Performance optimization for cognitive radio networks with RF energy harvesting capability

Dinh, Thai Hoang

2016

Dinh, T. H. (2016). Performance optimization for cognitive radio networks with RF energy harvesting capability. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/67022>

<https://doi.org/10.32657/10356/67022>

NANYANG TECHNOLOGICAL UNIVERSITY

Performance Optimization for Cognitive Radio Networks with RF Energy Harvesting Capability

Ph.D Thesis

By

Dinh Thai Hoang

Supervisor: Prof. Dusit Niyato

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of Doctor of Philosophy

January, 2016

Acknowledgments

First and foremost, I would like to express my profound gratitude to my supervisor, Prof. Dusit Niyato, for giving me the opportunity to pursue research works at Nanyang Technological University. I am truly privileged to have learned from his deep knowledge and research enthusiasm. Since the beginning of my research, he has inspired me and given me a passion for research. He always shows me great research directions and unique ideas, along with encouragement and unwavering support regardless of his time. As well, I would greatly thank Prof. Dong In Kim, Prof. Ping Wang, Prof. Le Bao Long, Prof. Ekram Hossain, and Prof. Zhu Han, for their helpful comments and advice.

In addition, I also want to thank all my colleagues in Nanyang Technological University and my friends for their numerous help. I gratefully acknowledge the scholarship towards my PhD from Nanyang Technological University. Last but not least, my deepest love and gratitude are devoted to all of my family members: my parents, my little sister, my brothers, and especially my wife Picky. Their loves will be endless motivation for me to strive and help me overcome difficult times in my life.

Contents

Acknowledgments	i
List of Figures	v
List of Tables	vii
Abstract	
1 Introduction	1
1.1 Cognitive Radio Networks and Wireless Energy Harvesting Techniques	1
1.1.1 Cognitive Radio Networks	1
1.1.2 Wireless Energy Harvesting Techniques	5
1.2 An Overview of Markov Decision Processes	10
1.2.1 The Markov Decision Process Framework	10
1.2.2 Solutions of MDPs	11
1.2.3 Extensions of MDPs and Complexity	14
1.3 Scope, Contributions and Organization	17
1.3.1 Research Motivations	17
1.3.2 Organization, Contributions and the Connection among Research Issues	19
2 Literature Review	22
2.1 Cognitive Radio Networks with RF Energy Harvesting	22
2.1.1 Channel Access Strategy for Secondary Users	22
2.1.2 Relaying	25
2.1.3 Time Scheduling for Secondary Users	31
2.1.4 Other Related Issues	33
2.2 Cognitive Radio Networks with Other Wireless Energy Harvesting Techniques	36
2.3 Research Trends, Scope and Novelty of the Thesis	37

3	Optimal Channel Access for Cognitive Users with Energy Harvesting Capability	42
3.1	System Model	43
3.2	Markov Decision Process Formulation	46
3.2.1	State Space and Action Space	46
3.2.2	Transition Probability Matrix	46
3.2.3	Optimization Formulation	50
3.2.4	Performance Measures	51
3.3	Learning Algorithm	52
3.3.1	Problem Formulation	52
3.3.2	Policy Gradient Method	56
3.3.3	An Idealized Gradient Algorithm	57
3.3.4	Learning Algorithm	57
3.4	Performance Evaluation	60
3.4.1	Parameter Setting	60
3.4.2	Numerical Results	61
3.5	Summary	71
4	Performance Optimization for Cooperative Multiuser Cognitive Radio Networks with RF Energy Harvesting Capability	74
4.1	System Model and Assumptions	76
4.2	A TDMA Learning Algorithm for RF Energy Harvesting Cognitive Radio Networks	78
4.2.1	Problem Formulation	79
4.2.2	Learning Algorithm Based on Policy Gradient Method	81
4.3	A Decentralized Solution for RF Energy Harvesting Cognitive Radio Networks	82
4.3.1	Optimization Formulation	82
4.3.2	Parameterization for DEC-POMDP	86
4.3.3	Lagrange Multiplier and Policy Gradient Method	87
4.3.4	Decentralized Online Learning Algorithm with Communications	89
4.4	Performance Evaluation	91
4.4.1	Simulation Setup	91
4.4.2	Simulation Results	92
4.5	Summary	101

5	Conclusions and Future Works	104
5.1	Conclusions	104
5.2	Future Works	105
5.2.1	Channel Feedback Information for Secondary Systems	105
5.2.2	Performance Optimization for Underlay RF-Powered CRNs . . .	106
5.2.3	Game Models in RF Powered CRNs	106
5.2.4	Integrating RF Powered CRNs with other Networks	107
5.2.5	Design and Implement on Hardware Devices	107
5.2.6	Economic Models in RF Powered CRNs	107
	Appendices	109
	References	126
	Author's Publications	140

List of Figures

1.1	A general cognitive radio network architecture.	4
1.2	A taxonomy of challenges in cognitive radio networks.	5
1.3	Inductive coupling.	7
1.4	Magnetic resonance coupling.	9
1.5	RF energy harvesting.	10
1.6	Extensions of Markov decision models [1].	15
1.7	The relation between Chapter 3 and Chapter 4.	21
2.1	Relay-assisted cognitive radio networks.	26
2.2	The comparison of the data frame structure between TDBC and MABC protocol in two-way relay cognitive radio network (TWRCN).	28
2.3	Two-way cognitive relay network.	29
2.4	Research trends of CRNs with energy harvesting capability.	38
2.5	Summary information of the percentage of problems in (a) general wire- less powered CRNs and (b) RF powered CRNs.	38
3.1	System model.	44
3.2	The learning model for the secondary user in a cognitiver radio network.	52
3.3	Policy of the secondary user.	62
3.4	The convergence of the learning algorithm when the number of channels is varied.	64
3.5	(a) Policy at 10^3 and (b) 10^7 iterations of the learning algorithm. . . .	65
3.6	The average throughput of the system when (a) the packet arrival prob- ability is varied and (b) the idle channel probability for channel 1 is varied.	66
3.7	Throughput under different successful harvesting probabilities for chan- nel 1 (a) and successful data transmission probabilities (b).	68
3.8	The average throughput of the system when (a) the false alarm prob- ability and (b) the miss detection probability is varied.	69

3.9	(a) The average number of packets in the data queue and (b) the packet blocking probability.	71
4.1	System model.	77
4.2	The convergence of (a) the TDMA learning algorithm and (b) the decentralized learning algorithm.	94
4.3	The policy of secondary user 1 obtained by the TDMA learning algorithm for the cases that (a) it is not scheduled and (b) it is scheduled to access a time slot.	96
4.4	The policy of secondary user 2 obtained by the TDMA learning algorithm for the cases that (a) it is not scheduled and (b) it is scheduled to access a time slot.	97
4.5	The policy of secondary users (a) 1 and (b) 2 obtained by the decentralized learning algorithm.	98
4.6	The average throughput of the system under different algorithms with 2 channels.	99
4.7	The average throughput with (a) two channels with low idle probability and one channel with high idle probability and (b) two channels with high idle probability and one channel with low idle probability.	100
4.8	Collision probability with two channels.	101

List of Tables

1.1	The comparison among wireless energy harvesting techniques.	10
1.2	The Worst Case Complexity of Markov Models.	17
2.1	Problems exploiting ambient energy sources in cognitive radio networks	40
2.2	Problems exploiting ambient energy sources in cognitive radio networks	41
3.1	Summary of key notations of Chapter 3	73
4.1	Summary of key notations of Chapter 4	103

List of Abbreviations

AES	Ambient Energy Source
AP	Access Point
BS	Base Station
CRN	Cognitive Radio Network
D2D	Device-to-Device
DEC-POMDP	Decentralized Partially Observable Markov Decision Process
EHT	Energy Harvesting Technique
FCC	Federal Communication Commission
HPPPs	Homogeneous Poisson Point Processes
MDP	Markov Decision Process
MICT	Magnetic Inductive Coupling Technique
MIMO	Multiple Input Multiple Output
MRICT	Magnetic Resonance Inductive Coupling Technique
PR	Primary Receiver
PT	Primary Transmitter
PU	Primary User
QoS	Quality-of-Service
RF	Radio Frequency
SR	Secondary Receiver
ST	Secondary Transmitter
SU	Secondary User
SWIPT	Simultaneous Wireless Information and Power Transfer
TDMA	Time Division Multiple Access
WEHT	Wireless Energy Harvesting Technique

Summary

A cognitive radio is an intelligent radio network in which unlicensed users (i.e., secondary users) can opportunistically access idle channels when such channels are not occupied by licensed users (i.e., primary users). The main purpose of cognitive radio networks (CRNs) is to utilize the available spectra which become more and more scarce due to the boom of wireless communication systems. In CRNs, wireless nodes can operate based on an AC power supply. However, this is feasible only for fixed wireless nodes. A mobile node instead has to operate on the energy from a battery. Accordingly, the battery has to be charged or replaced regularly to ensure sufficient energy supply for the mobile nodes. Alternatively, energy harvesting techniques have been introduced as a promising solution to perpetuate operations of the mobile nodes. With wireless energy transfer, batteries can be replenished without using any physical connection for charging or battery replacement.

Recently, radio frequency (RF) energy harvesting techniques with high efficiency have been introduced. Such techniques allow a wireless node to harvest and convert electromagnetic waves from ambient RF sources (e.g., TV, radio towers and cellular base stations) into energy which can be used for data transmissions. With the RF energy harvesting capability, the wireless node can continue its operation without physically changing or recharging its battery. In RF powered CRNs, secondary users can harvest energy from radio signals by using RF energy harvesting devices and then use such energy to transmit data to the primary channels. To obtain enough energy and spectrum access opportunities for data transmissions, the secondary users must search for not only idle channels to transmit their packets, but also busy channels to harvest RF energy. The channel access, which determines the channel to transmit packets or to harvest RF energy, is a crucial step to achieve optimal performance for

the secondary system. Therefore, in this thesis, we mainly focus on channel access strategies in RF powered CRNs with the goal of optimizing for the secondary network performance while still guaranteeing the quality-of-service for primary systems.

The thesis presents two major contributions. Firstly, we study the channel access problem in which a secondary user selects a channel to access for packet transmissions or to harvest RF energy. We formulate the channel access problem as a Markov decision process (MDP) and adopt the linear programming technique to maximize the throughput for the secondary user. We then propose an online learning algorithm to obtain the optimal channel access policy for the secondary user. With the proposed learning algorithm, the secondary user can observe and adapt the channel access decision to achieve its objective. Secondly, we study the scenario with multiple secondary users coexisting in the same network and they want to cooperate to maximize the joint objective. In this case, we introduce an approach using a round robin scheduling and each secondary user is equipped with an online learning algorithm in order to explore the surrounding environment and make optimal decisions. After that, we introduce a decentralized partially observable MDP framework to model the cooperation among secondary users without coordination and propose a decentralized learning algorithm through using the policy gradient and the Lagrange multiplier method.

In summary, this thesis mainly focuses on the channel access problem for RF powered CRNs which is one of the most important issues of CRNs as well as wireless powered CRNs. In all research works studying channel access strategies in wireless powered CRNs in the literature, they considered only one licensed channel. To the best of our knowledge, this is the first work that studies the performance optimization problem for RF powered CRNs with multiple licensed channels taken into considerations. In addition, in this thesis, we also present novel models as well as solutions to address the channel access problem.

Chapter 1

Introduction

1.1 Cognitive Radio Networks and Wireless Energy Harvesting Techniques

1.1.1 Cognitive Radio Networks

1.1.1.1 Definition and characteristics

The rapid development of wireless technologies led to the introduction of many new wireless communication networks such as machine-to-machine communications [2], body sensor networks [3], and mobile cloud computing [4]. Consequently, there is a boom in using wireless devices to meet human demands. According to the report of ABI research center [5], the number of wireless devices in 2014 was around 16 billion, and it is expected to exceed 40 billion by 2020. However, in contrast with the development of wireless devices, the spectrum bandwidth capacity is limited and non-expandable which is the main reason hindering the development of wireless networks. Therefore, we need to find appropriate solutions to overcome this issue.

A cognitive radio is an intelligent radio network in which unlicensed users (i.e., secondary users) can opportunistically access idle channels when such channels are not occupied by licensed users (i.e., primary users)¹. The main purpose of cognitive radio networks (CRNs) is to utilize effectively the available spectra which are more and more scarce due to the boom of wireless communication systems. Basically, we can define a cognitive radio as *a radio that can change its transmitter parameters based*

¹In the thesis, terms, i.e., unlicensed users, secondary users, and cognitive users, are used interchangeably. Similarly, terms, i.e., licensed users, and primary users, are also used interchangeably.

on interaction with the environment it operates in [6]. From this definition, there are two major characteristics of cognitive radios that are different from traditional communication systems, namely, cognitive capability and reconfigurability [7].

- *Cognitive capability:* This characteristic enables cognitive users to obtain necessary information from their environment.
- *Reconfigurability:* After gathering the environment information, with the reconfigurability characteristic, cognitive users can dynamically adjust their operating modes to environment variations in order to achieve optimal performance.

With these characteristics, there are four main functions which need to be implemented on cognitive users.

- *Spectrum sensing:* The goal of the spectrum sensing function is to determine the status of the spectrum and the activity of the licensed users by periodically sensing the target frequency band.
- *Spectrum analysis:* The information obtained from sensing the spectrum is used to schedule and plan for the spectrum access process of the cognitive users.
- *Spectrum access:* After a spectrum access decision is made based on the spectrum analysis, the spectrum holes are accessed by the cognitive users.
- *Spectrum mobility:* Spectrum mobility is a function related to the change of operating frequency bands of the cognitive users. The spectrum change has to ensure that the data transmission process by cognitive users can continue in a new spectrum band.

Under these functions, cognitive users can utilize the limited spectrum resource in a more efficient and flexible way.

1.1.1.2 Architecture, applications and challenges

Architecture: A general cognitive radio network architecture can be illustrated as in Figure 1.1 with two main networks, called primary networks and cognitive networks, which coexist in the same environment to utilize the joint spectra. The joint spectra are divided into sub-bands to provide for different kinds of users. In general, the spectrum bands can be divided into two catalogs, i.e., licensed and unlicensed bands. Licensed bands are reserved for particular objects with specific applications, e.g., TV channels, radio, or mobile communications, while unlicensed bands can be used by anyone without requiring any permission. In cognitive radio networks, licensed users are collected to the primary base stations (BSs) through indicated licensed channels, while secondary users are collected to secondary BSs or other cognitive devices (e.g., ad-hoc cognitive networks) through unlicensed channels or licensed channels when such channels are not occupied by primary users. The crucial objective of this architecture is to maximize the efficiency of using spectra for wireless users.

Applications, implementations and standardizations: Cognitive radios can be used in many wireless systems in practice to improve the efficiency of using bandwidth for such systems. For example, cognitive radio networks can be adopted in wireless body area networks and machine-to-machine communications to help a huge number of wireless devices in these networks to communicate with each other or with other devices through using licensed/unlicensed channels. Also, there are many applications of cognitive radios in other conventional wireless systems, e.g., cellular networks, mesh networks, and vehicular networks, as showed in [8].

There are some applications of cognitive radios which have been already implemented in practice. In particular, in 2011, the Federal Communication Commission (FCC) in the United States authorized the use of cognitive radios in TV spectrum [9]. This policy will allow unlicensed users to transmit signals to the broadcast television spectrum (also known as white space spectrum) at locations where the spectrum is not being used by licensed users. At the same time, the first well known global wireless

Figure 1.1: A general cognitive radio network architecture.

standard based on cognitive radios, i.e., IEEE 802.22, was released [10]. The IEEE 802.22 standard enables broadband wireless access using cognitive radio techniques in white spaces and this makes a significant amount of spectrum available for business and other services. More applications and standardizations of cognitive radios in wireless systems can be found in [11, 12].

Challenges and issues: Although cognitive radios are promising solutions to address effectively the bandwidth scarcity problem for wireless networks, there have been still relatively many challenges which hinder the development of cognitive radios in practice. Such challenges can stem from the implementation, the technology, or regulatory issues. In Fig. 1.2, we provide a taxonomy of issues along with challenges for the development of cognitive radio technology.

Figure 1.2: A taxonomy of challenges in cognitive radio networks.

1.1.2 Wireless Energy Harvesting Techniques

1.1.2.1 Motivations and the development of energy harvesting techniques

In wireless communication networks, wireless nodes can operate based on an AC power supply. However, this is feasible only for fixed wireless nodes. A mobile node instead has to operate on a battery. Accordingly, the battery has to be charged or replaced regularly to ensure the sufficient energy supply for the mobile nodes. Alternatively, energy harvesting techniques have been introduced as a promising solution to perpetuate operations for mobile nodes. With the wireless energy transfer, batteries can be replenished without using any physical connection for charging or the battery replacement.

Energy harvesting is a process which enables external energy sources (e.g., solar power, wind power, and thermal power) converted into electrical energy and supply for electric devices. Converting ambient energy into electrical energy has received a lot of attentions and there are a lot of practical applications for human beings. For example, dams have been implemented in many countries in the world to provide an essential energy source for human daily life. Alternatively, solar power systems also

have been implemented widely as green energy sources to serve human demands.

Although conventional energy harvesting techniques have been ubiquitously applied in practice, they are not really attractive to wireless communication systems for various reasons. First, energy sources used in conventional energy harvesting techniques are often dynamic and limited by locations, and thus they cannot be applied for general cases in wireless systems. In addition, to harvest energy from the environment, wireless nodes need to be equipped with special conversion devices (e.g., solar cells or dams) which are expensive and relatively big in size, making them inconvenient for deploying along with wireless nodes. Therefore, some wireless energy harvesting techniques which can be used more efficiently in wireless communication systems have been introduced.

Wireless energy harvesting, also known as wireless power transfer [13, 14], is a technique which enables power transferred wirelessly from a power source to electrical devices with the aim to help such devices operate without depending on traditional energy sources (e.g., fuel, solar, and battery). Different from conventional energy harvesting techniques, e.g., thermoelectrics, piezoelectric, and photovoltaic, the power supply sources used in wireless energy harvesting techniques are often stable, controllable, and very efficient. For example, recently Samsung has released a new product, called fast wireless charging pad (EP-PN920TBEGUS) [15], which can support wireless charging for multiple Samsung wireless devices, e.g., Galaxy Note 5 and Galaxy S6 Edge+, very quickly at multiple locations simultaneously. Therefore, wireless energy harvesting techniques have been receiving a lot of attentions recently.

1.1.2.2 Classification of wireless energy harvesting techniques

The development of wireless energy harvesting techniques leads to two major research directions, namely, radiative wireless harvesting and magnetic wireless harvesting. While the radiative wireless harvesting techniques use electromagnetic waves (typically radio frequency (RF) waves and microwaves) as a means to convey energy to the wireless nodes, the magnetic wireless harvesting techniques adopt magnetic fields

as a medium environment to transfer energy to the destination. Based on the transmission medium, i.e., radiative and magnetic medium, we also can classify wireless energy harvesting techniques into two classes, namely, far-field wireless energy harvesting and near-field wireless energy harvesting techniques, respectively. The reason is that while electromagnetic waves can be transmitted to far distances, the magnetic field between the source node and the destination node is limited to a short distance. In the following, we will present more details about these techniques.

Near-field wireless energy harvesting techniques: These techniques are developed based on the principle of magnetic field coupling (also called inductive coupling) between coupled coils and there are two potential techniques, called *magnetic inductive coupling* and *magnetic resonance inductive coupling*.

Magnetic inductive coupling: Magnetic inductive coupling [16] is a technique using magnetic fields to transfer energy between coupled wire coils based on Faraday's law of induction. In particular, when there is an electrical current moving through a wire coil (L1), it will generate a magnetic field and if we place a second wire coil (L2) in this magnetic field, the magnetic field can induce a current in the second wire coil as illustrated in Fig. 1.3. Such current then can be used to maintain the operation of the wireless node or be stored in the battery. Typically, the operating

Figure 1.3: Inductive coupling.

frequency of the magnetic inductive coupling technique (MICT) is in the KHz range and thus the transfer efficiency of this technique reduces significantly as the distance

between two coupled coils increases. It was shown in [17] that the effective energy transfer distance of MICT is few centimeters, but the transfer power can reach a few kilowatts. The advantages of MICT are safety to human health, high efficiency over short distances, convenience in implementing and operating. Thus, applications of MICT could be electric toothbrushes, mobile wireless chargers, and wireless powered medical implants.

Magnetic resonance inductive coupling: Basically, the operation principle of magnetic resonance inductive coupling technique (MRICT) is similar to that of MICT, i.e., based on the magnetic field between two coupled wire-coils. However, there is a difference in designing resonance circuits of the wire coils which makes differences for such techniques. As shown in Fig. 1.4, the coupled wire coils used in MRICT are equipped with resonance circuits (R1 and R2) which are able to resonate with their internal capacitance, i.e., C1 and C2, respectively. Then, by using resonance circuits, the coupled wire coils are able in turn to resonate at the same resonant frequency and thereby dramatically increase the coupling capability as well as the transferred power amount. The key idea behind this technique is that resonance circuits can exchange energy at a much higher rate and thus the same amount of energy can be conveyed to a greater distance [18]. In 2007, a group of researchers from MIT demonstrated that by using the MRICT we can transfer 60 Watts of energy over a distance of 2 meters with approximately 40% efficiency [19]. With the ability of transferring power to a few meters, MRICT can be applied to many areas such as portable charging devices, electric vehicles, and so on.

Far-field wireless energy harvesting techniques: Far-field wireless energy harvesting techniques are methods that enable energy transfer to a distant destination (up to a few kilometers) based on the electromagnetic radiation. Basically, the electromagnetic radiation is the emission of energy of electromagnetic waves when such waves are propagated to the space. According to Maxwell's theory [20], in electromagnetic waves, changes in the electrical field are always associated with a wave in

Figure 1.4: Magnetic resonance coupling.

the magnetic field in one direction, and vice versa. This process is taking place continuously in the direction of the wave propagation and enables electromagnetic waves to carry information as well as energy to a far distance. Radio frequency (RF) is a typical example of the electromagnetic waves, which is used widely in practice. In order to harvest energy from RF signals, a wireless node needs to be equipped with a receive antenna and a set of transforming devices, e.g., impedance matching, voltage multiplier, and capacitor, as illustrated in Fig. 1.5. It was shown in [21] that the efficiency of the RF energy harvesting techniques strongly depends on the ability of the receive antenna, the accuracy of the impedance matching circuit, and the proficiency between the voltage multiplier and the capacitor. The recent real implementation [22] has demonstrated that RF energy harvesting techniques can achieve the efficiency up to 84% with a 5.8 dBm input power. Although RF energy harvesting techniques can transmit energy to a long distance, the amount of transferred power is relatively small compared with MICT and MRICT techniques. Thus, these techniques are applicable to wireless sensor nodes and some mobile devices which require a moderate amount of energy.

1.1.2.3 Comparison

In Table 1.1, we provide a summary and comparisons among three different techniques, i.e., MICT, MRICT, and RF energy harvesting technique (EHT). From Table 1.1, we can see that different EHTs have different characteristics, and thus they will have different applications in practice. Among the three EHTs, RF EHT is an outstanding

Figure 1.5: RF energy harvesting.

candidate which is the most suitable for CRNs. Thus, in this thesis, we mainly focus on studying applications of RF EHT in CRNs.

Table 1.1: The comparison among wireless energy harvesting techniques.

	MICT	MRICT	RF
Strength	Very high	High	Low
Efficiency	Very high	High	Low
Multicast	Difficult	Yes	Yes
Mobility	No	No	Yes
Safety	Magnetic	Safe	Safe
Distance	Very short (few centimes)	Short (few centimes to few meters)	Long (up to few kilometers)
Applications	Mobile charging, electric toothbrush, and medical implant	Mobile charging, vehicle charging, and household electrical appliances	RIFD tags, wireless sensors, and unmanned planes

1.2 An Overview of Markov Decision Processes

1.2.1 The Markov Decision Process Framework

The MDP is defined by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{T} \rangle$ where,

- $\mathcal{S}$ is a finite set of states,
- $\mathcal{A}$ is a finite set of actions,
- $\mathcal{P}$ is a transition probability function from state s to state s' after action a is taken,

- $\mathcal{R}$ is the immediate reward obtained after action a is made, and
- $\mathcal{T}$ is the set of decision epoch which can be finite or infinite.

π denotes a “policy” which is a mapping from a state to an action. The goal of an MDP is to find an optimal policy to maximize or minimize a certain objective function. An MDP can be finite or infinite time horizon. For the finite time horizon MDP, we need to find an optimal policy π^* to maximize the expected total reward that is defined as follows:

$$\max \mathcal{V}_\pi(s) = \mathbb{E}_{\pi,s} \left[\sum_{t=1}^T \mathcal{R}(s'_t | s_t, \pi(a_t)) \right] \quad (1.1)$$

where s_t and a_t are the state and action at time t , respectively.

For the infinite time horizon MDP, the objective can be to maximize the expected discounted total reward or to maximize the average reward. The former is defined as follows:

$$\max \mathcal{V}_\pi(s) = \mathbb{E}_{\pi,s} \left[\sum_{t=1}^T \gamma^t \mathcal{R}(s'_t | s_t, \pi(a_t)) \right], \quad (1.2)$$

while the latter is expressed as follows:

$$\max \mathcal{V}_\pi(s) = \liminf_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_{\pi,s} \left[\sum_{t=1}^T \mathcal{R}(s'_t | s_t, \pi(a_t)) \right]. \quad (1.3)$$

Here, γ is the discounting factor and $\mathbb{E}[\cdot]$ is the expectation function.

1.2.2 Solutions of MDPs

Here we introduce solution methods for MDPs with the discounted total reward. The solutions for MDPs with the average reward are very similar to those of MDPs with the discounted total reward and they can be found in [23].

1.2.2.1 Solutions for finite time horizon Markov decision processes

In a finite time horizon MDP, the system operation takes place in a known period of time. In particular, the system starts at state s_0 and continues to operate in the next T periods. The optimal policy π^* is to maximize $\mathcal{V}_\pi(s)$ in (1.1). If we denote $v^*(s)$

as the maximum achievable reward at state s , then we can find $v^*(s)$ at every state recursively by solving the following *Bellman's optimal equations* [24]:

$$v_t^*(s) = \max_{a \in \mathcal{A}} \left[\mathcal{R}_t(s, a) + \sum_{s' \in \mathcal{S}} \mathcal{P}_t(s'|s, a) v_{t+1}^*(s') \right]. \quad (1.4)$$

Based on the optimal Bellman equations, two typical approaches for finite time horizon MDPs exist.

- *Backwards induction*: Also known as a dynamic programming approach, it is the most popular and efficient method for solving the Bellman's equations. Since the process will be stopped at a known period, we can first determine the optimal action and the optimal value function at the last time period. We then recursively obtain the optimal actions for earlier periods back to the first period based on the Bellman optimal equations.
- *Forward induction*: This forward induction method is also known as a value iteration approach. The idea is to divide the optimization problem based on the number of steps to go. In particular, given an optimal policy for $(t - 1)$ time steps to go, we calculate the Q-values for k steps to go. After that, we can obtain the optimal policy based on the following equations:

$$Q_t(s, a) = \mathcal{R}(s, a, s') + \sum_{s'} \mathcal{P}(s, a, s') v_{t-1}^*(s'),$$

$$v_t^*(s) = \max_{a \in \mathcal{A}} Q_t^*(s, a) \text{ and } \pi_t^*(s) = \arg \max_{a \in \mathcal{A}} Q_t^*(s, a),$$

where $v_t(s)$ is the value of state s and $Q_t(s, a)$ is the value of taking action a at state s . This process will be performed until the last period is reached.

Both approaches have the same complexity which depends on the time horizon of an MDP. However, they are used differently. Backward induction is especially useful when we know the state of MDPs in the last period. By contrast, forward induction is applied when we only know the initial state.

1.2.2.2 Solutions for infinite time horizon Markov decision processes

Solving an infinite time horizon MDP is more complex than that of a finite time horizon MDP. However, the infinite time horizon MDP is more widely used because in practice the operation time of systems is often unknown and assumed to be infinite. Many solution methods were proposed.

- *Value iteration (VI)*: This is the most efficiently and widely used method to solve an infinite time horizon discounted MDP. This method has many advantages, e.g., quick convergence, ease of implementation, and is especially a very useful tool when the state space of MDPs is very large. Similar to the forward induction method of a finite time horizon MDP, this approach was also developed based on dynamic programming. However, for infinite time horizon MDP, since the time horizon is infinite, instead of running the algorithm for the whole time horizon, we have to use a stopping criterion (e.g., $\|v_t^*(s) - v_{t-1}^*(s)\| < \epsilon$) to guarantee the convergence [24].
- *Policy iteration (PI)*: The main idea of this method is to generate an improving sequence of policies. It starts with an arbitrary policy and updates the policy until it converges. This approach consists of two main steps, namely policy evaluation and policy improvement. We first solve the linear equations to find the expected discounted reward under the policy π and then choose the improving decision policy for each state. Compared with the value iteration method, this method may take fewer iterations to converge. However, each iteration takes more time than that of the value iteration method because the policy iteration method requires solving a set of linear equations.
- *Linear programming (LP)*: Unlike the previous methods, the linear programming method aims to find a static policy through solving a linear program [25]. After the linear program is solved, we can obtain the optimal value $v^*(s)$, based on which we can determine the optimal policy $\pi^*(s)$ at each state. The linear programming method is relatively inefficient compared with the value and policy

iteration methods when the state space is large. However, the linear programming method is very useful for MDPs with constraints since the constraints can be included as linear equations in the linear program [26]. Also, there are many software tools developed to solve the linear program efficiently, e.g., Matlab, making LP more convenient to use.

- *Approximation method*: Approximate dynamic programming [27] was specifically developed for large MDPs. This method approximates the value functions (whether policy functions or value functions) by assuming that these functions can be characterized by a reasonable number of parameters. Thus, we can seek the optimal parameter values to obtain the best approximation, e.g., as given in [27, 28] and [29].
- *Online learning*: The aforementioned methods are performed in an offline fashion (i.e., when the transition probability function is provided). However, they cannot be used if the information of such function is unknown. Learning algorithms were proposed to address this problem [28, 30]. The idea of this solution is based on the simulation-based method that evaluates the interaction between an agent and its interaction system. Then, the agent can adjust its behavior to achieve its goal (e.g., trial and error).

Note that the solution methods for discrete time MDPs can be applied for continuous time MDPs through using uniformization techniques [31, 32]. The solutions of discrete time MDPs that solve the continuous time MDPs are also known as *semi-MDPs* (SMDPs).

1.2.3 Extensions of MDPs and Complexity

Next we present some extensions of an MDP, the relation of which is shown in Figure 1.6.

- MDP : Markov decision process
- POMDP : Partially observable Markov decision process
- MMDP : Multi-agent Markov decision process
- DEC-POMDP : Decentralized partially observable Markov decision process
- SG : Stochastic game

Figure 1.6: Extensions of Markov decision models [1].

1.2.3.1 Partially observable Markov decision processes (POMDPs)

In classical MDPs, we assume that the system state is fully observable by an agent. However, in many wireless systems, due to hardware limitations, environment dynamics, or external noise, the wireless nodes may not have full observability. Therefore, a POMDP [33] becomes an appropriate tool for such an incomplete information case. In POMDPs, the agent has to maintain the complete history of actions and observations in order to find an optimal policy, i.e., a mapping from histories to actions. However, instead of storing the entire history, the agent maintains a belief state that is the probability distribution over the states. The agent starts with an initial belief state b_0 , based on which it takes an action and receives an observation. Based on the action and the received observation, the agent then updates a new belief state. Therefore, a POMDP can be transformed to an MDP with belief state [34, 35]. Additionally, for a special case when the state space is continuous, parametric POMDPs [36] can be used.

1.2.3.2 Multi-agent Markov decision processes (MMDPs)

Unlike an MDP which is for a single agent, an MMDP allows multiple agents to cooperate to optimize a common objective [37]. In MMDPs, at each decision time, the agents stay at certain states and they choose individual actions simultaneously. Each agent is assumed to have a full observation of the system state through some information exchange mechanisms. Thus, if the joint action and state space of the

agents can be seen as a set of basic actions and states, an MMDP can be formulated as a classical MDP. Thus, the solution methods for MDPs can be applied to solve MMDP. However, the state space and action space will drastically grow as the number of agents increases. Therefore, approximate solution methods are often adopted.

1.2.3.3 Decentralized partially observable Markov decision processes (DEC-POMDPs)

Similar to MMDPs, DEC-POMDPs [38] are for multiple cooperative agents. However, in MMDPs, each agent has a full observation to the system. By contrast, in DEC-POMDPs, each agent observes only a part of the system state. In particular, the information that each agent obtains is local, making it difficult to solve DEC-POMDPs. Furthermore, in DEC-POMDPs, because each agent makes a decision without any information about the action and state of other agents, finding the joint optimal policy becomes intractable. Therefore, the solution methods for a DEC-POMDP often utilize special features of the models or are based on approximation functions to circumvent the complexity issue [39, 40]. Note that a decentralized Markov decision process (*DEC-MDP*) is a special case of a DEC-POMDP in which if all agents share their observations, they will have a global system state. In wireless systems, when the communication among cooperative wireless nodes is costly or impossible, the DEC-POMDP could be the best framework.

1.2.3.4 Stochastic games (SGs)

While MMDPs and DEC-POMDPs consider the cooperative interaction among agents, stochastic games (or Markov games) model the case where agents are non-cooperative and they aim to maximize their own payoffs rationally [41]. In particular, agents know states of all others in the system. However, due to the different objective functions that lead to the conflict among agents, finding an optimal strategy given the strategies of other agents is complex [42]. Note that the extension of stochastic games is known as a partial observable stochastic game [43] (*POSG*) which has a fundamental difference in observation. Specifically, in POSGs, the agents know only local

Table 1.2: The Worst Case Complexity of Markov Models.

Model	Complexity
MDP	P-complete
MMDP	P-complete
POMDP (finite time horizon)	PSPACE-complete
DEC-MDP (finite time horizon)	NEXP-complete
DEC-POMDP (finite time horizon)	NEXP-complete
POSG (finite time horizon)	NEXP-complete

states. Therefore, similar to DEC-POMDPs, POSGs are difficult to solve due to the incomplete information and decentralized decisions.

It was proved that both finite time and infinite time horizon MDPs can be solved in a complete polynomial time by dynamic programming [44, 45]. However, extensions of MDPs may have different computation complexity. For example, for POMDPs, the agents have incomplete information and thus they need to monitor and maintain a history of observations to infer the belief states. It was shown in [46] that the complexity of POMDPs can vary in different circumstances and the worst case complexity is PSPACE-complete [44, 46]. Since MMDPs can be converted to MDPs, its complexity in the worst case is P-complete. However, with multiple agents and partial observation (i.e., DEC-POMDP, DEC-MDP, and POSG), the complexity is dramatically increased. It was pointed out in [38] that even with just two independent agents, the complexity for both finite time horizon DEC-MDPs and DEC-POMDPs is NEXP-complete. Table 1.2 summarizes the worst case complexity. Note that partially observable problems are undecidable because infinite time horizon POMDPs are undecidable as shown in [47].

1.3 Scope, Contributions and Organization

1.3.1 Research Motivations

Recently, the development of wireless power transfer technologies has brought a new research direction for cognitive radio networks (CRNs), called CRNs with wireless energy harvesting, which has been receiving significant growing attention. By integrating wireless power transfer technologies into CRNs, secondary users are able to

not only access licensed channels opportunistically, but also harvest energy wirelessly to support their data transmissions. As a result, CRNs with wireless energy harvesting techniques can improve the spectrum utilization, the network performance, and also enhance the efficiency for both primary and secondary systems.

Among available wireless energy harvesting techniques, radio frequency (RF) energy harvesting is considered as a particularly appropriate solution for CRNs because of its outstanding advantages.

- *Pervasive environment:* RF sources are available almost everywhere, which provides an abundant energy resource for wireless nodes.
- *Long distance:* RF energy can propagate and transfer over long distances, which creates favorable conditions for mobile devices.
- *Multiple directions:* RF energy is broadcast in all directions allowing multiple users harvesting energy from the same source simultaneously.
- *Controllable power:* RF energy is controllable by adjusting transmit power at the RF sources.
- *Stability:* Compared with ambient energy sources, e.g., solar, wind, and heat, RF energy is much more stable, and it does not depend on environment conditions, e.g., day/night and rain/shine, as well as the wireless nodes' locations, e.g., indoor or outdoor.

Because of the advantages of using RF energy harvesting techniques in CRNs, there are many research exploring this topic. In this thesis, we focus on the performance optimization problems for CRNs with RF energy harvesting capability.

In this thesis, we adopt a Markov decision process (MDP) framework as a key tool to formulate and solve optimization problems. First, MDPs have a long history of successful applications in many areas such as population, agriculture, and manufacturing [48]. Many applications of MDPs have been implemented in wireless communication systems [49] and in the real world [50]. However, applications of MDPs

in CRNs with RF energy harvesting capability have still not yet been well investigated and this makes room for us to study on this topic. Furthermore, models and solutions of MDPs are very diverse and flexible which bring great opportunities for applying MDPs to solve stochastic optimization problems, especially in the dynamic environments like CRNs and wireless energy transfer.

1.3.2 Organization, Contributions and the Connection among Research Issues

The organization and the main contributions of the thesis are summarized as follows.

- **Chapter 1:**

- We introduce some fundamental backgrounds about cognitive radio networks, wireless energy harvesting techniques (WEHTs). Furthermore, applications and comparisons of WEHTs are also presented.
- We provide an overview of Markov decision processes including the formulation, models, and solutions. In addition, extensions of MDP models along with their solutions and complexities are given.

- **Chapter 2:**

- We discuss research works related to problems of harvesting wireless energy sources in CRNs together with analysis and comparisons among them.
- We highlight the significance and novelty of our works compared with other related works in the literature.

- **Chapter 3:**

- We design a channel access model for an SU in a CRN with RF energy harvesting capability in which the SU is able to harvest energy from PUs' signals and then use such energy to transmit packets to an idle channel.
- We introduce an optimization formulation based on an MDP and adopt the linear programming method to obtain the optimal channel access policy for the SU given the probability distribution function of the channel state.

- We propose an online learning algorithm to help the SU adapt with the change of the environment and to obtain the optimal policy without the knowledge of the channel state.
- We also provide the case with complete information about the channel state and use this case as a benchmark to evaluate our proposed solutions.
- We present extensive performance evaluation to show the convergence as well as the efficiency of the proposed learning algorithm. In addition, comparisons and analysis are provided to demonstrate the efficiency of the proposed learning algorithm in a dynamic environment.
- The proof of the convergence of the learning algorithm is also provided.

- **Chapter4:**

- For the cooperative problem among SUs, we first propose an approach using a round robin scheduling and each SU is equipped with an online learning algorithm in order to help SUs explore their environment and to make optimal decisions.
- We adopt the decentralized partially observable MDP framework to model the cooperation among SUs without coordination. We then study a decentralized learning algorithm through using the policy gradient and the Lagrange multiplier method to help SUs make local optimal decisions without full information from others.
- Extensive performance evaluations and comparisons are performed to demonstrate the efficiency as well as the convergence of the learning algorithms.
- The proof of the convergence of the decentralized learning algorithm is also provided.

- **Chapter5:**

- We present conclusions for the thesis and propose some potential research directions.

Chapter 3 and Chapter 4 present two typical scenarios in CRNs with RF energy harvesting capability, i.e., optimizing the performance for independent SUs and for cooperative SUs, respectively. To find the optimal policy for an independent SU, we consider three cases, when the SU has complete, incomplete, and unknown information about environment, and propose corresponding solutions using MDP frameworks. For the the scenario with multiple SUs that want to cooperate, obtaining full and complete information from all SUs to find the centralized optimal policy is intractable. Therefore, we propose two approaches, namely, TDMA-learning and decentralized learning algorithm, which can reduce significantly the communication overhead and avoid the computation complexity of centralized solutions. The relation between Chapter 3 and Chapter 4 is illustrated in Fig. 1.7.

Figure 1.7: The relation between Chapter 3 and Chapter 4.

Chapter 2

Literature Review

In this chapter, we first discuss problems studied in cognitive radio networks (CRNs) with RF energy harvesting techniques. We then present some related works which exploit general energy harvesting techniques in CRNs. Finally, we identify research trends in this topic and introduce the scope and the novelty of the thesis.

2.1 Cognitive Radio Networks with RF Energy Harvesting

2.1.1 Channel Access Strategy for Secondary Users

Similar to cognitive radio networks (CRNs), the channel access problem in RF powered CRNs has also received a lot of attention from researchers in the literature. In [51], a secondary user (SU) is equipped with a finite energy buffer to harvest energy from primary signals when it is close enough to primary transmitters. In addition, the SU can transmit data to its destination when it is sufficiently far away from primary transmitters. Based on these assumptions the authors proposed a stochastic-geometry model for the CRN where primary users (PUs) and SUs are distributed as independent homogeneous Poisson point processes (HPPPs) [52]. In this model, PUs are protected from interference from the SU by a *guard zone* and they transfer a significant amount of RF energy to the SU when it lies in a *harvesting zone*. Based on the stochastic-geometry model, the authors adopted the Markov-chain theory to derive the transmission probability of the SU and then used a Poisson approximation method to derive the maximum throughput for the SU. An important result found

in this paper is that the maximum throughput of the SU is linearly increasing with increasing PUs' density. Furthermore, the PUs' density is inversely proportional to the transmission probability of the SU.

In [51], the authors assumed that all SUs have to be fully charged before they can transmit data. This implies that all SUs transmit data with the same power which may limit the network capacity for the secondary system. Therefore, the authors in [53] proposed the energy-based opportunistic spectrum access (OSA) strategy which allows the SUs to use the variable power mode to transmit data instead of waiting until full charging as in [51]. With the proposed OSA strategy, an SU can transmit data when it is outside of the guard zone and its harvested energy level is greater than a predefined threshold. As a result, the reliability and stability of the secondary system can be significantly improved.

Park *et. al.* [54] studied an optimal mode selection policy for an SU in a CRN with RF energy harvesting capability. It was assumed that the SU can harvest energy when the PU transmits data or transmit data when the PU is idle. However, the SU cannot harvest RF energy and access the primary channel simultaneously, thus it has to opt between *access mode* and *harvest mode* at the beginning of each time slot. Alternatively, it is assumed that the SU does not have any information about the current state of the channel in advance, but it knows a part of channel state (e.g., idle channel probability). Therefore, the partially observable MDP (POMDP) framework was adopted to obtain the optimal mode selection policy for the SU. In order to improve the energy efficiency as well as the spectrum usage efficiency of the proposed model, the authors then proposed appropriate solutions in [55] and [56]. In particular, in [55], the authors developed a method to adjust the energy causality constraint and the collision constraint, which results in increasing the probability of accessing the occupied channel. While in [56], a theoretical upper bound on the maximum achievable throughput given the aforementioned constraints was derived with the main aim of deeply examining impacts of the temporal correlation of the primary system to the secondary system. Then, in [57] Park *et. al.* designed an optimal spectrum access strategy in order to obtain the theoretical maximum achievable throughput as

presented in [56]. The numerical results then verified that by taking the temporal correlation of the primary system into account, the proposed strategy can achieve high efficiency in using the harvested energy.

Some variations of [54] were studied in [58] and [59]. Specifically, in [58], based on the current energy level, the SU has three actions to choose from at each time epoch, i.e., sensing, accessing without sensing, or sleeping, instead of using two actions as in [54], i.e., sensing or sleeping. Then the optimization problem was formulated as a linear program which can be efficiently solved by using software tools, e.g., Matlab. The main aim of this paper is to prove that the spectrum sensing process of the SU is not always useful especially when the SU is very limited by energy. Different from [54, 58], [59] considered a scenario in which the secondary system consists of a secondary transmitter (AP) and a set of secondary receivers (SRs). Then, a cooperative spectrum sensing strategy including three stages was proposed. In the first state, all SRs use their energy detection to make local decisions (i.e., harvesting energy or transmitting data) and send requests to the AP in the second stage. After that, in the third stage, the AP makes the final decision, i.e., choosing one of the SRs for the data transmission or harvesting energy for all SRs. It is clear that with the cooperative spectrum sensing, the sensing accuracy will be higher and the network performance of the secondary system will be greater, but this kind of network requires more cooperative SRs and this may cause a communication overhead.

In addition to channel access strategies for CRNs with RF energy harvesting techniques, there were also some related works studying these strategies for correlated networks, e.g., wireless body area networks and device-to-device communications. In particular, in [60], the cognitive radio network with RF energy harvesting capability for body area networks was studied. In particular, physiological information of a patient is gathered by sensors and sent to the hospital through primary channels when these channels are not occupied by primary users. Sensors are assumed to be powered by RF energy scavenging and then they use this energy to access the channels opportunistically. Alternatively, the authors discussed challenges in designing and

implementing the physical, MAC, and the network layer and introduced some potential solutions. Finally, a practical scenario was examined to evaluate the considered solutions for the proposed model in utilizing the harvested RF to power sensor nodes. Through the obtained results, the authors concluded that RF energy harvesting is a promising future for the power supply in wireless body area networks.

The authors in [61] introduced a model for cognitive device-to-device (D2D) communications using RF energy harvesting in cellular networks. In this model, D2D transmitters are able to harvest RF energy from ambient interference caused by transmissions of macro base stations and cellular users, and then perform spectrum sensing to opportunistically access a predefined nonexclusive D2D channel. To deal with the multi-channel and the coexistence of the cellular and D2D users, two different spectrum access channel policies for cellular communication were presented. Under these channel access policies, a packet is transmitted successfully from a D2D transmitter to its receiver if the D2D channel is free and the signal-to-interference-plus-noise ratio (SINR) at the D2D receiver satisfies the required threshold. The authors then adopted tools from stochastic geometry in order to evaluate the performance of the aforementioned system. One of the most important findings in this paper is that under the low density of base stations, downlink channels have better performance than uplink channels. However, in an environment with high density of base stations, uplink channels are preferable for D2D transmissions.

2.1.2 Relaying

In relaying CRNs with RF energy harvesting technique, relay nodes can be used to improve performance for such networks through supporting source nodes to transmit data to their far destination nodes. There are two typical scenarios considering in such networks corresponding to two main functions of the relay nodes, i.e., relay for secondary systems and primary systems.

2.1.2.1 Relaying for secondary systems

[62] studied an application of RF energy harvesting technique in a cognitive amplify-and-forward relaying network. In particular, there is a cognitive radio network with

one PU and three SUs, called a source SU (SS), a relay SU (RS), and a destination SU (DS), as illustrated in Fig. 2.1. The RS forwards the information received from SS to the DS by using the energy harvested from RF signals of the SS with the aim to maximize the throughput of the secondary system. At the beginning of each time slot, the SUs sense the primary channel. If the primary channel is busy, SUs use the IDLE mode. On the other hand, if the primary channel is idle, SUs can perform the transmit data process which consists of two phases, i.e., from SS to RS and from RS to DS (each phase lasts for $\frac{1}{2}$ time slot). Given the proposed system, the authors formulated the throughput maximization problem and then adopted an approximation method using the upper bound to mitigate the complexity of the optimization problem. In addition, a suboptimal algorithm was developed in order to achieve the near-optimal throughput performance. The results demonstrated that the throughput obtained by the proposed approximation algorithm is close to that of the optimal solution, and it has a significant gain compared with the separate management algorithm. A similar model was investigated in [63], but under the underlay CRN. Moreover, in [63], the authors focused on deriving the outage probability for the secondary system with the aim of improving the energy conversion efficiency for such a secondary system.

Figure 2.1: Relay-assisted cognitive radio networks.

In [64], a similar relay cognitive radio model with RF energy harvesting techniques was studied, but different from [62, 63], in [64] the relay node harvests energy from the primary transmitter's signals instead of the secondary transmitter's signals. In

addition, different from [62] which considers an overlay spectrum sharing cognitive radio with the amplify-and-forward relaying technique, in [64], the authors took an underlay spectrum sharing cognitive radio with the decode-and-forward relaying technique into account. The aim of [64] is to demonstrate that the use of relay CRNs with simultaneous wireless information and power transfer (SWIPT) will not cause any loss of diversity gain, although it can reduce the outage performance of the system. Thus, it was concluded that, to improve the network performance, one of the possible ways is adopting the MIMO technique for SWIPT CRNs.

In [65, 66], the authors extended the scenario introduced in [62] by considering a set of secondary relay nodes instead of only one secondary relay node as in [62]. With the coexistence of multiple relay nodes in the secondary system, the problem now becomes how to select the best relay node and how to control the transmission power at the secondary transmitter and the relay nodes. There are some differences in designing relay powered CRNs which lead to differences in finding the joint optimal relay selection and power allocation scheme in [65, 66]. For example, in [65], the authors considered a two-way amplify-and-forward relaying method while in [66] a one-way decode-and-forward method for the secondary system was examined. Then, corresponding optimal solutions were presented to maximize the throughput of the secondary system under interference constraints for the primary system.

Another scenario was introduced in [67] in which the secondary relay node is powered by the secondary users' signals instead of the primary users' signals. In [67], the authors proposed a framework for CRNs using the decode-and-forward two-way relaying technique. In this network, there are two SUs, namely, A and B , communicating through a secondary relay node (R) by two-way communications. Node R is assumed to be able to harvest RF energy from both A and B , and it then uses such harvested energy to forward packets for both A and B , i.e., $A \rightarrow R \rightarrow B$ and $B \rightarrow R \rightarrow A$. To improve the network performance for the secondary system, there are two relaying mechanisms, namely, multiple access broadcast (MABC) and time division broadcast (TDBC), and two power transfer policies, called, dual-source (DF)

and single-fixed source (SFS), were proposed. The differences between the two protocols are presented in Fig. 2.2. In addition, in the data transmission phase, to protect the PU's signals, the authors defined the maximum tolerance interference parameter to control the transmit power of SUs. Through numerical results, the authors showed that the MABC protocol can achieve better performance than that of the TDBC protocol, while the DS policy always gets better efficiency than that of the SFS policy in both relaying protocols.

Figure 2.2: The comparison of the data frame structure between TDBC and MABC protocol in two-way relay cognitive radio network (TWRCN).

2.1.2.2 Relaying for primary systems

[68] considered a scenario in which there are two primary users, denoted by S and D, who want to exchange information, but the distance between them is too far and thus they cannot communicate directly with each other. However, there is a secondary user R who volunteers to relay signals for primary users S and D. At the same time, the secondary relay node R also has its own information and wants to transmit to the secondary user C as illustrated in Fig. 2.3. Consequently, the node R has to transmit the primary relay information and its own information simultaneously. It is assumed that when the relay node R receives signals from the primary node, it is able to extract information and energy from the received RF signals. Then, by using the harvested energy, the relay node can transmit information to primary and secondary nodes. For the proposed aforementioned network, the authors formulated the outage probability expressions for the primary system and derived lower/upper

bounds of outage probability for the secondary system. From analysis results, it was demonstrated that the proposed protocol has better outage performance than direct transmission without spectrum sharing. More numerical results can be found in [69].

Figure 2.3: Two-way cognitive relay network.

A similar model was also considered in [70]. However, instead of using the decode-and-forward protocol as in [68], the authors in [70] adopted the amplify-and-forward protocol for the relay process. In addition, in [70], the authors proposed three schemes for the energy and information cooperation problem between PUs and SUs. For each scheme, the authors introduced corresponding optimization solutions, and they also indicated that the scheme based on the power splitting can achieve larger rate region than that of the time splitting scheme if the harvested energy is sufficient. Nevertheless, in [70], the authors did not consider the time constraint in which a secondary relay node needs to forward the whole received PU's signals to the primary receiver. To overcome this issue, the authors in [71] proposed a time-divided power splitting scheme taking both the influence of time division proportion and the power splitting into account. By using this scheme, it was proved that the relay SU will have enough time to forward all received data from the PU if the energy supply is sufficient.

In [72], the authors also considered using an SU as a relay node to help a PU to transmit data to an access point (AP). However, different from all aforementioned works, in [72], the authors considered the case when RF energy is harvested at the

primary node. In particular, the authors assumed that the AP can transfer power to its PU over the forward link. Furthermore, to enhance the energy harvesting efficiency, the SU who is nearest to the PU will be chosen to transmit RF energy beside the harvested power from the AP. For the data transmission efficiency, the potential SU who has the best channel quality will be selected to forward data to the AP for the PU. As a redemption, a part of the available primary bandwidth will be used for data transmissions of SUs. Numerical and simulations results verified that the throughput of the primary system can be significantly improved through the proposed system.

[73] extended the scenario in [68] with a set of secondary relay nodes instead of only one as in [68]. Due to the multiple relay nodes, the authors proposed a protocol with three phases to address the relay selection and the resource allocation for the RF powered cognitive relay system. In the first phase, all relay nodes harvest energy from the PU's signals. Then, in the second phase, the PU transmits data to relay nodes and one of the relay nodes will be selected to forward data to the primary receiver based on an optimal relay selection algorithm with the aim of maximizing the throughput for the primary system. Finally, in the third phase, the selected secondary relay node transmits the PU's received signals together with its own signals to the corresponding destinations. Numerical results showed that the proposed relay selection algorithm is more efficient in the low SNR scenario.

Different from all above works which use SUs as relay nodes, in [74], the authors studied a scenario in which there is a relay node in the primary system to help a PU to transmit data to its distance destination. In particular, the authors introduced an underlay two-hop CRN including a primary transmitter P_t , a primary receiver P_r and relay node R to help P_t communicate with P_r , which coexists with a secondary system including a secondary base station (SBS) and multiple SUs. In the first hop, when P_t transmits data to the R , SUs can harvest energy from the nearby PU' signals. Then, in the second hop when the R relays data to the P_r , SUs can use the harvested energy to communicate with their SBS through adopting the dynamic spectrum access technique. The authors showed that the proposed framework not only greatly

enhances the secondary system's performance, but also helps the primary system get more profits for both energy transfer and spectrum sharing.

2.1.3 Time Scheduling for Secondary Users

In CRNs using energy harvesting techniques, SUs not only find opportunities to access primary channels, but also have to harvest energy. The problem is that if the energy harvesting time is too long, the data transmission time will be reduced, while if the energy harvesting time is insufficient, the harvested energy will not be adequate for the data transmission phase. This leads to the problem of how to balance between the energy harvesting time and the data transmission time for SUs to maximize their performance.

In [75], the authors proposed an online solution to find the optimal trade-off time between the energy harvesting phase (EHP) and the data transmission phase (DTP) for an underlay CRN. In the first phase, i.e., the EHP, the SU harvests energy from a PU's signals and uses this energy to transmit data in the second phase, i.e., the DTP, under interference constraints with the primary system. Then, to find the optimal trade-off time between the EHP and the DTP, the authors adopted the convex optimization technique and derived the optimal value for the time-sharing ratio in a similar way as in [76]. With the proposed solution, the authors showed that the interference constraints with the primary system are under control and the average achievable rate of the SU is maximized.

The same scenario was also examined in [77], but in [77] the authors also considered a cooperation scenario between the primary system and the secondary system. Specifically, in the second phase, the SU can opt to transmit data to its destination or relay data for the primary system. Consequently, the SU now has to determine not only how much time on the EHP, but also how much power for data relay or data transmission to allocate. For both cases, i.e., cooperation and non-cooperation, the authors proposed corresponding optimal solutions by using close-form solutions with numerical analysis. Simulation results showed that the proposed solutions can achieve

better performance than that of the stochastic cooperation protocol and the optimal underlay transmission protocol.

In underlay CRNs, SUs can access primary channels even when such channels are currently occupied by primary users as long as the interference level of the primary system is under control. However, in overlay CRNs, SUs can only access the primary channels when such channels are idle, i.e., not occupied by PUs. Consequently, sensing is a compulsory process in overlay CRNs. For SUs in overlay CRNs with energy harvesting capability, if they spend more time on the EHP, they may have sufficient energy for the spectrum sensing phase (SSP) and DTP phase, but the remaining time is less resulting in low throughputs. On the other hand, if the SUs spend too much time for the SSP and DTP phase, they can get more available channels and longer time to transmit data, but the harvested energy is insufficient to serve for such tasks. Thus, there were some solutions proposed to overcome this problem.

[78] introduced a strategy, called, one-step-ahead spectrum sensing, with the aim of balancing the time allocation among three phases, i.e., EHP, SSP, and DTP. For the proposed strategy, based on the information about the current system state, e.g., available channels, the current energy level, and idle channel probabilities, and the estimation about the next state, the SU will make the best decision to maximize its throughput. However, the proposed strategy just can obtain a myopic solution, and thus an optimal saving-sensing-transmitting structure was proposed in [79,80] to maximize the average throughput for the SU. In particular, the authors first formulated the time allocation optimization problem as a mixed integer non-linear programming problem and then employed a heuristic algorithm developed from the different evolution algorithms [81] to derive the optimal time structure for three phases. Simulation results verified the efficiency of the proposed solution compared with a stochastic sensing strategy.

2.1.4 Other Related Issues

2.1.4.1 Power allocation

[82] studied the power allocation problem for a secondary network in which the secondary receiver (SR) is able to harvest RF energy from its secondary transmitter (ST) as well as from a primary transmitter (PT). It was assumed that the secondary system adopts the orthogonal frequency division multiplexing (OFDM) modulation for transmitting data from ST to SR. Thus the main goal of this paper is to find the optimal power allocation policy on each subcarrier and the power splitting ratio (i.e., the power ratio between the energy harvesting process and the information decoding process) at the ST in order to maximize the energy efficiency of the secondary system. The optimization power allocation problem is transformed to an equivalent convex problem that is solved efficiently by using an iterative algorithm. Simulation results showed that the iterative algorithm can achieve a great tradeoff between the energy efficiency and high SINR regions.

Differently from [82], [83] examined the power allocation problem for the primary system and then analyzed its impacts on both the primary system and the secondary system. First, in the secondary system, the secondary transmitter is allowed to transmit packets to the primary channel iff the channel is not occupied by the PU. In addition, ST can harvest energy from the primary transmitter's signals as well as from the environment. Then, the authors introduced a power allocation scheme for the primary system in which the PT sets the transmission power at time slot t based on the following equation $P_p^*(t) = \frac{N_o W (2^{R_p} - 1)}{h_{ppd}(t)}$ where N_o is the additive white Gaussian noise power spectral density, W is the channel bandwidth, R_p is the targeted primary spectral efficiency, and h_{ppd} is the instantaneous channel gain. Under the proposed power allocation policy, the authors analyzed its impacts to both systems and concluded that by using the proposed policy at the primary system and implementing the RF energy harvesting at the secondary system, we can enhance the throughput for both systems.

2.1.4.2 Scheduling and security

In [84], the authors considered a primary system coexisting with an underlay secondary system which consists of one secondary transmitter (ST) and multiple secondary receivers (SU-Rx_N). Each SU-Rx_n is able to harvest RF energy from the PU's signals or decode information received from the ST at a time. It means that SU-Rx_n cannot carry out both processes concurrently. To avoid the collision among SU-Rx_n's transmissions, there is only one SU-Rx selected to decode information from the ST. Consequently, the SU-Rx with the best channel condition will be scheduled to decode information from the ST, while other SU-Rx will harvest RF energy from the ST's transmission. Then, in order to maximize the network throughput for the secondary system, a threshold condition using the Max-SNR scheduling [85] was adopted to find the best SU-Rx at each time epoch. Numerical results showed that the proposed scheduling strategy can obtain a desirable QoS for the secondary system.

Similar to [84], the authors in [86] also investigated the scheduling problem for the secondary system by choosing only one secondary receiver at each time slot for data transmission. However, in [86], to guarantee the communication security in the secondary system, the authors employed a resource allocation algorithm which treats both idle secondary receivers and primary receivers as potential eavesdroppers. In addition, different from [84], a non-convex multi-objective optimization problem is formulated with the aim of jointly minimizing the total transmission power of the multi-antenna secondary transmitter and maximizing the efficiency of harvesting energy, while guaranteeing to minimize the interference power leakage-to-transmit power ratio for the primary system. Numerical results then figured out an interesting tradeoff between the considered conflicting system design objectives.

Similar to [86], in [87], the authors considered not only scheduling strategies, but also the security transmission problem for the secondary system. In particular, the authors introduced a secure cognitive device-to-device (D2D) communication in cellular networks. In this network, there is a D2D transmitter (Alice) who wants to transmit data to D2D receivers (Bobs) under an environment with the presence of

D2D eavesdroppers (Eves). Alice is an energy-constrained device which needs to be powered by power beacons (PBs) before it can transmit data to Bobs. The spatial topology of the proposed network is modeled using HPPPs [52]. Each time frame is divided into two time slots. In the first time slot, Alice is scheduled to harvest RF energy from the strongest PB. In the second time slot, Alice uses the harvested energy to transmit packets to one of Bobs based on one of two schemes, so-called, the best receiver selection scheme (BRS), i.e., select the Bob with the best channel, and the nearest receiver selection scheme (NRS), i.e., select the nearest Bob. Based on these two schemes, the authors treated the secrecy throughput as a metric to characterize the security performance for the D2D communications between Alice and Bobs. Through analytical and numerical results, the authors showed that although the BRS achieves the better performance in terms of higher secrecy throughput than that of NRS, it causes more communication overhead than NRS.

2.1.4.3 Performance analysis

There were a couple of works studying the feasibility of CRNs with RF energy harvesting techniques through using Markov chain models [88, 89]. In particular, in [88] a Markov chain model was adopted to analyze the feasibility of CRNs where SUs can harvest RF energy from only one source, i.e., signals from a PU. In this model, at the beginning of each time slot, the SU senses the PU's channel and if the channel is not occupied by the PU, the SU will attempt to transmit data to the channel in the rest of the time slot if it has sufficient energy. For such kind of network, there are two constraints which need to be taken into consideration. The first constraint is the energy causality constraint which expresses the relation between the amount of the harvested energy and the ability to transmit data at the SU, while the second constraint is the collision constraint with the PU's signals which can happen when the SU's sensing process is wrong, e.g., miss detection. Accordingly, the state of the system is defined as the joint state of the battery levels of the SU and the channel state of the PU. Through using the Markov chain model, many analytical results were

derived with the aim to prove that RF energy harvesting CRNs is a promising solution for delay-tolerant sensor networks.

In [89], the authors also want to show the feasibility of the RF energy harvesting CRNs under the energy constraint. However, in this paper, the Markov chain model just takes the energy constraint into account and thus the system state is interpreted as the energy level at the SU only. Furthermore, in [89], there were two channel access schemes considered, namely, random access and sensing-based access scheme, and the network performance was thus analyzed according to these two schemes. A similar conclusion as in [88] was also presented, i.e., the RF powered CRN is appropriate to low data rate applications.

In [90], the authors presented analytical results for a RF powered CRN in the comparison with the RF powered non-CRNs. In this paper, a similar model as in [51] was considered in which SUs and PUs are distributed as an independent homogeneous Poisson process [52]. However, different from [51] which mainly focuses on channel access strategies, in [90] the authors tended to analyze the gain estimation and the signal to noise ratio for SUs with the aim to demonstrate that compared with non-CRNs, the number of active SUs in CRNs is linearly greater as the number of SUs increases. This implies that by integrating RF energy harvesting techniques into CRNs, we can significantly improve the efficiency for secondary systems.

2.2 Cognitive Radio Networks with Other Wireless Energy Harvesting Techniques

In addition to using RF energy harvesting techniques, there were also many research works exploiting ambient energy sources (AESs), e.g., solar, wind, and so on, for CRNs. The major difference in using the RF energy harvesting techniques and techniques exploiting AESs in CRNs is that techniques exploiting AESs are much dependent on the environment conditions, while RF energy harvesting techniques depend on radio power transmitter sources. Consequently, in CRNs using AESs, the energy process and the data transmission process are usually assumed independently, while

in RF powered CRNs these processes have a strong relation. In particular, both the energy harvesting process and the data transmission process in RF powered CRNs use radio signals as a means to convey data and power and thus they can cause interference if both processes occur concurrently at a wireless node. On the other hand, for CRNs using AESs, the energy harvesting process and the data transmission process can happen simultaneously and the energy harvesting processes are often assumed to be followed by a random process, e.g., Poisson process, which is completely independent with data transmission processes and has no impact on the transmission processes. Consequently, problems considered in CRNs using AESs are often less complicated than that of RF powered CRNs.

Similar to RF powered CRNs, there are also some common problems which are well investigated such as channel access strategies, relaying, time scheduling, and some other related issues, e.g., power control, performance analysis, security, and PU detection. To avoid overlaps with problems discussed in Section 2.1 and because this is not our main concern, in this Section, we just summarize research works studying problems related to CRNs using AESs in Table 2.1 and Table 2.2.

2.3 Research Trends, Scope and Novelty of the Thesis

In Fig. 2.4, we summarize research works in the literature and show the research trends in RF powered CRNs as well as in general wireless powered CRNs. As shown in Fig. 2.4, applications of wireless energy harvesting techniques in CRNs have received a lot of attention recently. In particular, there were just a couple of works in 2012, but they have been increasing doubly every year. Compared with general wireless powered CRNs, RF powered CRNs have been receiving more attention because of their many outstanding advantages as presented in Chapter 1 (Section 1.3). As a result, in this thesis, we mainly focus on studying applications of *RF energy harvesting techniques for CRNs*.

As shown in Fig. 2.5, channel access strategies and relaying problems are the most interesting topics which account for nearly 60% of research works in CRNs

Figure 2.4: Research trends of CRNs with energy harvesting capability.

with energy harvesting capability. This fact is quite clear when the channel access strategy is the most important problem in CRNs as well as wireless powered CRNs, while relay cognitive radio networks have been receiving a lot of attention recently because of potential advantages of both CRNs and relay networks. There are some other research topics that have also received attention from researchers such as power allocation, time scheduling, scheduling and security, but they are not key issues in wireless powered CRNs and thus we will not take them into account in this thesis.

Figure 2.5: Summary information of the percentage of problems in (a) general wireless powered CRNs and (b) RF powered CRNs.

In this thesis, we mainly focus on the *channel access problem* for RF powered CRNs which is the most important issue of CRNs as well as general wireless powered CRNs. In all research works studying the channel access strategies in CRNs with wireless energy harvesting capability in the literature, they considered only one primary channel. The channel access problem for wireless powered CRNs will become

more complicated when there is more than one channel for SUs to choose from which will lead to the channel selection problem for SUs. This is a common problem that we have to face in CRNs with energy harvesting in practice when SUs always have a set of licensed channels to choose from and thus they need to find the best channel to sense with the main aim of maximizing their throughput. To the best of our knowledge, our work is the first work that studies the performance optimization problem in RF powered CRNs with the presence of multiple licensed channels. In addition, in this thesis, we also present novel models as well as solutions to address the aforementioned problem.

Table 2.1: Problems exploiting ambient energy sources in cognitive radio networks

P	Article	Objective	Solution
Channel Access	[91] [92] [93]	Maximize the SU's throughput under the PU' QoS constraint	Propose a feedback-based access and sensing scheme
	[94]	Establish and analyze the stability region of the system	Use the stochastic dominance technique [95] and Lyoynes' theorem [96]
	[97]	Maximize the SU's throughput	MDP framework with the value iteration algorithm
	[98]	Maximize the SU's throughput under energy neutrality constraint and fading channel condition	POMDP framework and the value iteration algorithm
	[99]	Ensure energy efficiency as well as throughput performance for the secondary system	Propose an energy level based MAC protocol
	[100]	Maximize the SU's throughput	Markov chain and Matlab's solver
Relaying	[101]	Maximize the SU's throughput	Cooperative cognitive relaying protocol
	[102]	Maximize the SU's throughput	Markov chain and Alamouti coding scheme
	[103]	Maximize the SU's throughput under PU's QoS requirement	Design a flexible and dynamic relaying cooperative protocol
	[104]	Maximize the secondary expected utility	Q-learning algorithm
	[105]	Optimize jointly relay selection and power allocation	Lagrangian multiplier method
	[106]	Find the stable throughput region	Use principle of stochastic dominance [95]
Time Scheduling	[107]	Design a practical circuit model and transmission protocol	Propose a save-then-transmit protocol
	[108]	Maximize the SU's throughput under the stability of the PU and SU	The linear-fractional programming [109]
	[110]	Maximize the SU's throughput	Optimize jointly the sensing duration and the energy detector's sensing threshold

Table 2.2: Problems exploiting ambient energy sources in cognitive radio networks

P	Article	Objective	Solution
Power Control	[111]	Optimize transmission power of the secondary transmitter	Employing the sliding window approach
	[112]	Optimize transmission power of the secondary transmitter	Divide the optimization problem into two sub-problems and use analysis to find optimal solution
	[113]	Optimize the SU's power allocation policy under PU's interference constraint	Propose the geometric water-filling with peak power constrains and recursion machinery
	[114]	Optimize the SU's power allocation policy under PU's interference and SU's QoS constraints	Develop a non-convex approximation and a successive approximation approach
Performance Analysis	[115]	Explore the feasibility of the energy harvesting empowered CRNs in small cellular networks	Design a comprehensive framework to characterize the performance of the cognitive metro-cellular networks powered by solar energy harvesting
Security	[116]	Optimize the secrecy rate of the PU transmitter	Use the mixed integer non-linear program and a polynomial time algorithm
	[117]	Optimize SU's throughput under jamming attacks	Use MDP framework and learning algorithms
PU Detection	[118] [119]	Improve the primary user detection performance	Use hidden Markov model and a 2-D spectrum and power sensing scheme

Chapter 3

Optimal Channel Access for Cognitive Users with Energy Harvesting Capability

Recently, a radio frequency (RF) energy harvesting technique with high efficiency has been introduced. Such a technique allows a wireless node to harvest and convert electromagnetic waves from ambient RF sources (e.g., TV, radio towers and cellular base stations) into energy which can be used for data transmission. A few recent studies have shown the feasibility of the RF energy harvesting (e.g., [120], [121], [122]). For example, the study in [121] showed that with the transmit power of 0.5W by a mobile phone, 40mW, 1.6mW, and 0.4mW of power can be harvested at the distance of 1, 5, and 10 meters, respectively. With the RF energy harvesting capability, the wireless especially mobile node can perpetuate its operation without physically changing or recharging its battery. Cognitive radio can utilize the RF energy harvesting, where the secondary user is equipped with the RF energy harvesting device. To obtain enough energy and spectrum opportunity for packet transmission, the secondary user must search for not only an idle channel (i.e., spectrum opportunity) to transmit its packets, but also a busy channel to harvest RF energy. The channel access, which determines the channel for the secondary user to transmit a packet or to harvest RF energy, is the crucial component for the secondary user to achieve optimal performance.

In this chapter, we consider the aforementioned cognitive radio network. Specifically, we consider the channel access problem in which the secondary user selects the channel to access for packet transmission or RF energy harvesting. The channel

access policy which provides a mapping from the state of the secondary user to an action (i.e., a channel to select) is the solution. We first introduce the optimization formulation based on a Markov decision process (MDP) for the secondary user to obtain the channel access policy without knowing the current channel status.¹ The objective is to maximize the throughput of the secondary user. However, the optimization requires the secondary user to have model parameters and also solving the optimization entails considerable computation resource, which may not be feasible in practice. Therefore, we propose an online learning algorithm for the secondary user to obtain the channel access policy. With the learning algorithm, the secondary user can observe and adapt the channel access action to achieve the objective. Additionally, we also consider the complete information case, where the secondary user fully knows the status of all the channels. We present the optimization formulation for this case, where the policy is able to achieve the highest throughput. This performance can serve as a benchmark for the secondary user.

The rest of this chapter is organized as follows. Section 3.1 describes the system model and assumptions used in this chapter. Section 3.2 considers the optimization formulation based on an MDP for the secondary user with some statistic information from primary channels. In Section 3.3, we study the online learning algorithm for the secondary user to obtain the channel access policy without the model parameters. Finally, evaluation results are examined in Section 3.4 and we summarize this chapter in Section 3.5.

3.1 System Model

We consider a cognitive radio network with N primary users and a secondary user (Fig. 3.1). Without loss of generality, the primary user n is allocated with the non-overlapping channel n . The primary users use such channels to transmit data on a time slot basis and all the primary users align to the same time slot structure. Therefore, in one time slot, a channel can be idle or busy (i.e., free or occupied by the

¹By the “channel status” we mean the channel is idle or busy (i.e., occupied by the primary user).

Figure 3.1: System model.

primary user for data transmission). We consider the secondary user with RF energy harvesting capability. The secondary user performs channel access by selecting one of the channels. If the selected channel is busy, the secondary user can harvest RF energy from the primary user's transmission. The probability that the secondary user harvests one unit of RF energy successfully from channel n is denoted by γ_n . The harvested energy is stored in the energy storage, whose maximum size is E units of energy. By contrast, if the selected channel is idle, the secondary user can transmit the packet retrieved from its data queue. The secondary user requires W units of energy for data transmission in a time slot. The probability of successful packet transmission on channel n is denoted by σ_n . The probability of a packet arrival at the secondary user in a time slot is denoted by α . The arriving packet is buffered in the data queue of the secondary user. The maximum capacity of the data queue is Q packets. Note that the batch packet arrival and transmission can be incorporated straightforwardly. Also, the system model consideration and optimization can be extended for multiple secondary users by recalculating the successful packet transmission σ_n based on a contention mechanism.

We also consider spectrum sensing errors of the secondary user on the selected channel. The miss detection happens when the actual channel status is busy, but the secondary user senses it to be idle. By contrast, the false alarm happens when the

actual channel status is idle, but the secondary user senses it to be busy. The miss detection probability and the false-alarm probability are denoted by m_n and f_n for channel n , respectively.

We assume that the channel is modeled as a two-state Markov chain. The transition probability matrix of channel n is denoted by

$$\mathbf{C}_n = \begin{bmatrix} C_{0,0}(n) & C_{0,1}(n) \\ C_{1,0}(n) & C_{1,1}(n) \end{bmatrix} \begin{matrix} \leftarrow \text{idle} \\ \leftarrow \text{busy} \end{matrix} \quad (3.1)$$

where “0” and “1” represent idle and busy states, respectively. The probability of the channel n to be idle is denoted by $\eta_n = \frac{1-C_{1,1}(n)}{C_{0,1}(n)-C_{1,1}(n)+1}$.

The channel access of the secondary user is based on the policy, which could be obtained in the following cases.

- The secondary user has some knowledge about the environment which we refer to as the model parameters. Examples of the model parameters are the probabilities of successful packet transmission and successful RF energy harvesting, and the probability of the channel to be idle. However, the secondary user does not know the current channel status (i.e., whether it is idle or busy) when it selects the channel to sense. We will propose the optimization formulation based on a Markov decision process to obtain the channel access policy. The detail of the formulation is given in Section 3.2.
- The secondary user does not have the knowledge about the environment (i.e., the model parameters). Therefore, the secondary user has to observe the environment and select the channel to access according to its own information. We will study the online learning algorithm to obtain the channel access policy. The detail of the learning algorithm is given in Section 3.3.
- Similar to the first case, in addition to having the model parameters, the secondary user is also assumed to know the current status of all the channels. The optimization is formulated to obtain the channel access policy. Although this case might not be common in practice, the optimization can yield the upper-bound performance for benchmarking. The detail of the optimization formulation for this case is given in Appendix C.

3.2 Markov Decision Process Formulation

In this section, we present the optimization based on a Markov decision process (MDP) to obtain the channel access policy for the secondary user. Firstly, we define the state and action spaces. Next, we derive the transition probability matrix and describe the optimization formulation. Then, we obtain the performance measures.

3.2.1 State Space and Action Space

We define the state space of the secondary user as follows:

$$\Theta = \left\{ (\mathcal{E}, \mathcal{Q}); \mathcal{E} \in \{0, 1, \dots, E\}, \mathcal{Q} \in \{0, 1, \dots, Q\} \right\} \quad (3.2)$$

where $\mathcal{E}$ and $\mathcal{Q}$ represent the energy level of the energy storage and the number of packets in the data queue of the secondary user, respectively. E is the maximum capacity of the energy storage and Q is the maximum data queue size. The state is then defined as a composite variable $\theta = (e, q) \in \Theta$, where e and q are the energy state and number of packets in the data queue, respectively. The action space of the secondary user is defined as follows: $\Delta = \{1, \dots, N\}$, where the action $\delta \in \Delta$ is the channel to be selected by the secondary user for transmitting a packet or harvesting RF energy.

3.2.2 Transition Probability Matrix

We express the transition probability matrix given action $\delta \in \Delta$ of the secondary user as follows:

$$\mathbf{P}(\delta) = \begin{bmatrix} \mathbf{B}_{0,0}(\delta) & \mathbf{B}_{0,1}(\delta) & & & \\ \mathbf{B}_{1,0}(\delta) & \mathbf{B}_{1,1}(\delta) & \mathbf{B}_{1,2}(\delta) & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{B}_{Q,Q-1}(\delta) & \mathbf{B}_{Q,Q}(\delta) & \end{bmatrix} \begin{matrix} \leftarrow q = 0 \\ \leftarrow q = 1 \\ \vdots \\ \leftarrow q = Q \end{matrix} \quad (3.3)$$

where each row of matrix $\mathbf{P}(\delta)$ corresponds to the number of packets in the data queue (i.e., the queue state). The matrix $\mathbf{B}_{q,q'}(\delta)$ represents the queue state transition from q in the current time slot to q' in the next time slot. Each row of the matrix $\mathbf{B}_{q,q'}(\delta)$ corresponds to the energy level in the energy storage.²

²Note that the empty elements in the transition probability matrices are zeros or zero matrices with appropriate sizes.

There are two cases for deriving matrix $\mathbf{B}_{q,q'}(\delta)$, i.e., $q = 0$ and $q > 0$. For $q = 0$, there is no packet transmission, since the data queue is empty. As a result, the energy level will never decrease. Let $\mathbf{B}_0(\delta)$ denote a common matrix for $q = 0$. We have

$$\mathbf{B}_0(\delta) = \begin{bmatrix} 1 - \eta_\delta^\circ m_\delta^\circ \gamma_\delta & \eta_\delta^\circ m_\delta^\circ \gamma_\delta & & & \\ & \ddots & \ddots & & \\ & & 1 - \eta_\delta^\circ m_\delta^\circ \gamma_\delta & \eta_\delta^\circ m_\delta^\circ \gamma_\delta & \\ & & & 1 & \\ & & & & \end{bmatrix} \begin{array}{l} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = E - 1 \\ \leftarrow e = E \end{array} \quad (3.4)$$

where $\eta_\delta^\circ = 1 - \eta_\delta$ and $m_\delta^\circ = 1 - m_\delta$. Each row of the matrix $\mathbf{B}_0(\delta)$ corresponds to the energy level e . In this matrix, the energy level of the energy storage increases only when the selected channel δ is busy, there is no miss detection, and the secondary user successfully harvests RF energy. Then, $\mathbf{B}_{0,0}(\delta) = \mathbf{B}_0(\delta)\alpha^\circ$ and $\mathbf{B}_{0,1}(\delta) = \mathbf{B}_0(\delta)\alpha$, where $\alpha^\circ = 1 - \alpha$, for when there is no and there is a packet arrival, respectively.

For $q > 0$, we have three sub-cases, i.e., when the number of packets decreases, remains the same, and increases. The number of packets decreases, when the selected channel is idle (with probability η_δ), there is no false alarm (with probability $f_\delta^\circ = 1 - f_\delta$) and no packet arrival (with probability α°), the packet transmission is successful (with probability σ_δ), and there is enough energy in the energy storage, i.e., $e \geq W$. The corresponding matrix is defined as follows:

$$\mathbf{B}_{q,q-1}(\delta) = \begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ \hline \eta_\delta f_\delta^\circ \alpha^\circ \sigma_\delta & & & & \\ & \ddots & & & \\ & & \eta_\delta f_\delta^\circ \alpha^\circ \sigma_\delta & \cdots & 0 \end{bmatrix} \begin{array}{l} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = W - 1 \\ \leftarrow e = W \\ \vdots \\ \leftarrow e = E \end{array} \quad (3.5)$$

The first W rows of the matrix $\mathbf{B}_{q,q-1}(\delta)$ correspond to the energy level $e = 0, \dots, W - 1$, which is not sufficient for the secondary user to transmit a packet. As a result, there is no change of the number of packets in the data queue. Accordingly, all the elements in these rows are zero. The first $\eta_\delta f_\delta^\circ \alpha^\circ \sigma_\delta$ appears at the row for the energy level of W , which is for the case when there is sufficient energy for packet transmission. Therefore, the number of packets in the data queue can decrease and the energy level decreases by W units.

The number of packets in the data queue can remain the same. The transition matrix is expressed as follows:

$$\mathbf{B}_{q,q}(\delta) = \left[\begin{array}{cccc} \alpha^\circ b^\dagger(\delta) & \alpha^\circ b^\dagger(\delta) & & \\ & \ddots & \ddots & \\ & & \alpha^\circ b^\dagger(\delta) & \alpha^\circ b^\dagger(\delta) \\ b_{W,0}^\diamond(\delta) & \cdots & b_{W,W}^\diamond(\delta) & b_{W,W+1}^\diamond(\delta) \\ & \ddots & \ddots & \ddots \\ & & b_{E,E-W}^\diamond(\delta) & \cdots & b_{E,E}^\diamond(\delta) \end{array} \right] \begin{array}{l} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = W - 1 \\ \leftarrow e = W \\ \vdots \\ \leftarrow e = E \end{array} \quad (3.6)$$

Again, the first W rows correspond to the case of not having enough energy for packet transmission without packet arrival. Therefore, the energy level can remain the same with the probability $b^\dagger(\delta)$ or can increase with the probability $b^\dagger(\delta)$, but cannot decrease. The energy level increases if the selected channel is busy, there is no miss detection and RF energy is successfully harvested, i.e., $b^\dagger(\delta) = \eta_\delta^\circ m_\delta^\circ \gamma_\delta$. Accordingly, the energy level remains the same with probability $b^\dagger(\delta) = 1 - \eta_\delta^\circ m_\delta^\circ \gamma_\delta$.

The rows from $e = W$ to $e = E$ of the matrix $\mathbf{B}_{q,q}(\delta)$ correspond to the case of having enough energy for packet transmission. When the number of packets in queue remains the same, we have the following cases.

- Firstly, we derive the probability that the energy level decreases by W units, denoted by $b_{e,e-W}^\diamond(\delta)$. This case happens when
 - the channel is idle with no false alarm, no packet arrival, and unsuccessful packet transmission or
 - the channel is idle with no false alarm, a packet arrival and successful packet transmission, or
 - the channel is busy with miss detection (i.e., the secondary user transmits and collides with the primary user), and no packet arrival.

Therefore, we have $b_{e,e-W}^\diamond(\delta) = \eta_\delta f_\delta^\diamond(\sigma_\delta^\circ \alpha^\circ + \sigma_\delta \alpha) + \eta_\delta^\circ m_\delta \alpha^\circ$, where $\sigma_\delta^\circ = 1 - \sigma_\delta$.

- Secondly, we derive the probability that the energy level remains the same, denoted by $b_{e,e}^\diamond(\delta)$. This case happens when

- the channel is busy with no miss detection, no energy successfully harvested and no packet arrival, or
- the channel is idle with false alarm (i.e., the secondary user defers transmission), and no packet arrival.

Therefore, we have $b_{e,e}^\diamond(\delta) = \eta_\delta^\diamond m_\delta^\diamond \gamma_\delta^\diamond \alpha^\diamond + \eta_\delta f_\delta \alpha^\diamond$.

- Thirdly, we derive the probability that the energy level increases by one unit, denoted by $b_{e,e+1}^\diamond(\delta)$. This case happens when the channel is busy with no miss detection, energy successfully harvested, and no packet arrival. Therefore, we have $b_{e,e+1}^\diamond(\delta) = \eta_\delta^\diamond m_\delta^\diamond \gamma_\delta^\diamond \alpha^\diamond$.

Note that for $e = E$, the energy level cannot increase more than the capacity of the energy storage, and hence $b_{E,E}^\diamond(\delta) = b_{E-1,E-1}^\diamond(\delta) + b_{E-1,E}^\diamond(\delta)$.

The number of packets in the data queue can increase. The transition matrix is expressed as follows:

$$\mathbf{B}_{q,q+1}(\delta) = \begin{bmatrix} \alpha b^\dagger(\delta) & \alpha b^\ddagger(\delta) & & & \\ & \ddots & \ddots & & \\ & & \alpha b^\dagger(\delta) & \alpha b^\ddagger(\delta) & \\ b_{W,0}^+(\delta) & \cdots & b_{W,W}^+(\delta) & b_{W,W+1}^+(\delta) & \\ & \ddots & \ddots & \ddots & \\ & & b_{E,E-W}^+(\delta) & \cdots & b_{E,E}^+(\delta) \end{bmatrix} \begin{matrix} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = W - 1 \\ \leftarrow e = W \\ \vdots \\ \leftarrow e = E \end{matrix} \quad (3.7)$$

The first W rows (i.e., not enough energy to transmit a packet) is similar to that of $\mathbf{B}_{q,q}(\delta)$, but with a packet arrival. Similarly, there are three cases for the rows from $e = W$ to $e = E$, when the number of packets in the data queue increases.

- Firstly, we derive the probability that the energy level decreases by W units, denoted by $b_{e,e-W}^+(\delta)$. This case happens when
 - the channel is idle with no false alarm, unsuccessful packet transmission, and a packet arrival or
 - the channel is busy with miss detection, and a packet arrival.

Therefore, we have $b_{e,e-W}^+(\delta) = \eta_\delta f_\delta^\diamond \sigma_\delta^\diamond \alpha + \eta_\delta^\diamond m_\delta \alpha$.

- Secondly, we derive the probability that the energy level remains the same, denoted by $b_{e,e}^+(\delta)$. This case happens when
 - the channel is busy with no miss detection, no energy successfully harvested and a packet arrival, or
 - the channel is idle with false alarm, and a packet arrival.

Therefore, we have $b_{e,e}^+(\delta) = \eta_\delta^\circ m_\delta^\circ \gamma_\delta^\circ \alpha + \eta_\delta f_\delta \alpha$.

- Thirdly, we derive the probability that the energy level increases by one unit, denoted by $b_{e,e+1}^\circ(\delta)$. This case happens when the channel is busy with no miss detection, energy successfully harvested, and a packet arrival. Therefore, we have $b_{e,e+1}^\circ(\delta) = \eta_\delta^\circ m_\delta^\circ \gamma_\delta \alpha$.

Again, for $e = E$, the energy level cannot increase more than the capacity of the energy storage, and hence $b_{E,E}^+(\delta) = b_{E-1,E-1}^+(\delta) + b_{E-1,E}^+(\delta)$.

For the case when the data queue is full, the transition matrix is obtained as follows: $\mathbf{B}_{Q,Q}(\delta) = \mathbf{B}_{Q-1,Q-1}(\delta) + \mathbf{B}_{Q-1,Q}(\delta)$.

3.2.3 Optimization Formulation

Then we formulate the optimization based on an MDP. Specifically, we will obtain the optimal channel access policy for the secondary user denoted by χ^* to maximize the throughput of the secondary user. The policy is a mapping from the state to the action to be taken by the secondary user. In other words, given the data queue and energy states, the policy will determine the channel to select. The optimization problem is expressed as follows:

$$\max_{\chi} \quad \mathcal{J}(\chi) = \liminf_{t \rightarrow \infty} \frac{1}{t} \sum_{k=1}^t \mathbb{E}(\mathcal{T}(\theta_k, \delta_k)) \quad (3.8)$$

where $\mathcal{J}(\chi)$ is the average throughput of the secondary user and $\mathcal{T}(\theta_k, \delta_k)$ is an immediate throughput function given state $\theta_k \in \Theta$ and action $\delta_k \in \Delta$ at time step k . Again, the state variable is defined as $\theta = (e, q)$. The immediate throughput function is defined as follows:

$$\mathcal{T}(\theta, \delta) = \begin{cases} \eta_\delta f_\delta^\circ \sigma_\delta, & (e \geq W) \text{ and } (q > 0) \\ 0 & \text{otherwise.} \end{cases} \quad (3.9)$$

The secondary user successfully transmits a packet if there is enough energy, the queue is not empty, the selected channel is idle, and there is no false alarm.

Then, we obtain the channel access policy from the optimization by formulating and solving an equivalent linear programming (LP) problem [23]. The LP problem is expressed as follows:

$$\begin{aligned}
 \max_{\zeta(\theta, \delta)} \quad & \sum_{\theta \in \Theta} \sum_{\delta \in \Delta} \zeta(\theta, \delta) \mathcal{T}(\theta, \delta) \\
 \text{s.t.} \quad & \sum_{\delta \in \Delta} \zeta(\theta', \delta) = \sum_{\theta \in \Theta} \sum_{\delta \in \Delta} \zeta(\theta, \delta) P_{\theta, \theta'}(\delta), \quad \theta' \in \Theta \\
 & \sum_{\theta \in \Theta} \sum_{\delta \in \Delta} \zeta(\theta, \delta) = 1, \quad \zeta(\theta, \delta) \geq 0
 \end{aligned} \tag{3.10}$$

where $P_{\theta, \theta'}(\delta)$ denotes the element of matrix $\mathbf{P}(\delta)$ as defined in (3.3) where $\theta = (e, q)$ and $\theta' = (e', q')$. Let the solution of the LP problem be denoted by $\zeta^*(\theta, \delta)$. The channel access policy of the secondary user obtained from the optimization is obtained as follows:

$$\chi^*(\theta, \delta) = \frac{\zeta^*(\theta, \delta)}{\sum_{\delta' \in \Delta} \zeta^*(\theta, \delta')}, \text{ for } \theta \in \Theta. \tag{3.11}$$

3.2.4 Performance Measures

Given that the optimization is feasible, we can obtain the channel access policy of the secondary user. The following performance measures can be obtained.

Average number of packets in the data queue is obtained from

$$\bar{q} = \sum_{\delta \in \Delta} \sum_{q=0}^Q \sum_{e=0}^E q \zeta^*((e, q), \delta). \tag{3.12}$$

Average throughput is obtained from

$$\tau = \sum_{\delta \in \Delta} \sum_{q=1}^Q \sum_{e=W}^E \eta_{\delta} f_{\delta}^{\circ} \sigma_{\delta} \zeta^*((e, q), \delta). \tag{3.13}$$

Average delay can be obtained using Little's law as follows:

$$\bar{d} = \frac{\bar{q}}{\tau} \tag{3.14}$$

where τ is the effective arrival rate which is the same as the throughput.

Note that the optimization formulation presented in this section is based on the assumption that the secondary user does not know the channel status when it selects the channel. Appendix C presents the extended optimization formulation assuming that the secondary user knows the status of all the channels in advance. In other words, the channel status is incorporated as part of the state of the secondary user.

3.3 Learning Algorithm

In Section 3.2, we presented the optimization formulation with known model parameters. However, in practice, the model parameters may not be available to formulate the optimization and obtain the channel access policy for the secondary user. Therefore, in this section, we study the learning algorithm to obtain the channel access policy in an online fashion.

3.3.1 Problem Formulation

Figure 3.2: The learning model for the secondary user in a cognitive radio network.

Fig. 3.2 shows the schematic of the learning algorithm implemented for the secondary user to obtain the channel access policy. The secondary user is composed of the learning algorithm and controller. The learning algorithm is used to update and maintain the channel access policy, while the controller instructs the secondary user to take the action. In the first step, the learning algorithm indicates to the controller based on the current policy to select the channel to access. In the second step, the controller observes the channel status (i.e., idle or busy) to make the decision to

transmit a packet or harvest RF energy based on the observed channel status. Finally, the controller monitors the result (i.e., whether the packet is successfully transmitted or not and whether the RF energy is harvested successfully or not) and reports to the learning algorithm. This result is used by the learning algorithm to update the channel access policy.

In Section 3.2, to derive the transition probability matrix for the secondary user, we need to know environment parameters, such as idle channel probability, miss detection probability, false alarm sensing error probability, successful packet transmission probability, successful RF energy harvesting probability, and so on. However, in practice, it is not easy and even impossible to obtain these probabilities for secondary users. Therefore, we consider a learning algorithm that is developed based on the simulation-based method [123]. The basic idea of the simulation-based method is based on a “simulator” that can simulate an environment by generating parameters (e.g., idle channel probability) given the system process. After the simulation, the simulator yields the results which will be used by the secondary users to obtain channel access policies. Based on received observations (e.g., a packet is successfully transmitted or a unit of energy is successfully harvested), the secondary users can update their local information (i.e., the energy level and the number of packets in the energy queue and data queue, respectively) accordingly.

With the simulation-based learning algorithm, for a given control policy Ψ , the transition probability function $\mathcal{P}$ can be derived from the transition probability of the local state of the secondary users (i.e., a queue state, and energy state) as follows:

$$P(\theta(t+1)|\theta(t), \Psi) = \mathcal{P}_{\text{env}} P\left((e(t+1), q(t+1))|(e(t), q(t)), \Psi\right), \quad (3.15)$$

where $\mathcal{P}_{\text{env}}$ is the probability function of environment parameters that can be generated by the simulator. $\theta(t) \in \Theta$ denotes the state of the secondary user at time slot t . $q(t)$ and $e(t)$ denote the queue state and the energy state at time slot t , respectively. $P\left((e(t+1), q(t+1))|(e(t), q(t)), \Psi\right)$ is the transition probability of the secondary

user and this probability can be derived as follow:

$$P\left((e(t+1), q(t+1)) | (e(t), q(t)), \Psi\right) = \begin{cases} P(e(t), q(t))P(a(t)), & \text{if } e(t+1) = \mathcal{E}^*, \quad q(t+1) = \mathcal{Q}^* \\ 0, & \text{otherwise,} \end{cases} \quad (3.16)$$

where $\mathcal{Q}^* = \min\left([q(t) - q^{tr}(t)]^+ + \lambda(t), Q\right)$, and $\mathcal{E}^* = \min\left([e(t) - e^{tr}(t)]^+ + e^{har}(t), E\right)$. Again, Q is the maximum size of the data queue and E is the maximum size of the energy storage of the secondary user. Here, $q^{tr}(t)$ is the number of packets transmitted at time step t , $\lambda(t)$ is the number of arriving packets, $e^{har}(t)$ is the amount of energy harvested, and $e^{tr}(t)$ is the amount of energy used at time slot t . Furthermore, $[x]^+ = \max(x, 0)$.

To update and maintain the channel access policy for the secondary user, we consider a randomized parameterized policy μ that is well known in the literature [124–126]. Specifically, the randomized policy μ can be parameterized by vector Φ , which is called the randomized parameterized policy μ_Φ . At each time step, when the state of the secondary user is θ , the secondary user will select channel δ (i.e., the action) with the following probability

$$\mu_\Phi(\theta, \delta) = \frac{\exp(\phi_{\theta, \delta})}{\sum_{\delta' \in \Delta} \exp(\phi_{\theta, \delta'})} \quad (3.17)$$

where Φ is the parameter vector of the learning algorithm, defined as

$\Phi = [\cdots \quad \phi_{\theta, \delta} \quad \cdots]^\top$, that is used to help the secondary user to make decisions given the current state. This parameter vector will be adjusted at each time slot after the secondary user observes the results obtained from interactions with the environment. In (3.17), the probability of selecting channel δ is normalized. Furthermore, the parameterized randomized policy $\mu_\Phi(\theta, \delta)$ must not be negative and

$$\sum_{\delta \in \Delta} \mu_\Phi(\theta, \delta) = 1. \quad (3.18)$$

Under the randomized parameterized policy $\mu_\Phi(\theta, \delta)$, the transition probability function will be parameterized as follows:

$$P(\theta, \theta', \Psi(\Phi)) = \sum_{\delta \in \Delta} \mu_\Phi(\theta, \delta) P_\delta(\theta, \theta') \quad (3.19)$$

for all $\theta, \theta' \in \Theta$, where recall that $P_\delta(\theta, \theta')$ is the transition probability from state θ to state θ' when action δ is taken. The objective of the parameterized control policy, which is denoted by $\Psi(\Phi)$, is to maximize the average throughput of the secondary user under the randomized parameterized policy $\mu_\Phi(\theta, \delta)$.

Similarly, we have the parameterized immediate throughput function defined as follows:

$$\mathcal{T}_\Phi(\theta) = \sum_{\delta \in \Delta} \mu_\Phi(\theta, \delta) \mathcal{T}(\theta, \delta). \quad (3.20)$$

Then we need to make some necessary assumptions as follows.

Assumption 3.1 *The Markov chain is aperiodic and there exists a state θ^* which is recurrent for every of such Markov chain.*

Assumption 3.2 *For every state $\theta, \theta' \in \Theta$, the transition probability function $P_\Phi(\theta, \theta')$ and the immediate throughput function $\mathcal{T}_\Phi(\theta)$ are bounded, twice differentiable, and have bounded first and second derivatives.*

Assumption 3.1 implies that the system that we consider has a Markov property. Assumption 3.2 ensures that the transition probability function and the immediate reward function depend “smoothly” on ϕ . This assumption is important when we apply gradient methods for adjusting ϕ .

Then, we can define the parameterized average throughput (i.e., the throughput under the parameter vector Φ) by

$$\psi(\Phi) = \lim_{t \rightarrow \infty} \frac{1}{t} \mathbb{E}_\Phi \left[\sum_{k=0}^t \mathcal{T}_\Phi(\theta_k) \right] \quad (3.21)$$

where θ_k is the state of the secondary user at time step k . $\mathbb{E}_\Phi[\cdot]$ is the expectation of the throughput. Note that this average throughput is defined similar to $\mathcal{J}_T(\cdot)$ given in (3.8). However, $\psi(\Phi)$ is the objective function for the learning algorithm of the secondary user.

Under Assumption 3.1, the average throughput $\psi(\Phi)$ is well defined for every Φ , and does not depend on the initial state θ_0 . Moreover, we have the following balance

equations

$$\begin{aligned} \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) P(\theta, \theta', \Psi(\Phi)) &= \pi_{\Phi}(\theta'), \text{ for } \theta' \in \Theta \\ \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) &= 1 \end{aligned} \quad (3.22)$$

where $\pi_{\Phi}(\theta)$ is the steady-state probability of state θ under the parameter vector Φ . These balance equations have a unique solution defined as a vector $\Pi_{\Phi} = [\cdots \pi_{\Phi}(\theta) \cdots]^{\top}$. Then, the average throughput can be expressed as follows:

$$\psi(\Phi) = \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \mathcal{J}_{\Phi}(\theta). \quad (3.23)$$

3.3.2 Policy Gradient Method

We define the differential throughput $d(\theta, \Phi)$ at state θ by

$$d(\theta, \Phi) = \mathbb{E}_{\Phi} \left[\sum_{k=0}^{T-1} (\mathcal{J}_{\Phi}(\theta_k) - \psi(\Phi)) \mid \theta_0 = \theta \right] \quad (3.24)$$

where $T = \min\{k > 0 \mid \theta_k = \theta^*\}$ is the first future time that state θ^* is visited. Here, we need to note that, the main aim of defining the differential throughput $d(\theta, \Phi)$ is to represent the relation between the average throughput and the immediate throughput at state θ , instead of the recurrent state θ^* . Additionally, under Assumption 3.1, the differential throughput $d(\theta, \Phi)$ is a unique solution of the following Bellman equation:

$$d(\theta, \Phi) = \mathcal{J}_{\Phi}(\theta) - \psi(\Phi) + \sum_{\theta' \in \Theta} P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi) \quad (3.25)$$

for all $\theta \in \Theta$. Then, by using the differential throughput $d(\theta, \Phi)$, we can obtain the gradient of the average throughput $\psi(\Phi)$ with respect to the parameter vector Φ as stated in Proposition 3.1.

Proposition 3.1 *Let Assumption 3.1 and Assumption 3.2 hold, then*

$$\nabla \psi(\Phi) = \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \left(\nabla \mathcal{J}_{\Phi}(\theta) + \sum_{\theta' \in \Theta} \nabla P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi) \right). \quad (3.26)$$

Proposition 3.1 represents the gradient of the average throughput $\psi(\Phi)$ and the proof of Proposition 3.1 is provided in Appendix A.

3.3.3 An Idealized Gradient Algorithm

Using Proposition 3.1, we can formulate the idealized gradient algorithm based on the form proposed in [127] given as follows:

$$\Phi_{k+1} = \Phi_k + \rho_k \nabla \psi(\Phi_k) \quad (3.27)$$

where ρ_k is a step size. In this policy gradient method, we start with an initial parameter vector Φ_0 . The parameter vector Φ is updated at every time step based on (3.27). Under a suitable condition (from Assumption 3.3) of a step size ρ_k and Assumption 3.2, it was proved that $\lim_{k \rightarrow \infty} \nabla \psi(\Phi_k) = 0$ and thus $\psi(\Phi_k)$ converges [127].

Assumption 3.3 *The step size ρ_m is deterministic, nonnegative and satisfies the following condition,*

$$\sum_{m=1}^{\infty} \rho_m = \infty, \text{ and } \sum_{m=1}^{\infty} (\rho_m)^2 < \infty. \quad (3.28)$$

In other words, the value of the step size approaches zero when the time step goes to infinity.

3.3.4 Learning Algorithm

The idealized gradient algorithm can maximize the average throughput $\psi(\Phi_k)$, if we can calculate the gradient of the function $\psi(\Phi_k)$ with respect to Φ at each time step. However, if the system has a large state space, it is impossible to compute the exact gradient of the average throughput $\psi(\Phi_k)$. Therefore, we alternatively consider an approach that can estimate the gradient $\psi(\Phi_k)$ and update parameter vector Φ accordingly in an online fashion.

From (3.18), we have $\sum_{\delta \in \Delta} \mu_{\Phi}(\theta, \delta) = 1$, so we can derive that $\sum_{\delta \in \Delta} \nabla \mu_{\Phi}(\theta, \delta) = 0$ for every Φ . From (3.20), we have

$$\nabla \mathcal{J}_{\Phi}(\theta) = \sum_{\delta \in \Delta} \nabla \mu_{\Phi}(\theta, \delta) \mathcal{T}(\theta, \delta) \quad (3.29)$$

$$= \sum_{\delta \in \Delta} \nabla \mu_{\Phi}(\theta, \delta) (\mathcal{T}(\theta, \delta) - \psi(\Phi)) \quad (3.30)$$

since

$$\sum_{\delta \in \Delta} \nabla \mu_{\Phi}(\theta, \delta) = 0. \quad (3.31)$$

Moreover, we have

$$\sum_{\theta' \in \Theta} \nabla P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi) = \sum_{\theta' \in \Theta} \sum_{\delta \in \Delta} \nabla \mu_{\Phi}(\theta, \delta) P_{\delta}(\theta, \theta') d(\theta', \Phi) \quad (3.32)$$

for all $\theta \in \Theta$.

Therefore, along with Proposition 3.1, we can derive the gradient of $\psi(\Phi)$ as follows:

$$\nabla \psi(\Phi) = \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \left(\nabla \mathcal{J}_{\Phi}(\theta) + \sum_{\theta' \in \Theta} \nabla P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi) \right) \quad (3.33)$$

$$\begin{aligned} &= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \left(\sum_{\delta \in \Delta} \nabla \mu_{\Phi}(\theta, \delta) (\mathcal{J}(\theta, \delta) - \psi(\Phi)) \right. \\ &\quad \left. + \sum_{\theta' \in \Theta} \sum_{\delta \in \Delta} \nabla \mu_{\Phi}(\theta, \delta) P_{\delta}(\theta, \theta') d(\theta', \Phi) \right) \end{aligned} \quad (3.34)$$

$$\begin{aligned} &= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \sum_{\delta \in \Delta} \nabla \mu_{\Phi}(\theta, \delta) \left((\mathcal{J}(\theta, \delta) - \psi(\Phi)) \right. \\ &\quad \left. + \sum_{\theta' \in \Theta} P_{\delta}(\theta, \theta') d(\theta', \Phi) \right) \end{aligned} \quad (3.35)$$

$$= \sum_{\theta \in \Theta} \sum_{\delta \in \Delta} \pi_{\Phi}(\theta) \nabla \mu_{\Phi}(\theta, \delta) q_{\Phi}(\theta, \delta), \quad (3.36)$$

where

$$\begin{aligned} q_{\Phi}(\theta, \delta) &= \left(\mathcal{J}(\theta, \delta) - \psi(\Phi) \right) + \sum_{\theta' \in \Theta} P_{\delta}(\theta, \theta') d(\theta', \Phi) \\ &= \mathbb{E}_{\Phi} \left[\sum_{k=0}^{T-1} (\mathcal{J}(\theta_k, \delta_k) - \psi(\Phi)) | \theta_0 = \theta, \delta_0 = \delta \right]. \end{aligned} \quad (3.37)$$

Here θ_k and δ_k are the state and action at time step k , respectively, and $T = \min\{k > 0 | \theta_k = \theta^*\}$ is the first future time that the current state θ^* is visited. $q_{\Phi}(\theta, \delta)$ can be interpreted as the differential throughput if action δ is taken based on policy μ_{Φ} at state θ . Then, we present Algorithm 1 that updates the parameter vector Φ at the visit to the recurrent state θ^* .

In Algorithm 1, κ is a positive constant and ρ_m is a step size that satisfies Assumption 3.3. The term $F_m(\Phi_m, \tilde{\psi}_m)$ represents the estimated gradient of the average throughput and it is calculated by the cumulative sum of the total estimated gradient

Algorithm 1 Algorithm to update parameter vector Φ at the visit to the recurrent state θ^*

At the time step k_{m+1} of the $(m+1)$ th visit to state θ^* , we update the parameter vector Φ and the estimated average throughput $\tilde{\psi}$ as follows:

$$\Phi_{m+1} = \Phi_m + \rho_m F_m(\Phi_m, \tilde{\psi}_m), \quad (3.38)$$

$$\tilde{\psi}_{m+1} = \tilde{\psi}_m + \kappa \rho_m \sum_{k'=k_m}^{k_{m+1}-1} \left(\mathcal{T}(\theta_{k'}, \delta_{k'}) - \tilde{\psi}_m \right) \quad (3.39)$$

where

$$F_m(\Phi_m, \tilde{\psi}_m) = \sum_{k'=k_m}^{k_{m+1}-1} \tilde{q}_{\Phi_m}(\theta_{k'}, \delta_{k'}) \frac{\nabla \mu_{\Phi_m}(\theta_{k'}, \delta_{k'})}{\mu_{\Phi_m}(\theta_{k'}, \delta_{k'})}, \quad (3.40)$$

$$\tilde{q}_{\Phi_m}(\theta_{k'}, \delta_{k'}) = \sum_{k=k'}^{k_{m+1}-1} \left(\mathcal{T}(\theta_k, \delta_k) - \tilde{\psi}_m \right). \quad (3.41)$$

of the average throughput between two successive visits (i.e., m th and $(m+1)$ th visits) to the recurrent state θ^* . Furthermore, $\nabla \mu_{\Phi_m}(\theta_{k'}, \delta_{k'})$ expresses the gradient of the randomized parameterized policy function that is provided in (3.17). Algorithm 1 enables us to update the parameter vector Φ and the estimated average throughput $\tilde{\psi}$ iteratively. Accordingly, we derive the following convergence result for Algorithm 1.

Proposition 3.2 *Let Assumption 3.1 and Assumption 3.2 hold, and let $(\Phi_0, \Phi_1, \dots, \Phi_\infty)$ be the sequence of the parameter vectors generated by Algorithm 1 with a suitable step size ρ satisfying Assumption 3.3, then $\psi(\Phi_m)$ converges and*

$$\lim_{m \rightarrow \infty} \nabla \psi(\Phi_m) = 0, \quad (3.42)$$

with probability one.

The proof of Proposition 3.2 is given in Appendix B.

In Algorithm 1, to update the value of the parameter vector Φ at the next visit to the state θ^* , we need to store all values of $\tilde{q}_{\Phi_m}(\theta_n, \delta_n)$ and $\frac{\nabla \mu_{\Phi_m}(\theta_n, \delta_n)}{\mu_{\Phi_m}(\theta_n, \delta_n)}$ between two successive visits. However, this method could result in slow processing. Therefore, we modify Algorithm 1 to improve the efficiency. First, we rewrite $F_m(\Phi_m, \tilde{\psi}_m)$ as

follows:

$$\begin{aligned}
 F_m(\Phi_m, \tilde{\psi}_m) &= \sum_{k'=k_m}^{k_{m+1}-1} \tilde{q}_{\Phi_m}(\theta_{k'}, \delta_{k'}) \frac{\nabla \mu_{\Phi_m}(\theta_{k'}, \delta_{k'})}{\mu_{\Phi_m}(\theta_{k'}, \delta_{k'})} \\
 &= \sum_{k'=k_m}^{k_{m+1}-1} \frac{\nabla \mu_{\Phi_m}(\theta_{k'}, \delta_{k'})}{\mu_{\Phi_m}(\theta_{k'}, \delta_{k'})} \sum_{k=k'}^{k_{m+1}-1} (\mathcal{T}(\theta_k, \delta_k) - \tilde{\psi}_m) \\
 &= \sum_{k'=k_m}^{k_{m+1}-1} (\mathcal{T}(\theta_k, \delta_k) - \tilde{\psi}_m) z_{k+1},
 \end{aligned} \tag{3.43}$$

where

$$z_{k+1} = \begin{cases} \frac{\nabla \mu_{\Phi_m}(\theta_k, \delta_k)}{\mu_{\Phi_m}(\theta_k, \delta_k)}, & \text{if } k = k_m, \\ z_k + \frac{\nabla \mu_{\Phi_m}(\theta_k, \delta_k)}{\mu_{\Phi_m}(\theta_k, \delta_k)}, & k = k_m + 1, \dots, k_{m+1} - 1. \end{cases} \tag{3.44}$$

The detail of the algorithm can be expressed as in Algorithm 2, where κ is a positive constant and ρ_k is the step size of the algorithm.

Algorithm 2 Algorithm to update Φ at every time step

At time step k , the state is θ_k , and the values of Φ_k , z_k , and $\tilde{\psi}(\Phi_k)$ are available from the previous iteration. We update z_k , Φ , and $\tilde{\psi}$ according to:

$$z_{k+1} = \begin{cases} \frac{\nabla \mu_{\Phi_k}(\theta_k, \delta_k)}{\mu_{\Phi_k}(\theta_k, \delta_k)}, & \text{if } \theta_k = \theta^* \\ z_k + \frac{\nabla \mu_{\Phi_k}(\theta_k, \delta_k)}{\mu_{\Phi_k}(\theta_k, \delta_k)}, & \text{otherwise,} \end{cases} \tag{3.45}$$

$$\Phi_{k+1} = \Phi_k + \rho_k (\mathcal{T}(\theta_k, \delta_k) - \tilde{\psi}_k) z_{k+1}, \tag{3.46}$$

$$\tilde{\psi}_{k+1} = \tilde{\psi}_k + \kappa \rho_k (\mathcal{T}(\theta_k, \delta_k) - \tilde{\psi}_k). \tag{3.47}$$

The computation complexity of the Algorithm 2 is provided in Appendix F of the thesis.

3.4 Performance Evaluation

3.4.1 Parameter Setting

We consider the secondary user whose data queue and energy storage have the maximum sizes of 10 packets and 10 units of energy, respectively. The secondary user requires 1 unit of energy for packet transmission. The packet arrival probability is 0.5. There are two channels licensed to primary users 1 and 2. Unless otherwise stated, the probabilities that the channels 1 and 2 will be idle are 0.1 and 0.9, respectively. The probability of successful packet transmission on both the channels is

0.95. The probabilities of successful RF energy harvesting with one unit of energy on channels 1 and 2 are 0.95 and 0.70, respectively.

For the learning algorithm, we use the following parameters for performance evaluation. We assume that at the beginning of Algorithm 2, the secondary will start with a randomized policy, where the secondary user will select channels 1 and 2 with the same probability of 0.5. We set the initial value of $\rho = 0.0005$ and it will be updated after every 18,000 iterations as follows: $\rho_{k+1} = 0.9\rho_k$. We also set $\kappa = 0.01$.

We evaluate and compare the performance of the secondary user adopting the following channel access policies: complete information, incomplete information, learning, and random policy. The policy of the complete information case is obtained by solving the optimization given in Appendix C. This policy yields the upper-bound performance as the secondary user has complete information about channel status. The policy of the incomplete information case is obtained from the optimization given in Section 3.2. In this case, the secondary user does not know the channel status when it selects the channel to access. The policy of the learning algorithm is obtained from executing Algorithm 2 given in Section 3.3. Finally, we consider the random policy in which the secondary user selects channels 1 and 2 randomly with the same probability of 0.5.

3.4.2 Numerical Results

3.4.2.1 Policies from MDP-based optimization formulations

We consider the policies obtained from the optimization with complete and incomplete information about the channel status, respectively. Fig. 3.3(a) shows the channel access policy of the incomplete information case. The z-axis represents the probability of choosing channel 1. When this probability is high, the secondary user is likely to select channel 1. By contrast, if the probability is small, the secondary user is likely to select channel 2. We observe that the secondary user selects channel 1 or 2 depending on the queue and energy states. In particular, the secondary user selects channel 1 when the energy level is low and the number of packets in the data queue is small. This is due to the fact that channel 1 is more likely to be busy (i.e., available for

(a) Incomplete information case.

Complete information: Channel 1 is idle and channel 2 is busy

(b) Complete information when channel 1 is idle and channel 2 is busy.

Complete information: Channel 1 is busy and channel 2 is idle

(c) Complete information when channel 1 is busy and channel 2 is idle.

Figure 3.3: Policy of the secondary user.

RF energy harvesting). By contrast, the secondary user selects channel 2 when the number of packets in the data queue is large and the energy level is high. This is from the fact that channel 2 has higher chance to be idle, which is favorable for the packet transmission by the secondary user. Note that the channel access policy favors the secondary user to select channel 1 more than channel 2 since the probability of successful RF energy harvesting of channel 2 is lower than that of channel 1.

Fig. 3.3(b) and (c) show the policies obtained with complete information, when channel 1 is idle and channel 2 is busy, and when channel 1 is busy and channel 2 is idle, respectively. From Fig. 3.3(b), we observe that when channel 1 is idle and channel 2 is busy, the secondary user almost always selects channel 1, except when the data queue or energy storage is empty. This is for the secondary user to transmit its packet. However, from Fig. 3.3(c), when channel 1 is busy and channel 2 is idle, the secondary user selects channel 1 only when the energy level is low and the number of packets in the data queue is small. This result is similar to that of the incomplete information case. Note that we omit showing the cases when both channels 1 and 2 are idle (or busy) as it is clear for the secondary user to select the channel with a higher probabilities of successful packet transmission (or higher probability of successful RF energy harvesting).

3.4.2.2 Convergence of the learning algorithm

Fig. 3.4 shows the convergence of the proposed learning algorithm in terms of the average throughput when the number of channels is varied from two channels to four channels. As shown in Fig. 3.4, when the number of channels is 2, the learning algorithm converges to the average throughput of 0.45 relatively fast within the first 10^6 iterations. When we increase the number of channels, the convergence rate of the learning algorithm is slightly slower and the average throughput of the secondary user is slightly lower. Specifically, when the number of channels is three and four, the average throughput converges to approximately 0.4491 and 0.4484, respectively, and the learning algorithm converges within from 2×10^6 to around 3×10^6 iterations. The reason is that with more number of channels, the secondary user has more actions to

Figure 3.4: The convergence of the learning algorithm when the number of channels is varied.

learn. Therefore, the secondary user cannot use an optimal policy during exploring choices of actions, leading to poorer performance. It is also worth noting that, when we vary the idle channel probability of channel 3 and channel 4 within the range from 0.1 to 0.9, the average throughput obtained for both cases will not be changed, i.e., 0.4491 and 0.4484, respectively. The reason is, when the secondary user wants to harvest energy, it will select the channel that has the highest busy channel probability (i.e., 0.1 - the idle channel probability of channel 1) and when the secondary user wants to transmit data, it will select the channel that has the highest idle channel probability (i.e., 0.9 - the idle channel probability of channel 2).

Different from conventional learning processes, the learning algorithm in our model is composed of two processes. In the first learning process, the algorithm needs to determine whether the secondary user should transmit data or harvest energy at a certain state. After that, in the second learning process, the algorithm needs to choose which channel to sense. For example, if the secondary user has large number of packets in the data queue waiting for transmission, the secondary user should sense the channel that has the highest idle channel probability. In fact, the second learning process converges quickly since the number of channels is not too many and it does not impact much to the convergence rate as shown in Fig. 3.4. However, the first learning process has a significant impact on the convergence rate due to the state space of the

secondary user. For example, in our model, the number of states is $11^2 = 121$ and this makes the algorithm converge relatively slow. If we can find the structure of the queues of the secondary user to reduce to the state space, we might speed up for the convergence rate for the learning algorithm significantly. This could be our future work.

(a)

(b)

Figure 3.5: (a) Policy at 10^3 and (b) 10^7 iterations of the learning algorithm.

Figs. 3.5(a) and (b) show the policies from the learning algorithm at 10^3 and 10^7 iterations, respectively. We observe that the former is initialized to be random, the

learning algorithm is able to update the policy so that it is similar to that obtained from the MDP-based optimization (i.e., in the incomplete information case). Therefore, if the secondary user does not have the model parameters to formulate and solve the optimization, the learning algorithm is the only viable solution to obtain the channel access policy.

3.4.2.3 Throughput performance

In this section, we examine the throughput performance of the secondary user obtained from different policies under variation of the parameters. Fig. 3.6(a) shows the

Figure 3.6: The average throughput of the system when (a) the packet arrival probability is varied and (b) the idle channel probability for channel 1 is varied.

average throughput of the system of algorithms when the packet arrival probability is varied. Under the small packet arrival probability (i.e., less than 0.4), all the policies

yield almost the same throughput. This is due to the fact that the secondary user can harvest enough RF energy and has sufficient opportunity to transmit its packet. However, at the high packet arrival probability, Fig. 3.6(a) shows that the policy from the optimization with complete information yields the highest throughput, followed by the policy from the optimization with incomplete information. This is due to the fact that the policy in the complete information case can fully exploit the channel status to achieve the optimal performance. However, in the incomplete information case, the secondary user is still able to optimize its channel access, which yields higher throughput than that of the random policy. We observe that the learning algorithm yields the throughput close to that of the policies obtained from the optimizations, and it is much higher than that of the random policy. When the arrival probability is higher than 0.5 the system reaches to the saturated state and thus the average throughput is not increased.

Fig. 3.6(b) considers the case when the idle probability of channel 1 is varied and the packet arrival probability is fixed at 0.5. Note that the packet arrival probability will be set at 0.5 for the rest of this section when we compare the average throughput of the secondary user under different input parameters. As the idle probability of channel 1 increases (i.e., becomes less busy), the throughput first increases, since the secondary user has more opportunity to transmit its packets. However, at a certain point, the throughput decreases. This decrease is due to the fact that channel 1 is mostly idle and the secondary user cannot harvest much RF energy. Therefore, there is not enough energy to transmit the packets, and hence the throughput decreases.

Fig. 3.7(a) shows the throughput of the secondary user under different probability of successful RF energy harvesting from channel 1. When the secondary user has higher chance to successfully harvest RF energy from one of the channels, the performance improves. Again, the secondary user with complete information achieves the highest throughput, and the channel access policy with incomplete information still yields the throughput higher than that of the random policy. The policy from the learning algorithm still achieves the throughput close to those from the optimizations. Similarly, when the successful data transmission probability increases, the

(a)

(b)

Figure 3.7: Throughput under different successful harvesting probabilities for channel 1 (a) and successful data transmission probabilities (b).

(a)

(b)

Figure 3.8: The average throughput of the system when (a) the false alarm probability and (b) the miss detection probability is varied.

system throughput is also increased and the average throughput obtained by the proposed learning algorithm approached to the incomplete information and the complete information case as indicated in Fig. 3.7(b).

We then consider the impacts of the miss detection and false alarm sensing error to the average throughput of the secondary user in Fig. 3.8. When the false alarm probability increases, the average throughput will decrease for all algorithms. Fur-

thermore, the average throughput of the proposed learning algorithm is close to that of the incomplete information and the complete information case, and it is still higher than the average throughput of the random policy. However, when the miss detection probability is varied, the incomplete information case and the complete information case are not always the best solutions. Specifically, when the miss detection probability increases from 0.1 to 0.2, the average throughput obtained by the incomplete information and complete information cases is greater than that of the learning algorithm. However, when the miss detection probability increases from 0.2 to 0.9, the average throughput of the incomplete information and complete information cases will be lower than that of the learning algorithm. The reason is that when the secondary user wants to harvest energy, it will sense channel 1 (in the case of the incomplete information and complete information) since channel 1 has higher busy channel probability providing better opportunity for the secondary user to harvest energy. However, when the miss detection probability is high, the sensing result is also highly incorrect. In particular, the channel is busy but the sensing result is idle and thus the secondary user will transmit data on the sensed channel. As a result, collision will occur and thus instead of harvesting energy, the secondary will lose energy due to accessing the wrongly sensed channel. The energy level in the energy storage will be low. Furthermore, for the policies of incomplete and complete cases shown in Figs. 3.3(a) and (c), respectively, when the energy level in the energy storage is low, the secondary user will select channel 1 to sense. This process will continue and lead to the reduced average throughput of the system as shown in Fig. 3.8(b).

Different from the incomplete information and the complete information cases, the learning algorithm is able to adapt with the different kinds of the environment and the average throughput obtained by the learning algorithm is always better than the random policy as shown in Fig. 3.8(b).

3.4.2.4 Other performance

We evaluate the average number of packets in the data queue and the packet blocking probability of the secondary user (Figs. 3.9(a) and (b), respectively). As the

Figure 3.9: (a) The average number of packets in the data queue and (b) the packet blocking probability.

packet arrival probability increases, the average number of packets and the blocking probability increase. If the packet arrival probability is too large, the data queue of the secondary user will be mostly full. Therefore, the secondary user cannot accept incoming packet, resulting in more blocked packets.

3.5 Summary

We have considered a cognitive radio network in which the secondary user is equipped with the RF energy harvesting capability. The secondary user can transmit a packet if the selected channel is not occupied by the primary user. Alternatively, the secondary user can harvest RF energy from the primary user's transmission if the selected channel is occupied. In this network, the secondary user has to perform channel access

to select the best channel given its state. We have first presented the optimization formulation based on Markov decision process to obtain the channel access policy. This formulation does not need the secondary user to know the current channel status. However, the optimization still requires the model parameters, which may not be available in practice. Therefore, we have adopted the online learning algorithm which is able to interact with the environment and take appropriate actions. Additionally, we have considered the optimization in the case where the channel status is known by the secondary user. This complete information case can yield the upper-bound performance for a benchmarking purpose. Finally, we have performed the performance evaluation, which shows the success of using learning algorithm in terms of efficiency and convergence.

Table 3.1: Summary of key notations of Chapter 3

Notation	Physical meaning
N	Number of primary channels
γ_n	The probability that a secondary user harvests one unit of RF energy successfully from channel n
E	The maximum energy queue size
W	The energy required for data transmission in a time slot
δ_n	The probability of successful packet transmission on channel n
α	The probability of a packet arrival at the secondary user
Q	The maximum data queue size
m_n	The miss detection probability
f_n	The false alarm probability of the secondary user
σ_n	The probability of successful packet transmission on channel n
$\mathbf{C}_n$	The transition probability matrix of channel n
η_n	The probability that channel n is idle
$\mathcal{E}$	The energy level of the energy storage
$\mathcal{Q}$	The number of packets in the data queue
θ	The state of the state space Θ
θ^*	The revisit state
T	The first future time that state θ^* is revisited
δ	The action of the action space Δ
$\mathbf{P}(\delta)$	The transition probability matrix of the system given action δ
e, q	The energy state and queue state
$\bar{q}, \bar{d}$	The average number of packets and delay
ζ	The solution of the linear programming problem
χ^*	The optimal channel access policy
$\mathcal{J}$	The average throughput
$\mathcal{T}$	The immediate throughput
ϕ	The steady state probability
μ	The parameterized randomized policy
Φ	The parameter vector
ψ	The average throughput (for the learning algorithm)
$\tilde{\psi}$	The estimated average throughput
Ψ	The control policy
π	The steady state probability
$d(\theta, \Phi)$	The differential throughput at state θ under the parameter vector Φ
$q_\Phi(\theta, \delta)$	The differential throughput at state θ under the parameter vector Φ and action δ is taken based on the policy μ_Φ
ρ	The step size
κ	The positive constant
F_m	The estimated gradient of the average throughput at the m -th visit

Chapter 4

Performance Optimization for Cooperative Multiuser Cognitive Radio Networks with RF Energy Harvesting Capability

For the RF energy harvesting cognitive radio network under consideration, there are four different cases depending on the number of secondary users and available channels in the network as follows:

- *Single secondary user and single channel (SUSC) case:* This is the simplest case where a secondary user will always sense a channel to find a chance to transmit data or harvest RF energy. In this case, the problem is straightforward. If the sensed channel is busy due to the transmission of the primary user, the secondary user has an opportunity to harvest and store RF energy in its energy storage. By contrast, if the sensed channel is idle, then the secondary user has a chance to transmit its packet.
- *Single secondary user and multiple channels (SUMC) case:* There is only one secondary user, but there are multiple channels. In practice, the secondary user cannot sense all channels at the same time (e.g., due to hardware limitation) and thus the secondary user faces the problem of selecting a channel to sense. Different channels may have different idle channel probability and primary user's signal strength. Therefore, the secondary user has to make a decision of channel selection to maximize its performance (e.g., maximum throughput or minimum

average number of packets waiting in the data queue). This case was studied in Chapter 3 with both online and offline methods.

- *Multiple secondary users and single channel (MUSC) case:* In this case, we have multiple secondary users sharing one common channel. If the channel is busy, then all secondary users can harvest RF energy from the busy channel. However, if the secondary users sense the channel to be idle, then the secondary users face the multiple access channel (i.e., to transmit a packet or not). To avoid the undesirable transmission collision caused when there is more than one secondary user transmitting data simultaneously, which degrades spectrum utilization and wastes energy, we can use multiple access schemes such as backoff algorithm [128] or time division multiple access (TDMA) technique [129]. More solutions for the channel multiple-access problem in wireless networks can be found in [12, 129].
- *Multiple secondary users and multiple channels (MUMC) case:* This is the most complex case and it is also the main focus of this chapter. In this case, we face not only the channel selection problem (as in the SUMC case), but also the multiple access channel problem (as in the MUSC case) at the same time. Additionally, along with incomplete information from the environment and multiple decision makers from secondary users, finding the optimal control policy for secondary users is challenging.

In this chapter, we consider the RF energy harvesting cognitive radio network with multiple secondary users and multiple available primary channels. In this network, secondary users are able to harvest RF energy from the busy channel and this harvested energy will be used to transmit data over an idle channel. Furthermore, we consider the case in which the secondary users cooperate in order to maximize the network performance in terms of the average throughput of the network. The secondary users are assumed to have no prior information of primary channels and they also cannot sense all channels simultaneously. Consequently, at decision epochs,

the secondary users need to select one of channels to sense such that the average throughput for the system is maximized. To address the cooperative optimization problem among the secondary users, we propose two approaches as follows. In the first approach, we consider the case where the secondary users are cooperative in a round-robin scheduling and each secondary user is equipped with an online learning algorithm in order to help secondary users explore the environment (i.e., primary channels) and then to make optimal decisions. In the second approach, secondary users are assumed to cooperate in a decentralized fashion without coordination. In this case, we formulate the cooperative optimal problem for secondary users as the decentralized partially observable Markov decision process model [38] with the aim of obtaining the decentralized optimal channel access policies for secondary users. Moreover, to deal with the curse of model and the curse of dimensionality (caused by multiple decision makers and incomplete information) for the decentralized system, we study a decentralized learning algorithm developed based on the policy gradient and the Lagrange multiplier methods. The learning algorithms can work without prior environment parameters (e.g., the idle channel probability, the channel sensing error probability, and the successful packet transmission probability). Extensive performance evaluation and comparisons are performed to show the efficiency as well as the convergence of the learning algorithms.

The rest of the chapter is organized as follows. Section 4.1 introduces the system model together with the assumptions. Section 4.2 studies a learning solution based on the TDMA technique and Section 4.3 investigates a decentralized learning algorithm. Experiments are performed and results are analyzed in Section 4.4. We then summarize the chapter in Section 4.5.

4.1 System Model and Assumptions

We consider a RF energy harvesting cognitive radio network with multiple secondary users and multiple channels as shown in Fig. 4.1. In this system model, there are N secondary users sharing M channels that are allocated to primary users. The primary

Figure 4.1: System model.

users use the channels to transmit data in a time slot basis. The secondary users are assumed to be equipped with a data queue and an energy storage. The data queue is used to store data generated or collected from other sensors and the maximum size of the data queue for secondary user n , for $n \in \{1, \dots, N\}$ is denoted by $\mathcal{D}_n$. The energy storage is a battery which can store RF energy harvested from radio signal¹ and the maximum capacity of the energy storage of secondary user n is denoted by $\mathcal{E}_n$.

In each time slot, the probability of a packet arriving at the data queue of secondary user n is denoted by λ_n . The secondary user utilizes the energy in its energy storage for packet transmission. The amount of harvested RF energy by the secondary user depends on the channel status. Specifically, if the secondary user n senses the channel m and the channel m is busy, then secondary user n can harvest a unit of energy with a certain probability (i.e., a successful RF energy harvesting probability). However, if the channel m is idle, then secondary user n cannot harvest any energy. Conversely, the secondary user n can transmit a packet on the channel m or do nothing. If the

¹The design and implementation of circuits, antenna, and process of RF energy harvesting are beyond the scope of this thesis. The details may refer to [121, 130] for more details.

receiver of the secondary user receives the transmitted packet successfully (i.e., no collision and no error), the receiver will send back an ACK to the secondary user. Upon receiving the ACK, the secondary user then removes this packet from its data queue. By contrast, if the packet transmission is unsuccessful, the packet is still in the data queue and the secondary user has to re-transmit the packet later. The unsuccessful packet transmission happens due to the collision with primary user's transmission or other secondary users' transmissions, or channel error.

The channel sensing can be in error. The false alarm sensing error happens when a channel is idle, but the secondary user senses it to be busy. The miss detection happens when a channel is busy, but the secondary user senses it to be idle. The probabilities of such sensing error events are called false alarm probability and miss detection probability.

In the following, we consider two multiple access schemes for the secondary users. Firstly, the secondary users can access a channel based on a TDMA fashion in which a round-robin scheduling is applied. However, this scheme requires coordination among secondary users, e.g., to determine the sequence of packet transmissions. Alternatively, the secondary users can perform random access in a decentralized fashion. However, in this scheme, collision among secondary users can happen. We derive the optimal channel access policies and perform performance comparisons for both schemes.

4.2 A TDMA Learning Algorithm for RF Energy Harvesting Cognitive Radio Networks

We first consider the TDMA method for multiple accesses of secondary users. In particular, the secondary users are scheduled for channel access in a round-robin fashion. We propose a learning algorithm, that is similar to the learning algorithm studied in Chapter 3, for the secondary users to make optimal decisions to adapt with environment conditions.

4.2.1 Problem Formulation

We define the state space of the secondary user n as follows²:

$$\mathcal{S} = \left\{ (e, d, \vartheta); e \in \{0, 1, \dots, \mathcal{E}\}, d \in \{0, 1, \dots, \mathcal{D}\}, \vartheta \in \{0, 1\} \right\}, \quad (4.1)$$

where e , d , and ϑ represent the energy level of the energy storage, the number of packets in the data queue, and the time schedule of the secondary user n , respectively. $\mathcal{E}$ is the maximum capacity of the energy storage and $\mathcal{D}$ is the maximum data queue size. For the time schedule of each secondary user, it has two values, i.e., 0 and 1. For $\vartheta = 0$ and $\vartheta = 1$, the secondary user n is not scheduled and is scheduled for transmitting a packet in the current time slot, respectively. The state of secondary user n is then defined as a composite variable $s = (e, d, \vartheta) \in \mathcal{S}$.

$\mathcal{A}$ is the action space of secondary user n , which is a set of available channels for secondary users to select from. Then, at each time slot, each secondary user has to make a decision $a \in \mathcal{A} = \{0, 1, \dots, M\}$ to select one of the channels to sense. $a = 0$ means that the secondary user n does not select any channel.

We define the immediate reward function for each secondary user as follows:

$$\mathcal{R}(s, a) = \begin{cases} 1, & \text{if a packet is successfully transmitted} \\ 0 & \text{otherwise.} \end{cases} \quad (4.2)$$

We then consider a randomized parameterized policy [124–126]. Under the randomized parameterized policy, when secondary user n is at state s , the secondary user will select action a with the probability $\mu_{\Theta}(s, a)$ given as follows:

$$\mu_{\Theta}(s, a) = \frac{\exp(\theta_{s,a})}{\sum_{a_i \in \mathcal{A}} \exp(\theta_{s,a_i})}, \quad (4.3)$$

where $\Theta = \{\theta_{s,a} \in \mathbb{R}\}$ is the parameter vector of secondary user n at state s . Additionally, every $\mu_{\Theta}(s, a)$ must not be negative and $\sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) = 1$.

Under the randomized parameterized policy $\mu_{\Theta}(s, a)$, the transition probability function will be parameterized as follows:

$$\mathbf{p}(s'|s, \Psi(\Theta)) = \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) \mathbf{p}(s'|s, a), \quad (4.4)$$

²The TDMA technique is used in this case to avoid the collision among secondary users, and thus, the problem formulation for every secondary user is the same. Therefore, we omit the indicator of secondary users in this section for brevity of presentation.

for all $s, s' \in \mathcal{S}$, where $\mathbf{p}(s'|s, a)$ is the transition probability from state s to state s' when action a is taken. Similarly, we have the parameterized immediate throughput function defined as follows:

$$\mathcal{T}(s, \Theta) = \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) \mathcal{T}(s, a). \quad (4.5)$$

The objective of the policy is to maximize the average throughput of the secondary user under the randomized parameterized policy $\mu_{\Theta}(s, a)$, which is denoted by $\Psi(\Theta)$.

Similar to the learning algorithm in Chapter 3, we need to make some necessary assumptions as follows.

Assumption 4.1 *The Markov chain is aperiodic and there exists a state s^* which is recurrent for each of such Markov chain.*

Assumption 4.2 *For every state $s, s' \in \mathcal{S}$, the transition probability $\mathbf{p}(s'|s, \Psi(\Theta))$ and the immediate throughput function $\mathcal{T}(s, \Theta)$ are bounded, twice differentiable, and have bounded first and second derivatives.*

Then, we can define the parameterized average throughput (i.e., the throughput under the parameter vector Θ) by

$$\psi(\Theta) = \lim_{t \rightarrow \infty} \frac{1}{t} \mathbb{E}_{\Theta} \left[\sum_{k=0}^t \mathcal{T}(s_k, \Theta_k) \right], \quad (4.6)$$

where s_k is the state of the secondary user at time step k . $\mathbb{E}_{\Theta}[\cdot]$ is the expectation under parameter vector Θ . Under Assumption 4.1, the average throughput $\psi(\Theta)$ is well defined for every Θ , and does not depend on the initial state Θ_0 . Moreover, we have the following balance equations

$$\begin{aligned} \sum_{s \in \mathcal{S}} \pi_{\Theta}(s) \mathbf{p}(s'|s, \Psi(\Theta)) &= \pi_{\Theta}(s'), \text{ for } s' \in \mathcal{S}, \\ \sum_{s \in \mathcal{S}} \pi_{\Theta}(s) &= 1, \end{aligned} \quad (4.7)$$

where $\pi_{\Theta}(s)$ is the steady-state probability of state s under the parameter vector Θ . These balance equations have a unique solution defined as a vector $\Pi_{\Theta} = [\cdots \pi_{\Theta}(s) \cdots]^{\top}$. Then, the average throughput can be expressed as follows:

$$\psi(\Theta) = \sum_{s \in \mathcal{S}} \pi_{\Theta}(s) \mathcal{T}(s, \Theta). \quad (4.8)$$

4.2.2 Learning Algorithm Based on Policy Gradient Method

To update the parameter vector Θ , we will use the algorithm based on the gradient method as introduced in [127] as follows:

$$\Theta_{k+1} = \Theta_k + \rho_k \nabla \psi(\Theta_k), \quad (4.9)$$

where ρ_k is a step size and $\nabla \psi(\Theta_k)$ is the gradient of average throughput. Under a suitable step size satisfying Assumption 4.3 and Assumption 4.1, it is proved that $\lim_{k \rightarrow \infty} \nabla \psi(\Theta_k) = 0$ and thus $\psi(\Theta_k)$ converges [127].

Assumption 4.3 *The step size ρ_k is deterministic, nonnegative and satisfies the following conditions,*

$$\sum_{k=1}^{\infty} \rho_k = \infty, \text{ and } \sum_{k=1}^{\infty} (\rho_k)^2 < \infty. \quad (4.10)$$

We then propose Proposition 4.1 to calculate the gradient of the average throughput as follows:

Proposition 4.1 *Let Assumption 4.1 and Assumption 4.2 hold, then*

$$\nabla \psi(\Theta) = \sum_{s \in \mathcal{S}} \pi_{\Theta}(s) \left(\nabla \mathcal{T}(s, \Theta) + \sum_{s' \in \mathcal{S}} \nabla \mathbf{p}(s'|s, \Psi(\Theta)) d(s', \Theta) \right), \quad (4.11)$$

where $d(s', \Theta)$ is the differential throughput at state s' . In general, we can define the differential throughput at state s as follows:

$$d(s, \Theta) = \mathbb{E}_{\Theta} \left[\sum_{k=0}^{\mathbf{T}-1} (\mathcal{T}(s_k, \Theta) - \psi(\Theta)) | s_0 = s \right], \quad (4.12)$$

where $\mathbf{T} = \min\{k > 0 | s_k = s^*\}$ is the first future time that state s^* is visited. Here, we need to note that, the main aim of defining the differential throughput $d(s, \Theta)$ is to represent the relation between the average throughput and the immediate throughput at state s , instead of the recurrent state s^* . Additionally, under Assumption 4.1, the differential throughput $d(s, \Theta)$ is a unique solution of the following Bellman equation defined as follows:

$$d(s, \Theta) = \mathcal{T}(s, \Theta) - \psi(\Theta) + \sum_{s' \in \mathcal{S}} \mathbf{p}(s'|s, \Psi(\Theta)) d(s', \Theta), \quad (4.13)$$

for all $s \in \mathcal{S}$.

Proposition 4.1 presents the gradient of the average throughput $\psi(\Theta)$ and the proof of Proposition 4.1 can be done in a similar way as Appendix A.

We then propose a learning algorithm that allows secondary users to update the parameter vector Θ at each time step based on estimating the gradient of average throughput (Algorithm 1).

Algorithm 1 Algorithm to update Θ at every time step

At time step k , the state is s_k , and the values of Θ_k , z_k , and $\tilde{\psi}(\Theta_k)$ are available from the previous iteration. We update z_k , Θ_k , and $\tilde{\psi}$ according to:

$$z_{k+1} = \begin{cases} \frac{\nabla \mu_{\Theta_k}(s_k, a_k)}{\mu_{\Theta_k}(s_k, a_k)}, & \text{if } s_k = s^* \\ z_k + \frac{\nabla \mu_{\Theta_k}(s_k, a_k)}{\mu_{\Theta_k}(s_k, a_k)}, & \text{otherwise,} \end{cases} \quad (4.14)$$

$$\Theta_{k+1} = \Theta_k + \rho_k (\mathcal{T}(s_k, a_k) - \tilde{\psi}_k) z_{k+1}, \quad (4.15)$$

$$\tilde{\psi}_{k+1} = \tilde{\psi}_k + \kappa \rho_k (\mathcal{T}(s_k, a_k) - \tilde{\psi}_k). \quad (4.16)$$

In Algorithm 1, κ is a positive constant and ρ_k is the step size of the algorithm. The convergence of Algorithm 1 can be proved in a similar way as Appendix B. Furthermore, the computation complexity of the Algorithm 1 is provided in Appendix F of the thesis.

4.3 A Decentralized Solution for RF Energy Harvesting Cognitive Radio Networks

In this section, we consider the case where secondary users cooperate in a decentralized manner. We first formulate the cooperative optimization problem as decentralized partially observable Markov decision process [38] (DEC-POMDP) and then examine a decentralized learning algorithm to obtain optimal policies for secondary users.

4.3.1 Optimization Formulation

We formulate the optimization problem for the RF energy harvesting cognitive radio network with multiple secondary users and multiple channels (i.e., the MUMC case)

as a DEC-POMDP in a discrete time system. A general DEC-POMDP model can be defined as a tuple $\langle N, \mathcal{S}, \mathcal{A}, \mathbf{p}, g, \mathcal{O}, \mathbf{o} \rangle$, where

- N is the total number of secondary users,
- $\mathcal{S}$ is a finite set of states and it is known as the global state space of the network,
- $\mathcal{A}$ is a finite set of joint actions,
- $\mathbf{p}$ is a joint transition probability function,
- g is the global immediate reward function,
- $\mathcal{O}$ is a finite set of joint observations, and
- $\mathbf{o}$ is a joint observation probability function.

4.3.1.1 State space

We define $\mathcal{S} \triangleq (\mathcal{S}_1 \times \cdots \times \mathcal{S}_n \times \cdots \times \mathcal{S}_N)$, as the global system state space where $\mathcal{S}_n$ is the local state space of secondary user n .

We define the state space of the secondary user n as follows:

$$\mathcal{S}_n = \left\{ (e_n, d_n); e_n \in \{0, 1, \dots, \mathcal{E}_n\}, d_n \in \{0, 1, \dots, \mathcal{D}_n\}, \right\} \quad (4.17)$$

where e_n , and d_n represent the energy level of the energy storage, and the number of packets in the data queue, respectively. Again, $\mathcal{E}_n$ is the maximum capacity of the energy storage and $\mathcal{D}_n$ is the maximum data queue size. The state of secondary user n is then defined as a composite variable $s_n = (e_n, d_n) \in \mathcal{S}_n$.

4.3.1.2 Action space

The joint action space $\mathcal{A}$ is a composition of sets of local action spaces from secondary users. The joint action space can be defined as follows

$$\mathcal{A} \triangleq (\mathcal{A}_1 \times \cdots \times \mathcal{A}_n \times \cdots \times \mathcal{A}_N), \quad (4.18)$$

where $\mathcal{A}_n$ is the local state space of secondary user n that is a set of available channels for secondary users to select from. Then, at each time slot, each secondary user has to make a decision $a_n \in \mathcal{A}_n = \{0, 1, \dots, M\}$ to select one of the channels to sense. $a_n = 0$ means that the secondary user n does not select any channel.

4.3.1.3 Transition probability function

The transition probability matrix for secondary users n can be expressed as follows:

$$\mathbf{P}_n(a_n) = \begin{bmatrix} \mathbf{B}_{0,0}^n(a_n) & \mathbf{B}_{0,1}^n(a_n) & & & \\ \mathbf{B}_{1,0}^n(a_n) & \mathbf{B}_{1,1}^n(a_n) & \mathbf{B}_{1,2}^n(a_n) & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{B}_{\mathcal{D}_n, \mathcal{D}_n-1}^n(a_n) & \mathbf{B}_{\mathcal{D}_n, \mathcal{D}_n}^n(a_n) & \end{bmatrix} \begin{matrix} \leftarrow d_n = 0 \\ \leftarrow d_n = 1 \\ \vdots \\ \leftarrow d_n = \mathcal{D}_n \end{matrix}, \quad (4.19)$$

where each row of matrix $\mathbf{P}_n(a_n)$ corresponds to the number of packets in the data queue. Each element of matrix $\mathbf{P}_n(a_n)$, i.e., matrix $\mathbf{B}_{d_n, d'_n}^n(a_n)$, corresponds to the transition of the data queue state from d_n in the current time slot to d'_n in the next time slot. Then, similar to $\mathbf{P}_n(a_n)$, we will construct matrix $\mathbf{B}_{d_n, d'_n}^n(a_n)$, where each element of matrix $\mathbf{B}_{d_n, d'_n}^n(a_n)$ represents the transition of the energy level from state e_n to state e'_n . There are two cases to derive matrix $\mathbf{B}_{d_n, d'_n}^n(a_n)$, i.e., for $d_n = 0$ and $d_n > 0$. For $d_n = 0$, since the data queue is empty, the energy level will never decrease and it only increases when the sensed channel is busy, no miss detection sensing error, and the secondary user harvests energy successfully. In the case of $d_n > 0$, we have to consider three sub-cases, namely, when the number of packets decreases, remains the same, or increases. The number of packets decreases when the sensed channel is idle, there is no false alarm sensing error, there is no arrival packet, the energy storage is not empty, and the packet is transmitted successfully. Similarly, for the rest of the cases, we will consider specific sub-cases and construct corresponding matrices for these cases. Then we can construct a joint transition probability matrix for the system from all transition probability matrices of secondary users.

However, to derive a joint transition probability matrix for the whole system, we need to know the transition probability matrix of each secondary user. Furthermore, to derive transition probability matrices for secondary users, we need to know environment parameters, e.g., idle channel probability, miss detection probability, false alarm sensing error probability, successful packet transmission probability, and successful RF energy harvesting probability. However, in practice, it is not easy and even impossible to obtain these probabilities for secondary users. Therefore, similar to the

learning algorithm in Chapter 3, we propose a learning algorithm that is developed based on the simulation-based method [123]. With the simulation-based learning algorithm, for a given control policy Ψ , the joint transition probability function $\mathbf{p}$ can be derived from the transition probability of the local state of the secondary users (i.e., a queue state, and energy state) as follows:

$$\mathbf{p}(s(t+1)|s(t), \Psi) = \mathbf{p}_{\text{env}} \mathbf{p}\left((d(t+1), e(t+1)) | (d(t), e(t)), \Psi\right), \quad (4.20)$$

where $\mathbf{p}_{\text{env}}$ is the probability function of environment parameters that can be generated by the simulator. $s(t) \in \mathcal{S}$ denotes the joint state of the system at time slot t . $d(t)$ and $e(t)$ denote the joint queue state and the joint energy state at time slot t , respectively. $\mathbf{p}\left((d(t+1), e(t+1)) | (d(t), e(t)), \Psi\right)$ is the joint transition probability of secondary users and this probability can be derived from the transition probability of the local states (i.e., queue state and energy state of secondary users) as follows:

$$\begin{aligned} & \mathbf{p}\left((d(t+1), e(t+1)) | (d(t), e(t)), \Psi\right) \\ &= \begin{cases} \prod_{n=1}^N \mathbf{p}(d_n(t), e_n(t)) \mathbf{p}(a_n(t)), & \text{if } d_n(t+1) = \mathcal{D}_n^*, \quad e_n(t+1) = \mathcal{E}_n^* \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (4.21)$$

where $\mathcal{D}_n^* = \min\left([d_n(t) - \phi_n(t)]^+ + \kappa_n(t), \mathcal{D}_n\right)$, and $\mathcal{E}_n^* = \min\left([e_n(t) - c_n(t)]^+ + y_n(t), \mathcal{E}_n\right)$. Again, $\mathcal{D}_n$ is the maximum size of the data queue and $\mathcal{E}_n$ is the maximum size of the energy storage of secondary user n . Here, $\phi_n(t)$ is the number of packets transmitted by secondary user n at time step t , $\kappa_n(t)$ is the number of arriving packets, $y_n(t)$ is the amount of energy harvested, and $c_n(t)$ is the amount of energy used at time slot t . Furthermore, $[x]^+ = \max(x, 0)$.

4.3.1.4 Global immediate reward function

In this model, we consider the case when all secondary users cooperate to maximize the network throughput. Thus, the global immediate reward function can be defined as follows:

$$g(s(t), a(t)) = \sum_{n=1}^N g(s_n(t), a_n(t)) = \sum_{n=1}^N r_n(t). \quad (4.22)$$

Here, the global immediate reward is the number of packets transmitted successfully at time slot t .

4.3.1.5 Observations and observation probability function

In our system model, the observation of each secondary user is its local information from the data queue and the energy storage. Therefore, the observation of each secondary user is identical to its local state and thus the observation probability function is defined in the same way as in (4.21).

4.3.2 Parameterization for DEC-POMDP

Similar to Section 4.2, we use a parameterized randomized policy [124–126]. Under the parameterized randomized policy, when secondary user n is at state s_n , the secondary user will select action a_n with the probability $\mu_{\Theta_n}(s_n, a_n)$ given as follows:

$$\mu_{\Theta_n}(s_n, a_n) = \frac{\exp(\theta_{s_n, a_n})}{\sum_{a_i \in \mathcal{A}_n} \exp(\theta_{s_n, a_i})}, \quad (4.23)$$

where $\Theta_n = \{\theta_{s_n, a_n} \in \mathbb{R}\}$ is the parameter vector of secondary user n at state s_n . Moreover, every $\mu_{\Theta_n}(s_n, a_n)$ must not be negative and $\sum_{a_n \in \mathcal{A}_n} \mu_{\Theta_n}(s_n, a_n) = 1, \forall n$.

Under the parameterized randomized policies of the secondary users, the joint transition probability function and the average cost criterion can be parameterized as follows:

$$\mathbf{p}(s'|s, \Psi(\Theta)) = \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) \mathbf{p}(s'|s, a), \quad (4.24)$$

$$g(s, \Theta) = \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) g(s, a), \quad (4.25)$$

where $\mu_{\Theta}(s, a) = \prod_{n=1}^N \mu_{\Theta_n}(s_n, a_n)$ and Θ is a joint parameter vector of the system.

Then, we can define the parameterized average throughput by

$$\bar{\mathcal{C}}(\Theta) = \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_{\Psi(\Theta)} \left[\sum_{t=1}^T g(s(t), \Theta) \right] \quad (4.26)$$

where T is the total time horizon.

We consider the case when the secondary users cooperate to obtain a joint optimal stationary control policy $\Psi(\Theta)$ to maximize the average throughput for the system. However, the actions of secondary users are made based on their local information. As a result, the optimization problem is partially observable and the control policy $\Psi_n(\Theta_n)$ is a function of a local state only. Furthermore, when we want to control

energy consumption for sensing process of secondary user n , we can impose the following constraint that on average the energy consumption for spectrum sensing must not exceed a threshold $\mathcal{Q}_n^*$ per time slot, i.e.,

$$\overline{\mathcal{Q}}_n \leq \mathcal{Q}_n^*, \quad \forall n. \quad (4.27)$$

Consequently, the optimization problem with the constraint can be defined as follows:

$$\begin{aligned} \max_{\Theta} \quad & \overline{\mathcal{C}}(\Theta) = \sum_{n=1}^N \overline{\mathcal{C}}_n(\Theta) = \lim_{T \rightarrow \infty} \sup \frac{1}{T} \mathbb{E}_{\Psi(\Theta)} \left[\sum_{t=1}^T g(s(t), \Theta) \right], \\ \text{s.t.} \quad & \overline{\mathcal{Q}}_n(\Psi(\Theta)) = \overline{\mathcal{Q}}_n(\Theta) \leq \mathcal{Q}_n^*, \quad \forall n, \end{aligned} \quad (4.28)$$

where $g(s(t), \Theta)$ is the parameterized common immediate throughput generated at the joint state $s(t)$ after the joint action $a(t)$ is made, and $\overline{\mathcal{C}}(\Theta)$ is the parameterized average total throughput. $\overline{\mathcal{Q}}_n$ denotes the average energy consumption for sensing channels of secondary user n and $\mathcal{Q}_n^*$ denotes the target threshold that we want to control.

4.3.3 Lagrange Multiplier and Policy Gradient Method

To solve the optimization problem with constraints as defined in (4.28), we define a Lagrange function based on the *Lagrange multiplier* method as follows:

$$\mathcal{L}(\Theta, \gamma) = \sum_{n=1}^N \left(\overline{\mathcal{C}}_n(\Theta) + \gamma_n (\overline{\mathcal{Q}}_n(\Theta) - \mathcal{Q}_n^*) \right), \quad (4.29)$$

where γ_n is a Lagrange multiplier for the constraint of secondary user n . If we denote $\mathcal{G}(s, a) = \sum_{n=1}^N \left(g(s_n, a_n) + \gamma_n (\mathcal{Q}_n - \mathcal{Q}_n^*) \right)$ as the immediate value function, then the parameterized immediate value function will be

$$\mathcal{G}(s, \Theta) = \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) \mathcal{G}(s, a). \quad (4.30)$$

We have the following balance equations:

$$\begin{aligned} \sum_{s \in \mathcal{S}} \pi_s(\Theta) \mathbf{p}(s' | s, \Psi(\Theta)) &= \pi_{s'}(\Theta), \quad \text{for } s' \in \mathcal{S}, \\ \sum_{s \in \mathcal{S}} \pi_s(\Theta) &= 1, \end{aligned} \quad (4.31)$$

where $\pi_s(\Theta)$ is the steady-state probability of the joint state s under the parameter vector Θ . These balance equations have a unique solution defined as a vector $\Pi_\Theta = [\dots \pi_s(\Theta) \dots]^\top$. Then, the Lagrange function can be represented as follows:

$$\mathcal{L}(\Theta, \gamma) = \sum_{s \in \mathcal{S}} \pi_s(\Theta) \mathcal{G}(s, \Theta). \quad (4.32)$$

Then, to solve the Lagrange function, we use *Karush-Kuhn-Tucker* (KKT) conditions [131] to find a local optimal solution Θ^* that satisfies the following conditions:

$$\begin{aligned} \nabla_{\Theta} \mathcal{L}(\Theta^*, \gamma^*) &= \mathbf{0}, \\ \overline{\mathcal{Q}}_n(\Theta^*) - \mathcal{Q}_n^* &\leq 0, \\ \gamma_n^* (\overline{\mathcal{Q}}_n(\Theta^*) - \mathcal{Q}_n^*) &= 0, \forall n, \\ \gamma_n^* &\geq 0. \end{aligned} \quad (4.33)$$

To obtain the gradient of the Lagrange function $\mathcal{L}(\Theta, \gamma)$, we first define the differential cost $q(s, a, \Theta)$ at state s under a control action a as follows:

$$q(s, a, \Theta) = \mathbb{E}_{\Psi(\Theta)} \left[\sum_{t=0}^{\mathbf{T}-1} \left(\mathcal{G}(s(t), a(t)) - \mathcal{L}(\Theta, \gamma) \right) \middle| s(0) = s, a(0) = a \right], \quad (4.34)$$

where $\mathbf{T} = \min\{t > 0 | s(t) = s^\dagger\}$ is the first future time that state $s^\dagger = (s_1^\dagger, \dots, s_N^\dagger)$ is visited and $\mathcal{G}(s(t), a(t)) = \sum_{n=1}^N \left(g(s_n(t), a_n(t)) + \gamma_n (\mathcal{Q}_n(t) - \mathcal{Q}_n^*) \right)$. $s^\dagger$ can be selected randomly from state space $\mathcal{S}$. Here, $g(s_n(t), a_n(t))$ is the number of packets in the data queue and $\mathcal{Q}_n(t)$ is the used energy at time slot t of secondary user n . In (4.34), $q(s, a, \Theta)$ can be expressed as the differential cost if action a is made based on policy μ_Θ at state s . Then, we can obtain the gradient of the Lagrange function $\mathcal{L}(\Theta, \gamma)$ as in Proposition 4.2.

Proposition 4.2 *The gradient of the Lagrangian function is determined as follows:*

$$\nabla_{\Theta_n} \mathcal{L}(\Theta, \gamma) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi_s(\Theta) \mu_\Theta(s, a) \frac{\nabla_{\Theta_n} \mu_{\Theta_n}(s_n, a_n)}{\mu_{\Theta_n}(s_n, a_n)} q(s, a, \Theta), \quad (4.35)$$

where $\pi_s(\Theta)$ is steady state probability of state $s \in \mathcal{S}$ and $\mu_\Theta(s, a) = \prod_{n=1}^N \mu_{\Theta_n}(s_n, a_n)$.

The proof of Proposition 4.2 is provided in Appendix D.

Then, based on the Proposition 4.2, secondary users will update their parameter vectors based on the idealized gradient algorithm at each time step t as follows [127]:

$$\Theta_n^{t+1} = \Theta_n^t + \rho_t \nabla_{\Theta_n} \mathcal{L}(\Theta, \gamma), \quad \forall n \quad (4.36)$$

where ρ_t is a suitable step size that satisfies Assumption 4.4.

Assumption 4.4 *The step size ρ_t is deterministic, nonnegative and satisfies the following conditions,*

$$\sum_{t=1}^{\infty} \rho_t = \infty, \text{ and } \sum_{t=1}^{\infty} (\rho_t)^2 < \infty. \quad (4.37)$$

In other words, the value of the step size to update the parameter vectors approaches zero when the time step goes to infinity.

4.3.4 Decentralized Online Learning Algorithm with Communications

With the idealized gradient method, secondary users can update their parameter vectors iteratively by using (4.36) to find a joint optimal solution. However, with the idealized gradient method, the secondary users need to know the Lagrange function $\mathcal{L}(\Theta, \gamma)$ to calculate their partial differential equation for updating their parameter vector Θ_n . Furthermore, we need to calculate the gradient of the Lagrange function $\mathcal{L}(\Theta, \gamma)$ with respect to Θ_n at every time step, which is impossible to compute if the system has a large state space. Therefore, in this chapter, we propose a decentralized online learning algorithm with a small communication overhead that can estimate the gradient of the Lagrange function instead of computing its exact value. Then, the secondary users can update their parameter vectors independently and parallelly as shown in Algorithm 2.

In Step (iv) of Algorithm 2, to reduce the communication overhead, we just need to send a state synchronization signal when the current state of the secondary user is the recurrent state of that user, thereby reducing the communication for the whole system.³

³In the case when state $s^\dagger$ is visited, each secondary user will receive all messages from all other users to notify that they are in recurrent state $s_n^\dagger$.

Algorithm 2 Decentralized Online Learning Algorithm with Communications

- (i) **Initialization:** Each secondary user determines the local parameter vector Θ_n^0 .
- (ii) **Sensing and Decision Epoch:** At the beginning of each time slot, the secondary user makes a decision to sense a target channel based on the information from its local state (i.e., the number of packets and the energy level). If the sensed channel is busy and the energy storage of the secondary user is not full, the secondary user will harvest RF energy. By contrast, if the sensed channel is idle, the energy queue and the data queue are not empty, then the secondary user will transmit a packet. Otherwise, the secondary user does nothing.
- (iii) **Channel Processing:** After decisions are made, the secondary users perform RF energy harvesting or packet transmission according to the decisions made in the sensing and decision phases.
- (iv) **Information Sharing:** At the end of each time slot, each secondary user determines its local current state and shares the following information, $\mathcal{I}_n = g(s_n(t), a_n(t)) + \gamma_n(\mathcal{Q}_n(t) - \mathcal{Q}_n^*)$. If the current state of secondary user n is $s_n^\dagger$, the user will also send a state synchronization signal v_n to the other secondary users.
- (v) **Updating parameter Θ_n :** Each secondary user updates the local parameter Θ_n as follows:

$$\Theta_n^{t+1} = \Theta_n^t - \alpha(t) \left(\mathcal{I}_G - \tilde{\mathcal{L}}^t \right) z_n^t, \quad (4.38)$$

where $\mathcal{I}_G = \sum_{n=1}^N \mathcal{I}_n$ is the current total value of the Lagrange functions of secondary users and $\tilde{\mathcal{L}}^t$ is the estimated total value of Lagrange functions and it is updated as follows:

$$\tilde{\mathcal{L}}^{t+1} = \tilde{\mathcal{L}}^t - \alpha(t) \left(\mathcal{I}_G - \tilde{\mathcal{L}}^t \right), \quad (4.39)$$

where $\alpha(t)$ is the step size satisfying Assumption 4.5, and

$$z_n^{t+1} = \begin{cases} \frac{\nabla_{\Theta_n} \mu_{\Theta_n}(s_n, a_n)}{\mu_{\Theta_n}(s_n, a_n)}, & \text{if } s^\dagger \text{ is visited,} \\ z_n^t + \frac{\nabla_{\Theta_n} \mu_{\Theta_n}(s_n, a_n)}{\mu_{\Theta_n}(s_n, a_n)}, & \text{otherwise.} \end{cases} \quad (4.40)$$

- (vi) **Updating Lagrangian multiplier γ_n :** Each secondary user updates the local Lagrangian multiplier $\{\gamma_n\}$ as follows:

$$\gamma_n^{t+1} = \max \left(\gamma_n^t + \beta(t) (\mathcal{Q}_n(t) - \mathcal{Q}_n^*), 0 \right), \quad (4.41)$$

where $\beta(t)$ is the step size satisfying Assumption 4.5.

Assumption 4.5 *The step sizes $\alpha(t)$ for updating parameter vectors and $\beta(t)$ for updating Lagrangian multipliers are deterministic, nonnegative and satisfy the following conditions:*

$$\sum_{t=1}^{\infty} \alpha(t) = \sum_{t=1}^{\infty} \beta(t) = \infty, \quad \sum_{t=1}^{\infty} (\alpha^2(t) + \beta^2(t)) < \infty,$$

$$\text{and } \frac{\beta(t)}{\alpha(t)} \rightarrow 0.$$

The last condition, i.e., $\frac{\beta(t)}{\alpha(t)} \rightarrow 0$, implies that the sequence $\{\beta(t)\} \rightarrow 0$ faster than the sequence $\{\alpha(t)\}$. For example, we can choose $\alpha(t) = \frac{1}{t^{2/3}}$ and $\beta(t) = \frac{1}{t}$ or $\alpha(t) = \frac{1}{t}$ and $\beta(t) = \frac{1}{1+t \log t}$, and so on [132].

With Algorithm 2, the secondary users can make decisions based on their local information and exchanged messages (i.e., $\mathcal{I}_n$ and v_n defined in Algorithm 2). Appendix E provides the analysis and the proof of the convergence for Algorithm 2. Moreover, the computation complexity of the Algorithm 2 can be analyzed in a similar way in Appendix F of the thesis.

4.4 Performance Evaluation

In this section, we perform simulations to evaluate the performance of the RF energy harvesting cognitive radio network.

4.4.1 Simulation Setup

We perform simulations through using MATLAB to evaluate the performance of the network under different parameters and scenarios. First, we consider the case with two primary channels and two secondary users. The secondary users cooperate to maximize the network throughput. In this case, we will show the convergence of the learning algorithms together with their optimal policies. We then increase the number of cooperative secondary users and compare the network throughput of the proposed algorithms with two other schemes, namely, greedy policy and threshold policy. For the greedy policy, the secondary users will transmit a packet when they meet certain conditions (i.e., the energy queue and the data queue is not empty). Otherwise the secondary users will harvest energy. For the threshold policy, the secondary users will

transmit data when the data queue is not empty and the energy level in the energy queue is higher than a safety level. We set the safety level for the threshold policy for secondary user n as $\lfloor \mathcal{E}_n/2 \rfloor$ where $\lfloor \cdot \rfloor$ is the floor function. With the threshold policy, secondary users can reserve a certain energy in order to serve for data transmission when the sensed channel is idle, and also avoid collisions when the number of available channels is few. For both the greedy policy and threshold policy, we assume that secondary users know the channel idle probabilities of all the channels in advance. Thus if the secondary user wants to transmit a packet, it will sense the channel that has higher idle probability, whereas when the secondary user wants to harvest energy, it will choose the channel that has lower idle probability.

Finally, we increase the number of channels and consider two scenarios. In the case with two secondary users and two channels, we assume that the channel idle probabilities of channel 1 and channel 2 are 0.2 and 0.8, respectively. When we increase the number of channels to 3, the channel idle probability of channel 3 is 0.2 in the first scenario and 0.8 in the second scenario. The successful packet transmission and the successful energy harvesting probabilities for all cases are set to be 0.95. We set the false alarm probability and the miss detection probability to be 0.01. The maximum number of packets in the data queues of secondary users is 5 packets and the maximum levels of the energy queues are 5 units of energy. The packet arrival probability at data queues is 0.5.

For the parameter vector Θ , at the beginning of the learning algorithms, secondary users will choose to sense channels with the same probabilities. For example, if we have two channels to sense, the probabilities to sense channel 1 or channel 2 are equal, i.e., 0.5.

4.4.2 Simulation Results

In the first case, we show the convergence of the learning algorithms and the optimal policies obtained by using the proposed algorithms with two secondary users and two primary channels. Figs 4.2 (a) and (b) show the convergence of the TDMA learning algorithm (i.e., Algorithm 1) and the decentralized learning algorithm (i.e.,

Algorithm 2) respectively, through the average throughput of the network. As shown in Fig. 4.2, the average throughput of the Algorithm 1 converges to around 0.7 after 2×10^6 iterations, while the average throughput of the Algorithm 2 converges to approximately 0.46 after 4×10^6 iterations. For the TDMA learning algorithm, the secondary users have to experience two learning processes. In the first learning process, a secondary user needs to determine whether it should transmit a packet or harvest energy given the current state. In the second learning process, the secondary user needs to choose a channel to sense to maximize the throughput. For example, when the secondary user wants to transmit a packet, it will choose the channel with the highest channel idle probability. However, for the decentralized learning algorithm, the secondary user has three learning processes. The secondary user has to learn not only when to transmit a packet and which channel to sense, but also the behaviors of other secondary users due to random access process. As a result, the convergence rate of the decentralized learning algorithm will be slower than that of the TDMA learning algorithm as shown in Fig. 4.2.

In Fig. 4.2, the average throughput for both secondary users obtained by the TDMA learning algorithm is nearly equal, while the average throughput of secondary users obtained by the decentralized learning algorithm has noticeable difference. The reason can be explained from the policies obtained from the learning algorithm. Figs. 4.3 and 4.4 present the policies obtained by the TDMA learning algorithm and Fig. 4.5 presents the policies obtained by the decentralized learning algorithm for both secondary users. In these figures, the x-axis, y-axis, and z-axis represent the number of packets, the energy levels, and the probability of sensing channel 1 of the secondary user, respectively. In Figs. 4.3 and 4.4, the policies obtained by both secondary users are almost the same, while in Fig. 4.5, the policies obtained by the decentralized learning algorithm for both secondary users are different. For the TDMA learning algorithm, the secondary users are cooperative in a round-robin scheduling, and thus, at each time slot there is just only one secondary user interacting with the environment (i.e., primary channels). As a result, with the same environment, the secondary users will have the same policies to optimize their throughput. The policies

(a)

(b)

Figure 4.2: The convergence of (a) the TDMA learning algorithm and (b) the decentralized learning algorithm.

obtained by the TDMA learning algorithm can be explained as follows. The secondary users will sense channel 1 (higher channel busy probability) to harvest energy when they are not scheduled to access a time slot (i.e., not allowed to transmit a packet) and the secondary users will sense channel 2 which has higher idle probability to transmit a packet when they are scheduled (allowed to transmit a packet). However, for the decentralized learning algorithm, the secondary users interact not only with the channels to explore the environment, but also with other secondary users to learn their behaviors. The policies applied for the secondary users are randomized parameterized policies, and thus, actions are selected randomly. Consequently, in the network, the policies for the secondary users can be different as long as their objective (i.e., the average throughput) is maximized. We need to note that, although the policies for the secondary users could be different, the average network throughput obtained by the decentralized learning algorithm still converges to the same point.

We then vary the number of secondary users to evaluate the network performance. We also compare the performance with other schemes to demonstrate the efficiency of the proposed algorithms. As shown in Fig. 4.6, when the number of secondary users increases from 2 to 5, the average total throughput obtained by the TDMA learning algorithm is the highest and it increases when the number of secondary users increases and becomes saturated. In contrast, the average throughput obtained by the decentralized learning algorithm, the greedy policy and the threshold policy decrease as the number of secondary users increases. Here, although the average throughput of the decentralized algorithm is greater than those of the greedy policy and the threshold policy, it is still much lower than that of the TDMA learning algorithm. The reason is due to the energy harvesting process of secondary users. In the TDMA learning algorithm, when any secondary users are not scheduled to access a time slot, they can harvest energy. This is not the case for the decentralized learning algorithm since the secondary users have to contend for transmission according to their states, losing opportunity to harvest energy. Additionally, the transmissions in the decentralized learning algorithm can result in collision, lowering the network throughput. Finally,

(a)

(b)

Figure 4.3: The policy of secondary user 1 obtained by the TDMA learning algorithm for the cases that (a) it is not scheduled and (b) it is scheduled to access a time slot.

(a)

(b)

Figure 4.4: The policy of secondary user 2 obtained by the TDMA learning algorithm for the cases that (a) it is not scheduled and (b) it is scheduled to access a time slot.

(a)

(b)

Figure 4.5: The policy of secondary users (a) 1 and (b) 2 obtained by the decentralized learning algorithm.

Figure 4.6: The average throughput of the system under different algorithms with 2 channels.

for the greedy policy and the threshold policy, they are not optimized, and hence, their performance is inferior to those of the learning algorithms.

We now investigate the case when there are three available primary channels. Fig. 4.7 (a) shows the results for the case of two channels with high idle probability (i.e., 0.8) and one with low idle probability (i.e., 0.2). In Fig. 4.7 (a), the performance obtained by the algorithms is similar to the case with one channel having high idle probability and one channel having low idle probability as shown in Fig. 4.6. However, the average throughput obtained in the former case is greater than that in the latter case. The reason is that the secondary users now have more chances to harvest energy and transmit packets. Fig. 4.7 (b) presents the results for the case of two channels with low idle probability and one channel with high idle probability. In Fig. 4.7 (b), the average throughput obtained by the algorithms increases significantly except the TDMA learning algorithm. The reason is because the TDMA learning algorithm works in a round-robin scheduling. When the number of channels increases, the network performance will not be impacted. However, for other algorithms, when the number of idle channels increases, there are more chances for packets to be transmitted successfully in a time slot.

(a)

(b)

Figure 4.7: The average throughput with (a) two channels with low idle probability and one channel with high idle probability and (b) two channels with high idle probability and one channel with low idle probability.

Figure 4.8: Collision probability with two channels.

The above results lead to a conclusion that the TDMA learning algorithm is the most effective solution when the number of channels is small and there are many secondary users. In fact, when the number of channels becomes large and the number of secondary users is small enough, the average throughput obtained by the decentralized learning algorithm could be greater than that of the TDMA learning algorithm.

Additionally, we also show the average collision probability of the algorithms in Fig. 4.8. Note that the collision probability of the TDMA learning algorithm is very small. The collision in TDMA learning algorithm only occurs when a secondary user has a miss detection sensing error event. However, the average collision probabilities obtained by other algorithms are relatively high. This is because the collision can happen due to simultaneous transmissions by multiple secondary users.

4.5 Summary

This chapter has introduced the RF energy harvesting cognitive radio network. The secondary users in the network can harvest RF energy from a busy channel occupied by a primary user, and transmit their packets on an idle channel. The main focus is to solve the performance optimization problem for the RF energy harvesting cog-

nitive radio network with multiple secondary users and multiple channels. We have studied two solutions, namely, a TDMA learning algorithm and decentralized learning algorithm to obtain optimal channel access policies for secondary users in an environment with incomplete information. The simulation results have clearly shown the convergence of the learning algorithms as well as their efficiency in terms of network throughput.

Table 4.1: Summary of key notations of Chapter 4

Notation	Physical meaning
N	The total number of secondary users
M	The total number of primary channels
$\mathcal{D}_n$	Maximum size of the data queue for secondary user n
$\mathcal{E}_n$	Maximum size of the energy queue for secondary user n
λ_n	The probability of a packet arriving to secondary user n
s, a	The state and action of the state space $\mathcal{S}$ and action space $\mathcal{A}$, respectively
$s^*, s^\dagger$	The revisit state of the secondary user in Section 4.2 and of the whole system in Section 4.3, respectively
$\mathbf{T}$	The first future time that state s^* or $s^\dagger$ is revisited
d, e and ϑ	The energy level, the number of packets and the time schedule
$\mathcal{T}$	The immediate throughput function
μ	The randomized parameterized policy
Θ	The parameter vector
$\mathbf{p}$	The transition probability function
Ψ	The control policy
ψ	The average throughput of the secondary user in Section 4.2
$\tilde{\psi}$	The estimated average throughput of ψ
$\pi(s)$	The steady state probability of state s
ρ	The step size
κ	The positive constant
g	The global immediate reward function
$\phi_n(t)$	The number of packets transmitted by secondary user n at time slot t
$\kappa_n(t)$	The number of packets arrived to secondary user n at time slot t
$y_n(t)$	The amount of energy harvested by secondary user n at time slot t
$c_n(t)$	The amount of energy used by secondary user n at time slot t
$\mathcal{C}(\Theta)$	The parameterized average throughput
$\mathcal{Q}_n$	The immediate energy consumption of secondary user n
$\bar{\mathcal{Q}}_n$	The average energy consumption of secondary user n
$\mathcal{Q}_n^*$	The energy consumption threshold of secondary user n
$\mathcal{L}$	The Lagrange multiplier function
$\mathcal{G}$	The immediate value function of the Lagrange multiplier function
$q(s, a, \Theta)$	The differential cost at state s under a control action a and a parameter vector Θ
$d(s, \Theta)$	The differential cost at state s under the parameter vector Θ
$\mathcal{I}_n$ and $\mathcal{I}_G$	The Lagrange function value of secondary user n and of the system
v_n	The synchronization signal of secondary user n
$\tilde{\mathcal{L}}$	The estimated total value of Lagrange function
α and β	The pair of step sizes satisfying Assumption 4.5

Chapter 5

Conclusions and Future Works

In this chapter, we recapitulate the chapters and highlight key points which are the major contributions for this thesis. We then present potential research directions of RF powered CRNs.

5.1 Conclusions

The main contents of the thesis can be summarized as follows:

- **Chapter 1:** We provide fundamental backgrounds about cognitive radio networks, wireless energy harvesting techniques, and Markov decisions processes.
- **Chapter 2:** We discuss research works related to wireless powered CRNs. We also present the current research trends and show the scope as well as the novelty of the thesis.
- **Chapter 3:** We have considered a cognitive radio network in which the secondary user is equipped with the RF energy harvesting capability. The secondary user can transmit a packet if the selected channel is not occupied by the primary user. Alternatively, the secondary user can harvest RF energy from the primary user's transmission if the selected channel is occupied. In this network, the secondary user has to perform the channel sensing process to select the best channel given its current state. We have first presented the optimization formulation based on a Markov decision process to obtain the channel access policy.

This formulation does not need the secondary user to know the current channel status. However, the optimization still requires the model parameters, which may not be available in practice. Therefore, we have adopted the online learning algorithm which is able to interact with the environment and take appropriate actions. Additionally, we have considered the optimization for the case where the channel status is known by the secondary user. This complete information case can yield the upper-bound performance for benchmarking purposes. Finally, we have performed the performance evaluation, which shows the success of using the learning algorithm in terms of efficiency and convergence.

- **Chapter 4:** The main focus of this chapter is to solve the performance optimization problem for the RF energy harvesting cognitive radio network with multiple secondary users and multiple channels. We have studied two solutions, namely, the TDMA learning algorithm and the decentralized learning algorithm to obtain optimal channel access policies for secondary users in the environment without complete information. The simulation results have clearly shown the convergence of the learning algorithms as well as their efficiency in terms of network throughput.

5.2 Future Works

As shown in Chapter 2, it is clear that RF powered CRNs have been receiving a lot of attention recently and with a double up-trend every year (since 2012), this topic will be a key research direction in the next few years. However, as shown in Fig. 2.5 (b) (Chapter 2), there are some research issues which still have not been well investigated in this topic. This makes spaces for us to improve the network performance for RF powered CRNs.

5.2.1 Channel Feedback Information for Secondary Systems

In the current work, we assume that at each time slot, the channel state is either idle or busy, and the channel quality is stable in every time slot. However, in practice,

the channel quality could be different in different time slots. As a result, the channel feedback for secondary systems needs to be taken into account. For example, in the current time slot, the channel state is idle, but the channel quality (feedbacked from secondary receiver) is not good. Thus, the secondary transmitter may not transmit packets to the secondary receiver because the successful transmission probability is low. Consequently, to obtain the optimal decision for the secondary transmitter, we need to take not only its current state (i.e., the energy level and the number of packets) and the channel state (i.e., idle or busy), but also the channel feedback information into considerations simultaneously.

5.2.2 Performance Optimization for Underlay RF-Powered CRNs

In this thesis, we mainly focus on the optimization problem for the overlay cognitive radio networks (CRNs). In the future work, we can study the optimization problem for underlay RF-powered CRNs. Different from overlay CRNs where the secondary users (SUs) are allowed to access the primary channel when the channel is not occupied by the primary users, in underlay CRNs, the SUs are allowed to transmit packets to the primary channel even when the primary channel is occupied by primary users as long as the interference to the primary system is under a predefined threshold. Therefore, the optimization problem for the SUs becomes more complicated because the SUs have to choose not only one of the channels to sense, but also how much time and energy that they should spend for energy harvesting process and data transmission process. In addition, for the optimization problem in underlay CRNs, the interference constraints for primary systems must be taken into considerations.

5.2.3 Game Models in RF Powered CRNs

In this thesis, though many problems are discussed in the energy harvesting CRNs, most of them are optimization problems, e.g., minimizing delay, maximizing throughput or performance. However, in practice, wireless nodes are not always willing to cooperate and thus we need to adopt game theory in analyzing the competition among

them. In addition, game models can support distributed decision making by using local information, and thus wireless nodes can avoid communication overhead and minimize energy consumption. For example, by applying repeated games [133] with punishment mechanisms, we can encourage wireless nodes to cooperate in a distributed way and avoid network disruption due to selfish behavior.

5.2.4 Integrating RF Powered CRNs with other Networks

With many advantages presented in this thesis, RF powered CRNs have become more and more popular, and they can be integrated into many other networks, e.g., Internet-of-Things (IoT), body area networks, and machine-to-machine communications. For example, sensors/actuators in IoT have to gather and transmit data to the centralized controller for further processing. In order to reduce human intervention while still guaranteeing the QoS for IoT services, sensors/actuators must be smarter. In particular, they are able not only to harvest energy to serve for their operation, but also to choose efficient channels to transmit data. Thus, it is clear that the RF powered CRNs will bring many benefits for the development of other wireless systems in the near future.

5.2.5 Design and Implement on Hardware Devices

Most of current research works have been focusing on developing applications of RF powered CRNs, while how to design and implement such applications in practice has not received sufficient attention. The implementation of RF energy harvesting circuits along with transceivers on the same device can cause negative impacts for the device, e.g., reduce the energy harvesting capability and the data transmission performance. However, there is still no research work investigating the implementation issues for wireless nodes in RF powered CRNs. Therefore, this is an important step which needs to be studied before RF Powered CRNs can be deployed widely in practice.

5.2.6 Economic Models in RF Powered CRNs

In addition to technical problems, economic issues are also important aspects for the development of RF powered CRNs. It is clear that the development of RF powered

CRNs will not only bring many advantages to cognitive users, but also open new business opportunities for service providers as well as network operators. For example, in [134], an economic model based on the oligopoly price competition and the Bertrand game was proposed to address the spectrum pricing problem for a CRN. Through using the proposed economic model, the author demonstrated the potential services in cognitive networks which can bring great profits for spectrum service providers. In RF powered CRNs, we can jointly consider spectrum services and energy supply services for providers. In this way, the pricing problem of service providers will become more complex due to the interaction of the spectrum service and the energy supply service, which requires more sophisticated economic models to meet with new requirements.

Appendix A

The proof of Proposition 3.1

This is to show the gradient of the average throughput. In (3.22), we have $\sum_{\theta \in \Theta} \pi_{\Phi}(\theta) = 1$, so $\sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) = 0$.

Recall that

$$d(\theta, \Phi) = \mathcal{T}_{\Phi}(\theta) - \psi(\Phi) + \sum_{\theta' \in \Theta} P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi)$$

$$\text{and } \psi(\Phi) = \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \mathcal{T}_{\Phi}(\theta).$$

Then, we derive the following results:

$$\nabla \psi(\Phi) = \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \nabla \mathcal{T}_{\Phi}(\theta) + \sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) \mathcal{T}_{\Phi}(\theta) \quad (5.1)$$

$$= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \nabla \mathcal{T}_{\Phi}(\theta) + \sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) \mathcal{T}_{\Phi}(\theta) - \psi(\Phi) \sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) \quad (5.2)$$

$$= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \nabla \mathcal{T}_{\Phi}(\theta) + \sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) (\mathcal{T}_{\Phi}(\theta) - \psi(\Phi))$$

$$= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \nabla \mathcal{T}_{\Phi}(\theta) + \sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) \left(d(\theta, \Phi) - \sum_{\theta' \in \Theta} P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi) \right).$$

We define

$$\nabla \left(\pi_{\Phi}(\theta) P(\theta, \theta', \Psi(\Phi)) \right) = \nabla \pi_{\Phi}(\theta) P(\theta, \theta', \Psi(\Phi)) + \pi_{\Phi}(\theta) \nabla P(\theta, \theta', \Psi(\Phi)) \quad (5.3)$$

and from (3.22), $\pi_{\Phi}(\theta') = \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) P(\theta, \theta', \Psi(\Phi))$. Then we have the derivations as given in (5.4)-(5.8).

The proof is completed.

$$\begin{aligned} \nabla\psi(\Phi) &= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \nabla \mathcal{J}_{\Phi}(\theta) + \\ &\quad \sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) \left(d(\theta, \Phi) - \sum_{\theta' \in \Theta} P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi) \right) \end{aligned} \quad (5.4)$$

$$\begin{aligned} &= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \nabla \mathcal{J}_{\Phi}(\theta) + \sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) d(\theta, \Phi) \\ &\quad + \sum_{\theta, \theta' \in \Theta} \left(\pi_{\Phi}(\theta) \nabla P(\theta, \theta', \Psi(\Phi)) - \nabla \left(\pi_{\Phi}(\theta) P(\theta, \theta', \Psi(\Phi)) \right) \right) d(\theta', \Phi) \end{aligned} \quad (5.5)$$

$$\begin{aligned} &= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \nabla \mathcal{J}_{\Phi}(\theta) + \sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) d(\theta, \Phi) \\ &\quad + \sum_{\theta, \theta' \in \Theta} \pi_{\Phi}(\theta) \nabla P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi) - \\ &\quad \sum_{\theta' \in \Theta} \nabla \left(\sum_{\theta \in \Theta} \pi_{\Phi}(\theta) P(\theta, \theta', \Psi(\Phi)) \right) d(\theta', \Phi) \end{aligned} \quad (5.6)$$

$$\begin{aligned} &= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \nabla \mathcal{J}_{\Phi}(\theta) + \sum_{\theta \in \Theta} \nabla \pi_{\Phi}(\theta) d(\theta, \Phi) \\ &\quad + \sum_{\theta, \theta' \in \Theta} \pi_{\Phi}(\theta) \nabla P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi) - \sum_{\theta' \in \Theta} \nabla \pi_{\Phi}(\theta') d(\theta', \Phi) \end{aligned} \quad (5.7)$$

$$= \sum_{\theta \in \Theta} \pi_{\Phi}(\theta) \left(\nabla \mathcal{J}_{\Phi}(\theta) + \sum_{\theta' \in \Theta} \nabla P(\theta, \theta', \Psi(\Phi)) d(\theta', \Phi) \right). \quad (5.8)$$

Appendix B

The proof of Proposition 3.2

We will prove the convergence of the Algorithm 1. The update equations of Algorithm 1 can be rewritten in the specific form as in (5.9).

We define the vector $\mathbf{r}^{k_m} = \begin{bmatrix} \Phi_m & \tilde{\psi}_m \end{bmatrix}^\top$, then (5.9) becomes

$$\mathbf{r}^{k_{m+1}} = \mathbf{r}^{k_m} + \rho_m \mathbf{H}_m, \quad (5.10)$$

where

$$\mathbf{H}_m = \begin{bmatrix} \sum_{k'=k_m}^{k_{m+1}-1} \left(\sum_{k=k'}^{k_{m+1}-1} (\mathcal{T}(\theta_k, \delta_k) - \tilde{\psi}_m) \right) \frac{\nabla \mu_{\Phi_m}(\theta_{k'}, \delta_{k'})}{\mu_{\Phi_m}(\theta_{k'}, \delta_{k'})} \\ \kappa \sum_{k'=k_m}^{k_{m+1}-1} (\mathcal{T}(\theta_k, \delta_k) - \tilde{\psi}_m) \end{bmatrix}. \quad (5.11)$$

Let $\mathcal{F} = \{\Phi_0, \tilde{\psi}_0, \theta_0, \theta_1, \dots, \theta_m\}$ be the history of the Algorithm 1. Then from Proposition 2 in [125], we have

$$\mathbb{E}[\mathbf{H}_m | \mathcal{F}_m] = \mathbf{h}_m = \begin{bmatrix} \mathbb{E}_\Phi[T] \nabla \psi(\Phi) + \mathcal{V}(\Phi) (\psi(\Phi) - \tilde{\psi}(\Phi)) \\ \kappa \mathbb{E}_\Phi[T] (\psi(\Phi) - \tilde{\psi}(\Phi)) \end{bmatrix}, \quad (5.12)$$

where

$$\mathcal{V}(\Phi) = \mathbb{E}_\Phi \left[\sum_{k'=k_{m+1}}^{k_{m+1}-1} (k_{m+1} - k') \frac{\nabla \mu_{\Phi_m}(\theta_{k'}, \delta_{k'})}{\mu_{\Phi_m}(\theta_{k'}, \delta_{k'})} \right].$$

Recall that $T = \min\{k > 0 | \theta_k = \theta^*\}$ is the first future time that the recurrent state θ^* is visited.

Consequently, (5.9) has the following form

$$\mathbf{r}^{k_{m+1}} = \mathbf{r}^{k_m} + \rho_m \mathbf{h}_m + \varepsilon_m, \quad (5.13)$$

where $\varepsilon_m = \rho(\mathbf{H}_m - \mathbf{h}_m)$ and note that $\mathbb{E}[\varepsilon_m | \mathcal{F}_m] = 0$. Since ε_m and ρ_m converge to zero almost surely, along with $\mathbf{h}_m$ is bounded, we have

$$\lim_{m \rightarrow \infty} (\mathbf{r}^{k_{m+1}} - \mathbf{r}^{k_m}) = 0. \quad (5.14)$$

$$\begin{aligned}
 \Phi_{m+1} &= \Phi_m + \rho_m \left(\sum_{k'=k_m}^{k_{m+1}-1} \left(\sum_{k=k'}^{k_{m+1}-1} (\mathcal{T}(\theta_k, \delta_k) - \tilde{\psi}_m) \right) \frac{\nabla \mu_{\Phi_m}(\theta_{k'}, \delta_{k'})}{\mu_{\Phi_m}(\theta_{k'}, \delta_{k'})} \right), \\
 \tilde{\psi}_{m+1} &= \tilde{\psi}_m + \kappa \rho_m \sum_{k'=k_m}^{k_{m+1}-1} (\mathcal{T}(\theta_k, \delta_k) - \tilde{\psi}_m)
 \end{aligned} \tag{5.9}$$

After that, based on Lemma 11 in [125], it was proved that $\psi(\Phi)$ and $\tilde{\psi}(\Phi)$ converge to a common limit. This means the parameter vector Φ can be represented in the following way

$$\Phi_{m+1} = \Phi_m + \rho_m \mathbb{E}_{\Phi_m}[T](\nabla \psi(\Phi_m) + e_m) + \epsilon_m, \tag{5.15}$$

where e_m is an error term that converges to zero and ϵ_m is a summable sequence. (5.15) is known as the gradient method with diminishing errors [132, 135]. Therefore, following the same way in [132, 135], we can prove that $\nabla \psi(\Phi_m)$ converges to 0, i.e., $\nabla_{\Phi} \psi(\Phi_{\infty}) = 0$.

Appendix C

Markov Decision Process: Complete Information Case

We extend the optimization formulation based on an MDP presented in Section 3.2. We consider the complete information case, where the secondary user knows the status of all the channels before selecting the channel to access. The optimal channel access policy of the complete information case will determine the upper-bound performance for the secondary user. In practice, the complete information case can be realized when there is a dedicated infrastructure (e.g., a spectrum sensing network) to perform spectrum sensing and share the channel status to the secondary user.

State space and action space

We define the state space of the secondary user as follows:

$$\Theta = \left\{ (\mathcal{C}, \mathcal{E}, \mathcal{Q}); \mathcal{E} \in \{0, 1, \dots, E\}, \mathcal{Q} \in \{0, 1, \dots, Q\} \right\} \quad (5.16)$$

where E is the maximum capacity of the energy storage and Q is the maximum data queue size. $\mathcal{C}$ is the composite channel state defined as follows:

$$\mathcal{C} = \left\{ (\mathcal{C}_1, \dots, \mathcal{C}_N); \mathcal{C}_n \in \{0, 1\} \right\}. \quad (5.17)$$

$\mathcal{C}_n$ is the state (i.e., status) of channel n . Its value is “0” if the channel is idle and “1” if the channel is busy. The state is then defined as a composite variable $\theta = ((c_1, \dots, c_N), e, q) \in \Theta$, where c_n is the status of channel n , e is the energy level of the energy storage, and q is the number of packets in the data queue of the secondary user, respectively. The action space of the secondary user is defined as follows: $\Delta = \{1, \dots, N\}$, where again the action δ is the channel selected for transmitting a packet or harvesting RF energy.

Transition probability matrix

We denote the transition probability matrix given action $\delta \in \Delta$ of the secondary user by $\mathbf{P}(\delta)$ whose element is obtained from $C_{(c_1, \dots, c_N), (c'_1, \dots, c'_N)} \mathbf{Q}(\delta, c_\delta)$. $C_{(c_1, \dots, c_N), (c'_1, \dots, c'_N)}$ is the probability that the status of all channels are $(c_1, \dots, c_N)$ in the current time slot and change to $(c'_1, \dots, c'_N)$ in the next time slot. This probability is the element of the matrix $\mathbf{C} = \mathbf{C}_1 \otimes \dots \otimes \mathbf{C}_n \otimes \dots \otimes \mathbf{C}_N$, where $\otimes$ is a Kronecker product and the matrix $\mathbf{C}_n$ is given in (3.1). Let $\mathbf{Q}(\delta, c_\delta)$ be the transition matrix of the queue and energy states when the status of the selected channel δ is $c_\delta \in \{0, 1\}$. This matrix can be expressed as follows:

$$\mathbf{Q}(\delta, c_\delta) = \begin{bmatrix} \mathbf{B}_{0,0}(\delta, c_\delta) & \mathbf{B}_{0,1}(\delta, c_\delta) & & & \\ \mathbf{B}_{1,0}(\delta, c_\delta) & \mathbf{B}_{1,1}(\delta, c_\delta) & \mathbf{B}_{1,2}(\delta, c_\delta) & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{B}_{Q,Q-1}(\delta, c_\delta) & \mathbf{B}_{Q,Q}(\delta, c_\delta) & \end{bmatrix} \begin{matrix} \leftarrow q = 0 \\ \leftarrow q = 1 \\ \vdots \\ \leftarrow q = Q \end{matrix} \quad (5.18)$$

where again each row of the matrix $\mathbf{Q}(\delta, c_\delta)$ corresponds to the number of packets in the data queue. Each row of the matrix $\mathbf{B}_{q,q'}(\delta)$ corresponds to the energy level in the energy storage.

The selected channel is idle $c_\delta = 0$: For $q = 0$, there is no packet transmission, since the data queue is empty. As a result, the energy level will always be the same (i.e., cannot transmit a packet and cannot harvest RF energy). Therefore, we have $\mathbf{B}_0(\delta, c_\delta = 0) = \mathbf{I}$. Then, $\mathbf{B}_{0,0}(\delta, c_\delta = 0) = \mathbf{B}_0(\delta, c_\delta = 0)\alpha^\circ$ and $\mathbf{B}_{0,1}(\delta, c_\delta = 0) = \mathbf{B}_0(\delta, c_\delta = 0)\alpha$, where $\alpha^\circ = 1 - \alpha$, for when there is no and there is a packet arrival, respectively.

For $q > 0$, when the number of packets in the data queue decreases, the transition probability matrix is expressed as follows:

$$\mathbf{B}_{q,q-1}(\delta, c_\delta = 0) = \begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ f_\delta^\circ \alpha^\circ \sigma_\delta & & & & \\ & \ddots & & & \\ & & f_\delta^\circ \alpha^\circ \sigma_\delta & \cdots & 0 \end{bmatrix} \begin{matrix} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = W - 1 \\ \leftarrow e = W \\ \vdots \\ \leftarrow e = E \end{matrix} \quad (5.19)$$

When the number of packets remains the same, the transition probability matrix $\mathbf{B}_{q,q}(\delta, c_\delta = 0)$ is expressed as follows:

$$\mathbf{B}_{q,q}(\delta, c_\delta = 0) = \left[\begin{array}{cccc} \alpha^\circ & & & \\ & \ddots & & \\ & & \alpha^\circ & \\ b_{W,0}^\circ(\delta) & \cdots & b_{W,W}^\circ(\delta) & \\ & \ddots & & \ddots \\ & & b_{E,E-W}^\circ(\delta) & \cdots & b_{E,E}^\circ(\delta) \end{array} \right] \begin{array}{l} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = W - 1 \\ \leftarrow e = W \\ \vdots \\ \leftarrow e = E \end{array} \quad (5.20)$$

where $b_{e,e-W}^\circ(\delta) = f_\delta^\circ(\sigma_\delta^\circ \alpha^\circ + \sigma_\delta \alpha)$ and $b_{e,e}^\circ(\delta) = f_\delta^\circ \alpha^\circ$. In this case, $b_{e,e-W}^\circ(\delta)$ is for when the secondary user transmits a packet, while $b_{e,e}^\circ(\delta)$ is for when the secondary user does not transmit a packet.

When the number of packets increases, the transition probability matrix $\mathbf{B}_{q,q+1}(\delta, c_\delta = 0)$ is expressed as follows:

$$\mathbf{B}_{q,q+1}(\delta, c_\delta = 0) = \left[\begin{array}{cccc} \alpha & & & \\ & \ddots & & \\ & & \alpha & \\ b_{W,0}^+(\delta) & \cdots & b_{W,W}^+(\delta) & \\ & \ddots & & \ddots \\ & & b_{E,E-W}^+(\delta) & \cdots & b_{E,E}^+(\delta) \end{array} \right] \begin{array}{l} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = W - 1 \\ \leftarrow e = W \\ \vdots \\ \leftarrow e = E \end{array} \quad (5.21)$$

where $b_{e,e-W}^+(\delta) = f_\delta^\circ \sigma_\delta^\circ \alpha$ and $b_{e,e}^+(\delta) = f_\delta \alpha$. Again, $b_{e,e-W}^+(\delta)$ is for when the secondary user transmits a packet (but not successful), while $b_{e,e}^+(\delta)$ is for when the secondary user does not transmit a packet.

The selected channel is busy $c_\delta = 1$: For $q = 0$, there is no packet transmission. As a result, the energy level will never decrease. Let $\mathbf{B}_0(\delta, c_\delta = 1)$ denote a common matrix for $q = 0$. We have

$$\mathbf{B}_0(\delta, c_\delta = 1) = \left[\begin{array}{cccc} 1 - m_\delta^\circ \gamma_\delta & m_\delta^\circ \gamma_\delta & & \\ & \ddots & & \\ & & 1 - m_\delta^\circ \gamma_\delta & m_\delta^\circ \gamma_\delta \\ & & & 1 \end{array} \right] \begin{array}{l} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = E - 1 \\ \leftarrow e = E \end{array} \quad (5.22)$$

Then, $\mathbf{B}_{0,0}(\delta, c_\delta = 1) = \mathbf{B}_0(\delta, c_\delta = 1) \alpha^\circ$ and $\mathbf{B}_{0,1}(\delta, c_\delta = 1) = \mathbf{B}_0(\delta, c_\delta = 1) \alpha$ for when there is no and there is a packet arrival, respectively.

For $q > 0$, there are only two sub-cases, i.e., the number of packets remains the same or increases (i.e., no successful transmission if the channel is busy). When the

number of packets remains the same, the transition probability matrix is expressed as follows:

$$\begin{aligned} \mathbf{B}_{q,q}(\delta, c_\delta = 1) &= \\ &= \begin{bmatrix} \alpha^\circ(1 - m_\delta^\circ \gamma_\delta) & \alpha^\circ m_\delta^\circ \gamma_\delta & & & \\ & \ddots & \ddots & & \\ & & \alpha^\circ(1 - m_\delta^\circ \gamma_\delta) & \alpha^\circ m_\delta^\circ \gamma_\delta & \\ b_{W,0}^\circ(\delta) & \cdots & b_{W,W}^\circ(\delta) & b_{W,W+1}^\circ(\delta) & \\ & \ddots & \ddots & \ddots & \\ & & b_{E,E-W}^\circ(\delta) & \cdots & b_{E,E}^\circ(\delta) \end{bmatrix} \begin{matrix} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = W - 1 \\ \leftarrow e = W \\ \vdots \\ \leftarrow e = E \end{matrix} \end{aligned} \quad (5.23)$$

where $b_{e,e-W}^\circ(\delta) = m_\delta \alpha^\circ$, $b_{e,e}^\circ(\delta) = m_\delta^\circ \gamma_\delta^\circ \alpha^\circ$, and $b_{e,e+1}^\circ(\delta) = m_\delta^\circ \gamma_\delta \alpha^\circ$. In this case, $b_{e,e-W}^\circ(\delta)$ is for when the secondary user transmits a packet (i.e., due to miss detection¹), $b_{e,e}^\circ(\delta)$ is for when the secondary user unsuccessfully harvests RF energy, and $b_{e,e+1}^\circ(\delta)$ is for when the secondary user successfully harvests RF energy.

When the number of packets increases, the transition probability matrix $\mathbf{B}_{q,q+1}(\delta, c_\delta = 1)$ is expressed as follows:

$$\begin{aligned} \mathbf{B}_{q,q+1}(\delta, c_\delta = 1) &= \\ &= \begin{bmatrix} \alpha(1 - m_\delta^\circ \gamma_\delta) & \alpha m_\delta^\circ \gamma_\delta & & & \\ & \ddots & \ddots & & \\ & & \alpha(1 - m_\delta^\circ \gamma_\delta) & \alpha m_\delta^\circ \gamma_\delta & \\ b_{W,0}^+(\delta) & \cdots & b_{W,W}^+(\delta) & b_{W,W+1}^+(\delta) & \\ & \ddots & \ddots & \ddots & \\ & & b_{E,E-W}^+(\delta) & \cdots & b_{E,E}^+(\delta) \end{bmatrix} \begin{matrix} \leftarrow e = 0 \\ \vdots \\ \leftarrow e = W - 1 \\ \leftarrow e = W \\ \vdots \\ \leftarrow e = E \end{matrix} \end{aligned} \quad (5.24)$$

where $b_{e,e-W}^+(\delta) = m_\delta \alpha$, $b_{e,e}^+(\delta) = m_\delta^\circ \gamma_\delta^\circ \alpha$, and $b_{e,e+1}^\circ(\delta) = m_\delta^\circ \gamma_\delta \alpha$. These elements are similar to those in the case when the number of packets in the data queue remains the same.

Optimization formulation

The optimization can be formulated with the same objective as that in (3.8) to maximize the throughput of the secondary user. In this case, the state is defined as $\theta \in \Theta$,

¹Here, we note that the term “complete channel information” is defined from the perspective of the secondary user. However, the secondary user still needs to sense the target channel to check the real channel status. This is in order to ensure that its transmission will not cause interference to primary user transmissions. Consequently, the sensing error probabilities are still applied to the secondary user.

where $\theta = ((c_1, \dots, c_N), e, q)$. The immediate throughput function in the complete information case is defined as follows:

$$\mathcal{T}(\theta, \delta) = \begin{cases} f_\delta^\circ \sigma_\delta, & (e \geq W) \text{ and } (q > 0) \text{ and } (c_\delta = 0) \\ 0 & \text{otherwise.} \end{cases} \quad (5.25)$$

The secondary user successfully transmits a packet if the selected channel is idle, there is enough energy, the queue is not empty.

The similar method of transforming and solving the equivalent LP problem is applied to obtain the optimal channel section policy for the complete information case.

Performance measures

Let $\zeta^*(((c_1, \dots, c_N), e, q), \delta)$ denote the optimal solution of the equivalent LP problem for the optimization of the complete information case. The following performance measures can be obtained.

Average number of packets in the data queue is obtained from

$$\bar{q} = \sum_{c_1=0}^1 \cdots \sum_{c_N=0}^1 \left(\sum_{\delta \in \Delta} \sum_{q=0}^Q \sum_{e=0}^E q \zeta^*(((c_1, \dots, c_N), e, q), \delta) \right). \quad (5.26)$$

Throughput is obtained from

$$\tau = \sum_{c_1=0}^1 \cdots \sum_{c_N=0}^1 \left(\sum_{\delta \in \Delta} \sum_{q=1}^Q \sum_{e=W}^E \eta_\delta f_\delta^\circ \sigma_\delta \zeta^*(((c_1, \dots, c_N), e, q), \delta) \right). \quad (5.27)$$

Average delay can be obtained using Little's law as

$$\bar{d} = \frac{\bar{q}}{\tau}. \quad (5.28)$$

Appendix D

The proof of Proposition 4.2

From (4.32), we have

$$\mathcal{L}(\Theta, \gamma) = \sum_{s \in \mathcal{S}} \pi_s(\Theta) \mathcal{G}(s, \Theta), \quad (5.29)$$

where $\pi_s(\Theta)$ is the steady state probability, $\mathcal{L}(\Theta, \gamma)$ is the Lagrange function and $\mathcal{G}(s, \Theta)$ is the value of Lagrange function at state s . All are parameterized by parameter vector Θ .

Then, we take gradient for the Lagrange function $\mathcal{L}(\Theta, \gamma)$ with respect to the vector Θ as follows

$$\begin{aligned} \nabla_{\Theta} \mathcal{L}(\Theta, \gamma) &= \sum_{s \in \mathcal{S}} \pi_s(\Theta) \nabla_{\Theta} \mathcal{G}(s, \Theta) + \sum_{s \in \mathcal{S}} \nabla_{\Theta} \pi_s(\Theta) \mathcal{G}(s, \Theta), \\ &= \sum_{s \in \mathcal{S}} \pi_s(\Theta) \nabla_{\Theta} \mathcal{G}(s, \Theta) + \\ &\quad \sum_{s \in \mathcal{S}} \nabla_{\Theta} \pi_s(\Theta) \mathcal{G}(s, \Theta) - \mathcal{L}(\Theta, \gamma) \sum_{s \in \mathcal{S}} \nabla_{\Theta} \pi_s(\Theta), \\ &= \sum_{s \in \mathcal{S}} \pi_s(\Theta) \nabla_{\Theta} \mathcal{G}(s, \Theta) + \\ &\quad \sum_{s \in \mathcal{S}} \nabla_{\Theta} \pi_s(\Theta) \left(\mathcal{G}(s, \Theta) - \mathcal{L}(\Theta, \gamma) \right), \end{aligned} \quad (5.30)$$

We define the differential cost at state s by

$$d(s, \Theta) = \mathbb{E}_{\Psi(\Theta)} \left[\sum_{t=0}^{\mathbf{T}-1} (\mathcal{G}(s, \Theta) - \mathcal{L}(\Theta, \gamma)) \mid s(0) = s \right], \quad (5.31)$$

where $\mathbf{T} = \min\{t > 0 \mid s(t) = s^{\dagger}\}$ is the first future time that state $s^{\dagger}$ is visited. The main objective of defining $d(s, \Theta)$ is to express the relation between the average value of the Lagrange function and the immediate value of the Lagrange function at state s . Furthermore, $d(s, \Theta)$ is a unique solution of the following Bellman equation,

$$d(s, \Theta) = \mathcal{G}(s, \Theta) - \mathcal{L}(\Theta, \gamma) + \sum_{s' \in \mathcal{S}} \mathbf{p}(s' \mid s, \Psi(\Theta)) d(s', \Theta). \quad (5.32)$$

$$\begin{aligned}
 \nabla_{\Theta} \mathcal{L}(\Theta, \gamma) &= \sum_{s \in \mathcal{S}} \pi_s(\Theta) \nabla_{\Theta} \mathcal{G}(s, \Theta) + \sum_{s \in \mathcal{S}} \nabla_{\Theta} \pi_s(\Theta) \left(d(s, \Theta) - \sum_{s' \in \mathcal{S}} \mathbf{p}(s'|s, \Psi(\Theta)) d(s', \Theta) \right), \\
 &= \sum_{s \in \mathcal{S}} \pi_s(\Theta) \nabla_{\Theta} \mathcal{G}(s, \Theta) + \sum_{s \in \mathcal{S}} \nabla_{\Theta} \pi_s(\Theta) d(s, \Theta) \\
 &\quad + \sum_{s, s' \in \mathcal{S}} \left(\pi_s(\Theta) \nabla_{\Theta} \mathbf{p}(s'|s, \Psi(\Theta)) - \nabla_{\Theta} \left(\pi_s(\Theta) \mathbf{p}(s'|s, \Psi(\Theta)) \right) \right) d(s', \Theta), \\
 &= \sum_{s \in \mathcal{S}} \pi_s(\Theta) \nabla_{\Theta} \mathcal{G}(s, \Theta) + \sum_{s \in \mathcal{S}} \nabla_{\Theta} \pi_s(\Theta) d(s, \Theta) \\
 &\quad + \sum_{s, s' \in \mathcal{S}} \pi_s(\Theta) \nabla_{\Theta} \mathbf{p}(s'|s, \Psi(\Theta)) d(s', \Theta) - \sum_{s' \in \mathcal{S}} \nabla_{\Theta} \left(\sum_{s \in \mathcal{S}} \pi_s(\Theta) \mathbf{p}(s'|s, \Psi(\Theta)) \right) d(s', \Theta), \\
 &= \sum_{s \in \mathcal{S}} \pi_s(\Theta) \nabla_{\Theta} \mathcal{G}(s, \Theta) + \sum_{s \in \mathcal{S}} \nabla_{\Theta} \pi_s(\Theta) d(s, \Theta) + \sum_{s, s' \in \mathcal{S}} \pi_s(\Theta) \nabla_{\Theta} \mathbf{p}(s'|s, \Psi(\Theta)) d(s', \Theta) \\
 &\quad - \sum_{s' \in \mathcal{S}} \nabla_{\Theta} \pi_{s'}(\Theta) d(s', \Theta) \\
 &= \sum_{s \in \mathcal{S}} \pi_s(\Theta) \left(\nabla_{\Theta} \mathcal{G}(s, \Theta) + \sum_{s' \in \mathcal{S}} \nabla_{\Theta} \mathbf{p}(s'|s, \Psi(\Theta)) d(s', \Theta) \right).
 \end{aligned} \tag{5.33}$$

Then, we replace $\mathcal{G}(s, \Theta) - \mathcal{L}(\Theta, \gamma)$ in (5.30) by $d(s, \Theta) - \sum_{s' \in \mathcal{S}} \mathbf{p}(s'|s, \Psi(\Theta)) d(s', \Theta)$ in (5.32) and have the following results:

From (4.30), we have $\mathcal{G}(s, \Theta) = \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) \mathcal{G}(s, a)$ and we take gradient for function $\mathcal{G}(s, \Theta)$ under the parameter vector Θ_n as follows:

$$\begin{aligned}
 \nabla_{\Theta_n} \mathcal{G}(s, \Theta) &= \sum_{a \in \mathcal{A}} \nabla_{\Theta_n} \mu_{\Theta}(s, a) \mathcal{G}(s, a), \\
 &= \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) \frac{\nabla_{\Theta_n} \mu_{\Theta_n}(s_n, a)}{\mu_{\Theta_n}(s_n, a)} \mathcal{G}(s, a) \\
 &= \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) \frac{\nabla_{\Theta_n} \mu_{\Theta_n}(s_n, a)}{\mu_{\Theta_n}(s_n, a)} \left(\mathcal{G}(s, a) - \mathcal{L}(\Theta, \gamma) \right)
 \end{aligned} \tag{5.34}$$

From (4.24), we have $\mathbf{p}(s'|s, \Psi(\Theta)) = \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) \mathbf{p}(s'|s, a)$, and we also take gradient for function $\mathbf{p}(s'|s, \Psi(\Theta))$ under the parameter vector Θ_n as follows:

$$\nabla_{\Theta_n} \mathbf{p}(s'|s, \Psi(\Theta)) = \sum_{a \in \mathcal{A}} \mu_{\Theta}(s, a) \frac{\nabla_{\Theta_n} \mu_{\Theta_n}(s_n, a)}{\mu_{\Theta_n}(s_n, a)} \mathbf{p}(s'|s, a). \tag{5.35}$$

Then, we replace (5.34) and (5.35), into (5.33), and we derive the following results:

$$\begin{aligned}
 \nabla_{\Theta_n} \mathcal{L}(\Theta, \gamma) &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi_s(\Theta) \mu_{\Theta}(s, a) \frac{\nabla_{\Theta_n} \mu_{\Theta_n}(s_n, a_n)}{\mu_{\Theta_n}(s_n, a_n)} \\
 &\quad \left(\mathcal{G}(s, a) - \mathcal{L}(\Theta, \gamma) + \sum_{s' \in \mathcal{S}} \mathbf{p}(s'|s, a) d(s', \Theta) \right), \\
 &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi_s(\Theta) \mu_{\Theta}(s, a) \frac{\nabla_{\Theta_n} \mu_{\Theta_n}(s_n, a_n)}{\mu_{\Theta_n}(s_n, a_n)} q(s, a, \Theta),
 \end{aligned} \tag{5.36}$$

where

$$\begin{aligned}
 q(s, a, \Theta) &= \mathcal{G}(s, a) - \mathcal{L}(\Theta, \gamma) + \sum_{s' \in \mathcal{S}} \mathbf{p}(s'|s, a) d(s', \Theta), \\
 &= \mathbb{E}_{\Psi(\Theta)} \left[\sum_{t=0}^{\mathbf{T}-1} \left(\mathcal{G}(s(t), a(t)) - \mathcal{L}(\Theta, \gamma) \right) \middle| s(0) = s, a(0) = a \right].
 \end{aligned} \tag{5.37}$$

The proof now is completed.

Appendix E

The convergence proof of Algorithm 2 in Chapter 4

We analyze and then derive the proof for the convergence of the Algorithm 2 presented in Section 4.3. In Algorithm 2, the parameter vectors and the Lagrange multipliers are updated at two different time scales, namely, $\alpha(t)$ and $\beta(t)$, respectively. At the time scale $\beta(t)$, the value of the Lagrange multipliers are updated by a function of $\beta(t)$ as follows:

$$\gamma_n^{t+1} = \gamma_n^t + \mathcal{F}(\beta(t)) \quad \forall n. \quad (5.38)$$

Furthermore, under the Assumption 4.5, time scale $\beta(t)$ can be represented as a function of $\alpha(t)$ as follows:

$$\gamma_n^{t+1} = \gamma_n^t + \mathcal{F}(\beta(t)) = \gamma_n^t + f(\alpha(t)) \quad \forall n. \quad (5.39)$$

Consequently, we can conclude that (4.38) views (4.41) as quasi-static (i.e., almost a constant), while (4.41) views (4.38) as almost equilibrated (as shown in Chapter 6 [132]). Therefore, we have to establish the proof of convergence for Algorithm 2 at different time scales separately.

We will first prove the convergence of the algorithm in the first time scale (i.e., $\alpha(t)$). In other words, we will prove that $\nabla_{\Theta} \mathcal{L}(\Theta^\infty(\gamma), \gamma) = 0$.

Proof: We define vector $\mathbf{r}^t = \begin{bmatrix} \Theta^t & \tilde{\mathcal{L}}^t \end{bmatrix}^\top$. Then, the update equations as given in (4.38) and (4.39) of the Algorithm 2 can be rewritten in the following form,

$$\mathbf{r}^{t+1} = \mathbf{r}^t + \alpha(t)\mathbf{H}^t, \quad (5.40)$$

where

$$\mathbf{H}^t = \begin{bmatrix} -(\mathcal{I}_G - \tilde{\mathcal{L}}^t)z_n^t \\ -(\mathcal{I}_G - \tilde{\mathcal{L}}^t) \end{bmatrix}. \quad (5.41)$$

Let t_m be the time when the recurrent state $s^\dagger$ is revisited at m -th time and thus (5.40) can be presented as follows:

$$\mathbf{r}^{t_{m+1}} = \mathbf{r}^{t_m} + \sum_{t=t_m}^{t_{m+1}-1} \alpha(t) \mathbf{H}^t. \quad (5.42)$$

If we denote $\alpha(m) = \sum_{t=t_m}^{t_{m+1}-1} \alpha(t)$, then we have $\epsilon_m = \sum_{t=t_m}^{t_{m+1}-1} \alpha(t) (\mathbf{H}^t - \mathbf{h}(\mathbf{r}^{t_m}))$ where

$$\mathbf{h}(\mathbf{r}^{t_m}) = \begin{bmatrix} -\mathbb{E}_\Theta[T] \nabla_\Theta \mathcal{L}(\Theta) - \mathcal{V}(\Theta) (\mathcal{L}(\Theta) - \tilde{\mathcal{L}}(\Theta)) \\ -\mathbb{E}_\Theta[T] (\mathcal{L}(\Theta) - \tilde{\mathcal{L}}(\Theta)) \end{bmatrix}, \quad (5.43)$$

for $\mathcal{V}(\Theta) = [\mathcal{V}_1(\Theta), \dots, \mathcal{V}_N(\Theta)]$ and

$$\mathcal{V}_n(\Theta) = \mathbb{E}_\Theta \left[\sum_{t'=t_m+1}^{t_{m+1}-1} (t_{m+1} - t') \frac{\nabla_{\Theta_n} \mu_{\Theta_n}(s_{t'}, a_{t'})}{\mu_{\Theta_n}(s_{t'}, a_{t'})} \right].$$

Then, (5.42) becomes

$$\mathbf{r}^{t_{m+1}} = \mathbf{r}^{t_m} + \alpha(m) \mathbf{h}(\mathbf{r}^{t_m}) + \epsilon_m. \quad (5.44)$$

Next, we will show that

$$\sum_{m=1}^{\infty} \alpha(m) = \infty, \text{ and } \sum_{m=1}^{\infty} \alpha^2(m) < \infty \text{ with probability one.} \quad (5.45)$$

From Assumption 4.5, we have $\sum_{m=1}^{\infty} \alpha(m) = \sum_{t=1}^{\infty} \alpha(t) = \infty$. Additionally, under Assumption 4.5, the sequence $\alpha(t)$ is non-increasing, and thus, $\alpha(m) = \sum_{t=t_m}^{t_{m+1}-1} \alpha(t) \leq \alpha(t_m)(t_{m+1} - t_m)$. Then, we derive $\sum_{m=1}^{\infty} \alpha^2(m) \leq \sum_{m=1}^{\infty} \alpha^2(t_m)(t_{m+1} - t_m)^2 < \sum_{t=1}^{\infty} \alpha^2(t) < \infty$.

Now, we can transform the update equations into the same form as presented in (25) in [125]. By using Lemma 2 and Lemma 3 in [125], we have the sequence $\mathbf{h}(\mathbf{r}^{t_m})$ bounded and the sequence $\sum_{m=1}^{\infty} \epsilon_m$ converges almost surely. As a result, we can conclude that

$$\lim_{m \rightarrow \infty} (\mathbf{r}^{t_{m+1}} - \mathbf{r}^{t_m}) = 0 \text{ with probability one.} \quad (5.46)$$

After that, based on Lemma 11 and Section C of Appendix 1 in [125], we can prove that the sequence $\mathcal{L}(\Theta^{t_m})$ and $\tilde{\mathcal{L}}(\Theta^{t_m})$ converge to a common limit, and thus, $\nabla_\Theta \mathcal{L}(\Theta^{t_m})$ converges to zero, i.e., $\nabla_\Theta \mathcal{L}(\Theta^\infty) = 0$. The proof for the convergence in the first time scale now is completed.

Then, we will prove the convergence of the algorithm in the second time scale (i.e., $\beta(t)$). It means, we will prove that $\lim_{t \rightarrow \infty} \gamma^t = \gamma^\infty$ with probability 1 and γ^∞ satisfies the constraints in (4.27).

Proof: From Assumption 4.5, since the time scale $\beta(t)$ is very small compared with the time scale $\alpha(t)$, this makes γ^t quasi-static compared with Θ^t and this also has an effect similar to fixing γ^t and running (4.41) to infinity. In turn, γ^t views Θ^t as a converged approximation to $\Theta^*(\gamma^t)$ [136], and thus, we can rewrite the equation in (4.41) as follows:

$$\gamma_n^{t+1} = \max \left(\gamma_n^t + \beta(t) \left(\overline{\mathcal{Q}}_n(\Theta^*(\gamma)) - \mathcal{Q}_n^* + \mathcal{Q}_n(t) - \overline{\mathcal{Q}}_n(\Theta^*(\gamma)) \right), 0 \right). \quad (5.47)$$

We denote $w_n^{t+1} = \mathcal{Q}_n(t) - \overline{\mathcal{Q}}_n(\Theta^*(\gamma))$ and let $\mathcal{F} \triangleq \sigma(\gamma^l, w_n^l, l \leq t)$ be the σ -algebra generated by $\{\gamma_n^l, w_n^l, l \leq t\}$. We can see that the sequence $\{w_n^t\}$ is a martingale difference sequence since its expectation with regard to the past is zero, i.e., $\mathbb{E}[w_n^{t+1} | \mathcal{F}^t] = 0$. Moreover, we can always find an appropriate constant K such that $\mathbb{E}[|w_n^{t+1}|^2 | \mathcal{F}^t] \leq K(1 + |\gamma_n^t|^2)$. Hence, by using the standard stochastic approximation in [132], the Lagrange update equation for secondary users n can be represented by the ordinary differential equation (ODE) as follows:

$$\dot{\gamma}_n^t = \overline{\mathcal{Q}}_n(\Theta^*(\gamma^t)) - \mathcal{Q}_n^*. \quad (5.48)$$

We define

$$\mathcal{J}(\gamma) = \sum_{n=1}^N \left(\overline{\mathcal{C}}_n(\Theta^*(\gamma)) + \gamma_n (\overline{\mathcal{Q}}_n(\Theta^*(\gamma)) - \mathcal{Q}_n^*) \right). \quad (5.49)$$

Then, based on the chain rule, we have

$$\frac{\partial \mathcal{J}(\gamma)}{\partial \gamma_n} = \overline{\mathcal{Q}}_n(\Theta^*(\gamma)) - \mathcal{Q}_n^*, \quad (5.50)$$

and thus

$$\dot{\gamma}_n^t = \frac{\partial \mathcal{J}(\gamma)}{\partial \gamma_n}. \quad (5.51)$$

Consequently, from Proposition 3 in [135], we have $\lim_{t \rightarrow \infty} \frac{\partial \mathcal{J}(\gamma)}{\partial \gamma_n} = 0$. In other words, $\overline{\mathcal{Q}}_n(\Theta^*(\gamma^\infty)) - \mathcal{Q}_n^* = 0$ which satisfies the constraints in (4.27).

Now, the proof of the convergence of Algorithm 2 is completed.

Appendix F

Complexity analysis of the online learning algorithms

In this section, we will show that our proposed online learning algorithms are very efficient in using resource in terms of time and storage. This is due to the fact that the state and action variables in our algorithms can be updated iteratively in an online fashion without storing and using any information from history. We will analyze the resource consumption for Algorithm 1 (Chapter 4) in detail as follows.

Algorithm 1 Algorithm to update Θ at every time step

At time step k , the state is s_k , and the values of Θ_k , z_k , and $\tilde{\psi}(\Theta_k)$ are available from the previous iteration. We update z_k , Θ_k , and $\tilde{\psi}$ according to:

$$z_{k+1} = \begin{cases} \frac{\nabla \mu_{\Theta_k}(s_k, a_k)}{\mu_{\Theta_k}(s_k, a_k)}, & \text{if } s_k = s^* \\ z_k + \frac{\nabla \mu_{\Theta_k}(s_k, a_k)}{\mu_{\Theta_k}(s_k, a_k)}, & \text{otherwise,} \end{cases} \quad (4.14)$$

$$\Theta_{k+1} = \Theta_k + \rho_k (\mathcal{T}(s_k, a_k) - \tilde{\psi}_k) z_{k+1}, \quad (4.15)$$

$$\tilde{\psi}_{k+1} = \tilde{\psi}_k + \kappa \rho_k (\mathcal{T}(s_k, a_k) - \tilde{\psi}_k). \quad (4.16)$$

First, for the storage complexity, in Algorithm 1, we just need to store the values of three variables at each step, i.e., Θ_k , z_k , and $\tilde{\psi}(\Theta_k)$, where Θ_k is the parameter vector of the user that we need to optimize its performance, z_k is an auxiliary variable used to compute the value of $\frac{\nabla \mu_{\Theta_k}(s_k, a_k)}{\mu_{\Theta_k}(s_k, a_k)}$ before it is used to update the value of Θ_k , and $\tilde{\psi}(\Theta_k)$ is the estimated value of the average throughput. Because Algorithm 1 works in an online fashion, these variables will be updated at each step and we do not need to store all other values in the past. Thus, we can avoid the curse-of-storage which happens in the algorithms using values in history to make decisions.

Second, for the time complexity, each user needs to do four steps at each time slot. Specifically, in the first step, based on the value of Θ in the previous step, the user calculates the value of $\mu_{\Theta}(s, a)$ as in equation (4.3) (i.e., $\mu_{\Theta}(s, a) = \frac{\exp(\theta_{s,a})}{\sum_{a_i \in \mathcal{A}} \exp(\theta_{s,a_i})}$)

and decides which channel to sense. In the second step, the user computes the value of $\frac{\nabla \mu_{\Theta_k}(s_k, a_k)}{\mu_{\Theta_k}(s_k, a_k)}$ and updates the value of z as in equation (4.14). In the last two steps, the user updates the values of Θ and $\tilde{\psi}(\Theta_k)$ as in equations (4.15) and (4.16), respectively. Additionally, we note a very interesting point here for the calculation of equation (4.14). Because of the special structure of $\mu_{\Theta}(s, a)$ as shown in equation (4.3), instead of calculating the value of $\frac{\nabla \mu_{\Theta_k}(s_k, a_k)}{\mu_{\Theta_k}(s_k, a_k)}$ directly, we can use some mathematical manipulation to transform it into the equivalent form by $1 - \mu_{\Theta}(s, a)$. This can reduce the computation time considerably. Therefore, as we can see from the aforementioned analysis, at each computing step, the user just needs to perform simple calculations without any complex functions, and thus the time complexity at each time slot is negligible.

We then can apply the similar analysis to Algorithm 2 in Chapter 3, and Algorithm 2 in Chapter 4. In summary, we can confirm the high efficiency of the proposed algorithms in term of time and storage complexity.

References

- [1] M. A. Alsheikh, D. T. Hoang, D. Niyato, H.-P. Tan, and S. Lin, “Markov decision processes with applications in wireless sensor networks: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1239–1267, April 2015.
- [2] A.-H. Carles and M. Dohler, *Machine-to-Machine (M2M) communications: Architecture, performance and applications*. Woodhead Publishing, Limited, 2015.
- [3] S. Movassaghi, M. Abolhasan, J. Lipman, D. Smith, and A. Jamalipour, “Wireless body area networks: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1658–1686, January 2014.
- [4] D. T. Hoang, C. Lee, D. Niyato, and P. Wang, “A survey of mobile cloud computing: Architecture, applications, and approaches,” *Wireless Communications and Mobile Computing*, vol. 13, no. 18, pp. 1587–1611, December 2013.
- [5] ABI. (2010) The internet of things will drive wireless connected devices to 40.9 billion in 2020. [Online]. Available: <https://www.abiresearch.com/press/the-internet-of-things-will-drive-wireless-connect/>
- [6] E. FCC, “Docket no. 03-322, notice of proposed rule making and order,” December 2003.
- [7] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,” *IEEE Journal of Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, February 2005.
- [8] A. Khattab, D. Perkins, and M. Bayoumi, *Cognitive Radio Networks: From Theory to Practice*. Springer Science & Business Media, 2012.
- [9] E. FCC, “Et docket no. 04-186, unlicensed operation in the tv broadcast bands,” November 2011.

- [10] IEEE, “Ieee 802.22 working group on wireless regional area networks,” November 2011.
- [11] Z. K. Alenezi, “Practical applications of cognitive radio,” Ph.D. dissertation, Southern Methodist University, 2012.
- [12] E. Hossain, D. Niyato, and Z. Han, *Dynamic Spectrum Access and Management in Cognitive Radio Networks*. Cambridge University Press, 2009.
- [13] J. I. Agbinya, *Wireless power transfer*. River Publishers, 2012.
- [14] S. Bi, C. K. Ho, and R. Zhang, “Wireless powered communication: opportunities and challenges,” *IEEE Communications Magazine*, vol. 53, no. 4, pp. 117–125, April 2015.
- [15] Samsung. (2015) Fast charge wireless charging pad. [Online]. Available: <http://www.samsung.com/us/mobile/cell-phones-accessories/EP-PN920TBEGUS>
- [16] K. Van Schuylenbergh and R. Puers, *Inductive powering: Basic theory and application to biomedical systems*. Springer Science & Business Media, 2009.
- [17] X. Wei, Z. Wang, and H. Dai, “A critical review of wireless power transfer via strongly coupled magnetic resonances,” *Energies*, vol. 7, no. 7, pp. 4316–4341, July 2014.
- [18] A. Karalis, J. D. Joannopoulos, and M. Soljačić, “Efficient wireless non-radiative mid-range energy transfer,” *Annals of Physics*, vol. 323, no. 1, pp. 34–48, January 2008.
- [19] A. Kurs, A. Karalis, R. Moffatt, J. D. Joannopoulos, P. Fisher, and M. Soljačić, “Wireless power transfer via strongly coupled magnetic resonances,” *Science*, vol. 317, no. 5834, pp. 83–86, July 2007.
- [20] J. C. Maxwell, *A treatise on electricity and magnetism*. Clarendon press, 1881, vol. 1.
- [21] S. Ladan, N. Ghassemi, A. Ghiotto, and K. Wu, “Highly efficient compact rectenna for wireless energy harvesting application,” *IEEE Microwave Magazine*, vol. 14, no. 1, pp. 117–122, January 2013.

- [22] V. Kuhn, C. Lahuec, F. Seguin, and C. Person, “A multi-band stacked rf energy harvester with rf-to-dc efficiency up to 84%,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 5, pp. 1768–1778, May 2015.
- [23] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ: Wiley, 1994.
- [24] R. Bellman, *Dynamic programming*. Princeton University Press, 1957.
- [25] F. D. Epenoux, “A probabilistic production and inventory problem,” *Management Science*, vol. 10, no. 1, pp. 98–108, October 1963.
- [26] E. Altman, *Constrained Markov decision processes*. Chapman and Hall/CRC Press, 1999.
- [27] W. B. Powell, *Approximate dynamic programming: Solving the curses of dimensionality*. Wiley-Interscience, 2011.
- [28] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. MIT Press Cambridge, 1998.
- [29] D. P. D. Farias and B. V. Roy, “The linear programming approach to approximate dynamic programming,” *Operations Research*, vol. 51, no. 6, pp. 850–865, December 2003.
- [30] O. Sigaud and O. Buffet, *Markov decision processes in artificial intelligent: MDPs, beyond MDPs and applications*. Wiley-IEEE Press, 2010.
- [31] R. G. Gallager, *Discrete stochastic processes*. Kluwer Academic Publisher, 1995.
- [32] X. Guo and O. Hernandez-Lerma, *Continuous-time Markov decision processes: Theory and applications*. Springer, 2009.
- [33] G. E. Monahan, “State of the art - A survey of partially observable Markov decision processes: Theory, models, and algorithms,” *Management Science*, vol. 28, no. 1, pp. 1–16, January 1982.
- [34] I. Chelsea C. White, “A survey of solution techniques for the partially observed markov decision process,” *Journal of Annals of Operations Research*, vol. 32, no. 1, pp. 215–230, December 1991.

- [35] W. S. L., “A survey of algorithmic methods for partially observed Markov decision processes,” *Journal of Annals of Operations Research*, vol. 28, no. 1, pp. 47–65, April 1991.
- [36] A. M. Brooks, “Parametric POMDPs for planning in continuous state spaces,” Ph.D. dissertation, School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, 2007.
- [37] C. Boutilier, “Sequential optimality and coordination in multiagent systems,” in *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, August 1999, pp. 478–485.
- [38] D. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, “The complexity of decentralized control of markov decision processes,” *Journal of Mathematics of Operations Research*, vol. 27, no. 4, pp. 819–840, November 2002.
- [39] F. A. Oliehoek, “Value-based planning for teams of agents in stochastic partially observable environments,” Ph.D. dissertation, Informatics Institute, University of Amsterdam, 2010.
- [40] C. Amato, G. Chowdhary, A. Geramifard, N. K. Ure, and M. J. Kochenderfer, “Decentralized control of partially observable Markov decision processes,” in *Proceedings of the IEEE 52nd Annual Conference on Decision and Control*, December 2013, pp. 2398–2405.
- [41] L. S. Shapley, “Stochastic games,” in *Proceedings of the National Academy of Science USA*, October 1953, pp. 1095–1100.
- [42] A. Neyman and S. Sorin, *Stochastic games and applications*. Springer-Verlag, 2003.
- [43] E. A. Hansen, D. S. Bernstein, and S. Zilberstein, “Dynamic programming for partially observable stochastic games,” in *Proceedings of the 19th National Conference on Artificial Intelligence*, July 2004, pp. 709–715.
- [44] H. P. Christos and J. N. Tsitsiklis, “The complexity of markov decision processes,” *Journal of Mathematics of Operations Research*, vol. 12, no. 3, pp. 441–450, August 1987.

- [45] M. L. Littman, T. L. Dean, and L. P. Kaelbling, "On the complexity of solving Markov decision problems," in *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, June 1995, pp. 394–402.
- [46] J. Goldsmith and M. Mundhenk, "Complexity issues in markov decision processes," in *Proceedings of the IEEE conference on Computational Complexity*, June 1998, pp. 272–280.
- [47] O. Madani, S. Hanks, and A. Condon, "On the undecidability of probabilistic planning and related stochastic optimization problems," *Journal of Artificial Intelligence*, vol. 147, no. 1, pp. 5–34, July 2003.
- [48] D. J. White, "A survey of applications of markov decision processes," *The Journal of the Operational Research Society*, vol. 44, no. 11, pp. 1073–1096, November 1993.
- [49] E. Altman, "Applications of markov decision processes in communication networks," in *Handbook of Markov decision processes*. Springer, 2002, pp. 489–536.
- [50] D. J. White, "Real applications of markov decision processes," *Interfaces*, vol. 15, no. 6, pp. 73–83, December 1985.
- [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, September 2013.
- [52] J. F. C. Kingman, *Poisson Processes*. Oxford University Press, 1993.
- [53] Y. Yao, X. Song, C. Yin, and S. Huang, "Opportunistic energy harvesting and energy-based opportunistic spectrum access in cognitive radio networks," in *Cognitive Radio Oriented Wireless Networks*. Springer International Publishing, 2015, pp. 187–198.
- [54] S. Park, J. Heo, B. Kim, and W. Chung, "Optimal mode selection for cognitive radio sensor networks with rf energy harvesting," in *IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, September 2012, pp. 2155–2159.
- [55] 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, March 2013.

- [56] S. Park and D. Hong, “Achievable throughput of energy harvesting cognitive radio networks,” *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 1010–1022, February 2014.
- [57] —, “Optimal spectrum access for energy harvesting cognitive radio networks,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 12, pp. 6166–6179, December 2013.
- [58] R. E. Ali, F. F. Digham, K. G. Seddik, M. Nafie, A. KEl-Keyi, and Z. Han, “A probabilistic mac for cognitive radio systems with energy harvesting nodes,” in *IEEE Global Conference on Signal and Information Processing*, December 2014, pp. 143–147.
- [59] A. A. Alkheir and H. T. Mouftah, “Cooperative spectrum sensing for rf-energy harvesting cognitive radio networks,” in *IEEE International Conference on Communications*, June 2015, pp. 7492–7497.
- [60] N. Barroca, J. M. Ferro, L. M. Borges, J. Tavares, and F. J. Velez, “Electromagnetic energy harvesting for wireless body area networks with cognitive radio capabilities,” in *Proceedings of the URSI Seminar of the Portuguese Committee, Lisbon, Portugal*, November 2012, pp. 1–7.
- [61] A. Hamdi and E. Hossain, “Cognitive and energy harvesting-based d2d communication in cellular networks: stochastic geometry modeling and analysis,” *IEEE Transactions on Communications*, vol. 63, no. 5, pp. 1867–1880, May 2015.
- [62] X. Lu, W. Xu, S. Li, J. Lin, and Z. He, “Simultaneous information and power transfer for relay-assisted cognitive radio networks,” in *IEEE International Conference on Communications Workshops*, June 2014, pp. 331–336.
- [63] S. A. Mousavifar, Y. Liu, C. Leung, M. El Kashlan, and T. Q. Duong, “Wireless energy harvesting and spectrum sharing in cognitive radio,” in *IEEE 80th Vehicular Technology Conference*, September 2014, pp. 1–5.
- [64] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, “Outage performance of cognitive relay networks with wireless information and power transfer,” *IEEE Transactions on Vehicular Technology*, June 2015.

- [65] X. Lu, W. Xu, S. Li, Z. Lin, and J. Lin, "Simultaneous wireless information and power transfer for cognitive two-way relaying networks," in *IEEE 25th International Symposium on Personal, Indoor and Mobile Radio Communications*, September 2014, pp. 748–752.
- [66] V. D. Nguyen, D. V. Son, and O. S. Shin, "Opportunistic relaying with wireless energy harvesting in a cognitive radio system," in *IEEE Wireless Communications and Networking Conference*, April 2015, pp. 97–92.
- [67] D. K. Nguyen, M. Matthaiou, T. Q. Duong, and H. Ochi, "Rf energy harvesting two-way cognitive df relaying with transceiver impairments," in *IEEE International Conference on Communications*, June 2015, pp. 1970–1975.
- [68] Z. Wang, Z. Chen, Y. Yao, B. Xia, and H. Liu, "Wireless energy harvesting and information transfer in cognitive two-way relay networks," in *IEEE Global Communications Conference*, December 2014, pp. 3465–3470.
- [69] Z. Wang, Z. Chen, L. Luo, Z. Hu, B. Xia, and H. Liu, "Outage analysis of cognitive relay networks with energy harvesting and information transfer," in *IEEE International Conference on Communications*, June 2014, pp. 4348–4353.
- [70] G. Zheng, Z. Ho, E. A. Jorswieck, and B. Ottersten, "Information and energy cooperation in cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 9, pp. 2290–2303, May 2014.
- [71] Q. Gao, T. Jing, X. Xing, X. Cheng, Y. Huo, and D. Chen, "Simultaneous energy and information cooperation in mimo cooperative cognitive radio systems," in *IEEE Wireless Communications and Networking Conference*, April 2015, pp. 351–356.
- [72] C. Zhai, J. Liu, and L. Zheng, "Cooperative spectrum sharing with wireless energy harvesting in cognitive radio network," *IEEE Transactions on Vehicular Technology*, July 2015.
- [73] Y. Wang, W. Lin, R. Sun, and Y. Huo, "Optimization of relay selection and ergodic capacity in cognitive radio sensor networks with wireless energy harvesting," *Pervasive and Mobile Computing*, vol. 22, pp. 33–45, June 2015.

- [74] K. Janghel and S. Prakriya, "Outage performance of dynamic spectrum access systems with energy harvesting transmitters," in *IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication*, September 2014, pp. 648–652.
- [75] V. Rakovic, D. Denkovski, Z. Hadzi-Velkov, and L. Gavrilovska, "Optimal time sharing in underlay cognitive radio systems with rf energy harvesting," in *IEEE International Conference on Communications*, June 2015, pp. 7689–7694.
- [76] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 1, pp. 418–428, January 2014.
- [77] S. Yin, E. Zhang, Z. Qu, L. Yin, and S. Li, "Optimal cooperation strategy in cognitive radio systems with energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 4693–4707, September 2014.
- [78] D. Li, S. Yin, and S. Li, "One-step-ahead spectrum sensing in cognitive radio systems with wireless energy harvesting," in *IEEE Global High Tech Congress on Electronics*, November 2013, pp. 130–134.
- [79] S. Yin, E. Zhang, L. Yin, and S. Li, "Optimal saving-sensing-transmitting structure in self-powered cognitive radio systems with wireless energy harvesting," in *IEEE International Conference Communications*, June 2013, pp. 2807–2811.
- [80] —, "Saving-sensing-throughput tradeoff in cognitive radio systems with wireless energy harvesting," in *IEEE Global Communications Conference*, December 2013, pp. 103–1037.
- [81] Y. C. Lin, F. S. Wang, and K. S. Hwang, "A hybrid method of evolutionary algorithms for mixed-integer nonlinear optimization problems," in *Proceedings of the Congress on Evolutionary Computation*, July 1999, pp. 2159–2166.
- [82] B. Li, W. Xu, and X. Gao, "Energy-efficient simultaneous information and power transfer in ofdm-based crns," in *IEEE Vehicular Technology Conference*, May 2015, pp. 11–14.
- [83] A. E. Shafie, M. Ashour, T. Khattab, and A. Mohamed, "On spectrum sharing between energy harvesting cognitive radio users and primary users," in *International Conference on Computing, Networking and Communications*, February 2015, pp. 214–220.

- [84] L. Sibomana, H.-J. Zepernick, and H. Tran, “Wireless information and power transfer in an underlay cognitive radio network,” in *International Conference on Signal Processing and Communication Systems*, December 2015, pp. 1–7.
- [85] D. Gesbert and M. S. Alouini, “How much feedback is multi-user diversity really worth?” in *IEEE International Conference on Communications*, June 2004, pp. 234–238.
- [86] D. W. K. Ng, E. S. Lo, and R. Schober, “Multi-objective resource allocation for secure communication in cognitive radio networks with wireless information and power transfer,” *IEEE Transactions on Vehicular Technology*, May 2015.
- [87] Y. Liu, L. Wang, S. A. R. Zaidi, M. Elkashlan, and T. Q. Duong, “Secure d2d communication in large-scale cognitive cellular networks with wireless power transfer,” in *IEEE International Conference on Communications*, June 2015, pp. 4309–4314.
- [88] A. O. Ercan, M. O. Sunay, and S. Pollin, “Analysis of energy harvesting for green cognitive radio networks,” in *Cognitive Cellular Systems (CCS), 2014 1st International Workshop on*, September 2014, pp. 1–5.
- [89] S. K. Nobar, K. A. Mehr, and J. M. Niya, “Rf-powered green cognitive radio networks: Architecture and performance analysis,” *IEEE Communications Letter*, November 2015.
- [90] T. S. John, “Performance measure and energy harvesting in cognitive and non-cognitive radio networks,” in *IEEE International Conference on Innovations in Information, Embedded and Communication Systems*, March 2015, pp. 1–5.
- [91] A. E. Shafie and A. Sultan, “Optimal random access and random spectrum sensing for an energy harvesting cognitive radio,” in *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, October 2012, pp. 403–410.
- [92] —, “Optimal random access for a cognitive radio terminal with energy harvesting capability,” *IEEE Communications Letters*, vol. 17, no. 6, pp. 1128–1131, June 2013.

- [93] M. Ashour, A. E. Shafie, A. Mohammed, and T. Khattab, "Power-optimal feedback-based random spectrum access for an energy harvesting cognitive user," in *International Conference on Computing, Networking and Communications*, February 2015, pp. 462–468.
- [94] N. Pappas, J. Jeon, A. Ephremides, and A. Traganitis, "Optimal utilization of a cognitive shared channel with a rechargeable primary source node," *Journal of Communications and Networks*, vol. 14, no. 2, pp. 162–168, April 2012.
- [95] R. Rao and A. Ephremides, "On the stability of interacting queues in a multiple-access system," *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 918–930, September 1988.
- [96] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 58, no. 3, pp. 497–520, July 1962.
- [97] A. Sultan, "Sensing and transmit energy optimization for an energy harvesting cognitive radio," *IEEE Communications Letters*, vol. 1, no. 5, pp. 500–503, October 2012.
- [98] J. J. Pradha, S. S. Kalamkar, and A. Banerjee, "Energy harvesting cognitive radio with channel-aware sensing strategy," *IEEE Communications Letters*, vol. 18, no. 7, pp. 1171–1174, July 2014.
- [99] Y. Kim, C. W. Park, and T. J. Lee, "Mac protocol for energy-harvesting users in cognitive radio networks," in *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, January 2014, pp. 59–62.
- [100] A. E. Shafie, , M. Ashour, A. Mohamed, and T. Khattab, "Optimal spectrum access for a rechargeable cognitive radio user based on energy buffer state," in *International Conference on Computing, Networking and Communications*, February 2015, pp. 242–247.
- [101] A. E. Shafie and T. Khattab, "Maximum throughput of a cooperative energy harvesting cognitive radio user," in *IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication*, September 2014, pp. 1067–1072.

- [102] A. E. Shafie, "Space-time coding for an energy harvesting cooperative secondary terminal," *IEEE Communications Letters*, vol. 18, no. 9, pp. 1571–1574, September 2014.
- [103] A. E. Shafie, N. Al-Dhahir, and R. Hamila, "Spacetime coding for an energy harvesting cooperative secondary terminal," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3118–3131, September 2015.
- [104] A. E. Shafie, , T. Khattab, H. Saad, and A. Mohamed, "Optimal cooperative cognitive relaying and spectrum access for an energy harvesting cognitive radio: Reinforcement learning approach," in *International Conference on Computing, Networking and Communications*, February 2015, pp. 221–226.
- [105] D. Jiang, H. Zheng, D. Tang, and Y. Tang, "Relay selection and power allocation for cognitive energy harvesting two-way relaying networks," in *International Conference on Electronics Information and Emergency Communication*, May 2015, pp. 163–166.
- [106] K. Kulkarni and A. Banerjee, "Stable throughput tradeoffs in cognitive radio networks with cooperating rechargeable nodes," in *IEEE Wireless Communications and Networking Conference*, April 2015, pp. 1742–1747.
- [107] S. Luo, R. Zhang, and T. J. Lim, "Optimal save-then-transmit protocol for energy harvesting wireless transmitters," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1196–1207, March 2013.
- [108] A. E. Shafie and A. Sultan, "Optimal selection of spectrum sensing duration for an energy harvesting cognitive radio," in *IEEE Global Communications Conference*, December 2013, pp. 1020–1025.
- [109] S. Boyd and L. Vandenberghe, "Convex optimization problems," in *Convex optimization*. Cambridge University Press, 2008, pp. 127–214.
- [110] W. Chung, S. Park, S. Lim, and D. Hong, "Spectrum sensing optimization for energy-harvesting cognitive radio systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2601–2613, May 2014.
- [111] X. Gao, W. Xu, S. Li, and J. Lin, "An online energy allocation strategy for energy harvesting cognitive radio systems," in *International Conference on Wireless Communications & Signal Processing*, October 2013, pp. 1–5.

- [112] W. Chung, S. Park, S. Lim, and D. Hong, "Optimal transmit power control for energy-harvesting cognitive radio system," in *IEEE Vehicular Technology Conference*, September 2013, pp. 1–5.
- [113] P. He and L. Zhao, "Optimal power control for energy harvesting cognitive radio networks," in *IEEE International Conference on Communications*, June 2015, pp. 92–97.
- [114] S. Gong, L. Duan, and P. Wang, "Robust optimization of cognitive radio networks powered by energy harvesting," in *IEEE Conference on Computer Communications*, May 2015, pp. 612–620.
- [115] S. A. R. Zaidi, A. Afzal, M. Hafeez, M. Ghogho, and D. C. McLernon, "Solar energy empowered 5g cognitive metro-cellular networks," *IEEE Communications Magazine*, vol. 53, no. 7, pp. 70–77, July 2015.
- [116] L. Chen, L. Huang, H. Xu, C. Yang, Z. Sun, and X. Wang, "Primary secrecy is achievable: Optimal secrecy rate in overlay crns with an energy harvesting secondary transmitter," in *International Conference on Computer Communication and Networks*, August 2015, pp. 1–6.
- [117] D. T. Hoang, D. Niyato, P. Wang, and D. I. Kim, "Performance analysis of wireless energy harvesting cognitive radio networks under smart jamming attacks," *IEEE Transactions on Cognitive Communications and Networking*, October 2015.
- [118] Y. Zhang, W. Han, D. Li, P. Zhang, and S. Cui, "Power versus spectrum 2-d sensing in energy harvesting cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 23, pp. 6200–6211, December 2015.
- [119] —, "Two-dimensional sensing in energy harvesting cognitive radio networks," in *IEEE International Conference on Communications*, June 2015, pp. 2029–2034.
- [120] C. Mikeka, H. Arai, C. Georgiadis, and A. Collado, "Dtv band micropower rf energy-harvesting circuit architecture and performance analysis," in *IEEE International Conference on Proceedings of RFID-Technologies and Applications*, September 2011, pp. 561–567.

- [121] H. Ostaffe. (2010) Power out of thin air: Ambient rf energy harvesting for wireless sensors. [Online]. Available: <http://powercastco.com/PDF/Power-Out-of-Thin-Air.pdf>
- [122] H. J. Visser and R. J. M. Vullers, “Rf energy harvesting and transport for wireless sensor network applications: Principles and requirements,” *Proceedings of the IEEE*, vol. 101, no. 6, pp. 1410–1423, June 2013.
- [123] A. Gosavi, *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*. Springer Press, 2003.
- [124] O. Buffet, A. Dutech, and F. Charpillet, “Shaping multi-agent systems with gradient reinforcement learning,” *Journal of Autonomous Agents and Multi-Agent System*, vol. 15, no. 2, pp. 197–220, January 2007.
- [125] P. Marbach and J. N. Tsitsiklis, “Simulation-based optimization of markov reward processes,” *IEEE Transactions on Automatic Control*, vol. 46, no. 2, pp. 191–209, February 2001.
- [126] J. Baxter, P. L. Barlett, and L. Weaver, “Experiments with infinite-horizon, policy-gradient estimation,” *Journal of Artificial Intelligence Research*, vol. 15, no. 1, pp. 351–381, November 2001.
- [127] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.
- [128] D. T. Hoang and D. Niyato, “Performance analysis for cognitive machine-to-machine communications,” in *IEEE International Conference on Communication Systems*, November 2012, pp. 245–249.
- [129] M. Schwartz, *Mobile wireless communications*. Cambridge University Press, 2005.
- [130] C. Mikeka and H. Arai, “Design issues in radio frequency energy harvesting system,” in *Sustainable Energy Harvesting Technologies - Past, Present and Future*. INTECH Open Access Publisher, 2011, pp. 235–256.
- [131] H. W. Kuhn and A. W. Tucker, “Nonlinear programming,” in *Proceedings of 2nd Berkeley Symposium on Mathematical Statics and Probability*, August 1950, pp. 481–492.

- [132] V. S. Borkar, *Stochastic Approximation: A Dynamic Systems Viewpoint*. Cambridge University Press, 2008.
- [133] D. T. Hoang, X. Lu, D. Niyato, P. Wang, D. I. Kim, and Z. Han, “Applications of repeated games in wireless networks: A survey,” *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, pp. 2102–2135, June 2015.
- [134] D. Niyato and E. Hossain, “Competitive pricing for spectrum sharing in cognitive radio networks: Dynamic game, inefficiency of nash equilibrium, and collusion,” *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 192–202, January 2008.
- [135] D. P. Bertsekas and J. N. Tsitsiklis, “Gradient convergence in gradient methods with errors,” *Society for Industry and Applied Mathematics Journal on Optimization*, vol. 10, no. 3, pp. 627–642, February 1999.
- [136] V. R. Konda and J. N. Tsitsiklis, “Convergence rate of linear two-time-scale stochastic approximation,” *The Annals of Applied Probability*, vol. 14, no. 2, pp. 796–819, April 2004.

Author's Publications (Published)

Journal

- [1] **D. T. Hoang**, D. Niyato, P. Wang, and D. I. Kim, "Opportunistic channel access and RF energy harvesting in cognitive radio networks," *IEEE Journal of Selected Areas in Communications*, vol. 32, issue 11, pp. 2039-2052, Nov. 2014.
- [2] **D. T. Hoang**, D. Niyato, P. Wang, and D. I. Kim, "Performance optimization for cooperative multiuser cognitive radio networks with RF energy harvesting capability," *IEEE Transactions on Wireless Communications*, vol. 14, issue 7, pp. 3614-3629, March. 2015.
- [3] **D. T. Hoang**, D. Niyato, P. Wang, and D. I. Kim, "Performance analysis of wireless energy harvesting cognitive radio networks under smart jamming attacks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 2, pp. 200-216, Oct. 2015.
- [4] **D. T. Hoang**, X. Lu, D. Niyato, P. Wang, D. I. Kim and Z. Han, "Applications of repeated games in wireless networks: A survey," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, pp. 2102-2135, June. 2015.
- [5] M. A. Alsheikh, **D. T. Hoang**, D. Niyato, S. Lin, and H.-P. Tan, "Markov decision processes with applications in wireless sensor networks: A survey," *IEEE Communications Surveys and Tutorials*, vol. 17, issue 3, pp. 1239-1267, April. 2015.
- [6] D. Niyato, **D. T. Hoang**, N. C. Luong, P. Wang, D. I. Kim, and Z. Han, "Smart data pricing models for Internet-of-Things (IoT): A bundling strategy approach," *IEEE Network Magazine*, vol. 30, no.2, pp. 18-25, Mar. 2016.

Conference

- [1] **D. T. Hoang**, D. Niyato, and L. B. Long, “Simulation-based optimization for admission control of mobile cloudlets,” *IEEE International Conference on Communications*, pp. 3764-3769, Sydney, Australia, June. 2014.
- [2] **D. T. Hoang**, D. Niyato, and P. Wang, D. I. Kim, “Optimal decentralized control policy for wireless communication systems with wireless energy transfer,” *IEEE International Conference on Communications*, pp. 2835-2840, Sydney, Australia, June. 2014.
- [3] **D. T. Hoang**, D. Niyato, and D. In Kim, “Cooperative bidding of data transmission and wireless energy transfer,” to be presented in *IEEE Wireless Communications and Networking Conference*, pp. 1597-1602, Istanbul, Turkey, April. 2014.
- [4] **D. T. Hoang**, D. Niyato, and N. T. Hung, “Optimal energy allocation policy for wireless networks in the sky,” *IEEE International Conference on Communications*, pp. 3204-3209, London, UK, June. 2015.
- [5] **D. T. Hoang**, D. Niyato, “Information service pricing competition in Internet-of-Vehicle (IoV),” *IEEE International Conference on Computing, Networking and Communications*, pp. 1-5, Hawaii, USA, February, 2016.