



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

**MACHINE LEARNING-BASED RADIO RESOURCE
MANAGEMENT FOR COGNITIVE RADIO NETWORKS**

DISSERTATION

for the Degree of

DOCTOR OF PHILOSOPHY
(Electrical, Electronic and Computer Engineering)

HOANG THI HUONG GIANG

MAY 2021

**Machine Learning-based Radio Resource Management for
Cognitive Radio Networks**

Supervisor: Professor In-Soo Koo

DISSERTATION

Submitted in Partial Fulfillment
of the Requirements for the
Degree of

DOCTOR OF PHILOSOPHY
(Electrical, Electronic and Computer Engineering)

at the

UNIVERSITY OF ULSAN

by

Hoang Thi Huong Giang
May 2021

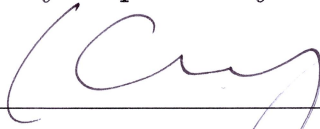
Publication No. _____

©2021 - Hoang Thi Huong Giang

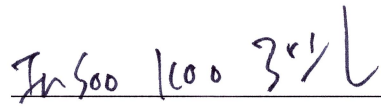
All rights reserved.

Machine Learning-based Radio Resource Management for
Cognitive Radio Networks

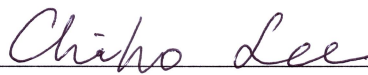
Approved by Supervisory Committee:



Prof. Hyung-Yun Kong, Chair



Prof. In-Soo Koo, Supervisor



Dr. Chi-Ho Lee



Prof. Young-Tae Noh



Prof. Sang-Jo Choi

Department of Electrical, Electronic and Computer Engineering

University of Ulsan, South Korea

Date: May, 2021

VITA

Hoang Thi Huong Giang was born in Nam Dinh City, Vietnam, in 1990. She received her bachelor's degree in Electronics and Telecommunications Engineering from Ton Duc Thang University, Ho Chi Minh City, Vietnam, in 2013, and her master's degree from Graduate Institute of Digital Mechatronic Technology, College of Engineering, in Chinese Culture University, Taiwan, in 2015. Since August 2015, she has been working as a lecturer at the Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Vietnam.

Since September 2016, she has been pursuing her Ph.D. degree in Electrical, Electronic and Computer Engineering at the University of Ulsan (UOU), South Korea, under the supervision of Professor Insoo Koo. Her current research focuses on POMDP, reinforcement learning, deep neural network and their applications to cognitive radio networks under energy constraint.

*Dedicated to my dearest family and friends
for
their endless love, support and encouragement*

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my dear parents for giving me the strength and education to reach for the stars and chase my dreams. I am also grateful to my little sister who has always been available to exchange confidences whenever I need in spite of different time zone. In addition, I thank my love and also my colleague, Pham Duy Thanh, who was always companion with me during my study and always encouraged me, helped me when I was in tough time with my research.

I owe my deepest gratitude to my advisor, *Professor Insoo Koo*, for offering me the opportunity to pursue my Ph.D. and be a part of his research group. His kindness, constant support, encouragement, and persistent guidance are invaluable since it has helped me a lot in doing research, especially when dealing with problems.

I would like to thank to the members of my Ph.D. supervisory committee for their valuable and useful comments that help a lot to improve the quality of this dissertation.

I am grateful to all members of the multimedia communications system laboratory (MCSL) for their friendship, enthusiastic help, and cheerfulness during my study in Korea. Especially, I would like to thank Dr. Vu Van Hiep, Dr. Tran Nhut Khai Hoan for their valuable discussion, collaboration, and useful guidance throughout my Ph.D. study. I am grateful to my little friend, Linh, for her kind support, whenever I am in trouble to deal with any problem related to Korean interpretation. I also thank Carla, Mario, Iqra, Thien, Dung, and Toan for all that we spent together.

Last but not least, I gratefully appreciate the BK21 Plus for financial support during my study in University of Ulsan.

Hoang Thi Huong Giang

Ulsan, South Korea, May - 2021.

ABSTRACT

Machine Learning-based Radio Resource Management for Cognitive Radio Networks

by

Hoang Thi Huong Giang

Supervisor: Prof. In-Soo Koo

Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy (Electrical, Electronic and Computer Engineering)

May 2021

Spectrum scarcity is one of the essential issues in fifth-generation (5G) and beyond communication systems. Moreover, in the last few decades, the number of dramatically increasing mobile applications led to surging demand for radio resources. In order to tackle with spectrum inefficiency issue, the dynamic spectrum access techniques (i.e., cognitive radio (CR)), ambient backscatter communication, and non-orthogonal multiple access (NOMA) are studied. In cognitive radio networks (CRNs), cognitive users (CUs) are able to utilize the licensed spectrum bands of the primary users (PUs) while either the interference caused by the cognitive users is acceptable or the PUs are inactive at that time. On the other hand, ambient backscatter communication is emerging technique for green communication, where its key idea is to transmit data from a transmitter to its corresponding receiver by backscattering the signals via an ambient radio frequency (RF) source. In addition, NOMA allows multiple users to use the same frequency and time resources for their data transmissions. The integration of these techniques is capable of further advancing the spectrum efficiency in wireless communication systems.

Along with rapid developments of mobile devices, energy management also becomes a crucial issue since most of the smart mobile devices require long-term operation to meet their high energy consumption applications, but the battery capacity is still limited. In recent works, wireless communications powered by external harvested energy have become a promising technique to solve the energy-constrained problem. Radio frequency (RF)-harvested energy in a CRN is one of the potential solutions for energy-constrained issue in

wireless networking, where the wireless devices can harvest energy from ambient RF signals. In addition, the wireless devices can also harvest ambient energy for their rechargeable batteries from perpetual non-RF sources (i.e., solar, wind. . .).

Nowadays, dynamic resource allocation algorithms for the energy harvesting CRNs are carefully being investigated due to the crucial effect of resource management on long-term system performance. Motivated by the aforementioned survey, this dissertation will focus on these remaining issues for CRNs as follows:

Firstly, we investigated jamming attacks in the physical layer against cooperative communications networks, where a jammer tries to block the data communication between the source and destination. An energy-constrained relay is able to assist the source to forward the data to the destination even when the jammer tries to block the direct link. Due to a limited capacity battery of the relay, a non-radio frequency energy harvester equipped in the relay helps to prolong its operation. We propose a scheme based on a partially observable Markov decision process (POMDP) to find the optimal action for the source such that we can maximize the achievable throughput of cooperative communications networks. Under this scheme, the source dynamically selects the appropriate action mode for its transmission in order to obtain maximum throughput under the jamming attack. Simulation results verify that the proposed scheme is superior to the myopic scheme where only current throughput is taken into account for making decisions.

Secondly, wireless energy harvesting enables wireless-powered communications to accommodate data services in a self-sustainable manner over a long operational time. Along with energy harvesting, an ambient backscatter technique helps a secondary transmitter reflect existing RF signal sources to communicate with a secondary receiver when the primary channel (PC) is utilized. However, secondary system performance is significantly affected by factors such as the availability of the primary channel, imperfect spectrum sensing, and energy-constrained problems. Therefore, we propose a novel approach for wireless-powered CRNs to improve the transmission performance of secondary systems. To reduce the dependence of the secondary system on RF sources, in the paper, we provide a new paradigm by integrating ambient backscattering with both RF and non-RF wireless-powered communications to facilitate secondary communications. Based on the sensing result in a time slot, the secondary transmitter can dynamically select the operational action: 1) backscattering, 2) harvesting or 3) transmitting to maximize the long-term achievable data transmission rate at the secondary receiver. In addition, the optimal action set for cognitive

radio networks with wireless-powered ambient backscatter is selected by the POMDP, which maximizes an expected transmission rate calculated over a number of subsequent time slots. The proposed scheme aims to improve long-term transmission rate of CRNs with wireless-powered ambient backscatter in comparison with conventional schemes where an action is taken only to maximize the immediate reward in every single time slot.

Thirdly, we consider an uplink NOMA cognitive system, where the SUs can jointly transmit data to the cognitive base station (CBS) over the same spectrum resources. Thereafter, successive interference cancellation (SIC) is applied at the CBS to retrieve signals transmitted by the SUs. In addition, the energy-constrained problem in wireless networks is taken into account. Therefore, we assume that the SUs are powered by a wireless energy harvester to prolong their operations; meanwhile, the CBS is equipped with a traditional electrical supply. Herein, we propose an actor-critic reinforcement learning approach to maximize the long-term throughput of the cognitive network. In particular, by interacting and learning directly from the environment over several time slots, the CBS can optimally assign the amount of transmission energy for each SU according to the remaining energy of the SUs and the availability of the primary channel. As a consequence, the simulation results verify that the proposed scheme outperforms other conventional approaches (such as Myopic NOMA and OMA), so the system reward is always maximized in the current time slot, in terms of overall throughput and energy efficiency.

Then, a hybrid NOMA/OMA scheme is considered for uplink wireless transmission systems where multiple cognitive users (CUs) can simultaneously transmit their data to a cognitive base station (CBS). We adopt a user-pairing algorithm in which the CUs are grouped into multiple pairs, and each group is assigned to an orthogonal sub-channel such that each user in a pair applies NOMA to transmit data to the CBS without causing interference with other groups. Subsequently, the signal transmitted by the CUs of each NOMA group can be independently retrieved by using successive interference cancellation (SIC). The CUs are assumed to harvest solar energy to maintain operations. Moreover, joint power and bandwidth allocation is taken into account at the CBS to optimize energy and spectrum efficiency in order to obtain the maximum long-term data rate for the system. To this end, we propose a deep actor-critic reinforcement learning (DACRL) algorithm to respectively model the policy function and value function for the actor and critic of the agent (i.e., the CBS), in which the actor can learn about system dynamics by interacting with the environment. Meanwhile, the critic can evaluate the action taken such that the

CBS can optimally assign power and bandwidth to the CUs when the training phase finishes. Numerical results validate the superior performance of the proposed scheme, compared with other conventional schemes.

Next, we consider an uplink solar-powered cognitive radio networks (CRNs) where multiple secondary users (SUs) transmit data to a secondary base station (SBS) by sharing a licensed channel of a primary system. A deep Q-learning (DQL) algorithm, which combines non-orthogonal multiple access (NOMA) and time division multiple access (TDMA) techniques, is proposed to maximize the long-term throughput of the system. By using our scheme, the agent (i.e. the SBS) can obtain the optimal decision by interacting with the environment to learn about system dynamics. Simulation results validate the superiority of the performance under the proposed scheme, compared with traditional schemes.

Consequently, we end up this dissertation by summarizing its main contributions and opening a new door for deep reinforcement learning and its applications in future wireless networks.

Contents

Supervisory Committee	ii
Vita	iii
Dedication	iv
Acknowledgments	v
Abstract	vi
Table of Contents	x
List of Figures	xiii
Nomenclature	xv
1 Introduction	1
1.1 Background	1
1.1.1 Cognitive Radio Network	1
1.1.2 Machine Learning	2
1.2 Thesis Motivation and Objective	2
1.3 Thesis Outline	4
2 POMDP-Based Throughput Maximization for Cooperative Communications Networks with Energy-Constrained Relay under Attack in the Physical Layer	6
2.1 Introduction	6
2.2 System Model	11
2.3 Optimal Mode Decision Policy Based on POMDP	17
2.3.1 Relay-assisted Transmission Mode	18
2.3.2 Direct Transmission Mode	20
2.4 Simulation Results	22
2.5 Conclusion	25
3 A POMDP-based Long-Term Transmission Rate Maximization for Cognitive Radio Networks with Wireless-Powered Ambient Backscatter	27
3.1 Introduction	27
3.2 System Description	32
3.2.1 Network Model	32
3.2.2 Non-RF Energy Harvesting Model	35
3.2.3 Spectrum Sensing	35

3.3	Problem Formulation	36
3.4	Proposed Scheme	37
3.4.1	Proposed Scheme Description and Observations	37
3.4.1.1	Backscattering	38
3.4.1.2	Harvesting	39
3.4.1.3	Transmitting	41
3.4.2	Overall Expected Reward	43
3.4.3	Optimal Mode Decision Policy	45
3.5	Simulations	46
3.6	Conclusion	52
4	Uplink NOMA-based Long-Term Throughput Maximization Scheme for Cognitive Radio Networks: An Actor-Critic Reinforcement Learning Approach	53
4.1	Introduction	53
4.2	System Model	56
4.2.1	Network Model	56
4.2.2	Energy Harvesting and Primary User Models	59
4.2.3	Imperfect Spectrum Sensing	60
4.3	Problem Formulation	61
4.4	Actor-Critic Reinforcement Learning-Based Algorithm for Uplink NOMA in Cognitive Radio Networks	62
4.4.1	Markov Decision Process	62
4.4.2	Actor-Critic Reinforcement Learning Algorithm	63
4.4.2.1	Silent Mode (Ω_1)	66
4.4.2.2	Transmission Mode	66
4.5	Simulation Results	68
4.6	Conclusion	76
5	Hybrid NOMA/OMA-Based Dynamic Power Allocation Scheme Using Deep Reinforcement Learning in 5G Networks	78
5.1	Introduction	78
5.2	System Model	82
5.2.1	Energy Arrival and Primary User Models	85
5.3	Long-Term Transmission Rate Maximization Problem Formulation	86
5.4	Deep Reinforcement Learning-Based Resource Allocation Policy	87
5.4.1	Markov Decision Process	88
5.4.1.1	Silent Mode	89
5.4.1.2	Transmission Mode	90
5.4.2	Deep Actor-Critic Reinforcement Learning Algorithm	92
5.4.2.1	The Critic with a DNN	93
5.4.2.2	The Actor with a DNN	93
5.5	Simulation Results	97
5.6	Conclusions	102

6	Deep Q-learning-based Resource Allocation for Solar-powered Users in Cognitive Radio Networks	103
6.1	Introduction	103
6.2	System Model	108
6.2.1	Network Model	108
6.2.2	Energy Arrival and Primary Channel Models	110
6.3	Long-term Throughput Maximization Problem Formulation	111
6.4	Deep Q-Learning-Based Resource Allocation Policy	112
6.4.1	Decision-making process	112
6.4.2	Observations	113
6.4.2.1	Silent Mode	114
6.4.2.2	Transmission Mode	114
6.4.3	Deep Q-Learning	116
6.5	Simulation Results	118
6.5.1	Simulation Setting	118
6.5.2	Results and Discussion	119
6.6	Conclusion	124
7	Summary of Contributions and Future Works	125
7.1	Introduction	125
7.2	Summary of Contributions	125
7.3	Future Directions	127
	Publications	129
	Bibliography	131

List of Figures

2.1	The system model of the proposed scheme.	11
2.2	Frame structure.	12
2.3	A Markov chain model of the jammer.	16
2.4	Flowchart of the proposed scheme.	16
2.5	Average throughput versus required transmitted energy.	23
2.6	Average throughput versus capacity of battery.	24
2.7	Average throughput versus capacity of the battery when detection probability $P_d = 0.4, 0.6,$ and 0.9	24
2.8	Average throughput versus detection probability P_d when transmission energy $E_{tr} = 4, 6,$ and 8	25
3.1	System model of the considered network.	32
3.2	Schematic structure of secondary transmitter.	33
3.3	The time frame structure of the secondary user.	33
3.4	Markov chain model of the PU.	34
3.5	The flowchart of the proposed scheme.	38
3.6	The long-term transmission rate of the secondary system under various values for harvested non-RF energy.	48
3.7	The energy efficiency of the secondary system under various values of harvested non-RF energy.	48
3.8	The selected action statistics of the secondary system under various values of harvested non-RF energy.	49
3.9	The long-term transmission rate of the secondary system according to different communications ranges.	50
3.10	The energy efficiency of the secondary system according to different communications ranges.	51
3.11	The selected action statistics of the secondary system according to different communications ranges.	52
4.1	System model of the proposed scheme.	57
4.2	Time frame of the three phases in the secondary users's operations.	57
4.3	Illustration of SIC detection of the signals at the CBS.	59
4.4	Markov chain model of the primary user.	60

4.5	The flowchart of the proposed scheme.	64
4.6	The actor-critic learning process.	65
4.7	The convergence process of the actor-critic according to different values of learning step-size.	70
4.8	Average throughput of the secondary system under various values of harvested energy.	71
4.9	Energy efficiency of the secondary system for various values of harvested energy.	72
4.10	The selected action statistics of each secondary user using the actor-critic NOMA approach for various values of harvested energy.	73
4.11	The selected action statistics of each secondary user using the Myopic NOMA approach for various values of harvested energy.	73
4.12	Average throughput for different values of h_1 and h_2	74
4.13	Energy efficiency according to the channel gain between the CBS and SU_1	75
4.14	Average throughput according to the noise variance.	75
4.15	Energy efficiency according to the noise variance.	76
5.1	System model of the proposed scheme.	82
5.2	Time frame of the cognitive users' operations.	84
5.3	Markov chain model of the primary channel.	86
5.4	The agent-environment interaction process.	88
5.5	The structure of deep actor-critic reinforcement learning.	91
5.6	The deep neural network in the critic.	92
5.7	The deep neural network in the actor.	94
5.8	The convergence rate of the proposed actor-critic deep reinforcement learning with different training steps in each episode.	98
5.9	The convergence rate of the proposed actor-critic deep reinforcement learning according to different learning rate values.	99
5.10	Average transmission rate according to different values for mean harvested energy.	99
5.11	Energy efficiency according to different values of harvested mean energy.	100
5.12	Average transmission rate according to noise variance.	101
5.13	Energy efficiency according to noise variance.	101
6.1	The considered network model.	107
6.2	Time frame structure.	108
6.3	Two-state Markov discrete-time model for states of the primary channel.	110
6.4	Overall structure of the proposed DQL-based power allocation scheme.	113
6.5	Structure of the neural network used for DQL.	116
6.6	The convergence behavior of the proposed scheme.	121
6.7	The average long-term throughput according to various values for mean harvested energy.	121
6.8	Energy efficiency according to various values for mean harvested energy.	122
6.9	Average throughput according to various values for noise variance.	123
6.10	Energy efficiency according to various values for noise variance.	123

Nomenclature

Notation Description

AWGN	Additive White Gaussian Noise
CBS	Cognitive Base Station
CRN	Cognitive Radio Network
CU	Cognitive User
CSI	Channel State Information
CSS	Cooperative Spectrum Sensing
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
DSA	Dynamic Spectrum Access
IoT	Internet of Thing
MDP	Markov Decision Process
ML	Machine Learning
NOMA	Non Orthogonal Multiple Access
PBS	Primary Base Station
POMDP	Partially Observable Markov Decision Process
PU	Primary User
RF	Radio Frequency
RL	Reinforcement Learning
SBS	Secondary Base Station
SNR	Signal-to-Noise Ratio
SIC	Successive Interference Cancellation
SR	Secrecy Rate
SU	Secondary User
TDMA	Time Division Multiple Access

Chapter 1

Introduction

1.1 Background

1.1.1 Cognitive Radio Network

Over the last 2 decades, we have been witnesses for the enormous number of successes of wireless communications technology and an impressive increase of the data traffic. Consequently, the limited spectrum resource conflicts with growing demands due to tremendous increase of mobile devices. According to the report of Federal Communications Commission, spectrum scarcity currently becomes largely because of the physical shortage of the spectrum [1]. In recent years, cognitive radio (CR) has emerged as the most expected candidate to deal with the spectrum scarcity and enhance the spectrum utilization, which is also known as dynamic spectrum access (DSA). In a cognitive radio network (CRN), the main components are the primary network and the secondary network. The primary network is the initial licensed network that possesses the spectrum. The secondary network is the unlicensed network, which intends to access the spectrum. In order to access the licensed spectrum, the secondary users (SUs) need to perform the spectrum sensing (SS) to identify the available spectrum bands, where temporally no primary users (PUs) are active. Regarding several works [2, 3], the network performance would be reasonably weakened due to the imperfect SS. For this reason, the SUs should perform cooperative spectrum sensing (CSS) techniques to improve the accuracy of decisions such as the detection probability is higher and the false alarm probability is smaller. In CSS, a number of SUs individually sense the licensed spectrum band and send their local sensing result to a fusion center where

the local sensing information is collected using specific rules to determine a final decision about the state of the primary channel. Thereafter, the fusion center broadcasts the final decision to the SUs in CRNs.

1.1.2 Machine Learning

Machine learning (ML) is a part of computer science, it is a process that trains a computer to learn from its past experiences to tackle a given problem. Nowadays, ML has been significantly attracted interest thanks to the concept that ML is able to solve problems faster than human. Moreover, it is possible to process and analyze a numerous amount of data to explore insights among the data that the people are not enable to observe obviously. The intelligent decision of the machine is based on different algorithms, which enables the machine to learn from its experiences in order to produce good judgments. ML can be classified into 4 main categories according to their purpose, such as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. In this dissertation, we focus on reinforcement learning (RL).

RL is a prominent solution to deal with Markov decision processes (MDPs), which allows an agent to learn the optimal decision policy when the agent has no information about the surrounding environment. In RL, the agent periodically selects actions, observe the results, and then automatically adjust its strategy to obtain the optimal policy. However, the learning process of RL takes a lot of time to converge because it needs time to explore and gain knowledge from the environment. Thus, this makes it inefficient and inapplicable into large-scale networks. Recently, deep reinforcement learning (DRL) has emerged as an advanced version of RL, which is combined of RL and deep neural network (DNN). Consequently, DRL can overcome the limitations of RL, and thus provides better solutions to large-scale and sophisticated problems.

1.2 Thesis Motivation and Objective

In the past few decades, with the explosive development of mobile communications and multimedia services, the traffic demand has been increasing dramatically. Consequently, more and more radio spectrum resources are needed to satisfy these demands. Nevertheless, most of allocated spectrum are not efficiently utilized. From this perspective, CRN has been

proposed as a promising solution that allows SUs to opportunistically access the licensed spectrum when it is temporally unoccupied by PUs [4].

Moreover, the SUs has limited capacity and are often powered by energy harvester modules, which are enable to gather energy from ambient sources, such as wind power, solar power, and RF energy. The devices with finite capacity always harvest and store the energy in their rechargeable batteries to carry out the operations of the network. Hence, it is essential to allocate the energy resource based on machine learning to improve the network lifetime and long-term network performance.

In general, the SUs are only enabled to utilize the licensed spectrum when they do not make any interference to the primary network. Therefore, SS plays a crucial role in the PUs' activities detection procedure. Inherently, SS is the signal detection, and it can be applied to identify the presence of the PUs. However, the performance of single detector may be reduced in the practical environment due to shadowing, fading, and hidden nodes issues. In order to deal with these issues, CSS is proposed to improve the detection performance and identify the state of PUs more sensitively.

The main objective of this dissertation is to solve the aforementioned issues by using machine learning-based methods, such as value iteration-based dynamic programming, reinforcement learning, and deep learning. The contributions of this dissertation are summarized as follows:

- (i) First, we investigate jamming attacks in the physical layer against cooperative communications networks to maximize the achievable throughput of the networks under the jamming attack.
- (ii) Second, we adopts the ambient backscatter for the secondary transceiver communication in wireless-powered CRN, in which SUs are powered by both non-radio frequency (RF) and RF energy harvesters to deal with energy-constrained problem.
- (iii) Third, we propose the uplink NOMA technique, where the SUs can simultaneously transmit data on the same channel and in the same time slot, with the objective being spectral efficiency.
- (iv) Next, we study joint power and bandwidth allocation to optimize energy and spectrum efficiency in order to obtain the maximum long-term data rate for the system.

- (v) Finally, we propose deep reinforcement learning for resource allocation, which aims to maximize the long-term throughput of the system under energy constraints of the SUs.

1.3 Thesis Outline

The rest of this dissertation is organized as follows. Chapter 2 presents a cooperative communication scheme with energy-constrained relay under jamming attack. Chapter 3 introduces the ambient backscatter for the secondary transceiver communication in wireless-powered CRN with both non-RF and RF energy harvesting. Chapter 4 studies the uplink NOMA for CRNs. Chapter 5 describes a model of a hybrid NOMA/OMA uplink CRN adopting energy harvesting at the CUs. Chapter 6 considers a resource allocation for solar-powered users in CRNs. Chapter 7 discusses about the future directions of research. A brief description of each chapter is given below.

Chapter 2 considers the average throughput maximization of cooperative communications networks under a jamming attack. Moreover, the energy-constraint problem of a relay is also considered. In this network, the source and cooperate with a relay to enhance the achievable throughput under the presence of jamming. The network is assumed to follow a synchronous in a time-slotted manner. At the beginning of each time slot, the source performs SS to detect the state of jammer. After that, the source can decide whether to cooperate with a relay or not based on the sensing result and state of network. The problem is formulated and solved by the partially observable Markov decision process (POMDP).

Chapter 3 studies an optimal transmit power decision policy for maximizing the long-term transmission rate of wireless-powered CRNs utilizing ambient backscatter. In this network, the self-sustainable communication scheme for secondary users is investigated by combining ambient backscatter technology and wireless energy harvesting technology in CRNs. In order to share the spectrum resource with PUs, the SUs first performs SS to identify if the primary channel is free or not. Subsequently, based on the sensing result, the secondary transmitter will select appropriate mode whether it is backscattering, RF energy harvesting, or data transmitting. Then, the problem can be solved by using a value-iterations method.

Chapter 4 proposes an uplink NOMA-based transmission power allocation scheme for CRNs. Herein, the SUs employed the NOMA technique are able to simultaneously transmit their data to the cognitive base station (CBS). We formulate the problem of

throughput maximization based on a MDP. Afterward, the actor-critic reinforcement learning approach is adopted to tackle the problem, where the CBS can interact with the environment and allocate the optimal amount of energy for each user in order to maximize the long-term performance of the network.

Chapter 5 studies efficient joint power and bandwidth allocation by adopting hybrid NOMA/OMA in uplink CRNs. In particular, solar energy-powered CUs are assigned the proper transmission power and bandwidth to transmit data to a CBS in every time slot in order to maximize the long-term data transmission rate of the system. We propose a deep actor-critic reinforcement learning framework, which is a combination of DNN and reinforcement learning, to allocate the appropriate transmission power and bandwidth to the CUs by directly interacting with the environment.

Chapter 6 employs a deep Q-learning algorithm in order to optimal policy for the system from trial-and-error interactions with the environment after training. We investigate a NOMA/TDMA-based deep Q-learning approach to maximize the long-term throughput of a secondary system. Throughout the training phase, the proposed scheme does not have prior knowledge of the harvested energy distribution of SUs. However, that information can be learned, and then an optimal decision policy is achieved.

Chapter 7 concludes this dissertation and gives a discussion on future research directions.

Chapter 2

POMDP-Based Throughput Maximization for Cooperative Communications Networks with Energy-Constrained Relay under Attack in the Physical Layer

2.1 Introduction

Cooperative communications are used to effectively improve the quality of a wireless network. The reliability and capacity of wireless communications are substantially increased by deploying the cooperative communication technique. In a cooperative communication system, each user can directly transmit data and collaborate with other users (i.e. relays) to transmit its data to a destination for enhancing the quality of transmissions [5]. In this case, an intermediate relay is used to support the transmissions between the source and the destination. Cooperative communications can offer remarkable advantages for wireless networks such as high energy efficiency and extended network lifetime [6, 7]. Recently, several studies have showed that cooperative communications can help to enhance the capacity and reliability of the wireless networks [8–10].

The physical layer is the lowest layer in the Open Systems Interconnection (OSI)

model. It can be used to verify the physical properties of a transmission in the network. However, the broadcast nature of wireless communications leaves the physical layer vulnerable to threats, e.g. eavesdropping, node tampering, hardware hacking, and jamming attacks [11]. In an eavesdropping attack, the eavesdropper can overhear the confidential information of legitimate users and occupy the data in the transmission area of a node. In node tampering, the attacker can replace the entire physical node or part of the node. Hardware hacking can damage nodes via malicious entities such that the nodes can lose their expected functionality, leaving them vulnerable to other risks. In jamming attacks, a jammer attempts to prevent users from accessing wireless network resources and reduces network availability by generating interference signals on the channels. This exhausts the energy of the nodes in the network [12, 13].

A cooperative communications network is particularly vulnerable to malicious attacks in the physical layer. Moreover, jamming is one of the more serious attacks that greatly degrade network performance. In order to tackle the jamming attacks, frequency hopping spread spectrum and direct sequence spread spectrum are widely utilized [14]. However, the same sequence can be used by the jammer to attack its target if the hopping sequence is exposed. Thus, the random rendezvous [15] and the uncoordinated frequency hopping [16] are used to safely share the hopping sequence. Nevertheless, these techniques result in the time wastes for the communications. Therefore, other secret sharing protocols are proposed such as public key cryptography, certificate and authentication protocol but they cause the large overheads and computational [17]. Desmedt [18] proposed an efficient coding method that provides protection against malicious users. Popper et al. [19] applied the uncoordinated spread spectrum (USS) techniques to prevent jamming of the communications between transmitter and receiver. USS achieved effective anti-jamming by discarding the require secrets before sharing, at the expense of a decreased communications throughput. However, USS techniques require the complex frequency synthesizers. Chorti derived optimal power allocation policies for transmitter and receiver pairs, where the active jammer is formulated as a one-shot zero-sum game for anti-jamming in secret key-generation systems [20]. Almost all the previous works on anti-jamming focus on how to design physical layer technologies (e.g. spread spectrum) [14, 16, 19, 21–23]. If the signals are widely spread, it will become harder for the jammer to interrupt the transmission link; meanwhile, the complexity in spread spectrum technique may be enlarged and thus it is not easy to deploy in the reality.

Recently, various jammer localization schemes have been proposed in wireless communications [24, 25]. The authors in [24] use spatial information as the basis for detecting the attacks to verify the number of attackers and localize the ambient adversaries. In [25], the authors consider the multiple jammers scenario in which multiple jammers can attack the network at the same time to achieve a better jamming effect. They propose the jamming detection scheme by developing x-rayed jammed-area localization algorithm. However, measurement data collection and position information sharing will bring more challenges. In [26], the authors investigate the multi-hop multi-channel cognitive radio network in the presence of multiple jammers. To deal with the energy-constrained problem, two novel algorithms are proposed to maximize the energy efficiency of data transmission from the source to the destination under the jamming attacks. Although the simulation results can verify the effectiveness of the proposed scheme, the applicable metrics may be limited due to the high overheads and algorithm complexity when the number of intermediate relays in the network is large. The digital feedback scheme is proposed to improve the speed of transceivers and it is also proved to be robust to noise in the feedback channel [27]. In order to deal with the energy-constrained issue and resource scarcity, the authors developed the joint controller and the related supporting access protocol to maximize both the energy and bandwidth efficiency of the vehicular access network, which is guaranteed to be reliable and safe in the wireless communication [28].

Recently, energy harvesting has become one of the appealing techniques for solving the energy-constraint problem in wireless networks. Practically, user equipment units (UEs) often are equipped with a limited-capacity battery. That results in degradation of network performance due to limited energy for the operation. Thus, energy harvesting technique can provide permanent energy for the battery without any physical replacements. Fortunately, UEs can harvest energy from non-radio frequency (RF) signals (e.g. solar, wind, heat, etc.) [29] or from RF signals [30, 31], which are available in ambient environments.

In this chapter, we consider the jamming attack scenario in cooperative communication system that consists of a source, a destination, a relay, and a jammer where the jammer intends to inject interference signals to block the transmission link (from the source to destination). The jammer in this chapter is assumed to always broadcast enough the interference (*power*) to block the communication in its communication range when it is activated. We define the jammer as “absolute jammer”. The behavior of the jammer is assumed to follow the Markov chain model.

In order to deal with the challenge of the jammer, we propose to use a relay to help the source to forward the data to the destination. However, the relay is assumed to be a small and movable device (for easy set up). Subsequently, the relay has a small battery that can allow the device to work in a limited time. The energy harvesting technique is applied to deal with energy-constrained problem at the relay; however, the limited energy arrival rate is also taken into account where the relay can harvest non-RF energy from the ambient environment with a limited amount of energy in each time slot [26, 29]. Moreover, the imperfection of the spectrum sensing mechanism [24, 25, 32] on the jamming detection at the source is also considered.

This chapter aims to maximize the long-term achievable throughput of the cooperative communication system in which the source will make the optimal decision on whether or not the source should cooperate with the relay to transmit the data to the destination securely with the purpose of degrading the jamming effect. We formulate the problem based on the POMDP framework [32] and propose a novel scheme to obtain the optimal policy such that the source can select the best action in every single time slot by considering long-term throughput maximization. The main contributions of this chapter are summarized as follows

- We investigate the throughput maximization of the cooperative communication system under the jamming attack, where an intermediate relay equipped with a limited-capacity battery is deployed to securely facilitate the transmission from the source to the destination. Meanwhile, due to the limited energy of the jammer, the jamming attack operation is assumed to follow the Markov chain model.
- We consider the imperfection of the spectrum sensing as well as the non-RF energy harvesting model at the relay for long-term operation, which greatly affects the network performance in practice.
- We propose a POMDP-based scheme at the source node to determine the optimal policy in the cooperation with the relay to improve the long-term achievable throughput of the network in the presence of the jamming attack. As a result, according to the optimal policy, the optimal action in every single time slot operation can be obtained using the proposed scheme.
- We evaluate the performance of the proposed scheme in comparison with traditional schemes such as *Myopic* and *Direct Link Only* schemes via Matlab simulation under

various network conditions. The numerical results are given to show the superior of the proposed scheme as compared with others according to the various network parameters of the cooperative communication network.

The remainder of the chapter is organized as follows. In Section 2.2, we introduce the system model and the Markov chain model of the jammer. In Section 2.3, we describe the optimal policy for direct-transmission and relay-assisted transmission modes. Section 2.4 evaluates our proposed scheme via simulation results. Finally, we conclude this paper in Section 2.5.

Table 2.1: The notation list

Symbol	Description
h_{SD}, h_{SR}, h_{RD}	Channel coefficients between S and D , S and R , and R and D
E_{ca}	Total capacity of the battery
E_{tr}	Transmission energy
$E_h(t)$	Harvested energy at the relay in the timeslot t^{th}
e_{mean}^h	Mean value of the harvested energy
$P_{E_h}(t)$	The probability of the harvested energy in time slot t^{th}
T	Total time frame
τ_s	The sensing time
$\tau_{SD}, \tau_{SR}, \tau_{RD}$	Transmission time between S and D , S and R , and R and D , respectively
P_S, P_R	Transmission power at the source and the receiver
x_s	The signal transmitted from the source
β_r	Amplify scale factor
γ	SNR of the channel between the source and the jammer
α	Discount factor
P_d, P_f	Probability of detection and false alarm

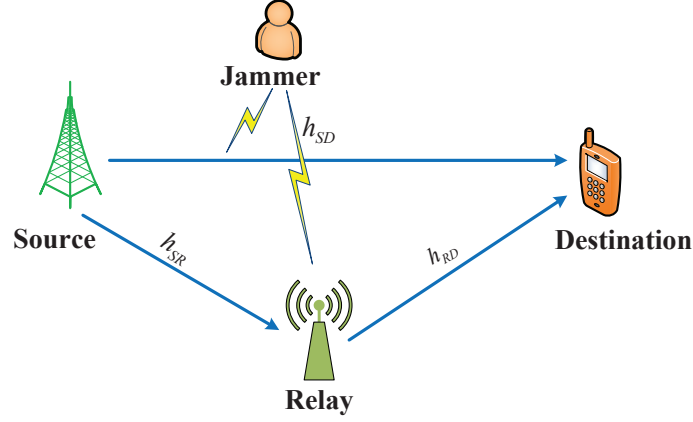


Figure 2.1: The system model of the proposed scheme.

2.2 System Model

As shown in Fig. 2.1, we consider a cooperative communication network consisting of a source, S , a destination, D , a relay, R , and a jammer, J . The source, destination, and jammer are assumed to have a fixed power supply such that they always have enough energy for transmission, reception, and jamming. By using the cooperative technique, S can cooperate with R to maximize the achievable throughput in the presence of the attack performed by J . The network is assumed to follow a synchronous, time-slotted model with time slot duration T . The channel coefficients between S and D , S and R , and R and D are denoted by h_{SD} , h_{SR} and h_{RD} , respectively.

In this chapter, we consider the “absolute jammer” who always has enough energy to transmit the interference signals to destroy the channel in target transmission link in a whole time slot duration. Therefore, when the jammer attacks the channel, the direct transmission link from the source to the destination will be blocked; and thus, D can not receive the data transmitted from S . Fortunately, R can help the source to forward its data to the destination in this case. However, the relay is assumed to have a limited-capacity battery without any fix powered supplies. Hence, the relay needs to harvest non-RF energy to maintain its long- term operation. The relay is assumed to scavenge energy during a whole time slot T and the harvested energy is stored in a battery with a finite capacity, E_{ca} . Therefore, at time slot t^{th} , the amount of energy $E_h(t)$ (energy units) that is harvested by the relay can be expressed as $E_h(t) \in \{e_1^h, e_2^h, e_3^h, \dots, e_\nu^h\}$, where $e_1^h, e_2^h, e_3^h, \dots, e_\nu^h$ are harvested energy levels, and $0 < e_1^h < e_2^h < e_3^h < \dots < e_\nu^h < E_{ca}$.

The probability mass function (PMF) of the harvested energy is given as:

$$P_{E_h}(k) = \Pr [E_h(t) = e_k^h], k = 1, 2, 3, \dots, \nu. \quad (2.1)$$

where $P_{E_h}(t)$ is the probability of the harvested energy in time slot t^{th} . In this chapter, we assume that the harvested energy of the relay follows a Poisson distribution where $E_h(t)$ is a Poisson random variable with mean value e_{mean}^h , and the PMF in (1) can be rewritten as follows:

$$P_{E_h}(k) \approx \frac{e^{-e_{mean}^h} (e_{mean}^h)^k}{k!}, k = 1, 2, \dots, \nu. \quad (2.2)$$

In order to deal with the jamming attack problem, at the beginning of each time slot, the source needs to determine whether it should use the direct transmission mode or relay-assisted transmission mode to transmit its data to the destination.

Fig. 2.2a depicts the time frame structure of the direct transmission mode. For this mode, the frame is divided into two phases: sensing and data transmission. In the sensing phase, the source performs spectrum sensing to detect jamming signals infected by the jammer. In the data transmission phase, the source transmits the data to the destination without any help from the relay.

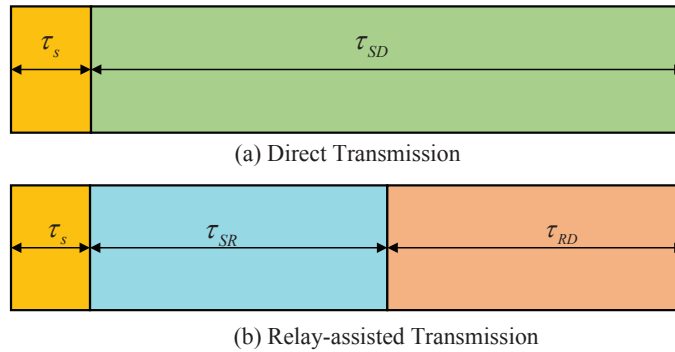


Figure 2.2: Frame structure.

Fig. 2.2b illustrates the time frame structure of the relay-assisted transmission mode. For this mode, the frame is divided into three phases: sensing, data transmission from the source to the destination ($S - R$), and data forwarding from the relay to the destination ($R - D$). Unlike direct transmission, after the sensing phase, the source will transmit the

data to the relay and then, the relay will forward the data to the destination. τ_s represents the sensing time of the source. τ_{SD}, τ_{SR} , and τ_{RD} represent the data transmission times of the links: S to D , S to R , and R to D , respectively. Note that the duration of sensing phase in relay-assisted transmission mode is the same as direct transmission mode meanwhile the duration $\tau_{SR} = \tau_{RD} = \frac{1}{2}\tau_{SD}$. For the direct transmission mode, the received signal of the destination at the end of a time slot can be expressed as follows

$$y_D = \sqrt{P_S}h_{SD}x_s + n_D. \quad (2.3)$$

where P_S is the transmission power at the source; x_s represents the signal transmitted from the source, h_{SD} represents the channel coefficient between S and D , n_D is white Gaussian noise (AWGN) with zero- mean and variance σ^2 at the destination. For the relay-assisted transmission mode, the received signals of the relay after the S-R phase can be expressed as follows

$$y_R = \sqrt{P_S}h_{SR}x_s + n_R. \quad (2.4)$$

where h_{SR} denotes the channel coefficient between S and R , n_R denotes the white Gaussian noise with zero- mean and variance σ^2 at the relay. This chapter adopts an amplify-and-forward (AF) relaying protocol to forward data to the destination. Hence, the relay amplifies the signal by using a scale factor, β_r , which can be calculated as follows:

$$\beta_r = \frac{\sqrt{P_R}}{\sqrt{P_S|h_{SR}|^2 + \sigma^2 + \sigma_J^2}}. \quad (2.5)$$

where P_R represents the transmission power at the relay, σ_J^2 is the noise variance that is created by the jammer.

In phase 2, the received signal at the destination is given by

$$\begin{aligned} y_D &= \beta_r h_{RD} y_R + n_D \\ &= \beta_r h_{RD} (\sqrt{P_S} h_{SR} x_s + n_R + n_J) + n_D \\ &= \beta_r h_{RD} \sqrt{P_S} h_{SR} x_s + \beta_r h_{RD} (n_R + n_J) + n_D. \end{aligned} \quad (2.6)$$

where n_J denotes jamming noise with zero-mean and variance σ_J^2 . Note that in the case that the jammer does not attack, the n_J will be zero. As a result, the signal-to-interference-plus-noise ratio (SINR) at D is denoted as φ_0 , φ_1 , and φ_2 for the direct transmission without jamming, relay-assisted transmission with jamming and relay-assisted transmission without jamming, respectively, obtained as follows:

$$\varphi_0 = \frac{P_S |h_{SD}|^2}{\sigma^2}, \quad (2.7)$$

$$\begin{aligned} \varphi_1 &= \frac{P_S P_R |h_{SR}|^2 |h_{RD}|^2}{(P_S |h_{SR}|^2 + \sigma^2 + \sigma_J^2) \left(\frac{P_R |h_{RD}|^2 (\sigma^2 + \sigma_J^2)}{P_S |h_{SR}|^2 + \sigma^2 + \sigma_J^2} + \sigma^2 \right)}, \\ &= \frac{P_S P_R |h_{SR}|^2 |h_{RD}|^2}{P_R |h_{RD}|^2 (\sigma^2 + \sigma_J^2) + (P_S |h_{SR}|^2 + \sigma^2 + \sigma_J^2) \sigma^2} \end{aligned}, \quad (2.8)$$

$$\begin{aligned} \varphi_2 &= \frac{P_S P_R |h_{SR}|^2 |h_{RD}|^2}{(P_S |h_{SR}|^2 + \sigma^2) \left(\frac{P_R |h_{RD}|^2 \sigma^2}{P_S |h_{SR}|^2 + \sigma^2} + \sigma^2 \right)}, \\ &= \frac{P_S P_R |h_{SR}|^2 |h_{RD}|^2}{(P_R |h_{RD}|^2 + P_S |h_{SR}|^2 + \sigma^2) \sigma^2} \end{aligned}. \quad (2.9)$$

According to the transmission mode, the average throughput can be calculated as

$$R = \begin{cases} \frac{T - \tau_s}{T} C_0 (1 - P_f) \Pr(\bar{J}) & (2.10a) \\ \frac{T - \tau_s - \tau_{SR}}{T} C_1 P_d \Pr(J) & (2.10b) \\ \frac{T - \tau_s - \tau_{SR}}{T} C_2 P_f \Pr(\bar{J}) & (2.10c) \end{cases}, \quad (2.10)$$

where

$$C_0 = \log_2(1 + \varphi_0), \quad (2.11)$$

$$C_1 = \log_2(1 + \varphi_1), \quad (2.12)$$

$$C_2 = \log_2(1 + \varphi_2). \quad (2.13)$$

P_f and P_d are the probability of false alarm and the probability of detection of the sensing mechanism, respectively, according to sensing time duration τ_s . $\Pr(\bar{J})$ and $\Pr(J)$ denote

the probability of no jamming and the probability of jamming in the network, respectively. C_0 , C_1 , and C_2 represent achievable throughput at the destination under the different cases in (2.10), i.e. direct transmission without jamming (2.10a), relay-assisted transmission with jamming (2.10b), and relay-assisted transmission without jamming (2.10c).

The probability of detection and false alarm can be estimated as follows [32]:

$$P_d = Q \left(\frac{\vartheta - M \times (\gamma + 1)}{\sqrt{2M \times (2\gamma + 1)}} \right), \quad (2.14)$$

and

$$P_f = Q \left(\frac{\vartheta - M}{\sqrt{2M}} \right), \quad (2.15)$$

where ϑ denotes energy threshold, M represents for number of sensing samples and can be calculated as $M = 2\tau_s f_s$ (f_s is sensing bandwidth), γ is signal-to-noise ratio (SNR) of the sensing channel (i.e., the channel between source and jammer). There are some available researches that propose methods to estimate the SNR value. Therefore, in this chapter we assume that the value is available at the source. The probability of false alarm also can be achieved by

$$P_f = Q \left(\sqrt{2\gamma + 1} Q^{-1}(P_d) + \sqrt{\tau_s f_s \gamma} \right). \quad (2.16)$$

In this chapter, we assume that the states of the jammer follow a Markov chain model. The states of the jammer changes between the two states, presence (J) and absence (\bar{J}), shown in Fig. 2.3. The transition probabilities of the jammer from state J to state \bar{J} and from state J to itself are denoted as $P_{J\bar{J}}$ and P_{JJ} , respectively [33].

We assume that the source always has a data packet to transmit to the destination. At the beginning of a time slot, the information about remaining energy of the relay (e^{re} , $0 \leq e^{re} \leq E_{ca}$) is assumed to be available at the source.

Fig. 2.4 shows the operation process of the system. First of all, the source performs sensing to identify the states (“presence” or “absence”) of the jammer. If the sensing engine provides the result “absence”, i.e. there is no jamming signal in the current time slot (not always true due to the imperfect sensing), the source will trust the result and then

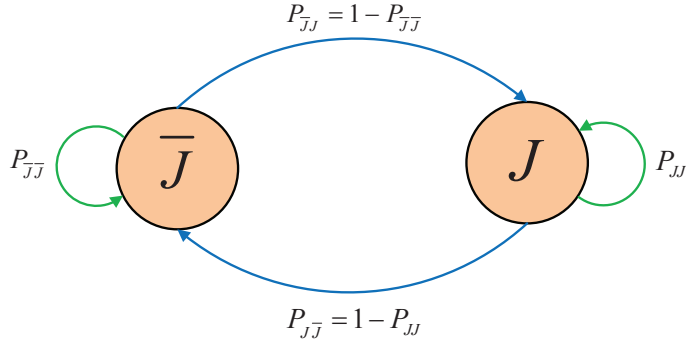


Figure 2.3: A Markov chain model of the jammer.

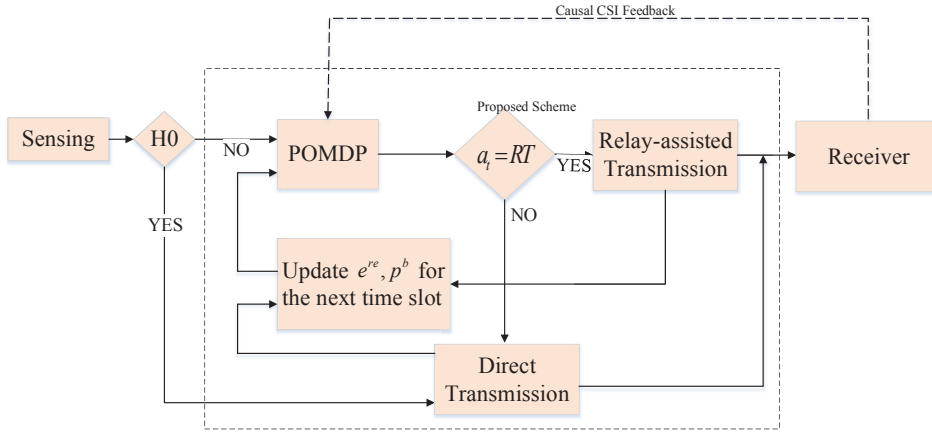


Figure 2.4: Flowchart of the proposed scheme.

transmits its data directly to the destination. If the source receives an acknowledge (ACK) message after the transmission phase, then the reward is calculated as

$$R = \frac{T - \tau_s}{T} C_0. \quad (2.17)$$

The belief probability p_{t+1}^b , which represents the probability of the jammer being present in the next time slot, will be updated as

$$p_{t+1}^b = P_{\bar{J}J}. \quad (2.18)$$

The remaining energy in the battery of the relay can be updated as

$$e_{t+1}^{re} = \min\{e_t^{re} + E_h(t), E_{ca}\}. \quad (2.19)$$

where $E_h(t)$ is the amount of harvested energy of the relay in time slot t^{th} .

If the source does not receive an ACK (or receive NACK) after the transmission phase, the reward will be zero (i.e. $R = 0$). Besides that, the updated belief in the next time slot, p_{t+1}^b , can be calculated as

$$p_{t+1}^b = P_{JJ}. \quad (2.20)$$

The transition probability is given as

$$\Pr(e_t^{re} \rightarrow e_{t+1}^{re}) = \Pr[E_h(t) = e_k^h]. \quad (2.21)$$

If the result obtained from the sensing engine is “*presence*”, then the proposed scheme, based on a partially observable Markov decision process (POMDP), will be applied to select the optimal action (i.e. either performs the direct transmission mode or relay-assisted transmission mode). The proposed scheme will be presented in more detail in the next section.

2.3 Optimal Mode Decision Policy Based on POMDP

In this scheme, we apply POMDP to obtain an optimal mode decision policy to maximize the throughput in a cooperative communications network in the presence of the jamming attack. In this system, there are two operation modes for the source: direct transmission (DT) and relay-assisted transmission (RT), $a_t = \{RT, DT\}$.

In direct transmission mode, the relay will not assist the source to forward the data to the destination (i.e. the relay is inactive for this case). That means the destination will receive the data transmitted directly from the source. In the relay-assisted transmission mode, the relay will help the source to forward the data to the destination (i.e. the relay is active for this case). Due to the energy constrained problem in the relay as well as the imperfect spectrum sensing, the source will consider the long-term reward to efficiently cooperate with the relay to optimize the network performance.

In order to formulate the framework of POMDP, we define the state space of the system as $s = \{e_t^{re}, p_t^b\}$ where e_t^{re} and p_t^b are the remaining energy of the relay and the probability of the presence of the jammer in time slot t^{th} , respectively. Value function $V_{(e^{re}, p^b)}$ represents the maximum total discounted throughput of the system, which is given by

$$V_{(e^{re}, p^b)} = \max_{a_k} \left(\sum_{t=k}^{\infty} \alpha^{t-k} R(e_t^{re}, p_t^b, a_t) \mid e_k^{re} = e^{re}, p_k^b = p^b \right). \quad (2.22)$$

where $0 < \alpha < 1$ denotes the discount factor and it is chosen to adjust the impact of future action to current action. More specifically, if the value of alpha is large, the reward of the future action will more affect to the reward of the current action and vice versa, $R(e_t^{re}, p_t^b, a_t)$ is the achieved throughput of system in time slot t^{th} when action a_t is performed at the state $s = \{e_t^{re}, p_t^b\}$.

2.3.1 Relay-assisted Transmission Mode

In the relay-assisted transmission mode, the relay will help the source forward data to the destination. In this mode, the destination can always receive data packet, so it will be difficult to distinguish the presence of jamming. From the received signals at the destination, we can realize whether the jammer actually attacks the channel or not. That is because the signal strength from received data packet when the jammer attacks will become stronger than a normal received data packet (i.e. without jamming). Therefore, the destination can recognize whether the original signal contains the jamming signal or not, in terms of the predefined jamming threshold χ_{jam} such as

$$J = \begin{cases} \text{Presence,} & \text{if } P_D \geq \chi_{jam}; \\ \text{Absence,} & \text{otherwise.} \end{cases} \quad (2.23)$$

where P_D is the received signal energy at the destination.

Observation 1 (Φ_1): The sensing result indicates the presence of the jammer, the source transmits data packet via the relay and the jammer actually attacks the channel. In this case, jammer attack is well detected, and corresponding achieved throughput is given by

$$R(e_t^{re}, p_t^b \mid \Phi_1) = \frac{T - \tau_s - \tau_{SR}}{T} C_1. \quad (2.24)$$

The probability that case Φ_1 happens can be calculated as

$$\Pr(\Phi_1) = \frac{p_t^b P_d}{p_t^b P_d + (1 - p_t^b) P_f}. \quad (2.25)$$

The updated belief for the next time slot is computed as

$$p_{t+1}^b = P_{JJ}. \quad (2.26)$$

The updated remaining energy for the next time slot is

$$e_{t+1}^{re} = \min\{e_t^{re} + E_h(t) - E_{tr}, E_{ca}\}. \quad (2.27)$$

where E_{tr} is the required energy for transmission from the relay to the destination.

The transition probability can be calculated as

$$\Pr(e_t^{re} \rightarrow e_{t+1}^{re} | \Phi_1) = \Pr[E_h(t) = e_k^h] \Pr(\Phi_1). \quad (2.28)$$

Observation 2 (Φ_2): In this case, the sensing result indicates the presence of the jammer, the source transmits data via the relay, and the jammer actually does not attack the channel. Hence, we recognize the false alarm happens in this case. The achieved throughput is given as

$$R(e_t^{re}, p_t^b | \Phi_2) = \frac{T - \tau_s - \tau_{SR}}{T} C_2. \quad (2.29)$$

The updated belief for the next time slot is given as

$$p_{t+1}^b = P_{JJ}. \quad (2.30)$$

The remaining energy for the next time slot can be updated as

$$e_{t+1}^{re} = \min\{e_t^{re} + E_h(t) - E_{tr}, E_{ca}\}. \quad (2.31)$$

The probability that case Φ_2 happens is

$$\Pr(\Phi_2) = \frac{(1 - p_t^b)P_f}{p_t^b P_d + (1 - p_t^b)P_f}. \quad (2.32)$$

The transition probability if the case Φ_2 happens is

$$\Pr(e_t^{re} \rightarrow e_{t+1}^{re} | \Phi_2) = \Pr[E_h(t) = e_k^h] \Pr(\Phi_2). \quad (2.33)$$

2.3.2 Direct Transmission Mode

In direct transmission mode, the source directly transmits the data to the destination (without a help from the relay). According to whether the source receives ACK from destination, the following two observations can be described as follows.

Observation 3 (Φ_3): The source transmits the data directly to the destination and receives an ACK. The achieved throughput is given by

$$R(e_t^{re}, p_t^b | \Phi_3) = \frac{T - \tau_s}{T} C_0. \quad (2.34)$$

The probability that the case Φ_3 occurs is computed as

$$\Pr(\Phi_3) = \frac{(1 - p_t^b) P_f}{p_t^b P_d + (1 - p_t^b) P_f}. \quad (2.35)$$

The belief that the jammer will be present in the next time slot can be updated as

$$p_{t+1}^b = P_{JJ}. \quad (2.36)$$

In this case, although the relay does not receive and forward data to the destination, it still harvests energy for future use. The updated remaining energy of the relay for the next time slot is

$$e_{t+1}^{re} = \min(e_t^{re} + E_h(t), E_{ca}). \quad (2.37)$$

The transition probability that the case Φ_3 happens is computed as

$$\Pr(e_t^{re} \rightarrow e_{t+1}^{re} | \Phi_3) = \Pr[E_h(t) = e_k^h] \Pr(\Phi_3), \quad (2.38)$$

Observation 4 (Φ_4): The source transmits data directly to the destination, but it does not receive an ACK. In this case, there is no achieved throughput such that we have $R(e_t^{re}, p_t^b | \Phi_4) = 0$. The probability that case Φ_4 happens can be calculated as follows:

$$\Pr(\Phi_4) = \frac{p_t^b P_d}{p_t^b P_d + (1 - p_t^b) P_f}. \quad (2.39)$$

The updated belief for the next time slot can be given as

$$p_{t+1}^b = P_{JJ}. \quad (2.40)$$

The remaining energy in the relay is updated in the same way as Eq. (2.37) under Φ_3 . The transition probability that case Φ_4 occurs is given by

$$\Pr(e_t^{re} \rightarrow e_{t+1}^{re} | \Phi_4) = \Pr[E_h(t) = e_k^h] \Pr(\Phi_4), \quad (2.41)$$

for $k = 1, 2, 3, \dots, \nu$.

Based on these observations, we can calculate the expected value function, and further we can find the optimal operation mode, a_k . Therefore, the value function in (22) can be rewritten as follows

$$V_{(e^{re}, p^b)} = \max_{a_k} \left\{ \begin{array}{l} \sum_{t=k}^{\infty} \alpha^{t-k} \sum_{\Phi_i \in a_t} \Pr(\Phi_i) \\ \sum_{e_{t+1}} \Pr(e_t^{re} \rightarrow e_{t+1}^{re} | \Phi_i) \\ R(e_t^{re}, p_t^b, a_t | \Phi_i) | e_k^{re} = e^{re}, p_k^b = p^b \end{array} \right\} \quad (2.42)$$

In order to solve problem in Eq. (2.42), a numerical method is used [34]. The solution to the problem provides the optimal policy of the system. The complexity of algorithm can be analyzed based on the amount of computation space such as number of states, actions, transition probabilities and observations. Based on the Bellman's equation,

Table 2.2: Simulation Parameters

Symbol	Description	Value
γ	SNR of the channel between the source and the jammer	-10 dB
T	Total time frame	0.02 s
f_s	Sensing bandwidth	0.2×10^6 Hz
P_d	Probability of detection	0.8
P_{H_1}	Initial belief that the jammer is present	0.7
P_{JJ}	Transition probability from state J to itself	0.6
$P_{\bar{J}J}$	Transition probability from state \bar{J} to state J	0.8
E_{ca}	Total capacity of the battery	10
e_{mean}^h	Mean value of the harvested energy	2
E_{tr}	Transmission energy	6
α	Discount factor	0.9

the optimal policy is chosen by solving the value function using iteration-based dynamic programming. Let us denote \mathbb{Z} , \mathbb{S} be the action set and the possible state set at the beginning of each time slot, respectively. The algorithm complexity can be defined according to the action and state space of the system. In the POMDP, the agent has to control the process at each time step to maximize the long-term reward. Therefore, the number of $O(|\mathbb{Z}| |\mathbb{S}|^2)$ operations is required in each iteration to calculate total number of the transition probabilities from one state $s(t)$ to other state $s'(t)$ after performing an action $a(t)$.

2.4 Simulation Results

In order to evaluate the effectiveness of the proposed scheme, we implemented a simulation using MATLAB. In this section, we present the performance comparisons among the proposed scheme, the *Myopic* scheme and the *Direct Link Only* scheme. In the *Myopic* scheme, we only considered the throughput for the current time slot to select optimal action. In the *Direct Link Only* scheme, the source always uses the direct link to transmit data

packet to the destination. The parameters used for our simulation are shown in **Table 2.2**.

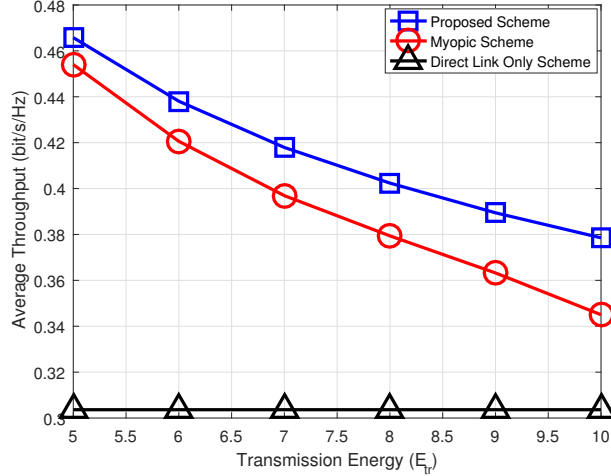


Figure 2.5: Average throughput versus required transmitted energy.

In Fig. 2.5 shows the average throughput of the system according to the required transmission energy of relay node. According to the Fig. 2.5, the throughput of the system decreases as the required transmission energy increases. The reason is as following: For a large amount of the required transmission energy, the source will have fewer opportunities to transmit the data packet via relay when the jammer appears since the relay lacks energy for the forwarding process. It is obvious that the increase in the required transmission energy does not affect the average throughput of the *Direct Link Only* scheme. The figure verifies that the proposed scheme outperforms the *Myopic* scheme and the *Direct Link Only* scheme.

Fig. 2.6 shows the relation between average throughput and battery capacity in the relay. We can see that the average throughput of the system increases as the battery capacity of the relay increases. The reason is why the relay has more energy to assist communication between the source and destination. On the other hand, the battery capacity of the relay does not affect the *Direct Link Only* scheme, and corresponding throughput is not changed. The figure shows that the proposed scheme can provide higher throughput than the *Myopic* and the *Direct Link Only* schemes.

Fig. 2.7 shows average throughput of the proposed scheme according to the battery capacity of relay node for different values of detection probability P_d . It is observed that average throughput of the proposed scheme increased as the battery capacity of relay node increases for a fixed value of P_d . However, the average throughput of the proposed scheme

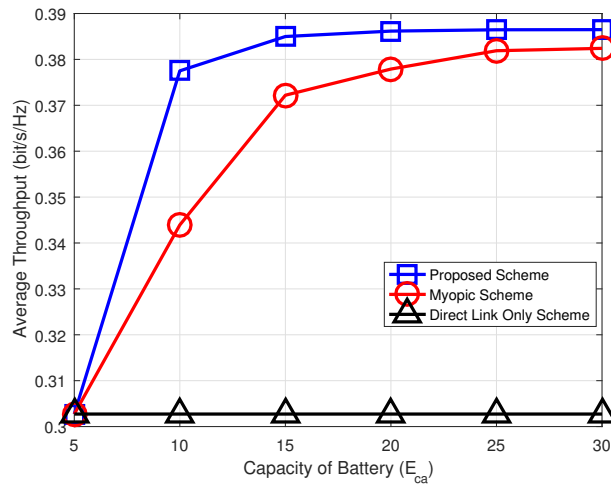


Figure 2.6: Average throughput versus capacity of battery.

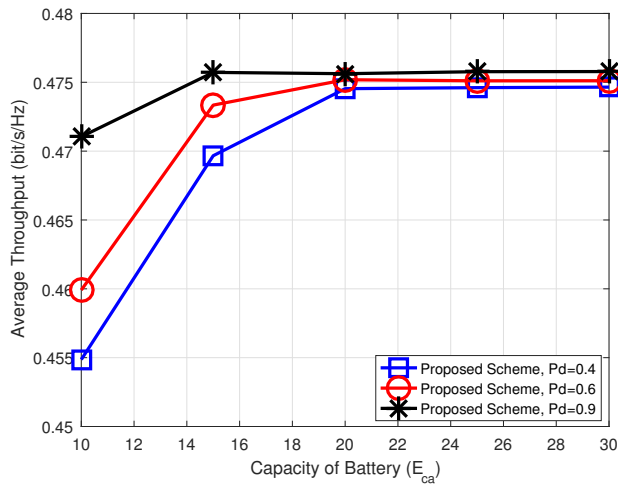


Figure 2.7: Average throughput versus capacity of the battery when detection probability $P_d = 0.4, 0.6,$ and 0.9 .

goes into a saturation mode for a certain value of the battery capacity. That is, as the battery capacity reaches a certain value, the average throughput of the proposed scheme cannot be enhanced. Fig. 2.7 also shows that more detection probability of jammer, the more the average throughput. To do this, however, we need more accurate sensing scheme at the source node.

Finally, Fig. 2.8 shows the average throughput of the proposed scheme according

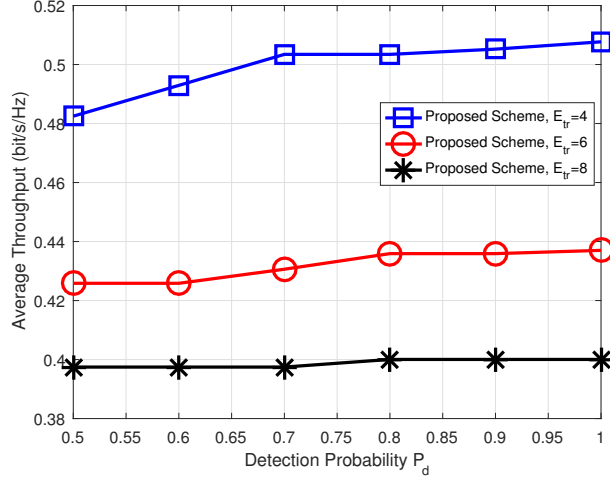


Figure 2.8: Average throughput versus detection probability P_d when transmission energy $E_{tr} = 4, 6,$ and 8 .

to the detection probability of jammer for different values of the required transmission energy of the relay node. As like the previous observation, the more required transmission energy of the relay node, the less the average throughput. For a fixed value of the required transmission energy of the relay node, the average throughput of the proposed scheme is improved as the detection probability of jammer increases.

2.5 Conclusion

In this chapter, we investigated the average throughput maximization of cooperative communications networks when under a jamming attack. In addition, the energy-constraint problem was taken into account. We proposed a POMDP-based scheme to achieve the optimal mode decision policy to maximize the long-term throughput by taking into account future reward. Simulation results confirmed that the proposed scheme can improve the overall throughput in cooperative communications networks and outperforms a *Myopic* scheme and *Direct Link Only* scheme under the jamming attack. In the future, we would like to investigate the joint relay and channel selection scheme in multiple relays and multiple channels to enhance the overall throughput of the network. Moreover, multiple jammers should be considered and an actor-critic-based scheme should be studied to determine the optimal policy in the cooperation communications networks with energy-constrained relay

under multiple jammers. Therefore, it is the key challenge in the future to find the optimal solution for the source to select the best relay and channel in the multiple channels, relays and jammers.

Chapter 3

A POMDP-based Long-Term Transmission Rate Maximization for Cognitive Radio Networks with Wireless-Powered Ambient Backscatter

3.1 Introduction

Cognitive radio networks (CRNs) are intelligent radio networks where a secondary user (SU) shares the available spectrum of the primary user (PU) because of the spectrum scarcity and the increasing growth in wireless communications demand [35–39]. In recent work, wireless communications powered by external harvested energy has become a promising technique to deal with the energy-constrained problem. In addition, radio frequency energy-harvesting systems developed along with a new network generation have been introduced, which are radio frequency powered CRNs [40–43].

RF-harvested energy in a CRN is a key solution for energy-constrained issues in wireless networking [44–46]. In an RF-powered CRN, wireless devices can harvest energy from ambient RF signals and use it for their operations. The harvested energy is stored in the battery of the SU without manually changing or recharging the battery. Recently,

rectifying antenna design has become highly efficient, and will become more efficient at harvesting energy from RF signals in the future [47, 48]. Thus, more and more researchers are concerned with investigating wireless-powered communications networks [49–51]. Along with RF energy harvesting, non-RF energy sources (solar, wind, etc.) can provide perpetual energy for rechargeable batteries of wireless users [29, 52].

In an underlay RF-powered CRN, the secondary transmitter (ST) opportunistically harvests energy from ambient primary signals to replenish its battery, and then transmits information to the secondary receiver (SR) such that interference does not significantly affect the primary user when the channel is busy [53]. Unlike the underlay RF-powered CRN, for the overlay RF-powered CRN protocol, the ST can harvest RF energy when the primary channel is busy, and transmits data only when the channel is free in the next frame time [54]. Nevertheless, secondary network performance depends on the activity of the PU on the primary channel and the total amount of harvested energy, which can degrade the achievable data transmission rate of secondary systems. For example, the issue occurs in the case of imperfect characteristics in the spectrum sensing mechanism, from poor quality of the energy harvesting circuit, or if the channel's idle period in the overlay CRN is too short. This leads to few transmitted bits at the secondary receiver due to the small amount of harvested energy from ambient RF signals. Therefore, there is growing interest in finding alternative solutions to overcome this limitation and to enhance secondary system performance [41, 55].

Stockman first introduced modulated backscatter technology [56], and it rapidly became a promising technique for advanced low-power wireless communication systems. In modulated backscatter systems, a backscatter transmitter is able to modulate and reflect the received RF signals to transmit its own data instead of producing RF signals by itself [57–59]. Nowadays, several useful applications have been integrated with the backscattering technique in practice such as radio-frequency identification (RFID), remote switches, tracking devices, and low-cost sensor networks [58, 60]. However, conventional backscatter communications may not be easily carried out for data-intensive communications systems due to some limitations [61]. For example, the distance between the backscatter transmitter and an RF source is limited to a short range; thereby, it results in limitations on the coverage of user communications. Thus, application of the conventional backscattering technique is still restricted in practical wide-range communications scenarios.

Backscatter communications is categorized into three main classes: mono-static backscatter communications systems (MBCSs), bistatic backscatter communications systems

(BBCSs), and ambient backscatter communications systems (ABCSs) [62]. In mono-static backscatter communication systems, a backscatter receiver that also acts as an RF source will emit RF signals for backscatter communications to activate the backscatter transmitter (e.g. an RFID tag). Because the RF source and backscatter receiver are placed in the same device, this leads to the issue of round-trip path loss from the modulated signals [63, 64]. In order to avoid round-trip path loss, the carrier emitter and backscatter receiver are separated in bi-static backscatter communications systems. As such, the performance of a bi-static backscatter communications systems can be greatly improved by setting optimal locations for the carrier emitters [62, 64]. Unlike bi-static backscatter communications systems, the ambient backscatter communications systems utilize ambient radio frequency sources, such as TV towers, cellular base stations, Wi-Fi access points, and so forth, instead of using the normal dedicated carrier emitters used in bi-static backscatter communications systems [65, 66]. Moreover, they do not require deploying the dedicated RF sources, and thereby, cost and energy consumption can be significantly reduced.

Recently, ambient backscatter communications have attracted a lot of attention from researchers [67–69]. The reason is that the ambient backscatter technique enables a passive device to transmit its data to a receiver by using ambient RF signals without resorting to an energy supply. Moreover, it does not require high-powered RF to the backscatter device, compared with RFID systems. Consequently, ambient backscatter technology is considered a promising solution for some existing issues, such as the energy-constrained problem, inefficient spectrum usage, etc. Parks *et al.* [70] used multiple antennas to improve data backscatter performance and the communications range. As a result, their experiments demonstrated that the backscatter rate and the backscatter communications range can reach up to 1 Mbps and 20 m, respectively. Pérez-Penichet *et al.* investigated a new coding approach to maximize the data transmission rate of ambient backscatter communications [71], wherein multiple bits are encoded in a single symbol. Suboptimal signal detection and bit error rate analysis for ambient backscatter communications systems were studied [72]. Along with methods increasing data transmission rates, a secure data backscatter protocol against eavesdropping in the physical layer of ambient backscatter communication systems was proposed [73].

Useful applications for ambient backscatter technology with RF-powered technology have been investigated [55, 74, 75]. Hoang *et al.* proposed a solution to maximize the performance of secondary systems in RF-powered CRNs with ambient backscatter by finding

the optimal splitting time for harvesting and backscattering by the SU [55]. Kim *et al.* adopted hybrid operational modes for ambient backscatter and bistatic scatter to increase the transmission range, and proposed uniform rate distribution of RF-powered communications networks [74]. Similar to the work in [55], Park *et al.* investigated the optimal allocated time for energy harvesting, backscattering, and transmitting to maximize the secondary data transmission rate in RF-powered CRNs [75]. In [76], the authors proposed a novel hybrid harvest-then-transmit and backscatter communications scheme in a cognitive wireless powered communication network for the Internet of Things applications. The closed-form optimal solution for a single CU case and the optimal combination of working modes are derived to maximize the throughput of secondary communication systems. In [77], the optimal control policy for RF-Powered Backscatter communication is investigated to maximize the throughput of the network. More specific, the optimal trade-off between the sleep and active states and the optimal reflection coefficient are provided and demonstrated with the superiority through the numerical results.

Motivated by the aforementioned literature, this chapter proposes an energy-efficient transmission approach of a secondary transceiver in the PU activity proximity to improve the long-term transmission performance of the secondary system in wireless-powered CRNs using ambient backscatter communications. In this chapter, we consider both non-RF and RF energy harvesting to overcome the energy-constrained issue of wireless secondary users. Specifically, the secondary transmitter equipped with both non-RF and RF energy harvesters can opportunistically scavenge the energy from the ambient non-RF source (e.g. solar, wind, etc.) and RF source (primary transmitter power). The former (non-RF harvester) performs energy harvesting in every processing time slots while the later (RF harvester) is only implemented according to the presence of the primary transmitter. In other words, this chapter considers a wireless-powered CRN that enables the ST not only harvest the non-RF energy but also can perform an action 1) harvest the RF energy from RF sources 2) backscatter the data to the SR when the sensing results indicate the channel is busy; or 3) transmit data to the SR when the sensing results indicate the channel is free. The system performance is significantly affected by the factors such as the availability of the PU, imperfect sensing, uncontrollability of the wireless sources (transition probability of primary users, the distribution of non-RF sources, etc.). Therefore, this chapter investigates the long-term secondary system rate maximization in wireless-powered CRN using ambient backscatter technique, where the ST selects the optimal action following the infinite time

horizon in every single time slots. The optimal policy is obtained by adopting the POMDP framework, such that the ST can efficiently use energy to maximize the overall long-term reward.

The main contributions of this chapter are summarized as follows:

- This chapter adopts the ambient backscatter for the secondary transceiver communication in wireless-powered CRN. Both non-RF and RF energy harvesting are considered to deal with energy-constrained problem of the secondary transmitter.
- The novel self-sustainable communication scheme for secondary users investigated by combining ambient backscatter technology and wireless energy harvesting technology in CRNs. Particularly, powered by an RF and non-RF energy harvesting circuit, the secondary transmitter can simultaneously harvest non-RF energy from the ambient environment and perform backscattering/RF harvesting/transmitting for the secondary receiver in the data communication phase via the Rayleigh fading channel.
- To maximize the accumulative discounted reward of the secondary system in the infinite time horizon, the maximization problem formulation is presented according to the framework of a partially observable Markov decision process (POMDP). Especially, this chapter also takes the spectrum sensing imperfection of the secondary system into account. Accordingly, we provide a POMDP-based scheme in a time-slotted fashion for the secondary system to achieve the optimal policy such that the ST can dynamically and efficiently use its remaining energy in each time slot to maximize the long-term transmission rate.
- We assess the proposed scheme performance in comparison with those of other conventional schemes under various network parameter variations via Matlab simulation. The valuable insights into the effect of network parameter variation on the secondary system performance are given throughout the numerical results.

The rest of this chapter is organized as follows. In Section 3.2, we describe system model of the ambient backscattering-assisted wireless-powered CRN communications. In Section 3.3, we present the problem formulation, and further describe the proposed scheme to obtain the optimal-action policy in Section 3.4. The simulation results and discussion are given in Section 3.5. Finally, in Section 3.6, we conclude this work.

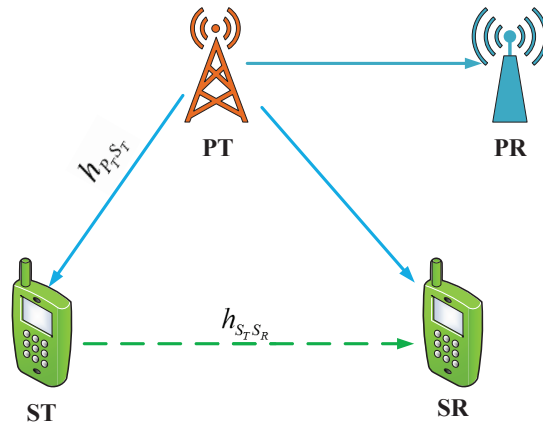


Figure 3.1: System model of the considered network.

3.2 System Description

3.2.1 Network Model

In this chapter, we consider a CRN consisting of a pair of primary transmitter (PT)-primary receiver (PR) and a pair of a secondary transmitter-secondary receiver. Fig. 3.1 shows the architecture of the ambient wireless-powered backscatter CRN where ST can harvest energy or backscatter data via the signals emitted from the PT when the PT transmits signals. Otherwise, the ST can transmit data using the energy stored in ST's battery when the PT does not perform its operations. As like Fig. 3.2, the ST is equipped with a non-RF energy harvesting circuit, an RF energy harvesting circuit, a backscattering circuit, and a controller. The controller decides the optimal operation mode (i.e. backscattering, RF energy harvesting, or data transmitting). The RF harvester, non-RF harvester, and ambient backscatter circuits can harvest RF and non-RF energy, and then stores it in the battery, which will be used to transmit data when the PT is absent, and backscatter the data to the SR, respectively. The primary channel is licensed for a pair of primary users, i.e. the PT and the PR. Meanwhile, in order to share the spectrum resource with primary users, the ST is assumed to opportunistically use the primary channel to send its messages to the SR. In particular, at the beginning of a time slot, the ST performs spectrum sensing to check if the primary channel is free or not (i.e. to check whether the PT is using the primary channel to transmit data to the PR or not). Subsequently, based

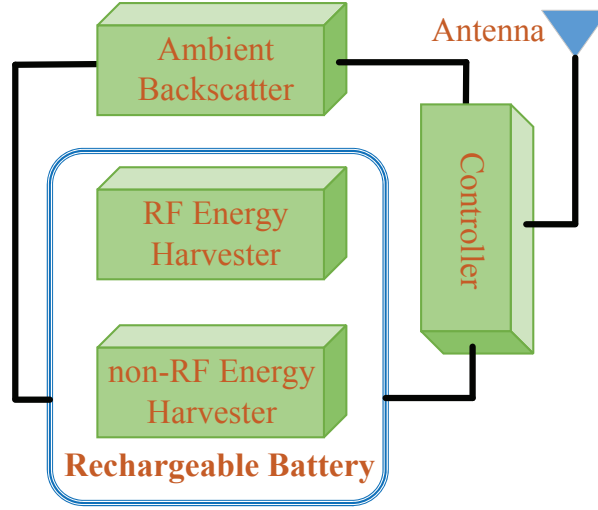


Figure 3.2: Schematic structure of secondary transmitter.

on the sensing results, the ST will choose whether it should backscatter the data to the ST, harvest RF energy from the transmission power of the PT, or transmit data to the SR. The operation of the secondary system is composed of three phases: the sensing phase, the decision phase, and the data transmission phase, which includes three modes: backscattering, RF energy harvesting, or data transmitting (as shown on Fig. 3.3).

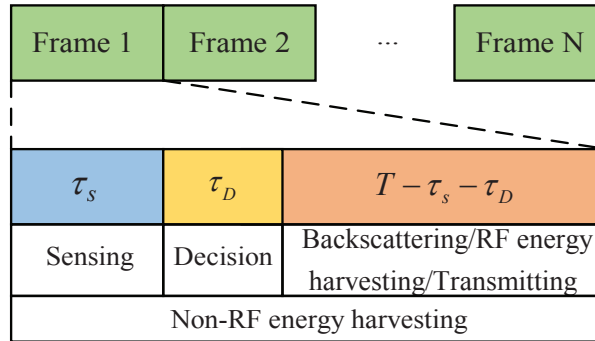


Figure 3.3: The time frame structure of the secondary user.

The state of the primary user is denoted as P or A , which respectively represent the presence or absence of the PU on the primary channel in a time slot. Fig. 3.4 shows a state transition of the PU from the current time slot to the next time slot, which follows a two-state discrete time Markov chain process, where $P_{ij} : i, j \in \{P, A\}$ denotes the transition probability from state i to state j [52]. The action of the ST depends on the results of the

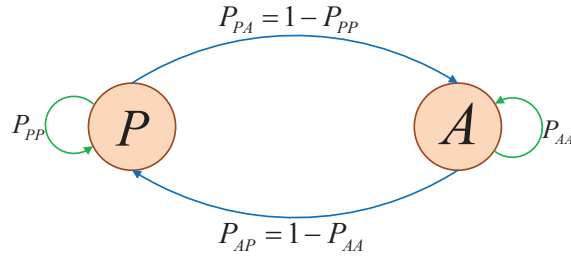


Figure 3.4: Markov chain model of the PU.

sensing mechanism.

If the sensing result indicates the presence of the PT in a time slot, the ST will trust the sensing mechanism and backscatter data to the SR or harvest energy from RF power of the PT. When the ST performs RF energy harvesting in a time slot, the total amount of harvested RF energy can be determined as follows [55]:

$$E_h = (T - \tau_s - \tau_D) \delta P_P h_{P_T S_T}, \quad (3.1)$$

where δ represents the energy harvesting efficiency; P_P is the transmission power of the PT, and T , τ_s , and τ_D are the whole frame time, the sensing time, and the decision time, respectively; $h_{P_T S_T}$ represents the channel gain from PT to ST, which is defined in a Friis equation as follows:

$$h_{P_T S_T} = \frac{G_T G_t \lambda^2}{(4\pi d_{P_T S_T})^2}, \quad (3.2)$$

where G_T and G_t are the antenna gain of the PT and the ST, respectively; λ is the carrier frequency wavelength, and $d_{P_T S_T}$ is the distance between the PT and the ST. When the ST performs backscattering in a time slot, the transmission rate of the ST is given by

$$R_B = \frac{T - \tau_s - \tau_D}{T} B_b, \quad (3.3)$$

where B_b is the achievable backscatter rate of the backscattering action.

If the sensing result indicates the absence of the PT in a time slot, the ST can transmit data packets to the SR. The received signal, y_{SR} , at the secondary receiver is given by

$$y_{SR}(t) = \sqrt{P_S} h_{S_T S_R} x(t) + n_{SR}, \quad (3.4)$$

where P_S is the transmission power of ST, and $h_{S_T S_R}$ represents the channel gain between the ST and the SR. In this chapter, we assume that channel gain $h_{S_T S_R}$ is a block-fading and frequency non-selective parameter that is constant over each time slot and follows a Rayleigh fading distribution; $x(t)$ is the normalized information signal of the ST, i.e., $\mathbb{E}\{|x(t)|^2\} = 1$, and n_{S_R} is additive white Gaussian noise at the SR. The transmission rate of the ST is given as

$$R_{Tr} = \frac{T - \tau_s - \tau_D}{T} W \log_2 \left(1 + \frac{P_S |h_{S_T S_R}|^2}{\sigma^2} \right), \quad (3.5)$$

where W is bandwidth of the primary channel, σ^2 is noise power at the secondary receiver.

3.2.2 Non-RF Energy Harvesting Model

Herein, we assume that the ST always harvests non-RF energy over the whole time slot from ambient sources (e.g., solar, wind, thermal...), which is shown in Fig. 3.3. As such, the ST can automatically and separately harvest non-RF energy in each of the sensing, decision, and implementation phases to replenish its battery in all of the time slots. The non-RF energy is assumed to follow a stochastic Poisson process with mean value e_{mean}^{hv} . The value of e^{hv} in time slot t can be described as follows:

$$e^{hv}(t) \in \left\{ e_1^{hv}, e_2^{hv}, \dots, e_\xi^{hv} \right\}. \quad (3.6)$$

where $0 \leq e_1^{hv} < e_2^{hv} < \dots < e_\xi^{hv} \leq E_{ca}$, and E_{ca} denotes the battery capacity of the secondary transmitter. The probability mass function (PMF) of e_{mean}^{hv} can be computed as

$$p^{hv}(k) = \Pr \left(e^{hv} = e_k^{hv} \right) = \frac{e^{-e_{mean}^{hv}} \left(e_{mean}^{hv} \right)^k}{k!}, k = 1, 2, \dots, \xi. \quad (3.7)$$

3.2.3 Spectrum Sensing

This chapter considers the imperfect spectrum sensing model. The probability of detection in the sensing scheme is P_d , whereas the probability of false alarm is P_f . P_d is the probability that the SU correctly detects the presence of the PU on the PC; meanwhile, P_f is the probability that the SU detects a signal's presence on the PC, but the PU is actually absent from the PC. More specifically, in this chapter, the ST performs spectrum sensing to identify any activity by the PT on the primary channel in every single time slot. The

sensing performance of the ST can be evaluated based on the value of P_d and P_f , which are given as follows [78]:

$$P_d = Q\left(\frac{\chi - M(\gamma + 1)}{\sqrt{2M(2\gamma + 1)}}\right), \quad (3.8)$$

and

$$P_f = Q\left(\frac{\chi - M}{\sqrt{2M}}\right), \quad (3.9)$$

where $M = 2\tau_s f_s$ is the number of sensing samples, in which τ_s is the sensing time duration and f_s denotes the sampling frequency of the ST, χ is energy threshold, and γ denotes average channel gain from the PT to the sensing device. The value for probability of detection P_d is set according to the maximum allowable probability that the ST transmission collides with the PT on the primary channel [52]. Hence, probability of false alarm P_f , according to sensing time τ_s , can be calculated as follows [78]:

$$P_f = Q\left(\sqrt{2\gamma + 1}Q^{-1}(P_d) + \sqrt{\tau_s f_s \gamma}\right). \quad (3.10)$$

3.3 Problem Formulation

In this chapter, we aim to improve the long-term performance of the secondary system in a wireless-powered cognitive radio network by using backscatter technology. The imperfection of the sensing mechanism and the energy-constrained problem are taken into account. By applying POMDP, we propose an approach to maximize the overall reward for secondary users. Based on the remaining energy of the ST and the belief regarding the absence of the PT on the primary channel of a time slot, the ST will determine the optimal action (e.g. backscattering, harvesting, or transmitting) to maximize its long-term transmission rate. We define the set of actions as $\mathbb{A} = \{BS, HV, TM\}$ where BS, HV, and TM represent backscattering, harvesting, and transmitting, respectively. Therefore, the reward of the secondary system can be defined as follows:

$$\begin{aligned} R = & \arg \max_{a^{opt}(t), e_{tr}^{opt}(t)} \sum_{k=t}^{\infty} R(t) \\ \text{s.t. } & 0 \leq e_{tr}(t) \leq e_{tr}^{\max} \end{aligned} \quad (3.11)$$

where

$$R(t) = \begin{cases} \beta R_B, & \text{if } a^{opt}(t) = BS \\ (1 - \beta)R_{Tr}, & \text{if } a^{opt}(t) = TM \\ 0, & \text{if } a^{opt}(t) = HV \end{cases}$$

and

$$\beta = \begin{cases} 1, & \text{if PC is actually busy} \\ 0, & \text{otherwise} \end{cases}$$

$a^{opt}(t) \in \mathbb{A}$ represents the optimal action of the secondary transmitter in time slot t , and $e_{tr}^{opt}(t)$ is the optimal amount of transmission energy to transmit data in time slot t if the ST selects to transmit; otherwise, $e_{tr}^{opt}(t) = 0$ if backscattering or harvesting is chosen. R_B and R_{Tr} are given by Eqs. (3.3) and (3.5), respectively. Note that although the reward will be zero when the ST executes the harvesting action in time slot t , the battery of the ST has a chance to be replenished by harvested RF energy from the PT's transmission.

3.4 Proposed Scheme

Efficiently utilizing the limited energy of a wireless secondary user, such that the secondary system can achieve a high long-term transmission rate, is quite challenging in CRNs. In addition, spectrum sensing imperfection also affects overall network performance. Therefore, we propose a novel scheme in order to obtain the optimal solution for the ST to efficiently share the spectrum with the PT despite the energy-constrained problem.

3.4.1 Proposed Scheme Description and Observations

Let us define the belief that the PT is absent from the primary channel as p . When the ST completes the processing time in time slot t , belief p will be updated according to the possible observations. In other words, the ST will update its remaining energy and the belief for time slot $t + 1$ after performing the selected action in time slot t . The state of the ST in the t^{th} time slot is denoted as $s(t) = \{e_{re}(t), p(t)\}$, where $e_{re}(t)$ is the amount of remaining energy in the battery, and $p(t)$ is the belief that the PT is absent from the channel in that time slot. In this chapter, we assume the energy consumption for backscattering and harvesting is small and negligible. We next present the possible observations based on the set of actions \mathbb{A} , as follows.

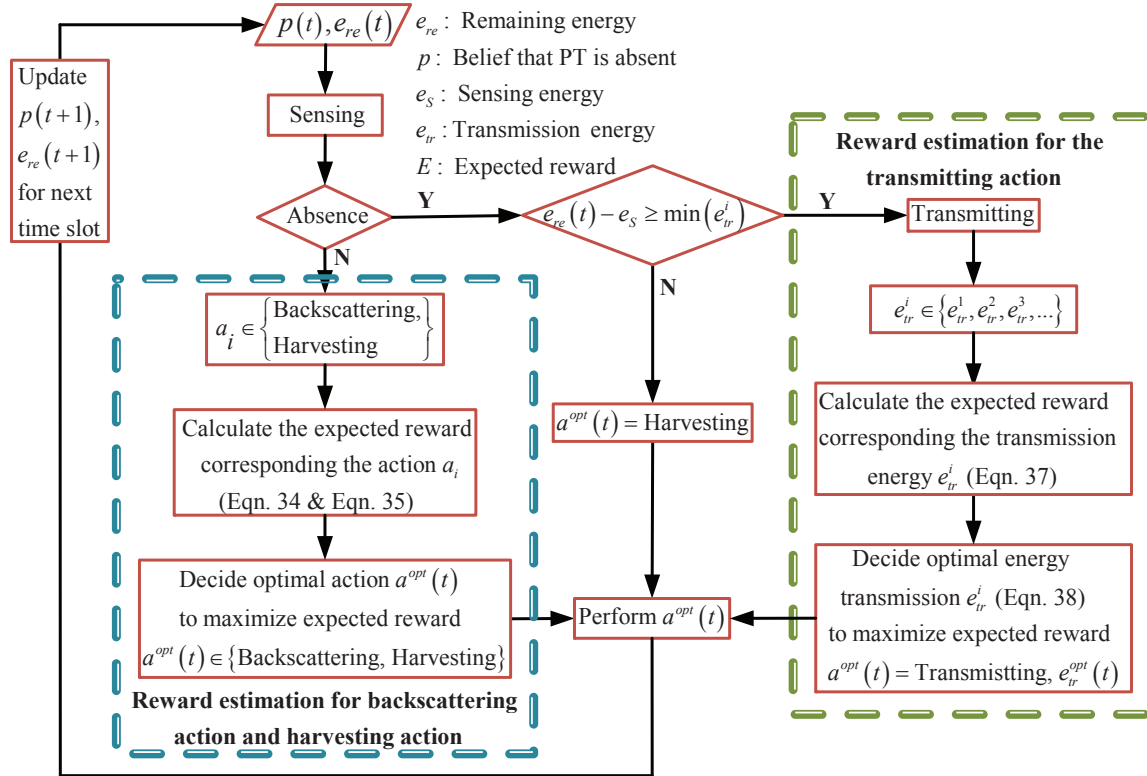


Figure 3.5: The flowchart of the proposed scheme.

3.4.1.1 Backscattering

If the ST executes BS when sensing result is present, there are two observations, $\Omega_1 \rightarrow \Omega_2$, as follows.

Observation 1 (Ω_1): If the ST backscatters data to the SR and receives an acknowledgement (ACK). That implies the PU is really present in the current time slot, the transmission rate is given by

$$R(e_{re}(t), p(t) | \Omega_1) = R_B. \quad (3.12)$$

The probability that event Ω_1 happens is

$$\Pr(\Omega_1) = (1 - p(t)) P_d. \quad (3.13)$$

The belief for the next time slot can be updated as

$$p(t+1) = P_{PA}. \quad (3.14)$$

The updated remaining energy for next time slot is calculated by

$$e_{re}(t+1) = \min \left\{ \max \left\{ e_{re}(t) - e_s + e^{hv}(t), 0 \right\}, E_{ca} \right\}. \quad (3.15)$$

The energy transition probability is computed as

$$\Pr \{ e_{re}(t) \rightarrow e_{re}(t+1) | \Omega_1 \} = \Pr \left(e^{hv}(t) = e_k^{hv} \right). \quad (3.16)$$

Observation 2 (Ω_2): If the ST backscatters data to the SR and does not receive an ACK or receives a negative ACK (NACK). It means that the PU is absent in the current time slot, there is no achieved data transmission at the SR, i.e. $R(e_{re}(t), p(t) | \Omega_2) = 0$. The probability that Ω_2 happens can be calculated as follows:

$$\Pr(\Omega_2) = p(t)P_f. \quad (3.17)$$

The updated belief that the PT will be absent in the next time slot is given as

$$p(t+1) = P_{AA}. \quad (3.18)$$

The updated remaining energy of the ST in the next time slot and the energy transition probability are given in Eqs. (3.15) and (3.16), respectively.

3.4.1.2 Harvesting

There are two cases in which the ST executes HV, as follows.

Case 1

If the ST harvests RF energy when sensing result is present, there are two observations: $\Omega_3 \rightarrow \Omega_4$.

Observation 3 (Ω_3): The sensing result indicates the presence of the PT, so the ST harvests RF energy from the PT and gets a certain amount. We realize that the PU is actually present in the current time slot. Hence, the achieved data transfer in this case is zero, i.e. $R(e_{re}(t), p(t) | \Omega_3) = 0$.

The probability that event Ω_3 happens is

$$\Pr(\Omega_3) = (1 - p(t))P_d. \quad (3.19)$$

The belief for the next time slot can be updated as

$$p(t+1) = P_{PA}. \quad (3.20)$$

The updated remaining energy for the source node is

$$e_{re}(t+1) = \min \left\{ \max \left\{ e_{re}(t) + E_h + e^{hv}(t) - e_s, 0 \right\}, E_{ca} \right\}. \quad (3.21)$$

The energy transition probability is given in Eq. (3.16).

Observation 4 (Ω_4): The sensing result indicates the presence of the PT, so the ST harvests RF energy from the PT but does not get any. That means the PU is really absent in the current time slot. There is no achieved data transfer, i.e. $R(e_{re}(t), p(t) | \Omega_4) = 0$.

The probability that Ω_4 happens can be calculated as follows:

$$\Pr(\Omega_4) = p(t)P_f. \quad (3.22)$$

The updated belief that the PT will be absent in the next time slot is given as

$$p(t+1) = P_{AA}. \quad (3.23)$$

The updated remaining energy for the ST is given as

$$e_{re}(t+1) = \min \left\{ \max \left\{ e_{re}(t) - e_s + e^{hv}(t), 0 \right\}, E_{ca} \right\}. \quad (3.24)$$

The energy transition probability is calculated with Eq. (3.16).

Case 2

If sensing result is absent but the ST does not have enough energy to transmit data to the SR, the ST executes HV instead of TM. There are two observations $\Omega_5 \rightarrow \Omega_6$ for this case.

Observation 5 (Ω_5): The sensing result indicates the absence of the PT, but the ST does not have enough energy to execute TM; then, the ST harvests RF energy from the PT and gets a certain amount. This means the PU actually performs its operation in the current time slot. Hence, the achieved data transfer in this case is zero, i.e. $R(e_{re}(t), p(t) | \Omega_5) = 0$.

The probability that the event occurs can be calculated as

$$\Pr(\Omega_5) = (1 - p(t))(1 - P_d). \quad (3.25)$$

The belief, the remaining energy, and the energy transition probability for the next time slot are updated using Eqs. (3.20), (3.21), and (3.16), respectively.

Observation 6 (Ω_6): The sensing result indicates the absence of the PT, but the ST does not have enough energy to execute TM; then, the ST harvests RF energy from the PT but does not get any. That implies there is no appearance of PU in the current time slot. Hence, the achieved data in this case transfer is zero, i.e. $R(e_{re}(t), p(t) | \Omega_6) = 0$.

The probability that Ω_6 happens can be calculated as

$$\Pr(\Omega_6) = p(t)(1 - P_f). \quad (3.26)$$

The belief, the remaining energy, and the energy transition probability for the next time slot are updated with Eqs. (3.23), (3.24) and (3.16), respectively.

3.4.1.3 Transmitting

If the sensing result indicates the PT is absent, and the ST has enough energy to transmit data to the SR, then the ST executes TM. There are two observations, $\Omega_7 \rightarrow \Omega_8$, in this case.

Observation 7 (Ω_7): If the sensing result indicates the PT is absent, and the ST currently has enough energy to transmit data to the SR, then the ST performs TM and receives an ACK. We realize there is no operation of PU in the current time slot. The transmission rate will be calculated as

$$R(e_{re}(t), p(t) | \Omega_7) = R_{Tr}. \quad (3.27)$$

The probability that Ω_7 happens is computed by

$$\Pr(\Omega_7) = p(t)(1 - P_f). \quad (3.28)$$

The belief for the next time slot is updated as

$$p(t+1) = P_{AA}. \quad (3.29)$$

The remaining energy of the ST in the next time slot is given as

$$e_{re}(t+1) = \min \left\{ \max \left\{ e_{re}(t) - e_s - e_{tr}(t) + e^{hv}(t), 0 \right\}, E_{ca} \right\}. \quad (3.30)$$

The energy transition probability is computed in a way similar to Eq. (3.16).

Algorithm 3.1 Proposed Scheme to Find the Optimal Decision for the ST in Processing Time Slots

```

1: Input:  $e_{re}(t)$ ,  $p(t)$ ,  $\mathbb{A} = \{\text{BS, HV, TM}\}$ 
2: for ST performs spectrum sensing.
3:   if The sensing result is absent.
4:     if  $e_{re}(t) < \min(e_{tr}^i)$ 
5:        $a^{opt}(t) = \text{Harvesting}$ ,  $e_{tr}^{opt}(t) = 0$ ; ST performs harvesting.
6:       if The ST successfully harvests RF energy.
7:         Update belief  $p$  and remaining energy  $e_{re}$  with Eqs. (3.20) and (3.21).
8:       else Update belief  $p$  and remaining energy  $e_{re}$  with Eqs. (3.23) and (3.24).
9:       end if
10:      else Calculate the overall expected reward corresponding  $e_{tr}^i \in \{e_{tr}^1, e_{tr}^2, \dots, e_{tr}^{\max}\}$ ,
      by using Eq. (3.37).
11:      Define  $e_{tr}^{opt}(t)$ , with Eq. (3.38), then transmit with  $e_{tr}^{opt}(t)$ .
12:      if The ST successfully transmits data to the SR.
13:        Update belief  $p$  and remaining energy  $e_{re}$  with Eq. (3.29) and Eq. (3.30).
14:      else Update belief  $p$  and remaining energy  $e_{re}$  with Eq. (3.32) and Eq. (3.30).
15:      end if
16:      end if
17:      else Calculate the overall expected reward with Eq. (3.34) or Eq. (3.35) if the ST
      executes BS or HV, respectively.
18:      Find  $a^{opt}(t) = \{\text{Backscattering, Harvesting}\}$  by using Eq. (3.36).
19:      if  $a^{opt}(t) = \text{Backscattering}$ , then execute backscattering.
20:        if The ST successfully backscatters data.
21:          Update belief  $p$  and remaining energy  $e_{re}$  with Eq. (3.14) and Eq. (3.15).
22:        else Update belief  $p$  and remaining energy  $e_{re}$  with Eq. (3.18) and Eq. (3.15).
23:        end if
24:      else  $a^{opt}(t) = \text{Harvesting}$ ; then go to step 5.
25:      end if
26:      end if
27:      Continue until the ST does not have data to transmit to the SR.
28: end for
29: Output:  $a^{opt}(t)$ ,  $e_{tr}^{opt}(t)$ .

```

Observation 8 (Ω_8): If the sensing result indicates the PT is absent, and the ST currently has enough energy to transmit data to the SR, then the ST executes TM and does not receive an ACK (or gets a NACK). That implies, in fact, the PU is present in the current time slot. The achieved data transfer at the SR will be zero, i.e. $R(e_{re}(t), p(t) | \Omega_8) = 0$.

The probability that Ω_8 happens is

$$\Pr(\Omega_8) = (1 - p(t))(1 - P_d). \quad (3.31)$$

The belief for the next time slot can be updated as

$$p(t+1) = P_{PA}. \quad (3.32)$$

The updated remaining energy and the energy transition probability are given as Eqs. (3.30) and (3.16), respectively.

3.4.2 Overall Expected Reward

The final decision of the secondary transmitter depends on the maximum value of the total discounted expected reward, called the value function, which is calculated by following the POMDP framework. In this chapter, the ST utilizes the value function to calculate the overall expected reward in order to decide the optimal action. Value function $V(e_{re}(t), p(t))$, according to remaining energy $e_{re}(t)$ and belief $p(t)$, starting from time slot t is given as follows:

$$V(e_{re}(t), p(t)) = \arg \max_{a(t) \in A} \left\{ \begin{array}{l} \sum_{k=t}^{\infty} \alpha^{k-t} \sum_{\Omega_i \in a(k)} \Pr[\Omega_i] \\ \times \sum_{e_{re}(k+1)} \Pr(e_{re}(k) \rightarrow e_{re}(k+1) | \Omega_i) \\ \times R(e_{re}(k), p(k), a(k) | \Omega_i) | e_{re}(k) = e, p(k) = p \end{array} \right\} \quad (3.33)$$

where k is the index of the time slot, t denotes the current time slot; α is the discount factor, and Ω_i denotes the possible observation of the action, $a(k)$; $R(e_{re}(k), p(k), a(k) | \Omega_i)$ represents the estimated reward when the remaining energy is $e_{re}(k)$, the belief is $p(k)$, and the taken action is $a(k)$ with corresponding observation Ω_i .

Let $E_a(e_{re}(t), p(t))$ be the overall expected reward if the PT is present on the primary channel. $E_{a_B}(e_{re}(t), p(t))$ and $E_{a_H}(e_{re}(t), p(t))$ denote the overall expected reward when the PT is present and the ST performs backscattering and harvesting, respectively.

The overall expected reward if the ST executes backscattering in time slot t can be computed by

$$\begin{aligned}
 E_{a_B}(e_{re}(t), p(t)) = & \\
 P_{ACK}^B \times \left(R_{a_B}(t) + V_{a_B} \left(e_{re}^{B,ACK}(t+1), p^{B,ACK}(t+1) \right) \right) & \quad (3.34) \\
 + P_{NACK}^B \times V_{a_B} \left(e_{re}^{B,NACK}(t+1), p^{B,NACK}(t+1) \right) &
 \end{aligned}$$

where P_{ACK}^B and P_{NACK}^B represent the probability that backscattering is executed successfully and unsuccessfully, respectively. $R_{a_B}(t)$ is the expected immediate reward in time slot t if the ST executes backscattering. $V_{a_B} \left(e_{re}^{B,ACK}(t+1), p^{B,ACK}(t+1) \right)$ and

$$V_{a_B} \left(e_{re}^{B,NACK}(t+1), p^{B,NACK}(t+1) \right)$$

represent the expected reward from time slot $t+1$ after executing backscattering in time slot t successfully and unsuccessfully, respectively; $e_{re}^{B,ACK}(t+1)$ and $e_{re}^{B,NACK}(t+1)$ represent the updated remaining energy for time slot $t+1$ at the ST after executing backscattering successfully or unsuccessfully, respectively. In addition, $p^{B,ACK}(t+1)$ and $p^{B,NACK}(t+1)$ represent the belief for the next time slot after executing backscattering successfully or unsuccessfully, respectively. The overall expected reward if the ST harvests energy in time slot t can be computed as

$$\begin{aligned}
 E_{a_H}(e_{re}(t), p(t)) = & \\
 P_{ACK}^H \times V_{a_H} \left(e_{re}^{H,ACK}(t+1), p^{H,ACK}(t+1) \right) & \quad (3.35) \\
 + P_{NACK}^H \times V_{a_H} \left(e_{re}^{H,NACK}(t+1), p^{H,NACK}(t+1) \right) &
 \end{aligned}$$

where P_{ACK}^H and P_{NACK}^H represent the probability that harvesting action is executed successfully and unsuccessfully, respectively. $V_{a_H} \left(e_{re}^{H,ACK}(t+1), p^{H,ACK}(t+1) \right)$ and $V_{a_H} \left(e_{re}^{H,NACK}(t+1), p^{H,NACK}(t+1) \right)$ represent the expected reward from time slot $t+1$ after harvesting energy in time slot t successfully and unsuccessfully, respectively, and $e_{re}^{H,ACK}(t+1)$ and $e_{re}^{H,NACK}(t+1)$ are the updated remaining energy for time slot $t+1$ at the ST after harvesting energy successfully or unsuccessfully, respectively. In addition, $p^{H,ACK}(t+1)$ and $p^{H,NACK}(t+1)$ represent the updated belief for time slot $t+1$ after harvesting energy successfully or unsuccessfully, respectively.

The overall expected reward, $E_a(e_{re}(t), p(t))$, if the PT is present in time slot t can be obtained as follows:

$$E_a(e_{re}(t), p(t)) = \max(E_{a_B}(e_{re}(t), p(t)), E_{a_H}(e_{re}(t), p(t))) \quad (3.36)$$

When the PT is absent in time slot t and the ST has enough energy to transmit data, if the ST uses transmission energy $e_{tr}^i \in \{e_{tr}^1, e_{tr}^2, \dots, e_{tr}^{\max}\}$ to transmit data to the SR, the overall expected reward in time slot t is given as follows:

$$\begin{aligned} E_{e_{tr}^i}(e_{re}(t), p(t)) = & \\ P_{ACK}^{Tr} \times & \left(R_{e_{tr}^i}(t) + V_{e_{tr}^i} \left(e_{re}^{Tr,ACK}(t+1), p^{Tr,ACK}(t+1) \right) \right) \\ + P_{NACK}^{Tr} \times & V_{e_{tr}^i} \left(e_{re}^{Tr,NACK}(t+1), p^{Tr,NACK}(t+1) \right) \end{aligned} \quad (3.37)$$

where P_{ACK}^{Tr} and P_{NACK}^{Tr} represent the probability that transmitting is performed successfully and unsuccessfully, respectively. $R_{e_{tr}^i}(t)$ indicates the expected immediate reward in time slot t if the ST uses an amount of energy, e_{tr}^i , to transmit data to the SR successfully. $V_{e_{tr}^i} \left(e_{re}^{Tr,ACK}(t+1), p^{Tr,ACK}(t+1) \right)$ and $V_{e_{tr}^i} \left(e_{re}^{Tr,NACK}(t+1), p^{Tr,NACK}(t+1) \right)$ are the expected rewards from time slot $t+1$ after transmitting data with transmission energy e_{tr}^i in time slot t successfully and unsuccessfully, respectively; $e_{re}^{Tr,ACK}(t+1)$ and $e_{re}^{Tr,NACK}(t+1)$ represent the updated remaining energy for time slot $t+1$ at the ST after transmitting successfully or unsuccessfully, respectively. $p^{Tr,ACK}(t+1)$ and $p^{Tr,NACK}(t+1)$ are the updated beliefs for time slot $t+1$ after transmitting successfully or unsuccessfully, respectively. $E_{e_{tr}^i}(e_{re}(t), p(t))$ is the overall expected reward corresponding the different value of transmission energy e_{tr}^i in the time slot t .

By solving a sub-optimal problem related to Eq. (3.37), we achieve the optimal amount of transmission energy as follows:

$$e_{tr}^{opt}(t) = e_{tr}^{i*} = \arg \max_{e_{tr}^i \in \{e_{tr}^1, e_{tr}^2, \dots, e_{tr}^{\max}\}} \left(E_{e_{tr}^i}(e_{re}(t), p(t)) \right) \quad (3.38)$$

In Eq. (3.38), we are going to find out the optimal value of transmission energy e_{tr}^{opt} regarding the current state $(e_{re}(t), p(t))$ for the time slot t .

3.4.3 Optimal Mode Decision Policy

Fig. 3.5 shows a flowchart of the proposed scheme. By applying the proposed scheme, the system's operations can be summarized as follows. At the beginning of time slot t , the secondary transmitter senses the primary channel. Subsequently, if sensing result indicates the PT present, the ST will choose the optimal action as either backscattering or harvesting, based on the overall expected reward for current time slot t in the Eqs. (3.34), (3.35), and (3.36). On the other hand, if the PT is absent, the ST will harvest energy from

ambient RF signals (i.e from the transmission power of the PT) if the remaining energy of the ST is not enough to transmit data. Conversely, if the ST has enough energy to transmit data (i.e. $e_{re}(t) \geq e_{tr}^{\min}$) when the PT is absent, it will decide the optimal amount of transmission energy to use to transmit data to the SR. The optimal transmission energy for time slot t is decided based on the overall expected reward, obtained by Eqs. (3.37) and (3.38). After finding out the optimal action, $a^{opt}(t)$, the ST then implements that action for the actual operation. Then, the remaining energy and the belief will be updated for time slot $t + 1$, based on observations after finishing the selected action. Subsequently, the process will continue to finish the total number of considered time slots. The overall flow to find optimal decision for the ST in each processing time slot is shown in **Algorithm 3.1**. The computation complexity of algorithm can be analyzed regarding the number of actions, states, transition probability, and observations. The optimal policy is chosen according to the iteration-based dynamic value function based on the Bellman's equation. In the POMDP, the algorithm complexity can be decomposed as the number of $O(|\mathbb{A}| |\mathbb{S}|^2)$ operations [79], where $|\mathbb{A}|$ and $|\mathbb{S}|$ are the possible action set and state set at the beginning of each time slot, respectively. It is mainly required for calculating the transition probabilities from one state ($s(t) \in \mathbb{S}$) to another state ($s'(t) \in \mathbb{S}$) after implementing an action ($a(t) \in \mathbb{A}$).

3.5 Simulations

In this section, we present simulation results and discussions to verify the efficiency of secondary system performance with the proposed scheme under various conditions in the network. In addition, we carried out the performance comparisons with other conventional schemes, namely, the Myopic-B scheme, the Myopic-H scheme, the Myopic-R scheme, Reference scheme [80] and the random scheme. In all myopic schemes, the ST always transmits data to the SR with the maximum transmission energy level when the sensing result indicates the PT is absent. However, when the PT is present, in the case of the Myopic-B scheme, the ST always executes backscattering. Similarly in the case of the Myopic-H scheme, the ST always executes harvesting. In the case of the Myopic-R scheme, the ST chooses either backscattering or harvesting randomly when the PT is present. Finally, in the case of the random scheme, the ST chooses randomly available backscattering or harvesting if sensing shows the PT is present, and if sensing shows the PT is absent, it will randomly select an amount of energy to transmit data to the SR. Table 3.1 shows the

Table 3.1: Simulation Parameters

Parameter	Description	Value
N	Number of time slots	10^4
T	Time slot duration	40 <i>ms</i>
τ_s	Sensing duration	0.4 <i>ms</i>
E_{ca}	Battery capacity	30 μJ
E_{tr}	Transmission energy	7, 11, 16, 20 μJ
e_{mean}^h	Mean value of harvested non-RF energy	7 μJ
p	Initial belief that the PU is absent	0.5
P_{AA}	Transition probability of the PU from state A to itself	0.8
P_{PA}	Transition probability of the PU from state P to state A	0.2
δ	Energy-harvesting efficiency	0.6
d_{PTST}	Distance between PU transmitter and SU transmitter	100 <i>m</i>
d_{STSR}	Distance between SU transmitter and SU receiver	10 <i>m</i>
P_d	Probability of detection	0.9
P_f	Probability of false alarm	0.1
P_P	Transmission power of the PU	10 <i>dBm</i>
G_T, G_t	Antenna gain of the PT and the ST	6 <i>dB</i>
σ^2	Noise variance at the SU receiver	0.01
α	Discount factor	0.9
W	Bandwidth	14 <i>MHz</i>
f_c	Frequency	2.15 <i>GHz</i>

parameter settings of our simulation.

Unless otherwise stated, we assume that the time for the ST to make the decision, τ_D , is short and negligible. The PT is assumed to be a cellular base station. Bandwidth and frequency of the RF signals are set at 14 MHz and 2.15 GHz, respectively [55]. The

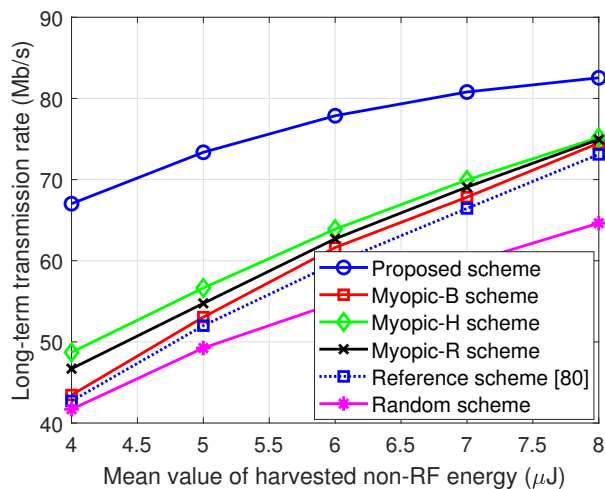


Figure 3.6: The long-term transmission rate of the secondary system under various values for harvested non-RF energy.

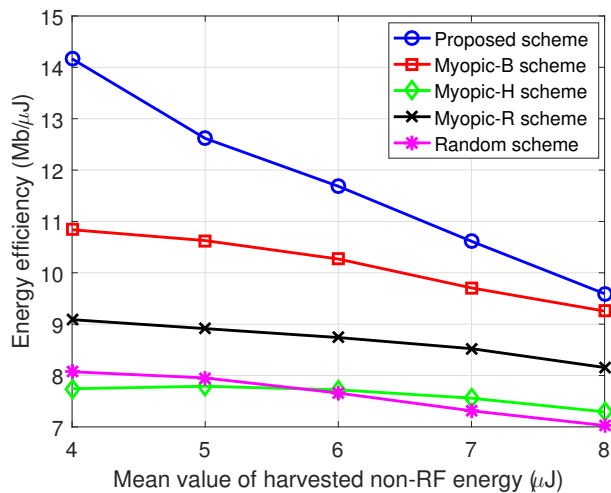


Figure 3.7: The energy efficiency of the secondary system under various values of harvested non-RF energy.

backscatter rate is 33 kbps [55]. The distance between the ST and the SR is 10 meters. In addition, we set the path loss exponent at 3, and the step size of the belief is 0.01 within the range (0,1).

We first show the impact of non-RF harvested energy on secondary system performance for all considered schemes. The simulation results for the various mean values of harvested non-RF energy are illustrated in Fig. 3.6, Fig. 3.7, and Fig. 3.8. In Fig. 3.6, we

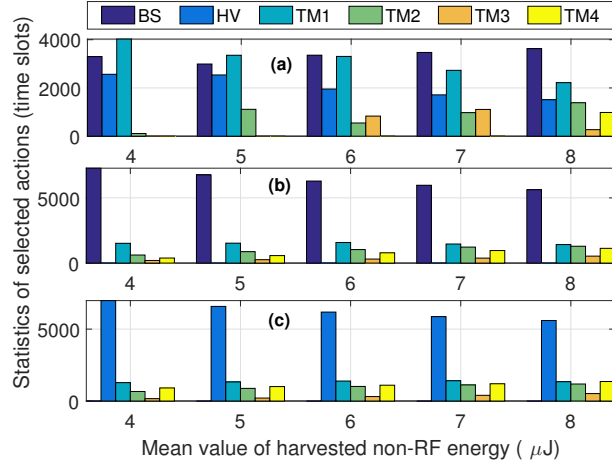


Figure 3.8: The selected action statistics of the secondary system under various values of harvested non-RF energy.

inspect the long-term transmission rate of the schemes with increasing values of harvested energy per time slot. The reference scheme [80] is adopted by optimizing one-step instant reward without using backscatter or RF-harvesting when the sensing outcome is busy. The curves show that the average long-term transmission rate of the ST greatly improves as the mean of harvested non-RF energy increases. This is because the ST has more energy to use for the transmission phase when the total amount of harvested non-RF energy becomes larger. Besides, the reward obtained in reference scheme is less than the scheme myopic-H scheme and Myopic-B scheme. It is because the system does not leverage backscattering and RF harvesting techniques. For instance, when $e_{mean}^h = 4\mu J$, the transmission rate of the proposed scheme provides improvements of 36.7%, 50.6%, and 53.1% for the Myopic-H scheme, the Myopic-B scheme and Reference scheme, respectively. The random scheme provides the lowest transmission rate due to the random selection when sensing determines the PT is both present and absent from the channel.

In Fig. 3.7, we investigate the energy efficiency of the system according to different mean values of harvested non-RF energy. The energy efficiency in this chapter is defined as an average long-term transmission rate over the total of harvested non-RF and RF energy (in μJ unit) of the ST during its operation over N time slots $\left(EE = \frac{\sum_{t=1}^N R(t)}{\sum_{t=1}^N (E_h(t) + e^{hv}(t))} \right)$. We can see that the energy efficiency of all the schemes degrades as e_{mean}^h increases. The reason is as following: The more energy the secondary transmitter harvests, the larger the

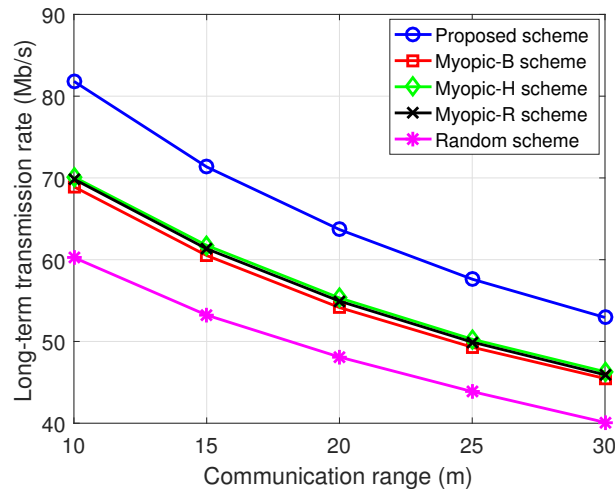


Figure 3.9: The long-term transmission rate of the secondary system according to different communications ranges.

overflow energy the ST suffers from in its operations. The curves show the superiority of the proposed scheme as compared with other schemes.

The statistics on action utilization for the proposed scheme, the Myopic-B scheme, and the Myopic-H scheme over 10^4 time slots are illustrated in Fig. 3.8(a), Fig. 3.8(b), and Fig. 3.8(c), respectively. For simplicity, the amount of transmission energy is divided into four levels, i.e $7\mu J$, $11\mu J$, $16\mu J$, and $20\mu J$ with corresponding notations TM1, TM2, TM3, and TM4. The results show that the Myopic-B and the Myopic-H schemes have fewer chances for ST data transmission with the ST’s own energy because it always uses the maximum transmission energy whenever there is enough energy for transmission. Moreover, when sensing indicates the PT is present, the ST always chooses backscattering or harvesting in the Myopic-B scheme and the Myopic-H scheme, which will lower the long-term rewards of the secondary system. As such, this results in the poor performance shown in Fig. 3.6 and Fig. 3.7. We see in Fig. 3.8(a) that when e_{mean}^h is small, the ST in the proposed scheme only uses a small amount of transmission energy to maximize the total achievable throughput. Therefore, by dynamically choosing the optimal action in each time slot, the proposed scheme provides more opportunities for the ST to transmit its data to the SR when the PT is sensed as absent from the primary channel. As a result, the performance of the wireless-powered CRN system can be significantly enhanced.

In the rest of the simulation section, we change the communication range between

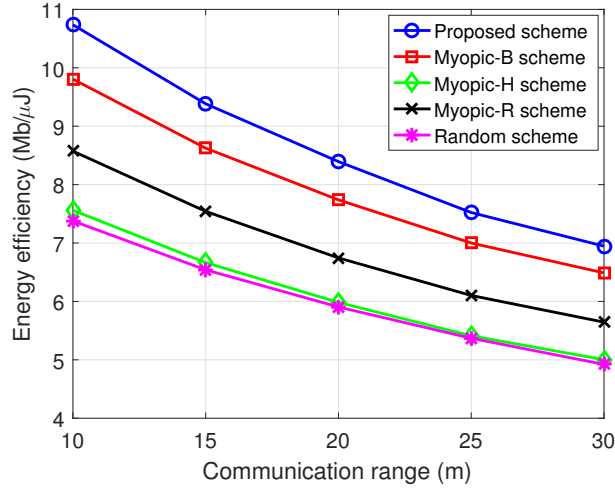


Figure 3.10: The energy efficiency of the secondary system according to different communications ranges.

the ST and the SR to observe the performance of the considered schemes, which is illustrated in Fig. 3.9, Fig. 3.10, and Fig. 3.11. Fig. 3.9 shows the long-term transmission rate of the secondary transmitter with different distances between the ST and the SR. The results in the figure show that a greater communication range provides a lower transmission rate in the secondary system. This is because the transmitted signals experience more path loss attenuation at a greater distance. Consequently, this degrades the total amount of transmitted data that are successfully decoded at the SR.

In Fig. 3.10, we plot energy efficiency according to the different communication ranges. We can see that the Myopic-B scheme give the highest performance, compared with the other conventional schemes, since it always uses backscattering when the PT is sensed as present on the primary channel. In other words, it results in lower energy consumption by the Myopic-B scheme, in comparison with other conventional schemes. However, the proposed scheme still is superior with a 10% improvement over the Myopic-B scheme.

In Fig. 3.11, the statistics of actions utilized by (a) the proposed scheme, (b) the Myopic-B scheme, and (c) the Myopic-H scheme are presented based on the various distances between the ST and the SR. For all ranges from 10m to 30m, the proposed scheme uses all the actions except TM4 (transmitting with the maximum amount of transmission energy). That is because the proposed scheme always chooses the optimal policy by considering the total expected rewards over the future time slots, not like other conventional schemes that

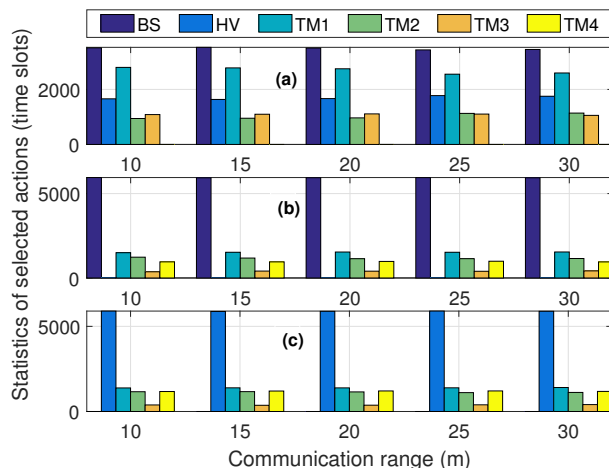


Figure 3.11: The selected action statistics of the secondary system according to different communications ranges.

only consider maximizing the immediate reward in the current time slot. As a result, the small changes regarding the number of selected actions by the proposed scheme can enhance system performance when the communication range varies, which is shown in Fig. 3.9 and Fig. 3.10.

3.6 Conclusion

In this chapter, we proposed a scheme to maximize the long-term transmission rate of wireless-powered CRNs utilizing ambient backscatter. We adopted the POMDP framework to obtain optimal action to maximize the long-term transmission rate of the considered system. Taking the energy-constrained issue of wireless networks into account, we presented the optimal policy for the secondary system in CRNs by combining RF and non-RF energy harvesting for secondary users. As a result, a secondary transmitter equipped with non-RF and RF energy harvesters can apply the proposed scheme to significantly improve network performance. By using a Matlab simulation, we verified the effectiveness of the proposed scheme in comparison with other schemes under various network conditions. As a result, it is showed that the proposed scheme can provide a high long-term transmission rate for the secondary system due to efficient utilization of energy harvesting from the wireless environment.

Chapter 4

Uplink NOMA-based Long-Term Throughput Maximization Scheme for Cognitive Radio Networks: An Actor-Critic Reinforcement Learning Approach

4.1 Introduction

Spectrum scarcity is one of the critical issues in fifth-generation (5G) communications systems and for future wireless networks, because the lack of accessible spectrum is hindering the application of novel communications technologies [81–83]. However, in [84], the authors revealed that the licensed spectrum remains under utilized. In order to deal with spectrum inefficiency, the dynamic spectrum access techniques are studied, with cognitive radio (CR), known as the key enabling technology [85]. In a CR network, the unlicensed secondary users (SUs) can access and utilize the unused spectrum of the licensed primary users (PUs) [4, 86, 87].

Nowadays, the CR network paradigm can broadly be categorized into three main models [88–90]: underlay, overlay, and interweave. In the underlay CR model, the SUs can perform their operations if and only if the interference caused by all SUs is lower than a

given threshold. In the overlay CR model, the SUs assist as relays for the PUs, and jointly transmit their signals using a portion of the licensed spectrum. In the interweave CR model, the SUs can only transmit when the primary channel is not occupied by any PU. With this model, vacant spectrum is temporarily available over certain time instants, such that an SU can opportunistically transmit data when the PU is inactive. In order to reduce collisions with the PUs and ensure energy-efficient utilization, the SUs sense the surrounding spectrum to verify the availability of the primary channel in order to transmit their data.

As another potential technique for the next generation of wireless networks, non-orthogonal multiple access (NOMA) has lately gotten noticeable attention, enabling multiple users to simultaneously access the spectrum, and it has become an important fundamental to designing radio access techniques for future wireless networks [91–94]. The key with NOMA is allowing multiple users to access the same spectrum resource block together, with the objective being spectral efficiency. NOMA is generally classified into two major approaches: power-domain NOMA [95–97] and code-domain NOMA [98–101]. In power-domain NOMA, different power levels are used to jointly serve multiple users at the same time using the channel frequency under different channel conditions. At the receiver, the signals of the different transmitters are superposed and then decoded via successive interference cancellation (SIC).

Moreover, by introducing the two aforementioned concepts, NOMA can be combined with a CR network in order to improve spectral efficiency. Liu *et al.* [102] proposed a stochastic geometry model for a large-scale CR network in order to depict the outage performance from the paradigm of integrated NOMA and CR. In [103], spectrum efficiency was enhanced by developing a NOMA-based secure transmission scheme in CR networks. A cooperative NOMA spectrum-sharing network over the Nakagami fading channel was investigated in [104]. Besides, multicast NOMA is also adopted in 5G systems in terms of user scheduling in order to improve network performance [105]. It has been pointed out that higher spectral efficiency can be promised by combining NOMA with CR networks.

In recent years, prolonging the long-term operation of the network is also one of the nearly essential purposes of wireless systems [106]. Using renewable energy sources for wireless users is considered a potential solution for dealing with the energy constraints of wireless devices. In particular, energy for the SUs can be harvested from natural ambient sources (solar [107, 108], wind [109, 110], radio frequency [?], etc.). Hence, the battery of a wireless user can replenish itself without manual recharging. Nevertheless, the harvested

energy is still restricted to users. That means finding a way for SUs to effectively utilize the harvested energy needs to be carefully investigated. For that reason, Celik *et al.* [111] proposed hybrid energy harvesting in a heterogeneous CR network to enhance spectrum efficiency while reducing the energy consumption of the system.

Furthermore, a dynamic power allocation algorithm is carefully being investigated owing to the significant role of power allocation for wireless users in uplink NOMA (i.e. the effect of power allocation on the rate of each user) [112–114]. In [112], the authors studied both power control and beamforming methods in order to maximize the sum rate of the system for millimeter-wave communications. The joint optimization problem for sum-throughput maximization under transmission power constraints, the minimum rate requirements of users, and SIC constraints were formulated in [113] for both uplink and downlink NOMA in a cellular system. In [114], the authors took into account channel assignment and power control to maximize the sum rate for a NOMA-based uplink network. They mathematically derived a more tractable form of the formulated problems as a maximum weighted independent set issue, and then used graph theory to deal with them.

In this chapter, we study an uplink actor–critic learning-based transmission power allocation scheme that allows multiple SUs access on the same channel by adopting NOMA in order to maximize the long-term throughput of the network. Herein, the SUs employ the NOMA technique to simultaneously transmit their information to the cognitive base station (CBS), and then the CBS can exploit SIC to decode the information. The key contributions of this chapter can be outlined as follows.

- We consider a CR network with uplink NOMA, where the SUs are allowed to concurrently access the same primary channel when it is not used by the PU. Specifically, by adopting NOMA, the SUs can transmit data on the same channel and in the same time slot when the sensing result indicates the primary channel is free. However, the SUs are equipped with a limited-capacity battery. Therefore, solar energy harvesting is executed by the SUs such that they can externally harvest energy to replenish the battery for use in long-term data transmission. In addition, the energy-constrained problem and sensing error issues of the SUs are also taken into account.
- To do this, we first formulate the problem of throughput maximization based on a Markov decision process (MDP). Afterward, the actor–critic reinforcement learning approach is adopted such that the CBS can adaptively interact with the environment

and dynamically assign the optimal amount of energy for each user in every time slot without prior information about the harvested energy model of the SUs, which is normally needed by some kinds of partially observable MDP schemes.

- Simulation results show that the proposed scheme, in terms of average throughput and energy efficiency, outperforms other conventional schemes under various network parameter variations.

The rest of the chapter is structured as follows. The system model is outlined in Section 4.2. In Section 4.3, we describe the problem formulation. The proposed power allocation algorithm is discussed in Section 4.4, and simulation results are provided in Section 4.5. Finally, conclusions are drawn in Section 4.6.

4.2 System Model

4.2.1 Network Model

In this chapter uplink NOMA in a cognitive radio network (CRN) is considered, as shown in Fig. 4.1, which comprises a CBS, a pair of PU transceivers and a set of SUs denoted by $\mathbb{N} = \{SU_1, SU_2, \dots, SU_N\}$. Although the PUs have priority to use the licensed spectrum, the SUs are allowed to simultaneously and opportunistically access the licensed spectrum of the PU when the sensing result indicates that the primary channel is free. In the network, the CBS and SUs are equipped with a single antenna to receive and transmit signals in a time slot on the currently free primary channel. In particular, at the beginning of a time slot, the SUs will share the primary channel to concurrently transmit data to the CBS if the sensing result for the primary channel is free, and then, the CBS will decode all data sent from the SUs by using the SIC technique. In this chapter, the SUs are equipped with a finite battery capacity E_{ca} , and they can replenish energy by themselves using an integrated solar energy harvester.

The operation of the considered network consists of three phases: the sensing and decision phase, the data transmission phase, and the energy information update phase, as shown in Fig. 4.2. In the first phase, with duration τ_{ss} , the SUs perform their individual sensing, and then, they report their local decisions to the CBS; afterward, the CBS gives its global sensing decision (about the state of the primary channel) and its global action decisions on the actions assigned to all SUs. The second phase, with duration τ_{tr} , is the

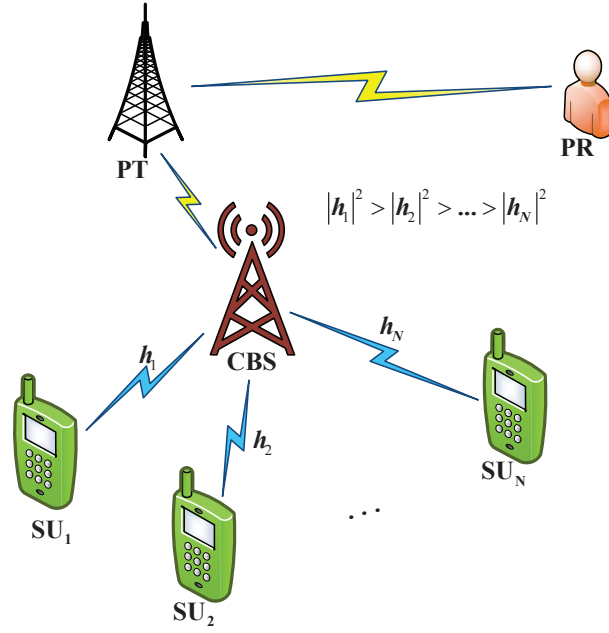


Figure 4.1: System model of the proposed scheme.

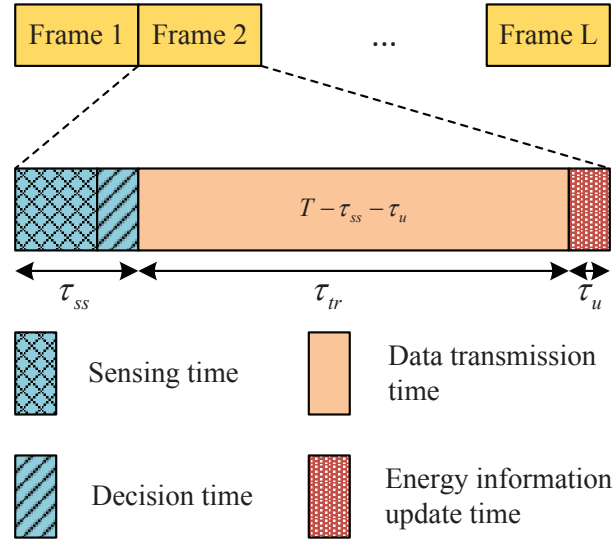


Figure 4.2: Time frame of the three phases in the secondary users's operations.

time for the SUs to transmit their data to the CBS. In the last phase, the SUs will send their remaining information to the CBS. Herein, it is assumed that the SUs always have information available to transmit. Furthermore, each transmission session may last several time frames, until all the information is successfully transmitted.

In this chapter, we adopt cooperative spectrum sensing for the SUs in the network. More specifically, the SUs perform spectrum sensing at the beginning of a time slot to check whether the licensed spectrum is occupied by the PU or not. There are several major sensing approaches, such as matched filtering, energy detection, and the cyclostationary method [115]. Energy detection is one of the most effective methods due to its low computational complexity [116–118]. All sensing results from the SUs are then gathered and sent to the CBS. After that, the CBS makes a global sensing decision about the activity or silence of the PU on the primary channel, and then decides whether the SUs should transmit data to the CBS or stay silent. Normally, the global sensing decision is done by following a combination rule at the CBS [119–122]. However, in this chapter, we do not focus on cooperative spectrum sensing, which has been widely investigated in the literature. Thus, we mainly study a power allocation algorithm for the SUs in order to efficiently use energy to transmit data to the CBS.

When the CBS determines that the PU is absent in the current time slot, all SUs can concurrently transmit their signals to the CBS. The received signal at the CBS is given as follows

$$y(t) = h_1x_1(t) + h_2x_2(t) + \dots + h_Nx_N(t) + \omega, \quad (4.1)$$

where h_i is the channel gain between the CBS and SU_i , $i \in \{1, 2, \dots, N\}$, $x_i(t) = \sqrt{P_i(t)}s_i(t)$, and $|h_1|^2 > |h_2|^2 > \dots > |h_N|^2$, when $s_i(t)$ is the signal transmitted by SU_i ($\mathbb{E}\{|s_i(t)|^2\} = 1$) with transmission power $P_i(t) = e_i^{tr}(t)/\tau_{tr}$, in which $e_i^{tr}(t)$ is the transmission energy assigned to SU_i for the t^{th} time slot; and ω is the additive white Gaussian noise (AWGN) at the CBS with zero mean and variance σ_ω^2 .

Fig. 4.3 illustrates the SIC detection process of the received signals at the CBS, where $|h_1|^2 > |h_2|^2 > \dots > |h_N|^2$. In uplink NOMA, an SU with the strongest channel gain will definitely have the priority for decoding by the CBS, and then, it vanishes from received signals y at the CBS, which continues decoding the other SU signals. Consequently, the attainable throughput of SU_1 is affected by interference from other users (SU_2, SU_3, \dots, SU_N), and meanwhile, the throughput of the lowest channel gain user (i.e SU_N) is obtained without any interference from the other SUs because interference from stronger signals is eliminated by the SIC technique. Thereby, the throughput for SU_i , $\forall i \in \{1, 2, \dots, N\}$ in uplink NOMA

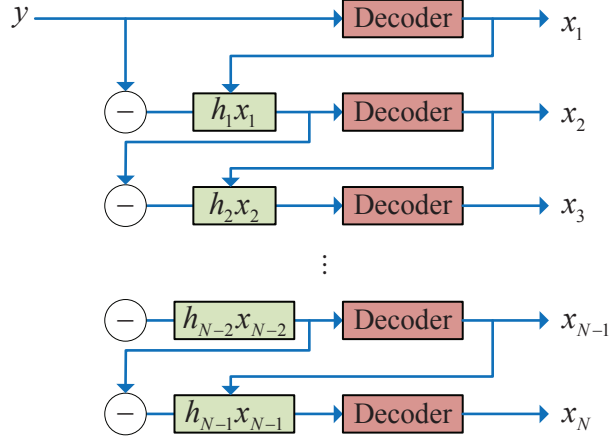


Figure 4.3: Illustration of SIC detection of the signals at the CBS.

can be computed as

$$R_i(t) = \frac{T - \tau_{ss} - \tau_u}{T} \log_2 \left(1 + \frac{P_i(t) |h_i|^2}{\sum_{j=i+1}^N P_j(t) |h_j|^2 + \sigma_\omega^2} \right), \quad (4.2)$$

where T , τ_{ss} , and τ_u denote the whole-frame time, the sensing and decision time, and the energy information update time, respectively. The total received throughput at the CBS can be given by

$$R(t) = \sum_{i=1}^N R_i(t). \quad (4.3)$$

4.2.2 Energy Harvesting and Primary User Models

Herein, we assume that the SUs always harvest energy during the whole of time slot T , and the amount of harvested energy is stored in their finite capacity batteries. Since the SUs perform the energy harvesting process in the same environment, it is also worth noting that they have the same distribution. The amount of harvested energy, $e^{hv,i}$, of SU_i in each time slot follows a Poisson distribution process with mean value e_{mean}^{hv} . The value for $e^{hv,i}$ in time slot t can be expressed as $e^{hv,i}(t) = \{e_1^{hv}, e_2^{hv}, e_3^{hv}, \dots, e_\nu^{hv}\}$ where $0 < e_1^{hv} < e_2^{hv} < e_3^{hv} < \dots < e_\nu^{hv} < E_{ca}$. The probability mass function for harvested energy can be given as [123]:

$$p_{hv}(k) = \Pr(e^{hv} = e_k^{hv}) = \frac{e^{-e_{mean}^{hv}} (e_{mean}^{hv})^k}{k!}, k = 1, 2, \dots, \nu. \quad (4.4)$$

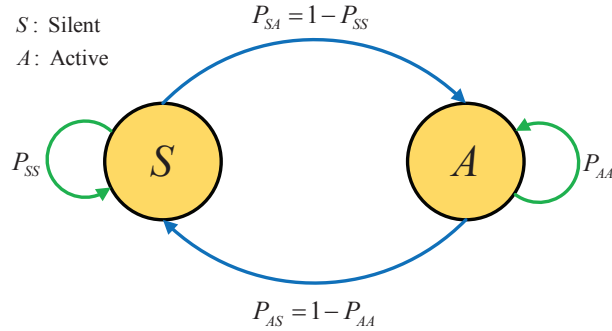


Figure 4.4: Markov chain model of the primary user.

In each time slot, cooperative sensing is performed by the secondary network to predict the state of the PU. At the beginning of each time slot, the PU activity on the licensed channel may switch between silent (S) and active (A) states according to a two-state Markov discrete-time process, which is assumed stationary during the entire time slot, T . The state transition probability for two contiguous time slots is given by $P_{ij} | i, j \in \{S, A\}$ as shown in Fig. 4.4. For example, P_{SA} refers to the probability that the PU transfers from the silent state in the current time slot to the active state in the next time slot.

4.2.3 Imperfect Spectrum Sensing

In the network, the SUs need to perform spectrum sensing in every single time slot to determine the activity of the PU on the primary channel, and then, they report these local sensing decisions to the CBS. The global sensing decision is assumed to be obtained by the soft combination scheme from [122]. However, the sensing engine may induce sensing errors in practice, which results in low transmission performance by the users. Accordingly, we consider the imperfect spectrum-sensing model for the CR network. The sensing performance can be evaluated principally by two probabilities: a detection probability P_d and false alarm probability P_f , which are defined as

$$P_d = \Pr(H_A(t) = A | A) \text{ and } P_f = \Pr(H_A(t) = A | S), \quad (4.5)$$

respectively. P_d represents the probability that the PU is correctly found to be active, whereas P_f is the probability that the PU is found active but is actually silent. $H_A(t)$ denotes the state of the PU (i.e. the global sensing decision at the CBS) in time slot t . As such, the value of P_d is set according to the maximum acceptable probability that collisions

between the secondary transmission and the primary transmission can happen [78, 124]. Besides, we further assume in this chapter that the value of P_d and P_f are available to the CBS.

4.3 Problem Formulation

The objective of this chapter is to enhance the long-term throughput of uplink NOMA at the CBS. The power allocation problem for throughput maximization of an uplink NOMA system in time slot t can be formulated as follows:

$$\begin{aligned} \arg \max_{\mathbf{A}(t) \in \mathbb{A}} & \sum_{k=t}^{\infty} \sum_{i=1}^N R_i(k) \\ \text{s.t.} & 0 \leq e_i^{tr}(t) \leq e^{tr, \max} \end{aligned}, \quad (4.6)$$

where $\mathbf{A}(t) = \begin{bmatrix} \mathbf{a}(t) \\ \mathbf{e}^{tr}(t) \end{bmatrix}$ is the global action that the CBS assigns to the SUs in time slot t , $\mathbf{a}(t)$ and $\mathbf{e}^{tr}(t)$ represent the assigned action mode vector and the assigned transmission energy vector for the SUs, respectively. The assigned action mode, and the transmission energy for the SUs are described in the row vectors with the same dimension. The index of each element in these vectors represents the index of the corresponding SU. Particularly, $\mathbf{a}(t) = [a_1(t), a_2(t), \dots, a_N(t)]$ includes the assigned actions of all SUs in time slot t , where $a_i(t) = \{“SL”, “TM”\}$ denotes the different action modes for SU_i , in which SL and TM stand for silent mode and transmission mode, respectively. Meanwhile, $\mathbf{e}^{tr}(t) = [e_1^{tr}; e_2^{tr}; \dots; e_N^{tr}]$ represents the assigned amount of transmission energy for the SUs in time slot t , where $e_i^{tr}(t) \in \{0, e^{tr,1}, e^{tr,2}, \dots, e^{tr,\psi}\}$ denotes the transmission energy of SU_i , in which $e^{tr,j} | j \in \{1, 2, \dots, \psi\}$ represents the transmission energy level. $e^{tr, \max}$ is the maximum transmission power at the SUs. The constraint in Eq. (4.6) is to guarantee that the assigned transmission power at each SU should not exceed the value of $e^{tr, \max}$. In the next section, we propose an actor-critic reinforcement learning approach to maximizing the overall reward from uplink NOMA. In particular, at the start of time slot t , the CBS will determine the most appropriate action (i.e. silent mode or transmission mode with different transmission energy levels) for each SU based on the remaining energy in each of the SUs and the belief that the PU will be inactive in the current time slot. The actor-critic framework will learn

and interact directly with the environment to obtain the optimal solution for the problem in Eq. (4.6) after a large enough number of time slots.

4.4 Actor–Critic Reinforcement Learning–Based Algorithm for Uplink NOMA in Cognitive Radio Networks

In this section, the actor–critic reinforcement learning approach is presented with the goal of allocating the optimal action for the SUs such that the maximum long-term throughput of uplink NOMA can be achieved according to information directly collected via practical interactions with the environment. If an SU does not have enough energy for data transmission in a time slot, it has to stay silent to save energy for the next time slot regardless of the active or inactive states of the primary user. If the channel is sensed as active, the SUs have to stay silent; otherwise, they will be assigned to concurrently transmit data with the corresponding amount of transmission energy on the channel, which is described in the following subsection.

4.4.1 Markov Decision Process

The actor–critic approach is a type of MDP [125], that can be defined as a quintuple $\langle \mathbb{S}, \mathbb{A}, \mathbb{P}, \mathbb{R}, \gamma \rangle$, where \mathbb{S} is the state space, \mathbb{A} represents the action space set, \mathbb{P} is the transition probability set in which the state of the agent changes from the current state to the next state when action \mathbf{A} is taken, \mathbb{R} is the reward space, and $\gamma \in [0, 1)$ denotes the discount factor.

- *State space*: The state of the network in time slot t is $s(t) = (\mu(t), \mathbf{e}^{re}(t))$, where $\mu(t)$ is the belief representing the probability that the PU is idle in this time slot, and $\mathbf{e}^{re}(t) = [e_1^{re}(t), e_2^{re}(t), \dots, e_N^{re}(t)]$ is a vector that includes the remaining energy of the SUs at the beginning of time slot t .
- *Action space*: The CBS assigns global action $\mathbf{A}(t)$, which comprises two vectors: $\mathbf{a}(t) = [a_1(t), a_2(t), \dots, a_N(t)]$ and $\mathbf{e}^{tr}(t) = [e_1^{tr}, e_2^{tr}, \dots, e_N^{tr}]$. Note that each element of these vectors is sorted by following the corresponding index of each SU in the network.
- *Reward*: Given the state of the system, $s(t)$, and the action, $\mathbf{A}(t)$, each SU performs

its own assigned action. As a result, the system can obtain an immediate reward which is defined as the summation of the SUs' throughput in the current time slot: $R(\mathbf{A}(t), s(t))$.

The state-value function $V(s(t))$ is the cumulative discounted reward from current state $s(t)$. If the CBS uses the policy, $\pi_t(\mathbf{A}(t) | s(t))$, the probability that the CBS will assign action $\mathbf{A}(t)$ for given state $s(t)$, then the state-value function, $V(s(t))$, can be expressed as follows:

$$V(s(t)) = \sum_{k=t}^{\infty} \gamma^{k-t} R(s(k), \mathbf{A}(k)). \quad (4.7)$$

The objective of the actor-critic reinforcement learning algorithm is to find an optimal policy $\pi_t^*(\mathbf{A}(t) | s(t))$ that maximizes the state-value function of each state $s(t)$ defined by Eq. (4.7). The optimal policy can be described by

$$\pi_t^*(\mathbf{A}(t) | s(t)) = \arg \max_{\mathbf{A}(t) \in \mathbb{A}} \left\{ \sum_{k=t}^{\infty} \gamma^{k-t} R(s(k), \mathbf{A}(k) | s(k) = s) \right\}. \quad (4.8)$$

4.4.2 Actor-Critic Reinforcement Learning Algorithm

In this chapter, we present the actor-critic approach as a model-free reinforcement learning framework to solve the MDP problem. The advantage of this algorithm is that it does not require any prior information from the dynamic environment (i.e. the harvested energy distribution). That is, it is worthwhile to utilize the actor-critic scheme in practice from the viewpoint that prior information from the environment is not easy to acquire. On the other hand, the system can directly interact with the environment to learn the information about the harvested energy.

Fig. 4.5 depicts the flowchart of the proposed scheme based on the actor-critic learning method. In a time slot, after cooperative spectrum sensing is executed, the CBS makes the global sensing decision about the existence of the PU on the licensed channel. If the global sensing decision indicates that the PU is active, the SUs accept this result and stay silent. Then, the SUs send to the CBS an update on their remaining energy, and the belief $\mu(t+1)$ can be calculated at the end of time slot t . Otherwise, if the global

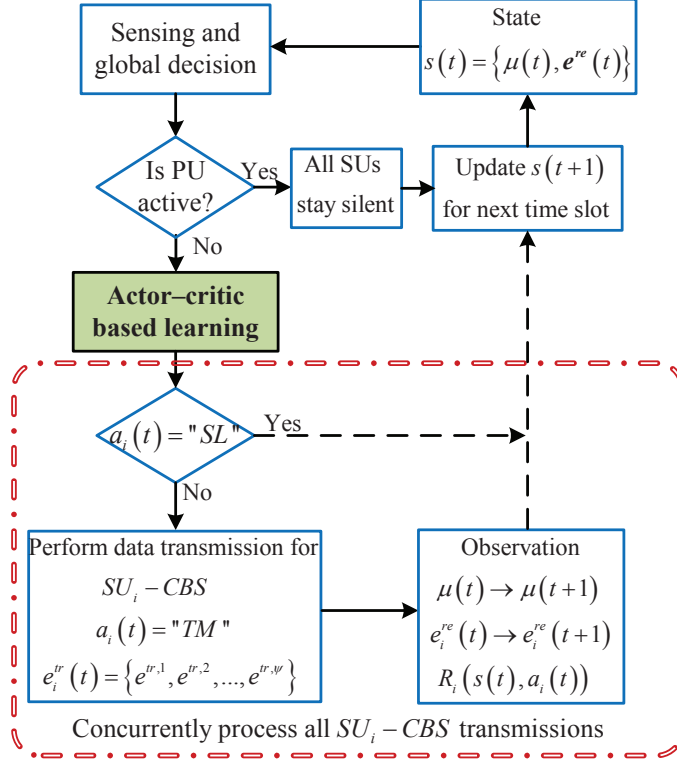


Figure 4.5: The flowchart of the proposed scheme.

sensing decision indicates the PU is inactive, the CBS can choose the possible actions from set \mathbb{A} by applying the proposed actor-critic learning approach. The belief can be updated by observing the successful/unsuccessful transmissions of the SUs if they are assigned to transmit data to the CBS.

The actor-critic learning process consists of two components (the *actor* and the *critic*), as shown in Fig. 4.6. The *actor* is used to define the policy and generate actions based on the observed environment state, while the *critic* learns the state-value function and criticizes the actions selected by the actor. At the start of each time slot, the actor employs an action, $\mathbf{A}(t) \in \mathbb{A}$, by following policy $\pi_t(\mathbf{A}(t) | s(t))$. The policy is calculated via Gibbs soft-max distribution [126]:

$$\pi_t(\mathbf{A}(t) | s(t)) = \frac{e^{h(s(t), \mathbf{A}(t))}}{\sum_{\mathbf{A} \in \mathbb{A}} e^{h(s(t), \mathbf{A})}}, \quad (4.9)$$

where $h(s(t), \mathbf{A}(t))$ indicates the tendency to select action $\mathbf{A}(t)$ in state $s(t)$.

At the end of a time slot, the system will update the immediate reward, $R(s(t), \mathbf{A}(t))$,

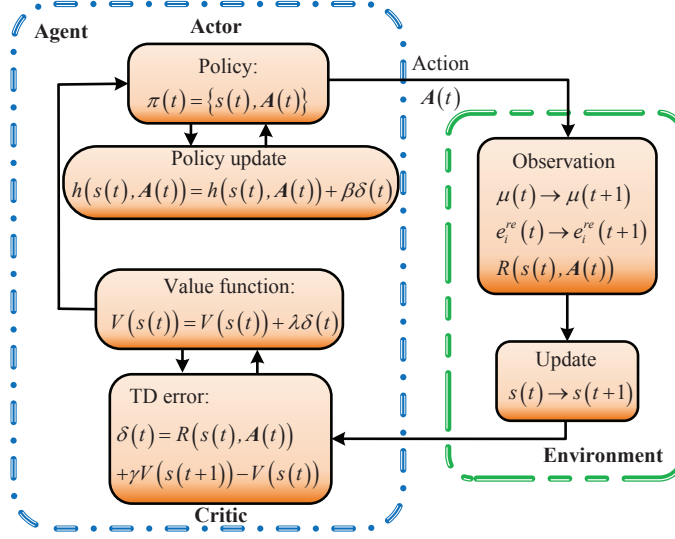


Figure 4.6: The actor–critic learning process.

and the next state $s(t+1)$. Subsequently, the critic will criticize the selected action and evaluate the policy by using the temporal difference (TD) error, which is computed as

$$\delta(t) = R(s(t), \mathbf{A}(t)) + \gamma V(s(t+1)) - V(s(t)), \quad (4.10)$$

where $\delta(t)$ denotes the difference between state–value function $V(s(t))$ from the preceding state and the state–value function after taking the selected action. Then, the critic uses the TD error to criticize the selected action as follows:

$$V(s(t)) = V(s(t)) + \lambda \delta(t), \quad (4.11)$$

where λ is the learning step-size of the critic. Thereafter, the TD error is fed back to the actor, and the tendency to select the action is upgraded as

$$h(s(t), \mathbf{A}(t)) = h(s(t), \mathbf{A}(t)) + \beta \delta(t), \quad (4.12)$$

where β is the learning step-size of the actor. Ultimately, the policy is updated by Eq. (4.9) and Eq. (4.12) for action selection in the subsequent time slots. The training process will be completed when state–value function $V(s(t))$ and policy $\pi_t(\mathbf{A}(t)|s(t))$ converge to $V^*(s(t))$ and $\pi_t^*(\mathbf{A}(t)|s(t))$ with probability 1 as $t \rightarrow \infty$ [127].

Hereafter, when the CBS assigns an action for each SU, one of the following observations may happen.

4.4.2.1 Silent Mode (Ω_1)

If the global sensing decision indicates that the PU is active in the current time slot. The CBS will trust this result and assign action SL to all SUs. In this case, no reward is achieved, i.e. $R(s(t), \mathbf{A}(t) | \Omega_1) = 0$. The belief in the current time slot can be updated using Bayes' rule [128] as follows:

$$\mu(t) = \frac{\mu(t) P_f}{\mu(t) P_f + (1 - \mu(t)) P_d}. \quad (4.13)$$

The updated belief for the next time slot is given as

$$\mu(t+1) = \mu(t) P_{SS} + (1 - \mu(t)) P_{AS}. \quad (4.14)$$

For simplicity in this work, we assume that the energy consumed for the information update of the SUs is tiny and can be ignored. Hence, the remaining energy of SU_i for the next time slot can be calculated as follows:

$$e_i^{re}(t+1) = \min \left\{ e_i^{re}(t) + e^{hv,i}(t) - e_{ss}, E_{ca} \right\}, \quad (4.15)$$

where e_{ss} denotes the energy consumed for spectrum sensing.

4.4.2.2 Transmission Mode

If the global sensing decision indicates that the PU is silent, then the CBS allows all SUs to transmit data with $a_i(t) = TM$ and the corresponding transmission energy level $e_i^{tr}(t)$. In this case, there are two observations: Ω_2 and Ω_3 .

Observation 2 (Ω_2): The CBS can successfully decode the signals transmitted by the SUs at the end of the time slot. In this case, the system recognizes that the PU was actually silent in the time slot. The reward can be computed as

$$R(s(t), \mathbf{A}(t) | \Omega_2) = \sum_{i=1}^N R_i(t), \quad (4.16)$$

where the throughput of the SU_i , $R_i(t)$ can be calculated with Eq. (4.2). Belief $\mu(t+1)$ for the next time slot can be updated as

$$\mu(t+1) = P_{SS}. \quad (4.17)$$

The remaining energy of SU_i can be updated as follows:

$$e_i^{re}(t+1) = \min \left\{ e_i^{re}(t) + e^{hv,i}(t) - e_{ss} - e_i^{tr}(t), E_{ca} \right\}, \quad (4.18)$$

Algorithm 4.1 Actor–critic reinforcement learning procedure of the transmission power decision policy for the SUs

- 1: **Input:** $\mathbb{S}, \mathbb{A}, \gamma, \lambda, \beta, \mathbf{e}^{re}(t), \mu(t), E_{ca}, e_{mean}^{hv}, T$. // Initial parameters
 - 2: Initialize state–value function $V(s(t))$, tendency $h(s(t), \mathbf{A}(t))$, and policy $\pi_t(\mathbf{A}(t) | s(t))$;
 - 3: **Repeat until convergence**
 - 4: **for** each time slot,
 - 5: Define the current state $s(t) \in \mathbb{S}$
 - 6: Choose an action, $\mathbf{A}(t) \in \mathbb{A}$, according to policy $\pi_t(\mathbf{A}(t) | s(t))$ in Eq. (4.9) after considering the sensing result and the remaining energy of the SUs.
 - 7:
 - 8: **Simultaneously excute the process for all SUs:**
 - 9: **if** $a_i(t) = \text{“SL”}$ // if SU_i is assigned to stay silent
 - 10: SU_i stays silent and only harvests solar energy.
 - 11: **else**// if SU_i is assigned to transmit data with the transmission energy $e_i^{tr}(t)$
 - 12: SU_i transmits data to CBS with assigned transmission energy and harvests solar energy.
 - 13: **end if**
 - 14: Compute instant reward $R_i(s(t), a_i(t))$; update network state $s(t+1)$.
 - 15:
 - 16: **Critic Process:**
 - 17: Calculate TD error $\delta(t)$ with Eq. (4.10).
 - 18: Update state–value function $V(s(t))$ with Eq. (4.11).
 - 19:
 - 20: **Actor Process:**
 - 21: Update tendency to select an action, $\mathbf{A}(t), h(s(t), \mathbf{A}(t))$, with Eq. (4.12).
 - 22: Update policy to choose action $\mathbf{A}(t)$ under the given state, $\pi_t(\mathbf{A}(t) | s(t))$, with Eq. (4.9).
 - 23: **end for**
 - 24: **Output:** Final policy $\pi_t^*(\mathbf{A}(t) | s(t))$. // Optimal action at given state.
-

where $e_i^{tr}(t)$ denotes the transmission energy of SU_i in time slot t .

Observation 3 (Ω_3): The CBS can not successfully decode the signals transmitted by the SUs due to collisions between the SUs and PUs transmissions. The system can infer

that misdetection occurred in this case. There is no reward: $R(s(t), \mathbf{A}(t) | \Omega_3) = 0$.

Belief $\mu(t+1)$ for the next time slot is updated as

$$\mu(t+1) = P_{AS}. \quad (4.19)$$

The remaining energy of SU_i for the next time slot is updated by:

$$e_i^{re}(t+1) = \min \left\{ e_i^{re}(t) + e^{hv,i}(t) - e_{ss} - e_i^{tr}(t), E_{ca} \right\}. \quad (4.20)$$

In the actor-critic algorithm, the state-value function and the policy parameters are sequentially and concurrently updated based on the action of the CBS over the time slots. The policy of the system can be dynamically obtained from a practical learning process, such that the local optimal policy can converge over a large number of time slots [129]. Finally, we summarize the learning process of the proposed actor-critic scheme in **Algorithm 4.1**.

4.5 Simulation Results

In this section, we analyze the performances of the proposed scheme under various conditions of the network by using simulation results based on MATLAB R2019a. In addition, we also compare the proposed scheme with the other conventional schemes, such as the Myopic NOMA scheme, the Myopic OMA scheme, and the Myopic Random scheme, where the term “Myopic” refers to the policy that only maximizes the instant reward of the system. In the Myopic NOMA scheme, if the sensing result indicates the absence of the PU, the SUs will simultaneously transmit data to the CBS at the highest transmission energy level, and then, similar to the proposed scheme the received signals will be decoded at the CBS by applying the SIC technique. For the Myopic OMA scheme, the decision is made by combining the myopic approach and TDMA-based technique. More specifically, the entire data transmission phase in a time slot is equally divided into sub-phases according to the number of SUs. After that, the SUs transmit their information in rapid succession during their respective sub-phases, one after the other. For the simulation ($N = 2$), we assumed each SU transmits data in half of the time for the transmission phase after sensing, and then, the achievable throughput of SU_i at the CBS can be computed as [130]

$$R_i^{OMA}(t) = \frac{\tau_{tr}}{2T} \log_2 \left(1 + \frac{P_i(t) |h_i|^2}{\sigma_\omega^2} \right), \quad (4.21)$$

Table 4.1: Simulation Parameters

Parameter	Description	Value
N	Number of SUs	2
T	Time slot duration	200 ms
τ_{ss}	Sensing duration	2 ms
τ_u	Update duration	1 ms
E_{ca}	Battery capacity	20 μJ
e_{ss}	Sensing cost	1 μJ
e^{tr}	Transmission energy	5, 10, 15 μJ
e_{mean}^{hv}	Mean value of harvested energy	6 μJ
μ	Initial belief that the PU is absent	0.5
P_{SS}	Transition probability of the PU from state S to itself	0.8
P_{AS}	Transition probability of the PU from state A to state S	0.2
P_d	Probability of detection	0.9
P_f	Probability of false alarm	0.1
h_1	Channel gain between SU_1 and the CBS	-20 dB
h_2	Channel gain between SU_2 and the CBS	-35 dB
σ_ω^2	Noise variance	-80 dB
γ	Discount factor	0.95
λ, β	Learning step-sizes of critic, actor	0.1, 0.1

In the Myopic Random scheme, the CBS randomly assigns NOMA/OMA to the SUs when the global sensing decision indicates that the PU is silent in the current time slot. For simplicity, the channel gain for each SU is fixed, and there are three levels for the transmission energy of the SUs: TM1 = 5 μJ , TM2 = 10 μJ , TM3 = 15 μJ . The span of each belief is 0.1. Furthermore, the performance of the proposed scheme was verified over 30,000 time slots, and the results were acquired by averaging 10 separate loops. **Table 4.1** shows the simulation parameters for the scheme proposed in this chapter.

In Fig. 4.7, we examine the convergence of the proposed scheme's algorithm over time slots with various step-size parameters, λ and β , based on the reward (throughput) of

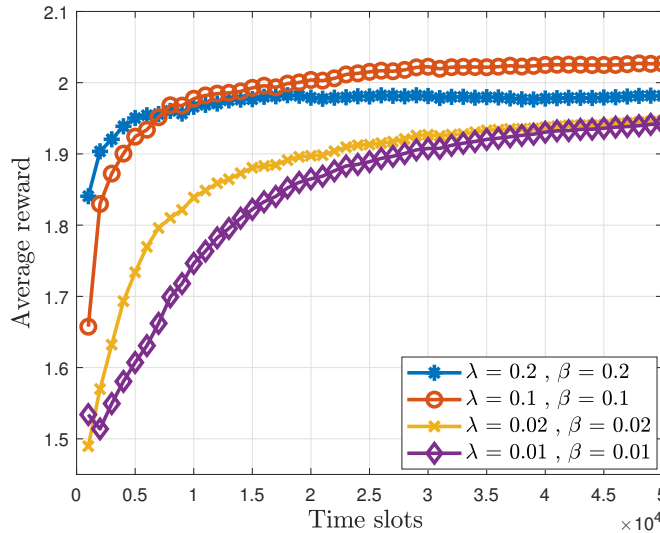


Figure 4.7: The convergence process of the actor–critic according to different values of learning step-size.

the system. In this chapter, the convergence condition for the proposed scheme was set at 10^{-3} . We can see that the system reward greatly increases during the first 15,000 time slots, and then gradually converges to the optimal value, which depends on the different values of λ and β . Obviously, if larger values of λ and β are used, the faster the convergence and the higher the throughput. However, from the figure, we can see that increasing the value of λ and β does not always guarantee a higher reward for the network due to underfitting, whereas the system might be prone to overfitting if we reduce the learning parameters. As a result, we set the value of actor and critic learning step-sizes as $\lambda = 0.1$, and $\beta = 0.1$, respectively, for the proposed scheme in the upcoming simulation.

In Fig. 4.8, we illustrate the effect of the amount of harvested energy of the SUs and the number of primary channels on the average throughput of the proposed actor–critic NOMA, compared with the other conventional schemes. We can see that when e_{mean}^{hv} increases, the SUs can collect more energy from the solar source, and can transmit at a higher transmission energy level, which leads to higher achievable throughput at the CBS. In addition, the performance of the proposed scheme outperforms the conventional schemes, since the conventional schemes disregard the impact of the current decision on future rewards. For that reason, whenever the PU is sensed as silent on the primary channel, then these

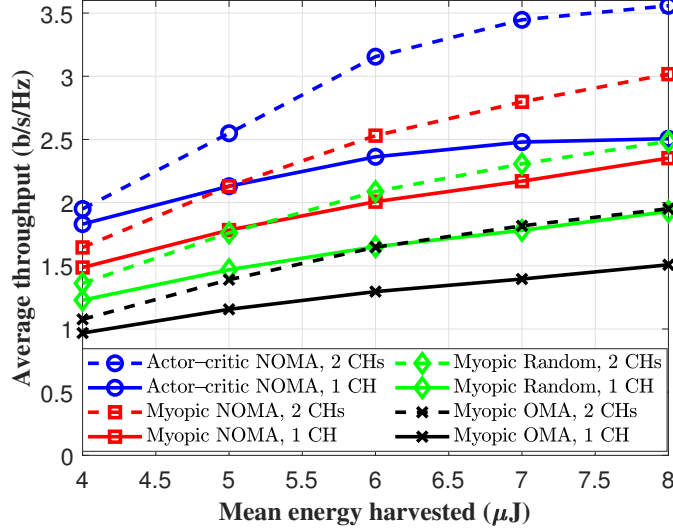


Figure 4.8: Average throughput of the secondary system under various values of harvested energy.

conventional schemes will allow SUs to use as much energy as possible to maximize the immediate throughput at the CBS. However, this action causes the SUs to stay silent longer than under the proposed actor-critic NOMA scheme owing to the limitations on battery capacity and harvested energy. It is also observed that average throughput achieved at the CBS in the case of two primary channels (2 CHs) is larger than that of a single channel case (1 CH). Obviously, with more primary channels, the SUs have more chances to transmit their data. As a consequence, the performance of the cognitive radio system is enhanced in the case of multiple channels. However, the proposed scheme provides the highest throughput in both cases of single and multiple channels.

Fig. 4.9 shows the energy efficiency of the system under various mean values of energy harvesting. In this chapter, we define energy efficiency as the average long-term throughput over the total energy-harvesting amount during the operations spanning M ($M = 20,000$) time slots $\left(EE = \frac{\sum_{t=1}^M \sum_{i=1}^N R_i(t)}{\sum_{t=1}^M \sum_{i=1}^N e^{hv,i}(t)} \right)$. In order to enhance the energy efficiency of the proposed scheme, we set the maximum transmission energy level for the SU if its battery is likely to overflow in each time slot [131]. From Fig. 4.9, we can see that the energy efficiency drops with increased levels of energy harvesting, e_{mean}^{hv} . The reason is that when e_{mean}^{hv} increases, the SUs can gather more energy for their operations but the total

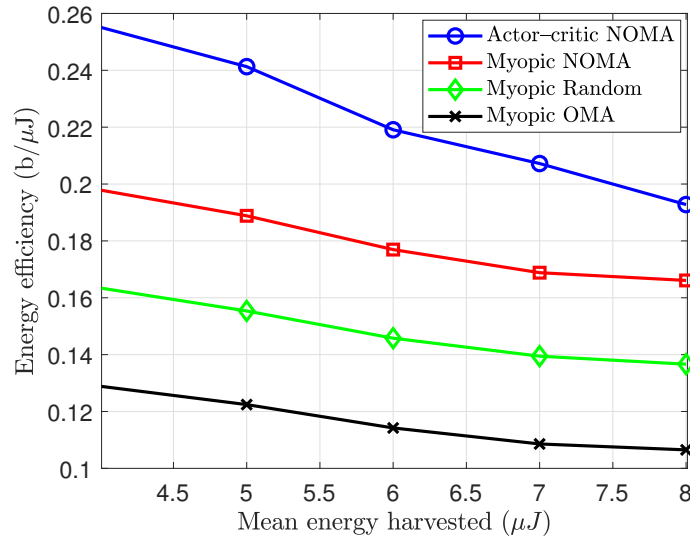


Figure 4.9: Energy efficiency of the secondary system for various values of harvested energy.

amount of overflow energy in the SUs increases concurrently. As a consequence, the figure shows that the proposed scheme is still superior to other conventional schemes under different amounts of harvested energy. For instance, when $e_{mean}^{hv} = 6\mu J$, the energy efficiency of the proposed scheme can provide 23.8%, 50.3%, and 91.8% in energy utilization improvement for the Myopic NOMA, Myopic Random, and Myopic OMA schemes, respectively. Myopic OMA brings the lowest result, because the SUs transmit data in turn during each half-phase of the data transmission duration, while other NOMA schemes allow the SUs to simultaneously transmit data over the entire data transmission phase.

Specific information about the number of actions selected for each SU under a change in e_{mean}^{hv} for the proposed scheme and the Myopic NOMA scheme is illustrated in Fig. 4.10 and Fig. 4.11, respectively. We can see that the proposed scheme normally assigns the proper amount of transmission energy for the SUs at low values for the harvested energy mean, e_{mean}^{hv} . Meanwhile Myopic NOMA scheme assigns the SUs the highest possible transmission energy at all values of e_{mean}^{hv} . This creates inefficiency in terms of both energy and throughput metrics, as presented in Fig. 4.8 and Fig. 4.9. It is obvious that although harvested energy may vary over time slots, the SUs in the proposed scheme usually utilize TM3 to obtain the highest throughput, provided that the PU is most likely absent, or energy overflow might happen.

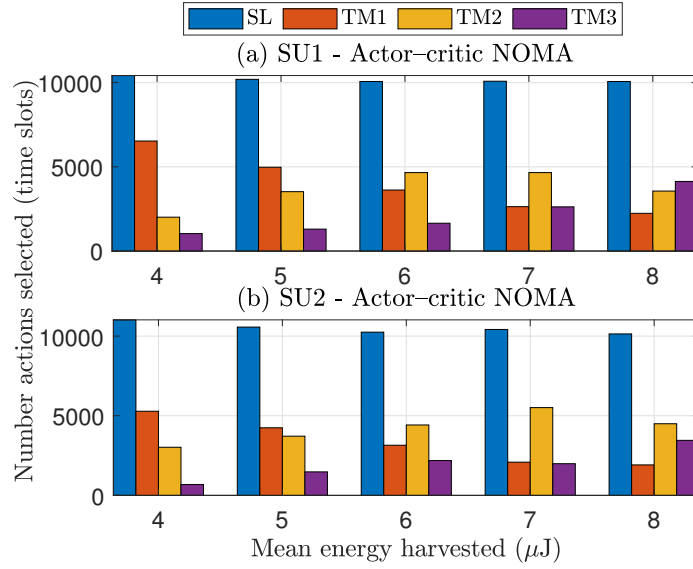


Figure 4.10: The selected action statistics of each secondary user using the actor-critic NOMA approach for various values of harvested energy.

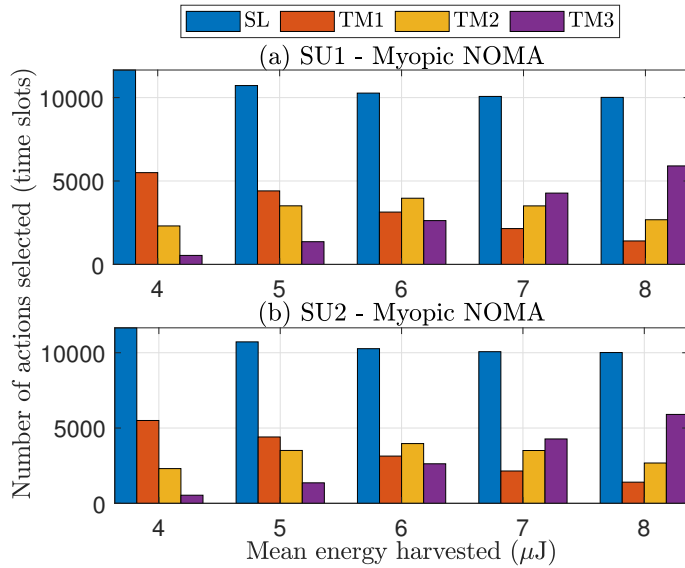


Figure 4.11: The selected action statistics of each secondary user using the Myopic NOMA approach for various values of harvested energy.

We further investigated the joint impact of channel gain between the CBS and SUs' throughput in the system, as shown in Fig. 4.12. It is evident that the performance of the system goes up with an increase in channel gain h_1 and h_2 . The reason is that the

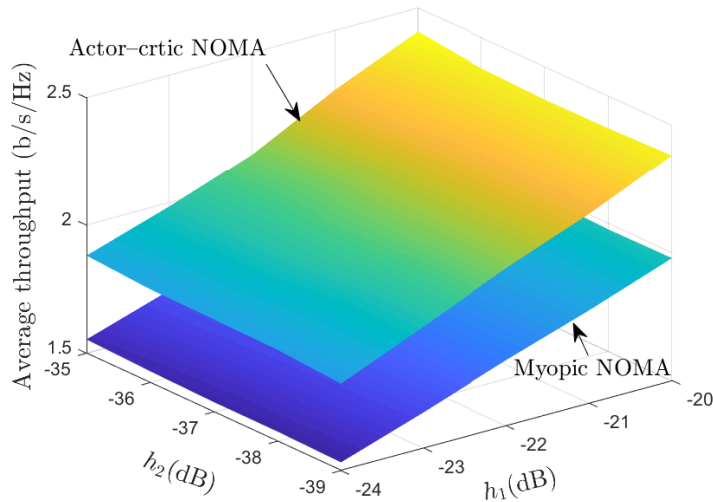


Figure 4.12: Average throughput for different values of h_1 and h_2 .

throughput of system throughput is dependent on the channel gain as shown in Eq. (2) and Eq. (3). Thus, the larger channel gains are, the higher throughput the system obtains. Specifically, when h_1 becomes larger, the average throughput of the system significantly increases; however, it only slightly increases when h_2 increases. That is because the value of h_2 is quite a bit smaller than the value of h_1 , and in this case, it has less impact on the signal of SU_1 and on total throughput of the system. Thus, increasing h_2 does not much influence the overall obtainable throughput at the CBS due to its small channel gain.

In Fig. 4.13, we investigate the energy efficiency of the proposed scheme versus the various values for channel gain between the CBS and SU_1 . The curves show that the energy efficiency of the system benefits from larger values of h_1 because with the same transmission power for SUs, the higher channel gain will bring more throughput at the CBS. Consequently, the proposed scheme is verified to be superior to other conventional schemes in terms of efficient energy utilization under the variation of channel gain.

In Fig. 4.14, we jointly study average system throughput of the schemes under the impact of various noise variances σ_w^2 and primary user activity which is expressed as transition probability of PUs from state *silent* to state *silent*. Fig. 14 shows that a large amount of noise variance can significantly degrade the obtained throughput. It can be explained as following: when the noise variance goes up, it will severely interfere with the

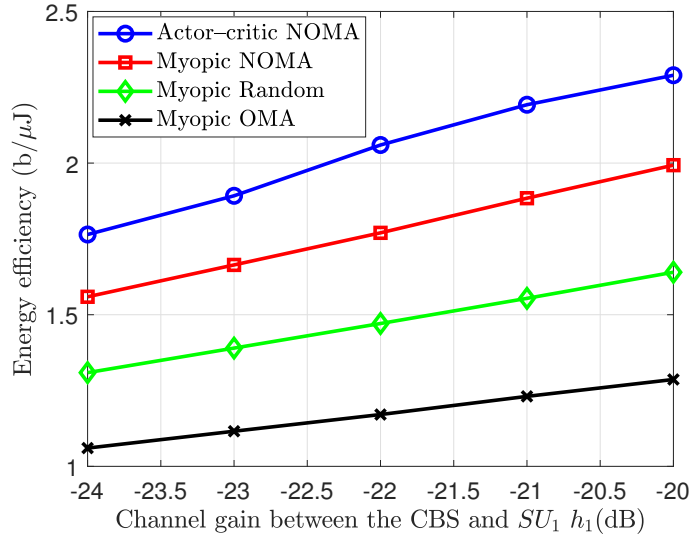


Figure 4.13: Energy efficiency according to the channel gain between the CBS and SU_1 .

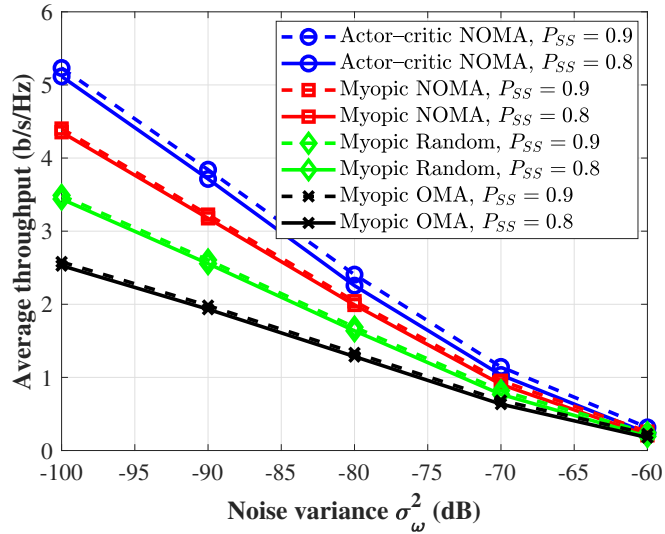


Figure 4.14: Average throughput according to the noise variance.

received signals at the CBS. According to Eq. (2) and Eq. (21), the noise variance, which is an interfering component in the denominator of the signal-to-interference-plus-noise ratio (SINR), will lower the system throughput as it increases, and vice versa. In addition, we can see that the performances of all schemes can get better as P_{SS} increases. The reason is that when the transition probability of PUs from state *silent* to itself rises, the probability that

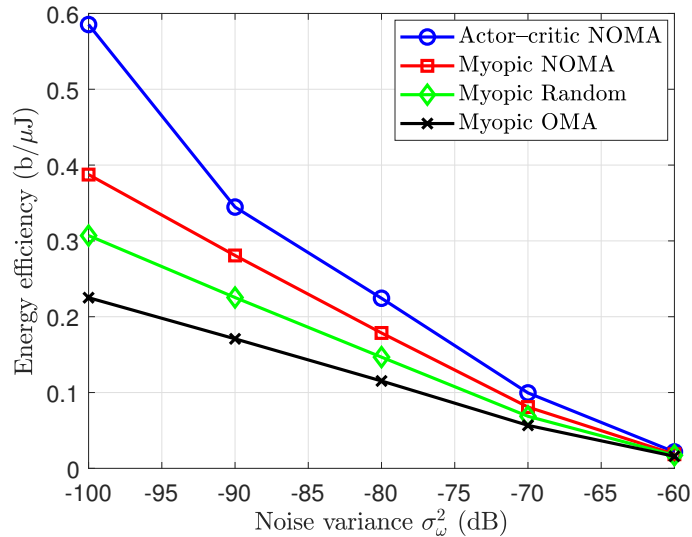


Figure 4.15: Energy efficiency according to the noise variance.

the primary channel is free also goes up, which results in more opportunities for the SUs to transmit data on the primary channel.

Finally, in Fig. 4.15, we examine the effect of noise variance on the energy efficiency of the schemes. It is observed that the energy efficiency of all schemes deteriorates as the noise variance increases. The reason for this is as following: the energy efficiency is calculated by the achieved long-term throughput over the total energy-harvesting during operational time. Hence, for the large value of the noise variance, the average long-term throughput tends to be reduced, which results in the low energy efficiency in the system. The figure shows that the energy efficiency of the proposed schemes dominates the other conventional schemes. Furthermore, the figure can verify the robustness of the proposed scheme with respect to noise variation at the CBS.

4.6 Conclusion

In this chapter, we propose an actor-critic reinforcement learning approach using uplink NOMA in a cognitive radio network. In the network, the solar energy-powered SUs can simultaneously transmit data to a cognitive base station. The energy-constrained and the imperfect-sensing problems are also taken into account. Consequently, the optimal policy can be obtained by using the proposed scheme, where the SUs can be assigned the

proper action mode (i.e. stay silent or transmit data) to maximize the long-term throughput of the secondary system. Simulation results demonstrate that the proposed scheme can improve both long-term throughput and the energy efficiency of the network, compared with conventional schemes.

Chapter 5

Hybrid NOMA/OMA-Based Dynamic Power Allocation Scheme Using Deep Reinforcement Learning in 5G Networks

5.1 Introduction

Recently, fourth-generation (4G) systems reached maturity, and will evolve into fifth-generation (5G) systems where limited amounts of new spectrum can be utilized to meet the stringent demands of users. However, critical challenges will come from explosive growth in devices and data volumes, which require more efficient exploitation of valuable spectrum. Therefore, non-orthogonal multiple access (NOMA) is one of the potential candidates for 5G and upcoming cellular network generations [91,92,132].

According to NOMA principles, multiple users are allowed to share time and spectrum resources in the same spatial layer via power-domain multiplexing, in contrast to conventional orthogonal multiple access (OMA) techniques consisting of frequency-division multiple access (FDMA) and time division multiple access (TDMA) [133]. Interuser interference can be alleviated by performing successive interference cancellation (SIC) on the receiver side. There has been a lot of research aimed at sum rate maximization, and the results showed that higher spectral efficiency (SE) can be obtained by using NOMA,

compared to baseline OMA schemes [134–137]. Zeng et al. [134] investigated a multiple-user scenario in which users are clustered and share the same transmit beamforming vector. Di et al. [135] proposed a joint sub-channel assignment and power allocation scheme to maximize the weighted total sum rate of the system while adhering to a user fairness constraint. Timotheou et al. [136] studied a decoupled problem of user clustering and power allocation in NOMA systems in which the proposed user clustering approach is based on exhaustive search with a high required complexity. Liang et al. [137] studied solutions for user pairing, and investigated the power allocation problem by using NOMA in cognitive radio (CR) networks.

Nowadays, energy consumption for wireless communications is becoming a major social and economic issue, especially with the explosive amounts of data traffic. However, limited efforts have been devoted to the energy-efficient resource allocation problem in NOMA-enabled systems [138–140]. The authors in [138] maximized energy efficiency subject to a minimum required data rate for each user, which leads to a nonconvex fractional programming problem. Meanwhile, a power allocation solution aiming to maximize the energy efficiency under users' quality of service requirements was investigated [139]. Fang et al. [140] proposed a gradient-based binary search power allocation approach for downlink NOMA systems, but it requires high complexity. NOMA was also applied to future machine-to-machine (M2M) communications in [141], and it was shown that the outage probability of the system can be improved when compared with OMA. Additionally, by jointly studying beamforming, user scheduling, and power allocation, the system performance of millimeter wave (mmWave) networks was studied [142].

On the other hand, CR (one of the promising techniques to improve SE), has been extensively investigated for decades. In it, cognitive users (CUs) can utilize the licensed spectrum bands of the primary users (PUs) as long as the interference caused by the CUs is tolerable [143–145]. Goldsmith et al. in [89] proposed three operation models (opportunistic spectrum access, spectrum sharing, and sensing-based enhanced spectrum sharing) to exploit the CR technique in practice. It is conceivable that the combination of CR with NOMA technologies is capable of further boosting the SE in wireless communication systems. Therefore, many studies on the performance of spectrum-sharing CR combined with NOMA have been analyzed [104, 105].

Along with the rapid proliferation of wireless communication applications, most battery-limited devices become useless if their battery power is depleted. As one of the

remedies, energy harvesting (EH) exploits ambient energy resources to replenish batteries, such as solar energy [123], radio frequency (RF) signals [146], and both non-RF and RF energy harvesting [90], etc. Among various kinds of renewable energy resources, solar power has been considered one of the most effective energy sources for wireless devices. However, solar power density highly depends on the environment conditions, and may vary over time. Thus, it is critical to establish proper approaches to efficiently utilize harvested energy for wireless communication systems.

Many early studies regarding NOMA applications have mainly focused on the downlink scenario. However, there are fewer contributions investigating uplink NOMA, where an evolved NodeB (eNB) has to receive different levels of transmitted power from all user devices using NOMA. Zhang et al. in [147] proposed a novel power control scheme, and the outage probability of the system was derived. Besides, the user-pairing approach was studied in many predefined power allocation schemes in NOMA communication systems [148] in which internet of things (IoT) devices first harvest energy from BS transmissions in the harvesting phase, and they then utilize the harvested energy to perform data transmissions using the NOMA technique during the transmission phase. The pricing and bandwidth allocation problem in terms of energy efficiency in heterogeneous networks was investigated in [149]. In addition, the authors in [123] proposed joint resource allocation and transmission mode selection to maximize the secrecy rate in cognitive radio networks. Nevertheless, most of the existing work on resource allocation assumes that the amount of harvested energy is known, or that traffic loads are predictable, which is hard to obtain in practical wireless networks.

Since the information regarding network dynamics (e.g., harvested energy distribution, primary user's behavior) is sometimes unavailable in the cognitive radio system, researchers usually formulate optimization problems as the framework of a Markov decision process (MDP) [90, 123, 150, 151]. Reinforcement learning is one of the potential approaches to obtaining the optimal solution for an MDP problem by interacting with the environment without having prior information about the network dynamics or without any supervision [152–154]. However, it is a big issue for reinforcement learning to have to deal with large-state-space optimization problems. For this reason, deep reinforcement learning (DRL) is being investigated extensively these days in wireless communication systems where deep neural networks (DNNs) work as function approximators and are utilized to learn the optimal policy [155–157]. Meng et al. proposed a deep reinforcement learning method for a

joint spectrum sensing and power control problem in a cognitive small cell [155]. In addition, deep Q-learning was studied for a wireless gateway that is able to derive the optimal policy to maximize throughput in cognitive radio networks [156]. Zhang et al. [157] proposed an asynchronous advantage, deep actor-critic-based scheme to optimize spectrum sharing efficiency and guarantee the QoS requirements of PUs and CUs.

To the best of our knowledge, there has been little research into resource allocation using deep reinforcement learning under a non-RF energy-harvesting scenario in uplink cognitive radio networks. Thus, we propose a deep actor-critic reinforcement learning framework for efficient joint power and bandwidth allocation by adopting hybrid NOMA/OMA in uplink cognitive radio networks (CRNs). In them, solar energy-powered CUs are assigned the proper transmission power and bandwidth to transmit data to a cognitive base station in every time slot in order to maximize the long-term data transmission rate of the system. Specifically, the main contributions of this chapter are as follows.

- We study a model of a hybrid NOMA/OMA uplink cognitive radio network adopting energy harvesting at the CUs, where solar energy-powered CUs opportunistically use the licensed channel of the primary network to transmit data to a cognitive base station using NOMA/OMA techniques. Beside that, a user-pairing algorithm is adopted such that we can assign orthogonal frequency bands to each NOMA group after pairing. We take power and bandwidth allocation into account such that the transmission power and bandwidth are optimally utilized by each CU under energy constraints and environmental uncertainty. The system is assumed to work on a time-slotted basis.
- We formulate the problem of long-term data transmission rate maximization as the framework of a Markov decision process (MDP), and we obtain the optimal policy by adopting a deep actor-critic reinforcement learning (DACRL) framework under a trial-and-error learning algorithm. More specifically, we use DNNs to approximate the policy function and the value function for the actor and critic components, respectively. As a result, the cognitive base station can allocate the appropriate transmission power and bandwidth to the CUs by directly interacting with the environment, such that the system reward can be maximized in the long run by using the proposed algorithm.
- Lastly, extensive numerical results are provided to assess the proposed algorithm performance through diverse network parameters. The simulation results of the proposed

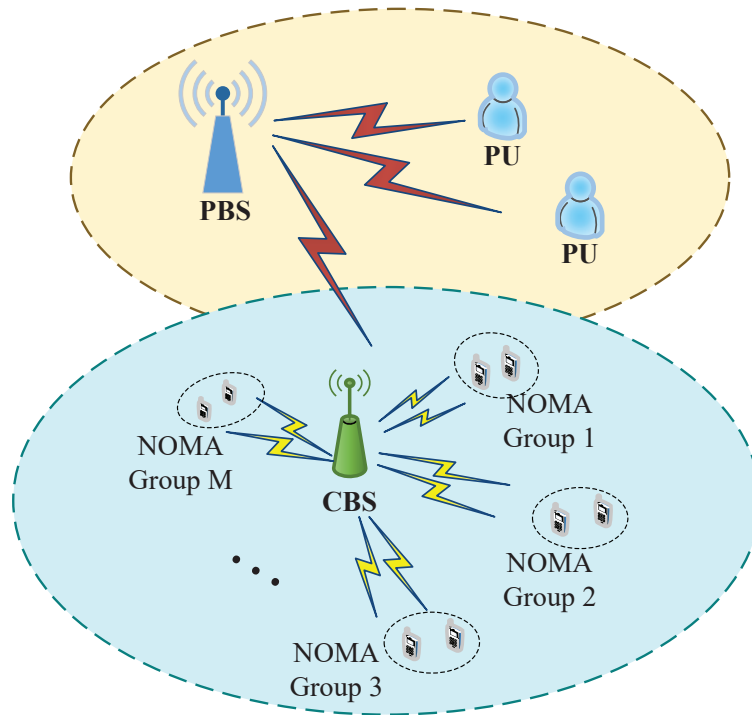


Figure 5.1: System model of the proposed scheme.

scheme are shown to be superior to conventional schemes where decisions on transmission power and bandwidth allocation are taken without long-term considerations.

The rest of this chapter is structured as follows. The system model is presented in Section 5.2. We introduce the problem formulation in Section 5.3, and we describe the deep actor-critic reinforcement learning scheme for resource allocation in Section 5.4. The simulation results and discussions are in Section 5.5. Finally, we conclude the chapter in Section 5.6.

5.2 System Model

We consider an uplink CRN that consists of a cognitive base station (CBS), a primary base station (PBS), multiple primary users, and $2M$ cognitive users as illustrated in Figure 5.1. Each CU is outfitted with a single antenna to transmit data to the CBS, and each is equipped with an energy-harvesting component (i.e., solar panels). The PBS and PUs have the license to use the primary channel at will. However, they do not always

Algorithm 5.1 User-pairing Algorithm

- 1: **Input:** channel gain, number of groups, M , number of CUs, $2M$.
 - 2: Sort the channel gain of all CUs in decending order: $g_1 \geq g_2 \geq \dots \geq g_{2M}$
 - 3: Define set of channel gains $U = \{g_1, g_2, \dots, g_{2M}\}$
 - 4: **for** $j = 1 : M$
 - 5: $G_j = \{\emptyset\}$
 - 6: $G_{\max} = \max \{U\}, G_{\min} = \min \{U\}$
 - 7: $G_j = G_j \cup G_{\max} \cup G_{\min}$
 - 8: $U = U \setminus G_{\max} \setminus G_{\min}$
 - 9: **end for**
 - 10: **Output:** Set of CU pairs.
-

have data to transmit on the primary channel. Meanwhile, the CBS and the CUs can opportunistically utilize the primary channel by adopting a hybrid NOMA/OMA technique when the channel is sensed as free. To this end, the CBS divides the set of CUs into pairs according to **Algorithm 5.1** where the CU having the highest channel gain will be coupled with the CU having the lowest channel gain, and one of available channels will be assigned to these pairs. More specifically, the CUs are arranged into M NOMA groups, and the primary channel is divided into multiple subchannels to apply hybrid NOMA/OMA for the transmissions between the CUs and the CBS, with $\mathcal{G} = \{G_1, G_2, G_3, \dots, G_M\}$ denoting the set of NOMA groups. Additionally, M NOMA groups are assigned to M orthogonal subchannels, $\mathcal{SC} = \{SC_1, SC_2, SC_3, \dots, SC_M\}$, of the primary channel such that the CUs in each NOMA group can transmit on the same subchannel and will not interfere with the other groups. In this chapter, successive interference cancellation (SIC) [158] is applied at the CBS for decoding received signals, which are transmitted from the CUs. Moreover, we assume that the CUs always have data to transmit, and the CBS has complete channel state information (CSI) of all the CUs.

The network system operation is illustrated in Figure 5.2. In particular, at the beginning of a time slot, with duration τ_{ss} , all CUs concurrently perform spectrum sensing and report their local results to the CBS. Based on these sensing results, the CBS first decides the global sensing result as to whether the primary channel is busy or not following the combination rule [122, 159], and then allocates power and bandwidth to all CUs for uplink data transmission. As a consequence, according to the allocated power and bandwidth

of the NOMA groups, the CUs in each NOMA group can transmit their data to the CBS through the same subchannel without causing interference with other groups within duration $\tau_{Tr} = T_{tot} - \tau_{ss}$, where T_{tot} is the total time slot duration. Information regarding the remaining energy in all the CUs is updated to the CBS at the end of each time slot. Each data transmission session of the CUs may take place in more than one time slot until all their data have been transmitted successfully.

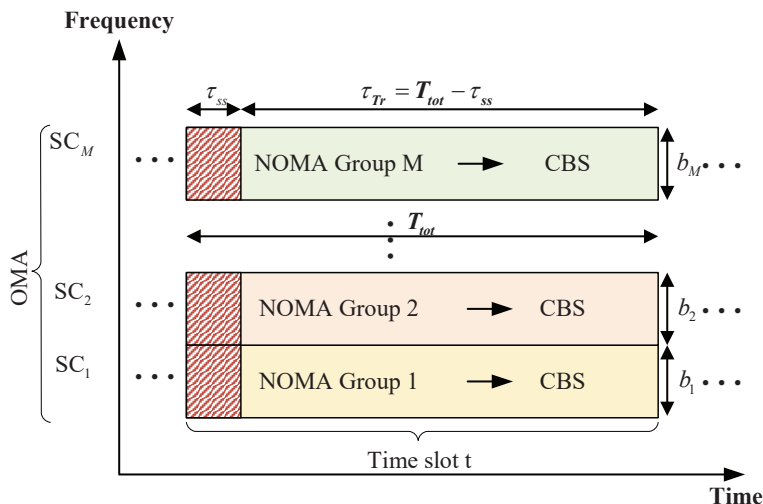


Figure 5.2: Time frame of the cognitive users' operations.

During data transmission, the received composite signal at the CBS on subchannel SC_m is given by

$$y_m(t) = \sqrt{P_{1m}(t)}x_{1m}(t)h_{1m} + \sqrt{P_{2m}(t)}x_{2m}(t)h_{2m} + \omega_m, \quad (5.1)$$

where $P_{im}(t) = e_{im}^{tr}(t)/\tau_{Tr}$, $i \in \{1, 2\}$, $m \in \{1, 2, \dots, M\}$ is the transmission power of CU_i in NOMA group G_m , in which $e_{im}^{tr}(t)$ is the transmission energy assigned for CU_{im} in time slot t ; $x_{im}(t)$ denotes the transmit signal of CU_{im} in time slot t , ($\mathbb{E}\{|x_{im}(t)|^2\} = 1$); ω_m is the additive white Gaussian noise (AWGN) at the CBS on subchannel SC_m with zero mean and variance σ^2 ; and h_{im} is the channel coefficient between CU_{im} and the CBS. The overall received signal at the CBS in time slot t is given by

$$y(t) = \sum_{m=1}^M y_m(t). \quad (5.2)$$

The received signals at the CBS on different sub-channels are independently retrieved from composite signal $y_m(t)$ using the SIC technique. In particular, the CU's signal

with the highest channel gain is firstly decoded, and then it will be removed from composite signal at the CBS, in a successive manner. The CU's signal with the lower channel gain in the sub-channel is treated as noise of the CU with the higher channel gain. We assume perfect SIC implementation at the CBS. The achievable transmission rate for the CUs in NOMA group G_m are

$$\begin{aligned} R_{1m}(t) &= \frac{\tau T_r}{T_{tot}} \times b_m(t) \times \log_2 \left[1 + \frac{P_{1m}(t)g_{1m}}{P_{2m}(t)g_{2m} + \sigma^2} \right] \\ R_{2m}(t) &= \frac{\tau T_r}{T_{tot}} \times b_m(t) \times \log_2 \left[1 + \frac{P_{2m}(t)g_{2m}}{\sigma^2} \right], \end{aligned} \quad (5.3)$$

where $b_m(t)$ is the amount of bandwidth allocated to subchannel SC_m in time slot t , $g_{im} = |h_{im}|^2$ is the channel gain of CU_{im} on subchannel m , and $g_{1m} \geq g_{2m}$. Since the channel gain of CU_{1m} , g_{1m} , is higher, CU_{1m} has a higher priority for decoding. Consequently, the signal of CU_{1m} is decoded first by treating the signal of CU_{2m} as interference. Next, user CU_{1m} is removed from signal $y_m(t)$, and the signal of user CU_{2m} is decoded as interference-free. The sum achievable transmission rate of NOMA group G_m can be calculated as:

$$R_m(t) = R_{1m}(t) + R_{2m}(t). \quad (5.4)$$

The sum achievable transmission rate at the CBS can be given as follows:

$$R(t) = \sum_{m=1}^M R_m(t). \quad (5.5)$$

5.2.1 Energy Arrival and Primary User Models

In this chapter, the CUs have a finite capacity battery, E_{bat} , which can be constantly recharged by the solar energy harvesters. Therefore, the CUs can perform their other operations and harvest solar energy simultaneously. For many reasons (such as the weather, the season, different times of the day), the harvested energy from solar resources may vary in practice. Herein, we take into account a practical case, where the harvested energy of CU_i in NOMA group G_m (denoted as e_{im}^h) follows a Poisson distribution with mean value ξ_{avg} , as studied in [160]. The arrival energy amount that CU_{im} can harvest during time slot t can be given as $e_{im}^h(t) \in \{e_1^h, e_2^h, \dots, e_v^h\}$ where $0 < e_1^h < e_2^h < \dots < e_v^h < E_{bat}$. The cumulative distribution function can be given as follows:

$$F\left(e_{im}^h(t); \xi_{avg}\right) = \sum_{k=0}^{e_{im}^h(t)} e^{-\xi_{avg}} \frac{(\xi_{avg})^k}{k!}. \quad (5.6)$$

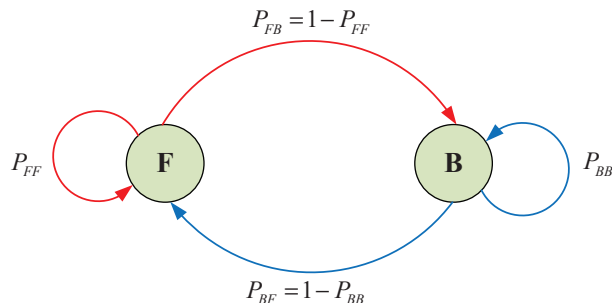


Figure 5.3: Markov chain model of the primary channel.

Herein, we use a two-state Markov discrete-time process to model the state of the primary channel, as depicted in Figure 5.3. We assume that the state of the primary channel does not change during the time slot duration, T_{tot} , and the primary channel can switch states between two adjacent time slots. The state transition probabilities between two time slots are denoted as $P_{ij} | i, j \in \{F, B\}$, in which F stands for the *free* state, and B stands for the *busy* state. In this chapter, we consider cooperative spectrum sensing, in which all CUs collaboratively detect spectrum holes based on an energy detection method, and they send these local sensing results to the CBS. Subsequently, the final decision on the primary users' activities is attained by combining the local sensing data at the CBS [159]. The performance of the cooperative sensing scheme can be evaluated based on probability of detection P_d and probability of false alarm P_f . P_d is denoted as the probability that the PU's presence is correctly detected (i.e., the primary channel is actually used by the PUs). Meanwhile, P_f is denoted as the probability that the PU's is detected to be active, but it is actually inactive (i.e., the sensing result of the primary channel is busy, but the primary channel is actually free) .

5.3 Long-Term Transmission Rate Maximization Problem Formulation

In this section, we aim at maximizing the long-term data transmission rate for uplink NOMA/OMA. The $2M$ users in the CRN can be decoupled into pairs according to their channel gain, as described in Algorithm 5.1. After user pairing, the joint power

allocation and bandwidth allocation problem can be formulated as follows:

$$\begin{aligned}
 \mathbf{a}^*(t) &= \arg \max_{\mathbf{a}(t)} \sum_{k=t}^{\infty} \sum_{m=1}^M R_m(k) \\
 s.t. & 0 \leq e_{im}^{tr} \leq e_{max}^{tr} \quad , \\
 & \sum_{m=1}^M \mathbf{b}_m(t) = W
 \end{aligned} \tag{5.7}$$

where $\mathbf{a}(t) = \{\mathbf{b}(t), \boldsymbol{\varepsilon}(t)\}$ represents the action that the CBS assigns to the CUs in time slot t ; $\mathbf{b}(t)$ indicates a vector of the allocated bandwidth portions assigned to the corresponding sub-channel, where $\mathbf{b}(t) = \{b_1(t), b_2(t), \dots, b_M(t)\} \left| \sum_{m=1}^M b_m(t) = W \right.$ is the assigned bandwidth amount for m^{th} sub-channel; $\boldsymbol{\varepsilon}(t) = [e_{11}^{tr}(t), e_{21}^{tr}(t), e_{12}^{tr}(t), e_{22}^{tr}(t), \dots, e_{1M}^{tr}(t), e_{2M}^{tr}(t)]$ refers to a transmission energy vector of the CUs, where $e_{im}^{tr}(t) \in \{0, e_1^{tr}, e_2^{tr}, \dots, e_{max}^{tr}\}$ is the transmission energy value for CU_{im} , and e_{max}^{tr} represents the upper-bounded value of transmission energy for each CU in time slot t .

5.4 Deep Reinforcement Learning-Based Resource Allocation Policy

In this section, we first reformulate the joint power and bandwidth allocation problem, which is aimed at maximizing the long-term data transmission rate of the system as the framework of an MDP. Then, we apply a DRL approach to solve the problem, in which the agent (i.e., the CBS) learns to create the optimal resource policy via trial-and-error interactions with the environment. One of the disadvantages of reinforcement learning is that the high computational costs can be imposed due to the long time learning process of a system with high state space and action space. However, the proposed scheme requires less computation overhead by adopting deep neural networks, as compared to other algorithms such as value iteration-based dynamic programming in partially observable Markov decision process (POMDP) framework [123] in which the transition probability of the energy arrival is required for obtaining the solution. Thus, the complex in formulation and computation can be relieved regardless of the dynamic properties of the environment by using the proposed scheme, as compared to POMDP scheme.

Furthermore, the advantage of a deep reinforcement learning scheme as compared with the POMDP scheme is that the unknown harvested energy distribution can be estimated

to create the optimal policy by interacting with the environment over the time horizon. In addition, the proposed scheme can work effectively in a large-state-and-space system by adopting deep neural networks. However, other reinforcement learning schemes such as Q-learning [161], actor-critic learning [162] might not be appropriate to large-state-and-space systems. In the proposed scheme, a deep neural network was trained to obtain the optimal policy where the reward of the system converges to optimal value. Then, the system can choose an optimal action at every state according to that policy learned from the training phase without re-training. Thus, deep actor-critic reinforcement learning can be more applicable to the wireless communication system.

5.4.1 Markov Decision Process

Generally, the purpose of reinforcement learning is for the agent to learn how to map each system state to an optimal action through a trial-and-error learning process. In this way, the accumulated sum of rewards can be maximized after a number of training time slots. Figure 5.4 illustrates the traditional reinforcement learning via agent–environment interaction. In particular, the agent observes the system state and then chooses an action at the beginning of a time slot. After that, the system receives the corresponding reward at the end of the time slot, and transfers to the next state based on the performed action. The system will be updated and will then go into the next interaction between agent and environment.

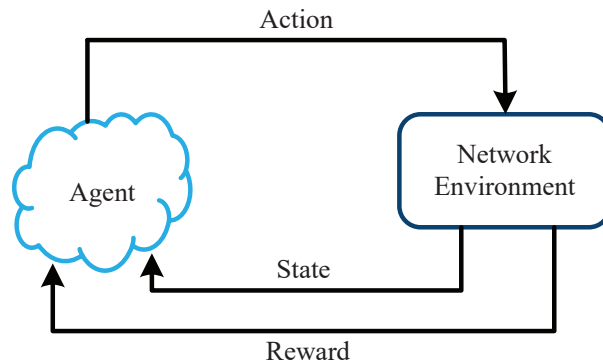


Figure 5.4: The agent–environment interaction process.

We denote the state space and action space of the system in this chapter as \mathbb{S} and \mathbb{A} , respectively; $s(t) = \{\mu(t), e^{re}(t)\} \in \mathbb{S}$ represents the state of the network in time slot

t , where $\mu(t)$ is the probability (*belief*) that the primary channel is free in that time slot, and $\mathbf{e}^{re}(t) = [e_{11}^{re}(t), e_{21}^{re}(t), e_{12}^{re}(t), e_{22}^{re}(t), e_{1M}^{re}(t), e_{2M}^{re}(t)]$ denotes a vector of remaining energy of CUs, where $0 \leq e_{im}^{re} \leq E_{bat}$ represents the remaining energy value of CU_{im} . The action at the CBS is denoted as $\mathbf{a}(t) = \{\mathbf{b}(t), \boldsymbol{\varepsilon}(t)\} \in \mathbb{A}$. In this chapter, we define the reward as the sum data rate of the system, as presented in Eq. (5.5).

The decision-making process can be expressed as follows. At the beginning of time slot t , the agent observes the state, $s(t) \in \mathbb{S}$, from information about the environment, and then chooses action $\mathbf{a}(t) \in \mathbb{A}$ following a stochastic policy, $\pi(a|s) = \Pr(a(t) = a | s(t) = s)$, which is mapped from the environment state to the probability of taking an action. In this work, the network agent (i.e, the CBS) determines the transmission power for each CU and decides whether to allocate the bandwidth portion to the NOMA groups in each time slot. Then, the CUs perform their operations (transmit data or stay silent) according to the assigned action from the CBS. Afterward, the instant reward, $R(t)$, which is defined in Eq. (5.5), is fed back to the agent, and the environment transforms to the next state, $s(t+1)$. At the end of the time slot, the CUs report information about the current remaining energy level in each CU to the CBS for network management. In the following, we describe the way to update information about the belief and the remaining energy based on the assigned actions at the CBS.

5.4.1.1 Silent Mode

The global sensing decision shows that the primary channel is busy in the current time slot, and thus, the CBS trusts this result and has all CUs stay silent. As a consequence, there is no reward in this time slot, i.e., $R(t) = 0$. The belief in current time slot t can be calculated according to Bayes' rule [128], as follows:

$$\mu(t) = \frac{\mu(t) P_f}{\mu(t) P_f + (1 - \mu(t)) P_d}. \quad (5.8)$$

Belief $\mu(t+1)$ for the next time slot is updated as follows:

$$\mu(t+1) = \mu(t) P_{FF} + (1 - \mu(t)) P_{BF}. \quad (5.9)$$

The remaining energy of CU_{im} for the next time slot is updated as

$$e_{im}^{re}(t+1) = \min\left(e_{im}^{re}(t) + e_{im}^h(t) - e_{ss}, E_{bat}\right), \quad (5.10)$$

where e_{ss} is the consumed energy from the spectrum sensing process.

5.4.1.2 Transmission Mode

The global sensing decision indicates that the primary channel is free in the current time slot, and then, the CBS assigns transmission power levels to the CUs for transmitting their data to the CBS. We assume that the data of the CUs will be successfully decoded if the primary channel is actually free; otherwise, no data can be retrieved due to collisions between the signals of the PUs and CUs. In this case, there are two possible observations, as follows.

Observation 1 (Φ_1): All data are successfully received and decoded at the CBS at the end of the time slot. This result means the primary channel was actually free during this time slot, and the global sensing result was correct. The total reward for the network is calculated as

$$R(s(t) | \Phi_1) = \sum_{m=1}^M R_m(t), \quad (5.11)$$

where the immediate data transmission rate of NOMA group G_m , $R_m(t)$, can be computed with Eq. (5.4). Belief $\mu(t+1)$ for the next time slot is updated as

$$\mu(t+1) = P_{FF}. \quad (5.12)$$

The remaining energy in CU_{im} for the next time slot will be

$$e_{im}^{re}(t+1) = \min\left(e_{im}^{re}(t) + e_{im}^h(t) - e_{ss} - e_{im}^{tr}(t), E_{bat}\right), \quad (5.13)$$

where $e_{im}^{tr}(t)$ denotes the transmission energy assigned to CU_{im} in time slot t .

Observation 2 (Φ_2): The CBS can not successfully decode the data from the CUs at the end of time slot t due to collisions between the signals of the CUs and the PUs. It implies that the primary channel was occupied, and misdetection happened. In this case, no reward is achieved, i.e., $R(s(t) | \Phi_2) = 0$. Belief $\mu(t+1)$ for the next time slot can be updated as

$$\mu(t+1) = P_{BF}. \quad (5.14)$$

The remaining energy in CU_{im} for the next time slot is updated by

$$e_{im}^{re}(t+1) = \min\left(e_{im}^{re}(t) + e_{im}^h(t) - e_{ss} - e_{im}^{tr}(t), E_{bat}\right). \quad (5.15)$$

In reinforcement learning, the agent is capable of improving the policy based on the recursive lookup table of state-value functions. The state-value function, $V^\pi(s)$, is defined as the maximum expected value of the accumulated reward starting from current state s with the given policy, which is written as [152]:

$$V^\pi(s) = E \left\{ \sum_{t=1}^{\infty} \gamma^t R(t) \mid s(t) = s, \pi \right\}, \quad (5.16)$$

where $E\{.\}$ denotes the expectation, in which $\gamma \in (0, 1)$ is the discount factor, which can affect the agent's decisions on myopic or foresighted operations; π is the stochastic policy, which maps environment state space \mathbb{S} to action space \mathbb{A} , $\pi(a|s) = \Pr(\mathbf{a}(t) = a \mid s(t) = s)$. The objective of the resource allocation problem is to find optimal policy π^* that provides the maximum discounted value function in the long run, which can satisfy the Bellman equation as follows [163]:

$$\pi^*(a|s) = \arg \max_{\pi} V^\pi(s). \quad (5.17)$$

The policy can be explored by using an ϵ -greedy policy in which a random action is chosen with probability ϵ , or an action can be selected based on the current policy with probability $(1 - \epsilon)$ during the training process [164]. As a result, the problem of joint power and bandwidth allocation in Eq. (5.7) can be rewritten as Eq. (5.17), and the solution to deep actor-critic reinforcement learning will be presented in the following section.

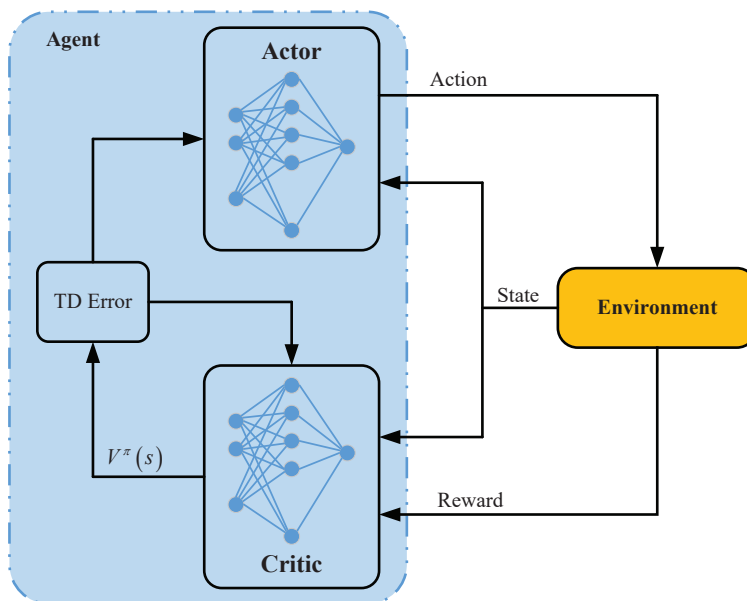


Figure 5.5: The structure of deep actor-critic reinforcement learning.

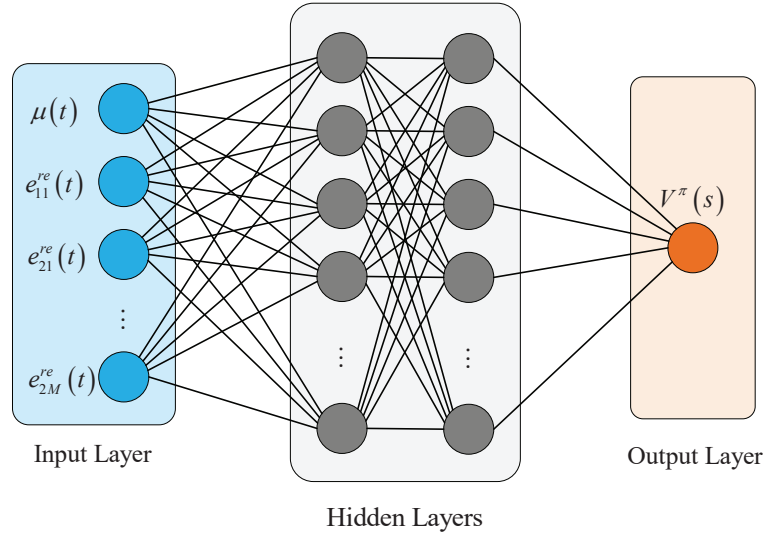


Figure 5.6: The deep neural network in the critic.

5.4.2 Deep Actor-Critic Reinforcement Learning Algorithm

The maximization problem in Eq. (5.17) can be solved by using the actor-critic method, which is derived by combining the value-based method [165] and the policy-based method [166]. The actor-critic structure involves two neural networks (actor and critic) and an environment, as shown in Fig. 5.5. The actor can determine the action according to the policy, and the critic evaluates the selected actions based on value functions and instant rewards that are fed back from the environment. The input of the actor is the state of the network, and the output is the policy, which directly affect how the agent chooses the optimal action. The output of the critic is a state-value function $V^\pi(s)$, which is used to calculate the temporal difference (TD) error. Thereafter, the TD error is used to update the actor and the critic.

Herein, both the policy function in the actor and the value function in the critic are approximated with parameter vectors θ and ω , respectively, by two sequential models of a deep neural network. Both value function parameter ω and policy parameter θ are stochastically initialized and updated constantly by the critic and the actor, respectively, during the training process.

5.4.2.1 The Critic with a DNN

Fig. 5.6 depicts the DNN at the critic, which is composed of an input layer, two hidden layers, and an output layer. The critic network is a feed-forward neural network that evaluates the action taken by the actor. Then, the evaluation of the critic is used by the actor to update its control policy. The input layer of the critic is an environment state, which contains $(2M + 1)$ elements. Each hidden layer is a fully connected layer, which involves H_C neurons and uses a rectified linear unit (ReLU) activation function [167, 168] as follows:

$$f_{ReLU}(z) = \max(0, z), \quad (5.18)$$

where $z = \sum_{i=1}^{2M+1} \omega_i s_i(t)$ is the estimated output of the layer before applying the activation function, in which $s_i(t)$ indicates the i th element of the input state, $s(t)$, and ω_i is the weight for the i th input. The output layer of the DNN at the critic contains one neuron and uses the linear activation function to estimate the state-value function, $V^\pi(s)$. In this chapter, the value function parameter is optimized by adopting stochastic gradient descent with a back-propagation algorithm to minimize the loss function, defined as the mean squared error, which is computed by

$$\mathcal{L}_\omega = \delta^2(t), \quad (5.19)$$

where $\delta(t)$ is the TD error between the target value and the estimated value, which is given by

$$\delta(t) = E[R(t) + \gamma V_\omega(s(t+1)) - V_\omega(s(t))], \quad (5.20)$$

and it is utilized to evaluate selected action $\mathbf{a}(t)$ of the actor. If the value of $\delta(t)$ is positive, the tendency to choose action $\mathbf{a}(t)$ in the future, when the system is in the same state, will be strengthened; otherwise, it will be weakened. The critic parameter can be updated in the direction of the gradient, as follows:

$$\Delta\omega = \alpha_c \delta(t) \nabla_\omega V_\omega^\pi(s(t)), \quad (5.21)$$

where α_c is the learning rate of the critic.

5.4.2.2 The Actor with a DNN

The DNN in the actor is shown in Fig. 5.7, which includes an input layer, two hidden layers, and an output layer. The input layer of the actor is the current state of the

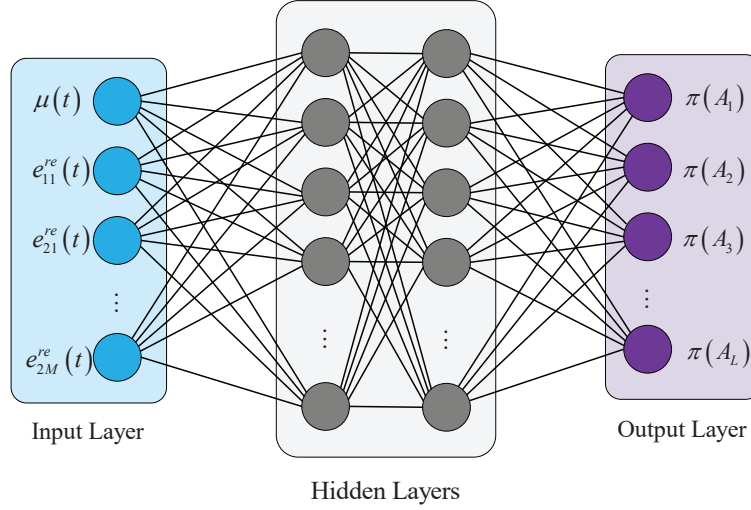


Figure 5.7: The deep neural network in the actor.

environment. There are two hidden layers in the actor, where each hidden layer is comprised of H_A neurons. The output layer of the actor provides the probabilities of selecting actions in a given state. Furthermore, the output layer utilizes the soft-max activation function [152] to compute the policy of each action in the action space, which is given as:

$$\pi_{\theta}(a|s) = \frac{e^{z_a}}{\sum_{\mathbf{a}' \in \mathcal{A}} e^{z_{\mathbf{a}'}}}, \quad (5.22)$$

where z_a is the estimated value for the preference of choosing action \mathbf{a} . In the actor, the policy can be enhanced by optimizing the state-value function as follows:

$$\begin{aligned} J(\pi_{\theta}) &= E[V^{\pi}(s)] \\ &= \sum_{s \in \mathcal{S}} d^{\pi}(s) V^{\pi}(s), \end{aligned} \quad (5.23)$$

where $d^{\pi}(s)$ is the state distribution. Policy parameter θ can be updated toward the gradient ascending to maximize the objective function [162], as follows:

$$\Delta \theta = \alpha_a \nabla_{\theta} J(\pi_{\theta}), \quad (5.24)$$

where α_a denotes the learning rate of actor network, and policy gradient $\nabla_{\theta} J(\pi_{\theta})$ can be computed by using the TD error [169]:

$$\nabla_{\theta} J(\pi_{\theta}) = E[\nabla_{\theta} \log \pi_{\theta}(s, \mathbf{a}) \delta(t)]. \quad (5.25)$$

Algorithm 5.2 The training procedure of the deep actor-critic reinforcement learning algorithm

```

1: Input:  $\mathbb{S}, \mathbb{A}, \gamma, \alpha_a, \alpha_c, \mathbf{e}^{re}(t), \mu(t), E_{ca}, \xi_{avg}, T, W, \epsilon_{min}, \epsilon_{max}, \epsilon_d$ .
2: Initialize network parameters of the actor and the critic:  $\boldsymbol{\theta}, \boldsymbol{\omega}$ .
3: Initialize  $\epsilon = \epsilon_{max}$ .
4: for each episode  $e = 1, 2, 3, \dots, L$  :
5:     Sample an initial state  $s \in \mathbb{S}$ .
6:     for each time step  $t = 0, 1, 2, 3, \dots, T - 1$  :
7:         Observe current state  $s(t)$ , and estimate state value  $V_{\boldsymbol{\omega}}^{\pi}(s(t))$ .
8:         Choose an action:
9:
10:        
$$\mathbf{a}(t) = \begin{cases} \arg \max \pi_{\boldsymbol{\theta}}(a(t) | s(t)) & \text{w.p } 1 - \epsilon \\ \text{random action } a(t) \in \mathbb{A} & \text{otherwise} \end{cases}$$

11:        Execute action  $\mathbf{a}(t)$ , observe current reward  $R(t)$ .
12:        Update epsilon rate  $\epsilon = \max(\epsilon, \epsilon_d, \epsilon_{min})$ 
13:        Update next state  $s(t+1)$ 
14:        Critic Process:
15:        Estimate next state value  $V_{\boldsymbol{\omega}}^{\pi}(s(t+1))$ .
16:        Critic calculates TD error  $\delta(t)$ 
17:        if episode is end at time slot  $t$ :
18:            
$$\delta(t) = R(t) - V_{\boldsymbol{\omega}}(s(t)).$$

19:        else
20:            
$$\delta(t) = R(t) + \gamma V_{\boldsymbol{\omega}}(s(t+1)) - V_{\boldsymbol{\omega}}(s(t)).$$

21:        end if
22:        Update parameter of critic network  $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} + \Delta\boldsymbol{\omega}$ 
23:        Actor Process:
24:        Update parameter of actor network  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ 
25:    end for
26: end for
27: Output: Final policy  $\pi_t^*(\mathbf{a}(t) | s(t))$ .

```

It is worth noting that TD error $\delta(t)$ is supplied by the critic. The training procedure of the proposed DACRL approach is summarized in **Algorithm 5.2**. In the

algorithm, the agent interacts with the environment and learns to select optimal action in each state. The convergence of the proposed algorithm depends on number of steps per episode, the number of training episodes and the learning rate, which is discussed in the following section.

Table 5.1: SIMULATION PARAMETERS

Parameter	Description	Value
M	Number of groups	3
T_{tot}	Time slot duration	200 ms
τ_{ss}	Sensing duration	2 ms
W	Total system bandwidth	1 Hz
E_{bat}	Battery capacity	30 μJ
e_{ss}	Sensing cost	1 μJ
e^{tr}	Transmission energy	0, 10, 20 μJ
ξ_{avg}	Mean value of harvested energy	5 μJ
μ	Initial belief that the primary channel is free	0.5
P_{FF}	Transition probability of the primary channel from state F to itself	0.8
P_{BF}	Transition probability of the primary channel from state B to state F	0.2
P_d	Probability of detection	0.9
P_f	Probability of false alarm	0.1
σ^2	Noise variance	-80 dB
γ	Discount factor	0.9
α_a	Learning rate of the actor	0.001
α_c	Learning rate of the critic	0.005
ϵ	Epsilon rate	1 \rightarrow 0.01
ϵ_d	Epsilon decay	0.9999
L	Number of episodes	300
T	Number of iterations per episode	2000

5.5 Simulation Results

In this section, we investigate the performance of uplink NOMA systems using our proposed scheme. The simulation results are compared with other myopic schemes [170] (Myopic-UP, Myopic-Random, and Myopic-OMA) in terms of average data transmission rate and energy efficiency. In the myopic schemes, the system only maximizes the reward in the current time slot, and the system bandwidth is allocated to the group only if it has at least one active CU in the current time slot. In particular, with the Myopic-UP scheme, the CBS arranges the CUs into different pairs based on **Algorithm 5.1**. In the Myopic-Random scheme, the CBS randomly decouples the CUs into pairs. In the Myopic-OMA scheme, the total system bandwidth is divided equally into sub-channels in order to assign them to each active CU without applying user pairing. In the following, we analyze the influence of the network parameters on the schemes through the numerical results.

In this chapter, we used Python 3.7 with the TensorFlow deep learning library to implement the DACRL algorithm. Herein, we consider a network based on different channel gain values between the CUs and the CBS, such as $h_1 = -20$ dB, $h_2 = -25$ dB, $h_3 = -30$ dB, $h_4 = -35$ dB, $h_5 = -40$ dB, $h_6 = -45$ dB, where $h_1, h_2, h_3, h_4, h_5, h_6$ are the channel gains between $CU_1, CU_2, CU_3, CU_4, CU_5, CU_6$ and the CBS, respectively. Two sequential DNNs are utilized to model the value function and the policy function in the proposed algorithm. Each DNN is designed with an input layer, two hidden layers and an output layer as described in Section 4. The number of neurons in each hidden layer of the value function DNN in the critic, and the policy function in the actor, are set at $H_C = 24$ and $H_A = 24$, respectively. For the training process, value function parameter ω and the policy parameter θ are stochastically initialized by using uniform Xavier initialization [171]. The other simulation parameters for the system are shown in **Table 5.1**.

We first examine the average transmission rates of the the DACRL scheme under different training iterations, T , while the number of episodes, L , increases from 1 to 400. We achieved the results by calculating the average transmission rate after separately running the simulation 20 times, as shown in Fig. 5.8. The curves sharply increase in the first 50 training episodes, and then gradually converge to the optimal value. We can see that the agent needs more than 350 episodes to learn the optimal policy at $T = 1000$ iterations per episode. However, with the increment in T , the algorithm begins to converge faster. For instance, the proposed scheme learns the optimal policy in less than 300 episodes when $T = 2000$.

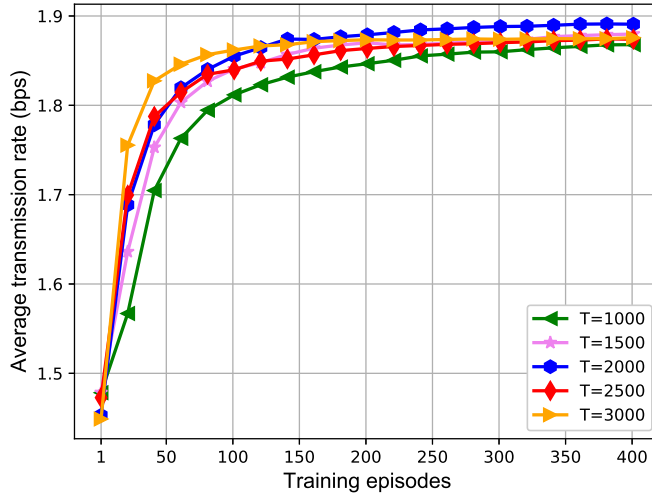


Figure 5.8: The convergence rate of the proposed actor-critic deep reinforcement learning with different training steps in each episode.

Nevertheless, it might take a very long time for the training process if each episode uses too many iterations, and the algorithm evenly converges to a locally optimal policy. As a result, the number of training iterations per episode and the number of training episodes should not be too large or too small. In the rest of the simulations, we set training episodes at 300 and training iterations at 2000.

Fig. 5.9 shows the convergence rate of the proposed scheme according to various values of actor learning rate α_a and critic learning rate α_c . The figure shows that the reward converges faster with increments in the learning rates. In addition, we can observe that the proposed scheme with actor learning rate $\alpha_a = 0.001$ and critic learning rate $\alpha_c = 0.005$ provides the best performance after 300 episodes. When the learning rates of the actor and the critic increase to $\alpha_a = 0.01$ and $\alpha_c = 0.005$, respectively, the algorithm converges very fast, but does not bring a good reward due to underfitting. Therefore, we set the actor and critic learning rates at $\alpha_a = 0.001$ and $\alpha_c = 0.005$, respectively, for the rest of the simulations.

Fig. 5.10 illustrates the average transmission rates under the influence of mean harvested energy. We can see that the average transmission rate of the system increases when the mean value of harvested energy grows. The reason is that with an increase in

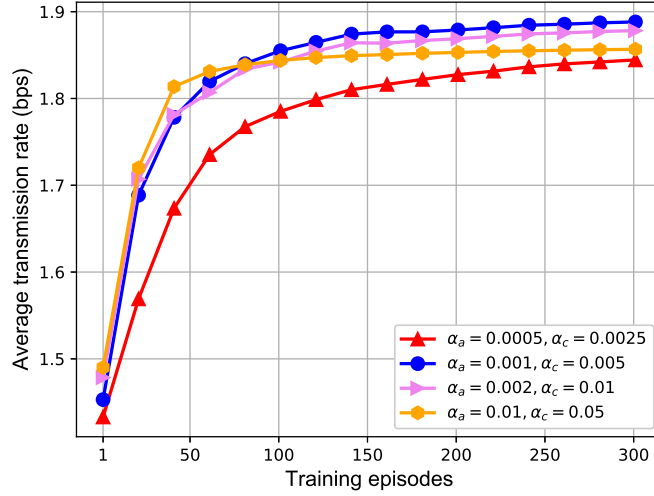


Figure 5.9: The convergence rate of the proposed actor-critic deep reinforcement learning according to different learning rate values.

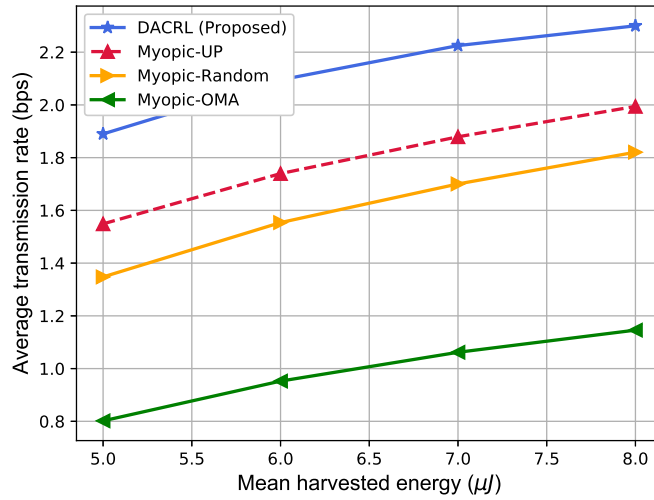


Figure 5.10: Average transmission rate according to different values for mean harvested energy.

ξ_{avg} , the CUs can harvest more solar energy, and thus, the CUs have a greater chance to transmit data to the CBS. In addition, the average transmission rate of the proposed scheme dominates the conventional schemes because the conventional schemes focus on maximizing

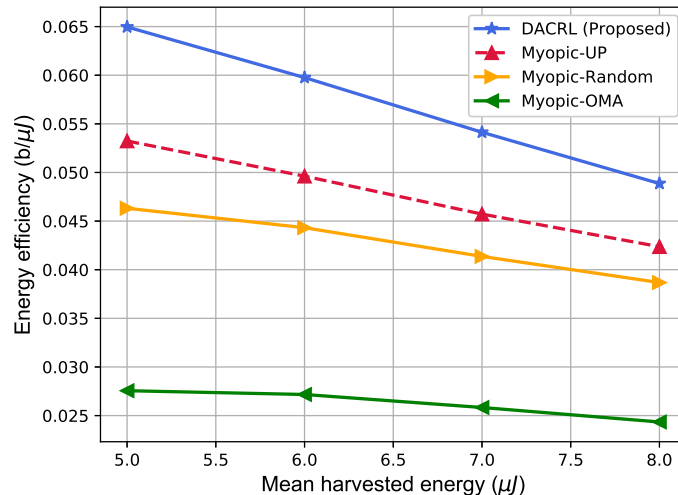


Figure 5.11: Energy efficiency according to different values of harvested mean energy.

the current reward, and they ignore the impact of the current decision on the future reward. Thus, whenever the primary channel is free, these conventional schemes allow all CUs to transmit their data by consuming most of the energy in the battery in order to maximize the instant reward. This makes the CUs stay silent in the future due to energy shortages. Although the Myopic-Random scheme had lower performance than the Myopic-UP scheme, it still had greater rewards than Myopic-OMA. This outcome demonstrates the efficiency of the hybrid NOMA/OMA approach, compared with the OMA approach, in terms of average transmission rate.

In Fig. 5.11, the energy efficiency of the schemes was compared with respect to the mean value of the harvested energy. In this chapter, we define energy efficiency as the transmission data rate obtained at the CBS over the total energy consumption of the CUs during the operations. We can see that the energy efficiency declines as ξ_{avg} rises. The reason is that when the harvested energy goes up, the CUs can gather more energy for their operations; however, the amount of energy overflowing the CUs' batteries also increases. The curves show that the performance of the proposed scheme outperforms the other conventional schemes because the DACRL agent can learn about the dynamic arrival of harvested energy from the environment. Thus, the proposed scheme can make proper decision in each time slot.

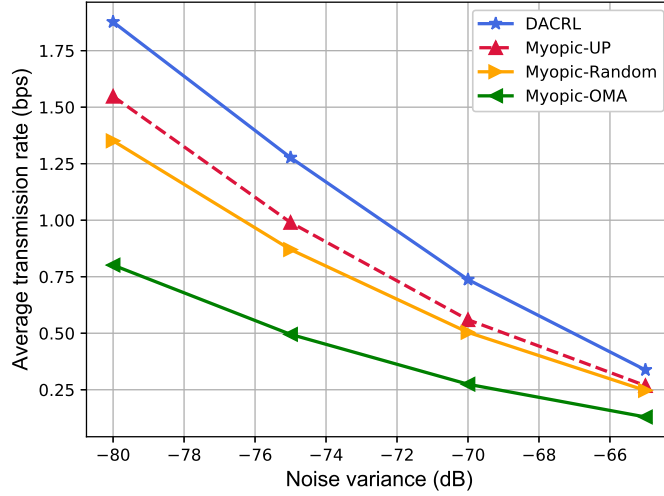


Figure 5.12: Average transmission rate according to noise variance.

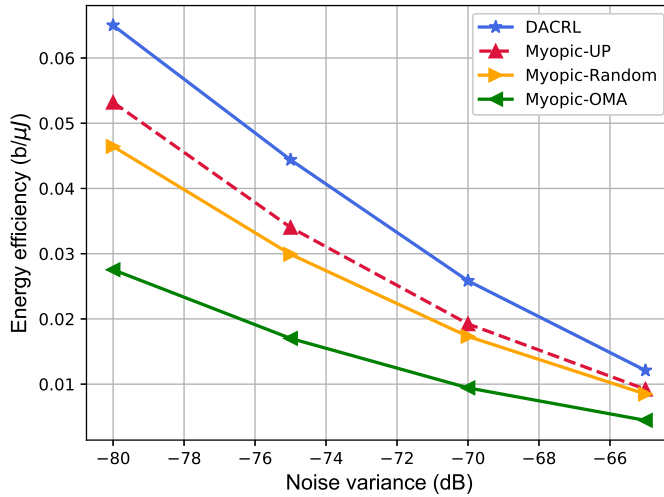


Figure 5.13: Energy efficiency according to noise variance.

In Figs. 5.12 and 5.13, we plot the average transmission rate and the energy efficiency, respectively, based on differing noise variance at the CBS. The curves show that system performance notably degrades when noise variance increases. To explain this, noise variance will degrade the data transmission rate, as shown in Eq. (5.3). As a consequence, energy efficiency also decreases with an increment in noise variance. Based on noise variance

at the CBS, the figures verify that the proposed scheme dominates the myopic schemes.

5.6 Conclusions

In this chapter, we investigated a deep reinforcement learning framework for joint power and bandwidth allocation by adopting both hybrid NOMA/OMA and user pairing in uplink CRNs. The DACRL algorithm was employed to maximize the long-term transmission rate under the energy constraint in the CUs. A DNN was applied to approximate the policy function and the value function such that the algorithm can work in the system with large state and action spaces. The agent of the DACRL can explore the optimal policy by interacting with the environment. As a consequence, the CBS can effectively allocate bandwidth and power to the CUs based on the current network state in each timeslot. The simulation results verified the advantages of the proposed scheme in improving network performance under various network conditions in the long run, compared to the conventional schemes.

Chapter 6

Deep Q-learning-based Resource Allocation for Solar-powered Users in Cognitive Radio Networks

6.1 Introduction

The fifth-generation (5G) network requires efficient enabling technologies to accommodate increasing requirements for high spectrum efficiency and high data rates in wireless communications. Therefore, the tremendous development of wireless-transmission and mobile-networking technologies has led to surging demand for resources in the last few decades. Both academic and industrial communities have been developing efficient resource management approaches to attain efficient spectrum utilization for the emerging mobile Internet [172]. Specifically, network virtualization [173] and software-defined networking (SDN) [174] are regarded as the key techniques to enhance network utility in terms of higher data rates, better resource utilization, and lower operational costs. Moreover, cognitive radio (CR) and NOMA have been considered potential techniques that allow secondary users to share licensed spectrum bands to improve spectrum efficiency and tackle the resource scarcity issues [175, 176].

There are several conventional multiple access schemes, such as TDMA, orthogonal frequency division multiple access (OFDMA), and code division multiple access (CDMA), to avoid interference among users. Nevertheless, owing to the rapid growth in the number of

mobile devices, these methods might not satisfy the requirements of users needing access to wireless communication systems. For that reason, NOMA is emerging as a critical principle in 5G networks for designing multi-access schemes, because it allows multiple users to utilize the same frequency resources at the same time [91]. Basically, the NOMA technique is categorized into two types: code domain and power domain. In this chapter, we focus on the power domain, where multiple users are assigned to use the same frequency and time resources for their data transmissions. In particular, the signals of multiple users are superposed to transmit over the same resources, and successive interference cancellation (SIC) is executed to decode the users' desired signals and remove interference at the receiver [130]. Many research works on NOMA schemes have been investigated in different communication systems, such as the Industrial Internet of Things (IIoT) [177, 178], machine-to-machine communications [179, 180], and cooperative communications [181, 182].

In reality, however, one of NOMA's drawbacks is that a massive NOMA-enabled communication systems might give rise to high computational complexity at the receiver. Besides, in designing the optimization schemes, numerous users applying NOMA will significantly degrade the performance of the system [183]. Indeed, NOMA can work well if it is multiplexed by a small number of users in a group [184]. Thus, combination of different multiple access techniques was proposed [185, 186]. In those studies, the authors proposed hybrid NOMA and OMA algorithms where the users switch between NOMA and OMA modes to improve network performance. In addition, the integration of NOMA with CR networks was studied to deal with 5G challenges, such as spectrum efficiency, and massive connectivity [102, 105, 187, 188]. The authors in [187] devised a taxonomy to categorize the literature according to operation paradigms, objectives, techniques, and optimization characteristics. Meanwhile, closed-form expressions of the outage probability for large-scale underlay cognitive networks were derived by using stochastic geometry [102]. In order to enhance the performance of both primary and secondary networks, the authors in [105] proposed an application of NOMA for cooperative multicast CRNs. Xu *et al.* [188] investigated optimal sensing duration adaptation, matching-theory user scheduling, and power allocation for cognitive OFDM-NOMA in order to boost system capacity.

Along with rapid developments of mobile devices, energy management is also a crucial issue. There have been a lot of model-based resource allocation schemes proposed to increase EE or other objectives in NOMA systems. The problem of power assignment was studied in [138], while the joint subcarrier assignment and power allocation algorithm

was proposed in [189]. However, these conventional approaches require complete network information, and induce high computational complexity, or even inapplicable, in practice. To address this problem, several studies have applied model-free deep learning to reduce computational complexity with available training data. Gui *et al.* [190] and Liu *et al.* [191] investigated resource allocation problems by using a neural network to train offline with simulated data first, and to then output results through well-trained networks. However, it is hard to obtain the correct data set or optimal solutions, and the training process is generally time-consuming. For the above issues, deep reinforcement learning (DRL) [192] has been emerging as a feasible option for real-time decision-making problems, since the requirements of the system model and the need for a priori data are significantly relaxed. Rather than optimizing current benefits only, DRL is able to generate an optimal decision policy that maximizes the long-term performance of systems through trial-and-error interaction with the environment. Deep Q-learning, considered one of the more famous DRL methods, applies a deep Q-network (DQN) that uses deep neural networks in conventional reinforcement learning (RL). Nowadays, DRL has been broadly used from many aspects, such as power allocation in NOMA systems [193], in heterogeneous networks [194], and IoT systems [195].

In recent years, one of most effective ways to enhance self-sustainability is to equip wireless devices with a rechargeable battery that is able to harvest ambient energy to enable long-term operation. There are many energy harvesting approaches from many natural resources for improving the lifetime of wireless users, such as solar energy [90, 123], wind power [196], and thermal power [197]. Among various types of energy harvesting resources, solar is regarded as one of the most effective sources for wireless users. Nevertheless, its effectiveness is highly dependent on the environment. For this reason, it is important to determine proper schemes to efficiently leverage harvested energy for wireless communication networks. Thanh *et al.* [123] proposed a framework of a partially observable Markov decision process (POMDP) to allocate both optimal frequency bands and optimal transmit power to solar-powered cognitive users.

To further improve performance, other techniques were studied in NOMA systems [198–202]. Ding *et al.* [198] introduced NOMA power allocation to acquire high sum data rates by grouping users with distinctive channel conditions. A resource allocation algorithm design was formulated as a non-convex optimization problem in which the authors considered the channel state information as well as quality of service (QoS) constraints [199]. Besides, Li *et al.* adopted a DQN to deal with the challenge of unknown system dynamics and maximize

the number of the successful transmissions from secondary systems without interfering with primary transmissions [200]. Furthermore, hybrid access mechanisms that include TDMA and NOMA technology were proposed [201,202]. In [201], a dynamic clustering approach was proposed to increase system throughput and balance power consumption, and the authors in [202] studied a method of power allocation and power splitting-ratio assignment for users to minimize transmission power under minimum rate and minimum-energy harvesting constraints.

In the aforementioned works, the authors in [201] proposed hybrid TDMA/NOMA-based dynamic clustering for machine-to-machine communications to improve the energy efficiency, throughput and lifetime of machine. However, the concept of energy harvesting was not considered. To facilitate the wireless communications, Al-Obiedollah *et al.* in [202] investigated hybrid TDMA/NOMA scheme by using the energy harvesting from radio frequency signal in order to minimize the transmit power. They utilized sequential convex approximation method to tackle with the non-convexity of the minimization problem. For improving radio spectrum and energy utilization efficiencies, the authors in [123] proposed the POMDP scheme to maximize the long-term secrecy rate of the solar-powered CRNs. Nevertheless, the approach requires the prior information of energy harvesting distribution, which is not easy to obtain in the practical scenarios. Motivated by the above analysis, in this chapter we focus on maximizing the long-term throughput of solar-powered secondary systems without prior knowledge of energy harvesting distribution by using NOMA/TDMA-based deep Q-learning scheme, in which secondary users can opportunistically use the licensed channel of the primary users. Consequently, by employing a deep Q-learning algorithm, the system is capable of obtaining optimal policy from trial-and-error interactions with the environment after training. The main contributions of this chapter are summarized as follows

- First, we investigate an uplink NOMA CRN in which multiple secondary users aim to transmit their data to a secondary base station (SBS), and in which secondary users are able to harvest solar energy for self-sustainability. In addition, we consider sensing error when the secondary system determines the status of the primary system through cooperative spectrum sensing.
- Second, we propose a NOMA/TDMA-based deep Q-learning approach to maximize the long-term throughput of a secondary system in which a deep neural network (DNN)

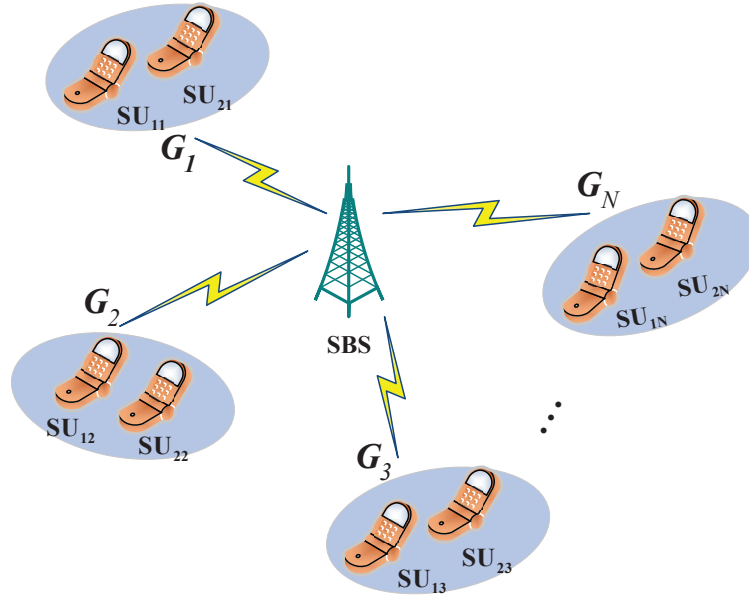


Figure 6.1: The considered network model.

is used to approximate the value function of every state-action pair. Throughout the training phase, the proposed scheme does not have prior information about the harvested energy distribution of the secondary users. Instead, that energy information can be learn, and then an optimal decision policy is achieved through trial-and-error interactions with the environment.

- We further present the impact of changed network parameters via the numerical simulation in which a performance comparison between the proposed scheme and conventional schemes is made based on the various parameters.

The remainder of this chapter is organized as follows. Section 6.2 describes the system model. The problem formulation is presented in Section 6.3. The DQL algorithm for power allocation is proposed in Section 6.4. Simulation results are discussed in Section 6.5. Finally, we conclude the chapter in Section 6.6.

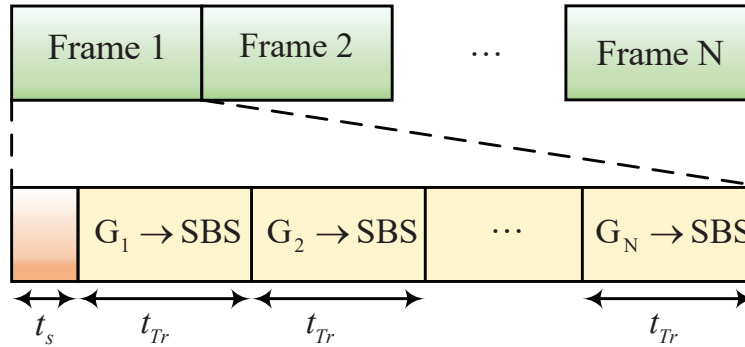


Figure 6.2: Time frame structure.

6.2 System Model

6.2.1 Network Model

This chapter considers an uplink CRN where multiple secondary users attempt to transmit data to an SBS by opportunistically using a licensed channel of the primary system. In particular, the network consists of an SBS and SUs, as illustrated in Fig. 6.1. Each SU is equipped with an energy-harvesting circuit that can harvest solar energy for its long-term operation. We assume the SUs always have data to transmit to the SBS and they can simultaneously harvest energy in the whole time frame while do other assigned actions. Hence, they opportunistically utilize the licensed channel of the primary system to transmit their data to the SBS by applying a joint NOMA and TDMA technique when the PUs are not active. To this end, in this chapter, the SBS assembles SUs into pairs according to a conventional near-far user pairing technique [203], in which the furthest user and the nearest user from the SBS can be coupled into a pair (i.e. a group of two users) based on their locations. In particular, the group $G_n = \{SU_{1n}, SU_{2n}\}$ is the n^{th} NOMA group which consists of 2 SUs, SU_{1n} and SU_{2n} , as shown in Fig. 6.1. Each user in a NOMA group can simultaneously transmit data by using the NOMA technique, and SIC from [130] is applied at the SBS to decode the received signals of each secondary user. Each group executes data transmission in turn during a whole time frame, such that there is no interference between each group. Each time frame of the system is divided into two phases (sensing and data transmission) as shown in Fig. 6.2. In this chapter, we adopt cooperative spectrum sensing in which the global sensing result is generated by the SBS. In the first phase, with a duration of t_s , all SUs simultaneously perform spectrum sensing in their local regions,

and then inform the SBS of their local results. Thereafter, the SBS determines the global sensing result on the status of the primary channel (vacant or occupied) by applying the combination rule from [159]. Subsequently, transmission power levels are assigned to all SUs for their data transmissions. As a result, the SUs in each NOMA group can transmit their data to the SBS on the primary channel based on the allocated transmission power without causing interference with other groups. Specifically, by adopting TDMA, each time frame is divided into N identical time slots according to the number of NOMA groups. Subsequently, each NOMA group respectively transmits data in its assigned time slot, with duration $t_{Tr} = \frac{T_{tot} - t_s}{N}$, where T_{tot} is time frame duration, t_s is spectrum sensing duration, and N is number of NOMA groups. The operation of the considered network model is illustrated in Fig. 6.2. The received signal at the SBS that is transmitted by n^{th} group, G_n , can be given as follows:

$$y_n(t) = \sum_{i=1}^2 \sqrt{P_{in}(t)} h_{in}(t) x_{in}(t) + z, \quad (6.1)$$

where $P_{in}(t) = e_{in}^{Tr}(t)/t_{Tr}$, $i \in \{1, 2\}$, $n \in \{1, 2, \dots, N\}$ is the transmission power allocated to the i^{th} SU of NOMA group n in time frame t ; $x_{in}(t)$ denotes the normalized signal of SU_{in} , ($\mathbb{E}\{|x_{in}(t)|^2\} = 1$); $h_{in}(t) = g_{in}(t) d_{in}^{-\kappa}$ denotes the channel gain between SU_{in} and the SBS in time frame t , in which $g_{in}(t)$ is the channel coefficient that follows the Rayleigh distribution, d_{in} denotes the distance between SU_{in} and the SBS, and κ and z are the path loss coefficient and the additive white Gaussian noise (AWGN) with zero mean and variance σ^2 , respectively. The overall received signal at the SBS in time frame t is expressed as

$$y(t) = \sum_{n=1}^N y_n(t). \quad (6.2)$$

Let $\Gamma_{in}(t) = h_{in}^2(t)/\sigma^2$ denote the channel-to-noise ratio (CNR) of SU_{in} . Without loss of generality, we assume the CNRs of the SUs in each NOMA group G_n are arranged in descending order: $\Gamma_{1n} \geq \Gamma_{2n}$. The SIC technique is applied to each NOMA group in order to retrieve the signals at the SBS. According to the NOMA principle, the received signal transmitted by the higher-CNR user is decoded first, and then it is removed from the composite signal at the SBS, which considers the signal of the another user as interference. After that, the interference-free signal of the lower-CNR user in NOMA group n is decoded.

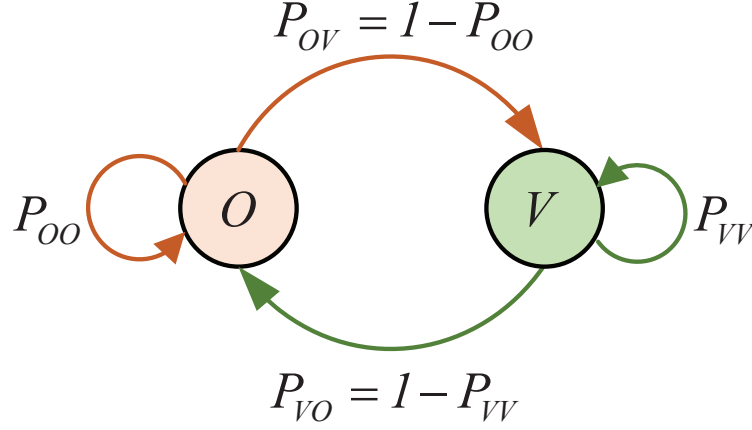


Figure 6.3: Two-state Markov discrete-time model for states of the primary channel.

As a result, the throughput of secondary users in NOMA group n can be expressed as follows:

$$\begin{aligned} R_{1n}(t) &= \frac{tT_r}{T_{tot}} \log_2 \left(1 + \frac{P_{1n}(t)\Gamma_{1n}(t)}{1+P_{2n}(t)\Gamma_{2n}(t)} \right), \\ R_{2n}(t) &= \frac{tT_r}{T_{tot}} \log_2 (1 + P_{2n}(t)\Gamma_{2n}(t)). \end{aligned} \quad (6.3)$$

The sum of the achievable throughput at the SBS can be calculated as

$$R(t) = \sum_{n=1}^N \sum_{i=1}^2 R_{in}(t). \quad (6.4)$$

6.2.2 Energy Arrival and Primary Channel Models

The SUs are powered by solar energy to prolong their operational lifetimes. The amount of harvested energy of SU_{in} during time frame t , is denoted as $e_{in}^h(t) \in \{e_1^h, e_2^h, \dots, e_v^h\}$, with $0 < e_1^h < e_2^h < \dots < e_v^h < E_c$, where E_c represents the battery capacity for the SUs. We assume that e_{in}^h follows a Poisson distribution with mean value e_{avg}^h .

Fig. 6.3 illustrates the two-state Markov discrete-time model for the states of the primary channel, in which O and V denote the states *occupied* and *vacant*, respectively. The probability that the primary channel changes its state between two adjacent time frames is denoted as $P_{ij} | i, j \in \{O, V\}$. In this chapter, an imperfection scenario for the spectrum sensing in the secondary system is taken into account. Thus, the performance of cooperative spectrum sensing is evaluated through two metrics: the detection probability, P_d , and the false alarm probability, P_f . P_d represents the probability that the *occupied* state of the primary channel is accurately detected, whereas P_f represents the probability that the primary channel is identified as *occupied* but it is actually *vacant*. It is obvious that the

performance of the system can be lowered by misdetections (i.e. the sensing result indicates that the primary channel is *vacant* but it is actually *occupied*).

6.3 Long-term Throughput Maximization Problem Formulation

In this chapter, the main goal is to obtain the maximum long-term throughput from an uplink NOMA/TDMA system. In particular, finding the optimal power allocation solution allows the SUs to transmit data to the SBS more efficiently, and hence, achieve higher throughput in long-term operation. Therefore, the problem formulation can be presented as follows:

$$\begin{aligned}
 \mathbf{a}^*(t) &= \arg \max_{\mathbf{a}(t) \in \mathbb{A}} \sum_{k=t}^{\infty} \sum_{n=1}^N \sum_{i=1}^2 R_{in}(k) \\
 \text{s.t.} \quad & \text{C1 : } 0 \leq \varepsilon_{in}^{tr}(t) \leq \varepsilon^{tr, \max} \quad , \\
 & \text{C2 : } 0 \leq \varepsilon_{in}^{tr}(t) \leq \varepsilon_{in}^{re}(t)
 \end{aligned} \tag{6.5}$$

where $\mathbf{a}(t) = \{\varepsilon_{11}^{tr}(t), \varepsilon_{21}^{tr}(t), \varepsilon_{12}^{tr}(t), \varepsilon_{22}^{tr}(t), \dots, \varepsilon_{1N}^{tr}(t), \varepsilon_{2N}^{tr}(t)\}$ is the allocated transmission power for all SUs in time frame t . Thus, $\varepsilon_{in}^{tr}(t) \in \{0, \varepsilon^{tr,1}, \varepsilon^{tr,2}, \dots, \varepsilon^{tr, \max}\}$ is the amount of transmission energy for SU_{in} ; $\varepsilon^{tr, \max}$ represents the upper bound of transmission energy for each SU in time frame t , and $\boldsymbol{\varepsilon}^{re}(t) = \{\varepsilon_{11}^{re}(t), \varepsilon_{21}^{re}(t), \varepsilon_{12}^{re}(t), \varepsilon_{22}^{re}(t), \dots, \varepsilon_{1N}^{re}(t), \varepsilon_{2N}^{re}(t)\}$ is a remaining-energy vector that consists of information about the remaining energy of SUs. $\varepsilon_{in}^{re}(t)$ denotes the value of the energy remaining in SU_{in} (i.e., the remaining energy of SU_i in group n). Constraint C1 shows that the transmission energy of SU_{in} does not exceed the maximum transmission value. In addition, constraint C2 guarantees that the transmission energy of each user is always less than or equal the remaining energy.

In order to solve problem (5), we can reformulate it into a Markov Decision Process (MDP) problem and obtain the solution by using iteration-based dynamic programming method. However, this approach requires the prior knowledge of the environment such as the harvested energy distribution, which is hard to achieve in practical scenarios. Furthermore, it also may impose a high computational overhead. To address problems without prior knowledge, the reinforcement learning has been considered as one of the potential approaches where the agent can obtain the optimal policy by directly interacting with the environment through trial-and-error [152, 204]. Unfortunately, it can be challenging to efficiently work in

such a large state-and-space system. Motivated by aforementioned things, we utilize a deep Q learning-based algorithm where Q-value for each action based on each state in Q-learning algorithm is approximated by using a neural network to deal with the formulated problem assuming that there is no prior knowledge of the environment. To achieve the solution to the problem in Eq. (6.5), we transform this problem into the Markov decision process (MDP) problem presented in the next section.

6.4 Deep Q-Learning-Based Resource Allocation Policy

In this section, we first reformulate the problem (5) into the MDP problem. Then, the observations based on the possible actions are introduced. Subsequently, a deep Q-learning approach is proposed to solve the MDP problem.

6.4.1 Decision-making process

Reinforcement learning is the framework for the MDP problem, which allows the agent to learn the optimal policy by interacting with an uncertain environment. The policy is regarded as guidance that tells the agent which action should be selected according to each state in order to obtain the maximum expected long-term throughput. Fig. 6.4 illustrates an interaction between agent and environment in the proposed deep reinforcement learning scheme. Herein, the agent can conceive of what happens in the environment after executing the selected action. Therefore, the optimal decision solution can be attained, as time goes on, through the learning process. Traditionally, an MDP is composed of five components in a tuple $\{\mathbb{S}, \mathbb{A}, \mathbb{P}, \mathbb{R}, \gamma\}$, where \mathbb{S} represents the state space, \mathbb{A} is the action space, \mathbb{P} indicates the state transition probability, \mathbb{R} refers to the reward function, and $\gamma \in [0, 1]$ is a discount factor that determines how much the selected action in the current time frame affects future rewards. It is worth noting that if $\gamma = 0$, the agent only maximizes the throughput in the current time frame without considering the impact of the current action on future rewards.

States: At the beginning of time frame t , the agent observes the system information in current state $s(t) \in \mathbb{S}$. The state is a combination of the belief and remaining energy vector, as follows:

$$s(t) = \{p(t), \epsilon^{re}(t)\}, \quad (6.6)$$

where $p(t)$ represents the probability that the primary channel is vacant in time frame t .

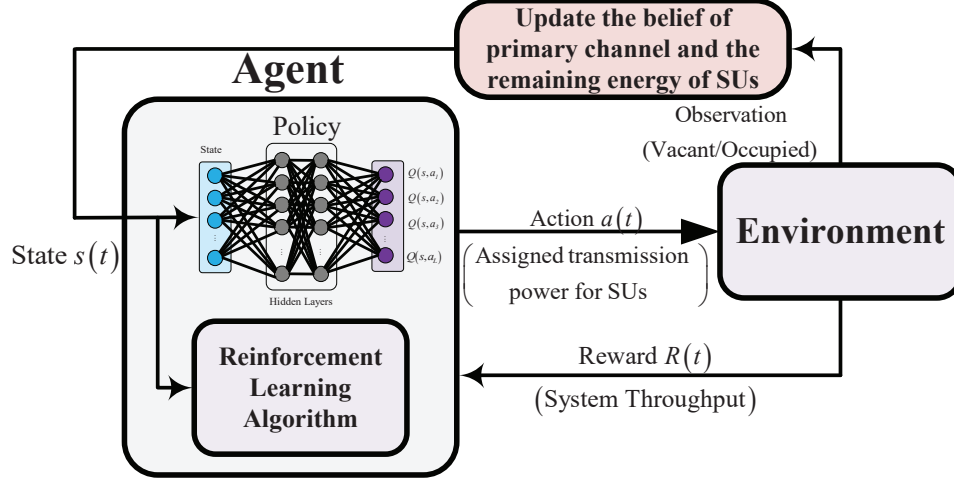


Figure 6.4: Overall structure of the proposed DQL-based power allocation scheme.

Actions: After observing current state $s(t)$, the agent selects an action $a(t) \in \mathbb{A}$, which is defined as

$$\mathbf{a}(t) = \{\varepsilon_{11}^{tr}(t), \varepsilon_{21}^{tr}(t), \varepsilon_{12}^{tr}(t), \varepsilon_{22}^{tr}(t), \dots, \varepsilon_{1N}^{tr}(t), \varepsilon_{2N}^{tr}(t)\} \quad (6.7)$$

where $\varepsilon_{in}^{tr}(t)$ refers to the allocated transmission power of SU_{in} in time frame t .

Transition Probability: The state transition probability represents the probability that the environment will transfer into a new state, $s(t+1)$, from current state $s(t)$ when selecting action $\mathbf{a}(t)$. However, in the reinforcement learning algorithm, the agent can deal with the problem of having no prior information about the state transition probability. Instead, the agent can directly interact with the environment, and then gradually learn the optimal solution without using the information on system uncertainty.

Reward function: In this chapter, the immediate reward is defined as the sum achievable throughput at the SBS in time frame t . Specifically, after employing selected action $\mathbf{a}(t)$ in current state $s(t)$, the immediate reward $R(t)$ is calculated with Eq. (6.4).

6.4.2 Observations

In this section, we describe the possible observations when the assigned action is executed by the SUs.

6.4.2.1 Silent Mode

If the SBS concludes that the primary channel is occupied after combining the local sensing results from all SUs, the SBS will assign the action “stay silent” to the SUs. In this case, all SUs do nothing until the end of the data transmission phase. Consequently, no reward is attained in this time frame, i.e., $R(t) = 0$. Belief $p(t)$ in the current time frame is estimated by Bayes’ rule, as follows:

$$p(t) = \frac{p(t) P_f}{p(t) P_f + (1 - p(t)) P_d}. \quad (6.8)$$

As a result, belief $p(t+1)$ for the next time frame can be updated as follows:

$$p(t+1) = p(t) P_{VV} + (1 - p(t)) P_{OV}. \quad (6.9)$$

The remaining energy of SU_{in} that can be used for the next time frame is updated by

$$\varepsilon_{in}^{re}(t+1) = \min \left(\varepsilon_{in}^{re}(t) + e_{in}^h(t) - e_s, E_c \right), \quad (6.10)$$

where e_s is the required energy for the sensing process at the beginning of each time frame.

6.4.2.2 Transmission Mode

After the global sensing decision, if the SBS decides that the primary channel is vacant, it allocates the transmission power levels to all SUs. Then, the SUs transmit their data based on the allocated transmission power levels. In this case, there are two possible observations.

Observation 1 (Ω_1): The SBS can not successfully decode the signals transmitted by the SUs because of collisions between SUs and PUs. This means that a misdetection occurred, since the sensing result was wrong. Hence, no reward is achieved in this slot: $R(s(t), a(t) | \Omega_1) = 0$. Belief $p(t+1)$ for the next time frame is updated by

$$p(t+1) = P_{OV}. \quad (6.11)$$

The remaining energy of SU_{in} is updated as

$$\varepsilon_{in}^{re}(t+1) = \min \left(\varepsilon_{in}^{re}(t) - \varepsilon_{in}^{tr}(t) + e_{in}^h(t) - e_s, E_c \right). \quad (6.12)$$

Observation 2 (Ω_2): The SBS successfully decodes all data from the SUs at the end of the time frame. This means that the primary channel was vacant during the time

frame. So the total reward of the system is computed as

$$R(s(t), a(t) | \Omega_2) = \sum_{n=1}^N \sum_{i=1}^2 R_{in}(t), \quad (6.13)$$

where the instant throughput for each SU is calculated by using Eq. (6.3). Belief $p(t+1)$ for the next time frame is updated as follows:

$$p(t+1) = P_{VV}. \quad (6.14)$$

The remaining energy in SU_{in} for the next time frame is updated by

$$\varepsilon_{in}^{re}(t+1) = \min \left(\varepsilon_{in}^{re}(t) - \varepsilon_{in}^{tr}(t) + e_{in}^h(t) - e_s, E_c \right). \quad (6.15)$$

The purpose of reinforcement learning is to find a policy that brings the maximum long-term cumulative reward, called the state action-value function. It indicates how good action \mathbf{a} is in state s . Hence, the expected state-action value function (or Q -value function) can be written as follows [152]

$$\begin{aligned} Q(s, a) &= \mathbb{E} \left[\sum_{k=t}^{\infty} \gamma^{k-t} R(k) \mid s(t) = s, a(t) = a \right], \\ &= \mathbb{E} [R(t) + \gamma Q(s', a') \mid s(t) = s, a(t) = a] \end{aligned} \quad (6.16)$$

where $\mathbb{E}[\cdot]$ represents the expectation, and $\gamma \in [0, 1]$ is the discount factor. Thus, our aim is to find the optimal action, \mathbf{a}^* , in the current time frame to maximize the state-action value function $Q(s, a)$, which is expressed as follows:

$$\mathbf{a}^* = \arg \max_{\mathbf{a} \in \mathbb{A}} \{Q(s, a)\}, \quad (6.17)$$

where \mathbb{A} is the action space of the system. In general, this equation can be tackled by utilizing reinforcement learning approaches (e.g. Q-learning, actor-critic learning). Nevertheless, the computational complexity for traditional Q-learning algorithms is too high if the system has a large state space and action space. Generally, the traditional Q-learning scheme updates its Q-table that consists of the total numbers of states and actions. As the spaces of state and actions are larger, the size of the Q-table will also get very larger. Subsequently, the number of computational operations for calculating the Q-value of each state will be significantly increased. Thus, the complexity of Q-learning notably depends on the number of state and action spaces of the system. Meanwhile, the deep Q-learning scheme can apply the deep neural network to approximate the Q-value of the states, which reduces the number of

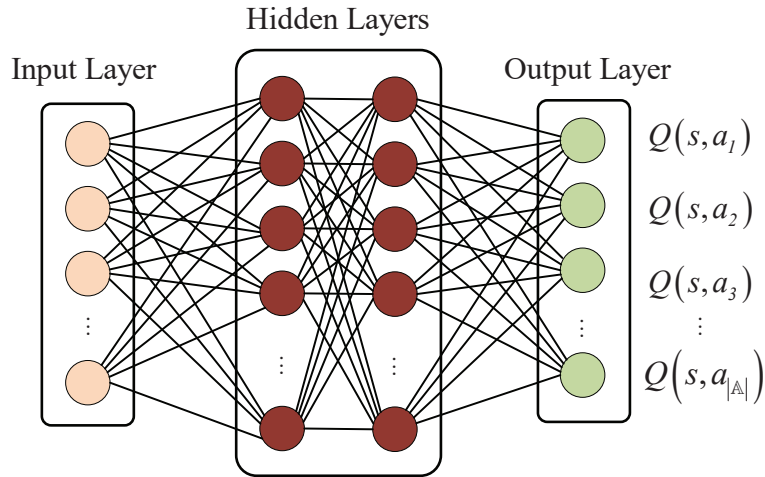


Figure 6.5: Structure of the neural network used for DQL.

computational operations for updating Q-values through each step of training process. Thus, the computational complexity can be significantly improved by using deep Q-learning. In the next subsection, we present deep Q-learning scheme to find the solution of the problem.

6.4.3 Deep Q-Learning

Generally, deep Q-Learning is an approach that combines a value-based method and a neural network. The neural network, considered a DQN, is used to approximate the *Q-value function* for each action based on each given state. Fig. 6.5 illustrates an example of the structure of a neural network including one input layer, two hidden layers, and one output layer. In this work, the input layer of the DQN represents the system state, which includes $(2N + 1)$ elements (a belief about the primary channel and $2N$ remaining energy values for the SUs). There are two hidden layers in the DQN such that each hidden layer is comprised of H neurons. A vector of size $|A|$ is the output of the network, in which each value of the *Q-value function* represents the probability that the system tends to select that action. That means the size of the output vector is the total number of possible actions that we defined in the system. A rectified linear unit (ReLU) function is used as an activation function [205] for each hidden layer, which is expressed as follows:

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{otherwise} \end{cases}, \quad (6.18)$$

Algorithm 6.1 The training process of the proposed DQL-based power allocation scheme

- 1: **Input:** $\mathbb{S}, \mathbb{A}, \alpha, \epsilon^{re}(t), p(t), E_c, e_{avg}^h, T, \epsilon_{ini}, \epsilon_{min}, \epsilon_d$.
 - 2: Initialize network parameter θ , replay memory M , exploration rate $\epsilon = \epsilon_{ini}$
 - 3: **Repeat until convergence**
 - 4: Randomly select an $s(t) \in \mathbb{S}$
 - 5: **for** each time frame $t = 1, 2, \dots, T$:
 - 6: Observe current state $s(t)$
 - 7: With probability ϵ select random action $a \in \mathbb{A}$
 - 8: Otherwise select $a = \arg \max_{a \in \mathbb{A}} Q(s, a, \theta)$
 - 9: Execute action a
 - 10: Attain immediate reward $R(s, a)$ by Eq. (6.4)
 - 11: Store transition $\langle s, a, R(s, a), s' \rangle$ in replay memory M
 - 12: Randomly sample mini-batches of transitions from M
 - 13: **for** each $\langle s, a, R(s, a), s' \rangle$ in mini-batches:
 - 14: **if** s' is the terminal state:
 - 15: $Q_{target} = R(t)$
 - 16: **else**
 - 17: $Q_{target} = R(s, a) + \gamma \max_{a' \in \mathbb{A}} Q(s', a', \theta)$
 - 18: **end if**
 - 19: Perform back-propagation to calculate loss function $L(\theta)$ in Eq. (6.19)
 - 20: Update Q -network parameter θ by minimizing the loss function
 - 21: **end for**
 - 22: Update next state s' , target network parameter θ , exploration rate $\epsilon = \max(\epsilon \cdot \epsilon_d, \epsilon_{min})$
 - 23: **end for**
 - 24: **Output:** Q -network parameter θ .
-

where $x = \sum_{i=1}^{2N+1} \theta_i s_i(t)$ is the estimated output of the layer before applying activation function, in which $s_i(t)$ denotes the i^{th} -element of input state $s(t)$, and θ_i refers to the weight value from the i^{th} input state. In this chapter, the weight parameter of the network, θ , is optimized by adopting stochastic gradient descent with a back-propagation algorithm

to minimize the loss function, which is defined as follows:

$$\mathbf{L}(\boldsymbol{\theta}) = E \left[\begin{array}{c} R(s, a) + \gamma \max_{a' \in \mathbb{A}} Q(s', a', \boldsymbol{\theta}) \\ -Q(s, a, \boldsymbol{\theta}) \end{array} \right]^2, \quad (6.19)$$

where $Q(s, a, \boldsymbol{\theta})$ can be updated as

$$Q(s, a, \boldsymbol{\theta}) \leftarrow Q(s, a, \boldsymbol{\theta}) + \alpha \left[\begin{array}{c} R(s, a) + \gamma \max_{a' \in A} Q(s', a', \boldsymbol{\theta}) \\ -Q(s, a, \boldsymbol{\theta}) \end{array} \right], \quad (6.20)$$

in which $\alpha \in (0, 1]$ is a learning rate. The parameter $\boldsymbol{\theta}$ is updated in the direction of the gradient as

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} Q(s, a, \boldsymbol{\theta}) \delta, \quad (6.21)$$

where δ denotes the temporal different (TD) error, which is calculated by

$$\delta = R(s, a) + \gamma \max_{a' \in A} Q(s', a', \boldsymbol{\theta}) - Q(s, a, \boldsymbol{\theta}) . \quad (6.22)$$

The details of the training process of the proposed DQN are expressed in **Algorithm 6.1**. In each learning step, the SBS selects an action, $\mathbf{a}(t)$, based on an ϵ -greedy method, and then observes and updates the next state, s' . The system will store the transition tuple $\langle s, a, R(s, a), s' \rangle$ in replay memory M . During the training phase, the agent updates the network parameters by using the TD error to enhance system performance. The training process repeats until convergence.

6.5 Simulation Results

6.5.1 Simulation Setting

Throughout the simulation, we considered a cognitive radio network consisting of $2N = 4$ SUs located at different distances from the SBS ($d_1 = 50m$, $d_2 = 100m$, $d_3 = 150m$, and $d_4 = 200m$). There were four layers in the DQN: an input layer, two hidden layers, and an output layer in which each hidden layer has 64 nodes. The learning rate of the Q-network was set at $\alpha = 0.005$. We utilized ReLU function as an activation function for the hidden layers, and a linear function as an activation function for the output layer of the DQN. An adaptive optimization algorithm (i.e. the Adam optimizer) was used to periodically

update the weights of the Q-network after each step. The main parameter settings in the experiments are presented in **Table 6.1**. In this section, we compare the performance of the proposed scheme and other baseline schemes: a Q-learning, a deep Q-learning TDMA only, a Myopic-NOMA/TDMA scheme, a Myopic-TDMA scheme, and a TDMA-Random scheme. These schemes are described as follows in more details.

- Q-learning: The SUs are clustered into a group based on a conventional near-far user pairing algorithm. In this scheme, we also adopt the Q-learning algorithm proposed in [204] that uses both NOMA and TDMA techniques.
- Deep Q-learning TDMA only: In this scheme, the system uses the proposed deep Q-learning algorithm but only TDMA technique without NOMA, where the data transmission duration is divided into the number of slots based on the number of SUs and they respectively transmit data according to their assigned slots.
- Myopic-NOMA/TDMA: The SUs are clustered into a group using a conventional near-far user pairing algorithm. In this scheme, the system uses both NOMA and TDMA techniques. Nevertheless, the SUs are always allocated the optimal transmission power to transmit their data when the primary channel is sensed as vacant, and without considering the future reward.
- Myopic-TDMA: In the data transmission phase of each time frame, the data transmission time is equally divided among all the SUs, and the SUs transmit their data in turn. Similar to the Myopic-NOMA/TDMA, the system only maximizes the instant reward, and thus, the SUs are assigned the maximum transmission power available in their batteries in each slot.
- TDMA-Random: The values for the transmission power of the SUs are randomly allocated, and the SUs transmit their data in turn within the equal data transmission time.

6.5.2 Results and Discussion

The convergence rates of the proposed DQL-based power allocation scheme and the Myopic-NOMA/TDMA scheme through 200 training episodes are shown in Fig. 6.6. The average long-term throughput of the proposed scheme rapidly increases during the first

Table 6.1: Simulation Parameters

Parameter	Description	Value
N	Number of groups	2
T	Time frame duration	200 ms
t_s	Sensing duration	2 ms
E_c	Battery capacity	20 μJ
e_s	Sensing cost	1 μJ
ε^{tr}	Transmission energy	0, 5, 10, 15 μJ
e_{avg}^h	Mean value of harvested energy	5 μJ
p_V	Initial belief that the primary channel is free	0.5
P_{VV}	Transition probability of the primary channel staying in state V	0.8
P_{OV}	Transition probability of the primary channel changing from state O to state V	0.2
P_d	Probability of detection	0.9
P_f	Probability of false alarm	0.1
σ^2	Noise variance	-80 dB
γ	Discount factor	0.9
α	Learning rate	0.001
ϵ	Epsilon rate	1 \rightarrow 0.01
ϵ_d	Epsilon decay	0.9999
L	Number of episodes	200
K	Number of iterations per episode	3000

50 episodes, and gradually converges to the optimal value. It is obvious that since the SBS in the Myopic-NOMA/TDMA scheme always maximizes the instant reward in the current time frame, the average long-term throughput of this scheme remains unchanged throughout the increasing number of training episodes.

In Fig. 6.7, we show the average throughput of the system according to harvested energy that varied from $5\mu J$ to $13\mu J$. We can see that, the performance of all schemes increases as e_{avg}^h goes up because the SUs have more opportunities to transmit data to

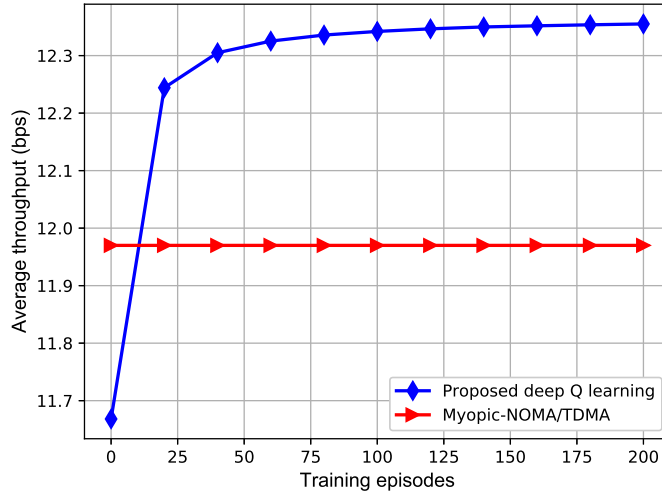


Figure 6.6: The convergence behavior of the proposed scheme.

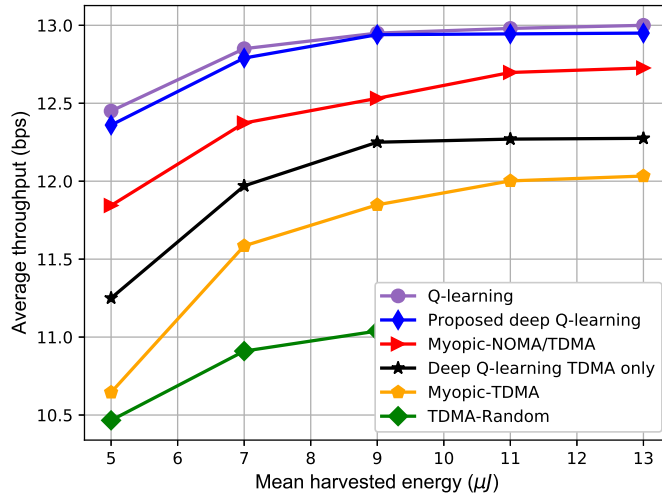


Figure 6.7: The average long-term throughput according to various values for mean harvested energy.

the SBS. Besides, the average throughput of the proposed scheme converges that of the Q-learning scheme. Thus, it can validate the good performance of the deep-Q learning in terms of approximating the Q-value function of Q-learning. Obviously, the proposed scheme outperforms the other schemes since it considers not only the immediate reward but also

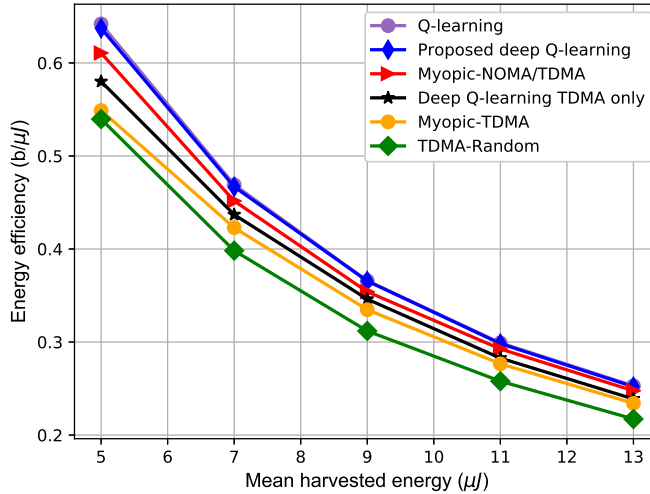


Figure 6.8: Energy efficiency according to various values for mean harvested energy.

the future evolution. On the other hand, the myopic schemes are easily affected by system uncertainties. Meanwhile, the proposed scheme effectively allocates transmission power to the SUs by learning network information through training. Besides, it is worthy that, as the mean harvested energy is high, the average throughput of myopic approaches to that of the proposed deep Q-learning methods. The reason is that, the benefit of power allocation method may be gradually degraded with the increment of available transmit power.

Fig. 6.8 shows the energy efficiency of the schemes according to different values of mean harvested energy. The curves show that the proposed scheme is superior to the other schemes, although the system becomes less efficient in energy utilization with an increment in harvested energy. Fig. 6.9 shows the average throughput of the system versus different values for noise variance and transition probability ($P_{VV} = 0.7$ and $P_{VV} = 0.8$) of the primary channel. The values of noise variance were varied from -80 dB to -65 dB with increments of 5 dB. Obviously, the greater the value for noise variance, the smaller the average throughput of the system. Similarly, as P_{VV} increases, the system throughput goes up. It is because the system has higher probability of detecting the status “vacant” of the primary channel, which leads to more primary channel utilizations of the SUs. Moreover, the proposed DQL-based power allocation approach can achieve about 4%, 14%, and 15% higher average throughput, compared with the Myopic-NOMA/TDMA, Myopic-TDMA,

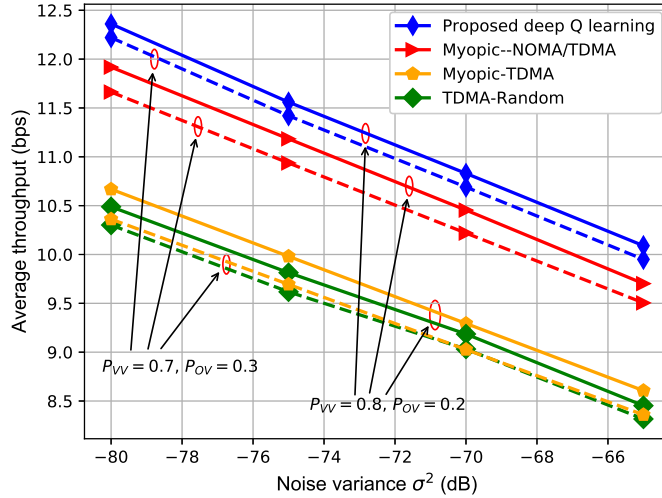


Figure 6.9: Average throughput according to various values for noise variance.

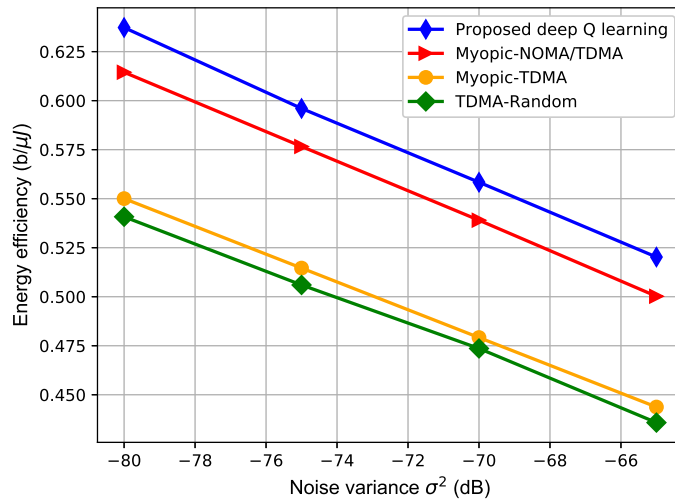


Figure 6.10: Energy efficiency according to various values for noise variance.

and TDMA-Random schemes, respectively.

Fig. 6.10 shows the energy efficiency of the system for the proposed DQL-based power allocation scheme, versus the Myopic-NOMA/TDMA, Myopic-TDMA, and TDMA-Random schemes at various values for noise variance. Similarly, a lower energy efficiency is acquired with increments of σ^2 . The reason is that higher noise variance can reduce more of

the total achievable throughput at the SBS, and thus, the system works less efficiently.

6.6 Conclusion

In this chapter, we investigated an uplink CRNs in which multiple secondary users attempt to transmit data to a secondary base station, and in which the users are powered by solar energy. Deep reinforcement learning for resource allocation is proposed for the secondary users by combining NOMA and TDMA. This chapter aims to maximize the long-term throughput of the system under energy constraints of the SUs. In the proposed scheme, a neural network is utilized to approximate the Q-value function with a large-scale system. The SBS can attain the optimal decision policy by applying the proposed scheme after a number of training time frames. Simulation results demonstrate the advantage of the proposed algorithm in improving the performance of solar-powered cognitive radio networks in long-term operation. In the future work, we aim to use another deep reinforcement learning technique, which is the combination between the value-based method and policy-based method to consider power allocation problem. Furthermore, beyond the NOMA/TDMA technique studied in this chapter, it can be interesting future work to apply for more advanced systems with multi-channel and multi-agent using other reinforcement learning algorithms.

Chapter 7

Summary of Contributions and Future Works

7.1 Introduction

Previous chapters have presented the research motivations, the problems, and solutions regarding the applications of the accomplished research . This chapter summarizes the main contributions of this dissertation and discusses future research directions.

7.2 Summary of Contributions

This dissertation discusses the applications of machine learning techniques, such as reinforcement learning and deep learning, in wireless communication networks, which intend to enhance the long-term performance of the network. The main contributions of this research are summarized as follows:

Firstly, we consider the jamming attack scenario in cooperative communication, where a jammer intends to block the direct transmission link. The behavior of the jammer is assumed to follow the Markov chain model. In addition, the relay is used to help the source to forward the data to the destination and it has definite capacity in its battery. Hence, the energy harvesting technique is applied to solve the energy-constrained problem at the relay. Moreover, the imperfection of SS mechanism to detect the jammer at the source is also considered. We propose a POMDP-based scheme at the source node to determine the operation mode of the relay considering its remaining energy and the sensing result to

improve the achievable throughput of the network. The objective of this work is to find the optimal action according to the optimal policy in order to maximize the long-term throughput of the network in the presence of the jamming attack.

Secondly, we studied the ambient backscatter technique for the SUs in wireless-powered CRN. In such a network, the ST can simultaneously harvest non-RF energy from the ambient environment and perform backscattering/RF harvesting/transmitting for the SR in the data communication phase. The ST usually performs spectrum sensing to check whether the primary channel is free or not. Subsequently, based on the sensing result, the ST can effectively select its proper operation and allocate power for data transmission to maximize the accumulative discounted reward of the secondary system. The problem was formulated according to the framework of POMDP in a time-slotted fashion for the secondary system to achieve the optimal policy. The simulation results validated that our proposed scheme can efficiently provide a high long-term transmission rate for the secondary system due to efficient utilization of energy harvesting from the wireless environment.

Thirdly, we investigated a CRN with uplink NOMA, where the SUs are allowed to simultaneously access the same channel at the same time. In addition, the energy-constrained and imperfect sensing issues of the SUs are also taken into account. We formulated the optimization problem as a Markov decision process. Afterward, an actor-critic reinforcement learning approach was employed such that the CBS can adaptively interact with the environment to find the optimal solution for maximizing the system rewards. The simulation results validated that our proposed scheme can efficiently improve both the throughput and energy utilization in the long run.

Next, we studied a model of a hybrid NOMA/OMA uplink CRN adopting energy harvesting at the CUs. We consider power and bandwidth allocation such that each CU is able to optimally utilize the power and bandwidth under energy constraint and uncertain environmental conditions. The problem was first formulated as the framework of MDP, then we applied a deep actor-critic reinforcement learning framework to obtain the optimal policy. More specifically, we exploited deep neural networks to approximate the policy and value functions, which allowed the algorithm to work with large state space and action space. Consequently, the CBS can allocate the appropriate transmission power and bandwidth to the CUs by interacting with the environment. The simulation results verified the advantages of our proposed scheme in improving network performance under various network conditions in the long run.

Finally, we consider an uplink CRN where multiple SUs attempt to transmit data to an SBS by opportunistically using a licensed channel of the primary system. The proposed NOMA/TDMA-based deep Q-learning approach was proposed to maximize the throughput of a secondary system in the long run. In particular, the DNN is used to approximate the value function of every state-action pair. After training, the agent of the DQL algorithm is able to allocate transmission powers to each CU based on the current state of the network. The simulation results showed the average throughput attained by the proposed scheme was significantly improved compared to that of conventional schemes.

7.3 Future Directions

For future research directions regarding to machine learning-based techniques for enhancing wireless network performance, we consider several aspects as follows:

Due to massive advancement in technologies such as large and distributed antenna arrays, ultra-dense network, software-based networks. The low latency and high reliability requirements are essential needs for future applications, which requires extensive research on machine learning applications. Thus, further studies must to be carried out to enhance the efficiency of RL techniques and guarantee the advantage in the high mobility of radio environments. However, improving low latency and high reliability under transmission rate guarantee results in many network challenges. These parameters are not able to be compatible since when one of them is improved, the other two will be detrimental. Deep reinforcement learning can be a promising solution for these issues by efficiently allocating resources to balance the data rate, reliability and latency trade off. More specifically, the DQL framework allows the system to learn knowledge about the network to select intelligent decisions for maximizing the network performance.

In mobile crowd sensing, mobile users supply their sensing data to a crowd sensing service provider and receive a reward in exchange. However, the mobile devices have to decide on whether upload their data to the provider or not, and also how much data to be uploaded due to resource limitation and scarcity. Hence, the provider has to determine the given amount of reward that should be appropriate to according actions of the mobile users. Due to enormous number of users and uncertain environment, DRL can be applied to obtain an optimal crowd sensing policy for wireless system.

Nowadays, the huge number of vehicles has induced traffic delay, safety risks

and accidents in the transportation system. It is hardly enable to control the traffic load due to the high traffic density. IoT-enable autonomous transportation system (ATS) has emerged as a promising solution to provide highly efficient, flexible, and smarter approach for transportation. The ATS operates its actions based on DRL and infrastructure of IoT. The agent of DRL, which is a vehicle, carries out certain action in the environment (a traffic scenario). After every action, the agent moves from one state to another state and receives a reward. The DRL is designed to enhance decision making parameters in emergency transportation. The objective of IoT-enable ATS is to minimize total moving time, energy consumption, and environment pollution while increasing the safety. It is really meaningful in case of emergency healthcare service, where it can reasonably increase delivery of services and also reducing the overall burden on healthcare organizations.

Publications

International Journals

- [1] Hoang Thi Huong Giang, Hiep Vu-Van, and Insoo Koo, "POMDP-Based Throughput Maximization for Cooperative Communications Networks with Energy-Constrained Relay under Attack in the Physical Layer," *Applied Sciences*, vol. 8, no. 10, Oct. 2018.
- [2] Hoang Thi Huong Giang, Tran Nhut Khai Hoan, Pham Duy Thanh, and Insoo Koo, "A POMDP-based Long-term Transmission Rate Maximization for Cognitive Radio Networks with Wireless-Powered Ambient Backscatter," *International Journal of Communication Systems*, vol. 32, no. 12, Aug. 2019.
- [3] Hoang Thi Huong Giang, Tran Nhut Khai Hoan, and Insoo Koo, "Uplink NOMA-based Long-Term Throughput Maximization Scheme for Cognitive Radio Networks: An Actor-Critic Reinforcement Learning Approach," *Wireless Networks*, vol. 27, no. 2, pp. 1319-1334, Jan. 2021.
- [4] Hoang Thi Huong Giang, Tran Nhut Khai Hoan, Pham Duy Thanh, and Insoo Koo, "Hybrid NOMA/OMA-Based Dynamic Power Allocation Scheme Using Deep Reinforcement Learning in 5G Networks," *Applied Sciences*, vol. 10, no. 12, June 2020.
- [5] Hoang Thi Huong Giang, Pham Duy Thanh, and Insoo Koo, "Deep Q-learning-based Resource Allocation for Solar-Powered Users in Cognitive Radio Networks," *ICT Express*, vol. 7, Issue 1, pp. 49-59, March 2021.
- [6] Pham Duy Thanh, Tran Nhut Khai Hoan, Hoang Thi Huong Giang, and Insoo Koo, "Cache-Enabled Data Rate Maximization for Solar-Powered UAV Communication Systems," *Electronics*, vol. 9, no. 11, pp. 2050-2059, Mar. 2020.

Conferences

- [7] Hoang Thi Huong Giang, Pham Duy Thanh, and Insoo Koo, "Dynamic Power Allocation Scheme for NOMA Uplink in Cognitive Radio Networks Using Deep Q Learning," *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2020, Jeju, Korea.

- [8] Pham Duy Thanh, Hoang Thi Huong Giang, and Insoo Koo, "UAV-assisted NOMA Downlink Communications Based on Content Caching," *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2020, Jeju, Korea.

Bibliography

- [1] S. Force, “Spectrum policy task force report,” *Federal Communications Commission ET Docket 02*, vol. 135, 2002.
- [2] A. Ali and W. Hamouda, “Low power wideband sensing for one-bit quantized cognitive radio systems,” *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 16–19, 2016.
- [3] O. Altrad, S. Muhaidat, A. Al-Dweik, A. Shami, and P. D. Yoo, “Opportunistic spectrum access in cognitive radio networks under imperfect spectrum sensing,” *IEEE Transactions on Vehicular Technology*, vol. 63, no. 2, pp. 920–925, 2014.
- [4] S. Haykin, “Cognitive radio: brain-empowered wireless communications,” *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, Feb. 2005.
- [5] A. Nosratinia, T. E. Hunter, and A. Hedayat, “Cooperative communication in wireless networks,” *IEEE Communications Magazine*, vol. 42, no. 10, pp. 74–80, Oct. 2004.
- [6] A. F. M. S. Shah and M. S. Islam, “A survey on cooperative communication in wireless networks,” *International Journal of Intelligent Systems and Applications(IJISA)*, vol. 6, no. 7, pp. 66–78, June 2014.
- [7] X. Tao, X. Xu, and Q. Cui, “An overview of cooperative communications,” *IEEE Communications Magazine*, vol. 50, no. 6, pp. 65–71, June 2012.
- [8] T. Cover and A. E. Gamal, “Capacity theorems for the relay channel,” *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, Sept. 1979.
- [9] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, “Cooperative diversity in wireless networks: Efficient protocols and outage behavior,” *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

- [10] J. N. Laneman and G. W. Wornell, "Distributed space-time coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003.
- [11] Y. Zou, J. Zhu, X. Wang, and L. Hanzo, "A survey on wireless security: Technical challenges, recent advances and future trends," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1727–1765, Sept. 2016.
- [12] I. Tomić and J. A. McCann, "A survey of potential security issues in existing wireless sensor network protocols," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1910–1923, Dec. 2017.
- [13] D. Panagiotis D, P. Koralia N, K. George K, X. Hong, and N. Arumugam, "Joint downlink/uplink design for wireless powered networks with interference," *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, vol. 3, no. 2, pp. 202–207, April 2013.
- [14] R. Pickholtz, D. Schilling, and L. Milstein, "Theory of spread-spectrum communications - a tutorial," *IEEE Transactions on Communications*, vol. 30, no. 5, pp. 855–884, May 1982.
- [15] J. Huang, G. Chang, and J. Huang, "Anti-jamming rendezvous scheme for cognitive radio networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 3, pp. 648–661, March 2017.
- [16] M. Strasser, C. Popper, S. Capkun, and M. Cagalj, "Jamming-resistant key establishment using uncoordinated frequency hopping," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008, pp. 64–78.
- [17] J. Paek, "Fast and adaptive mesh access control in low-power and lossy networks," *IEEE Internet of Things Journal*, vol. 2, no. 5, pp. 435–444, Oct. 2015.
- [18] Y. Desmedt, R. Safavi-Naini, Huaxiong Wang, C. Charney, and J. Pieprzyk, "Broadcast anti-jamming systems," in *IEEE International Conference on Networks. ICON '99 Proceedings (Cat. No.PR00243)*, 1999, pp. 349–355.

-
- [19] C. Popper, M. Strasser, and S. Capkun, “Anti-jamming broadcast communication using uncoordinated spread spectrum techniques,” *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 5, pp. 703–715, June 2010.
- [20] A. Chorti, “Overcoming limitations of secret key generation in block fading channels under active attacks,” in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2016, pp. 1–5.
- [21] B. Awerbuch, A. Richa, and C. Scheideler, “A jamming-resistant MAC protocol for single-hop wireless networks,” in *Proceedings of the Twenty-Seventh ACM Symposium on Principles of Distributed Computing*, 2008, pp. 45–54.
- [22] X. Liu, G. Noubir, R. Sundaram, and S. Tan, “Spread: Foiling smart jammers using multi-layer agility,” in *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, 2007, pp. 2536–2540.
- [23] V. Navda, A. Bohra, S. Ganguly, and D. Rubenstein, “Using channel hopping to increase 802.11 resilience to jamming attacks,” in *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, 2007, pp. 2526–2530.
- [24] Y. Chen, J. Yang, W. Trappe, and R. P. Martin, “Detecting and localizing identity-based attacks in wireless and sensor networks,” *IEEE Transactions on Vehicular Technology*, vol. 59, no. 5, pp. 2418–2434, June 2010.
- [25] T. Cheng, P. Li, and S. Zhu, “Multi-jammer localization in wireless sensor networks,” in *2011 Seventh International Conference on Computational Intelligence and Security*, 2011, pp. 736–740.
- [26] P. D. Thanh, H. Vu-Van, and I. Koo, “Efficient channel selection and routing algorithm for multihop, multichannel cognitive radio networks with energy harvesting under jamming attacks,” *Security and Communication Networks*, vol. 2018, pp. 1–13, March 2018.
- [27] H. Soleimani, S. Tomasin, T. Alizadeh, and M. Shojafar, “Cluster-head based feedback for simplified time reversal prefiltering in ultra-wideband systems,” *Physical Communication*, vol. 25, pp. 100–109, 2017.

- [28] N. Cordeschi, D. Amendola, M. Shojafar, and E. Baccarelli, "Distributed and adaptive resource management in cloud-assisted cognitive radio vehicular networks with hard reliability guarantees," *Vehicular Communications*, vol. 2, no. 1, pp. 1–12, Jan. 2015.
- [29] A. Bhowmick, S. D. Roy, and S. Kundu, "Performance of secondary user with combined RF and non-RF based energy-harvesting in cognitive radio network," in *2015 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, 2015, pp. 1–3.
- [30] A. Bhowmick, K. Yadav, S. D. Roy, and S. Kundu, "Throughput of an energy harvesting cognitive radio network based on prediction of primary user," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8119–8128, Sept. 2017.
- [31] C. Zhai, J. Liu, and L. Zheng, "Cooperative spectrum sharing with wireless energy harvesting in cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5303–5316, July 2016.
- [32] H. Vu-Van and I. Koo, "Optimal throughput for cognitive radio with energy harvesting in fading wireless channel," *The Scientific Word Journal*, vol. 2014, pp. 1–7, 2014.
- [33] W. Cadeau, X. Li, and C. Xiong, "Markov model based jamming and anti-jamming performance analysis for cognitive radio networks," *Communications and Network*, vol. 6, pp. 76–85, 2014.
- [34] D. P. Bertsekas, *Dynamic programming and optimal control 2nd edition*, 2001, vol. 1, no. 2.
- [35] R. Tandra, S. Mishra, and A. Sahai, "What is a spectrum hole and what does it take to recognize one?" *Proceedings of the IEEE*, vol. 97, no. 5, pp. 824–848, 2009.
- [36] Y. Hur, J. Park, W. Woo, K. Lim, C.-H. Lee, H. Kim, and J. Laskar, "A wideband analog multi-resolution spectrum sensing (mrss) technique for cognitive radio (cr) systems," in *2006 IEEE International Symposium on Circuits and Systems*, 2006, pp. 4–pp.
- [37] H. Wang, Y. Yao, X. Zhang, and H. Li, "Secondary user access control in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 11, pp. 2866–2873, 2016.

- [38] H. Salameh, M. Krunz, and O. Younis, "Cooperative adaptive spectrum sharing in cognitive radio networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 18, no. 4, pp. 1181–1194, 2010.
- [39] M. El Tanab and W. Hamouda, "Resource allocation for underlay cognitive radio networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1249–1276, 2017.
- [40] F. Wang and X. Zhang, "Resource allocation for multiuser cooperative overlay cognitive radio networks with RF energy harvesting capability," in *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6.
- [41] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 1, pp. 418–428, 2014.
- [42] S. Lee and R. Zhang, "Cognitive wireless powered network: Spectrum sharing models and throughput maximization," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 3, pp. 335–346, 2015.
- [43] S. Bi, Y. Zeng, and R. Zhang, "Wireless powered communication networks: an overview," *IEEE Wireless Communications*, vol. 23, no. 2, pp. 10–18, 2016.
- [44] S. Park, H. Kim, and D. Hong, "Cognitive radio networks with energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1386–1397, 2013.
- [45] A. Bhowmick, S. Roy, and S. Kundu, "Throughput of a cognitive radio network with energy-harvesting based on primary user signal," *IEEE Wireless Communications Letters*, vol. 5, no. 2, pp. 136–139, 2016.
- [46] X. Lu, P. Wang, D. Niyato, and E. Hossain, "Dynamic spectrum access in cognitive radio networks with RF energy harvesting," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 102–110, 2014.
- [47] U. Olgun, C. Chen, and J. Volakis, "Design of an efficient ambient wifi energy harvesting system," *IET Microwaves, Antennas Propagation*, vol. 6, no. 11, pp. 1200–1206, 2012.
- [48] R. Vullers, R. Schaijk, I. Doms, C. Hoof, and R. Mertens, "Micropower energy harvesting," *Solid-State Electronics*, vol. 53, no. 7, pp. 684 – 693, 2009.

- [49] Y. Shi, L. Xie, Y. Hou, and H. Sherali, "On renewable sensor networks with wireless energy transfer," in *2011 Proceedings IEEE INFOCOM*, 2011, pp. 1350–1358.
- [50] K. Huang and V. Lau, "Enabling wireless power transfer in cellular networks: Architecture, modeling and deployment," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 902–912, 2014.
- [51] S. Lee, R. Zhang, and K. Huang, "Opportunistic wireless energy harvesting in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4788–4799, 2013.
- [52] T. Hoan and I. Koo, "Multi-slot spectrum sensing schedule and transmitted energy allocation in harvested energy powered cognitive radio networks under secrecy constraints," *IEEE Sensors Journal*, vol. 17, no. 7, pp. 2231–2240, 2017.
- [53] V. Rakovic, D. Denkovski, Z. Hadzi-Velkov, and L. Gavrilovska, "Optimal time sharing in underlay cognitive radio systems with RF energy harvesting," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 7689–7694.
- [54] B. Nguyen, H. Jung, D. Har, and K. Kim, "Performance analysis of a cognitive radio network with an energy harvesting secondary transmitter under nakagami- m fading," *IEEE Access*, vol. 6, pp. 4135–4144, 2018.
- [55] D. Hoang, D. Niyato, P. Wang, D. Kim, and Z. Han, "Ambient backscatter: A new approach to improve network performance for RF-powered cognitive radio networks," *IEEE Transactions on Communications*, vol. 65, no. 9, pp. 3659–3674, 2017.
- [56] H. Stockman, "Communication by means of reflected power," *Proceedings of the IRE*, vol. 36, no. 10, pp. 1196–1204, 1948.
- [57] G. Vannucci, A. Bletsas, and D. Leigh, "A software-defined radio system for backscatter sensor networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2170–2179, 2008.
- [58] A. Bletsas, S. Siachalou, and J. Sahalos, "Anti-collision backscatter sensor networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 5018–5029, 2009.

- [59] J. Kimionis, A. Bletsas, and J. Sahalos, "Bistatic backscatter radio for power-limited sensor networks," in *2013 IEEE Global Communications Conference (GLOBECOM)*, 2013, pp. 353–358.
- [60] J. Griffin and G. Durgin, "Complete link budgets for backscatter-radio and RFID systems," *IEEE Antennas and Propagation Magazine*, vol. 51, no. 2, pp. 11–25, 2009.
- [61] P. Zhang and J. Gummesson, Ganesan, "Blink: A high throughput link layer for backscatter communication," in *In MobiSys*, 2012, pp. 99–112.
- [62] N. Huynh, D. Hoang, X. Lu, D. Niyato, P. Wang, and D. Kim, "Ambient backscatter communications: A contemporary survey," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 2889–2922, 2018.
- [63] J. Kimionis, A. Bletsas, and J. Sahalos, "Increased range bistatic scatter radio," *IEEE Transactions on Communications*, vol. 62, no. 3, pp. 1091–1104, 2014.
- [64] S. Choi and D. Kim, "Backscatter radio communication for wireless powered communication networks," in *2015 21st Asia-Pacific Conference on Communications (APCC)*, 2015, pp. 370–374.
- [65] B. Kellogg, A. Parks, S. Gollakota, J. Smith, and D. Wetherall, "Wi-fi backscatter: internet connectivity for RF-powered devices," in *SIGCOMM*, 2014.
- [66] V. Liu, A. Parks, V. Talla, S. Gollakota, D. Wetherall, and J. Smith, "Ambient backscatter: wireless communication out of thin air," in *SIGCOMM*, 2013.
- [67] D. Dobkin, *The RF in RFID: Passive UHF RFID in Practice*. Newton, MA, USA: Newnes, 2007, ISBN 9780750682091.
- [68] C. Boyer and S. Roy, "Backscatter communication and RFID: Coding, energy, and MIMO analysis," *IEEE Transactions on Communications*, vol. 62, no. 3, pp. 770–785, 2014.
- [69] G. Yang, Q. Zhang, and Y. Liang, "Cooperative ambient backscatter communications for green internet-of-things," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1116–1130, 2018.

- [70] A. Parks, A. Liu, S. Gollakota, and J. Smith, “Turbocharging ambient backscatter communication,” *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 619–630, 2014.
- [71] C. Pérez-Penichet, A. Varshney, F. Hermans, C. Rohner, and T. Voigt, “Do multiple bits per symbol increase the throughput of ambient backscatter communications?” in *Proceedings of the 2016 International Conference on Embedded Wireless Systems and Networks*, USA, 2016, pp. 355–360, ISBN 978-0-9949886-0-7.
- [72] J. Qian, F. Gao, and G. Wang, “Signal detection of ambient backscatter system with differential modulation,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 3831–3835.
- [73] J. You, G. Wang, and Z. Zhong, “Physical layer security-enhancing transmission protocol against eavesdropping for ambient backscatter communication system,” in *6th International Conference on Wireless, Mobile and Multi-Media (ICWMMN 2015)*, 2015, pp. 43–47.
- [74] S. Kim and D. Kim, “Hybrid backscatter communication for wireless-powered heterogeneous networks,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6557–6570, 2017.
- [75] K. Park, D. Munir, J. Kim, and M. Chung, “Integrating RF-powered backscatter with underlay cognitive radio networks,” in *2017 International Conference on Information Networking (ICOIN)*, 2017, pp. 288–292.
- [76] B. Lyu, H. Guo, Z. Yang, and G. Gui, “Throughput maximization for hybrid backscatter assisted cognitive wireless powered radio networks,” *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2015–2024, 2018.
- [77] B. Lyu, C. You, Z. Yang, and G. Gui, “The optimal control policy for RF-powered backscatter communication networks,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2804–2808, 2018.
- [78] Y. Liang, Y. Zeng, E. Peh, and A. Hoang, “Sensing-throughput tradeoff for cognitive radio networks,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, pp. 1326–1337, 2008.

- [79] C. Papadimitriou and J. Tsitsiklis, "The complexity of markov decision processes," *Math. Oper. Res.*, vol. 12, no. 3, pp. 441–450, 1987.
- [80] Y. Lu and A. Duel-Hallen, "Channel-adaptive sensing strategy for cognitive radio ad hoc networks," in *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, 2013, pp. 466–471.
- [81] F. A. Khan, T. Ratnarajah, and M. Sellathurai, "Multiuser diversity analysis in spectrum sharing cognitive radio networks," in *2010 Proceedings of the Fifth International Conference on Cognitive Radio Oriented Wireless Networks and Communications*, 2010, pp. 1–5.
- [82] C. Wang, F. Haider, X. Gao, X. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, Feb. 2014.
- [83] X. Hong, J. Wang, C. Wang, and J. Shi, "Cognitive radio in 5G: a perspective on energy-spectral efficiency trade-off," *IEEE Communications Magazine*, vol. 52, no. 7, pp. 46–53, July 2014.
- [84] B. Wang and K. J. R. Liu, "Advances in cognitive radio networks: A survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 1, pp. 5–23, Feb. 2011.
- [85] I. F. Akyildiz, W. Lee, M. C. Vuran, and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 40–48, April 2008.
- [86] J. Mitola and G. Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13–18, Aug. 1999.
- [87] E. Hossain, D. Niyato, and Z. Han, *Dynamic Spectrum Access and Management in Cognitive Radio Networks*. Cambridge University Press, 2009.
- [88] L. Lv, J. Chen, Q. Ni, Z. Ding, and H. Jiang, "Cognitive non-orthogonal multiple access with cooperative relaying: A new wireless frontier for 5G spectrum sharing," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 188–195, April 2018.

- [89] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 894–914, May 2009.
- [90] H. T. H. Giang, T. N. K. Hoan, P. D. Thanh, and I. Koo, "A POMDP-based long-term transmission rate maximization for cognitive radio networks with wireless-powered ambient backscatter," *International Journal of Communication Systems*, vol. 32, no. 12, p. e3993, 2019, e3993 dac.3993.
- [91] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. ElKashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, Feb. 2017.
- [92] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, Sept. 2015.
- [93] W. Zhiqiang, Y. Jinhong, D. W. K. Ng, M. ElKashlan, and D. Zhiguo, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," *ZTE Communications*, vol. 14, no. 4, pp. 17–25, Oct. 2016.
- [94] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, 2013, pp. 1–5.
- [95] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.
- [96] D. Wan, M. Wen, F. Ji, H. Yu, and F. Chen, "Non-orthogonal multiple access for cooperative communications: Challenges, opportunities, and trends," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 109–117, April 2018.
- [97] H. Tabassum, E. Hossain, and J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using poisson cluster processes," *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3555–3570, Aug. 2017.

- [98] R. Razavi, R. Hoshyar, M. A. Imran, and Y. Wang, "Information theoretic analysis of LDS scheme," *IEEE Communications Letters*, vol. 15, no. 8, pp. 798–800, Aug. 2011.
- [99] M. AL-Imari, M. A. Imran, and R. Tafazolli, "Low density spreading for next generation multicarrier cellular systems," in *2012 International Conference on Future Communication Networks*, 2012, pp. 52–57.
- [100] Y. Du, B. Dong, Z. Chen, J. Fang, and X. Wang, "A fast convergence multiuser detection scheme for uplink SCMA systems," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 388–391, Aug. 2016.
- [101] H. Nikopour, E. Yi, A. Bayesteh, K. Au, M. Hawryluck, H. Baligh, and J. Ma, "SCMA for downlink multiple access of 5G wireless networks," in *2014 IEEE Global Communications Conference*, 2014, pp. 3940–3945.
- [102] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10 152–10 157, Dec. 2016.
- [103] D. Wang and S. Men, "Secure energy efficiency for NOMA based cognitive radio networks with nonlinear energy harvesting," *IEEE Access*, vol. 6, pp. 62 707–62 716, Oct. 2018.
- [104] L. Lv, Q. Ni, Z. Ding, and J. Chen, "Application of non-orthogonal multiple access in cooperative spectrum-sharing networks over nakagami- m fading channels," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5506–5511, June 2017.
- [105] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Transactions on Communications*, vol. 65, no. 6, pp. 2641–2656, June 2017.
- [106] F. I. Simjee and P. H. Chou, "Efficient charging of supercapacitors for extended lifetime of wireless sensor nodes," *IEEE Transactions on Power Electronics*, vol. 23, no. 3, pp. 1526–1536, May 2008.
- [107] Z. Chen, M. Law, P. Mak, and R. P. Martins, "A single-chip solar energy harvesting IC using integrated photodiodes for biomedical implant applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, no. 1, pp. 44–53, Feb. 2017.

- [108] C. Wang, J. Li, Y. Yang, and F. Ye, "Combining solar energy harvesting with wireless charging for hybrid wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 3, pp. 560–576, March 2018.
- [109] J. Stuyts, G. Horn, W. Vandermeulen, J. Driesen, and M. Diehl, "Effect of the electrical energy conversion on optimal cycles for pumping airborne wind energy," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 1, pp. 2–10, Jan. 2015.
- [110] L. Zhao, L. Tang, J. Liang, and Y. Yang, "Synergy of wind energy harvesting and synchronized switch harvesting interface circuit," *IEEE/ASME Transactions on Mechatronics*, vol. 22, no. 2, pp. 1093–1103, April 2017.
- [111] A. Celik, A. Alsharoa, and A. E. Kamal, "Hybrid energy harvesting-based cooperative spectrum sensing and access in heterogeneous cognitive radio networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 1, pp. 37–48, March 2017.
- [112] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, "Joint power control and beamforming for uplink non-orthogonal multiple access in 5G millimeter-wave communications," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 6177–6189, Sept. 2018.
- [113] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.
- [114] D. Zhai and J. Du, "Spectrum efficient resource management for multi-carrier-based NOMA networks: A graph-based method," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 388–391, June 2018.
- [115] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Communications Surveys Tutorials*, vol. 11, no. 1, pp. 116–130, First Quarter 2009.
- [116] S. Shankar N., C. Cordeiro, and K. Challapali, "Spectrum agile radios: utilization and sensing architectures," in *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005.*, 2005, pp. 160–169.

- [117] P. Pawelczak, G. J. M. Janssen, and R. V. Prasad, "WLC10-4: Performance measures of dynamic spectrum access networks," in *IEEE Globecom 2006*, 2006, pp. 1–6.
- [118] X. Liu and S. S. N., "Sensing-based opportunistic channel access," *Mobile Networks and Applications*, vol. 11, no. 4, pp. 577–591, 2006.
- [119] K. Cichoń, A. Kliks, and H. Bogucka, "Energy-efficient cooperative spectrum sensing: A survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1861–1886, Third Quarter 2016.
- [120] Z. Quan, W. Ma, S. Cui, and A. H. Sayed, "Optimal linear fusion for distributed detection via semidefinite programming," *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2431–2436, April 2010.
- [121] F. C. Ribeiro, M. L. R. de Campos, and S. Werner, "Distributed cooperative spectrum sensing with adaptive combining," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 3557–3560.
- [122] W. Han, J. Li, Z. Li, J. Si, and Y. Zhang, "Efficient soft decision fusion rule in cooperative spectrum sensing," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 1931–1943, April 2013.
- [123] P. D. Thanh, T. N. K. Hoan, and I. Koo, "Joint resource allocation and transmission mode selection using a POMDP-based hybrid half-duplex/full-duplex scheme for secrecy rate maximization in multi-channel cognitive radio networks," *IEEE Sensors Journal*, vol. 20, no. 7, pp. 3930–3945, April 2020.
- [124] C. R. Stevenson, G. Chouinard, Z. Lei, W. Hu, S. J. Shellhammer, and W. Caldwell, "IEEE 802.22: The first cognitive radio wireless regional area network standard," *IEEE Communications Magazine*, vol. 47, no. 1, pp. 130–138, Jan. 2009.
- [125] R. H. Crites and A. G. Barto, "An actor/critic algorithm that is equivalent to Q-learning," in *Advances in Neural Information Processing Systems*, 1995, pp. 401–408.
- [126] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.

- [127] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, “Convergence results for single-step on-policy reinforcement-learning algorithms,” *Machine learning*, vol. 38, no. 3, pp. 287–308, 2000.
- [128] J. V. Stone, “Bayes’ rule: A tutorial introduction to bayesian analysis,” 2013.
- [129] V. R. Konda and J. N. Tsitsiklis, “Actor-critic algorithms,” in *Advances in neural information processing systems*. Citeseer, 2000, pp. 1008–1014.
- [130] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, “Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges,” *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 721–742, Second Quarter 2017.
- [131] P. D. Thanh, T. N. K. Hoan, H. Vu-Van, and I. Koo, “Efficient attack strategy for legitimate energy-powered eavesdropping in tactical cognitive radio networks,” *Wireless Networks*, vol. 25, no. 6, pp. 3605–3622, 2019.
- [132] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. Kwak, “Resource allocation for downlink NOMA systems: Key techniques and open issues,” *IEEE Wireless Communications*, vol. 25, no. 2, pp. 40–47, April 2018.
- [133] W. Yu, L. Musavian, and Q. Ni, “Link-layer capacity of NOMA under statistical delay QoS guarantees,” *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4907–4922, Oct. 2018.
- [134] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, “On the sum rate of MIMO-NOMA and MIMO-OMA systems,” *IEEE Wireless Communications Letters*, vol. 6, no. 4, pp. 534–537, Aug. 2017.
- [135] B. Di, L. Song, and Y. Li, “Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.
- [136] S. Timotheou and I. Krikidis, “Fairness for non-orthogonal multiple access in 5G systems,” *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.
- [137] W. Liang, Z. Ding, Y. Li, and L. Song, “User pairing for downlink non-orthogonal multiple access networks using matching algorithm,” *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5319–5332, Dec. 2017.

- [138] Y. Zhang, H. Wang, T. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2852–2857, March 2017.
- [139] W. Hao, M. Zeng, Z. Chu, and S. Yang, "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 782–785, Dec. 2017.
- [140] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3722–3732, Sept. 2016.
- [141] T. Lv, Y. Ma, J. Zeng, and P. T. Mathiopoulos, "Millimeter-wave NOMA transmission in cellular M2M communications for internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1989–2000, June 2018.
- [142] 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, March 2018.
- [143] W. S. H. M. W. Ahmad, N. A. M. Radzi, F. S. Samidi, A. Ismail, F. Abdullah, M. Z. Jamaludin, and M. N. Zakaria, "5G technology: Towards dynamic spectrum sharing using cognitive radio networks," *IEEE Access*, vol. 8, pp. 14 460–14 488, 2020.
- [144] M. Amjad, L. Musavian, and M. H. Rehmani, "Effective capacity in wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3007–3038, Four Quarter 2019.
- [145] F. Zhou, N. C. Beaulieu, Z. Li, J. Si, and P. Qi, "Energy-efficient optimal power allocation for fading cognitive radio channels: Ergodic capacity, outage capacity, and minimum-rate capacity," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2741–2755, April 2016.
- [146] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Wireless networks with RF energy harvesting: A contemporary survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 757–789, Second Quarter 2015.

- [147] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5G systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, March 2016.
- [148] Z. Ni, Z. Chen, Q. Zhang, and C. Zhou, "Analysis of RF energy harvesting in uplink-NOMA IoT-based network," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019, pp. 1–5.
- [149] C. M. Gabriel Gussen, E. V. Belmega, and M. Debbah, "Pricing and bandwidth allocation problems in wireless multi-tier networks," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2011, pp. 1633–1637.
- [150] S. Arunthavanathan, S. Kandeepan, and R. J. Evans, "A markov decision process-based opportunistic spectral access," *IEEE Wireless Communications Letters*, vol. 5, no. 5, pp. 544–547, Aug. 2016.
- [151] H. Xiao, K. Yang, X. Wang, and H. Shao, "A robust mdp approach to secure power control in cognitive radio networks," in *2012 IEEE International Conference on Communications (ICC)*, 2012, pp. 4642–4647.
- [152] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [153] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, "TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2000–2011, April 2014.
- [154] R. H. Puspita, S. D. A. Shah, G. Lee, B. Roh, J. Oh, and S. Kang, "Reinforcement learning based 5G enabled cognitive radio networks," in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, 2019, pp. 555–558.
- [155] X. Meng, H. Inaltekin, and B. Krongold, "Deep reinforcement learning-based power control in full-duplex cognitive radio networks," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–7.

- [156] K. S. H. Ong, Y. Zhang, and D. Niyato, "Cognitive radio network throughput maximization with deep reinforcement learning," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019, pp. 1–5.
- [157] H. Zhang, N. Yang, W. Huangfu, K. Long, and V. C. M. Leung, "Power control based on deep reinforcement learning for spectrum sharing," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 4209–4219, June 2020.
- [158] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.
- [159] J. Ma, G. Zhao, and Y. Li, "Soft combination and detection for cooperative spectrum sensing in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4502–4507, Nov. 2008.
- [160] P. Lee, Z. A. Eu, M. Han, and H. Tan, "Empirical modeling of a solar-powered energy harvesting wireless sensor node for time-slotted operation," in *2011 IEEE Wireless Communications and Networking Conference*, 2011, pp. 179–184.
- [161] Y. Kawamoto, H. Takagi, H. Nishiyama, and N. Kato, "Efficient resource allocation utilizing q-learning in multiple ua communications," *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 3, pp. 293–302, July 2019.
- [162] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in hetnets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 680–692, Jan. 2018.
- [163] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1291–1307, Nov. 2012.
- [164] Y. Wei, F. R. Yu, M. Song, and Z. Han, "Joint optimization of caching, computing, and radio resources for fog-enabled iot using natural actor-critic deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2061–2073, April 2019.

- [165] 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.
- [166] 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.” in *International Conference on Learning Representations (Poster)*, 2016.
- [167] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *The 27th International Conference on Machine Learning*, 2010, pp. 807–814.
- [168] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.
- [169] C. Zhong, Z. Lu, M. C. Gursoy, and S. Velipasalar, “A deep actor-critic reinforcement learning framework for dynamic multichannel access,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 4, pp. 1125–1139, Dec. 2019.
- [170] K. Wang, L. Chen, and Q. Liu, “On optimality of myopic policy for opportunistic access with nonidentical channels and imperfect sensing,” *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2478–2483, June 2014.
- [171] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9, 2010, pp. 249–256.
- [172] C. I. C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, “Toward green and soft: a 5G perspective,” *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66–73, Feb. 2014.
- [173] J. van de Belt, H. Ahmadi, and L. E. Doyle, “Defining and surveying wireless link virtualization and wireless network virtualization,” *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1603–1627, Third Quarter 2017.
- [174] T. Theodorou and L. Mamatras, “CORAL-SDN: A software-defined networking solution for the internet of things,” in *2017 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2017, pp. 1–2.

- [175] L. Zhang, M. Xiao, G. Wu, M. Alam, Y. Liang, and S. Li, "A survey of advanced techniques for spectrum sharing in 5G networks," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 44–51, Oct. 2017.
- [176] A. Ahmad, S. Ahmad, M. H. Rehmani, and N. U. Hassan, "A survey on radio resource allocation in cognitive radio sensor networks," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 888–917, Second Quarter 2015.
- [177] X. Liu, H. Ding, and S. Hu, "Uplink resource allocation for NOMA-based hybrid spectrum access in 6G-enabled cognitive internet of things," *IEEE Internet of Things Journal*, pp. 1–10, 2020.
- [178] L. P. Qian, B. Shi, Y. Wu, B. Sun, and D. H. K. Tsang, "NOMA-enabled mobile edge computing for internet of things via joint communication and computation resource allocations," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 718–733, Jan. 2020.
- [179] Z. Li and J. Gui, "Energy-efficient resource allocation with hybrid TDMA–NOMA for cellular-enabled machine-to-machine communications," *IEEE Access*, vol. 7, pp. 105 800–105 815, 2019.
- [180] Y. Wu, N. Zhang, and K. Rong, "Non-orthogonal random access and data transmission scheme for machine-to-machine communications in cellular networks," *IEEE Access*, vol. 8, pp. 27 687–27 704, 2020.
- [181] G. Liu, Z. Wang, J. Hu, Z. Ding, and P. Fan, "Cooperative NOMA broadcasting/multicasting for low-latency and high-reliability 5G cellular V2X communications," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7828–7838, Oct. 2019.
- [182] Q. Y. Liao and C. Y. Leow, "Successive user relaying in cooperative NOMA system," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 921–924, June 2019.
- [183] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. 98, no. 3, pp. 403–414, March 2015.
- [184] A. Zafar, M. Shaqfeh, M. Alouini, and H. Alnuweiri, "On multiple users scheduling using superposition coding over rayleigh fading channels," *IEEE Communications Letters*, vol. 17, no. 4, pp. 733–736, April 2013.

- [185] Y. Yuan, Z. Yuan, G. Yu, C. Hwang, P. Liao, A. Li, and K. Takeda, "Non-orthogonal transmission technology in LTE evolution," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 68–74, July 2016.
- [186] A. S. Marcano and H. L. Christiansen, "A novel method for improving the capacity in 5G mobile networks combining NOMA and OMA," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, 2017, pp. 1–5.
- [187] F. Zhou, Y. Wu, Y. Liang, Z. Li, Y. Wang, and K. Wong, "State of the art, taxonomy, and open issues on cognitive radio networks with NOMA," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 100–108, April 2018.
- [188] W. Xu, X. Li, C. Lee, M. Pan, and Z. Feng, "Joint sensing duration adaptation, user matching, and power allocation for cognitive OFDM-NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1269–1282, Feb. 2018.
- [189] C. Wang, T. Chen, Y. Chen, and D. Wu, "Low-complexity resource allocation for downlink multicarrier NOMA systems," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2018, pp. 1–6.
- [190] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8440–8450, Sept. 2018.
- [191] M. Liu, T. Song, and G. Gui, "Deep cognitive perspective: Resource allocation for NOMA-based heterogeneous IoT with imperfect SIC," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2885–2894, April 2019.
- [192] K. N. Doan, M. Vaezi, W. Shin, H. V. Poor, H. Shin, and T. Q. S. Quek, "Power allocation in cache-aided NOMA systems: Optimization and deep reinforcement learning approaches," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 630–644, Jan. 2020.
- [193] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019.

- [194] Y. Liu, X. Wang, J. Mei, G. Boudreau, H. Abou-Zeid, and A. B. Sediq, "Situation-aware resource allocation for multi-dimensional intelligent multiple access: A proactive deep learning framework," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 116–130, Jan. 2021.
- [195] Z. Li, M. Xu, J. Nie, J. Kang, W. Chen, and S. Xie, "NOMA-enabled cooperative computation offloading for blockchain-empowered internet of things: A learning approach," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2364–2378, Feb. 2021.
- [196] Y. K. Tan and S. K. Panda, "Optimized wind energy harvesting system using resistance emulator and active rectifier for wireless sensor nodes," *IEEE Transactions on Power Electronics*, vol. 26, no. 1, pp. 38–50, Jan. 2011.
- [197] A. Cuadras, M. Gasulla, and V. Ferrari, "Thermal energy harvesting through pyroelectricity," *Sensors and Actuators A: Physical*, vol. 158, no. 1, pp. 132–139, March 2010.
- [198] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.
- [199] Z. Wei, D. W. K. Ng, and J. Yuan, "Power-efficient resource allocation for MC-NOMA with statistical channel state information," in *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–7.
- [200] Y. Li, W. Zhang, C. Wang, J. Sun, and Y. Liu, "Deep reinforcement learning for dynamic spectrum sensing and aggregation in multi-channel wireless networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 464–475, June 2020.
- [201] A. B. Rozario and M. F. Hossain, "Hybrid TDMA-NOMA based M2M communications over cellular networks with dynamic clustering and 3D channel models," in *2019 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, 2019, pp. 1–6.
- [202] H. Al-Obiedollah, K. Cumanan, A. G. Burr, J. Tang, Y. Rahulamathavan, Z. Ding,

- and O. A. Dobre, “On energy harvesting of hybrid TDMA-NOMA systems,” in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.
- [203] M. B. Shahab, M. Irfan, M. F. Kader, and S. Young Shin, “User pairing schemes for capacity maximization in non-orthogonal multiple access systems,” *Wireless Communications and Mobile Computing*, vol. 16, no. 17, pp. 2884–2894, Dec. 2016.
- [204] X. Wang, T. Jin, L. Hu, and Z. Qian, “Energy-efficient power allocation and Q-learning-based relay selection for relay-aided D2D communication,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6452–6462, June 2020.
- [205] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.