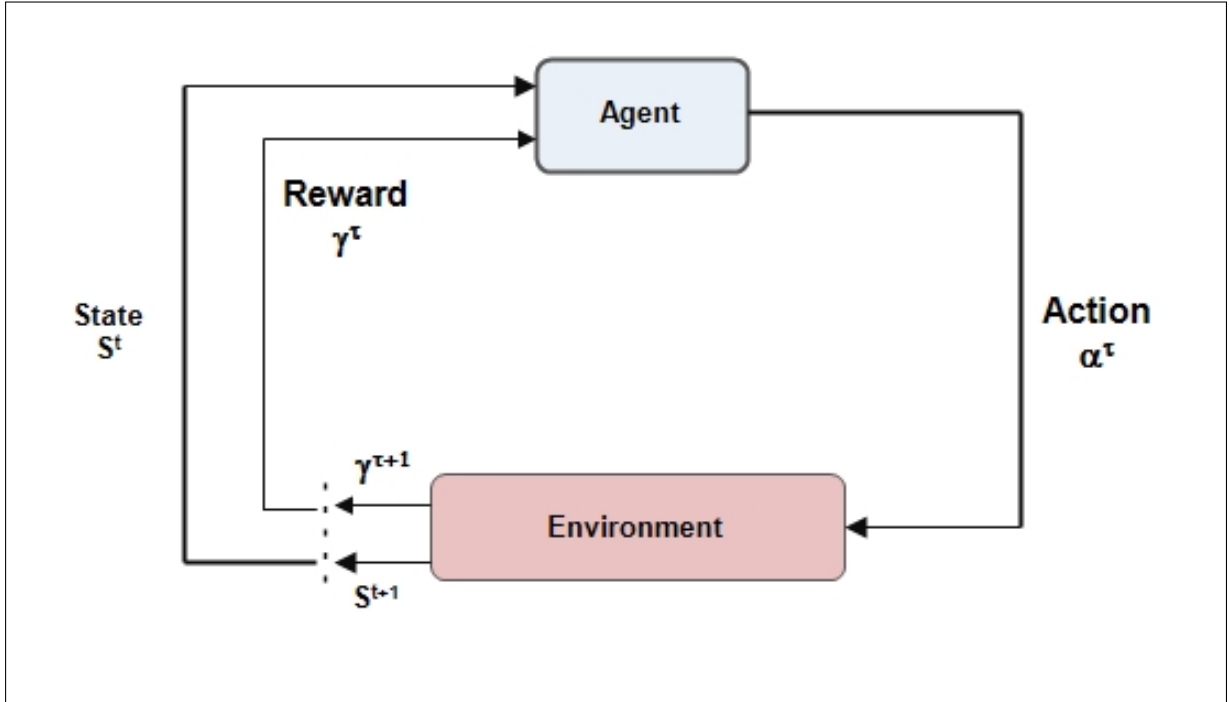


Figure 2.1: Description of RL Process.

based on this criteria, a policy can be classified as deterministic or stochastic [21]. The policy, in a random scenario, is expressed by $\pi(s^t, a^t) := \mathcal{P}(s^t | a^t)$. This expression signifies the likelihood of selecting a specific action a^t in a particular state s^t . On the other hand, policy, in a deterministic scenario, policy is expressed by $\pi(s^t) = a^t$. This expression signifies the likelihood of selecting a specific action a^t in a particular state s .

For the purpose of improving the clarity of our introductory explanation, we will concentrate on utilizing models that have distinct time values. The engagement of the agent in this specific model necessitates active participation with the surrounding environment through a sequence of distinct temporal intervals. The agent's aim is to acquire the ability to efficiently map states with corresponding actions. This involves the discovery of a policy aimed at optimizing the VF, defined as $\mathbb{V}_f^\pi(s^0)$, over any given state s^0 . The expected reward is achieved through the VF, defined as $\mathbb{V}_f^\pi(s^0)$, once a strategy is put into action, commencing from an starting state s^0 and is defined as follows:

$$\mathbb{V}_f^\pi(s^0) = \mathbb{E}_{u_{s^0} \sim \pi} \left[A(u_{s^0}) \right] \quad (2.1)$$

In (1) \mathbb{E}, u_{s^0} and $A(u_{s^0}$ denotes the expectation operator, sequence comprised of triplets and average reward, respectively. The triplets is represented as (s^t, a^t, γ^{t+1}) . Defining $\gamma^{t+1} = \gamma(s^t, a^t)$, $a^t \sim \pi(s^t, a^t)$ and $s^{t+1} \sim \mathcal{P}(s^{t+1}|s^t, a^t)$.

In addition to the VF, another crucial function is the QF $Q^\pi(s^0, a^0)$. This particular function is responsible for determining the anticipated reward upon performing a definite action a^0 in a specific state s^0 , followed by an adherence to a policy π . If the policy, denoted as π , is deemed to be the most optimal policy, it shall be represented by π^* . Additionally, the corresponding VF is defined as $\mathbb{V}_f^*(s^t)$, while the QF is similarly defined as $\mathbb{Q}^*(s^t, a^t)$. Now using Bellman optimal equations the VF can be expressed as follows [22]:

$$\mathbb{V}_f^*(s^t) = \mathbb{V}_{\pi^*} = \max_{a^t} \left[\gamma^{t+1} + \Gamma \sum_{s^{t+1}} \mathcal{P}(s^{t+1}|s^t, a^t) \mathbb{V}^*(s^{t+1}) \right]. \quad (2.2)$$

Similarly the QF can be expressed as follows:

$$\mathbb{Q}^*(s^t, a^t) = \gamma^{t+1} + \Gamma \sum_{s^{t+1}} \mathcal{P}(s^{t+1}|s^t, a^t) \max_{a^{t+1}} \mathbb{Q}(s^{t+1}, a^{t+1}). \quad (2.3)$$

The correlation between the VFs/QFs of the present state and the subsequent state is depicted by the Bellman equations. From both (2.2) and (2.3), it can be inferred that the anticipated reward is equivalent to the total of the instant reward and the maximum potential reward. After obtaining the anticipated future reward, one can effortlessly calculate the probable reward from the present state. The Bellman equations form the foundation of a significant category of RL techniques that utilize the "bootstrap" approach. An agent initiates the VFs/QFs with random values over the learning phase. Then, the algorithm continues to perform the phases of predicting the policy and evaluating the policy in an iterative manner until the VFs/QFs have converged. During the policy predicting stage, the agent selects an action depending on the current VFs/QFs, which leads to an instantaneous reward as well as a new state. In the policy evaluating stage, the agent proceeds to update the VFs/QFs in accordance with the Bellman equations (2.2) or (2.3) while taking into account the instant reward as well as the new state.

2.1.1.2 Fundamental of DL

DL is a distinct form of algorithm that is employed in the realm of ML. It heavily relies on artificial neural networks (ANN) to acquire knowledge from massive amounts of data set in a self-governing manner. It has the ability to excel in tasks such as regression as well as classification algorithms. The regression algorithm involves estimating a continuous parameter. On the contrary, the classification algorithm entails the estimation of the output based on a collection of limited categorical parameters. Let \mathbb{X}_i denotes the input data and \mathbb{Y}_o denotes the output data then NN models can be perceived as a precise mathematical framework that establishes a unique function $\mathcal{F} : \mathbb{X}_i \rightarrow \mathbb{Y}_o$. The NN's learning rule adjusts its parameters to enable the network to generate a preferred output $\tilde{\mathbb{Y}}_o$ that most closely resembles the target output value \mathbb{Y}_o , given a specific input \mathbb{X}_i .

Fig. 2.2 [20] shows the architecture of feed-forward NN (FFNN). It is created with three completely connected layers (input layer, hidden layer, and output layer). Every layer Each layer is made up of one or multiple neurons that represent unique non-linear nodes within NN structure. The k^{th} neuron located in the l^{th} layer of FFNN is equipped with a weight vector, represented as $\mathbb{W}W_k^l$, that corresponds to the connections from the preceding $(l-1)^{th}$ layer to its own layer, along with a bias value b_k^l . An Activation Function is used in FFNN for the activation of a particular neuron. This implies that the neuron's contribution to the network's input will be evaluated to determine its significance in making predictions, through the use of more straightforward mathematical operations. The main function of the Activation Function is to convert the combined weighted input from the node into an output value that is to be supplied to the next hidden layer. The activation functions (h) that are most frequently utilized include Sigmoid, Logistic, Tanh, and ReLU. Now k^{th} neuron's output in the l^{th} layer is defined as follows:

$$a_k^l = h(w_k^l a_{l-1}^T + b_k^l) \quad (2.4)$$

A FFNN lacks the concept of temporal sequence and solely focuses on the present set of input data it has been provided. A Recurrent Neural Network (RNN) is a distinct type of NN that has the capability to process input sequences by utilizing internal memory. In the context

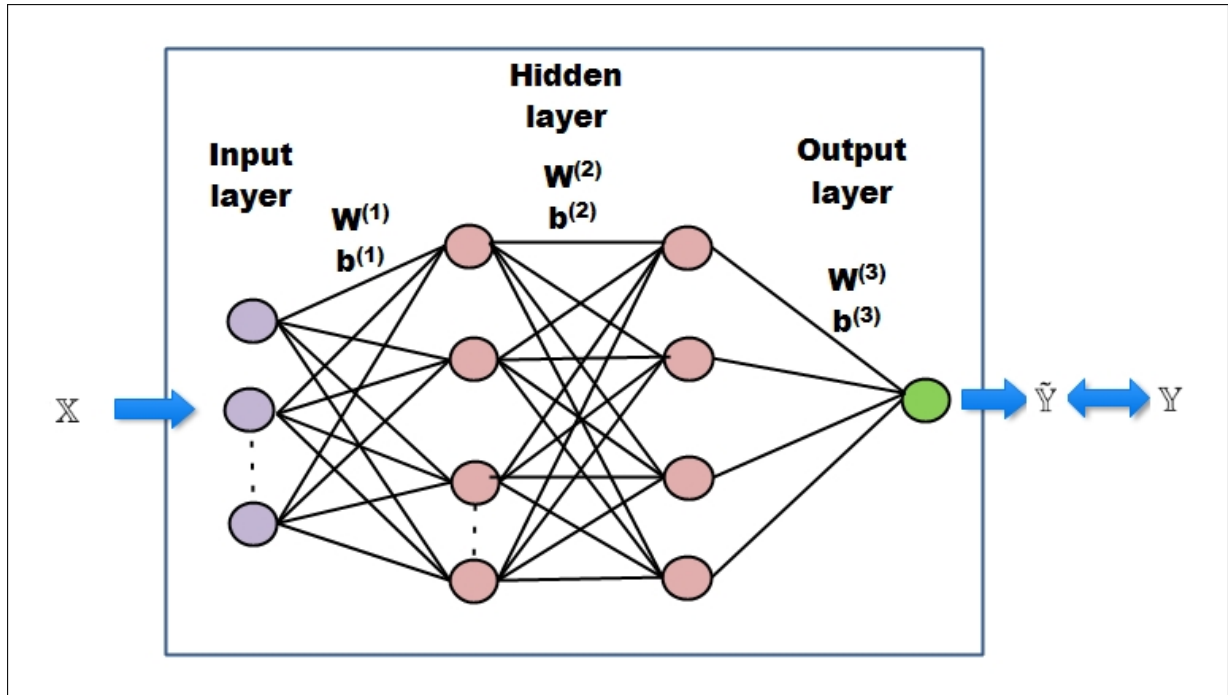


Figure 2.2: Feed forward NN.

of a RNN, the previous output generated by the neurons residing within the concealed has the potential to serve as an additional input, supplementing the existing input dataset. This capability allows the RNN to acquire knowledge from past events and improve its learning. The fundamental framework of the RNN is depicted in Fig. 2.3 [20]. Typically, within the realm of DL, a loss function (LF), represented by $\mathcal{L}_f(\phi = g(\tilde{Y}(\phi)Y))$, is utilized. This function is dependent upon both the output $\tilde{Y}(\phi)$ produced by the NN as well as the target output Y_o . The evaluation of the given data $Y = \mathcal{F}(X)$ is done by the loss function, which assesses the effectiveness of a particular NN with its current learned parameter values ϕ .

The objective of the NN is to diminish the loss function, specifically to achieve the minimum value of $\min_{\phi} \mathcal{L}_f(\phi)$. A technique known as gradient descent (GD) is used to modify the parameters ϕ in NNs for this particular objective. It is customary to utilize the simple gradient $\nabla_{\phi} \mathcal{L}_f(\phi) = \frac{\partial \mathcal{L}_f(\phi)}{\partial \phi}$ in order to update the parameters for a specified function $\mathcal{L}_f(\phi)$. The process of gradient descent commences by selecting a point ϕ_0 as the initial position. When a set of input data is given to the NN in the form of a mini-batch, it proceeds to compute the average loss function for all the data present within the batch. This value is subsequently employed to ascertain the minimum of $\mathcal{L}_f(\phi)$ by executing a step in the direction of descent, which is

defined as:

$$\phi \leftarrow \phi - \chi \nabla_{\phi} \mathcal{L}_f(\phi), \quad (2.5)$$

where $0 < \chi < 1$ is the range of χ and χ is referred to as the learning rate and is a hyperparameter. It is configured to figure out how quickly the values of the parameters shift towards the optimal direction with great efficiency. The previously mentioned procedure is executed in a repetitive manner, with supplementary mini-batches of input data being incorporated into the NN until it attains convergence.

Deriving the simple gradient $\nabla_{\phi} \mathcal{L}_f(\phi)$ is a straightforward task, however, it has been commonly observed that it may not be the most efficient approach for optimizing the loss function.

$$a_k^l = h(w_k^l a_{l-1}^T + b_k^l) \quad (2.6)$$

function. It is critical to create an appropriate step size value during the training process.

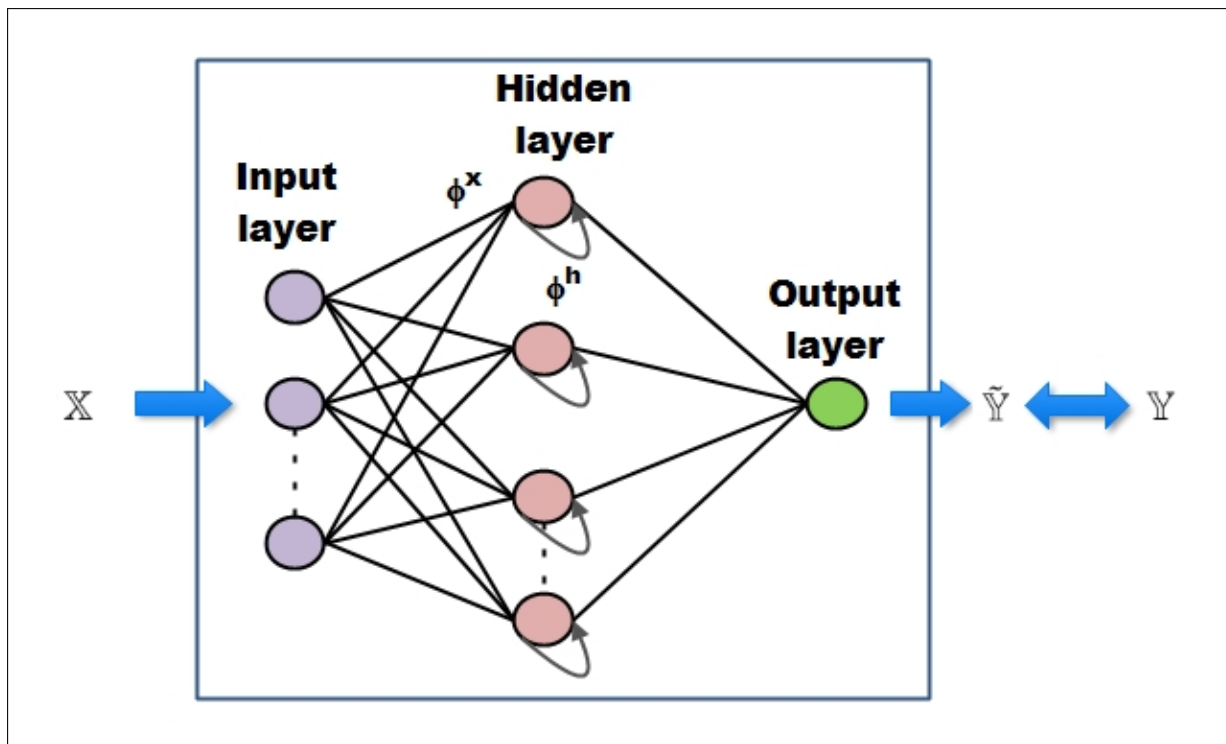


Figure 2.3: Recurrent NN.

This is because an excessively large value may hinder the ability to reach the local minimum, while an excessively small value may consume an excessive amount of time to arrive at the

local optimal point. Natural gradients, denoted as $\nabla_{\phi}^N \mathcal{L}_f(\phi)$, do not adhere to the conventional steepest path in the space of parameter, unlike straightforward gradients. Instead, they proceed in the direction of steepest descent within the distribution space, as determined by the Fisher metric. The Fisher information metric F [21] is customarily utilized to ascertain the magnitude of the step size for achieving $\nabla_{\phi}^N \mathcal{L}_f(\phi) = \nabla_{\phi} \mathcal{L}_f(\phi) F^{-1}$ in a precise manner.

2.2 Classification of DRL Algorithm

In this particular section, we begin by classifying the primary DRL algorithms into two comprehensive categories, specifically the value-based (VB) and policy gradient (PG) methodologies. The classification relies on the utilization of NNs for the approximation of value/Q-functions and policy functions, as demonstrated in Figure 2.4. The PG approaches are examined in greater detail in the current discourse by exploring them from three distinct perspectives.

- According to the distinct attributes of the policy functions: we propose two distinct methods - the stochastic policy gradient approach (SPGA) and the deterministic policy gradient approach (DPGA).
- According to the various policy evaluation approaches: we propose a comparative analysis of Monte Carlo PG with actor-critic approaches.
- According to the parameter updates in various learning techniques: we propose a comparison between the simple policy gradient with natural policy gradient (NPG) approaches.

After that, we present two categories of advanced DRL techniques, specifically, DRL utilizing partially observable Markov decision processes (POMDP) and DRL utilizing multi-agent (MA) approaches. These techniques are anticipated to be greatly beneficial in addressing the unresolved challenges in D2D-C.

2.2.1 Basic DRL Classifications

Over the years, a multitude of distinct DRL algorithms have been created to meet the varied demands of different D2D-C applications. The classification of DRL algorithms as value-based

or policy-based is determined by the extent to which the algorithm prioritizes reward or policy.

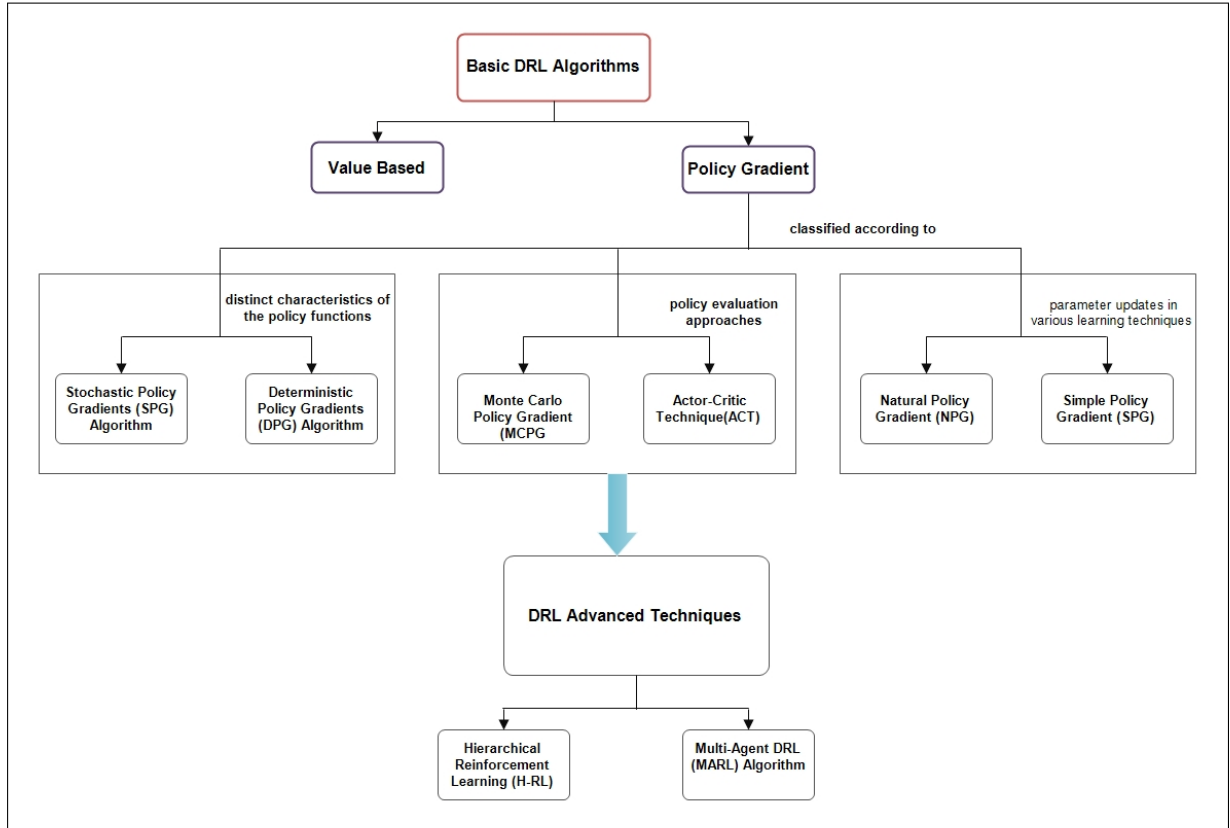


Figure 2.4: Basic DRL Classifications.

2.2.1.1 VB DRL Algorithms

VB algorithms strive to approximate the VF with the utmost precision. This particular function plays a vital role in determining the expected future benefit of either a state or a state-action combination. In the context of DRL, value-based methods utilize a particular approach as demonstrated in Fig. 2.5 [20]. Specifically, the inputs to NNs consist of either the states s^t or the pair of state-action (s^t, a^t) . Meanwhile, the QF $Q_\pi(s^t, a^t)$ or VF $V_\pi^F(s^t)$ are estimated through the utilization of parameters ϕ in said NNs. A NN is capable of returning the estimated value of either the QFs or the VFs for the input states or state-action pairs. The depicted Fig. 2.5 showcases the possibility of one output neuron or numerous output neurons. For a singular output neuron, the outcome can potentially manifest as either $V_\pi^F(s^t)$ or $Q_\pi(s^t, a^t)$, contingent upon whether the input is constituted of s^t or state-action pairs (s^t, a^t) . In the latter scenario,

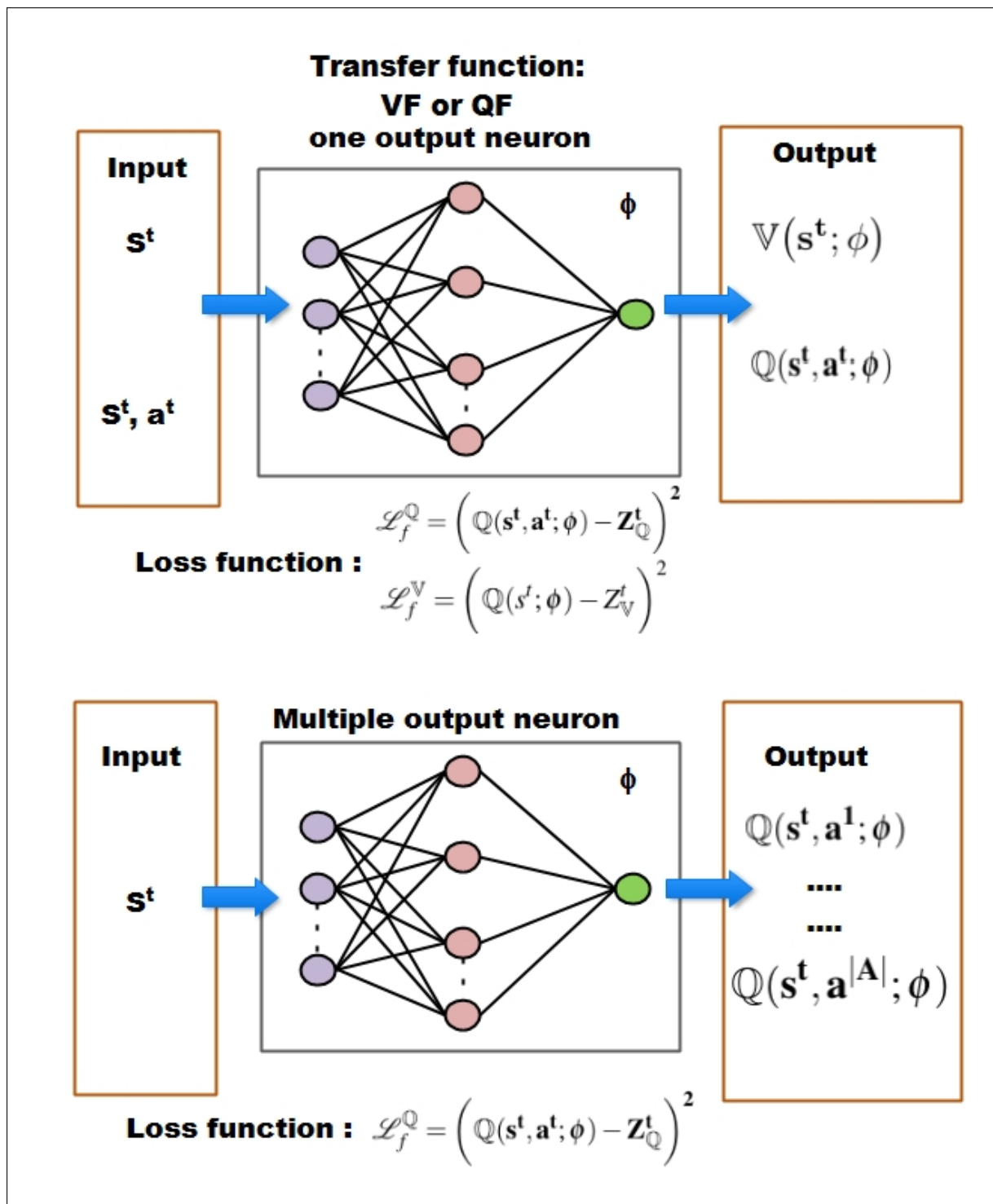


Figure 2.5: Generalized Value-Based Approaches for DRL.

the resulting outputs consist of the QFs for the state s^t in combination with each action. The output can be represented by $Q_\pi(s^t, a^1) \dots Q_\pi(s^t, a^A)$.

In order to derive the loss functions, it is necessary to define Z_Q^t and Z_V^t as the target values

of QFs and VFs, respectively. The loss function for QF is defined as follows:

$$\mathcal{L}_f^{\mathbb{Q}} = \left(\mathbb{Q}(s^t, a^t; \phi) - Z_{\mathbb{Q}}^t \right)^2 \quad (2.7)$$

Similarly, The loss function for VF is defined as follows:

$$\mathcal{L}_f^{\mathbb{V}} = \left(\mathbb{V}(s^t; \phi) - Z_{\mathbb{V}}^t \right)^2 \quad (2.8)$$

When employing value-based methodologies, these loss functions can be leveraged to evaluate the proficiency with which the neural network approximates QFs or VFs.

2.2.1.1.1 Deep Q-networks (DQN) Algorithm The formulation of the DQN algorithm was based on the concept of NN fitted QFs, with the primary objective of achieving exceptional proficiency in ATARI games [23]. The NN incorporated in DQN receives a state as its input and generates an estimate of QFs for all possible actions within the given state, as shown in Fig. 2.6 [20].

In the DQN framework, the initial step involves the stochastic initialization of network parameters as ϕ_0 . Now, using the Bellman equation, the target QF in DQN is defined as follows:

$$Z_{DQN}^t = \gamma^{t+1} + \beta \max_{a^{t+1}} \mathbb{Q}(s^{t+1}, a^{t+1}; \phi^t) \quad (2.9)$$

where γ and β represent reward and discount factor, respectively.

The updating of the parameters in DQN is achieved through the minimization of the loss function \mathcal{L}_{DQN} . It can be obtained from equation (1) by substituting $Z_{\mathbb{Q}}^t$ with Z_{DQN}^t . Now, through the utilization of the SGD method, the variables are modified as follows:

$$\phi \leftarrow \phi + \beta \left(Z_{DQN}^t - \mathbb{Q}(s^t, a^t; \phi) \right) \nabla_{\phi} \mathbb{Q}(s^t, a^t; \phi). \quad (2.10)$$

In Eq. (4) β represents the learning rate.

Freezing target networks and experience replay are the two crucial methods that are implemented in DQN to overcome the issues of DRL. To avoid the instability and incontestability

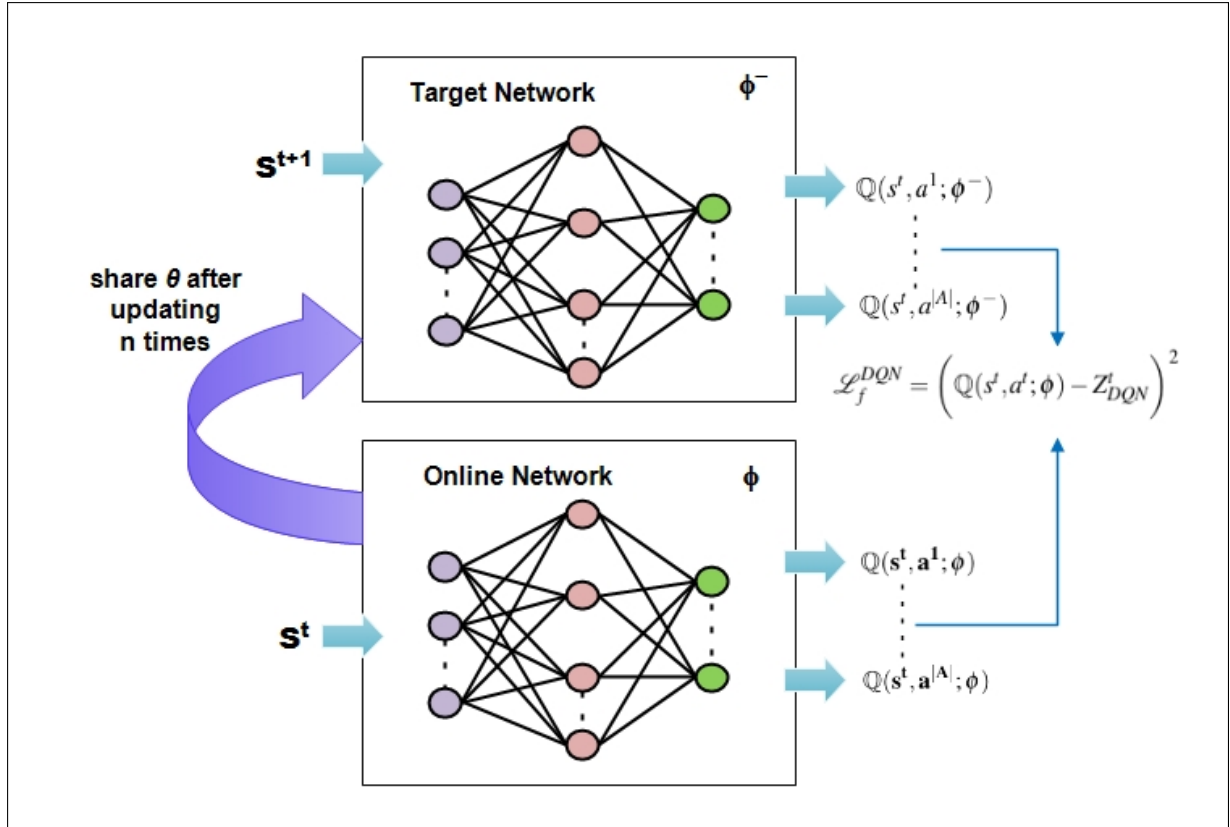


Figure 2.6: DQN with Target Network.

of the training process, the utilization of target networks is employed. The parameters, specifically ϕ^- , are maintained at a fixed state for a designated time period. These target networks are utilized to assess the QF of the subsequent state. Therefore, Eq. (3) is modified as follows:

$$Z_{DQN}^t = \gamma^{t+1} + \beta \max_{a^{t+1}} Q(s^{t+1}, a^{t+1}; \phi_-^t). \quad (2.11)$$

The variables pertaining to the internet-based network, denoted by ϕ_-^t , undergo modification following each iteration. Upon reaching a specified number of iterations, the online network proceeds to disseminate its variables to the target network. This practice mitigates the likelihood of divergence and serves as a safeguard against the instabilities that can arise from excessively rapid propagation.

In the method of experience replay, the agent's data set, comprising its experiences, is stored in a memory buffer. Subsequently, updates are executed on the aforementioned data set, thereby eliminating interdependent relationships in the observation sequence while simultaneously mit-

igating variances in the data set distribution. The utilization of this methodology enables the inclusion of modifications that cover a vast array of state-action spaces, thereby providing a greater potential for the implementation of more significant parameter updates.

2.2.1.1.2 Double DQN (DDQN) Algorithm It uses the target networks (TNs) to assess the QF. Target networks serves the dual purpose of selecting and evaluating an action. However, this approach may result in an overestimation of the QF for a particular action. The DDQN approach was introduced to effectively tackle this issue. This involves utilizing two separate parameter sets to estimate the target value \mathbb{Z}_{DDQN}^t , as illustrated in Fig. 2.7 [24]. Therefore, the target QF in DDQN is re-defined as follows:

$$\mathbb{Z}_{DDQN}^t = \gamma^{t+1} + \beta Q\left(s^{t+1}, \arg \max_{a^{t+1}} Q(s^{t+1}, a^{t+1}; \phi^t); \phi^-\right) \quad (2.12)$$

In (11), the process of selecting an action in the online network is determined by the parameters ϕ , while the assessment of the existing action is based on the variables ϕ^- in the target network. This implies that there will be a reduction in the overestimation of QF values and an increase in stability to enhance the performance of DRL approaches [21]. The DDQN algorithm has the capability to acquire the benefits of double Q-learning, all the while maintaining the remaining structure of the DQN algorithm. In addition to DQN and DDQN, there exist another VB techniques that have been built upon these methods and enhanced them in various ways. Examples of such techniques include DDQN-PP [25] and DDQN-DA [26].

Although the VB DRL approaches have been extensively utilized due to their perceived simplicity and commendable performance, it is worth noting that there exist some limitations associated with these techniques. Firstly, it is noteworthy that value-based methods are constrained in their ability to tackle RL problems with UMC. Secondly, it is incapable of resolving RL challenges wherein the ideal policy is random in nature and needs particular probabilities.

2.2.2 PG Based DRL Algorithm

NN are commonly used in PG approaches for the direct estimation of policy π based on the current state such as $\pi_\phi^t(s^t)$. As illustrated in Fig. 2.8 [20], the NNs receive the states as inputs

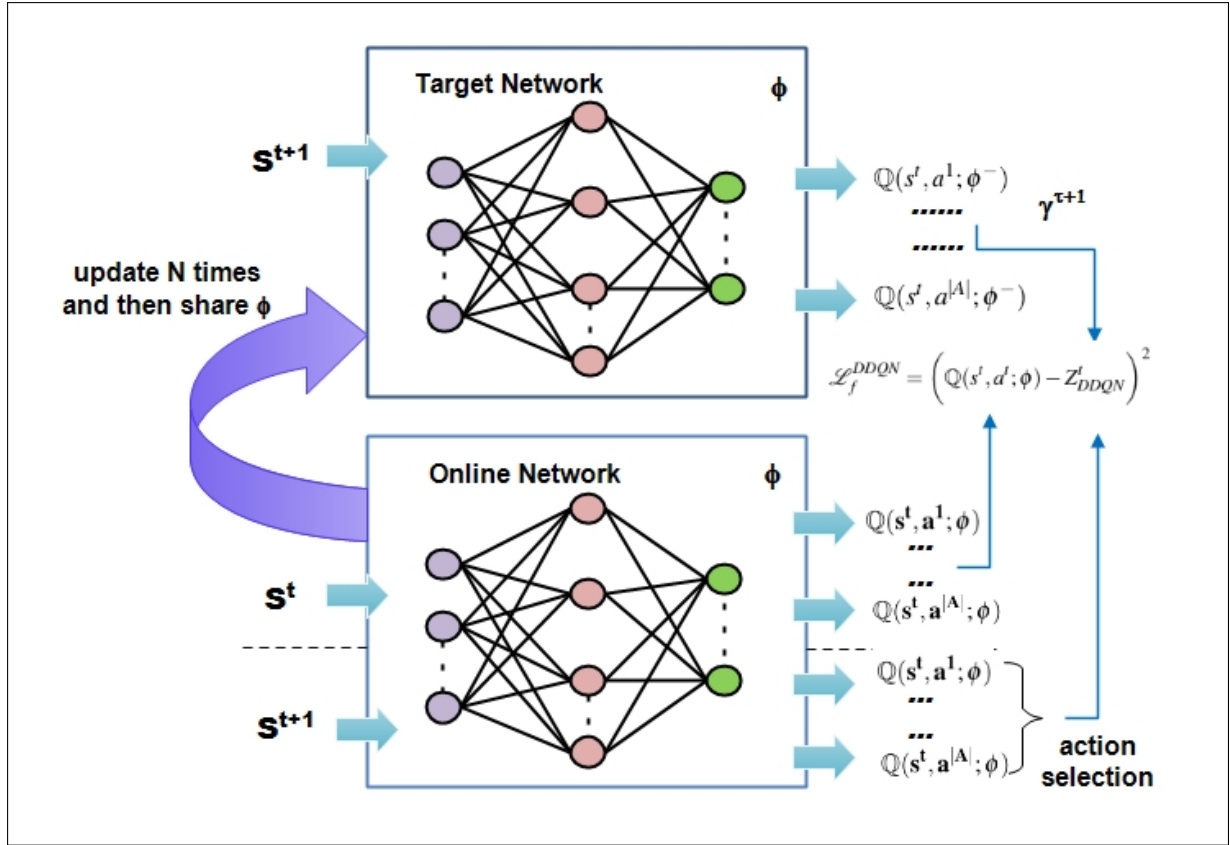


Figure 2.7: DDQN with Target Network.

and approximate the policy through the parameters of the networks. The objective function is utilized to assess the effectiveness of the existing policy and can be expressed as:

$$\mathbb{J}(\phi) = \mathbb{V}_f^\pi(s^0) = \mathbb{E}_{u_{s^0} \sim \pi} [A(u_{s^0})], \forall s_0 \in \mathcal{S}, \quad (2.13)$$

The fundamental concept underlying policy gradient approaches is to modify the parameters towards an increase in the predicted reward [27]. Now the loss function can be modified as follows:

$$\mathcal{L}^P(\phi) = -\mathbb{J}(\phi) = -\mathbb{V}^{\pi_\phi}(s_0) \quad (2.14)$$

For updating the parameters, it is necessary to represent the gradient of \mathbb{J} in relation to the parameter as a prediction derived from stochastic estimation based on (7). The Policy Gradient algorithm can be categorized into the following classifications.

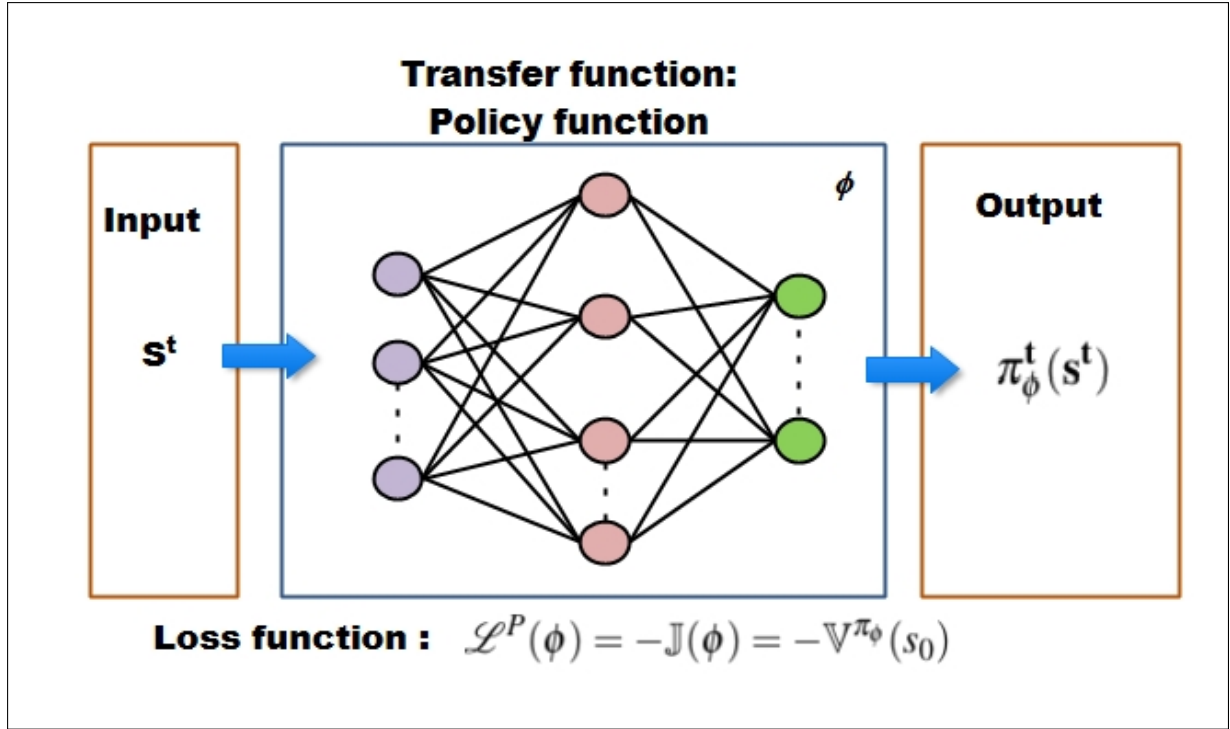


Figure 2.8: DRL Techniques Based on PG Methods.

2.2.2.1 Stochastic Policy Gradients (SPG) Algorithm

When DRL is implemented, the approximation of a SPG takes the form of $\pi_\phi = \pi(a^t | s^t; \phi)$. This enables the determination of the likelihood of a particular action being taken by the agent in a particular state s , when said agent go behind the policy that is specified by ϕ . The typically employed parameters of policy are both the weights as well as the bias of a particular NN [21]. When states and actions have continuous values, the Gaussian distribution is often utilized to proficiently describe the policy for a DRL framework. When states and actions have discrete values, the Soft max function is often utilized to proficiently describe the policy. An ANN is utilized to estimate the average value. On the contrary, the standard deviation of the Gaussian distribution is determined by the group of parameters [28]- [29]. Using theorem of PG, we can define:

$$\nabla \mathbb{E}_{u_{s^0} \sim \pi} [A(u_{s^0})] = \mathbb{E}_{u_{s^t} \sim \pi} [A(u_{s^t}) \nabla_\phi \log \pi(a^t | s^t; \phi)] \quad (2.15)$$

The parameters of the NN are modified through the utilization of stochastic gradient descent in the following manner:

$$\phi \leftarrow \phi + \beta A(u_{s^t}) \nabla_{\phi} \log \pi(a^t | s^t; \phi), \quad (2.16)$$

where β is the parameter ϕ 's learning-rate. Now, the SPG algorithm's loss function can be described as:

$$\mathcal{L}^{SPG}(\phi) = -A(u_{s^t}) \log \pi(a^t | s^t; \phi) \quad (2.17)$$

2.2.2.2 Deterministic Policy Gradients (DPG) Algorithm

The SPG algorithm represents the policy as a function of the probability distribution across actions. In contrast, the DPG model represents the policy as a deterministic decision and represented by the following condition $\pi_{\phi} = \pi(s^t; \phi)$. Using (7) and as per the DPG theorem, we can define:

$$\nabla_{\phi} \mathbb{J}(\phi) = \mathbb{E}_{s \sim \rho^{\pi_{\phi}}} \left[\nabla_{\phi} \pi(s^t; \phi) \nabla_a Q^{\pi_{\phi}}(s^t, a^t, \psi) |_{a=\pi(s^t; \phi)} \right]. \quad (2.18)$$

In (17), policy enhancement is achieved by breaking it down into two components: QF's gradient concerning actions, and policy's gradient in relation to the parameters of policy. $\rho^{\pi_{\phi}}$ represents the distribution of states that result from implementing policy π_{ϕ} . Therefore, updated parameters are expressed as follows:

$$\phi \leftarrow \phi + \beta \left[\nabla_{\phi} \pi(s^t; \phi) \nabla_a Q^{\pi_{\phi}}(s^t, a^t, \psi) |_{a=\pi(s^t; \phi)} \right]. \quad (2.19)$$

Now, the DPG algorithm's loss function can be expressed as follows:

$$\mathcal{L}^{DPG}(\phi) = -\pi(s^t; \phi) \nabla_a Q^{\pi_{\phi}}(s^t, a^t, \psi) |_{a=\pi(s^t; \phi)} \quad (2.20)$$

2.2.2.3 Monte Carlo Policy Gradient(MCPG) Algorithm

The fundamental architecture of MCPG approach is shown in Fig. 2.9 [20]. The goal of the MCPGA is to use MC simulation to determine the value of $A(\mu_{s^t})$. The RL algorithm proposed

in [30] is a prime example of the typical MCPG algorithm utilized in SPG methods. Using the MCPG approach, we first sample a trajectory μ_{s^0} by executing the current policy from a beginning state s^0 . After that the cumulative reward $A(\mu_{s^t})$ is computed for every time step. The cumulative reward is now multiplied with the PG $\nabla_{\phi} \log \pi(a^t | s^t; \phi)$ in order to effectively update the parameters ϕ as specified in (15). The aforementioned process is iterated through several runs, where each run entails sampling a distinct trajectory. In addition, to decrease the PG's variance, a baseline function $b(s^t)$ that is not reliant on a^t is incorporated. therefore, loss function for MCPG algorithm can be rewritten as:

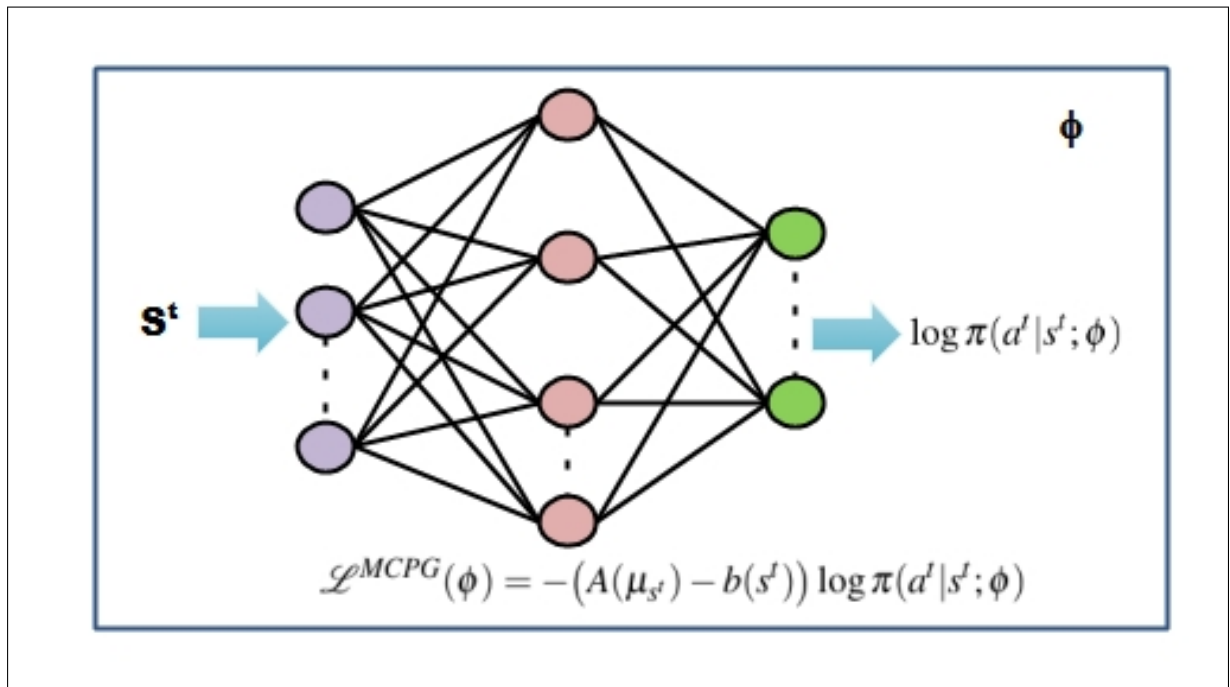


Figure 2.9: MCPG Method.

$$\mathcal{L}^{MCPG}(\phi) = -(A(\mu_{s^t}) - b(s^t)) \log \pi(a^t | s^t; \phi) \quad (2.21)$$

The MCPG algorithm in DRL offers a seamless correspondence between a particular state and specific action, resulting in superior convergence of the algorithm and enhanced efficiency in large continuous state and action spaces as compared to value-based DRL approaches. Additionally, it has the ability to learn SPG, which exhibits superior performance over DPG in certain scenarios. MCPG approaches, unfortunately, encounter issues with estimations that

have high variance.

2.2.2.4 Algorithm Based on Natural Policy Gradient (NPG)

All of the PG techniques that were previously described rely on updating the parameters of the NN utilizing a SG of the loss function $\nabla_{\phi} \mathcal{L}(\phi)$. The NPG techniques, on the other hand, use the natural gradient $\nabla_{\phi}^N \mathcal{L}(\phi)$ to modify the NN parameters. Therefore, NPG techniques provide a more efficient solution compared to the SPG technique [28]. The loss function utilized in NPG is equivalent to that utilized in SPG, with a generalized expression provided in (13). Therefore, the updated value of the parameter can be estimated as:

$$\phi \leftarrow \phi + F_{\phi}^{-1} \nabla_{\phi} \nabla^{\pi} \phi(s). \quad (2.22)$$

In (21), the symbol F_{ϕ} denotes the Fisher information matrix. This matrix serves the purpose of determining the appropriate step size for the update of NN parameters [33] and defined as:

$$F_{\phi} = \mathbb{E}_{\pi_{\phi}} \left[\nabla_{\phi} \log \pi(a^t | s^t; \phi) (\nabla_{\phi} \log \pi(a^t | s^t; \phi))^T \right]. \quad (2.23)$$

The NPG techniques introduces a novel step size format that determines the appropriate amount of adjustment for the parameters, resulting in a more consistent and efficient update process. However, the disadvantage of utilizing NPG arises when a complex NN is employed to estimate the policy with an extensive range of parameters. In such cases, it becomes impracticable to compute the Fisher information matrix or adequately storing them [21]. The upgrade techniques utilized by NPG provide a partial solution to the aforementioned problem, and are widely employed in practical DRL applications. Examples of such techniques include TRPO [29] and PPO [31].

2.2.3 DRL Advanced Techniques

2.2.3.1 Actor-Critic Technique(ACT)

ACTs are renowned for their ability to merge the benefits of MCPG and value-based approaches. These methods have been extensively researched in the field of DRL. As depicted in Fig. 2.10 [20], the implementation of an ACT usually involves a pair of NNs, specifically the actor network (AC-N) and the critic network (CR-N), that possess shared parameters. The AC-N bears resemblance to the NN employed in the PG approach, whereas the CR-N exhibits similarity to the VB approach's NN. Throughout the training procedure, the CR-N modifies the parameters of VFs ψ in accordance with the AC-N's policy. In the meantime, the AC-N modifies the policy parameters (ϕ) based on the VFs assessed by the critic. Usually, it is necessary to predefine two distinct learning rates for the updates of (ψ) and (ϕ) [36]. Now, the CR-N's loss function can be expressed as:

$$\mathcal{L}^{CR-N}(\psi) = \vartheta_t^2 \quad (2.24)$$

In (23), ϑ represents the TD error that is utilized to rectify the CR-N. This value can be computed by:

$$\vartheta_t = \gamma^{t+1} + \Gamma \nabla \pi^\phi(s^{t+1}; \psi) - \nabla \pi^\phi(s^t; \psi) \quad (2.25)$$

The updated parameters of CR-N are expressed as follows:

$$\psi \leftarrow \psi + \beta \vartheta_t \nabla \pi^\phi(s^t, \psi) \quad (2.26)$$

where β is the CR-N's learning-rate.

Now, the AC-N's loss function can be described as:

$$\mathcal{L}^{AC-N}(\phi) = -\vartheta_t \log \pi(a^t | s^t; \phi) \quad (2.27)$$

The updated parameters of AC-N are expressed as follows:

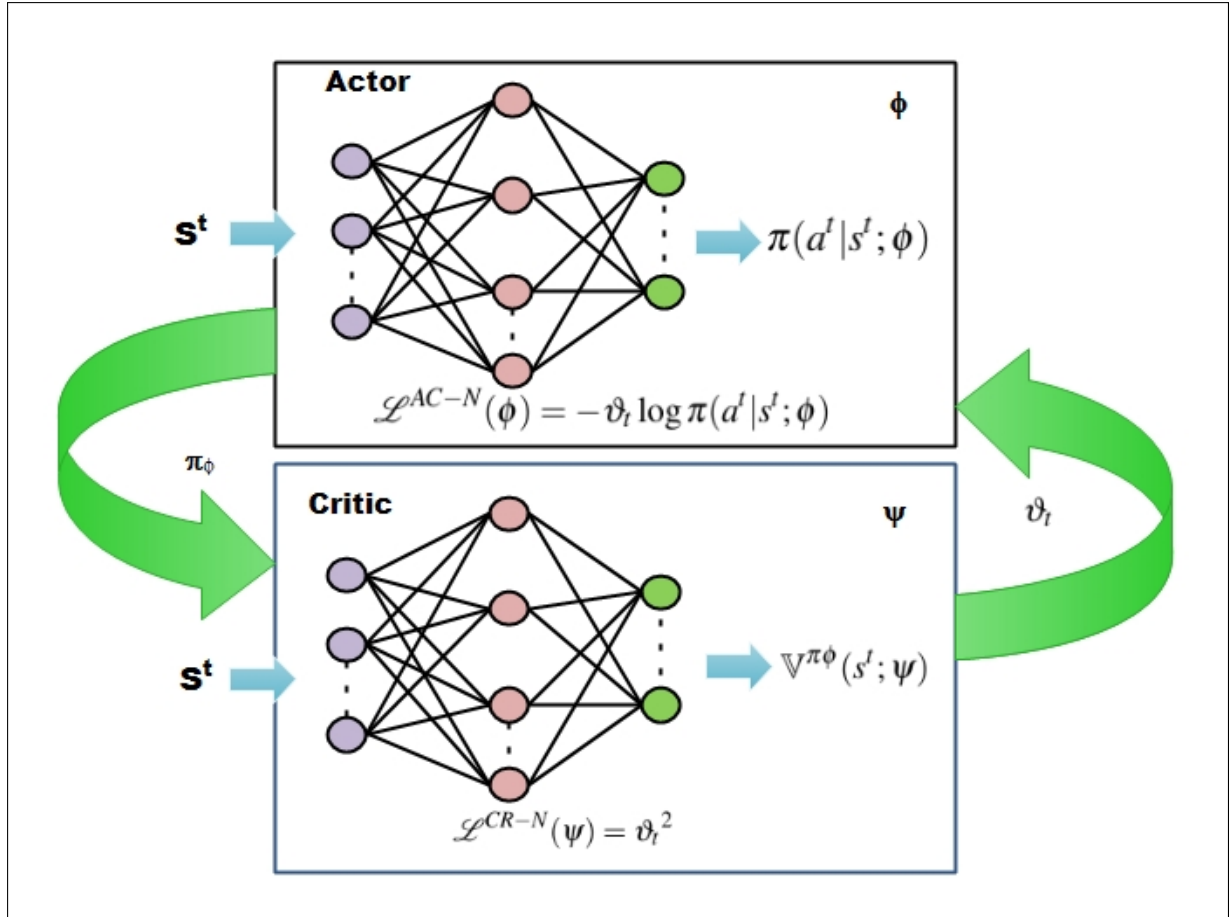


Figure 2.10: Actor-Critic Method.

$$\phi \leftarrow \phi + \beta \vartheta_t \nabla_\phi \log \pi(a^t | s^t; \phi), \quad (2.28)$$

where β is the AC-N's learning-rate.

When comparing the VB DRL structure to the AR/CR structure, it becomes evident that the latter is more suitable for addressing MDP issues that have a huge set of states and actions. Furthermore, the utilization of the CR-N can result in a decrease in variance and a higher level of sample efficiency. However, because of the repetitive usage of value estimations, AC/CR approaches are prone to instability. The AC/CR architecture, as a matter of fact, has been extensively embraced as the fundamental framework of subsequent algorithms, including D4PG [32], TD3 [33], MADDPG [34], RDPG [35]. Actor-critic [36], A2C, A3C [37], DPG [38], DPG [39].

2.2.3.2 Hierarchical Reinforcement Learning (H-RL)

When dealing with complex D2D-C systems that have limited feedback or huge set of states and actions, the utilization of the aforementioned DRL algorithms may result in significant scaling problems. The DRL algorithm extensively depends on the reward function for the purpose of updating the policy. The Semi-Aggregated Markov Decision Process (SAMDP)

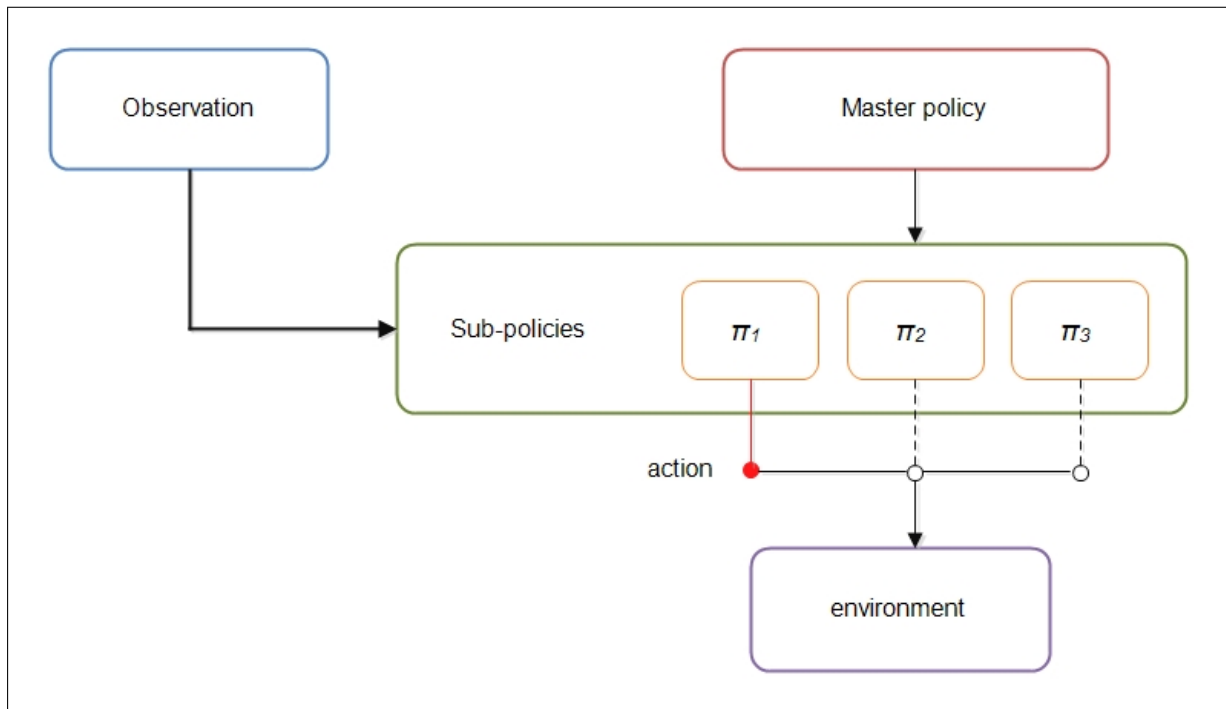


Figure 2.11: An example of a H-RL approach, wherein sub-policy π_1 is chosen to generate the environmental action.

framework has been proposed in order to mitigate the challenges posed by training complexity. SAMDP modeling is designed to enhance the comprehension of intricate behaviors by detecting temporal and spatial abstractions. The SAMDP modeling technique has been developed with HRL in order to improve the understanding of complex behaviors through the identification of temporal and spatial abstractions. Unlike other modeling techniques, SAMDP is developed in a transformed state-space that encodes problem dynamics. Unlike other modeling techniques, SAMDP has been developed using a transformed state-space that effectively encodes the problem dynamics. As a result, the framework of HRL was introduced, which involved acquiring sub-policies with either temporal or behavioral abstraction [7]. The fundamental concept behind HRL is to expand the range of actions to macro-actions that enable

high-level management of abstract objectives or long-term time-frames. The set of actions that are accessible at the lower level is determined by the high level policies and controls. Particularly, the environment's actions are produced by the policies and controls at the lowest level, similar to conventional DRL algorithms. Fig. 2.11 [22] shows that the primary policy chooses sub-policy π_1 to produce the environment's action.

HRL algorithm provides various advantages as it divide the main policy into several sub-policies. Primarily, it is important to note that problems related to decision-making can be combined and arranged in a hierarchical manner. This characteristic presents a promising avenue for the division of policies [7]. Additionally, a detailed and well-planned investigation that employs various sub-policies can greatly improve the ability to achieve generalization convergence, particularly in scenarios where rewards are scarce. This is due to the fact that a policy operating at a lower level is able to acquire knowledge and improve its performance by leveraging the intrinsic rewards offered by a higher-level policy, as opposed to relying solely on the infrequent rewards provided by the environment. Finally, HRL enables the utilization of transfer learning across diverse sub-policies to attain optimal learning outcomes .

2.2.3.3 Multi-Agent DRL (MA-DRL) Algorithm

The majority of the aforementioned algorithms have been constructed based on MDP framework, whereby the environmental states are readily accessible. However, in numerous practical issues, certain state information may be confidential or obtaining global information necessitates collaboration and communication among multiple agents, ultimately leading to significant latency. The use of conventional centralized approaches for POMDPs is impracticable In this particular scenario. As depicted in Fig 2.12 [22], the states of the environment remain concealed and DRL agents are limited to making decisions solely on the basis of their observations. The difficulty in solving POMDP arises due to the fact that it is notoriously challenging to solve, owing to its unobservability. MA-DRL is a significant sub field of RL that combines game theory approaches to RL approach and emphasizes on the enhancement of long-term performance in POMDPs [40]. To address the challenge of scaling, the MA-DRL algorithm transfers centralized control among several agents. Therefore, it reduces the burden on a single

agent. Recent MA-DRL research predominantly concentrates on attaining effective coordination amongst various RL agents and reaching a state of equilibrium. Generally, it can be asserted that MA-DRL poses significant technical and theoretical difficulties. Moreover, in addition to the aforementioned challenge related to inefficient learning, there exist novel issues in the domain of MA-DRL. For example, every individual agent within a given locality holds the potential to collaborate, contradict, and exert an impact on other agents, ultimately leading to a dynamic and ever-changing environment with a lack of consistency in performance. A policy that seems effective currently may not necessarily yield the desired outcomes in the future due

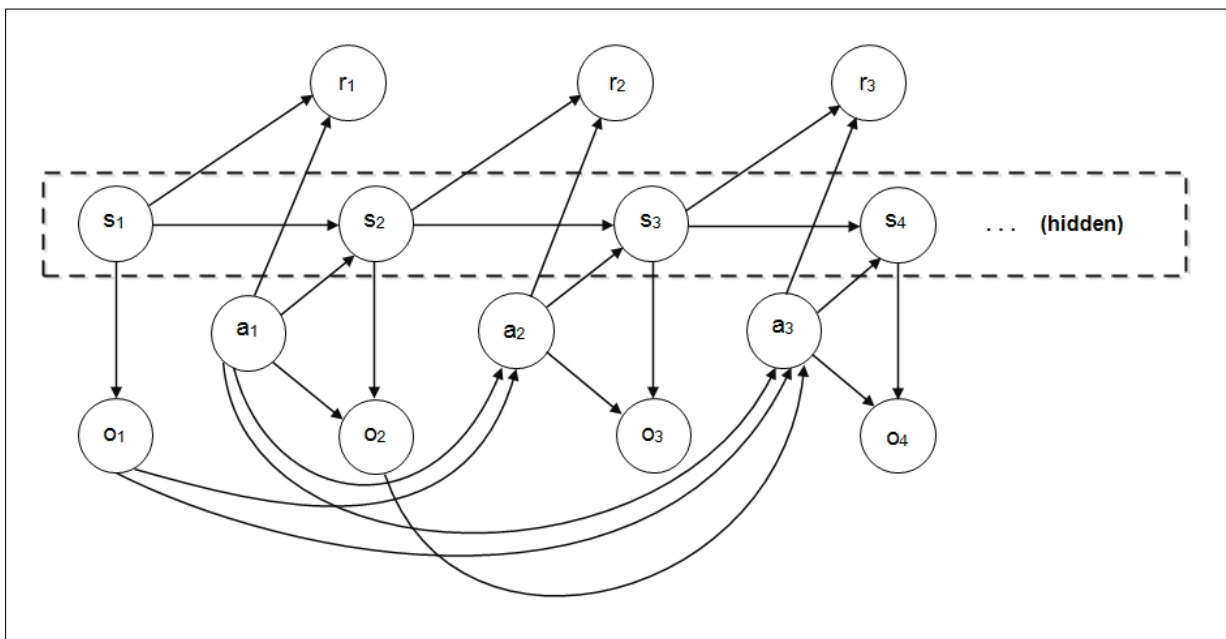


Figure 2.12: An exemplification of the POMDP model is provided, wherein an arrow is sketched from the reliant category to the corresponding dependence, signifying the perceivable information of the surroundings.

to updates made by other agents. As a result, achieving equilibrium among several agents can be a challenging task for MARL. Despite this, the MARL algorithm demonstrates significant potential for addressing complex MDPs as well as POMDPs with high dimensions, particularly in various D2D applications. Recently, certain investigations have attempted to fully capitalize on the potential of MARL in tackling POMDP issues. These efforts include the exploration of mean field MA=DRL [41] and MADDPG [42].

2.3 Deep Reinforcement Learning (DRL)

In this particular segment, we shall elaborate on the various variants of DRL techniques that have been implemented in the context of D2D-C and RIS. DRL is an efficient technique for embedded optimization and has the ability to take immediate action in WCNs. In DRL approaches, NN are trained off-line before being deployed. It utilizes the trained framework for computing resource allocation with relatively low complexity, to accurately assess the optimize transmit power management.

2.3.1 DRL with D2D-C

In D2D-C technology, two neighboring devices can share the data directly without the base station (BS). As a result, it enhances mobile users' quality of service by reducing the transmission delay. Also, in D2D-C, the D2D pairs (DDPs) reuse the same resources as used by CUs to boost the spectral efficiency. Despite these advantages, key challenges such as CR-CI and CO-CI, as well as ultra-massive connectivity (UMC), need to be investigated more for QoS provision to the end users. Various DRL-based schemes have been proposed to fully leverage the advantages of D2D with DRL. These schemes will be discussed in detail below.

2.3.1.1 FC-MAQ (Fuzzy clustering multi-agent Q-learning)

The objective of the author in this paper [43] was to maximize the throughput of D2D underlay cellular network via power control of D2D transmitters. To achieve this aim, the author initially grouped D2D users with significant differences in eigenvalues together. Subsequently, the problem was modeled as the RL problem, and each user group was treated as an agent. Finally, the combination of the MARL and the fuzzy clustering algorithm has been established to maximize the throughput. The simulation outcome indicates that the suggested algorithm outperforms in terms of MAQ, OP, and OCP schemes.

2.3.1.2 S-DDPG (Sharing-deep deterministic policy gradient)

In this article [44], the author presented two innovative strategies utilizing DDPG Scheme to effectively address the issues of distributed learning and power allocation in a D2D-based V2V wireless communication network. Through adjusting the input of the neural network, it becomes possible for all agents within the MA-DRL algorithm to employ a singular actor network as well as a single critic network. This technique results in improved performance and faster convergence, while also reducing computational complexity and memory storage needed drastically. Finally, upon completion of training the policy NN, the non-cooperative power allocation problem can be solved expeditiously in mere milliseconds. The experimental findings demonstrate exceptional potential compared to other conventional DRL methodologies.

2.3.1.3 MAAC-NAAC DRL (Multi-agent actor critic, neighbor-agent actor critic-DRL)

In this paper [45], the author presented two innovative frameworks for distributed spectrum allocation, namely MAAC and NAAC. These frameworks are trained in a centralized manner and executed in a distributed fashion. The MAAC framework effectively alleviates the instability and convergence challenges that arise in multi-agent environments during training by sharing the historical states, actions, and policies of all users in centralized training. On the other hand, the convergence of the training has been ensured by NAAC, which has resulted in a reduction of computing complexity. This approach is highly appropriate for a range of complex and dynamic communication scenarios. The experimental results show that the proposed scheme has superior performance in comparison to existing DRL techniques such as Q-learning, Actor-Critic (AC) and DQN.

2.3.1.4 ICWN-DRL (Information-centric wireless networking-DRL)

The author of this article [46] has delved into a pioneering approach to address the issue of allocating resources and controlling power in a fully integrated D2D-enabled MEC system by means of the policy gradient approach. To address the problem, the initial step involves formulating it as a model-free reinforcement learning challenge. Then the policy gradient method is utilized to regularly update the parameters of the NN so as to adeptly handle the variable trans-

mission power as well as uncertain channel conditions. Experimental results indicate that the method demonstrates excellent convergence when compared to the state-of-the-art schemes.

2.3.1.5 NC-DRL (Non cooperative-DRL)

Authors investigated the problem of power allocation in D2D communication wireless network and proposed a NC-DRL technique in this paper [47]. The NC-DRL algorithm relies on three distinct MARL algorithms, namely DQN, DDQN, and dueling deep Q-networks algorithms. In NC-DRL scheme, each DDT has the ability to optimize its power usage for transmitting information, which enables it to seamlessly adapt to the ever-changing dynamics of its environment. After a specific duration, all DDPs disseminate their tactics to their peers with the aim of assessing the network's performance. The outcomes exhibit great potential in regards to the scalability towards vast domains, exceptional performance, and reduced execution time.

2.3.1.6 ST-DRL (Social Trust-DRL)

In this article [48], the authors presented a ST scheme that effectively improves the security of mobile social networks. The ST scheme relies on DRL, wherein an agent is provided with a range of observations from the integrated network. These observations encompass the wireless channel conditions, the trust value of each node, the contents present in the cache, and the available computational capacity. The agent transmits these parameters to the DNN which then generates the most favorable actions. The operator's earnings are monitored and transmitted back to the agent as a form of reward. The iteration of the process continues until the attainment of optimal actions. The experimental findings demonstrate exceptional potential compared to other conventional DRL techniques.

2.3.1.7 C-MADRL-PER (Coordinated-multi-agent DRL-prioritized experience replay)

The authors conducted research on a joint RB and power allocation issue with the objective of enhancing the network's EE [49]. Additionally, they considered the URLLC, minimum data rate, and interference limitations in this paper. A distributed C-MADRL-PER has been proposed to address this issue. It ensures the strict reliability as well as latency requirements

of URLLC services in a social-aware environment. A QoS-aware reward function has been constructed to incorporate the overall EE and QoS demand during the learning phase. PER and coordinated learning have been implemented to enhance the efficiency of learning and accelerate the rate of convergence. Experimental results show that this approach surpasses other existing methodologies in meeting both the EE and QoS demands.

2.3.1.8 CSSCA-DRL (Constrained stochastic successive convex approximation-DRL)

In this paper [50], the authors explore the problems of joint task assignment, power control, and resource assignment in D2D transcoding systems enabled by blockchain technology. To address the concern, the authors initially created an assessment mechanism for transcoder selection, which is known as a transcoder evaluation mechanism. The main purpose of this mechanism is to provide a means of measuring the effectiveness of transcoder selection. In order to obtain the combined transcoder allocation along with the task assignment, dynamic power control, and RB assignment, authors designed a two-stage CSSCA-DRL algorithm for every epoch. According to experimental results, the suggested algorithm can successfully manage a variety of dynamic scenarios and produce high transcoding earnings while also satisfying the QoS criteria.

2.3.1.9 TP-DQN (Two parallel-deep Q networks)

This paper [51] focused on achieving maximum EE in an underlay D2D wireless communication network, while simultaneously fulfilling the system throughput constraints and QoS requirements of DDPs and CUES. To attain this objective, the authors first introduce an algorithm for dynamic power optimization that is based on DRL and incorporates dynamic rewards as well. Furthermore, a TP-DQN algorithm has been developed with the objective of optimizing the energy efficiency of the network under consideration. The proposed scheme yields a greater EE, all the while ensuring network throughput demands are met.

2.3.1.10 SRPR-DRL (Successive rounding and power refinement-DRL)

In this paper [52], the authors have undertaken a thorough investigation of the challenging issue of joint channel selection and power control for D2D wireless networks. The main objective of this investigation is to optimize the WSR, thereby enhancing the overall efficiency of the network. The authors initially devised an FP-based algorithm to address the problem. However, this algorithm necessitates instantaneous global network information, which renders it non-scalable. To address this challenge, a novel SRPR-DRL has been introduced for each D2D pair to acquire insights into the interrelationship between various network data. The simulation outcomes have illustrated that the proposed SRPR-DRL approach can attain a similar level of efficiency to that of the FP-based algorithm, even in the absence of immediate global CSI.

2.3.1.11 MSRA-DDPG (Mode selection and resource allocation DDPG)

In this article [53], the authors aimed to optimize the energy efficiency of all user (DDPs and CUEs) while simultaneously ensuring that the QoS needs of all users are met. However, transmission power, channel links, and RB allocation are all continuous values. In order to effectively handle these continuous variables, the authors have proposed the use of the MSRA-DDPG algorithm. This algorithm is specifically designed to identify the optimal policy within a continuous state and action space and utilizes the actor-critic approach. Simulation results indicate that the proposed MSRA-DDPG algorithm exhibits superior convergence characteristics and enhanced energy efficiency when compared to state-of-the-art schemes.

2.3.1.12 D-MARL (Deep-multi-agent RL)

The authors investigated the challenging issues of subcarrier assignment and power allocation jointly [54]. To address the challenging issue, they proposed D-MARL algorithm. The proposed algorithm relies on delayed CSI and QoS feedback from specific neighbors, thus leading to a significant reduction in the signaling overhead. Moreover, a remarkable characteristic of the D-MARL is its ability to direct self-interested agents towards accomplishing a comprehensive design objective through a series of observations of state-action-reward while

interacting with the existing environment. The experimental results demonstrated that DDPs possess the ability to acquire near-optimal SE, while keeping to strict interference temperature restrictions.

2.3.1.13 ADMM-DRL (Alternating direction method of multipliers based DRL)

This article [55] studied resource gathering and slicing in D2D assisted V2V network. The problem was tackled by the authors through their proposed framework, which utilized DRL technique. The three-stage layered framework of the slicing technique was developed to address the issues of resource gathering, and resource allocation. Its ultimate goal is to significantly enhance the total throughput of D2D links. Additionally, the author has developed a DRL algorithm based on ADMM, which provides faster convergence speed. In addition, simulation findings showed that the proposed algorithm are capable of enhancing resource utilization, improving slice satisfaction, and yielding larger throughput as compared to existing techniques.

2.3.1.14 C-DRL (Centralized-DRL)

This proposal [56] addressed the problem of power allocation in a dynamic environment within a D2D wireless network. Aiming to address this issue, the authors introduced a C-DRL to tackle the problem of power allocation in a D2D network in a dynamic scenario. The C-DRL acquires power control by optimizing network performance and quality of user experience, with special emphasis on the actual dynamic channels and DDT interference. The results of the simulation indicate that the proposed technique displays superior performance in comparison to the conventional method.

2.3.1.15 SOTPSR-DRL (Self-organization of transmission power and power splitting-DRL)

This proposal [57] examined the management of power in a wireless D2D underlaid cellular network in order to maximize EE. The approach involves numerous DDPs for implementing the SWIPT technique with a power split strategy. To address this particular issue, the authors

initially employed optimization-based iterative approaches, namely exhaustive search and gradient search with barriers, to identify the global optimum as well as the suboptimum. The SOTPSR-DRL approach is proposed to enable autonomous adjustment of the transmit power and power splitting ratio, thereby achieving energy efficiency optimization of the system.

2.3.1.16 BDRFL (Double-layer blockchain-based deep reinforcement federated learning)

The authors proposed a BDRFL scheme to handle the issues of data confidentiality and network security in D2D caching wireless network [58]. BDRFL scheme is a FL framework, which is further strengthened by a sophisticated double-layer blockchain system. This FL framework enables users to train models in a decentralized manner without the need for exchanging raw data. Moreover, the author devised a method to modify the FL parameters for both local and area models as well as the global model. The simulation results initially confirmed the convergence of the proposed algorithm. Subsequently, they exhibited that the caching scheme based on BDRFL led to a reduction in download latency.

2.3.1.17 D4SA (Double deep Q-network based D2D spectrum access)

The authors of this study [59] proposed a cutting-edge hybrid spectrum accessing technique that takes into account both orthogonal and non-orthogonal access scenario for D2D-C. The authors examined two different ways in which spectrum access could be utilized for D2D-C, based on the positioning of CUEs. Additionally, they introduced an "accessible area" as a means to ensure that the QoE for CUs located at the edge of the cell is maintained at a high level. The D4SA scheme proposed facilitates D2D pairs to independently acquire a supreme spectrum access policy that maximizes the aggregate throughput, even in the absence of any prior information. Since the allocation of resources among D2D pairs is subject to stochastic and unfair conditions, a fairness utility function has been devised to guide the agent in achieving maximum sum throughput as well as the fairness among DDPs.

2.3.1.18 AW-FDRL (Attention weighted federated -DRL)

In this article [60], the authors conducted an investigation into the matter of D2D-enabled Het-net cooperative edge caching in cellular network. They transformed the issue of joint node selection and cache replacing as a long-term mixed integer linear programming (LT-MILP) issue. Authors used the DQN framework to dynamically govern the joint decision-making procedure by considering both the network state and past data. It trains the DQN framework in a distributed fashion by retaining the information in the local users, while simultaneously dealing with the challenge of model aggregation among Het-net users. Simulation findings indicate that the proposed AWFDR framework can significantly enhance hit rate, decrease average delay, and alleviate traffic, relative to the current approaches.

2.3.1.19 CO-DDDPG (Conventional optimization scheme with DDPG)

This paper [61] aimed to explore the optimization of sum-rate and fairness in the context of NOMA-enabled CUs and DDPs, while taking into account the resource and power limitations of both BS and DDT. To address this joint issue, the initial step involves utilizing MDP to transform the formulated optimization problem into a MADRL problem. After that, the authors introduced three DRL techniques, namely multi-DQN, DDDPG, and CODDDPG, as potential solutions for the joint problem. The proposed schemes have been implemented to augment the efficiency of learning, eliminate any superfluous samples, and regulate the power of both the CUs and DDPs. The simulated outcomes have showcased the efficacy of the proposed approach in augmenting the combined network's sum rate. This is achieved by ensuring equal fairness for all users.

2.3.1.20 SGMA-DRL (Stackelberg game guided multi-agent DRL)

The authors investigated the challenging issues of channel allocation and power control jointly [62]. To address the challenging issue, they proposed SGMA-DRL approach. The proposed algorithm has been developed in a distributed manner with the aim of making smart D2D resource allocation judgments in time-varying environments, thereby enhancing network throughput. The utilization of Stackelberg Q-value in the proposed scheme enables a signifi-

cantly quicker convergence in the learning process as compared to the conventional MADRL approach. The simulation results show that the SGMA-DRL has the ability to yield suitable strategies for allocating D2D resources, which can enhance network performance in a time-effective manner.

2.3.1.21 MSPA-DRL (Mode selection integrated with power allocation-DRL)

In this manuscript [63], the authors have undertaken an investigation into the issue of collaborative cache placement in D2D assisted UAV network. To ensure the quality of experience in terms of delays for users, the authors have devised an optimization problem. The problem aims to minimize the total file access latency for users who request files by optimizing the placement of cache in both UAV and D2D users. To address the optimization problem, they presented a MSPA-DRL approach for determining the placement and caching of files. The simulation results demonstrate that the proposed approach has the potential to significantly decrease file access delay when compared to other existing approaches.

2.3.1.22 D3QN-UARA (Double-dueling-deep Qnetwork based joint user association and resource allocation)

This paper [64] presented a novel algorithm for UARA for DDPs in ultra-dense network, with the aim of optimizing sum data rate. Particularly when it comes to DDPs situated in the overlapping regions that exist between adjacent cells, the authors considered the joint problems of user association, subcarrier allocation, and power control of these DDPs. To tackle the problem of optimizing jointly, the authors have introduced an algorithm D3QN-UARA. Simulation results confirmed that proposed algorithm can attain performance that is almost optimal, and it surpasses other schemes.

2.3.1.23 MAOD-DRL (Multi-agent online distributed DRL)

The authors of this paper [65] conducted an investigation into the issue of selection of channels and power control for D2D-C, with a focus on sharing uplink RBs within a cellular network. To address the problem, the authors presented the MAOD-DRL learning algorithm in

DDPs. This algorithm aims to optimize the overall throughput in a dynamic wireless channel environment, all while ensuring the QoS for both DDPs and CUs. Authors established three metrics for assessing the proposed algorithm's effectiveness. The results of the simulation indicate that proposed approach outperforms the DQN-based approach in terms of collision probability, access rate, and time-average network throughput.

2.3.1.24 DSM-DRL (Dynamic spectrum matching-DRL)

The focus of this article [66] was on the dynamic channel accessing technique of DDPs, which aims to optimize the throughput of D2Dus in a D2D underlay cellular uplink network. The issue of matching distributed channels for DDPs is stated as a noncooperative game that is random in nature and involves multiple selfish players. These players are required to discover their optimal strategies by relying on locally observed actions, states, and rewards. A DSM-DRL algorithm is proposed on this foundation, allowing DDU to acquire optimal strategies through online learning. The simulation results demonstrate that the proposed approach exhibits superior performance when compared to other existing approaches.

2.3.1.25 PS-D3QN-DRL (Priority sampling based dueling double deep Q-network-DRL)

The authors investigated the challenging issues of resource allocation and power control jointly [67], which aims to optimize the throughput of D2DUs as well as CUs in a D2D underlay cellular uplink network. To address the challenging issue, they proposed DRL-based PS-D3QN approach. The utilization of PS-D3QN facilitated the acquisition of the predominant features by the agents. This was achieved through the incorporation of both Q-value and V-value during the training process of DRL. The outcomes of the simulation demonstrated that in comparison to the current DRL approaches, the PS-D3QN approach executes more efficient allocation of resources and control of power. As a result, it attains a greater throughput for both D2DUs and CUs.

2.3.1.26 FD-DRA-DRL (Federated learning aided decentralized resource allocation-DRL)

In this proposal [68], a distributed FD-DRA-DRL approach was proposed for D2D-assisted dynamic 5G and beyond networks. The primary objective of this approach was to maximize the sum capacity while simultaneously minimizing total power consumption. Furthermore, the approach must ensure that the QoS demands of D2DUs and CUs was met in order to achieve optimal results. The authors briefly presented three distinct technologies that served to support their proposal. Specifically, D2D-C was found to enhance network performance, while FL facilitated the protection of user privacy. Additionally, DRL afforded users the capacity to enact their own resource allocation strategies depending on the state of the network. The outcomes of the simulation revealed that the proposed algorithm substantially enhanced the network's overall performance in terms of both throughput and power consumption.

2.3.1.27 DSA-DRL-ToT (Dynamic spectrum access- DRL-Internet of Things)

This paper [69] aimed to explore the spectrum sharing and fairness problem in the context of NOMA-enabled CUs and D2Ds in both uplink and downlink scenarios. The authors introduced a DSA-DRL algorithm to address this challenge. In the proposed scheme, D2D nodes was able to autonomously obtain an optimal strategy for effective spectrum utilization, irrespective of any previous knowledge, with the aim of maximizing the normalized sum throughput. Furthermore, authors implemented a DDQN algorithm in order to investigate the fairness of resource allocation among D2DUs. The findings of the simulation indicated that the suggested approach has attained a near-optimal performance and exhibited excellent convergence behavior.

2.3.1.28 UAVs-SWIPT-MADQN-D2D

In this paper [70], the authors examined how to improve energy efficiency and throughput in D2D communication using UAVs with SWIPT. The goal was to increase both EE and sum-rate while using power splitting. To achieve this, a distributed MADQN approach was used and compared to traditional methods like DDQN and DQN. The simulation findings revealed that the proposed algorithm had superior performance in terms of EE and throughput, even with

different parameter variations. Additionally, the results showed that the number of D2D pairs deployed in the coverage area and the maximum altitude variation had an impact on EE. It was also crucial to obtain an optimal PS ratio for effective resource allocation.

2.3.1.29 DTD3-D2D (Dinkelbach combined twin delayed deep deterministic policy gradient)

This study [71] examined the management of resources in a D2D-NOMA enabled heterogeneous network, specifically in joint NOMA cluster association and power allocation. The objective is to optimize long-term cumulative energy efficiency while ensuring that both cellular users and NOMA users meet QoS requirements. The association problem is formulated as a perfect matching of weighted bipartite graph and addressed through the K-times iteration algorithm. Meanwhile, the power allocation problem is addressed through the proposed DRL framework, which employs the Dinkelbach-TD3 method. This method combines the Dinkelbach algorithm and TD3 to significantly reduce the action space and iterations required for power allocation. The DRL algorithm's flexibility and intelligence are also leveraged in this approach, further enhancing performance.

2.3.1.30 SACC-D2D (Social-aware cooperative caching)

In this paper [72], the author has introduced an innovative social-aware D2D caching mechanism that amalgamates the concept of social incentive and recommendation with D2D caching strategies. Firstly, the author has proposed a federated learning-based graph convolutional network and long short-term memory network (GLSTM) as a prediction model to capture the spatial-temporal feature of social networks in a privacy-preserving manner. Secondly, the predicted social relationship is encoded into the state of DRL. To make optimal decisions, we have proposed the DDQN-based D2D caching algorithm. The performance of prediction, DRL convergence, and caching scheme has been validated through simulation results. Based on the results, it can be inferred that the scheme enhances the offloading probability and reduces average delay consumption to some extent.

Table 2.1: Comparative Analysis of DRL in D2D

Ref.	Scenario	Approach	Problem	Model	Algorithm	Agent	Technique Used	Variant	Open Issues
[43]	Underlay	Distributed	Interference-CR-CI & CO-CI	MDP	MAQL	DDPs-Multi Agent	Channel Selection + power control	FC-MAQ	1. User's mobility 2. EE 3. Multicell scenario
[44]	Underlay	Distributed	Interference-CR-CI & CO-CI	MDP	DDPG	D2D Transmitter - Multi Agent	Power Allocation	S-DDPG	1. Multi-agent learning approaches 2. Low convergence rate 3. High training variance
[45]	Underlay	Centralized & Distributed	Interference-CO-CI	MDP	MAAC & NAAC	BS-Single Agent & DDPs-Multi agent	Spectrum Allocation	MAAC-NAAC-DRL	1. DRL for continuous-valued power control 2. Automatic selection of RB 3. Power allocation
[46]	Underlay & Overlay	Distributed	Interference-CR-CI & System Capacity	MDP	Policy Gradient	D2D Links-Multi Agent	Resource allocation + Power control	ICWN-DRL	1. Intracell interference 2. Energy Efficiency
[47]	Underlay	Distributed	Interference-CR-CI & CO-CI	MDP	MA-DDQL	DDT-Multi Agent	Power Allocation	NC-DRL	1. DRL for continuous action problems 2. Non-cooperative issues
[48]	Underlay	Centerized	Interference-CR-CI & CO-CI	MDP	DQN	BS- single agent	Power Optimization	ST-DRL	1. Multicell Scenario 2. Imperfect CSI
[49]	Underlay	Distributed	Interference-CR-CI & CO-CI	MDP	MADRL	D2D Links-Multi Agent	RB Allocation + Power Control	C-MADRL-PER	Learning model in indoor environments
[50]	Overlay	Distributed	Energy Consumption	MDP	DQN	BS - single agent & DDPs-Multi Agent	Power control + Resource Allocation	CSSCA-DRL	1. Multicell Scenario 2. Imperfect CSI
[51]	Underlay	Centralized	Energy Efficiency	MDP	DQN	DDP-Single Agent	Throughput + Power Control	TP-DQN	1. Imperfect CSI 2. Users mobility
[52]	Underlay	Distributed	Weighted Sum Rate	MDP	DQN	DDPs-Multi Agent	Channel Selection + Power Control	SRPR-DRL	1. User's mobility 2. EE 3. Multicell scenario
[53]	Underlay	Centralized	Energy Efficiency	MDP	Actor-Critic	BS-Single Agent	Mode Selection + Resource Allocation	MSRA-DDPG	1. CR-CI 2. CO-CI
[54]	Underlay & Overlay	Distributed	Interference-CR-CI & CO-CI	MDP	Double DQN	DDPs-Multi-Agent	Channel Assignment + Power Control	D-MARL	1. Het-Net 2. Multicell scenario
[55]	Underlay	Distributed	Interference-CR-CI	MDP	DDQN	DDPs-Multi agent	Sub carrier assignment + Power Allocation	ADMM-DRL	1. Multicell Scenario 2. Imperfect CSI
[56]	Underlay	Centralized	Cross & Co-channel Interference	MDP	DDPG	Macro cell-Single Agent	Spectrum Allocation + Power Allocation	Maximize Energy Efficiency	1. Multicell scenario 2. Imperfect CSI
[57]	Underlay	Distributed	Interference-CR-CI & CO-CI	MDP	DDPG	DDPs-Multi Agent	Power Allocation	SOTPSR-DRL	Imperfect Channel Coefficient Estimation

Ref.	Scenario	Approach	Problem	Model	Algorithm	Agent	Technique Used	Variant	Open Issues
[58]	Underlay	Distributed	Transmission Latency	MDP	Federated learning	DDPs-Multi-Agent	Data privacy + system security	BDRFL	1. CR-CI 2. CO-CI
[59]	Underlay	Centralized	Interference-CO-CI	MDP	Double DQN	DDT - Single Agent	Sum-rate + fairness	D4SA	Distributed multi-agent algorithm
[60]	Underlay	Distributed	D2D Sharing Traffic	MDP	DQN	BS-Single Agent	Node Selection + Cache Replacement	MAAC-NAAC-DRL	1. Imperfect CSI 2. Users mobility
[61]	Underlay	Distriubuted	Interference-CR-CI & CO-CI	MDP	DDPG	DDPs-Multi Agent	Resource allocation + Power control	CO-DDDPG	Multiple Cell Scenario
[62]	Underlay	Distributed	Interference-CR-CI & CO-CI	MDP	MADRL	DDUs-Multi Agent	Joint Channel Allocation + Power Control	SGMA-DRL	1. Energy Efficiency 2. Multiple Cell Scenario
[63]	Underlay	Centerlized	Cache placement of D2D & UAV	MDP	DDPG	UAV- single agent	Cooperative Cache Placement	MSPA-DRL	Distributed Multi-agent algorithm
[64]	Underlay	Distributed	Interference-CR-CI & CO-CI	MDP	D3QN	Central Controller-Single Agent	User Association + Resource Allocation	D3QN-UARA	1. Energy Efficiency 2. Users mobility
[65]	Overlay	Distributed	Throughput	MDP	v	DDTs - Multi Agent	Power control + Channel Selection	MAOD-DRL	1. User's mobility 2. Energy Efficiency 3. Multicell Scenario
[66]	Underlay	Distributed	Throughput	MDP	DDQN	DDP-Single Agent	Dynamic Channel Matching	DSM-DRL	1. Multicell Scenario 2. Imperfect CSI
[67]	Underlay	Distributed	Interference-CO-CI	MDP	D3QN	DDPs-Multi Agent	Power Control + Resource Allocation	PS-D3QN-DRL	Energy Efficiency Analysis with Federated Learning
[68]	Underlay	Centralized	Spectrum Utilization	MDP	FDL	All UEs	Sum Capacity + Power Consumption	FD-DRA-DRL	1. Mobility of DDU's 2. Interference Mitigation 3. Multicell Scenario
[69] & Overlay	Underlay	Distributed	Interference-CR-CI & CO-CI	MDP	DQN	DDPs-Multi-Agent	Channel Assignment + Power Control	DSA-DRL-D2D-IOT	1. Het-Net 2. Multicell scenario
[70]	Underlay	Distriubuted	Interference-CR-CI	MDP	DQN	UAVs-Multi agent	Joint Optimization of Power Splitting + Time	UAVs-SWIPT-MADQN	Energy Harvesting with UAV mobility
[71]	Underlay	Centralized	Energy Efficiency	MDP	TD3	BS- Single Agent	Joint Cluster Association + Power Allocatio	DTD3-D2D	1. CR-CI 2. CO-CI
[72]	Underlay	Centralized	Interference-CR-CI & CO-CI	MDP	FDL	BS- Single Agent	Social Incentive with D2D Caching	SACC-D2D	1. Collaborative computation offloading, 2. Security issues

2.3.2 DRL with RIS

RIS has turned out to be a suitable paradigm to address the needs of 5G and even beyond networks and fulfil the demand for an intelligent and programmable wireless environment [73]. The passive components of RIS reflect the incident signal towards the different users, and these elements are controlled via a programmable controller. In order to minimize the cost of building a new BS, the RIS is typically installed in high places (towers, buildings). However, RIS performance optimization continues to be a difficult task due to the huge number of programmable components and the controller's rectifying capabilities. Therefore, it is essential to develop new techniques in order to improve RIS performance. Various DRL-based schemes have been proposed to fully leverage the advantages of RIS with DRL. These schemes will be discussed in detail below.

2.3.2.1 DLR-DQN (Decaying learning rate deep Q-network)

The authors proposed an algorithm [74] based on DLR-DQN to address the joint trajectory of UAV and RIS phase shift optimization problems. In the proposed D-DQN based algorithm, DLR (decaying learning rate) is utilized as opposed to the traditional DQN algorithm. This approach allowed for a balance between expediting training speed and achieving convergence to the local optimal solution while also avoiding oscillation. The agent responsible for determining the UAV's trajectory and the passive beamforming design was chosen to be the central controller. Simulation results demonstrated that the utilization of the RIS can greatly diminish the energy consumption of the UAV. Additionally, it revealed that the RIS-NOMA scenario consumes less energy than the RIS-OMA scenario.

2.3.2.2 WL-DDPG (Water filling-deep deterministic policy gradient)

The issue of joint optimization of power allocation on SCs and Phase shifts of each element in RIS was thoroughly examined by the authors of this article [75]. To address the issue at hand, an initial approach involves the implementation of a WL algorithm designed to optimize the overall spectral efficiency through the allocation of power on the SCs. Subsequently, the phase shift of each element in RIS was controlled through the utilization of the DDPG algorithm. The

results of the simulation demonstrate that the WL-DDPG scheme can attain a SE performance that is nearly equivalent to the SCA algorithm. The SCA algorithm is known to be a solution that is optimal within a certain local context. Meanwhile, the WL-DDPG scheme accomplished this with significantly less computation delay.

2.3.2.3 EA-DDPG (Exploration attenuate-deep deterministic policy gradient)

The issue of joint optimization of power allocation in AP and Phase shifts of each element in RIS was thoroughly examined by the authors of this article [76]. To achieve the target, the authors proposed two ML algorithms. The environment trained-DL algorithm facilitated the training of the DL agent by means of the interaction environment, with the ultimate goal of obviating the need for a set of training data. On the other hand, the EA-DDPG algorithm has the ability to achieve consistent and deterministic management of phase shifts. The initial findings from the simulation indicated that the NOMA approach had achieved significant benefits in comparison to the OMA technique across a diverse range of scenarios.

2.3.2.4 PSD-DRL (Phase shift design-DRL)

The objective of the author in this paper [77] was to optimize the PS-RIS. A PSD-DRL technique has been proposed to achieve the said objective. Interestingly, this algorithm did not require any further parameter adjustments, in contrast to the non-optimized PS-RIS algorithm. Furthermore, when contrasted with the conventional DRL for high-definition mode, it substantially lowered the computational complexity, with a noteworthy enhancement in the rate, requiring fewer steps per episode.

2.3.2.5 MC-PD-NOMA-DRL (Multi carrier power domain non-orthogonal multiple access-DRL)

This proposal [78] investigated the challenge of minimizing the overall transmit power by carrying out a joint optimization of every UAV's velocity, phase shifts of RISs, subcarrier assignment, as well as active beamforming at each BS. To address this intricate issue, the primary problem has been partitioned into two sub-issues. The first sub-issue was solved by

the authors using the DDQN approach, while the second sub-issue was tackled through the application of the MC-PD-NOMA technique.

2.3.2.6 ERAP-DRL (Efficient resource allocation parallel-DRL)

The EE problem was formulated by the authors for a multi-UAV network assisted by RIS [79]. The problem takes into consideration power restrictions as well as the requirements of the RIS. The authors have proposed a centralized-DRL technique to effectively address the power distribution issue at the UAVs and the RIS's phase-shift set, in a joint manner, with the aim of enhancing the network performance in terms of EE. After that, the authors employed P-DRL to train every component to possess an intelligence model and minimize the latency in transmitting actions between the UAV and RIS. Finally, they introduced the PPO approach with an improved sampling method to further enhance the overall network's performance. The simulation findings clearly showed that the suggested methods effectively resolve the issue of joint optimization under time-varying environmental conditions and CSI, surpassing other benchmark techniques.

2.3.2.7 UCA-HSE-DRL (UAV collision avoidance-high sample efficient- DRL)

In this proposal [80], the authors collaboratively optimize the trajectory of the UAV, the configuration of the RIS, and the power uploading to enhance the sum rate while adhering to limits that assure the safety of the UAV's flight and meet the UAV and GU's minimal data rate demands. Firstly, a proficient HSA-DRL algorithm was introduced with the aim of tackling the sequential decision-making challenge. Secondly, the authors proposed an advanced distributed and robust DRL algorithm to tackle the uncertainties that arise from the unknown locations of obstacles. This scheme further strengthened the algorithm's robustness. The simulation findings have unequivocally shown that the suggested DRL-based algorithms have exceeded the traditional ones. This was evident not only in their learning efficiency but also in their ability to withstand difficulties.

2.3.2.8 DRPO-PGAC (Decomposition and relaxation-based precoding optimization-policy gradient Actor- Critic)

In this study [81], the authors developed DRL-based algorithms with the aim of enhancing the efficiency of the BS precoding matrix and the phase shifts of each RIS element in a vehicular network assisted by RIS technology. In an extremely changing and unpredictable wireless environment, a two-stage framework has been designed to adjust the BS and RIS beamforming matrices. To address the optimization of the BS precoding matrix, the authors initially presented the DRPO scheme. Subsequently, the PGAC scheme was introduced to fine-tune the phase shifts of individual elements within the RIS. The simulation findings indicated that the effectiveness of both approaches and demonstrated their capability to enhance dependability in the face of unforeseen obstructions.

2.3.2.9 SAC-AO-RIS (Soft actor-critic-alternating optimization-RIS)

The authors of this study [82] formulated the challenge of reducing the average episodic age of information's by jointly optimizing TR-UAV and PS-RIS. The authors broke the issue into two sub-issues to achieve their purpose. The SAC approach was used for optimizing the TR-UAV. On the other hand, the AO approach was used to provide solution for optimizing the PS-RIS. The simulation findings clearly showed that both of the suggested algorithms significantly reduce the age of information as compared to other baseline approaches.

2.3.2.10 WPT-DDPG (Wireless power transfer-DDPG)

The author's aim in this paper [83] was to jointly optimize the TR-UAV, PS-RIS and WPT scheduling of IoT terminals. The optimization problem was successfully solved through the utilization of DRL-based DDPG and PPO algorithms. The IoT terminals were capable of harvesting energy in a down-link scenario and transferring information to the UAVs in a up-link scenario with the assistance of RIS. The authors examined two UAV scenarios. In the first scenario, the UAV was observed hovering at the center of the cluster while concurrently rendering energy to the IoT terminals. The implementation of RIS served to mitigate the uplink interference. In second scenario, the UAV was deployed in an initial place and needed to identify an

improved communication point. The simulation findings indicated that the assistance provided by the RIS facilitated a more robust association, resulting in a noteworthy enhancement of the overall performance.

2.3.2.11 PSD-TD3-RIS (Primal-dual sub gradient descent-twin-delayed DDPG-RIS)

This research [84] investigated the simultaneous optimization of channel selection and phase shifts of individual elements in RIS. The objective was to enhance the system overall sum rate, whilst simultaneously fulfilling the BER requirements of all users. A model based on TD3 was developed with the aim of learning about RIS setup. As a result, the algorithm gains the capacity to enhance the overall system rate through adjustments to the data rates provided to the users. The experimental findings demonstrated that the suggested algorithm surpassed its non-learning counterparts in reference to the sum rate, thus proving its superiority.

2.3.2.12 MAML-DDPG (Model-agnostic-meta-learning-DDPG)

The authors of this proposal [85] optimized the PS-RIS and power distribution at BS in a joint manner. The objective was to improve the sum rate of overall network. The problem was divided into two sub-problems with precision and accuracy. algorithm was utilized to improve the NN's capacity for generalization while also achieving a rapid convergence rate for short-term sub-problem. On the contrary, the utilization of the DDPG approach was implemented to set up the PS-RIS and distribute power at the BS in the long-term sub-problem. The simulation findings clearly showed that both of the proposed DRL-based algorithms significantly reduce the age of information as compared to other baseline approaches. The numerical findings indicated that both of the proposed algorithms have remarkably enhanced the network in comparison to the traditional algorithms.

2.3.2.13 EH-R-DRL (Energy harvesting-robust-DRL)

In this paper [86], the authors examined how to minimize the battery power consumption in a UAV-supported RIS network, which in turn limits its service capabilities. During the process, a proposal was put forth for the RIS supported UAV system that promises to yield

Table 2.2: Comparative Analysis of DRL in RIS

Ref.	Scenario	Approach	Problem	Model	Algorithm	Agent	Technique Used	Variant	Open Issues
[74]	Underlay	Distributed	Energy Consumption Minimization	MDP	DDQN	Central Controller-Single Agent	Phase Shift Control + Trajectory Design	DLR-DQN	Optimization of the UAV Velocity
[75]	Underlay	Centralized	Spectrum Efficiency	MDP	Actor-Critic Method	BS - Single Agent	Joint Optimization of Power Allocation + Phase Shifts	WL-DDPG	1. Optimization of Phase Shift 2. Power Allocation 3. Multicell Scenario
[76]	Underlay	Centralized & Distributed	Effective Throughput	MDP	Actor-Critic Method	AP (single-antenna access point)	Joint Optimization of Power Allocation + Phase Shifts	EA-DDPG	1. Multiple-Antenna Transmission, 2. Overhead-Dependent 3. Joint Beamforming optimization
[77]	Underlay & Overlay	Centralized	Sum-Rate	MDP	DDPG	RIS Controller-Single Agent	Optimization of RIS Phase Shifts	PSD-DRL	Multi-User Scenario
[78]	Underlay	Distributed	Transmit Power Minimization	MDP	DDQN	UAV & RIS-Multi Agent	Resource Management for Transmit Power Minimization	MC-PD-NOMA-DRL	1. CO-CI Mitigation 2. CR-CI Mitigation
[79]	Underlay	Centralized	Energy Efficiency	MDP	DDPG & PPO	UAV-RIS-single Agent	Joint Optimization of Power Allocation + Phase Shifts	ERAP-DRL	1. Multiple-RIS 2. Cooperative Communication
[80]	Underlay	Centralized	Network Sum Rate	MDP	Actor-Critic Algorithm	Central Controller-Single Agent	Joint Optimization of UAV's Trajectory + RIS Configuration + Power Control	UCA-HSE-DRL	1. Multi-Cell Scenario 2. Inter-Cell Interference Mitigation
[81]	Underlay	Centralized & Distributed	Network Data Rate	MDP	Actor-Critic Algorithm	BS - single agent	Joint Optimization of Phase Shifts of the RIS + BS preceding Matrix	DRPO-PGAC	1. Multiple-RIS 2. Inter-Cell Interference Mitigation
[82]	Underlay	Distributed	UAV's Trajectory & Phase Shift of RIS	MDP	Soft Actor-Critic	UAV-RIS-Multi Agent	Optimization of Age of Information of IoT Devices	SAC-AO-RIS	1. Variation in Distribution of Buildings 2. Location of the RIS 3. Locations of IoTDs
[83]	Underlay	Centralized	Sum Rate	MDP	DDPG & PPO	UAV-Single Agent	Joint Optimization of the UAV's Trajectory + WPT Scheduling of IoT Terminals + RIS's Phase Shifts	WPT-DDPG	1. Distributed Model 2. Cooperative communications with Multiple UAVs

far-reaching results. This was achieved by partitioning the inactive reflected arrays within the geometric realm, to concurrently transfer information along with energy harvesting. The scheme that was suggested relied on the SWIPT scheme. Simulation outcomes indicated that the suggested energy harvesting technique is both efficient and effective in terms of minimizing power consumption in UAV-RIS systems.

Ref.	Scenario	Approach	Problem	Model	Algorithm	Agent	Technique Used	Variant	Open Issues
[84]	Underlay	Centralized	Energy Efficiency	MDP	PPO	BS-Single Agent	Joint Optimization of Location + Phase Shift of RIS	SCA-AC-PPO	Optimization of the UAV's Trajectory
[85]	Underlay	Centralized	Transmit Rate	MDP	DDPG	BS - Single Agent	Joint Optimization of Channel Selection and Phase Shifts	PSD-TD3-RIS	Interference Mitigation
[86]	Underlay	Centralized	Network Throughput	MDP	DDPG	BS-Single Agent	Joint Optimization of RIS phase Shifts and BS Power Allocation	MAML-DDPG	Imperfect Channel State Information
[87]	Underlay	Centralized	Harvesting Energy Efficiency	MDP	DDPG	RIS Controller-Single Agent	Optimization of Energy Harvesting Efficiency	EH-R-DRL	Multiple UAV-RIS Scenario
[88]	Underlay	Centralized	Energy Efficiency	MDP	DDPG	RIS Controller-Single Agent	Hybrid Joint Optimization for Sub-array Partition + RIS Beam-forming	ARSO-DDPG	Multi-cell Consideration
[89]	Underlay	Centralized	System Sum Rate	MDP	DDPG & PPO	Central Bontroller-single Agent	Joint Optimization of Phase Shift and Beam-forming of RIS	IS-UAV-TN-MO-DDPG	Multiple UAVs Trajectories Optimization
[90]	Underlay	Centralized	System Sum Rate	MDP	DDPG & PPO	Central Bontroller-single Agent	Joint Optimization of Phase Shift and Beam-forming of RIS	MU-MISO-PW-DRL	Joint Beam-forming Based on the Received Pilots without Relying on the CSI
[91]	Underlay	Centralized	Efficient Spectrum Sharing	MDP	DDPG	BS - single agent	Joint Optimization of Phase Shifts of the RIS + BS preceding Matrix	DDPG-TD3	1. Multiple-RIS 2. Inter-Cell Interference Mitigation
[92]	Underlay	Centralized	Secrecy Energy Efficiency	MDP	DDPG	UAV-Single Agent	Joint Optimization of Flight Trajectory + Beam Forming	TTD3-DRL	Optimization of the Phase Shift of RIS
[93]	Underlay	Centralized	Sum Rate	MDP	PPO	BS-Single Agent	Joint Optimization of the Phase-Shift Matrices of Two RIS	RAFD-6GV2X-LCPPO	Interference Mitigation

2.3.2.14 ARSO-DDPG (Active RIS subarray optimization scheme based on deep deterministic policy gradient)

In this proposal [87], the authors thoroughly investigated the problem of subarray partitioning and RIS beamforming. The long-term objective of optimization was to achieve average EE with the restrictions of the network's power and the minimal capacity needs of every user in the network. This was achieved through the utilization of proposed ARSO-DDPG algorithm and adaptable optimization of parameters. The findings from the numerical analysis have substantiated that average EE of the network can be enhanced to a great extent by means of delicate subarray partition design.

2.3.2.15 IS-UAV-TN-MO-DDPG (Integrated satellite-unmanned aerial vehicle-terrestrial network-multi objective-DDPG)

: In this manuscript [88], the authors presented an innovative model for the RIS-supported IS-UAV-TNs architecture. The objective of this model is to satisfy the practical demands of the 5G and beyond networks. A MO-DDPG algorithm was designed for real-time control TR-UAV for increasing the system sum-rate and minimize energy consumption of the UAV. Moreover, utilizing a UAV supported by RIS technology, satellite sends signals to the users via the NOMA technique in order to furnish a unified virtual line-of-sight (V-LoS) connection. Experimental finding confirmed that the efficacy of the suggested approach, and it has been demonstrated that the MO-DDPG approach is more flexible in its optimization strategy compared to conventional approaches.

2.3.2.16 MU-MISO-PW-DRL (Multi-user-multiple input single output-piecewise-DRL)

In this article [89], authors investigated the issue of joint beamforming within an RIS-supported MU-MISO network. This includes the optimization of the PS-RIS as well as the BS's transmission pre-coding matrices. An effective PW-DRL algorithm has been introduced for addressing this formidable issue. Particularly, the proposed algorithm effectively enhanced the network's sum rate with the DDPG approach. The simulation findings have demonstrated the excellence of the suggested approach when compared to the existing ones, in terms of both convergence and sum rate.

2.3.2.17 DDPG-TD3 (DDPG integrated twin delayed deep deterministic)

In this article [90], authors presented the combination of MIMO radar and CUs network with the assistance of RIS. They conducted research on optimization of the transmit signal of radar, RIS's phase shift and transmit precoder matrix issues jointly with the objective of optimal spectrum allocation. To address the joint optimization challenge, the TD3-based algorithm was enriched through the implementation of the DDPG technique. The algorithm that has been suggested aims to prioritize the maximization of reward by leveraging the knowledge that has been acquired from past experiences in conjunction with the existing state. The finding revealed

that the deployment of RIS effectively reduced interference in the system, thereby limiting the impact of RIS echoing pulses on radar detecting.

2.3.2.18 TTD3-DRL (Twin twin-delayed deep deterministic policy gradient based DRL)

This study [91] explored the joint optimization problem of both UAV trajectory and active beamforming, as well as passive beamforming in RIS. To tackle the joint problem, the authors implemented a cutting-edge algorithm known as the TTD3-based DRL approach with the aim of optimizing the expected cumulative reward. This has resulted in a significant increase in the SEE. The simulation outcomes verified that the suggested technique attains superior energy conservation in terms of secrecy than the conventional TD3-based DRL techniques.

2.3.2.19 SCA-AC-PPO (Successive convex approximation-actor critic-proximal policy optimization)

In this manuscript [92], the authors conducted research on an energy-efficient communication network with the assistance of multiple UAV-equipped RIS. In order to optimize EE, the authors developed a problem statement that focuses on the joint aspects of locations, phase shift, and power of multiple RIS. To tackle the challenge, the problem was broken down into three sub-problems: deployment, phase shift, and power control. The authors subsequently put forth the SCA technique, the AC-PPO technique, and the whale optimization technique as viable solutions to address aforementioned sub problems. The findings from the numerical analysis have substantiated that network's EE can significantly improve by means of integrating multiple RIS.

2.3.2.20 RAFD-6GV2X-LCPPO (RIS assisted full duplex-6G vehicle to everything-low complexity proximal policy optimization scheme)

In this paper [93], the authors explore the problems of phase-shift matrices of the two RIS. The main objective of this investigation was to enhance the attainable sum-rate while adhering to the discrete phase-shift constraint that corresponds to every reflective component of the two RISs. The authors proposed LCPPO approach that effectively operates through an

online fashion to address optimization problem. The LCPPO algorithm acquired knowledge from its interaction with the surrounding environment, thereby improving the course of action to procure the intended reward.

2.4 Summary

In this chapter, we explore the various DRL variants that are applicable in the context of 5G and beyond. The literature review is partitioned into three distinct sections. The first section of the literature review thoroughly examines the overview and categorization of DRL. Firstly, a comprehensive discussion is presented on the basics of RL and DL. Then we present the categorization of DRL schemes, which is based on the distinct characteristics of the policy functions, various policy evaluation approaches, and the parameter updates in various learning techniques. The second part of the chapter discussed the various DRL variants in the D2D-C. The third part of the literature review discussed the various DRL variants in the implementation of RIS technology.

2.5 Research Gaps

After carrying out an in-depth examination of the literature review, it is apparent that there exist a number of research deficiencies that necessitate attention during the execution of D2D-C.

- **Peer Discovery in Ultra Dense Networks:** In general peer discovery is categorized into two parts, which can be either user perspective or network perspective. User perspective is further categorized into two categories- restricted and open [94]. In restricted, without the permission of end users devices cannot be discovered, while in open, devices can be discovered when they lie in the coverage range of other users. Whereas, in network perspective peer discovery can be partially or fully controlled by the BS [94]. Hence, peer discovery is the very first step used for link establishment between DDPs. Therefore, to discover the peer UEs, an efficient method which has long coverage range and low power

consumption technique is required. In an ultra-dense networks, the cooperation between the adjacent BSs is difficult to establish which makes peer discovery a challenging task [4].

- **Interference management in multi-cell scenario** [5]: If the same RBs are assigned to both D2D users (DDUs) and CUs, interference will arise. During down-link transmission, CUs experience interference from DDTs, while DDU encounter interference from BS. On the contrary, during up-link transmission, the BS is susceptible to interference from the DDTS, while the DDU experience interference from the CU transmitters. Inter-user interference, which refers to interference among DDU, is another challenge that must be addressed for D2D-C, in addition to the aforementioned interference. Various authors have proposed different schemes to resolve this issue but the existing proposals do not have a capability to cope up with next generation wireless networks.
- **Resource allocation for more than one BSs**: In order to enhance the SE of the overall network, it is imperative to implement an appropriate resource allocation strategy among DDU and CU. In the context of D2D-C, there are two distinct approaches for resource allocation, namely orthogonal and non-orthogonal techniques. In orthogonal resource allocation, a portion of the RBs is assigned to the DDU while the remaining portion is utilized by the CU. On the other hand, non-orthogonal resource allocation entails the sharing of the same RBs between the DDU and CU. RB allocation in D2D-C can be done in either a centralized or distributed fashion. In schemes that are centralized [95], the BS functions as an agent and allocates the RBs to both DDU and CU. Centralized schemes can accommodate a considerable number of users; however, the complexity of the system is heightened. On the contrary, when employing a distributed approach [94], it is necessary to utilize a message passing algorithm to address interference, yet this results in a reduction in device complexity [96].
- **Mode selection to mitigate interference** [97]: It is a process which is performed when DDU formed a pair with each other. Since, in D2D-C users directly communicate with each other but still an optimal performance is not achieved. Mode selection is a problem

in which network or D2D paired users decide whether DDUs communicate directly or via cellular network. In D2D-C, there are three types of mode communication: 1) Overlay mode: In this, dedicated RBs are assigned to DDUs while the remaining ones are utilized by CUs. 2) Underlay mode: In this mode both DDUs and CUs reuse the same resources. 3) Cellular mode: In this CUs communicate with each other via BS. Out of the above three modes, spectral efficiency and interference are high in underlay mode while lower in overlay.

- **Power control for DDTs:** To optimize the use of RBs it is imperative to exercise control over the power of DDTs. If DDTs are deployed randomly within cellular networks, the network's performance may experience degradation as a result of CO-CI. To meet the SINR demands of CUs, it is necessary for DDTs to constrain their power. The implementation of power optimization methods has the potential to enhance both the network's throughput and EE. Power optimization methods can also be used to reduce the impact of CO-CI [4]. In the context of 5G and future networks, it is imperative to devise an efficient power control mechanism, particularly in a scenario involving multiple cells.
- **Security for DDUs and CUs:** In traditional communication system, location of the devices are first identified, authenticated and then encrypted to carry out the communication. This is a trusted system as it is based on the core network. On the other hand, in case of D2D-C, security is a challenge because in this devices communicate with each other with or without involvement of core network. Therefore, cryptographic solutions are required to pass the information through wireless channels. DDUs used the security schemes provided by cellular operators only when they lie under their range but when they are out of operator's range their security is an issue. In such a case, various authors [98] proposed relays mechanism to protect the users from malicious attacks but still an effective security scheme need to be developed.
- **Energy Efficiency [99]:** D2D-C improves the network's energy efficiency by means of close-proximity communication, however, both DDUs and CUS are restricted by their limited battery life. In general, D2D-C relies heavily on peer discovery and commu-

nication protocols. In situations where these protocols require UEs to quickly establish connections and frequently transmit data, the battery is drained at a faster rate. Alongside the aforementioned concern, altering and recharging device batteries at regular intervals is not a feasible option, leading to an increase in expenses. In order to extend the lifespan of DDU and CU, it is essential to implement a suitable and efficient energy management strategy between the UEs.

2.6 Objectives

After analyzing the current proposals and identifying the research gaps, we propose the following objectives in our proposal.

1. To design a resource allocation scheme for D2D users underlying cellular networks using machine learning.
2. To develop an energy efficient resource allocation scheme for D2D users using machine learning.

2.7 Methodology

The various steps should be followed for achieving the following objectives is shown in Fig. 2.13.

2.8 Methodology for Objective 1

In the future, wireless networks that have a high density of users and are subject to rapid environmental changes pose two challenges for resource allocation schemes: (i) Due to the significant signaling overhead, obtaining the accurate channel state information is unattainable for BS. (ii) these issues are usually represented as combinatorial optimization problems with non-linear constraints that are challenging to optimize using conventional optimization techniques. RL techniques are implemented to tackle these concerns by effectively handling poli-

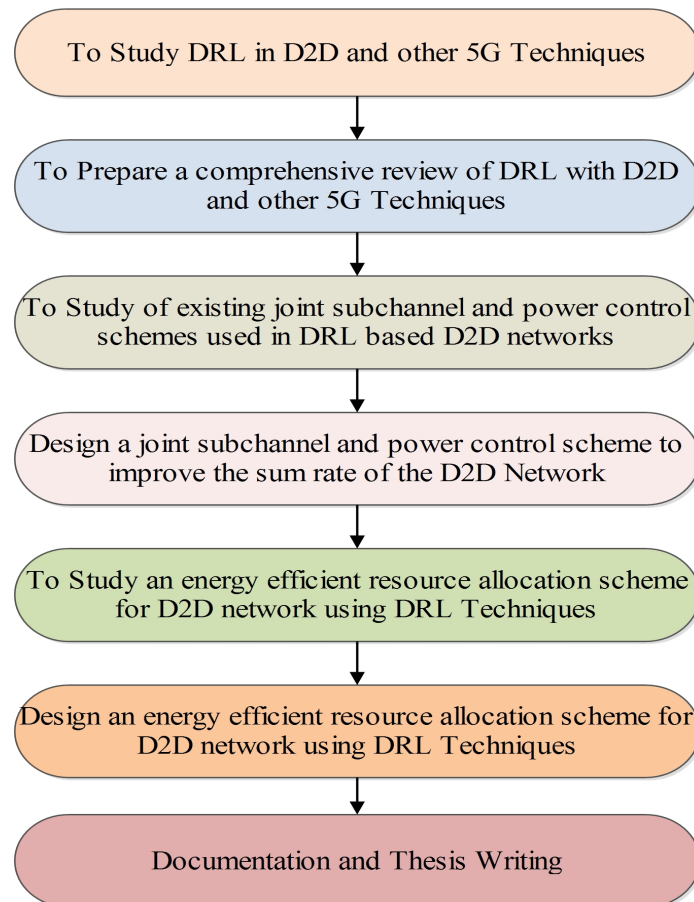


Figure 2.13: A system model to meet the objectives

cies and promoting intelligent decisions. It is correlated with the policies for decision-making and the rewards provided to attain the desired objective. Furthermore, within the realm of RL, every individual agent is responsible for selecting the optimal channel using the policy that has been acquired through learning. Nevertheless, traditional RL algorithms encounter issues with sluggish convergence rates and are not as effective when dealing with challenges that involve extensive state and action spaces. This particular objective has the aim of devising a resource allocation strategy that is highly efficient, utilizing the DRL techniques. To attain the intended objective, the integration of DRL with D2D-C can be employed, which offers the subsequent benefits: (i) mitigating the instability of the environment, and (ii) augmenting the reward of every agent. Finally, an evaluation and analysis of the proposed scheme has been conducted in comparison to the existing DRL schemes, with a focus on sum-rate.

2.9 Methodology for Objective 2

The main goal of this objective is to create an energy-efficient resource allocation scheme. To achieve this objective, we put forth a framework based on DRL for leveraging RIS in a UAV-assisted wireless network. RIS has turned out to be a suitable paradigm to address the needs of 5G and beyond networks and fulfill the demand for an intelligent and programmable wireless environment. However, RIS performance optimization continues to be a difficult task due to the huge number of programmable components and the controller's rectifying capabilities. Therefore, it is essential to develop new techniques in order to improve RIS performance. Various DRL-based schemes have been proposed to fully leverage the advantages of RIS with DRL. Our objective is to design an energy-efficient resource allocation algorithm with the aim of maximizing the overall EE of the networks. Firstly, in order to accomplish this, we will delve into a comprehensive analysis of the diverse energy-efficient schemes that have been suggested for RIS-UAV-assisted D2D underlay network through the utilization of DRL. Secondly, we will delve into the diverse facets of RIS and its numerous benefits when integrated with DRL-powered D2D-C. Thirdly, our next undertaking will involve the creation of a proposed scheme that entails the utilization of RIS-based D2D with the aid of DRL. Finally, we will thoroughly evaluate the effectiveness of the proposed scheme in comparison to the current schemes based on its EE.

Chapter 3

A Deep Reinforcement Learning Scheme for Sum Rate and Fairness Maximization Among D2D Pairs Underlying Cellular Network With NOMA

Following are the major contributions of this chapter.

- Firstly, sum rate and FUF are formulated for all users to maximise it while considering CUs and DDPs' resource and power allocations and QoS demands. To discover the solution that is optimal on a global scale, we bifurcated the problem into two sub-problems. Firstly, we focused on optimizing the power of the BS with the intent of decreasing CR-CI amongst CUs. Secondly, we concentrated on managing the power of the DDPs and CUs with the aim of reducing CO-CI and enhancing fairness.
- To mitigate the effect of CR-CI, we used the DDPG at the NOMA-enabled BS to control its power across each RB. Also, the presence of NOMA across each RB mitigates the intra-user interference due to SIC but results in unfairness across the CUs due to the different power levels. So, to improve fairness, we integrate AGMA with DDPG because it has the ability to divide the power equally among each RB.

- Also, it has been observed that the DDTs cannot be trained simultaneously due to their inability to access the instantaneous CSI, resulting in a decrease in fairness. To handle this problem, a multi-agent D3PG scheme is proposed, in which DDPs serve as an agent.
- Apart from fairness among the DDPs and CUs, CO-CI is another problem that arises because one DDP does not know the transmitting powers of other DDPs. Hence, it becomes extremely difficult for an optimizer to optimize the transmitting power of all DDPs when they transmit at the same time. To overcome this issue, the conventional optimization (CO) scheme known as successive convex approximation (SCA) is integrated with D3PG and known as CO-D3PG.
- The performance of the proposed scheme has been compared to that of the existing schemes. The experimental results have demonstrated that it offers a solution that is nearly optimal while maintaining a high convergence speed.

3.1 Network Model

This specific section delineates the diverse constituents of the proposed network structure, as depicted in Figure 3.1.

3.1.1 Network Architecture

The network architecture in Fig. 3.1 depicts a single cell down-link transmission scenario in an underlay cellular network consisting of CUs along with DDPs. The BS is located at the center and denoted by b , while CUs and DDPs are evenly distributed around the cell. Let the set of CUs be denoted as $\mathcal{C} = \{1, 2, \dots, c, \dots, C\}$ and the set of DDPs be denoted as $\mathcal{D} = \{1, 2, \dots, d, \dots, D\}$. The BS uses the NOMA approach to provide services to a group of CUs, and D2D transmitters (DDTs) provide services to D2D receivers (DDRs) using the orthogonal multiple access (OMA) technology. Let B represent the entirety of the network's available bandwidth, which has been segmented into a total of \mathcal{R} RBs. Let the set of RBs denoted as $\mathcal{R} = \{1, 2, \dots, r, \dots, R\}$, and the set of time slot denoted as $\mathcal{T} = \{1, 2, \dots, t, \dots, T\}$. In the

Table 3.1: A List of Symbols Utilized.

Notations	Description
C, D, R	Count of CUs, DDPs and RBs
P_B, P_c, P_d^r	Power of BS, CU and DDT
x_c^r	Transmitted symbol for c^{th} CU
g_c^r	Channel gain between c^{th} CU and BS
g_{d-c}^r	Channel gain between c^{th} CU and d^{th} DDP
g_d^r	Channel gain of DDR
g_{b-d}^r	Channel gain across BS and d^{th} DDP
$g_{d'-d}^r$	Channel gains across other DDPs to d^{th} DDP
t	Duration of time slot
T	Total time slots
ζ_c^r	AWGN
α_c^r, β_d^r	Channel coefficient for c^{th} CU and d^{th} DDP
$\Upsilon_c^r, \Upsilon_d^r$	SINR for c^{th} CU and d^{th} DDP over r^{th} RB
F, G	Scheduling policy and strategy
$\mathbb{U}_c^r, \mathbb{U}_d^r$	PUF for c^{th} CU and d^{th} DDP
$\mathbb{V}, \mathbb{E}\{.\}$	State value function and expectation operator
\mathbb{D}	Markov chain's static distribution
δ	Step-size parameter
S, A, γ	State space, action space, reward function
C, E	Size of replay buffer and samples
Q	Action-value function of RL
L	Loss function
y_e	Target value
ϑ	Temporal difference
\mathbb{P}_k^{er}	Probability of k^{th} sample
Φ_k	Priority assign to k^{th} sample
ϖ_1	Priority control factor
Φ_k	Prioritize sample weights
χ	Hyper-parameter
ϕ, ϕ'	Weights of critic, target critic network
θ, θ'	Weights of actor, target actor network
$\mathcal{O}_m^M, \mathcal{O}_m^A$	CC of the matrix process, activation function
X	Total layers
W, ρ_w^A	number of actor network layers, neurons in the w^{th} layer

proposed network architecture, CUs and DDPs share the same RB to enhance the SE of the network. In this study, we consider that multipath propagation, shadowing, and path loss cause rapid and slow fading effects on all channel-links. Furthermore, we employ the time-varying channel scenario, wherein each time slot experiences a variation in the global CSI. Also, the

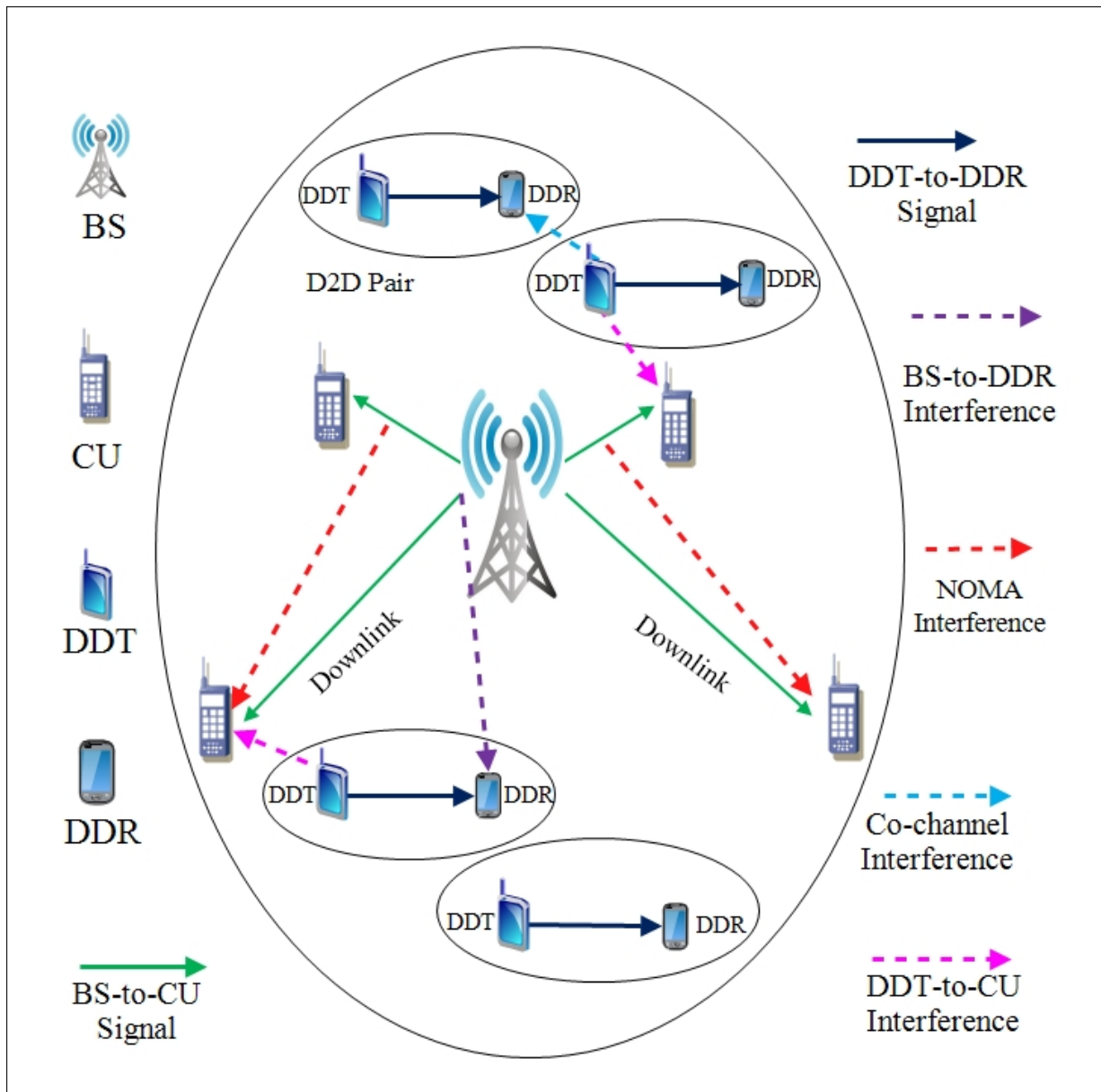


Figure 3.1: Network Architecture

location of the DDRs changes with reference to the DDT. Assuming that the Base Station (BS) possesses sufficient knowledge of the CSI for all users on the network, including DDPs) as well as CUs, and communication occurs across a quasi-static Rayleigh fading channel. Table 3.1 describes the symbols utilized throughout the paper.

3.1.2 Channel Model

3.1.2.1 CU Channel Model

Let power transmitted by the BS be denoted by P_B and power assigned to c^{th} CU be denoted by P_c . The message received at c^{th} CU by BS in the r^{th} RB during the t^{th} time slot is as follows [100]:

$$M_c^r(t) = \sqrt{P_c^r(t)}g_c^r(t)x_c^r + \sum_{c' \neq c, c' \in \mathcal{C}} \alpha_{c'}^r(t) \sqrt{P_{c'}^r(t)}g_{c'}^r(t)x_{c'}^r(t) + \sum_{d \in \mathcal{D}} \beta_d^r(t) \sqrt{P_d^r(t)}g_{d-c}^r(t)x_d^r(t) + \zeta_c^r(t), \quad (3.1)$$

where $x_c^r(t)$, $x_{c'}^r(t)$ and $x_d^r(t)$ represent the transmitted symbol for c^{th} CU, c'^{th} CU and d^{th} DDP in the r^{th} RB, respectively. g_c^r specifies the channel gain that exists between the c^{th} CU and b, while $g_{c'}^r$ specifies the channel gain that exists between the c'^{th} CU and b. Similarly, g_{d-c}^r specifies the channel gain existing between the c^{th} CU and d^{th} DDP. The power of d^{th} DDP is specified by P_d^r , and the AWGN for c^{th} CU is specified by ζ_c^r . Finally, the channel coefficients for c^{th} CU and d^{th} DDP are represented by α_c^r and β_d^r , respectively, and are defined as follows:

$$\alpha_c^r(t) = \begin{cases} 1 & \text{if } c^{th} \text{ CU is scheduled on the } r^{th} \text{ RB,} \\ 0 & \text{otherwise.} \end{cases} \quad (3.2)$$

$$\beta_d^r(t) = \begin{cases} 1 & \text{if } d^{th} \text{ DDP is scheduled on the } r^{th} \text{ RB,} \\ 0 & \text{otherwise.} \end{cases} \quad (3.3)$$

NOMA is a scheme in which more than one user shares the same spectrum resources at any instant but with different power levels. Let c^{th} CU wants to decrypt and eliminate interference from the combined signal of c'^{th} CU using the SIC technique in r^{th} RB. The obtained SINR of c^{th} CU for the j^{th} CU's signal is now greater or equal to the obtained SINR of j^{th} CU's own signal for effective interference cancellation [101], [102]. This condition can be mathematically

expressed as follows:

$$\begin{aligned} & \frac{P_c^r(t)|g_c^r(t)|^2}{\sum_{c' \neq c, c' \in \mathcal{C}} \alpha_{c'}^r(t)P_{c'}^r(t)|g_{c'}^r(t)|^2 + \sum_{d \in \mathcal{D}} \beta_d^r(t)P_d^r(t)|g_{d-c}^r(t)|^2 + \xi_c^r(t)} \\ & \geq \frac{P_c^r(t)|g_c^r(t)|^2}{\sum_{c' \neq c, c' \in \mathcal{C}} \alpha_{c'}^r(t)P_{c'}^r(t)|g_{c'}^r(t)|^2 + \sum_{d \in \mathcal{D}} \beta_d^r(t)P_d^r(t)|g_{d-c}^r(t)|^2 + \xi_c^r(t)}. \end{aligned} \quad (3.4)$$

Let $\mathcal{U}_{c,c'}^r$ is the SIC status indicator between c^{th} CU and c'^{th} CU, which may be describes as:

$$\mathcal{U}_{c,c'}^r = \begin{cases} 1 & \text{if } c^{th} \text{ CU successfully eliminate the interference} \\ & \text{across the } r^{th} \text{ RB,} \\ 0 & \text{otherwise.} \end{cases} \quad (3.5)$$

Now SINR for c^{th} CU is calculated as follows [103]:

$$\Upsilon_c^r(t) = \frac{P_c^r(t)|g_c^r(t)|^2}{\mathbb{I}\mathbb{F}_{c'-c}^r + \mathbb{I}\mathbb{F}_{d-c}^r + \xi_c^r(t)}, \quad (3.6)$$

where $\mathbb{I}\mathbb{F}_{c'-c}^r$ represents the interference (intra-user) caused by other CUs on c^{th} CU and can be defined as follows:

$$\mathbb{I}\mathbb{F}_{c'-c}^r = \sum_{c' \neq c, c' \in \mathcal{C}} \mathcal{U}_{c,c'}^r \alpha_{c'}^r(t)P_{c'}^r(t)|g_{c'}^r(t)|^2. \quad (3.7)$$

$\mathbb{I}\mathbb{F}_{d-c}^r$ represents the interference caused by d^{th} DDP on c^{th} CU and may be describes as:

$$\mathbb{I}\mathbb{F}_{d-c}^r = \sum_{d \in \mathcal{D}} \beta_d^r(t)P_d^r(t)|g_{d-c}^r(t)|^2. \quad (3.8)$$

Defining $|g_c^r(t)|^2 = |\widehat{g}_c^r|^2 z_c^{-\eta}$, $|g_{c'}^r(t)|^2 = |\widehat{g}_{c'}^r|^2 z_{c'}^{-\eta}$, and $|g_{d-c}^r(t)|^2 = |\widehat{g}_{d-c}^r|^2 z_{d-c}^{-\eta}$. Here, $|\widehat{g}_c^r| \sim \mathcal{C}\mathcal{N}(0, 1)$, $|\widehat{g}_{c'}^r| \sim \mathcal{C}\mathcal{N}(0, 1)$, $|\widehat{g}_{d-c}^r| \sim \mathcal{C}\mathcal{N}(0, 1)$. These coefficients are defined as small scale fading. The distance between the c^{th} CU and b , the $(c')^{th}$ CU and b and the c^{th} CU and d^{th} DDP are denoted by the symbols $z_c^{-\eta}$, $z_{c'}^{-\eta}$, and $z_{d-c}^{-\eta}$, respectively.

3.1.2.2 D2D Channel Model

The received message at d^{th} DDR over the r^{th} RB is given as [104]:

$$\begin{aligned}
M_d^r(t) &= \sqrt{P_d^r(t)} g_d^r(t) x_d^r(t) \\
&+ \sum_{d' \neq d, d' \in \mathcal{D}} \beta_{d'}^r(t) \sqrt{P_{d'}^r(t)} g_{d'}^r(t) x_{d'}^r(t) \\
&+ \sum_{d \in \mathcal{D}} \beta_d^r(t) \sqrt{P_d^r(t)} |g_{b-d}^r(t) x_d^r(t) + \zeta_d^r(t), \tag{3.9}
\end{aligned}$$

where g_d^r signifies the channel gain linking the d^{th} DDT and DDR, and $g_{d'}^r$ signifies the channel gain connecting the d'^{th} DDT and DDR. Likewise, the channel gain connecting b and the d^{th} DDP is signified by g_{b-d}^r .

$$\Upsilon_d^r = \frac{P_d^r(t) |g_d^r(t)|^2}{\mathbb{I}\mathbb{F}_{c'-c}^r + \mathbb{I}\mathbb{F}_{d'-d}^r + \mathbb{I}\mathbb{F}_{b-d}^r + \xi_d^r(t)}, \tag{3.10}$$

where $\mathbb{I}\mathbb{F}_{d'-d}^r$ is the interference (co-tier) created by other DDPs on d^{th} DDP which is given as:

$$\mathbb{I}\mathbb{F}_{d'-c}^r = \sum_{d' \neq d, d' \in \mathcal{D}} \beta_{d'}^r(t) P_{d'}^r(t) |g_{d'}^r(t)|^2. \tag{3.11}$$

The interference (cross-tier) created by BS on DDPs is represented by $\mathbb{I}\mathbb{F}_{b-d}^r$, which is defined as:

$$\mathbb{I}\mathbb{F}_{b-d}^r = \sum_{d \in \mathcal{D}} \beta_d^r(t) P_d^r(t) |g_{b-d}^r(t)|^2, \tag{3.12}$$

where g_{b-d}^r express the channel gain across BS and d^{th} DDP. Defining $|g_d^r(t)|^2 = |\widehat{g}_d^r|^2 z_d^{-\eta}$, $|g_{d'}^r(t)|^2 = |\widehat{g}_{d'}^r|^2 z_{d'}^{-\eta}$, and $|g_{b-d}^r(t)|^2 = |\widehat{g}_{b-d}^r|^2 z_{b-d}^{-\eta}$. Here, $|\widehat{g}_d^r| \sim \mathcal{C}\mathcal{N}(0, 1)$, $|\widehat{g}_{d'}^r| \sim \mathcal{C}\mathcal{N}(0, 1)$, $|\widehat{g}_{b-d}^r| \sim \mathcal{C}\mathcal{N}(0, 1)$. These coefficients are defined as small scale fading. The distance between the d^{th} DDT and DDR, the $(d')^{th}$ DDT and DDR and the d^{th} DDP and b are denoted by the symbols $z_d^{-\eta}$, $z_{d'}^{-\eta}$, and $z_{b-d}^{-\eta}$, respectively.

3.1.3 Network Sum Rate Calculation

The desirable sum rate for c^{th} CU over the r^{th} RB using (3.1) and (3.6) is given as [105]-[106]:

$$SR_c^r(t) = B \log_2 \left[1 + \frac{P_c^r(t) |g_c^r(t)|^2}{\mathbb{I}\mathbb{F}_{c'-c}^r + \mathbb{I}\mathbb{F}_{d-c}^r + \xi_c^r(t)} \right]. \quad (3.13)$$

Similarly, the desired sum rate for d^{th} DDP over the r^{th} RB is specified as:

$$SR_d^r(t) = B \log_2 \left[1 + \frac{P_d^b(t) |g_d^r(t)|^2}{\mathbb{I}\mathbb{F}_{d'-d}^r + \mathbb{I}\mathbb{F}_{b-d}^r + \xi_d^r(t)} \right]. \quad (3.14)$$

By utilizing equations (13) and (14), the comprehensive summation rate of the overall network can be articulated as stated in [107]:

$$SR_T^r(t) = \sum_{r=1}^R \left[\sum_{c=1}^C SR_c^r(t) + \sum_{d=1}^D SR_d^r(t) \right]. \quad (3.15)$$

3.1.4 Fairness Utility Function

Fairness plays a crucial role in the distribution of resources. It is a utility-based function that is used to increase each user's throughput. Compared to alternative scheduling algorithms, fairness has the capacity to provide superior system performance since it can resolve the conflict between throughput and fairness. In this paper, our objective is to enhance the overall network's data transmission rate by providing certain users with ample RBs whenever their channel conditions surpass the average. Additionally, we anticipate that RBs may be planned properly to prevent starving users with bad channel conditions. In the proposed model, system fairness and sum rate work together to provide D2D-C while simultaneously meeting the criteria for QoS on all links utilized by both DDPs and NOMA-based CUs. Here, to evaluate network fairness, we estimate the FUF and fair scheduling (FS) for both CUs and DDPs. The proportion of the aggregate sum rate to all CUs and DDPs in the network architecture for a time slot of long duration t is referred to as proportional fairness. Moreover, consider that each user's requirement for data rate remains identical, and it retains a boundless backlog of data with minimal SINR to ensure fairness between CUs and DDPs. According to [108], if a viable scheduling policy F exists, then a system is considered proportionately fair for the scheduling

strategy G.

$$\sum_{d=1}^D \frac{\overline{SR}_d^G - \overline{SR}_d^F}{\overline{SR}_d^F}, \quad (3.16)$$

where \overline{SR}_d^G and \overline{SR}_d^F denote the average sum rate of d^{th} DDP for scheduling G and F, respectively.

The FS F for the d^{th} DDP is defined as:

$$F = \arg \max_G \sum_{d=1}^D \overline{SR}_d^G. \quad (3.17)$$

The FUF for the d^{th} DDP assigned to a time slot with a duration of t is explicitly expressed as:

$$\mathbb{U}_d^r(t) = \frac{SR_{d,t}^r}{SR_{d,(t-1)}^r}, \quad (3.18)$$

where $SR_{d,(t-1)}^r$ represents the d^{th} DDP's average sum rate and $SR_{d,t}^r$ to a time slot of duration t is defined as:

$$SR_{d,t}^r = \begin{cases} \frac{(t-1)SR_{d,(t-1)}^r + SR_{d,t}^r}{t} & t > 1, \\ SR_{d,t}^r & t = 1. \end{cases} \quad (3.19)$$

According to [108], the value of F using (3.17) can be reformulated as:

$$F = \arg \max_G \frac{1}{(t-1)} \sum_{d=1}^D \frac{SR_{d,t}^G}{\overline{SR}_{d,(t-1)}^G}. \quad (3.20)$$

Since the current slot of time t is equal to the window's size slot of time $(t-1)$, as a result, (20) is rewritten as follows:

$$F = \arg \max_G \sum_{d=1}^D \frac{SR_{d,t}^G}{\overline{SR}_{d,(t-1)}^G}. \quad (3.21)$$

Now, (3.21) reveals that radio resources are scheduled to DDPs with a higher instantaneous sum-rate $SR_{d,t}^r$ and comparatively lower average sum-rate $SR_{b,(t-1)}^r$.

3.1.5 Total Fairness Utility Function Calculation

The FUF for the c^{th} CU is elucidated as:

$$\mathbb{U}_c^r = \frac{SR_{c,t}^r}{SR_{c,(t-1)}^r}, t > 1. \quad (3.22)$$

The FUF for the d^{th} DDP is elucidated as:

$$\mathbb{U}_d^r = \frac{SR_{d,t}^r}{SR_{d,(t-1)}^r}, t > 1. \quad (3.23)$$

Because there is no average data rate in the initial time-slot, FUF at $t = 1$ is not feasible. As a result, the case for $t > 1$ has been considered.

Now, the total FUF for all the users in the network (CUs and DDPs) is given as:

$$\mathbb{U}_{total}^r = \sum_{c=1}^C \left(\mathbb{U}_c^r + \sum_{d=1}^D \mathbb{U}_d^r \right). \quad (3.24)$$

3.1.6 Problem Formulation

The goal of this study is to optimize the overall network's sum rate and fairness among NOMA-enabled CUs and DDPs while considering the resource allocation, power allocation, and QoS demands of CUs and DDPs. The optimization problem is mathematically formulated as follows:

$$\begin{aligned} P.F. & : \max_{\alpha, \beta, P_c^r} [SR_T^r(t) + \mathbb{U}_{total}^r], & (3.25) \\ s.t. \quad \mathbb{V}_1 & : P_c^r \leq P_B, & \forall c \in \mathcal{C} \\ \mathbb{V}_2 & : P_d^r \leq P_d^{r, \max}, & \forall d \in \mathcal{D} \\ \mathbb{V}_3 & : P_c^r \geq 0, & \forall c \in \mathcal{C} \\ \mathbb{V}_4 & : P_d^r \geq 0, & \forall d \in \mathcal{D} \\ \mathbb{V}_5 & : 2 \leq \sum_{c=1}^C C_c^r + \sum_{d=1}^D D_d^r \leq |C| + |D|, \end{aligned}$$

$$\begin{aligned}
\mathbb{V}_6 & : \sum_{d=1}^D \beta_{d,c}^r P_d |g_{d,c}|^2 \leq IF_c^{\text{threshold}}, \quad \forall c \in \mathcal{C} \\
\mathbb{V}_7 & : \Upsilon_c^r \geq \Upsilon_c^{r,\min} \quad \forall c \in \mathcal{C} \\
\mathbb{V}_8 & : \Upsilon_d^r \geq \Upsilon_d^{r,\min} \quad \forall d \in \mathcal{D} \\
\mathbb{V}_9 & : 0 \leq \sum_{d=1}^d \beta_d^r \leq 1 \quad \forall r \in \mathcal{R} \\
\mathbb{V}_{10} & : \sum_{d=1}^D \frac{SR_{d,t}^G}{SR_{d,(t-1)}^G} \quad \forall d \in \mathcal{D},
\end{aligned}$$

where the constraint \mathbb{V}_1 assures that the power assigned to CU should be less than the transmitted power of BS and \mathbb{V}_2 assures that each DDT power is no more than the maximum transmission power. Constraints \mathbb{V}_1 and \mathbb{V}_2 ensure the acceptable QoS requirements of the network. Constraint \mathbb{V}_3 implies that the CUs power must take on values that are non-negative integers. Constraint \mathbb{V}_3 implies that the DDTs power must take on values that are non-negative integers. According to the limitation denoted as \mathbb{V}_5 , the aggregate count of users, comprising both CUs and DDPs, within the cell may range from 2 to the summation of $|\mathcal{C}|$ and $|\mathcal{D}|$. Constraint \mathbb{V}_6 ensures that the overall interference generated by all DDP does not above the maximum interference threshold. The constraints \mathbb{V}_7 and \mathbb{V}_8 indicate the c^{th} CU's and d^{th} DDP's lowest QoS requirements, respectively. Constraint \mathbb{V}_9 guarantees that each DDP get no more than one RB. \mathbb{V}_{10} denotes that the RB has been assigned to DDPs that have high instant data while having a poor average data rate.

3.2 PRELIMINARIES OF DRL

The system can be established through the utilization of two techniques, namely value search (VS) and policy search (PS), when employing DRL. The methodology of VS utilizes the difference in the rewards garnered from two distinct samples. The value function is subsequently adjusted in accordance with said difference. On the contrary, the technique of policy search directly discovers the policy for problems. To formulate the problem as discussed in the previous section, MDP is characterized by five tuples $(S, A, P, \gamma, \Gamma)$, where S defined as a set of an agent's finite states, A defined as a set of finite states taken by agents, P defined as a

probability matrix that expresses the probability of transition from one state to another state, γ defined as a set of rewards based on the current situation and the action taken, and Γ defined as a discount factor.

3.2.1 Value Function (VF)

The methodology of utilizing VF approaches relies on evaluating the value of a specific state. If policy π commences from state s , the computation of the state VF follows the policy in the following way:

$$\mathbb{V}_\pi(s) = \mathbb{E}\{\gamma|s_t, \pi\}. \quad (3.26)$$

In (24) $\mathbb{E}\{.\}$ denotes the expectation operator and is relies on the $\mathcal{P}_{ss'}(a) = \mathcal{P}(s'|s, a)$ transition function as well as the stochastic characteristic of policy π . Now state-VF $\mathbb{V}_*(s)$ according to optimized policy π_* can be defined:

$$\mathbb{V}_*(s) = \max_{\pi} \mathbb{V}_\pi(s), s \in S. \quad (3.27)$$

The agent in network performs an action a according to the best policy for maximizing the cumulative reward. The optimized policy is that which fulfills the Bellman equation, as stated by [109] and is given as:

$$\mathbb{V}_\pi(s) = \mathbb{V}_{\pi^*} = \max_{\pi} \{\mathbb{E}(\gamma, (s, a)) + \Gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}(a) \mathbb{V}_*(s')\}. \quad (3.28)$$

The agent is bestowed with a reward upon execution of an action a in a given state s , in adherence to the policy π . The action -VF has been formulated in accordance with the award that has been received, and it is described as:

$$\mathbb{Q}_\pi(s, a) = \mathbb{E}(\gamma, (s, a)) + \Gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}(a) \mathbb{V}_*(s'). \quad (3.29)$$

Now substituting $\mathbb{Q}_*(s, a) = \mathbb{Q}_{\pi^*}$ for the optimal policy

$$\mathbb{V}_\pi(s) = \max_{a \in \mathcal{A}} \mathbb{Q}_*(s, a). \quad (3.30)$$

3.2.2 Policy Search (PS)

The utilization of VF techniques centers on the assessment of the value in a state that is already known. On the contrary, by implementing policy search techniques, agents are able to autonomously determine the optimal policy. The policy gradient is the most commonly used policy search approach. A large set of parameters is selected in the policy gradient method for effective sampling. According to policy π , the reward function is formulated in the following manner:

$$\mathbb{J}(\phi_\pi) = \sum_{s \in \mathcal{S}} \mathbb{D}_\pi(s) \sum_{a \in \mathcal{A}} \pi_\phi(a|s) \gamma_\pi(s, a), \quad (3.31)$$

where policy-parameters vector is denoted by ϕ_π , the Markov chain's static distribution according to the policy π_ϕ is represented by $\mathbb{D}_\pi(s)$. GD is used to find the optimize policy for adjusting the parameters dependent upon on $\nabla_\phi \mathbb{J}(\phi_\pi)$. For any MDP, gradient is given as [110]

$$\nabla_\phi \mathbb{J}(\phi_\pi) = \sum_{s \in \mathcal{S}} \mathbb{D}_\pi(s) \sum_{a \in \mathcal{A}} \pi_\phi(a|s) \mathbb{Q}_\pi(s, a). \quad (3.32)$$

Now, (3.30) can be expressed in following manner:

$$\nabla_\phi \mathbb{J}(\phi_\pi) = \mathbb{E}_{\pi_\phi} [\nabla_\phi \ln \pi_\phi(s, a) \mathbb{Q}_\pi(s, a)]. \quad (3.33)$$

The RL algorithm estimates the return by adjusting the parameters ϕ_π using MCPG approaches. The attainment of optimized policy parameter ϕ_{π^*} can be realized in the following manner:

$$\phi_{\pi^*} = \arg \max_{\phi_\pi} \mathbb{E} \left[\sum_a \pi(a|s; \phi_\pi) r(s, a) \right]. \quad (3.34)$$

The GD is represented as:

$$\nabla \phi_\pi = \mathbb{E}_\pi [\nabla \phi_\pi \ln \pi(a|s; \phi_\pi) r(s, a) |_{s=s^t, a=a^t}] \quad (3.35)$$

The parameter ϕ_π can be updated by using gradient ascent, as:

$$\phi_\pi \leftarrow \phi_\pi + \delta \nabla \phi_\pi, \quad (3.36)$$

whereas δ refers to a parameter denoting the size of the step, it should be noted that the admissible range of δ falls within the interval of 0 to 1, exclusive. Now, based on the maximal probability, the most suitable course of action a^* can be taken as:

$$a^* = \arg \max_a \pi(a|s; \phi_\pi). \quad (3.37)$$

3.2.3 Actor Method

The actor network (AC-N) employs the PG approach to assess and boost parametric policies. When this approach is used with the GD, it provides enhancements to the policies while iteratively modifying the parameters of the policy network.

3.2.4 Critic Method

The CR-N has the important role of evaluating the AC-N's performance. It comprises of a pair of DNN networks that bear resemblance to the AC-N, namely the critic Q network (CR-QN) and the target Q network (T-QN). Traditional RL methods based on tables exhibit strong performance in scenarios where the state space is limited and the action space is categorical. Nevertheless, the act of approximating the VF involves estimating it through a specific set of parameters. The reduction of CC results in a decrease in the dimension of input samples, which leads to an enhancement of generality and eliminates the problem of over-fitting.

3.3 Proposed Scheme

First, the optimization issue mentioned in (23) is reformulated as an MDP model. The AGMA approach is then combined with DDPG to reduce CR-CI and enhance fairness among the CUs. However, it is revealed that the DDPs do not train concurrently because they are unable to access the real-time, instantaneous global CSI. Therefore, fairness is reduced, and CO-CI increases. To overcome these issues, a D3PG scheme is proposed. Co-channel interference occurs in D3PG because any DDT does not know the transmitting powers of other DDTs. Finally, the CO-D3PG scheme is proposed to overcome the problem of D3PG. The diagram

depicting the flow of the suggested approaches can be observed in Fig. 3.2.

3.3.1 MDP

In this article, we aim to maximize the sum-rate and fairness among NOMA-enabled CUs and DDPs. Hence, it is necessary to restructure the optimization problem as a Markov MDP for better optimization. The following is a description of the MDP for the optimization problem:

$$MDP = (S, A, P, \gamma, \Gamma). \quad (3.38)$$

The game using centralized optimization technique for the aforementioned model can be formulated as follows:

3.3.1.1 Agent

In the network architecture that is being proposed, the BS is positioned as an agent located at the cell's center. All the information is managed in a centralized fashion at the BS. At the start of every time-interval, the next action for each network element will be performed.

3.3.1.2 Network Space

To improve the sum rate of CUs and DDPs, the agent (BS) interacts with the parameters of the environment. Through interaction with environmental parameters like channel gains, interferences, and PUF, the agent obtains knowledge of all local information. The state space can be expressed in terms of environmental parameters as follows:

$$s_{b,t}^r = [g_c^r, g_{b-d}^r, g_{d-c}^r, g_{d'-d}^r, \mathbb{I}\mathbb{F}_{d-c}^r, \mathbb{I}\mathbb{F}_{b-d}^r, \mathbb{I}\mathbb{F}_{c'-c}^r, \mathbb{I}\mathbb{F}_{d'-d}^r, \mathbb{U}_c^r, \mathbb{U}_d^r]. \quad (3.39)$$

3.3.1.3 Network Action

In a network that integrates NOMA technology, our objective is to enhance the network's overall sum-rate in relation to (13). Therefore, the agent's overall action can be described as

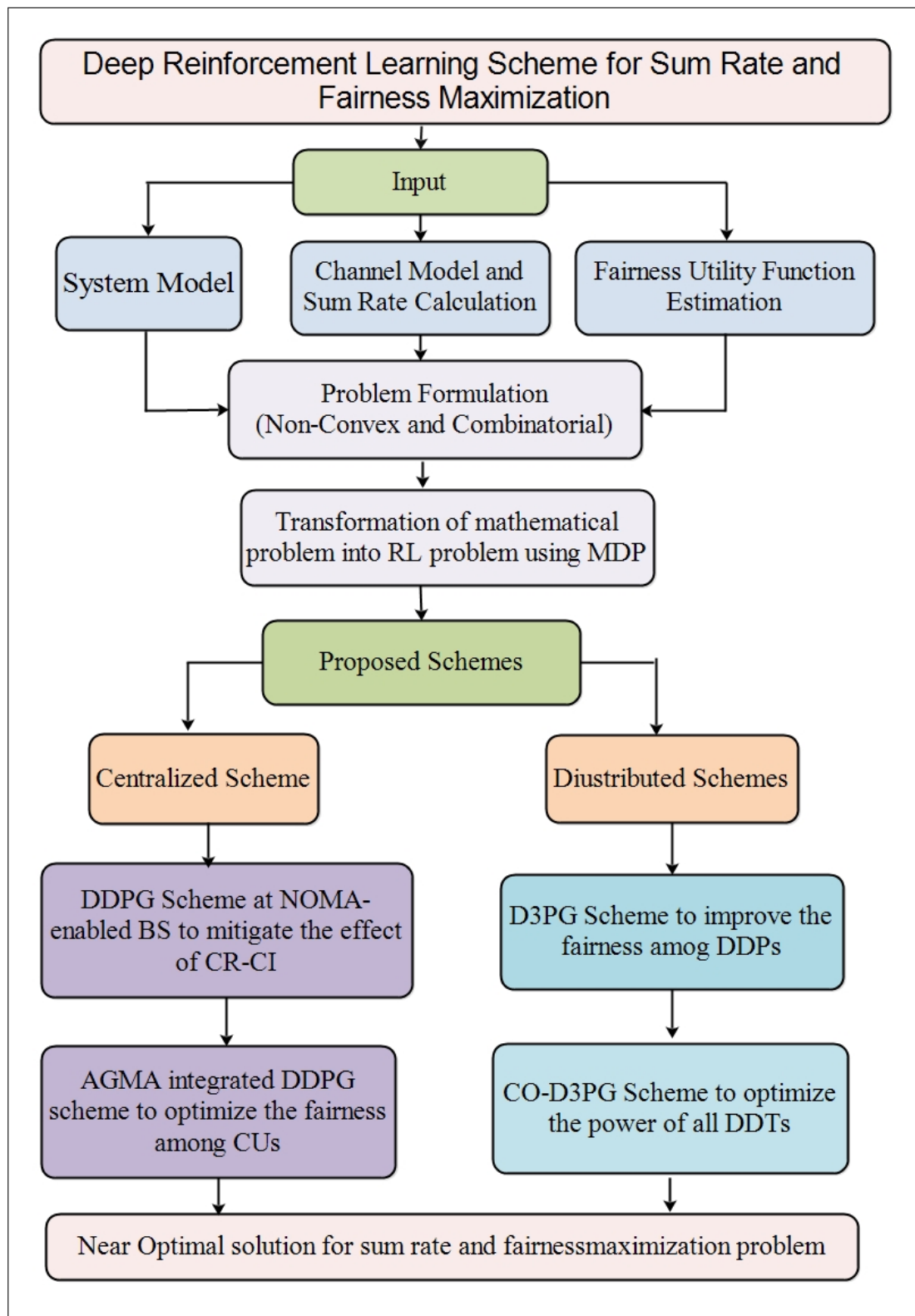


Figure 3.2: Flow Diagram of the Proposed Scheme.

follows:

$$a_{b,t}^r = [(P_1^r, P_c^r, \dots, P_C^r); (P_1^r, P_d^r, \dots, P_D^r), \alpha_c^r(t), \beta_d^r(t)]. \quad (3.40)$$

The agent executes the activity $a_{b,t}^r$ while in the state $s_{b,t}^r$ and subsequently advances to the following state $s_{b,t+1}^r$.

3.3.1.4 Network Reward function

The design of reward functions has a crucial role in the training phase. The decision-taking capability of a policy is analyzed by the reward function. With assistance from the interactions of its surroundings, agent (b) determines the optimal course of action to maximize its rewards. For the optimization the overall sum-rate and meet the QoS needs of CUs as well as DDPs, the network's immediate reward function with respect to constraints \mathbb{V}_7 and \mathbb{V}_8 over the time slot t is represented as follows:

$$\begin{aligned} \gamma(t) &= \sum_{r=1}^R \left[\sum_{c=1}^C SR_c^r(t) + \sum_{d=1}^D SR_d^r(t) \right] \\ &= S_T^r(t). \end{aligned} \quad (3.41)$$

The immediate reward function in (3.41) is composed of two factors. The first factor is the data rate of CUs and DDPs, and the second factor is the SINR needs of CUs and DDPs. The immediate reward function is designed in such a way that if SINR needs are satisfied, then the reward function has a positive value. If SINR needs are not satisfied, then the reward function has a negative value. Hence, the reward function can be mathematically defined as follows:

$$\gamma(t) = \begin{cases} \text{positive} & \text{if } \Upsilon_c^r \geq \Upsilon_c^{r,min}, \Upsilon_d^r \geq \Upsilon_d^{r,min} \\ \text{negative} & \text{otherwise.} \end{cases} \quad (3.42)$$

Algorithm 1 Sum Rate and Fairness Maximization for DDPs Using Centralized Optimization Technique.

Input

- Environment: (a) DDPs and CUs (b) BS with NOMA scheme.
- $\Upsilon_c^r \geq \Upsilon_c^{r,\min}$: Minimum SINR requirement of CUs
- $\Upsilon_d^r \geq \Upsilon_d^{r,\min}$: Minimum SINR Requirements of DDPs

Initialization:

- CR-N : $\mathbb{Q}(s, a, \phi)$.
- AC-N : $v(s; \theta)$.
- T-CR-N : $\mathbb{Q}'(s, a, \phi')$.
- T-AC-N : $v'(s; \theta')$.
- C = Replay memory of fixed size

- 1: **for** episode = 1, ..., Θ **do**
- 2: Initiate an action exploration process.
- 3: Retrieve the observation's starting state s^0 .
- 4: **for** iteration = 1, ..., \mathcal{T} **do**
- 5: Perform action $a_{b,t}^r$ that was obtained at state $s_{b,t}^r$
- 6: Renew γ_t in (40)
- 7: Select state $s_{b,t+1}^r$
- 8: Store the tuple $(s_{b,t}^r, a_{b,t}^r, \gamma_t, s_{b,t+1}^r)$ in C
- 9: From the C , take a random mini-batch E of tuples $(s_{b,t}^r, a_{b,t}^r, \gamma_t, s_{b,t+1}^r)$.
- 10: Modify L^f using (3.43)
- 11: Modify actor policy parameter $\nabla_{\theta, \mathbb{J}}$ using (3.46)
- 12: Renew the ϕ' and θ' according to (3.47) and (3.48)
- 13: Renew the state $s_{b,t}^r = s_{b,t+1}^r$
- 14: **end for**
- 15: **end for**
- 16: **Output:** α, β

3.3.2 DDPG For Centralized Optimization

We have formulated the sum rate game. Now, a DRL-based scheme called DDPG is proposed to allocate the resources to CUs with NOMA. The agent (BS) interacts with the environmental parameters in order to explore the best optimal policy. DDPG is an off-line, model-free method for training a network's continuous actions. DDPG is a combination of two components: an AC-N and a CR-N. The AC-N depends on value functionality, whereas the CR-N depends on PS. We utilize experience replay and slower-training target networks to enhance the convergence rate and reduce superfluous calculations. The experience replay buffer records the finished transition $(s_{b,t}^r, a_{b,t}^r, \gamma_{b,t}^r, s_{b,t+1}^r)$ with a storage capacity of C that is bounded. The recording ensures smooth and seamless operation. After gathering sufficient observations, we choose a tiny chunk of size E arbitrarily from the bounded storage capacity of C . The neural network is trained with these small batches. For updating newer samples and eliminating previous samples, storage C is allocated to a fixed size. Also, T-N are used to estimate the target value for the critic as well as for the AC-Ns.

Defining critic network as $\mathbb{Q}(s_{b,t}^r, a_{b,t}^r; \phi)$ with parameter ϕ and target network as $\mathbb{Q}'(s_{b,t}^r, a_{b,t}^r; \phi')$ with parameter ϕ' . Furthermore, defining actor network as $v(s_{b,t}^r; \theta)$ with θ and target actor network as $v'(s_{b,t}^r; \theta')$ with parameter θ' . To modify the CR-N for each learning phase, the loss function must be minimized, which is as follows:

$$L^f(\phi) = \frac{1}{E} \sum_{e=1}^E (y_e - \mathbb{Q}(s_e^r, a_e^r; \phi))^2. \quad (3.43)$$

where y_e is the target value and can be calculated as follows:

$$y_e = \gamma_e(s_e^r, a_e^r) + \Gamma \mathbb{Q}(s_{e+1}^r, a_{e+1}^r; \phi')|_{a_{e+1}^r} = v'(s_{e+1}^r; \theta'). \quad (3.44)$$

The temporal-difference error (TDE) can be represented as:

$$\vartheta(t) = y_t - \mathbb{Q}(s_{d,t}^r, a_{d,t}^r; \phi) \quad (3.45)$$

The following is the updated value of the actor network parameter:

$$\nabla_{\theta_t} \mathbb{J} \approx \frac{1}{E} \sum_{e=1}^E \nabla_{a_e^r} \mathbb{Q}(s_e^r, a_e^r; \phi) |_{a_e^r=v(s_e^r)} \nabla_{\theta} v(s_e^r; \theta). \quad (3.46)$$

Now, variable ϕ' and variable θ' are changed via soft target updates. Therefore, the updated value of the parameters can be estimated as:

$$\phi' \leftarrow \chi \phi + (1 - \chi) \phi' \quad (3.47)$$

$$\theta' \leftarrow \chi \theta + (1 - \chi) \theta', \quad (3.48)$$

where $0 < \chi < 1$ is the range of χ and χ is referred to as the learning rate and is a hyper-parameter.

DDPG is a model-free approach for training the continuous action of a system in an off-line manner. As a result, we introduce a noise procedure of $\mathcal{L}[0,1]$ for exploration and exploitation. Now, the target actor network can be formulated as:

$$v'(s_t^r; \phi_t') = v(s_t^r; \phi_t) + \mathcal{L} \chi (0, 1). \quad (3.49)$$

The details of the centralized optimization technique are shown in Algorithm 1, which is used to allocate power at the NOMA-integrated BS in the downlink situation. The count of maximal episodes is expressed by Θ , whereas the time interval is expressed by \mathcal{T} .

3.3.3 Arithmetic-Geometric Mean Approximation Scheme For CUs Power Allocation

In this section, the Arithmetic-Geometric Mean Approximation (AGMA) scheme is applied to improve fairness among CUs. To achieve the target, we are going to optimize the power of CUs. According to the work in [111], AGMA provides a solution to its non-convexity. AGMA proves to be a highly efficient technique when it comes to resolving power allocation issues [112, 113]. First, the proposed scheme recognized that all CUs have higher channel gains. Then power allocation is done with the interference part only comprised of the user's

Algorithm 2 AGMA-Based Power Allocation Algorithm for CUs.**Input**

- Environment: (a) DDPs (b) CUs (c) BS with NOMA scheme.
- $\Upsilon_c^r \geq \Upsilon_c^{r,\min}$: Minimum SINR requirement of CUs
- $\Upsilon_d^r \geq \Upsilon_d^{r,\min}$: Minimum SINR Requirements of DDPs

Initialization:

- CR-N: $\mathbb{Q}(s, a, \phi)$; AC-N: $v(s; \theta)$; T-CT-N: $\mathbb{Q}'(s, a, \phi')$; T-AC-N: $v'(s; \theta')$; and C = Replay memory of fixed size.

1: **for** episode = 1, ..., Θ **do**

2: Initiate an action exploration process.

3: Retrieve the observation's stating state s^0 .

4: **for** iteration = 1, ..., \mathcal{T} **do**

5: Perform action $a_{b,t}^r$ that was obtained at state $s_{b,t}^r$

6: Renew γ_t in (40)

7: Select state $s_{b,t+1}^r$

8: Store the tuple $(s_{b,t}^r, a_{b,t}^r, \gamma_t, s_{b,t+1}^r)$ in C

9: From the C , take a random mini-batch E of tuples $(s_{b,t}^r, a_{b,t}^r, \gamma_t, s_{b,t+1}^r)$.

10: Modify L^f using (3.43)

11: Modify actor policy parameter $\nabla_{\theta_t} \mathbb{J}$ using (3.46)

12: Renew the ϕ' and θ' according to (3.47) and (3.48)

13: Renew the state $s_{b,t}^r = s_{b,t+1}^r$

14: **end for**

15: initiate $t = 1$;

16: **repeat**

17: Formulate the coefficients $\bar{A}_c, \hat{A}_c, \tilde{A}_c$ $\bar{A}_c = \frac{P_c^r[t-1]|g_c^r|^2}{u_c^r(P[t-1])}$ $\hat{A}_c = \frac{\hat{\xi}_d^r(t)}{u_c^r(P[t-1])}$,

$$\tilde{A}_c = \frac{P_c^r[t-1]|g_c^r|^2}{u_c^r(P[t-1])}.$$

18: Formulate monomial

$$\underline{u}_c^r = \sum_{c' \in \mathcal{C}} \left(\frac{P_{c'}^r[t-1]|g_{c'}^r|^2}{\bar{A}_c} \right)^{\bar{A}_c} \left(\frac{\hat{\xi}_d^r(t)}{\hat{A}_c} \right)^{\hat{A}_c} X \left(\frac{P_c^r[t-1]|g_c^r|^2}{\tilde{A}_c} \right)^{\tilde{A}_c}. \quad (3.50)$$

19: Solve geometric program (3.55) using $u_c^r(P)[t]$.

20: Set $t = t + 1$;

21: **until** convergence of P

22: **end for**

23: **Output:** P_d^r

CUs with greater channel.

The formulation of power allocation for CUs is presented as follows:

$$\begin{aligned}
P.F.1 & : B \log_2 \left[1 + \frac{P_c^r(t) |g_c^r(t)|^2}{\mathbb{I}\mathbb{F}_{c'-c}^r + \mathbb{I}\mathbb{F}_{d-c}^r + \xi_d^r(t)} \right] \\
s.t. \ \forall_1 & : P_c^r \leq P_B, \quad \forall c \in \mathcal{C} \\
\forall_3 & : P_c^r \geq 0, \quad \forall c \in \mathcal{C}
\end{aligned} \tag{3.51}$$

The problem defined in (3.51) is a non-convex problem because it consists of the rate function as revealed in (3.13). To address this problem, we use the AGMA scheme.

Now the achievable sum rate of c^{th} CU can be reformulated as follows:

$$\begin{aligned}
SR_c^r(t) & = \\
B \log_2 & \left[\frac{\sum_{c' \in \overline{\mathcal{C}}(c)} P_{c'}^r(t) |g_{c'}^r(t)|^2 + \widehat{\xi}_d^r(t) + P_c^r(t) |g_c^r(t)|^2}{\sum_{c' \in \overline{\mathcal{C}}(c)} P_{c'}^r(t) |g_{c'}^r(t)|^2 + \widehat{\xi}_c^r(t)} \right],
\end{aligned} \tag{3.52}$$

where subset $\overline{\mathcal{C}}(c)$ is defined as $\overline{\mathcal{C}}(c) \triangleq \{c' \in \mathcal{C}\}$.

Since DDP's transmitting powers are constant, they can be considered as noise. Now, noise can be defined as:

$$\widehat{\xi}_c^r(t) = \sum_{d \in \mathcal{D}} \beta_d^r(t) P_d^r(t) |g_{d-c}^r(t)|^2 + \xi_c^r(t) \tag{3.53}$$

Problem (3.52) can be shown to be identical to

$$\begin{aligned}
P.F.2 & : \min_P \sum_{r \in \mathcal{R}} \sum_{m \in \mathcal{M}} B \log_2 \\
& \left[\frac{\sum_{c' \in \overline{\mathcal{C}}(c)} P_{c'}^r(t) |g_{c'}^r(t)|^2 + \widehat{\xi}_c^r(t)}{\sum_{c' \in \overline{\mathcal{C}}(c)} P_{c'}^r(t) |g_{c'}^r(t)|^2 + \widehat{\xi}_d^r(t) + P_c^r(t) |g_c^r(t)|^2} \right] \\
s.t. \ \forall_1 & : P_c^r \leq P_B, \quad \forall c \in \mathcal{C} \\
\forall_3 & : P_c^r \geq 0, \quad \forall c \in \mathcal{C}
\end{aligned} \tag{3.54}$$

let define $u_c^r(P) = \sum_{c' \in \overline{\mathcal{C}}(c)} P_{c'}^r(t) |g_c^r(t)|^2 + \widehat{\xi}_d^r(t) + P_c^r(t) |g_c^r(t)|^2$, then according to the arithmetic-geometric mean (AGM) inequality

$$u_c^r \geq \underline{u}_c^r = \sum_{c' \in \overline{\mathcal{C}}} \left(\frac{P_{c'}^r |g_c^r|^2}{\bar{A}_c} \right)^{\bar{A}_c} \left(\frac{\widehat{\xi}_d^r(t)}{\widehat{A}_c} \right)^{\widehat{A}_c} \left(\frac{P_c^r |g_c^r|^2}{\widetilde{A}_c} \right)^{\widetilde{A}_c}, \quad (3.55)$$

where \bar{A}_c is defined as $\bar{A}_c = \frac{P_{c'}^r |g_c^r|^2}{\underline{u}_c^r(P)}$, \widehat{A}_c is defined as $\widehat{A}_c = \frac{\widehat{\xi}_d^r(t)}{\underline{u}_c^r(P)}$, and \widetilde{A}_c is defined as $\widetilde{A}_c = \frac{P_c^r |g_c^r|^2}{\underline{u}_c^r(P)}$ for all $c \in \mathcal{C}, c' \in \overline{\mathcal{C}}$. The approximation problem below can be categorized as a geometric program.

$$\begin{aligned} P.F.3 : \min_P \sum_{r \in \mathcal{R}} \sum_{m \in \mathcal{M}} B \log_2 & \\ \left[\frac{\sum_{c' \in \overline{\mathcal{C}}(c)} P_{c'}^r(t) |g_c^r(t)|^2 + \widehat{\xi}_c^r(t)}{\underline{u}_c^r(P)} \right] & \\ s.t. \quad \mathbb{V}_1 : P_c^r \leq P_B, \quad \forall c \in \mathcal{C} & \\ \mathbb{V}_3 : P_c^r \geq 0, \quad \forall c \in \mathcal{C} & \end{aligned} \quad (3.56)$$

A geometric program, as in (3.56) can be converted into a convex issue [34] by utilizing a logarithmic variation of variables. The update of the approximate parameters involves utilizing the solution of the present iteration, which is defined as $P(t)$ in equation (3.55). Furthermore, when we solve (3.56), the accuracy improves with each iteration.

3.4 Proposed Scheme

3.4.1 Architecture of Multi-Agent Power Control Scheme With DRL

In this section, we used Multi-Agent Distributed Deep Deterministic Policy Gradient (*D3PG*) for DDPs as shown in Fig. 3.3. In algorithm 1, we used centralized DDPG to control the power. But as the number of DDPs increases, there may be a chance to increase the interference, which also increases the hardware complexity and burden of the BS. Furthermore, the transmitted power of DDTs is closely correlated with the instantaneous global CSI. DDTs, on the other

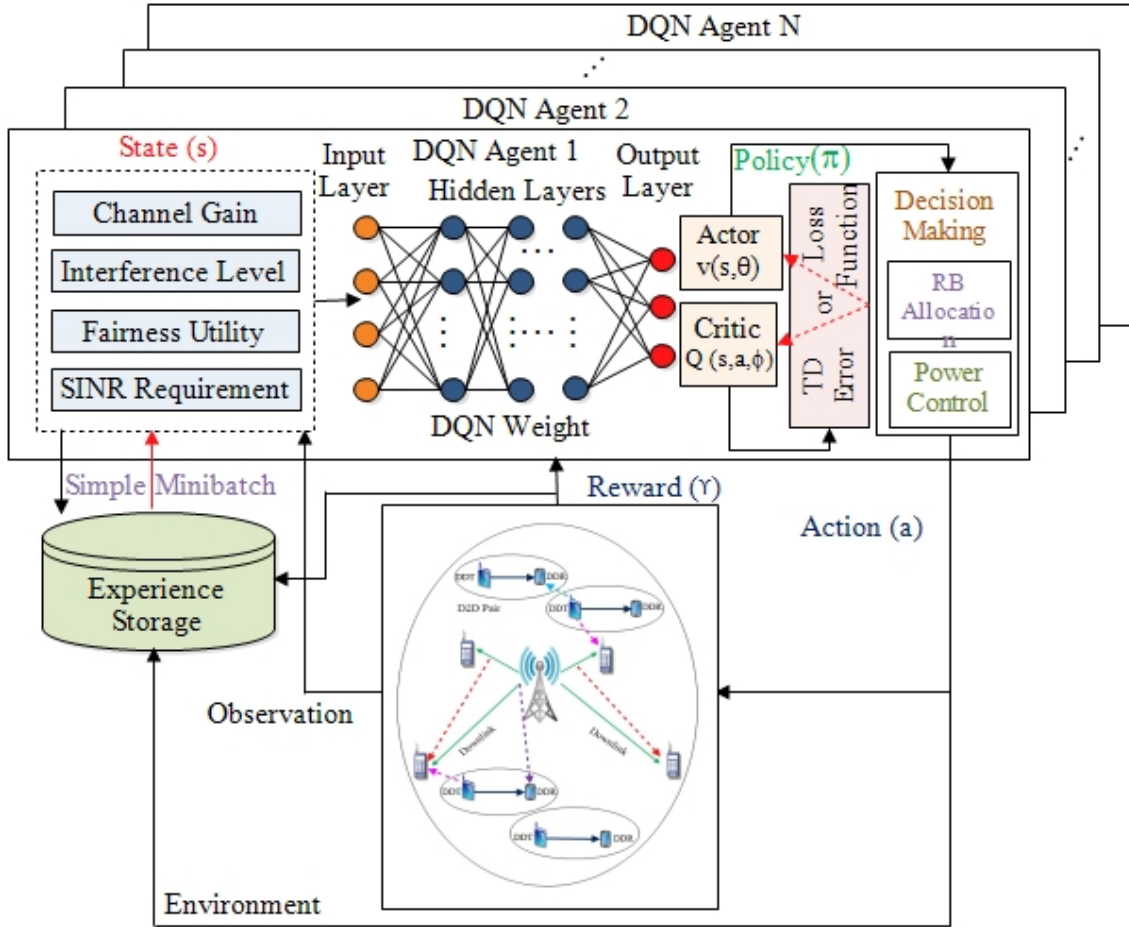


Figure 3.3: Multi-Agent Power Control Scheme With DRL

hand, do not have access to instantaneous global CSI. As a result, using conventional power control techniques to optimize transmit power is impractical for each DDT. So, to overcome these issues, we are going to use multi-agent *D3PG* [114].

We used the *D3PG* scheme for our system so that each DDP could work cooperatively. Now, each DDP is an agent; therefore, the state space is $s_{d,t}^r = [g_c^r, g_{b-d}^r, g_{d-c}^r, \mathbb{I}\mathbb{F}_{d-c}^r, \mathbb{I}\mathbb{F}_{b-d}^r, \mathbb{I}\mathbb{F}_{c'-c}^r, \mathbb{U}_{d,t}^r]$. The action space for each DDP is transmitted power in slot t , i.e., $a_{d,t}^r = \{\alpha, \beta, P_c\}$. In the t^{th} time slot, the modified reward function can be described as:

$$\begin{aligned} \gamma_{d,t}^r &= \mathbb{U}_{total}^r - Z_1 \sum_{c=1}^C (\Upsilon_c^r - \Upsilon_c^{r,min}) \\ &\quad - Z_2 \sum_{d=1}^D (\Upsilon_d^r - \Upsilon_d^{r,min}) \end{aligned} \quad (3.57)$$

The first term in (3.57) denotes the total FUF for all the users in the network (CUs and DDPs). The 2nd and 3rd expressions represent the minimal data rate requirements that have not been fulfilled by the CUs and DDPs, respectively. Z_1 and Z_2 represent the weights of the second and third terms, respectively.

We add a small amount of white noise $w_n \in (0, 1)$ to the exploration because exploration is very difficult in the training of agents if action spaces are continuous. After combining noise, the Gaussian exploratory policy for actor-based DNN can be constructed as follows:

$$v'(s_{d,t}^r) = v(a_{d,t}^r | s_{b,t}^r, \theta) + w_n. \quad (3.58)$$

By using temporal-difference error, the probability of k^{th} sample of experience replay buffer can be written as follows:

$$\mathbb{P}_k^{per} = \frac{\Psi_k^{\varpi_1}}{\sum_{k' \neq k}^K \Psi_{k'}^{\varpi_1}}, \quad (3.59)$$

where $\Psi_k = \frac{1}{rank(k)} > 0$ denoted as the priority given to the k th sample of experience replay, $rank(k)$ represents the experience replay's k^{th} sample rank, and ϖ_1 denoted the priority control factor.

The ϖ_1 is defined as follows:

$$\varpi_1 = \begin{cases} 1 & \text{for the priority condition,} \\ 0 & \text{for the uniform condition.} \end{cases} \quad (3.60)$$

The samples have a considerable TD-error rate, so a large number of repeats are required. The frequency of state transitions is increased as a result. The DQN learning process diverges as a result of this increase in transition and destabilizes the RL technique. We utilize prioritize sample weights to stabilize the RL schemes, which are expressed as follows:

$$\Phi_k = \frac{1}{C(\mathbb{P}_k^{per})^{\varpi_2}} \quad (3.61)$$

where C and $\bar{\omega}_2$ are the replay buffer size and corrective control value, respectively.

The critic is now modified by reducing the loss function in the following manner:

$$\bar{L}^f(\phi) = \sum_{k=1}^K \left(\Phi_k L_k^f(\phi) \right) \quad (3.62)$$

The updated value of AC-N parameter is expressed as follows:

$$\nabla_{\theta} \mathbb{J} = \sum_{k=1}^K \left(\sum_{s_k \in \mathcal{S}} D^{\pi}(s_k) \sum_{a_k \in \mathcal{A}} \pi_{\phi}(a_k | s_k) Q^{\pi}(s_k, a_k) \right). \quad (3.63)$$

Now using (3.61) and (3.62), the modified value of CR-N and AC-N can be stated as follows:

$$\phi_{(d,t+1)} = \phi_{(d,t)} + \delta_c(t) \vartheta(t) \nabla_{\phi} L_{(d,t)}^f. \quad (3.64)$$

$$\theta_{(d,t+1)} = \theta_{(d,t)} + \delta_a(t) \nabla_{\theta} \mathcal{K}. \quad (3.65)$$

Now, variable ϕ' and variable θ' are changed via soft target updates. Therefore, the defined parameters have been modified in the following manner::

$$\phi'_{(d,t+1)} \leftarrow \chi \phi_{(d,t)} + (1 - \chi) \phi'_{(d,t)} \quad (3.66)$$

$$\theta' \leftarrow \chi \theta + (1 - \chi) \theta', \quad (3.67)$$

3.4.2 Fair Resource Allocation Scheme Using CO-D3PG

CO-CI interference occurs in multi-agent D3PG because any DDT does not know the transmitting powers of other DDTs. To mitigate the co-channel interference, an optimizer is typically used to adjust the agent's power over an RB. However, when all DDTs transmit at the same time, the optimizer is unable to adjust each DDT's transmitting power. To overcome this issue, a conventional optimizing scheme such as successive convex approximation (SCA) is integrated with D3PG and termed CO-D3PG.

Algorithm 3 Multi-Agent D3PG Training Algorithm for Power Control of DDPs.**Input**

- Environment: (a) DDPs and CUs (b) BS with NOMA scheme.
- $\Upsilon_c^r \geq \Upsilon_c^{r,\min}$: Minimum SINR requirement of CUs
- $\Upsilon_d^r \geq \Upsilon_d^{r,\min}$: Minimum SINR Requirements of DDPs

Initialization:

- CR-N : $\mathbb{Q}(s, a, \phi)$.
- CC-N : $v(s; \theta)$.
- T-CR-N : $\mathbb{Q}'(s, a, \phi')$.
- T-AC-N : $v'(s; \theta')$.
- C = Replay memory of fixed size

```

1: for ( $d = 1; d \leq D; d ++$ ) do
2:   for ( $r = 1; r \leq R; r ++$ ) do
3:     for ( $c = 1; c \leq C; c ++$ ) do
4:       Each DDP notes the starting state  $s_t$ ;
5:       for ( $t = 1; t \leq T; t ++$ ) do
6:         Action  $a_{d,t}^r$  is taken by DDP ;
7:         Calculate  $SINR$  according to (3.10) ;
8:         Modify the reward  $\gamma_t$  according to (3.57);
9:         Choose state  $s_{d,t+1}^k$ ;
10:        Save  $c_{d,t}^r = (s_{b,t}^r, a_{b,t}^r, \gamma_t, s_{b,t+1}^r)$  in the replay buffer;
11:        if (Replay memory is full) then
12:          The minimal-utilized historic experience is relinquished;
13:          Choose the sample from replay buffer at random ;
14:          for ( $k = 1; K \leq K; k ++$ ) do
15:            Using (3.59), sample experience  $k$  having probability  $\mathbb{P}_k^{er}$ ;
16:            Calculate  $\Phi_k$  from (3.61);
17:          end for
18:          Modify  $\overline{L^f}(\phi)$  using (3.62)
19:          Modify  $\nabla_{\theta, \mathbb{J}}$  and  $\phi$  using (3.63) and (3.64);
20:          Update the  $\phi'$  and  $\theta'$  according to (3.66) and (3.67);
21:        end if
22:      end for
23:    end for
24:  end for
25: end for
26: Output:  $\beta, P_d^r$ 

```

Algorithm 4 CO-D3PG Training Algorithm for Fair Resource Allocation among DDPs.**Input**

- Environment: (a) DDPs and CUs (b) BS with NOMA scheme.
- $\Upsilon_c^r \geq \Upsilon_c^{r,\min}$: Minimum SINR requirement of CUs
- $\Upsilon_d^r \geq \Upsilon_d^{r,\min}$: Minimum SINR Requirements of DDPs

Initialization:

- Critic network: $\mathbb{Q}(s, a, \phi)$; Actor network: $v(s; \theta)$.
- Target critic network: $\mathbb{Q}'(s, a, \phi')$, and Target actor network: $v'(s; \theta')$.
- C = Replay memory of fixed size
- i = index of iteration, and I_{max} = Maximum number of iterations

```

1: for ( $d = 1; d \leq D; d ++$ ) do
2:   for ( $r = 1; r \leq R; r ++$ ) do
3:     for ( $e = 1; e \leq E; e ++$ ) do
4:       Each DDP notes the starting state  $s_t$  and  $\gamma_d^r$ ;
5:       for ( $t = 1; t \leq T; t ++$ ) do
6:         Action  $a_{d,t}^r$  is taken by DDP ;
7:         Calculate  $\Upsilon_d^r$  according to (10);
8:         Modify the reward  $\gamma_t$  according to (57);
9:         Choose state  $s_{d,t+1}^k$ ;
10:        Save  $e_{d,t}^r = (s_{b,t}^r, a_{b,t}^r, \gamma_t, s_{b,t+1}^r)$  in the replay buffer;
11:        if (Replay memory is full) then
12:          The minimal-utilized historic experience is relinquished;
13:          Choose the sample from replay buffer at random;
14:        end if
15:        for ( $i = 1; i \leq I_{max}; i ++$ ) do
16:          Using (3.59), sample experience  $k$  having probability  $\mathbb{P}_k^{per}$ ;
17:          Calculate  $\Phi_k$  from (3.61);
18:        end for
19:        while ( $\gamma_c^r(i) - \gamma_c^r(i-1)$ ) ||
20:        ( $|\gamma_d^r(i) - \gamma_d^r(i-1)| \geq \gamma^{th}$ ) do
21:          for ( $i = 1; i \leq I_{max}; i ++$ ) do
22:             $\gamma_c^r(i) = \gamma_c^r(i-1)$ ;
23:             $\gamma_d^r(i) = \gamma_d^r(i-1)$ ;
24:            Calculate  $F_1, F_2, \Lambda_1, \Lambda_2$  from (72)-(75)
25:            Calculate  $\delta(k)$ , where  $\delta_c(k) = 2^{\lambda_c(k)}$  and  $\delta_d(k) = 2^{\lambda_d(k)}$ ;
26:            Calculate  $\gamma_c^r(k)$  and  $\gamma_d^r(k)$  according to  $\delta_c(k)$  and  $\delta_d(k)$ ;
27:            Modify  $a_1, a_2$  and  $a_3$  according to (3.83), (3.84) and (3.85),
28:          end for
29:        end while
30:        Modify  $\bar{L}^f(\phi)$  using (3.62)
31:        Modify  $\nabla_{\theta, \mathbb{J}}$  and  $\phi$  using (3.62) and (3.64);
32:        Update the  $\phi'$  and  $\theta'$  according to (3.67) and (3.66);
33:      end for
34:    end for
35:  end for
36: end for
37: Output: Power Control of DDPs ( $P_d^r$ )

```


The optimization problem described in (3.25) is reformulated as:

$$\begin{aligned} & \arg \max_{P_d^r} \mathbb{U}_{total}^r, \\ s.t. \quad & : \mathbb{V}_1, \mathbb{V}_2 \text{ and } \mathbb{V}_7 \end{aligned} \quad (3.68)$$

The SCA approach [115] states that

$$\mathbb{U}_{total}^r = \log_2(1 + \gamma_c^r) + \log_2(1 + \gamma_d^r), \quad (3.69)$$

Where

$$\log_2(1 + \gamma_c^r) \geq F_1 \log_2 \gamma_c^r + \Lambda_1, \quad (3.70)$$

$$\log_2(1 + \gamma_d^r) \geq F_2 \log_2 \gamma_d^r + \Lambda_2. \quad (3.71)$$

With

$$F_1 = \frac{\hat{\gamma}_c^r}{1 + \hat{\gamma}_c^r}, \quad (3.72)$$

$$F_2 = \frac{\hat{\gamma}_d^r}{1 + \hat{\gamma}_d^r}, \quad (3.73)$$

$$\Lambda_1 = \log_2(1 + \hat{\gamma}_c^r) - \frac{\hat{\gamma}_c^r}{1 + \hat{\gamma}_c^r} \log_2 \hat{\gamma}_c^r, \quad (3.74)$$

$$\Lambda_2 = \log_2(1 + \hat{\gamma}_d^r) - \frac{\hat{\gamma}_d^r}{1 + \hat{\gamma}_d^r} \log_2 \hat{\gamma}_d^r. \quad (3.75)$$

Since $\hat{\gamma}_c^r \geq 0$ and $\hat{\gamma}_d^r \geq 0$, as a result his lower bound (LB) for $\gamma_c^r = \hat{\gamma}_c^r$ and $\gamma_d^r = \hat{\gamma}_d^r$ can be said to be tight.

The LB of the power control functions is stated as follows according to the inequality functions (3.70) and (3.71):

$$SR_c^r(\delta_c) + SR_d^r(\delta_d) \geq \hat{SR}_c^r(\delta_c) + \hat{SR}_d^r(\delta_d), \quad (3.76)$$

where $\hat{SR}_c^r(\delta_c)$ and $\hat{SR}_d^r(\delta_d)$ are represented in the following way:

$$\hat{SR}_c^r(\delta_c) = F_1 \log_2 \gamma_c^r + \Lambda_1, \quad (3.77)$$

$$\hat{S}R_d^r(\delta_d) = F_2 \log_2 \gamma_d^r + \Lambda_2. \quad (3.78)$$

Because (77) and (78) represent non-convex functions, we replace $\delta_c = 2^{\lambda_c}$ and $\delta_d = 2^{\lambda_d}$, such as $\lambda = [\lambda_c, \lambda_d]$. As a result, the LB maximization is rewritten as follows:

$$\begin{aligned} P.C. & : \max_{\lambda_1} \left\{ \hat{S}R_c^r(2^{\lambda_c}) + \hat{S}R_d^r(2^{\lambda_d}) \right\}, \\ s.t. & : 2^{\lambda_c} + 2^{\lambda_d} \leq 1, \\ & : \gamma_c \geq \gamma_c^{r,min}, \gamma_d \geq \gamma_d^{r,min}. \end{aligned} \quad (3.79)$$

The optimal power allocation between CU and DDT can now be achieved through the use of the Lagrangian dual decomposition method and the Karush Kuhn Tucker criteria. The Lagrangian associated with (3.69) is written as follows:

$$\begin{aligned} \mathcal{L}(\lambda, a_1, a_2, a_3) & = \left\{ \hat{S}R_c^r(2^{\lambda_c}) + \hat{S}R_d^r(2^{\lambda_d}) \right\} \\ & + a_1 \left(1 - (2^{\lambda_c} + 2^{\lambda_d}) \right) \\ & + a_2 \left(\gamma_c - \gamma_c^{r,min} \right) \\ & + a_3 \left(\gamma_d - \gamma_d^{r,min} \right), \end{aligned} \quad (3.80)$$

where the Lagrangian multipliers corresponding to constraints $\mathbb{V}_1, \mathbb{V}_2$ & \mathbb{V}_3 are a_1, a_2 , and a_3 , respectively. The dual analogous issue is obtained as follows [116] by using Lagrange dual decomposition:

$$\min_{a_1, a_2, a_3 \geq 0} \max_{\lambda} \mathcal{L}(\lambda, a_1, a_2, a_3). \quad (3.81)$$

The Karush-Kuhn-Tucker (KKT) conditions are a set of mathematical conditions that characterize the optimal solutions to nonlinear programming problems with inequality constraints. These conditions extend the Lagrange multiplier method to handle both equality and inequality constraints. The KKT conditions consist of three types: the stationarity condition, which ensures that the gradient of the objective function is a linear combination of the gradients of the constraint functions; the primal feasibility condition, ensuring that the inequality constraints are satisfied; and the dual feasibility condition, which ensures that the Lagrange multipliers

associated with the inequality constraints are non-negative. Together, these conditions provide a comprehensive framework for identifying and verifying optimal solutions in constrained optimization problems. The KKT criteria can be used to find the optimal value of λ^* . Therefore, to calculate λ^* , apply $\frac{\partial \mathcal{L}(\lambda^*, a_1, a_2, a_3)}{\partial \lambda^*} = 0$.

$$\lambda^* = \left[K \left(\frac{A_1}{1 + A_1 \lambda} + \frac{A_2}{1 + A_2 \lambda} \right) + a_1 \gamma_c^r + a_2 \gamma_d^r - a_3 \right]^+, \quad (3.82)$$

where, $K = \left(\frac{1}{\ln 2}\right)$, $A_1 = 2^{\lambda_c} \gamma_c^r$ and $A_2 = 2^{\lambda_d} \gamma_d^r$.

$$a_1(i+1) = \left[a_3(i) - \zeta \left(1 - \left(\sum_{c=1}^C \delta_c + \sum_{d=1}^D \delta_d \right) \right) \right]^+, \quad (3.83)$$

$$a_2(i+1) = \left[a_1(i) - \zeta (\gamma_c - \gamma_c^{r, \min}) \right]^+, \quad (3.84)$$

$$a_3(i+1) = \left[a_2(i) - \zeta (\gamma_d - \gamma_d^{r, \min}) \right]^+, \quad (3.85)$$

where, $[\cdot]^+ = \max\{0, \cdot\}$, the iteration index for the dual variables update is i . Therefore, every agent can update the variables a_1 , a_2 , and a_3 locally.

3.4.3 Complexity Analysis

3.4.3.1 Algorithm 1

The computational complexity (CC) for DRL-based schemes consists of two parts: the model training stage and the model conclusion stage [117]. The model training stage generates training complexity. Large volumes of data are fed into the DRL algorithm during the model training stage until the model's performance meets specific requirements. Then comes the conclusion stage, where the trained model may be used to make some predictions immediately at runtime. The overhead produced at this stage reflects conclusion complexity. The CC for every layer of the NN is derived through matrix processes and activation function estimations. Let X represent the number of layers, \mathcal{O}_m^M and \mathcal{O}_m^A represent the CC of the matrix process and activation function, respectively, for each layer with node number m . The training complexity can therefore be calculated as $(\mathcal{O}_m^M + \mathcal{O}_m^A) * X$.

3.4.3.2 Algorithm 2

:Algorithm 2 is the integration of DDPG and AGMA. The complexity of the AGMA is $\mathcal{O}(m^2)$. The accuracy of this approximation depends on the tightness of the inequality. Therefore, the CC of the Algorithm 3 is $[(\mathcal{O}_m^M + \mathcal{O}_m^A) * X + \mathcal{O}(m^2)]$.

3.4.3.3 Algorithm 3

Let W indicate the quantity of layers consisting of AC-N, and let ρ_w^A indicate the count of neurons present within the w^{th} layer. At the time of execution, the CC of the aAC-N's w^{th} layer can be expressed as $\mathcal{O}(\rho_{w-1}^A \rho_w^A + \rho_w^A \rho_{w+1}^A)$. As a result, the CC of the actor network at execution time is $\mathcal{O}(\sum_{w=2}^{W-1} (\rho_{w-1}^A \rho_w^A + \rho_w^A \rho_{w+1}^A))$. Assuming that Z signifies the count of layers present within the CR-N's, and ρ_z^C implies the tally of neurons that are situated in the z^{th} layer. The CC of the critic network's z^{th} layer and critic network is $\mathcal{O}(\rho_{z-1}^C \rho_z^C + \rho_z^C \rho_{z+1}^C)$ and $\mathcal{O}(\sum_{z=2}^{Z-1} (\rho_{z-1}^C \rho_z^C + \rho_z^C \rho_{z+1}^C))$, respectively. The training procedure involves actor network as well as critic network, hence its level of computational complexity can be estimated as $\mathcal{O}(\sum_{w=2}^{W-1} (\rho_{w-1}^A \rho_w^A + \rho_w^A \rho_{w+1}^A) + \sum_{z=2}^{Z-1} (\rho_{z-1}^C \rho_z^C + \rho_z^C \rho_{z+1}^C))$.

3.4.3.4 Algorithm 4

The Algorithm 4 is the combination of Algorithm 3 and Lagrangian method. Therefore, the CC of the Algorithm 4 is $\mathcal{O}(\sum_{w=2}^{W-1} (\rho_{w-1}^A \rho_w^A + \rho_w^A \rho_{w+1}^A) + \sum_{z=2}^{Z-1} (\rho_{z-1}^C \rho_z^C + \rho_z^C \rho_{z+1}^C) + (I_{max}(C + D)^2, \Omega))$. Here I_{max} is the maximum iteration, and Ω is the cost of evaluating the first and second derivatives of the objective and constraint functions.

3.5 Performance Assessment

In this section, the proposed DRL algorithm's performance is evaluated and compared to that of existing schemes such as DDPG, deep duelling DRL, and DQN Scheduling. It has three parts: (i) Simulation Parameters (ii) Baseline Schemes (iii) Experimental Graph and Discussion.

3.5.1 Simulation Parameters

In the proposed network architecture, we consider only one cell, in which CUs and DDPs are randomly allocated. The radius of the cell is 500 m. The transmission distance between the DDT and the DDR varies between 10 m and 50 m. We consider that several DDPs and one CU share the same RB. As a result, the count of RBs is equal to the count of CUs, which is equal to C . Furthermore, it is important to note that when a system is equipped with NOMA, the complexity of implementing SIC at the receiver escalates in direct proportion to the count of users that are allotted on the same RB. The complexity of implementing SIC represented by $\mathcal{O}(C^3)$, where C signifies the number of CUs [118]. To ensure that the complexity of the receiver remains reasonably simple, we make the assumption that only two users, namely NOMA-enabled CUs, can be allotted to an identical RB. The simulation channel model has been chosen based on 3GPP and IMT-2020 standards [119, 120]. We employed a protocol stack that included additional PHY, packet data convergence protocol (PDCP) layer, radio link control (RLC), and medium access control (MAC) to preserve direct transmission between the DDT and DDR. These protocols can also maintain the CUs' connectivity while communicating across a DDP link.

We created a DDQN structure with three completely connected layers (input layer, hidden layer, and output layer), each containing 150, 200, and 200 neurons, respectively. On the Python 3.13 platform, we utilized Keras 3.3.6 and TensorFlow 2.23 to evaluate the model. Table 3.2 summarizes the remaining simulation parameters.

3.5.2 Baseline Schemes

In this section, the baseline schemes are studied, with which we have compared the proposed scheme. The descriptions of the baseline schemes are as follows:

- **Deep Duelling DRL:** In [121], the authors used the deep duelling DRL to enhance network performance while simultaneously upholding QoS constraints in real-time setting
- **DQN Scheduling:** In [122], the authors introduced a parallel DQN approach aimed at optimizing network throughput by dynamically rewarding power optimization efforts.

- **DDPG:**In [123], the authors put forth a proposition for the simultaneous allocation of spectrum and regulation of power in a joint manner. The authors used DDPG to reduce the complexity and maximize the sum rate of the network.

Table 3.2: Simulation Parameters

Parameters	Values	Parameters	Values
Cell's radius	500 m	Actor network's training rate	0.01
DDP link distance	10-50 m	Critic network's training rate	0.01
Count of CUs	60	Q -function discount factor	0.9
Count of RBs, (\mathcal{C})	20	Starting exploration	1
Count of DDPs	20, 40, 60, ..., 180	Finishing exploration	0.01
Each RB Bandwidth	180 KHz	Count of exploratory steps	1000
Frequency of carrier	5 MHz	Replay storage capacity	1000
Noise power spectrum density	-174 dBm/Hz	Mini-batch size (E)	32
DDP link path loss exponent	4	Each epoch's number of steps	20
q_{\max}	2	Weights in Reward Function	1,1
Multi path fading	Unit Mean	level of discretization	10
Shadowing standard deviation	8 dB	Weights renew duration	10
Maximum Power (P_c^{\max})	25 dBm	Number of pre-training steps	900
Maximum Power (P_d^{\max})	10-25 dBm	Overall number of training steps	4500
SINR threshold, ($\Upsilon_c^{r,\min}$)	8 dB	Length of time slot (t)	15 s
SINR threshold, (Υ_d^r)	8 dB	Total number of time slots (T)	150
CU link path loss	$128.1 + 37.6 \log d$	Optimizer	Adam
DDP link path loss	$148 + 40 \log d$	Activation function	ReLu

3.5.3 Experimental Graph and Discussion

3.5.3.1 Comparative Metric

In this particular section, the network sum rate of the proposed scheme is evaluated and discussed through experimental analysis, with respect to various parameters.

The graph of the network sum rate versus network size is shown in Fig. 3.2 (a). The result reveals that as the network size increases, the sum rate decreases. This occurred because, as the size of the network increased, the influence of CR-CI and CO-CI increased along with an increase in the number of CUs and DDPs. Despite this, the result depicts that when the network

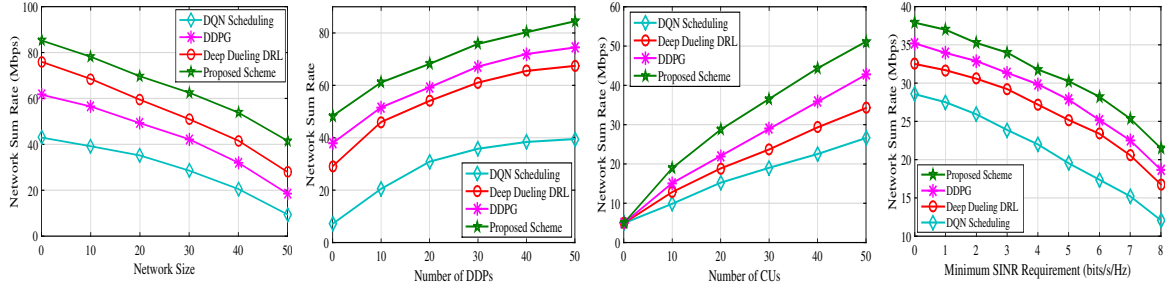


Figure 3.4: Comparative Metric (a) Network Sum Rate vs. Network Size (b) Network Sum Rate vs. Number of DDPs (c) Network Sum Rate vs. Number of CUs (d) Network Sum Rate vs. Minimum SINR requirement Requirement.

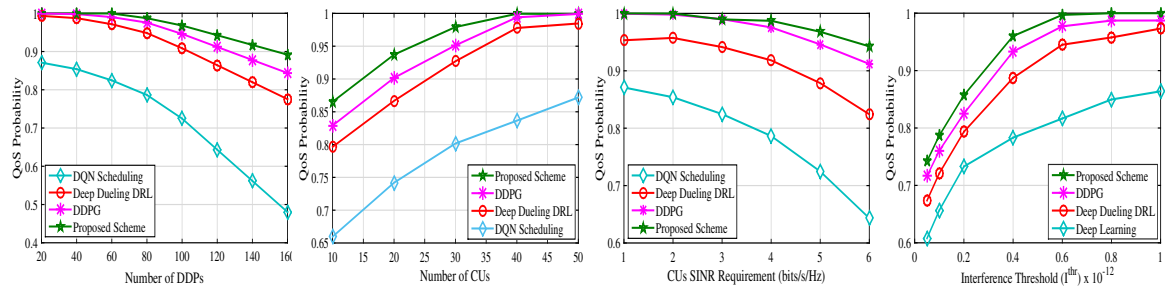


Figure 3.5: Probability Analysis (a) QoS Probability vs. Total Number of DDPs (b) QoS Probability vs. Total Number of CUs (c) QoS Probability vs. Minimum SINR Requirement for CUs (d) QoS Probability vs. Interference Threshold.

size reaches 30, the proposed scheme obtains 21.05%, 34.21%, and 49.8% higher sum rate than that of the DDPG, Deep dueling, and DQN scheduling. The reason behind this situation is that the proposed scheme controls the power of CUs by using a combination of AGMA and DDPG to handle the CR-CI. On the other hand, the power of DDTs is controlled by using the combination of SCALE and DDPG to limit the influence of CO-CI.

The graph in Fig. 3.2 (b) shows the relationship between the network sum rate and the interference threshold. The outcome demonstrates that the network sum rate falls as the number of DDPs increases in the cell. This occurred since the number of DDPs increased, which in turn increased the amount of CO-CI. Furthermore, it is evident from the graph that the proposed scheme achieves a superior sum- rate as compared to the baseline schemes. The cause for this predicament lies in the utilization of CO-D3PG by the proposed scheme. CO-D3PG is a combination of DDPG and SCALE. The DDPG converts the continuous policy into a deterministic one in real-time, and the SCALE optimize the power of DDTs, which not only

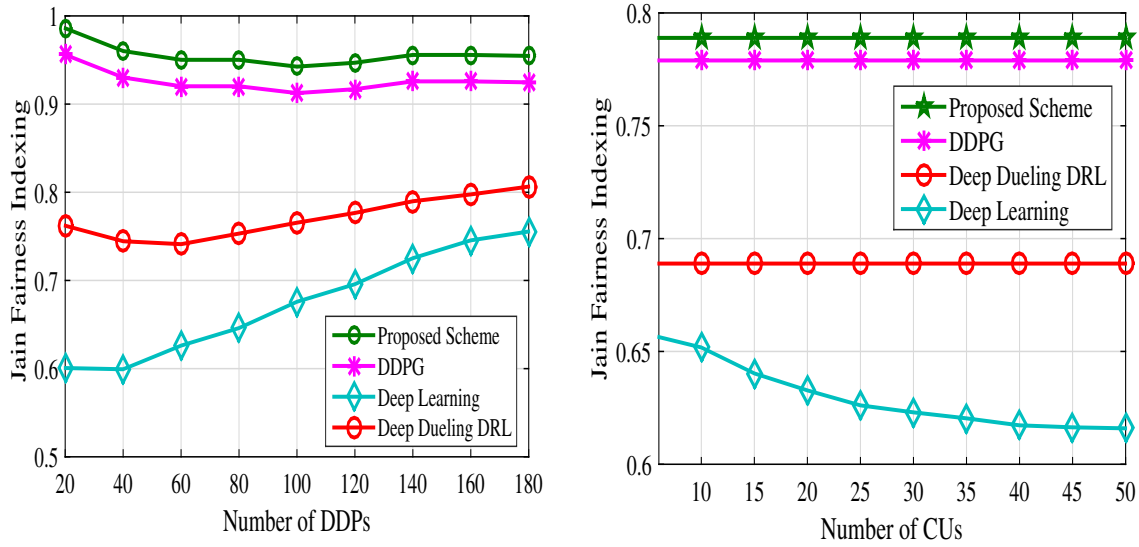


Figure 3.6: Jain Fairness Indexing (a) Jain Fairness Indexing vs. Total Number of DDPs (b) Jain Fairness Indexing vs. Total Number of CUs

mitigates the CO-CI but also enhances the QoS of the CUs and DDPs.

The graph of the network sum rate versus the number of CUs is shown in Fig. 3.2 (c). The result implies that when the CUs increase, the sum rate increases. This occurred since the number of DDPs increased, which in turn decreased the amount of CO-CI with an increase in the number of resources. Moreover, the result depicts that the proposed scheme achieves a higher sum rate than the baseline schemes.

The correlation between network sum rate and interference threshold is shown in Fig. 3.2 (d). This indicates that the network sum rate increases as the interference threshold increases. This occurred because, when an interference threshold increases, the DDTs are able to transfer power flexibly between the resources, resulting in an improvement in their SINR. On the other hand, due to the restricted transmit power budgets, the sum rate is observed to become constant after a certain period of time. The situation can be attributed to the fact that the proposed scheme effectively maximizes the power of both CUs and DDTs concurrently.

3.5.3.2 QoS Probability Analysis

In this particular section, an analysis is conducted on the probability of QoS in the proposed scheme, in comparison to the baseline schemes. The formula used to measure the QoS

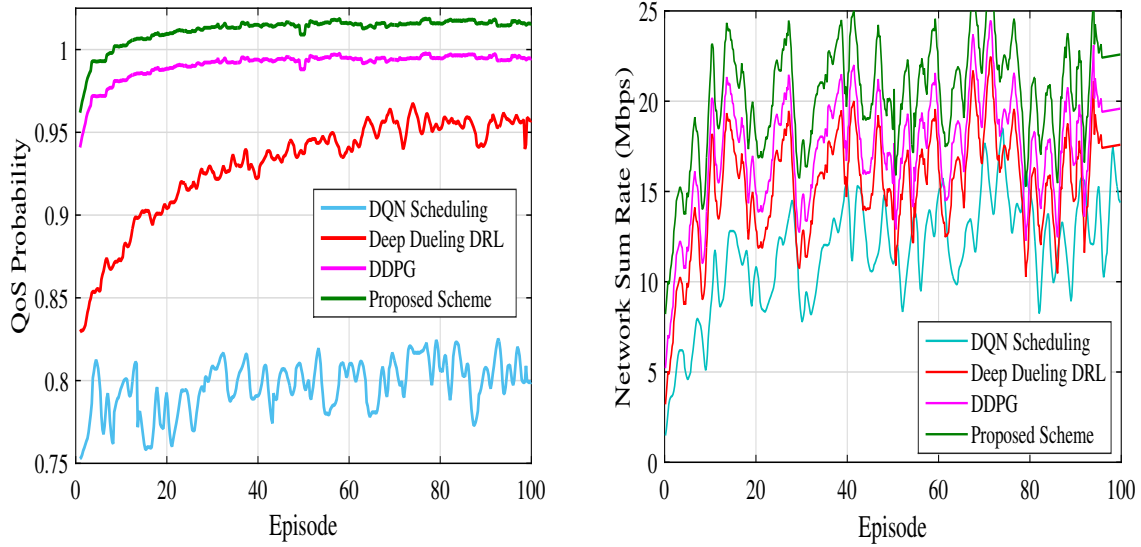


Figure 3.7: Convergence Behavior (a) QoS Probability vs. Episode (b) Network Sum Rate vs. Episode.

probability is estimated as follows.

$$\text{QoS Probability} = \frac{\text{Number of CUs}}{\text{Number of CUs and DDPs}} \quad (3.86)$$

Fig. 4.3(a) depicts how the number of DDPs affects the QoS probability. The graph shows that when DDPs in a cell are low, the QoS is high. As the number of DDPs in a cell approaches 60, the network's QoS starts to decrease. This occurred since the number of DDPs increased, which in turn increased the amount of co-channel interference with an increase in the number of resources. In order to address this issue, we used the SCALE, which maintained the power of the DDPs and resulted in a decrease in CO-CI. In this way, the proposed scheme attained a higher QoS compared to the baseline schemes.

Fig. 4.3(b) depicts the correlation of QoS probability with the count of CUs. The finding suggests that as the count of CUs in a cell increases, so does the probability of QoS achieving success. The cause behind this is that the number of CUs is proportional to the resources. Also, the presence of NOMA over the BS mitigates the receiver side interference due to SIC.

Fig. 4.3(c) illustrates the correlation of variation in QoS probability with minimum SINR threshold. The graph reveals that the possibility of QoS increases with an increase in minimum

rate demand for DDPs. The cause of this behavior is that as the SINR requirement increases, the DDT's transmission power increases as well, resulting in a higher CO-CI over the CUs. As a result, the QoS probability of DDPs gradually rises while that of CUs slowly reduces, resulting in a decrease in the total QoS network probability. Despite this situation, the proposed scheme achieves a higher QoS in comparison to the baseline schemes because we proposed the CO-D3PG scheme, which not only decreases the CO-CI but also maintains the QoS probability of the CUs across each resource.

Fig. 4.3(d) illustrates the variation in QoS probability as a function of the interference threshold. This occurs since, as the interference threshold increases, the chance of QoS probability also increases as well, but beyond a certain point, the probability remains constant. The explanation behind this behavior is that when the threshold increases, the transmission power of DDPs towards the CUs increases, which decreases their performance. Despite this, the proposed system outperforms the baseline schemes due to the presence of SIC over the CUs side, which overcomes the effect of interference produced by the DDPs.

3.5.3.3 Jain Fairness Indexing

In this section, the fairness of the proposed scheme is measured by changing the average data rates of the users. The suggested scheme's fairness is estimated using Jain's fairness indexing formula [61]:

$$\mathcal{F} = \sum_{c=1}^C \left(\mathbb{U}_c^b + \sum_{d=1}^D \mathbb{U}_d^r \right) / C \quad (3.87)$$

The impact on network fairness as a function of the number of DDPs is depicted in Fig. 3.6(a). The graph shows that the proposed scheme delivers greater fairness in comparison to the baseline schemes. Furthermore, the graph suggests that as the number of DDPs increases, there is a corresponding decrease in fairness, leading to an increase in CO-CI. To address this issue, the proposed scheme uses AGMA and SCALE to maintain the power of the CUs and DDTs. Moreover, the proposed scheme does not provide sufficient resources to the DDPs when it exceeds the CO-CI using the SCALE.

The graph of fairness versus number of CUs is revealed in Fig. 3.6(b). The result shows that as the number of CUs in the cell increases, the resources available for DDPs to reuse the CUs

also increase. CR-CI, CO-CI, and SIC interference all decrease as a result. This decrement in interference enhances fairness and QoS among the CUs.

3.5.3.4 Convergence Behavior

In this section, the proposed scheme's convergence behavior with respect to network sum rate and QoS probability is examined. Also, it demonstrates why the proposed approach converges faster than the baseline schemes.

The graph in Fig. 3.7(a) demonstrates Algorithm 4's convergence performance. The graph implies that the network sum rate of the proposed scheme has converged within 50 episodes. This happened because the proposed scheme optimizes the power of CUs as well as DDPS, which reduces CR-CT and CO-CI. As a result, each agent had to train themselves multiple times and rely on already trained agents to find the appropriate policy much faster.

The curve depicted in Figure 5(b) showcases the rate at which Algorithm 4 converges with respect to the QoS probability. The graph clearly demonstrates that the suggested scheme offers speedier QoS when compared to the baseline schemes. The explanation for this scenario lies in the fact that the suggested plan encompasses the utilization of AGMA and SCALE along with DDPG to optimize the power of the CU and DDPs, which not only decreases the CO-CI but also reduces the rate of repeated samples from learned samples.

3.6 Summary

This chapter investigates the sum-rate and fairness maximization among NOMA-enabled CUs and DDPs while considering the resource and power constraints of BS and DDT. To achieve the target, firstly, the centralized DDPG is used to allocate the resources to CUs with NOMA. Afterwards, to mitigate the CR-CI, CO-CI and improve fairness among the CUs, the AGMA technique is integrated with DDPG. The AGMA optimize the power of the CUs and maintains the power distribution among the CUs based on their channel gain. However, it is found that the DDPs do not train simultaneously because they cannot access the instantaneous global CSI in real time. This problem results in an increase in CO-CI and a decrease in fair-

ness. To address these problems, first of all, the D3PG is proposed to improve fairness. D3PG helps provide resources to DDPs by reusing the CUs' resources. Finally, to reduce CO-CI, CO-D3PG is proposed. CO-D3PG is the combination of CO and D3PG that controls the power of the DDPs. The experimental results reveal that the proposed scheme enhances the overall network's sum rate by maintaining fairness among the CUs and DDPs as compared to baseline schemes.

Chapter 4

Deep Reinforcement Learning Based Energy Consumption Minimization for Intelligent Reflecting Surfaces Assisted D2D Users Underlying UAV Network

The subsequent are the primary achievements that are presented by this chapter..

- We suggest the utilization of a RIS-integrated UAV wireless network in the up-link direction. This implementation involves the integration of a RIS for the purpose of enhancing connectivity in WCN and restricting the UAV' mobility . These measures are taken to ensure the long-term advantages of the aforementioned network architecture.
- To attain the objective, we collaboratively optimize the trajectory of the UAV (TR-UAV) and the phase shift of the RIS (PS-RIS) through a joint approach, which is formulated as an MDP problem. Through the implementation of the DRL-based $C - DDQN$ algorithm, we were able to effectively solve the optimization problem
- The proposed C-DDQN technique achieves a trade-off between an increasing learning rate and convergent local optimality and also prevents oscillation.

The paper's flow is delineated as follows: in Section II, there is a discussion of the system

model and problem formulation. The recommended approach is described in Section III. The experimental results are illustrated and discussed in Section IV. In the end, Section V concludes the article.

4.1 Network Model and Problem Formulation

4.1.1 Network Model

A RIS-added UAV-enabled uplink single-cell wireless communication system is taken into consideration, as presented in Fig. 4.1. The suggested network model has a single BS, a single UAV, multiple MUs, and multiple D2DPs. All MUs, D2DPs, and BS are positioned on the ground. The D2DPs and MUs both use the same SC. Consider the set of MUs and D2DP, which are identified by $\mathcal{M} = \{1, 2, \dots, m, \dots, M\}$ and $\mathcal{N} = \{1, 2, \dots, n, \dots, N\}$, respectively. Every MU sends its signal towards the BS over each orthogonal SC through OMA protocols. In each D2DP, the D2DT communicates with its respective D2DR through OMA protocols. The D2DPs and MUs both use the same SC. It is presumed that every single SC is taken up by a single MU. Let B be the total network capacity, divided into \mathcal{K} SCs. The set of SCs is specified as $\mathcal{K} = \{1, 2, \dots, k, \dots, K\}$, while the set of time slots is defined as $\mathcal{T} = \{1, 2, \dots, t, \dots, T\}$. However, because of the presence of obstacles and objects in some real-world scenarios, a direct connection between MU and BS is not feasible. In order to take advantage of the high mobility possibilities, a RIS-integrated UAV is implemented in the network. This aims to improve the quality of communication by establishing a direct link between the MU and the RIS-integrated UAV and RIS, as well as between the RIS-integrated UAV and the BS. A RIS has a homogeneous linear array (HLA) of $L \times L$ reflecting elements. Furthermore, a UAV is capable of regulating the phase shift of each individual element in the HLA. The BS is taken to be positioned in the X-Z plane with the coordinates $(x_{bs}, 0, H)$, where H is the BS's height. Let MU and D2DPs be located in the X-Y plane, and their 3D coordinates are $(x_m, y_m, 0)$ and (x_n, y_n) , respectively. Let us define the coordinates of UAV-RIS as $X_{UAV} = (x_l, y_l, z_l)$. The estimation of the distance between the MU and the l^{th} component of RIS is determined by $d_{M,L}(t)$ as $d_{M,L}(t) = \sqrt{(x_m - x_l)^2 + (y_m - y_l)^2 + (z_l)^2}$. Similarly, the distance between the l^{th}

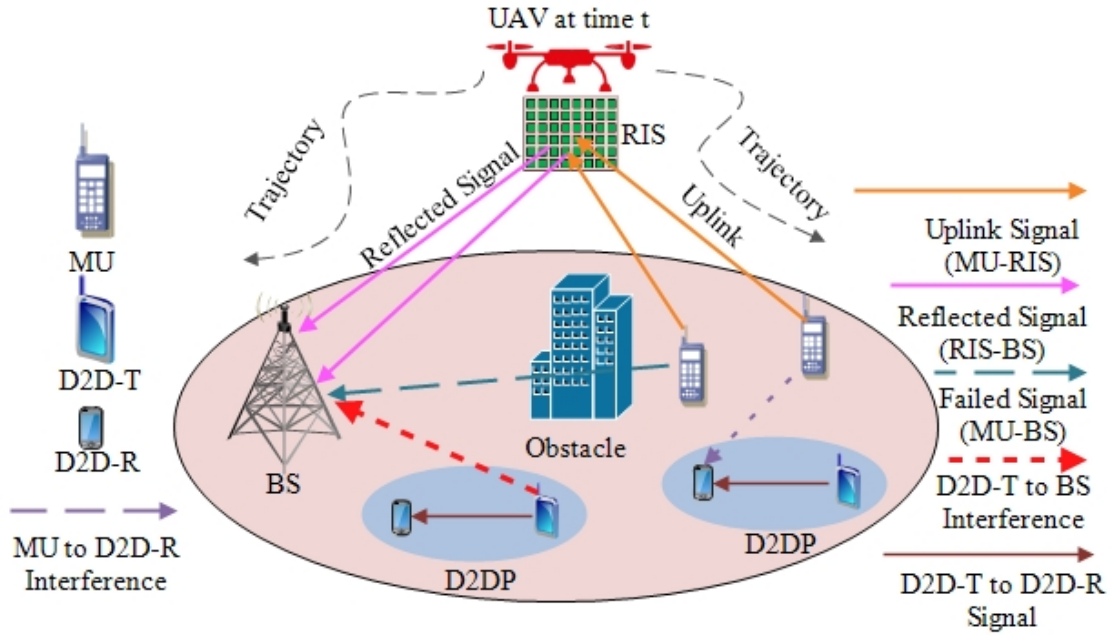


Figure 4.1: Network Architecture.

component of RIS and BS is estimated as: $d_{l,BS}(t) = \sqrt{(x_l - x_{bs})^2 + (y_l)^2 + (z_l - H)^2}$.

4.1.2 Channel Model

Consider that all channel links are expected to experience quasi-static flat fading in the proposed network architecture. Let $g_{m,l}^k \in \mathbb{C}^{L \times 1}$ and $\hat{g}_{l,bs}^k \in \mathbb{C}^{M \times 1}$ define the channels from MU to l^{th} component of RIS and from l^{th} component of RIS to BS in the k^{th} SC, respectively. Let $\tilde{g}_{mu,bs}^k \in \mathbb{C}^L$ and $h_{n,i}^k \in \mathbb{C}^L$ define the channels from m^{th} to BS and from n^{th} D2DT to i^{th} D2DR in the k^{th} SC, respectively. Using $\Theta \in \mathbb{C}^{L \times L}$ as the RIS's reflecting coefficient matrix, which can be defined as:

$$\Theta = \text{diag}\{u_1 e^{j\Phi_1}, u_2 e^{j\Phi_2}, \dots, u_L e^{j\Phi_L}\}, \quad (4.1)$$

where $u_L \in A^+$ and $\Phi_L \in A$ are the reflecting amplitude and phase shift of RIS's l^{th} element. Assume that $u_L \in (0, 1)$ and $\Phi_L \in (0, 2\pi)$.

4.1.2.1 MU-BS Channel Model

The BS gets two signals. The first is a direct signal (MU-BS) from MU, whereas the second is a reflected signal (MU-RIS-BS) through RIS. We assume that reflected signals are characterized by Rician fading, whereas direct links follow Rayleigh fading. We consider that direct links pursue Rayleigh fading, while RIS-assisted links are characterized by Rician fading. Now $g_{m,l}^k$, $\hat{g}_{l,bs}^k$, and $\tilde{g}_{m,bs}^k$ articulated in the following manner:

$$g_{m,l}^k = \sqrt{\omega_0 (d_{m,l}^k)^{-\phi_1}}$$

$$\left(\sqrt{\frac{V_{m,l}^k}{1+V_{m,l}^k}} g_{m,l}^{LoS,k} + \sqrt{\frac{1}{1+V_{m,l}^k}} g_{m,l}^{NLoS,k} \right), \quad (4.2a)$$

$$\hat{g}_{l,bs}^k = \sqrt{\omega_0 (d_{l,bs}^k)^{-\phi_2}}$$

$$\left(\sqrt{\frac{V_{l,bs}^k}{1+V_{l,bs}^k}} \hat{g}_{l,bs}^{LoS,k} + \sqrt{\frac{1}{1+V_{l,bs}^k}} \hat{g}_{l,bs}^{NLoS,k} \right), \quad (4.2b)$$

$$\tilde{g}_{m,bs}^k = \sqrt{\omega_0 (d_{m,bs}^k)^{-\phi_3}} \tilde{g}_{m,bs}^{NLoS,k}, \quad (4.2c)$$

where the reference distance's propagation loss and the exponent of path loss are shown by ω_0 and ϕ , respectively. Defining $V_{m,l}^k$ and $V_{l,bs}^k$ as Rician factors of the MU-RIS and RIS-BS links [47]. The non-line-of-sight (NLoS) elements, denoted by $g_{m,l}^{NLoS,k}$, $\hat{g}_{l,bs}^{NLoS,k}$, and $\tilde{g}_{m,bs}^{NLoS,k}$, can be described by Rayleigh fading. Each element belongs to $\mathcal{CN}(0, 1)$. Now $g_{u,m}^{LoS,n}$ and $f_{m,j}^{LoS,n}$ can be given as follows:

$$g_{m,l}^{LoS,k} = \left[1, \dots, e^{j(l-1)\pi \sin(a_{m,l})}, \dots, e^{j(L-1)\pi \sin(a_{m,l})} \right]^{\mathbb{T}} \quad (4.3a)$$

$$\hat{g}_{l,bs}^{LoS,k} = \left[1, \dots, e^{j(l-1)\pi \sin(a_{l,bs})}, \dots, e^{j(L-1)\pi \sin(a_{l,bs})} \right]^{\mathbb{T}}, \quad (4.3b)$$

where $\sin(a_{m,l}) = \frac{y_l - y_m}{\sqrt{(x_l - x_m)^2 - (y_l - y_m)^2}}$ and $\sin(a_{l,bs}) = \frac{y_l}{\sqrt{(x_l - x_{bs})^2 - (y_l)^2}}$.

The signal transmitted by the m^{th} MU to BS over the k^{th} SC can be expressed as follows:

$$Y_{m,bs}^n = \underbrace{\sqrt{p_m} \left(g_{m,l}^k \Theta \tilde{g}_{l,bs}^k + \tilde{g}_{m,bs}^k \right)}_{\text{Reflected + Direct Signal}} x_{m,bs}^k + \underbrace{\sum_{n \in \mathcal{N}} \alpha_n^r(t) \sqrt{p_n^k(t)} h_{n-bs}^k x_{n-bs}^k}_{\text{DT to BS Interference}} + \underbrace{\sigma^2}_{\text{AWGN}}, \quad (4.4)$$

where $\sqrt{p_m^k}$ and $\sqrt{p_n^k}$ are the m^{th} MU and n^{th} D2DT transmitted powers, respectively. The transmitted symbols for m^{th} MU and n^{th} D2DT are denoted by $x_{m,bs}^k$ and $x_{n,bs}^k$, respectively. The SC allocation coefficient for the m^{th} MU link and the n^{th} D2DT link is represented by $\alpha_{m,n}^k$. It represents the following:

$$\alpha_{m,n}^k = \begin{cases} 1 & \text{if } n^{th} \text{ D2DT occupies } m^{th} \text{ MU,} \\ 0 & \text{otherwise.} \end{cases} \quad (4.5)$$

The SINR over the k^{th} SC for the m^{th} MU is now defined as:

$$\Gamma_{m,bs}^k = \frac{p_m \left| g_{m,l}^k \Theta \tilde{g}_{l,bs}^k + \tilde{g}_{m,bs}^k \right|^2}{\sum_{n \in \mathcal{N}} p_n \left| h_{n-bs}^k \right|^2 + \sigma^2} \quad (4.6)$$

4.1.2.2 D2DT-D2DR Channel Model

The i^{th} D2DR's received signal from the n^{th} D2DT via the k^{th} RB is expressed as:

$$Y_{n,i}^k = \underbrace{\sqrt{p_n^k} h_{n,i}^k x_{n,i}^k}_{\text{Direct Signal}} + \underbrace{\sum_{i' \neq i, i' \in \mathcal{N}} \alpha_n^k \sqrt{p_{n'}^k} h_{n-i'}^k x_{n-i'}^k}_{\text{D2DT to D2DR interference}} + \underbrace{\sum_{m \in \mathcal{M}} \sqrt{p_m^k} |g_{m-i}^k| x_{m-i}^k}_{\text{MU to D2DR interference}} + \underbrace{\sigma^2}_{\text{AWGN}}, \quad (4.7)$$

Now SINR for i^{th} D2DR over k^{th} SC is defined as:

$$\Gamma_{n,i}^k = \frac{P_n |h_{n,i}^k|^2}{\sum_{i' \neq i, i' \in \mathcal{N}} \alpha_{m,i}^k P_n^k |h_{n-i'}^k|^2 + \sum_{m \in \mathcal{M}} P_m^k |g_{m-i}^k|^2} \quad (4.8)$$

4.1.3 Energy Efficiency Estimation

The transmission rate for the m^{th} MU can be given as:

$$\mathbb{R}_{m,bs}^k = B \log_2 \left[1 + \Gamma_{m,bs}^k \right]. \quad (4.9)$$

The transmission rate for the n^{th} D2DR can be given as:

$$\mathbb{R}_{n,i}^k = B \log_2 \left[1 + \Gamma_{n,i}^k \right] \quad (4.10)$$

The combined transmit rate of network can be presented as::

$$\mathbb{R}_{m,n}^k = \sum_{k \in \mathcal{K}} \left[\sum_{m \in \mathcal{M}} \mathbb{R}_{m,bs}^k + \sum_{n \in \mathcal{N}} \mathbb{R}_{n,i}^k \right]. \quad (4.11)$$

The following expression represents the network's overall power consumption:

$$\mathbb{P}_T^k = P_{UAV}^k + P_{RIS}^k + \sum_{m=1}^M P_m^k + \sum_{n=1}^N P_m^k, \quad (4.12)$$

In Eq. (4.12), the usage of power by RIS is denoted by term P_{RIS}^k while the usage of power by circuits is denoted by term P_{ckt}^k ; term $\sum_{n=1}^N P_m^k$ denotes the power usage of D2DTs; and the term

$\sum_{m=1}^M P_m^k$ denotes the power usage of MUs. Power usage for the UAV is denoted by P_{UAV}^k and is as follows:

$$P_{UAV}^k = P_l \left(1 + \frac{3v^2}{U_T^2} \right) + P_l \left(\sqrt{1 + \frac{v^4}{4v_l^4}} - \frac{v^2}{2v_l^2} \right)^{\frac{1}{2}} + \frac{1}{2} d_l \mu_s A v^3 \quad (4.13)$$

Furthermore, the limitations in terms of mobility that the UAV encounters can be articulated as :

$$X_{UAV}(t+1) = X_{UAV}(t) + V_{UAV}(t), \quad (4.14)$$

where V_{UAV} corresponding to the UAV-RIS's flying velocity.

Finally, EE may be computed in the following manner:

$$EE = \frac{\mathbb{R}_{m,n}^k}{\mathbb{P}_T^n} = \frac{\sum_{k \in \mathcal{K}} \left[\sum_{m \in \mathcal{M}} \mathbb{R}_{m,bs}^k + \sum_{n \in \mathcal{N}} \mathbb{R}_{n,i}^k \right]}{P_{UAV}^k + P_{RIS}^k + \sum_{m=1}^M P_m^k + \sum_{n=1}^N P_n^k} \quad (4.15)$$

4.1.4 Problem Formulation

The primary objective of this paper is to optimize the TR-UAV and PS-RIS with the aim of minimizing UAV's energy usage while ensuring preservation of SINR for both MUs and D2DPs. Accordingly, the optimization issue may be expressed as follows:

$$\begin{aligned} \mathcal{P}. \mathcal{F}. & : \min_{(X_{UAV}, \phi)} EE(t), & (4.16) \\ s.t. \quad \mathbb{Z}_1 & : \sum_{k=1}^K \alpha_{m,n}^k \leq 1, & \forall \mathcal{M}, \mathcal{N}, \\ \mathbb{Z}_2 & : \Gamma_{m,BS}^k \geq \Gamma_{m,BS}^{k,\min}, & \forall \mathcal{K}, \\ \mathbb{Z}_3 & : V_{UAV} \leq V_{UAV}^{\max}, \\ \mathbb{Z}_4 & : \phi \in [0, 2\pi], & \forall \mathcal{L}, \\ \mathbb{Z}_5 & : X_{UAV} \in \mathcal{Z}, \end{aligned}$$

The indent \mathbb{Z}_1 ensures that each MU is associated with a single D2DT. The minimal data rate demand for MU is represented by \mathbb{Z}_2 . \mathbb{Z}_3 defines that UAV's velocity should be less than its maximum velocity. \mathbb{Z}_4 limits the restricted range of PS-RIS. \mathbb{Z}_5 mandates that the UAV's position should be in the restricted area (\mathcal{Z}) for vertical and horizontal flight.

4.2 PROPOSED SCHEME

First, we optimize both the TR-UAV and PS-RIS in a joint manner by transforming problem (4.16) into an MDP. After that, we suggested the C-DDQN algorithm for joint optimization of TR-UAV and PS-RIS. The suggested C-DDQN algorithm also describes the component of MDP. The diagram depicting the flow of the suggested approach can be observed in Fig. 4.2.

4.2.1 Markov Decision Process

The presented issue is converted into an MDP predicament prior to the utilization of RL techniques. MDP is characterized by five tuples, namely $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. The tuples \mathcal{S} , \mathcal{A} , \mathcal{P} , \mathcal{R} , and γ stand for a collection of states, actions, mapping linkages between states s and s' , rewards, and discount factor, respectively. The following is a detailed description of the MDP model:

4.2.2 The C-DDQN Algorithm for the Joint Optimization of Both TR-UAV and PS-RIS.

In this subsection, a C-DDQN approach is introduced to accurately predict the TR-UAV and PS-RIS, all while ensuring that every user's data needs are met. In the proposed system model, a central controller acts as an agent. Both parameters are regulated by a centralized controller. The agent scrutinizes the state s_t^k during every time slot t to delineate the surroundings. The environment consists of the positions of all users (UAV and MUs) along with the RIS's phase shift. The agent acts in accordance with the present state and decision strategy D by selecting an action a_t^k , which includes the UAV's hovering directions as well as varying degrees of phase shift for every reflecting element. After performing the actions, an agent is granted a reward or a penalty denoted by r_t^k , which depends on the energy consumption of the overall network. A Q_t^k -value is computed for each time-slot in accordance with the present state and prior actions. As a result, the decision policy D is determined by the Q-function $Q(s_t^k, a_t^k)$, which stores the state (s_t^k), action(a_t^k), and Q_t^k -value. In contrast to attempting to optimize the reward within a particular time frame, the fundamental principle of the C-DDQN algorithm is to maximize the

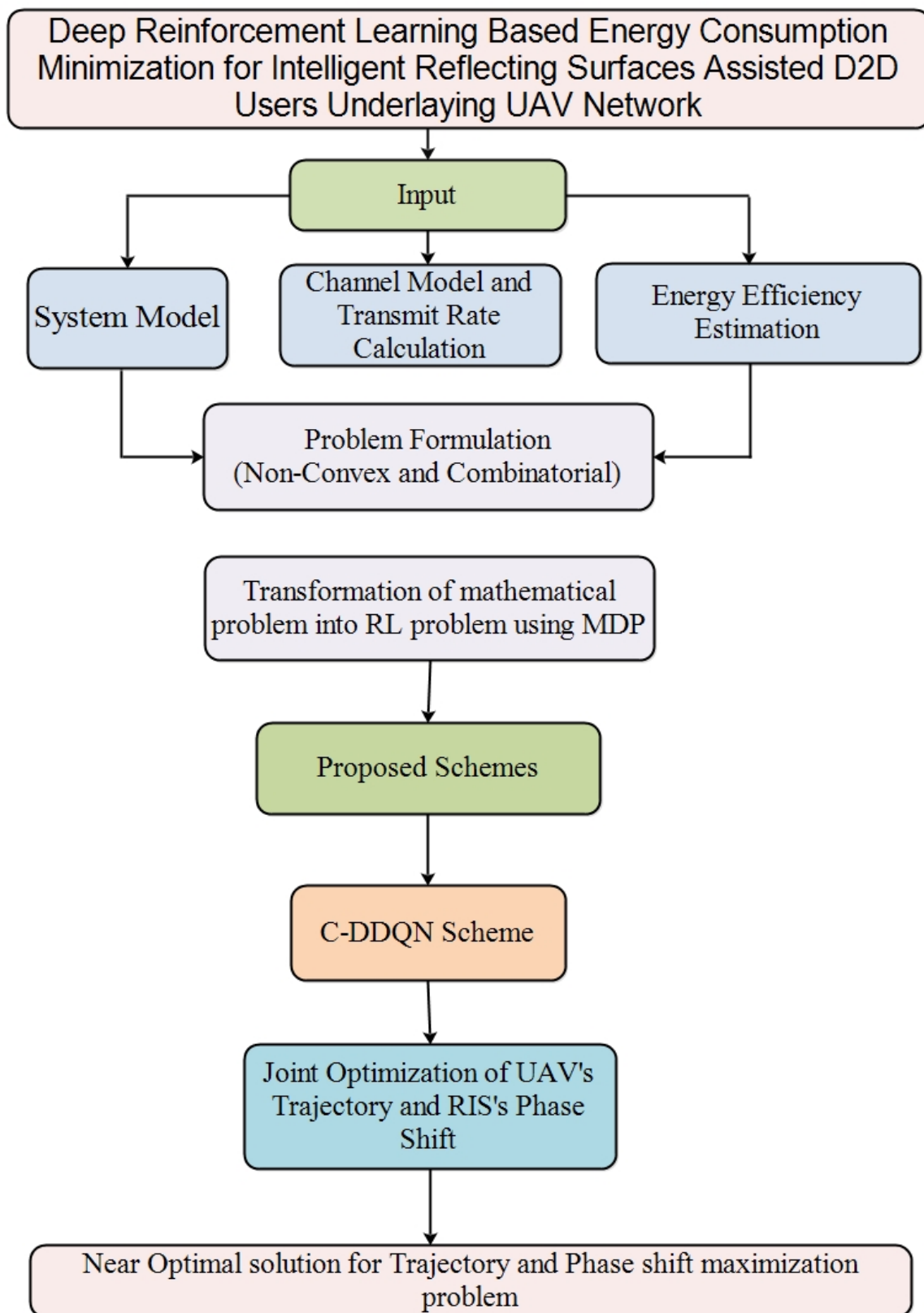


Figure 4.2: Flow Diagram of Proposed Scheme.

cumulative reward over the long term. Hence, the course of action selected by the C-DDQN algorithm may not be the optimal choice for the present moment, yet it could prove to be the most effective alternative for attaining long-term benefits.

Now, given the current state s_n^K along with action a_n^k , the state value function expresses the anticipated reward for a decision policy D and is explicitly characterized as:

$$Q_{t+1}(s_t^k, a_t^k) \leftarrow (1 - \beta)Q_t + \beta [r_t^k + \Upsilon \max_{a_t^k} Q_t(s_{t+1}^k, a_t^k)], \quad (4.17)$$

whereas β is denoted as the learning rate, Υ is referred to as the discount factor.

Now the optimized Q-value can be represented as follows::

$$Q^*(s_t^k, a_t^k) = \mathbb{E}_{s_t^k} [r_t^k + \Upsilon \max_{a_t^k} Q^*(s_t^k, a_t^k) | s_t^k, a_t^k] \quad (4.18)$$

4.2.2.1 Agent

In the suggested network model, a central controller acts as an agent.

4.2.2.2 State Space

The agent meticulously analyzes the current state with the intention of delineating the surroundings. This includes a comprehensive assessment of phase shifts, UAV-RIS positions, MU locations and the transmitted power of the MUs. The state space elements, when taking into account environmental parameters, are outlined in the following manner:

$$s_t^k = [\phi_{t,l}^k, X_{t,l}^k, X_{t,m}^k, P_{t,m}^k]. \quad (4.19)$$

4.2.2.3 Action Space

In the proposed model, the agent is UAV-RIS. Because UAV in every state have the ability to move in any of the following directions: upward, downward, backward, forward, left, and right. The UAV's speed is also considered a possible action. Define ζ_{UAV} as the UAV's acceptable action, which combines all of the UAV's movement directions along with speed. The agent's

entire action can be computed as:

$$a_t^k = \left[\zeta_{UAV}, \Delta\phi_{t,l}^k, \Delta p_{t,m}^k, V_{UAV} \right], \quad (4.20)$$

where $\Delta\phi_{t,l}^k$ is the varying value corresponding to every reflecting component's phase shift, $\Delta p_{t,m}^k$ is the varying value linked to the transmission power from every MU to the UAV, and V_{UAV} corresponding to the UAV-RIS's flying velocity. The agent performs the action a_t^k in the state s_t^k and subsequently transitions to the next state s_{t+1}^k upon completion of the aforementioned action.

4.2.2.4 Reward Function

The problem's objective is usually connected to the reward function. The expression for the reward function can be described as follows:

$$\mathbb{R} = \begin{cases} \frac{1}{P_T^k} & \text{if each constraint is met,} \\ 0 & \text{otherwise.} \end{cases} \quad (4.21)$$

The rewarding function mentioned above increases gradually as power consumption decreases. As a result, reward function may ultimately attain the over all networks' minimum energy consumption.

The outcome of the Q-table is processed by a CNN including weights θ . The suggested C-DDQN method employs memory replay of capacity Z in an effort to minimize sampling's correlation. The agent performs arbitrary execution actions in the initial stages of training and records its experiences in a replay buffer. The set of experiences can be used as training samples. It includes the states, actions, and rewards. CNN's goal is to reduce the loss function for every episode, which can be described as follows:

$$\begin{aligned} \mathbb{L}^F(\theta) &= \widehat{\mathbb{E}}_{\pi} \left[(\mathbb{R} + \Upsilon \max_{a_t^k \in \mathcal{A}} Q_{old}(s_t^k, a_t^k, \theta) - Q(s_t^k, a_t^k, \theta))^2 \right] \\ &= \sum \left[(y^{DQN} - Q(s_t^k, a_t^k, \theta))^2 \right] \end{aligned} \quad (4.22)$$

In Equation (4.22), the target Q network is represented by $Q(s_t^k, a_t^k, \theta)$. The DNN parameters of the target Q network are denoted by y^{DQN} and θ .

The C-DDQN algorithm employs the ε -greedy exploration technique to strike a balance between exploration and exploitation. The ε -greedy exploration method is based on the following concept:

$$\mathbb{P}_{\mathbb{R}} = \begin{cases} 1 - \varepsilon & \text{if } \bar{a}_t^k = \arg \max Q(s_t^k, a_t^k), \\ \frac{\varepsilon}{|a_t^k| - 1} & \text{otherwise.} \end{cases} \quad (4.23)$$

The application of the C-DDQN method facilitates the attainment of a balance between an increasing learning rate and convergent local optimality. The C-DDQN method also prevents oscillation. The declining learning rate is expressed as:

$$\beta_{n_e} = \frac{\beta_o}{1 + \eta n_e} \quad (4.24)$$

In Eq. (4.24) β_{n_e} stands for the training sets, β_o denotes rate of learning during the initial episode, and constant value η is used to calculate the declining rate.

The intricacies of the C-DDQN approach can be observed in Algorithm 1. This algorithm utilizes θ to signify the utmost number of episodes and \mathbb{T} to represent the time-slot.

4.3 Performance Assessment

4.3.1 Parameters for Simulation

The RIS's phase shift parameter is likewise arbitrarily provided at the beginning time-slot, and the UAV is initially positioned at a random location in the simulation. \mathcal{M} MUs and \mathcal{N} D2DPs are evenly distributed around the BS in a cell. We created a DQN structure with three completely connected layers (input layer, hidden layers, output layer) forms the basis of the DQN training model, each containing 500, 500, and 250 neurons. Table 4.1 enumerates the simulation parameters that still require attention.

Algorithm 5 *C – DDQN Method for UAV’s Trajectory and RIS’s Phase Shift Optimization.*

Input Environment: (a) UAV, MUs and D2DPs (b) BS with OFDMA scheme (uplink) (c) RIS; $\Gamma_{m,bs}^k \geq \Gamma_{m,bs}^{k,\min}$: Minimum SINR requirement of MUs; and $\Gamma_{n,i}^k \geq \Gamma_{n,i}^{k,\min}$: Minimum SINR Requirements of DDPs.

Initialization:

- Weights in the Q-network: θ .
- Target network’s weights: $\theta^* = \theta$.
- Value function: $Q(s_t^k, a_t^k)$.
- UAV-RIS random location: (x_l, y_l, z_l) .
- Replay buffer: Z

1: **repeat**2: **for** episode = 1, . . . , θ **do**3: The agent (central controller) selects an action (a_t^n) evenly by employing ε

4: greed policy ;

5: $a_t^k = \arg \max_{a_t^k \in \mathcal{A}} Q(s_t^n, a_t^n)$, when probability is $1 - \varepsilon$;6: $a_t^n = \text{randomly}\{a_t^k\}_{a_t^k \in \mathcal{A}}$, when probability is ε ;

7: Formulate the total EE as in (16)

8: Renew the reward function as in (21)

9: The C-DDQN model moves to state s_{t+1}^k ;10: Store transition $(s_t^k, a_t^k, \mathbb{R}_t^k, s_{t+1}^k)$ in experience replay buffer11: of capacity Z 12: Select randomly a small batch of transitions $(s_i^k, a_i^k, \mathbb{R}_i^k, s_{i+1}^k)$ from13: the replay buffer of capacity Z .14: **for** iteration = 1, . . . , \mathbb{T} **do**15: Set the value of $y^{DDQN}(i)$ according to (4.22).

16: Execute the gradient descent step

17: $\theta \leftarrow \theta - a_i^k [y^{DDQN}(i) - Q_{\theta}(s_i^k, a_i^k)] \delta_{\theta} Q_{\theta}(s_i^k, a_i^k)$;
 $\theta \leftarrow \theta^*$.

18: Compute declining learning rate using (4.24)

19: **until** convergence of state s 20: **end for**21: **end for**22: **Output:** X_{UAV}, Θ

4.3.2 Results and Discussion

The illustration in Figure 4.2(a) specifies the UAV’s overall energy usage during all episodes. In contrast, the technique known as C-DDQN possesses the ability to converge with the aid of the principle of faster convergence via NNs. It is apparent that the C-DDQN method proposed can successfully converge after roughly 50000 episodes when the learning rate is set to 0.005.

Table 4.1: Simulation Parameters

Parameters	Values
Cellular cell's Radius	500m
Carrier Frequency	1.5GHz
UAV's velocity	4m/s
Power of MU transmission	5W
Gain of Channel Power	-30dB
Density of noise power spectrum	-174 dBm/Hz
Exponent of path loss	4
Pathloss MU-RIS links	$148 + 40\log d$
D2DPs number	3
MUs number	5
Rician factor	6 dB
Factor of Discount	0.9
Starting learning rate	0.2
Declining learning rate	0.002
Replay storage capacity	1000
Small-batch Size	64
Steps in every Epoch	20
Episodes	100
Optimizer	Adam
Activation function	ReLU

This indicates that the DDQN model, with a learning rate of 0.005, exhibits superior performance in scenarios featuring elevated learning rates, as evidenced by its accelerated rate of convergence and reduced average energy usage.

The power transmitted in each time-slot is illustrated in Figure 4.2(b) as the combined aggregate. It is noticeable that the immediate attainable capacity of users diminishes as time elapses. In the designated design models, users tend to wander in various directions with different velocities, leading them to move farther away from their original position. In a stationary position of UAV, the distance between the UAV and the users is progressively widening, resulting in a decrease in the achievable rate. By strategically planning the trajectory of the UAV, it is evident in Figure 4.2(b) that the decline in achievable rate can be effectively minimized. Furthermore, it is evident from Fig.2(b) that the implementation of RIS results in a higher transmission rate per time-slot compared to the scenario where RIS is not utilized.

The energy usage of both the UAV's transmitting and receiving components is depicted in Figure 4.2(c). As depicted in Figure 2(c), increasing the transmission power results in re-

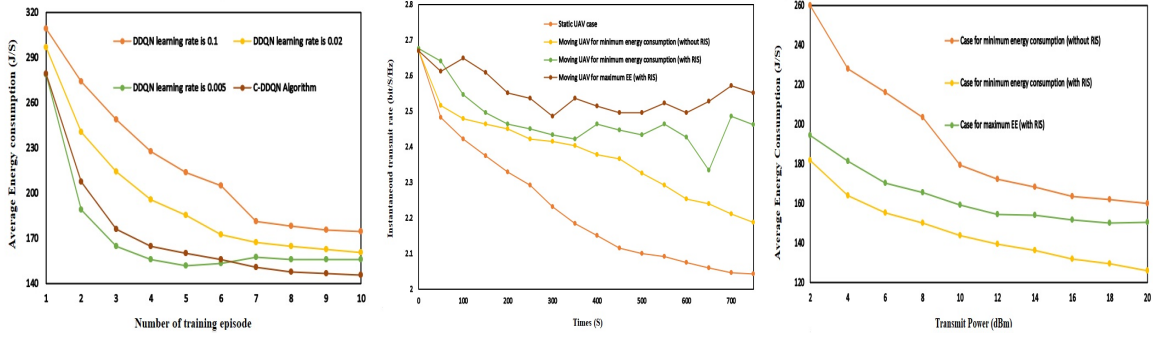


Figure 4.3: Comparative Analysis (a) e suggested C-DDQN algorithm’s convergence rate. (b) The instant rate of transmission over a period of time (c) Mean usage of energy in relation to transmission power.

duced energy consumption by the UAV. The UAV expends the greatest amount of energy in the absence of the reflective intelligent surface while attempting to minimize energy usage. This is due to the fact that the UAV needs to execute additional navigation to establish a LoS connection with the MUs without the aid of the RIS.

4.4 Summary

In this investigation, we delve into the communication system assisted by up-link RIS, while taking into account the presence of D2DPs. RIS is introduced into WCNs with the assistance of an UAV in order to solve the issues of CO-CI and CR-CI. A method known as the centralized declining deep-Q network (C-DDQN) has been proposed to accurately estimate the TR-UAV and PS-RIS, while simultaneously meeting the data demands of of DDU and CU. In the C-DDQN method, a central controller acts as an agent. The central controller governs both the TR-UAV and PS-RIS with precision.

Chapter 5

Conclusion and Future Scope

The proliferation of mobile users, smart gadgets, and multimedia applications generates an unprecedented growth of data traffic in 5G and beyond networks. This massive increase in data traffic puts a lot of burden on efficient spectrum utilization in the years to come. To address this problem, researchers recommended various D2D-C techniques. In D2D-C technology, two neighboring devices can share the data directly without the base BS. As a result, it enhances mobile users' quality of service by reducing the transmission delay. Also, in D2D-C, the DDPs reuse the same resources as used by CUs to boost the SE. Despite these advantages, key challenges such as CR-CI and CO-CI, as well as UMC, need to be investigated. Energy represents yet another significant challenge in D2D -C given the restricted capacity for energy storage within a battery. In this study, we have put forth the three approaches to resolve the previously mentioned concerns.

Firstly, a DRL Scheme for Sum Rate and Fairness Maximization Among D2D Pairs Underlying Cellular Network With NOMA is proposed. This scheme investigates the sum-rate and fairness maximization among NOMA-enabled CUs and DDPs while considering the resource and power constraints of BS and DDT. To achieve the target, firstly, the centralized DDPG is used to allocate the resources to CUs with NOMA. Afterwards, to mitigate the CR-CI, CO-CI and improve fairness among the CUs, the AGMA technique is integrated with DDPG. The AGMA optimize the power of the CUs and maintains the power distribution among the CUs based on their channel gain. However, it is found that the DDPs do not train simultaneously

because they cannot access the instantaneous global CSI in real time. This problem results in an increase in CO-CI and a decrease in fairness. To address these problems, first of all, the D3PG is proposed to improve fairness. D3PG helps provide resources to DDPs by reusing the CU' resources. Finally, to reduce CO-CI, CO-D3PG is proposed. CO-D3PG is the combination of CO and D3PG that controls the power of the DDPs. In future, we will extend this study to the multiple cell scenario in the environment of underlaid D2D communication cellular network. For a more realistic scenario having multiple BSs, the proposed scheme can be extended to a multi-agent reinforcement learning (MARL) scheme. Multiple BSs can be used as multiple agents in a MARL setting, where each agent will share the local information with other BSs or agents to collect global CSI values.

The second approach is DRL Based Energy Consumption Minimization for Intelligent Reflecting Surfaces (RIS) Assisted D2D Users Underlying UAV Network. In this investigation, we delve into the communication system assisted by up-link RIS, while taking into account the presence of D2DPs. RIS is introduced into WCNs with the assistance of an UAV in order to solve the issues of CO-CI and CR-CI. A method known as the C-DDQN has been proposed to accurately estimate the TR-UAV and PS-RIS, while simultaneously meeting the data demands of of DDU and CUs. In the C-DDQN method, a central controller acts as an agent. The central controller governs both TR-UAV and PS-RIS with precision. In order to further improve the effectiveness of UAV-RIS- assisted wireless networks, our future study will aim to optimize the UAV's velocity.

Bibliography

[1]

[2] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, “A survey of device-to-device communications: Research issues and challenges,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2133–2168, April 2018.

[3] A. Asadi, Q. Wang, and V. Mancuso, “A survey on device-to-device communication in cellular networks,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1801–1819, April 2014.

[4] P. Gandotra and R. K. Jha, “Device-to-device communication in cellular networks: A survey,” *Journal of Network and Computer Applications*, vol. 71, pp. 99–117, Aug. 2016.

[5] M. Noura and R. Nordin, “A survey on interference management for device-to-device (d2d) communication and its challenges in 5g networks,” *Journal of Network and Computer Applications*, vol. 71, pp. 130–150, Aug. 2016.

[6] C. Kai, H. Li, L. Xu, Y. Li, and T. Jiang, “Energy-efficient device-to-device communications for green smart cities,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1542–1551, April 2018.

[7] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, March 1998, vol. 135.

[8] M. J. Matari, “Reinforcement learning in the multi-robot domain,” *Robot colonies*, pp. 73–83, 1997.

- [9] I. Ghory, “Reinforcement learning in board games,” *Department of Computer Science, University of Bristol, Tech. Rep*, vol. 105, May 2004.
- [10] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, “Mastering the game of Go with deep neural networks and tree search,” *nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [11] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [12] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *BMVC 2015- Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.
- [13] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales, “When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition,” in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 142–150.
- [14] K. K. Nguyen, D. T. Hoang, D. Niyato, P. Wang, D. Nguyen, and E. Dutkiewicz, “Cyber-attack detection in mobile cloud computing: A deep learning approach,” in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, April 2018, pp. 1–6.
- [15] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.
- [16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [17] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

- [18] D. Bertsekas, *Dynamic programming and optimal control: Volume I*. Athena scientific, 2012, vol. 4.
- [19] R. E. Bellman and S. E. Dreyfus, *Applied dynamic programming*. Princeton university press, 2015, vol. 2050.
- [20] X. Wang, S. Wang, X. Liang, D. Zhao, J. Huang, X. Xu, B. Dai, and Q. Miao, “Deep reinforcement learning: A survey,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.
- [21] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, J. Pineau *et al.*, “An introduction to deep reinforcement learning,” *Foundations and Trends® in Machine Learning*, vol. 11, no. 3-4, pp. 219–354, 2018.
- [22] Kai, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
- [23] 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, 2015.
- [24] H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.
- [25] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, “Prioritized experience replay,” *arXiv preprint arXiv:1511.05952*, 2015.
- [26] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, “Dueling network architectures for deep reinforcement learning,” in *International conference on machine learning*. PMLR, 2016, pp. 1995–2003.
- [27] S.-i. Amari, “Natural gradient works efficiently in learning,” *Neural Computation*, vol. 10, no. 2, pp. 251–276, Feb. 1998.

- [28] S. M. Kakade, “A natural policy gradient,” *Advances in neural information processing systems*, vol. 14, 2001.
- [29] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.
- [30] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, pp. 229–256, 1992.
- [31] N. Heess, J. J. Hunt, T. P. Lillicrap, and D. Silver, “Memory-based control with recurrent neural networks,” *arXiv preprint arXiv:1512.04455*, 2015.
- [32] M. Zhao, Y. Wei, M. Song, and G. Da, “Power control for d2d communication using multi-agent reinforcement learning,” in *2018 IEEE/CIC International Conference on Communications in China (ICCC)*, Feb. 2018, pp. 563–567.
- [33] S. Fujimoto, H. Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.
- [34] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in neural information processing systems*, vol. 30, 2017.
- [35] N. Heess, J. J. Hunt, T. P. Lillicrap, and D. Silver, “Memory-based control with recurrent neural networks,” *arXiv preprint arXiv:1512.04455*, 2015.
- [36] V. Konda and J. Tsitsiklis, “Actor-critic algorithms,” *Advances in neural information processing systems*, vol. 12, 1999.
- [37] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” in *International conference on machine learning*. PMLR, 2016, pp. 1928–1937.

- [38] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” in *International conference on machine learning*. Pmlr, 2014, pp. 387–395.
- [39] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.
- [40] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, “Deep reinforcement learning for multi-agent systems: A review of challenges, solutions, and applications,” *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.
- [41] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, “Mean field multi-agent reinforcement learning,” in *International conference on machine learning*. PMLR, 2018, pp. 5571–5580.
- [42] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in neural information processing systems*, vol. 30, 2017.
- [43] M. Zhao, Y. Wei, M. Song, and G. Da, “Power control for d2d communication using multi-agent reinforcement learning,” in *2018 IEEE/CIC International Conference on Communications in China (ICCC)*, Aug. 2018, pp. 563–567.
- [44] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and L. D. Nguyen, “Distributed deep deterministic policy gradient for power allocation control in d2d-based v2v communications,” *IEEE Access*, vol. 7, pp. 164 533–164 543, Nov. 2019.
- [45] Z. Li, C. Guo, and Y. Xuan, “A multi-agent deep reinforcement learning based spectrum allocation framework for d2d communications,” in *2019 IEEE Global Communications Conference (GLOBECOM)*, Feb. 2019, pp. 1–6.
- [46] D. Wang, H. Qin, B. Song, X. Du, and M. Guizani, “Resource allocation in information-centric wireless networking with d2d enabled mec: A deep reinforcement learning approach,” *IEEE Access*, vol. 7, pp. 114 935–114 944, Aug. 2019.

- [47] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and M.-N. Nguyen, “Non-cooperative energy efficient power allocation game in d2d communication: A multi-agent deep reinforcement learning approach,” *IEEE Access*, vol. 7, pp. 100 480–100 490, July 2019.
- [48] Y. He, F. R. Yu, N. Zhao, and H. Yin, “Secure social networks in 5g systems with mobile edge computing, caching, and device-to-device communications,” *IEEE Wireless Communications*, vol. 25, no. 3, pp. 103–109, June 2018.
- [49] H. Yang, W.-D. Zhong, C. Chen, A. Alphones, and X. Xie, “Deep-reinforcement-learning-based energy-efficient resource management for social and cognitive internet of things,” *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5677–5689, March 2020.
- [50] M. Liu, Y. Teng, F. R. Yu, V. C. M. Leung, and M. Song, “A deep reinforcement learning-based transcoder selection framework for blockchain-enabled wireless d2d transcoding,” *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3426–3439, Feb. 2020.
- [51] Z. Ji, A. K. Kiani, Z. Qin, and R. Ahmad, “Power optimization in device-to-device communications: A deep reinforcement learning approach with dynamic reward,” *IEEE Wireless Communications Letters*, vol. 10, no. 3, pp. 508–511, March 2021.
- [52] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, “Deep reinforcement learning for joint channel selection and power control in d2d networks,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1363–1378, Feb. 2021.
- [53] T. Zhang, K. Zhu, and J. Wang, “Energy-efficient mode selection and resource allocation for d2d-enabled heterogeneous networks: A deep reinforcement learning approach,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1175–1187, Feb. 2021.
- [54] B. Gu, X. Zhang, Z. Lin, and M. Alazab, “Deep multiagent reinforcement-learning-based resource allocation for internet of controllable things,” *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3066–3074, March 2021.

- [55] G. Sun, G. O. Boateng, D. Ayepah-Mensah, G. Liu, and J. Wei, "Autonomous resource slicing for virtualized vehicular networks with d2d communications based on deep reinforcement learning," *IEEE Systems Journal*, vol. 14, no. 4, pp. 4694–4705, Dec. 2020.
- [56] Z. Bi and W. Zhou, "Deep reinforcement learning based power allocation for d2d network," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, June 2020, pp. 1–5.
- [57] S. Muy, D. Ron, and J.-R. Lee, "Energy efficiency optimization for swipt-based d2d-underlaid cellular networks using multiagent deep reinforcement learning," *IEEE Systems Journal*, vol. 16, no. 2, pp. 3130–3138, Aug. 2022.
- [58] R. Cheng, Y. Sun, Y. Liu, L. Xia, D. Feng, and M. A. Imran, "Blockchain-empowered federated learning approach for an intelligent and reliable d2d caching scheme," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 7879–7890, Aug. 2022.
- [59] J. Huang, Y. Yang, G. He, Y. Xiao, and J. Liu, "Deep reinforcement learning-based dynamic spectrum access for d2d communication underlay cellular networks," *IEEE Communications Letters*, vol. 25, no. 8, pp. 2614–2618, 2021.
- [60] X. Wang, R. Li, C. Wang, X. Li, T. Taleb, and V. C. M. Leung, "Attention-weighted federated deep reinforcement learning for device-to-device assisted heterogeneous collaborative edge caching," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 154–169, Nov. 2021.
- [61] I. Budhiraja, N. Kumar, and S. Tyagi, "Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay d2d communication," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3143–3156, Aug. 2021.
- [62] D. Shi, L. Li, T. Ohtsuki, M. Pan, Z. Han, and H. V. Poor, "Make smart decisions faster: Deciding d2d resource allocation via stackelberg game guided multi-agent deep reinforcement learning," *IEEE Transactions on Mobile Computing*, vol. 21, no. 12, pp. 4426–4438, June 2022.

- [63] D. Wang, Q. Liu, J. Tian, Y. Zhi, J. Qiao, and J. Bian, "Deep reinforcement learning for caching in d2d-enabled uav-relaying networks," in *2021 IEEE International Conference on Communications in China (ICCC)*, Nov. 2021, pp. 635–640.
- [64] W. Dong, Z. Li, and X. Chen, "Deep reinforcement learning-based adaptive clustering approach in short video sharing through d2d communication," in *2021 IEEE/CIC International Conference on Communications in China (ICCC)*, Dec. 2021, pp. 745–752.
- [65] Z. Sun and M. R. Nakhai, "Channel selection and power control for d2d communication via online reinforcement learning," in *ICC 2021 - IEEE International Conference on Communications*, June 2021, pp. 1–6.
- [66] Z. Li, Z. Liu, Y. Yuan, and H. Wang, "Dynamic channel matching based on deep reinforcement learning for d2d communications," in *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, vol. 1, June 2020, pp. 870–875.
- [67] H. Xiang, Y. Yang, G. He, J. Huang, and D. He, "Multi-agent deep reinforcement learning-based power control and resource allocation for d2d communications," *IEEE Wireless Communications Letters*, vol. 11, no. 8, pp. 1659–1663, April 2022.
- [68] Q. Guo, F. Tang, and N. Kato, "Federated reinforcement learning-based resource allocation in d2d-enabled 6g," *IEEE Network*, pp. 1–7, Sept. 2022.
- [69] J. Huang, Y. Yang, Z. Gao, D. He, and D. W. K. Ng, "Dynamic spectrum access for d2d-enabled internet of things: A deep reinforcement learning approach," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 793–17 807, March 2022.
- [70] M. A. Ouamri, G. Barb, D. Singh, A. B. M. Adam, M. S. A. Muthanna, and X. Li, "Non-linear energy-harvesting for d2d networks underlying uav with swipt using madqn," *IEEE Communications Letters*, vol. 27, no. 7, pp. 1804–1808, May 2023.
- [71] X. Wang, H. Shi, Y. Li, Z. Qian, and Z. Han, "Energy efficiency resource management for d2d-noma enabled network: A dinkelbach combined twin delayed deterministic policy gradient approach," *IEEE Transactions on Vehicular Technology*, pp. 1–16, March 2023.

- [72] Y. Bai, D. Wang, G. Huang, and B. Song, "A deep-reinforcement-learning-based social-aware cooperative caching scheme in d2d communication networks," *IEEE Internet of Things Journal*, vol. 10, no. 11, pp. 9634–9645, Jan. 2023.
- [73] I. Budhiraja, V. Vishnoi, N. Kumar, D. Garg, and S. Tyagi, "Energy-efficient optimization scheme for ris-assisted communication underlaying uav with noma," in *ICC 2022 - IEEE International Conference on Communications*, Aug 2022, pp. 1–6.
- [74] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in uav-ris wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2042–2055, July 2021.
- [75] W. Huang, Y. Chen, J. Wang, X. Li, and S. Jin, "Reconfigurable intelligent surface-enhanced broadband ofdm communication based on deep reinforcement learning," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, Dec. 2021, pp. 1–6.
- [76] R. Zhong, Y. Liu, X. Mu, Y. Chen, and L. Song, "Ai empowered ris-assisted noma networks: Deep learning or reinforcement learning?" *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 182–196, Nov. 2022.
- [77] A. Faisal, I. Al-Nahhal, O. A. Dobre, and T. M. N. Ngatched, "Deep reinforcement learning for optimizing ris-assisted hd-fd wireless systems," *IEEE Communications Letters*, vol. 25, no. 12, pp. 3893–3897, Oct. 2021.
- [78] A. Khalili, E. M. Monfared, S. Zargari, M. R. Javan, N. M. Yamchi, and E. A. Jorswieck, "Resource management for transmit power minimization in uav-assisted ris hetnets supported by dual connectivity," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1806–1822, March 2022.
- [79] K. K. Nguyen, S. R. Khosravirad, D. B. da Costa, L. D. Nguyen, and T. Q. Duong, "Reconfigurable intelligent surface-assisted multi-uav networks: Efficient resource allocation with deep reinforcement learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 358–368, April 2022.

- [80] J. Zhao, L. Yu, K. Cai, Y. Zhu, and Z. Han, "Ris-aided ground-aerial noma communications: A distributionally robust drl approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 4, pp. 1287–1301, Jan. 2022.
- [81] M. Alsenwi, M. Abolhasan, and J. Lipman, "Intelligent and reliable millimeter wave communications for ris-aided vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21 582–21 592, Nov. 2022.
- [82] X. Fan, M. Liu, Y. Chen, S. Sun, Z. Li, and X. Guo, "Ris-assisted uav for fresh data collection in 3d urban environments: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 1, pp. 632–647, Aug. 2023.
- [83] K. K. Nguyen, A. Masaracchia, V. Sharma, H. V. Poor, and T. Q. Duong, "Ris-assisted uav communications for iot with wireless power transfer using deep reinforcement learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 5, pp. 1086–1096, May 2022.
- [84] X. Yuan, S. Hu, W. Ni, R. P. Liu, and X. Wang, "Joint user, channel, modulation-coding selection, and ris configuration for jamming resistance in multiuser ofdma systems," *IEEE Transactions on Communications*, vol. 71, no. 3, pp. 1631–1645, Jan. 2023.
- [85] Y. Zou, Y. Liu, X. Mu, X. Zhang, Y. Liu, and C. Yuen, "Machine learning in ris-assisted noma iot networks," *IEEE Internet of Things Journal*, pp. 1–1, Feb. 2023.
- [86] H. Peng and L.-C. Wang, "Energy harvesting reconfigurable intelligent surface for uav based on robust deep reinforcement learning," *IEEE Transactions on Wireless Communications*, pp. 1–1, Feb. 2023.
- [87] T. Zhang, P. Ren, D. Xu, and Z. Ren, "Ris subarray optimization with reinforcement learning for green symbiotic communications in internet of things (iot)," *IEEE Internet of Things Journal*, pp. 1–1, April 2023.
- [88] K. Guo, M. Wu, X. Li, H. Song, and N. Kumar, "Deep reinforcement learning and noma-based multi-objective ris-assisted is-uav-tns: Trajectory optimization and beamforming design," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, April 2023.

- [89] J. Li, W. Wang, R. Jiang, X. Wang, Z. Fei, S. Huang, and X. Li, "Piecewise-drl: Joint beamforming optimization for ris-assisted mu-miso communication system," *IEEE Internet of Things Journal*, pp. 1–1, May 2023.
- [90] P. Saikia, K. Singh, O. Taghizadeh, W.-J. Huang, and S. Biswas, "Drl algorithms for efficient spectrum sharing in ris-aided mimo radar and cellular systems," in *MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM)*, Nov. 2022, pp. 55–60.
- [91] M.-L. Tham, Y. J. Wong, A. Iqbal, N. B. Ramli, Y. Zhu, and T. Dagiuklas, "Deep reinforcement learning for secrecy energy-efficient uav communication with reconfigurable intelligent surface," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, May 2023, pp. 1–6.
- [92] L. Zhang, A. Celik, S. Dang, and B. Shihada, "Energy-efficient trajectory optimization for uav-assisted iot networks," *IEEE Transactions on Mobile Computing*, vol. 21, no. 12, pp. 4323–4337, April 2022.
- [93] P. Saikia, S. Pala, K. Singh, S. K. Singh, and W.-J. Huang, "Proximal policy optimization for ris-assisted full duplex 6g-v2x communications," *IEEE Transactions on Intelligent Vehicles*, pp. 1–16, May 2023.
- [94] D. Feng, L. Lu, Y. Yuan-Wu, G. Li, S. Li, and G. Feng, "Device-to-device communications in cellular networks," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 49–55, 2014.
- [95] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlying cellular networks," *IEEE Transactions on Wireless communications*, vol. 10, no. 8, pp. 2752–2763, 2011.
- [96] B. Kaufman, J. Lilleberg, and B. Aazhang, "Spectrum sharing scheme between cellular users and ad-hoc device-to-device users," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1038–1049, 2013.

- [97] H. ElSawy, E. Hossain, and M.-S. Alouini, “Analytical modeling of mode selection and power control for underlay d2d communication in cellular networks,” *IEEE Transactions on Communications*, vol. 62, no. 11, pp. 4147–4161, 2014.
- [98] I. Cha, Y. Shah, A. U. Schmidt, A. Leicher, and M. V. Meyerstein, “Trust in m2m communication,” *IEEE Vehicular Technology Magazine*, vol. 4, no. 3, pp. 69–75, Sept 2009.
- [99] Y. Luo, P. Hong, R. Su, and K. Xue, “Resource allocation for energy harvesting-powered d2d communication underlying cellular networks,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 486–10 498, 2017.
- [100] I. Budhiraja, N. Kumar, S. Tyagi, S. Tanwar, and Z. Han, “An energy efficient scheme for wpcn-noma based device-to-device communication,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 11, pp. 11 935–11 948, July 2021.
- [101] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, “Optimal user scheduling and power allocation for millimeter wave noma systems,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1502–1517, Dec. 2018.
- [102] Z. Qin, X. Yue, Y. Liu, Z. Ding, and A. Nallanathan, “User association and resource allocation in unified noma enabled heterogeneous ultra dense networks,” *IEEE Communications Magazine*, vol. 56, no. 6, pp. 86–92, June 2018.
- [103] I. Budhiraja, S. Tyagi, S. Tanwar, N. Kumar, and M. Guizani, “Cross layer noma interference mitigation for femtocell users in 5g environment,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 4721–4733, Mar. 2019.
- [104] I. Budhiraja, N. Kumar, and S. Tyagi, “Cross-layer interference management scheme for d2d mobile users using noma,” *IEEE Systems Journal*, vol. 15, no. 2, pp. 3109–3120, July 2021.
- [105] ———, “Energy-delay tradeoff scheme for noma-based d2d groups with wpcns,” *IEEE Systems Journal*, vol. 15, no. 4, pp. 4768–4779, Aug. 2020.

- [106] I. Budhiraja, R. Gupta, N. Kumar, S. Tyagi, S. Tanwar, and J. J. P. C. Rodrigues, “Interference mitigation and secrecy ensured for noma-based d2d communications under imperfect csi,” in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–6.
- [107] I. Budhiraja, S. Tyagi, S. Tanwar, N. Kumar, and J. J. P. C. Rodrigues, “Diya: Tactile internet driven delay assessment noma-based scheme for d2d communication,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6354–6366, Apr. 2019.
- [108] J. Gu, S. J. Bae, S. F. Hasan, and M. Y. Chung, “Heuristic algorithm for proportional fair scheduling in d2d-cellular systems,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 769–780, Jan. 2016.
- [109] D. Bertsekas, “Dynamic programming and optimal control, i and ii, athena scientific, belmont, massachusetts,” *New York-San Francisco-London*, 1995.
- [110] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” in *Advances in neural information processing systems*, 2000, pp. 1057–1063.
- [111] D. T. Ngo, S. Khakurel, and T. Le-Ngoc, “Joint subchannel assignment and power allocation for ofdma femtocell networks,” *IEEE Transactions on Wireless Communications*, vol. 13, no. 1, pp. 342–355, Jan. 2014.
- [112] B. R. Marks and G. P. Wright, “A general inner approximation algorithm for nonconvex mathematical programs,” *Operations research*, vol. 26, no. 4, pp. 681–683, 1978.
- [113] M. Chiang, C. W. Tan, D. P. Palomar, D. O’neill, and D. Julian, “Power control by geometric programming,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 7, pp. 2640–2651, July 2007.
- [114] H. Yang, A. Alphones, W.-D. Zhong, C. Chen, and X. Xie, “Learning-based energy-efficient resource management by heterogeneous rf/vlc for ultra-reliable low-latency industrial iot networks,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5565–5576, Aug. 2020.

- [115] J. Papandriopoulos and J. S. Evans, “Scale: A low-complexity distributed protocol for spectrum balancing in multiuser dsl networks,” *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3711–3724, Aug 2009.
- [116] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [117] C. Zhang, P. Patras, and H. Haddadi, “Deep learning in mobile and wireless networking: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2224–2287, Mar. 2019.
- [118] I. Budhiraja, N. Kumar, H. Sharma, M. Elhoseny, Y. Lakys, and J. J. P. C. Rodrigues, “Latency-energy tradeoff in connected autonomous vehicles: A deep reinforcement learning scheme,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, Nov. 2022.
- [119] L. ETSI, “Evolved universal terrestrial radio access (eutra),” *User Equipment (UE) radio access capabilities (3GPP TS 36.306 version 10.4 Release 10)*, Mar. 2012.
- [120] M. Series, “Guidelines for evaluation of radio interface technologies for imt-2020,” *IMT*, Munich Germany, June 2017.
- [121] K. K. Nguyen, T. Q. Duong, N. A. Vien, N. A. Le-Khac, and M. N. Nguyen, “Non-Cooperative Energy Efficient Power Allocation Game in D2D Communication: A Multi-Agent Deep Reinforcement Learning Approach,” *IEEE Access*, vol. 7, pp. 100 480–100 490, July 2019.
- [122] Z. Ji, A. K. Kiani, Z. Qin, and R. Ahmad, “Power Optimization in Device-to-Device Communications: A Deep Reinforcement Learning Approach With Dynamic Reward,” *IEEE Wireless Communications Letters*, vol. 10, no. 3, pp. 508–511, March 2021.
- [123] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, “Deep reinforcement learning for joint channel selection and power control in d2d networks,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1363–1378, Feb. 2021.