

國立交通大學

電子工程學系 電子研究所

博士論文

自組式網路中的無線資源管理：分散式學習與穩當策略

Radio Resource Management in Self-organized Networks: Distributed
Learning and Robust Strategies

研究生：曾理銓

指導教授：黃經堯 教授

中華民國一〇二年十二月

自組式網路中的無線資源管理：分散式學習與穩當策略

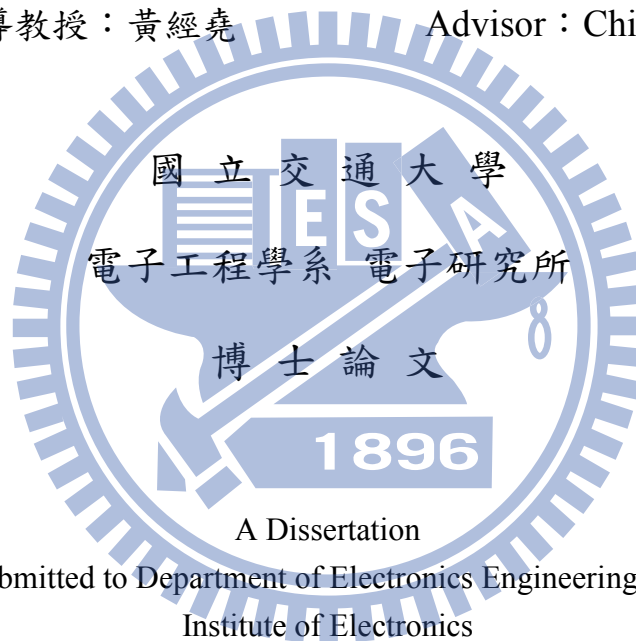
Radio Resource Management in Self-organized Networks: Distributed Learning and Robust Strategies

研究生：曾理銓

Student : Li-Chuan Tseng

指導教授：黃經堯

Advisor : Ching-Yao Huang



A Dissertation

Submitted to Department of Electronics Engineering and
Institute of Electronics

College of Electrical and Computer Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Electronics Engineering

December 2013

Hsinchu, Taiwan

中華民國一〇二年十二月

學生：曾理銓

指導教授：黃經堯

國立交通大學 電子工程學系 電子研究所 博士班

摘 要

由於其在頻譜使用上的彈性，自組式網路被視為滿足不斷增加的行動通訊流量需求的一個重要方案。在自組式網路中，共享頻譜的節點是分散的，必須由個別的節點進行無線資源管理。此外，各個節點的無線資源管理決定會影響彼此的效能，因此我們需要能考慮節點的相互作用的分散式無線資源管理方法。為達此目的，本論文將包括博弈論，信息論，隨機學習在內的多元數學工具，用於無線資源管理問題的建模與解決方案。雖然自組式網路中的當紅議題，如異構網路和無線感知網路等，已有了深入的研究，我們的工作的新穎性在於基於分散式學習演算法，各節點在資訊有限的條件下，仍具有自組與調整的能力。

本論文首先介紹相關的數學工具，包括賽局理論的基礎知識與隨機學習演算法的簡介。接著是關於無線感知網路的一份文獻探討。隨後，我們提供了四個應用實例。在每個例子中，我們針對一個在分散式網路中可能會遇到的無線資源管理問題，建構賽局理論模型。網路中的節點被視為具備自主學習能力的自動機，並能藉由個別行為-回報歷史，習得適當的資源管理策略。我們亦透過數值模擬，評估學習過程的收斂性及其性能。

關鍵詞：自組式網路、無線資源管理、賽局理論、隨機學習演算法

Radio Resource Management in Self-organized Networks: Distributed Learning and Robust Strategies

Student : Li-Chuan Tseng

Advisor : Dr. Ching-Yao Huang

Department of Electronics Engineering
& Institute of Electronics
National Chiao Tung University

ABSTRACT

Self-organized network (SoN) has been considered as an important solution to the increasing demand of mobile traffics, due to its flexibility in spectrum access. In SoNs, the nodes sharing the spectrum are located in a distributed manner, and the radio resource management (RRM) must be performed by individual nodes. Moreover, since the RRM decisions of the nodes affect the performance of each other, distributed RRM methods considering the interactions of nodes are desirable for SoNs. To this aim, a diversified class of mathematical tools including game theory, information theory, and stochastic learning are involved in this thesis, for the problem formulation and solution of the RRM in SoNs. While the rising topics of SoNs such as heterogeneous networks and cognitive radio networks (CRNs) have been intensively studied, the novelty of our work lies in the capability of self-organization and adjustment under limited information, based on distributed learning methods.

We start our presentation with the underlying mathematics, including game theory fundamentals and an introduction to the stochastic learning algorithm. A survey on CRNs follows. Four application examples are provided afterwards. In each example, game theoretical framework is adopted to formulate an RRM problem we may encounter in distributed networks. The nodes in networks are modeled as self-organized learning automata, which learn proper RRM strategies through individual action-reward history. The convergence of the learning procedure and its performance are evaluated via numerical simulations.

Keywords: Self-organized Networks, Radio Resource Management, Game Theory, Stochastic Learning Algorithm

誌 謝

隨著這本論文的完成，漫長的博士班生涯也即將畫下句點。首先感謝我的指導教授黃經堯博士，一路上的指導與支持，使我能順利完成論文。加上大學部專題與碩士班時期，近十年的提攜之情，畢生難忘。

博士班四年級時，我暫別交大，在一個大雪紛飛的清晨抵達巴黎並赴法國 Télécom Sudparis 研修。留法期間，承蒙 D. Zhaglache 與 A. Marzouki 兩位教授的協助與指導，讓我一年半的留學生涯過得十分充實。另外要特別感謝法國 Supélec 的 H. Tembine 教授，對研究方向的提點與數學理論的詳細解說，使我的研究有重要的突破。

由於一開始低估了讀博士的難度，我的研究過程難免挫折。所幸除了台法雙方的指導教授的支持，在各階段都有貴人相助。特別感謝電子所簡鳳村教授與中研院張佑榕博士、鍾偉和博士，對於我的關鍵論文的貢獻。三位在自身研究與教學工作繁忙之餘，仍仔細修改我的拙作並提供建議，這個過程著實獲益良多。當然也要謝謝與同窗好友冠穎、Robert、Maria、金鑫等，在研究工作上的互相切磋及在生活上的照顧。

最後，感謝我的家人，在我的求學階段關懷鼓勵與經濟支援，使我可以無後顧之憂，得以順利完成學業。

曾理銓 謹誌

中華民國一〇二年十二月

Contents

摘要	i
Abstract	ii
誌謝	iii
Contents	iv
List of Figures	ix
List of Tables	xii
1 Introduction	1
1.1 Background and Motivations	1
1.2 Thesis Outline and Contributions	4
1.3 Publications	7
I The Backgrounds	9
2 Stochastic Learning in Games	10
2.1 Introduction	10
2.2 Non-cooperative Game Theoretical Concepts	10
2.2.1 Game with External State	11
2.2.2 Mixed Strategy Extension	12
2.2.3 Potential Games	13
2.2.4 Achieving NE: Previous Methods	14

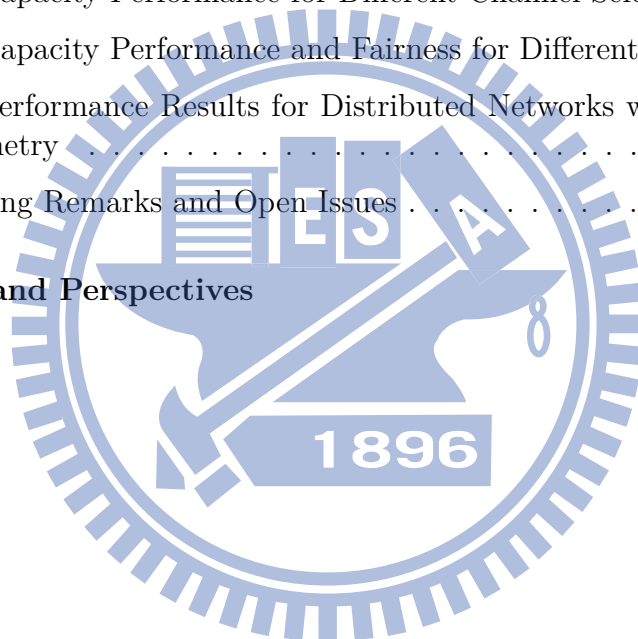


2.3	Evolutionary Game and Replicator Dynamics	14
2.3.1	Replicator Dynamics	15
2.3.2	Stochastic Game	16
2.4	Stochastic Learning Algorithm	17
2.4.1	Generic SLA Structure	19
2.5	Update Rules	20
2.5.1	Bush-Mosteller (BM) Update Rule	21
2.5.2	Multiplicative-weight Update Rule	22
2.6	Convergence of the Proposed Algorithm	24
2.6.1	Potential Games	25
2.6.2	Non-potential Games	28
2.7	Applications: Game Theoretic Modeling	28
	Appendix 2.A Assumptions for Stochastic Approximation	29
3	A Survey on the Spectrum Access of Cognitive Radio Networks	31
3.1	Cognitive Spectrum Access	31
3.2	Opportunistic Spectrum Access	33
3.3	Spectrum Trading with Single Seller	34
3.4	Spectrum Trading with Multiple Seller	34
3.4.1	Exclusive Access	35
3.4.2	Shared Access	36
3.5	CR Primary Access	37

II	Examples of Fully Distributed Learning	39
4	Network Selection in Cognitive Heterogeneous Networks	40
4.1	Introduction	40
4.1.1	Game-theoretic Problem Mapping	41
4.2	System Model	42
4.3	Self-Organized Network Selection	43
4.3.1	Game Model	44
4.3.2	Analysis of Nash Equilibrium	45
4.3.3	Stochastic Learning Procedure	46
4.4	Numerical Results	48
4.5	Concluding Remarks	50
5	Spectrum Trading in Multiple-Seller Cognitive Radio Networks	52
5.1	Introduction	53
5.2	System Model	55
5.2.1	Spectrum Trading Mechanism	55
5.2.2	Two-level Competition as a Stackelberg Game	56
5.3	Service Selection of Secondary Users	57
5.3.1	Game Model	58
5.3.2	Analysis of Nash Equilibrium	59
5.3.3	Stochastic Learning Procedure for Service Selection	61
5.3.4	Social Welfare and Price of Anarchy	63
5.4	Price Competition among Service Providers	64
5.4.1	Game Model	64
5.4.2	Stochastic Learning Procedure for Price Competition	66
5.5	Numerical Results	69
5.5.1	Convergence Behavior of the Lower-level Game	70
5.5.2	Performance Comparison in the Lower-level Game	72
5.5.3	Convergence Behavior of the Upper-level Game	75
5.5.4	Non-unique User Loads	78

5.6	Conclusion and and Open Issues	79
III Examples of Distributed Learning with Partial Cooperation		81
6	Self-organized Channel Assignment in Two-tier Distributed Networks	82
6.1	Introduction	83
6.1.1	Examples of Two-tier Distributed Networks	83
6.1.2	Contributions	85
6.1.3	Game-theoretic Problem Mapping	86
6.2	Related Works	86
6.2.1	Variations of Frequency Planning	87
6.2.2	Learning-based Methods	87
6.3	System Model	90
6.4	Game-theoretic Model	93
6.4.1	Problem Formulation and Game Model	94
6.4.2	Analysis of Nash Equilibrium	96
6.5	Stochastic Learning Procedure	97
6.6	Numerical Results	98
6.6.1	Convergence of the proposed SL-based learning algorithm	98
6.6.2	Capacity performance	102
6.7	Concluding Remarks	105
7	Distributed Channel Allocation in Network MIMO	106
7.1	Introduction	107
7.2	Related Works	108
7.2.1	Precoding with BS Cooperation	108
7.2.2	Spectrum Sharing	109
7.3	System Model	110
7.3.1	The Network MIMO System	110
7.3.2	Transmitter Precoding	112

7.4	Channel Selection for Network MIMO	115
7.4.1	Game-Theoretic Formulation	116
7.4.2	Existence of Nash Equilibrium	117
7.4.3	Acquisition of the Interference Information	119
7.5	Stochastic Learning-based Channel Selection Algorithm	119
7.5.1	Algorithm Description	120
7.5.2	Convergence Properties of the Proposed Algorithm	120
7.6	Numerical Results and Discussions	123
7.6.1	Convergence Behaviors of the Proposed Learning Algorithm	123
7.6.2	Capacity Performance for Different Channel Selection Strategies	124
7.6.3	Capacity Performance and Fairness for Different Precoding Schemes	126
7.6.4	Performance Results for Distributed Networks with Random Geometry	128
7.7	Concluding Remarks and Open Issues	131
8	Conclusion and Perspectives	133
	Bibliography	136

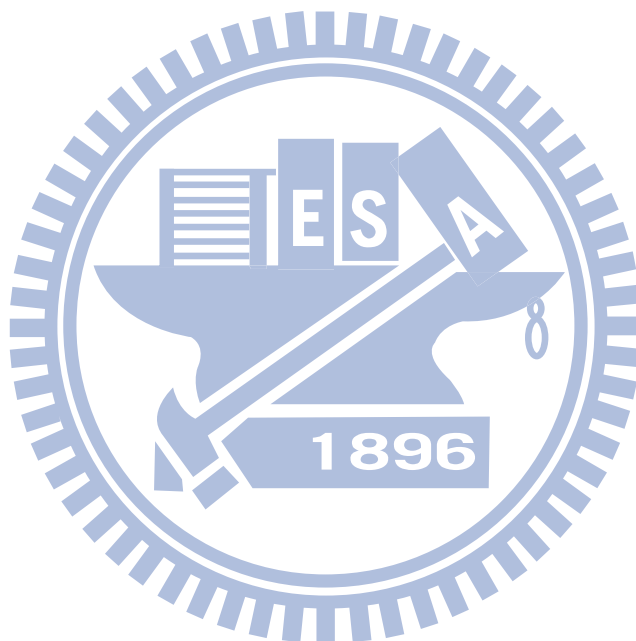


List of Figures

1.1	Two scenarios in cooperative communications.	3
2.1	Replicator dynamics in evolutionary game and stochastic game.	17
2.2	Generic SLA structure.	19
3.1	Cognitive radio network architecture.	32
4.1	An exemplary heterogeneous network with 2 SPs, 3 PUs, and 4 SUs. The filled and blank blocks in the licensed band of each SP denote the busy channels currently used by the PUs and the residual channels available for serving the SUs, respectively.	42
4.2	Evolution of the mixed strategies (choice probability of actions) of some players, using different learning rates.	48
4.3	Test of unilateral deviation from the resulting strategy profile of each of the 10 players, using different learning rates.	49
5.1	An exemplary cognitive radio network with 2 SPs, 3 PUs, and 4 SUs. The filled and blank blocks in the licensed band of each SP denote the busy channels currently used by the PUs and the residual channels available for serving the SUs, respectively.	55
5.2	Evolution of the mixed strategies (probability of taking different actions) of all players. Each pair of $p_{i,1}(j)$ and $p_{i,2}(j)$ shows the behavior of a player $i \in \mathcal{N}$	71
5.3	Test of unilateral deviation from the learned strategy profile of each of the $N = 6$ players, with learning rates $b = 0.3$ and $b = 0.5$	72
5.4	Evolution of the actions $a_i(j)$ for selected players.	73
5.5	Comparison of the average (normalized) utility per SU for different service selection schemes.	74
5.6	Comparison of the JFI using three service selection schemes.	74

5.7	Evolution of the mixed strategies (probability of taking different actions) of the $M = 2$ sellers.	76
5.8	Price and revenue dynamics of the $M = 2$ sellers.	76
5.9	Test of different strategies. For each seller, the four bars show its revenues when taking the four different pricing strategies, while its opponent sticks to the learned strategy.	77
5.10	Drift in user loads. For each seller, the dynamics of prices, revenues, and estimated revenues are shown.	79
6.1	Possible interference scenarios related to femtocell communications.	84
6.2	Dual-stripe deployment of sensor clusters.	91
6.3	Exemplary time slot allocation in a frame. In the first slot, cluster head A and C assign channels for sensor node A1 and C1, respectively.	92
6.4	Evolution of the mixed strategies (probability of taking different actions) of all players. Each pair of $p_{i,1}(t)$ and $p_{i,2}(t)$ shows the behavior of player i	100
6.5	Test of unilateral deviation from the resulting strategy profile of each of the 10 players.	100
6.6	Evolution of the actions $a_i(j)$ for some players.	101
6.7	Evolution of the mixed strategies (probability of taking different actions) of all players with active ratios of 50% and 75%. Each pair of $p_{i,1}(t)$ and $p_{i,2}(t)$ shows the behavior of a player $i \in \mathcal{N}$	102
6.8	Test of unilateral deviation from the resulting strategy profile of each of the 10 players.	103
7.1	Illustration of distributed channel selection with joint precoding in multicell networks. For MS_1 , $\mathcal{C}_1 = \{1, 3\}$ and $\mathcal{D}_1 = \{1\}$, where BS_1 and BS_3 both receive CSI feedback from MS_1 and perform interference mitigation but only BS_1 serves MS_1 . For MS_2 , $\mathcal{C}_2 = \{1, 2, 3\}$ and $\mathcal{D}_2 = \{2, 3\}$, where BS_2 and BS_3 jointly serve MS_2 while all three BSs perform interference mitigation. For MS_3 , $\mathcal{C}_3 = \{3\}$ and $\mathcal{D}_3 = \{3\}$, where only BS_3 serves MS_3	112
7.2	Evolution of the mixed strategies (probability of taking different actions) of four selected players when joint processing is adopted.	125
7.3	Evolution of the estimated cost of taking different actions for two selected players (marked by blue and red colors, respectively) when joint processing is adopted.	126
7.4	Cost and capacity for each player for the NE strategy and unilateral deviation from the NE strategy.	127

7.5	Comparison of the achievable capacity for three channel selection strategies when joint processing is adopted.	128
7.6	A snapshot of the nodes' positions and network topology. The link ID is shown in parenthesis next to the link.	129
7.7	Evolution of the mixed strategies of four selected players when local precoding is applied to the distributed network.	130
7.8	(a) Achievable capacity for different channel selection strategies. (b) Cost for each player for the NE strategy and unilateral deviation from the NE strategy.	131

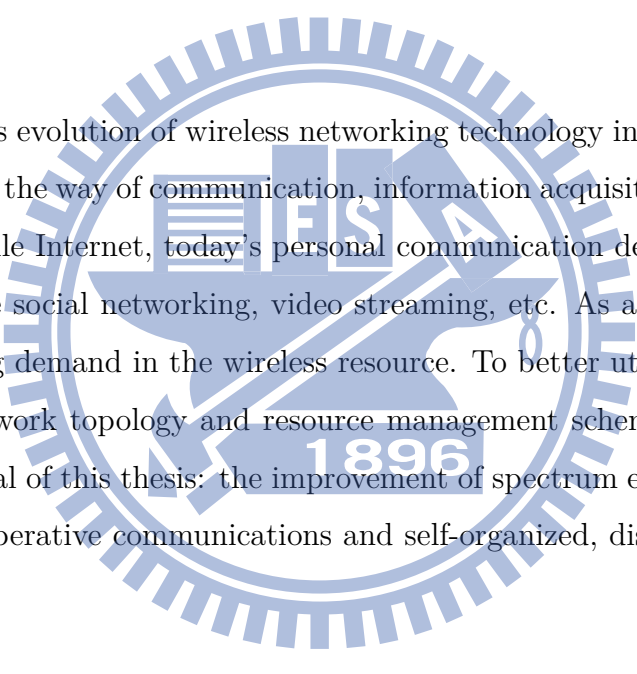


List of Tables

2.1	Comparison	16
2.2	Information Available to the Players	18
2.3	Summary of Notations in Game-theoretic Formulation	18
2.4	Summary of Symbols for Game-theoretic Formulation	29
4.1	Mapping to game-theoretic formulation.	42
4.2	Summary of Symbols for Game-theoretic Formulation	44
4.3	Comparison of the achievable expected system throughput of three network selection schemes	50
5.1	Elements in a Stackelberg game	57
5.2	Summary of Notations for Game-theoretic Formulation	58
5.3	Simulation Parameters	70
6.1	Mapping to game-theoretic formulation.	86
6.2	Summary of Notations in Game-theoretic Formulation	94
6.3	Simulation Parameters	99
6.4	Comparison of the capacity and fairness for different channel assignment schemes	104
7.1	Summary of Notations in Game-theoretic Formulation	116
7.2	The Simulation Setup	124
7.3	Capacity per MS (bps/Hz) for Different Combinations of Channel Selection and Precoding Schemes	127
7.4	JFI (7.31) for Different Combinations of Channel Selection and Precoding Schemes	128

Chapter 1

Introduction

The logo of Fudan University is a circular seal with a gear-like outer edge. Inside the seal, there is a shield with a book and a torch, and the year '1896' is written at the bottom. The letters 'F S A' are also visible in the center.

The continuous evolution of wireless networking technology in the last decade has significantly changed the way of communication, information acquisition, and entertainment. Through the mobile Internet, today's personal communication devices provide more and more services, like social networking, video streaming, etc. As a consequence, there has been an increasing demand in the wireless resource. To better utilize the shared wireless medium, new network topology and resource management scheme are important. This constitutes the goal of this thesis: the improvement of spectrum efficiency in wireless systems through cooperative communications and self-organized, distributed radio resource management.

1.1 Background and Motivations

ACHIEVING reliable and high data rate communications over wireless links remains a challenging problem. In fact, the inherent nature of the wireless medium has created a number of new research topics. Compared to the wire-line communications, the wireless medium is a ubiquitous resource which is accessible simultaneously by multiple transmissions. The sharing of the medium by multiple links results in a mutually interfered environment, and gives rise to challenges in resource management. In conventional cellular

networks consisting of multiple base stations, frequency planning is adopted. However, we have to consider universal frequency reuse. The reasons are two-fold. First, frequency reuse factor larger than one limits the spectrum efficiency in that only a fraction of spectrum is utilized by each cell regardless of the actual interference condition. Second, in newly developed network topology, the base stations can be deployed in a distributed manner, which makes cell planning hard. Obviously, universal frequency reuse among nearby cells results in inter-cell interference (ICI) and degrades the performance. This statement, though straightforward, lies at the basis of many research topics within wireless communications. Let us mention two examples as follows.

■ Cooperative communications.

The broadcast nature of wireless communications suggests that a receiver node can *overhear* the source signal transmitted towards a neighboring nodes. Instead of treating the overheard information as interference and trying to mitigate the negative effect, cooperative communication takes advantage of the proximity of nodes to create spatial diversity, thereby to improve the spectrum efficiency and reliability. In practice, the cooperation can be implemented in different ways. In the relay (multi-hop) networks, the signal is received and processed at the surrounding nodes, then re-transmitted towards the destination. On the other hand, when multi-antenna system is considered, signal processing techniques can be applied to transmit the signal simultaneously from multiple nodes. In this case, the signal to be transmitted is pre-processed to suppress the ICI and obtain the diversity gain. Assuming perfect back-haul connection, the network consisting of multiple cells can be viewed, and we end up with a *virtual* MIMO system. The two scenarios are shown in Figure 1.1.

■ Self-organized resource management.

The limitations on coordination of distributed networks gives rise to new challenges for resource management. On top of that, self-organized network (SoN) capability has received much attention because, unlike the negotiation-based approaches, it does not suffer from the information exchange overhead. SoN has been considered

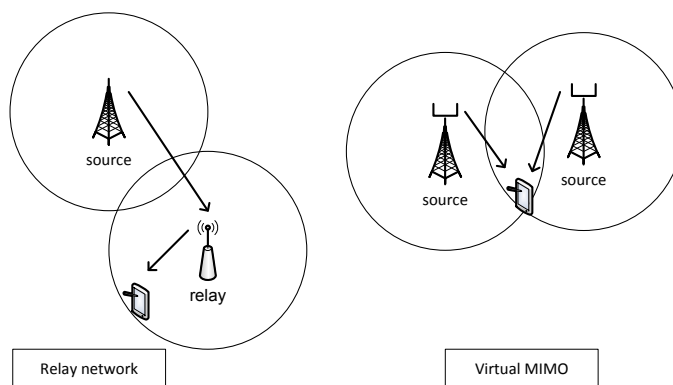


Figure 1.1: Two scenarios in cooperative communications.

in different examples. From the spectrum utilization perspective, dynamic spectrum access (DSA) suggests a distributed decision-making mechanism with consideration on a possibly varying environment. Another example is the heterogeneous networks, in which the spectrum is owned by multiple service providers, and users need proper network selection. Two fundamental mathematical tools frequently involved in SoN are the game theory and the reinforcement learning (RL). Game theory investigates the interaction among self-acting agents, in either cooperative or non-cooperative ways. Game theoretic formulation defines possible solution concept of *equilibrium* at which unilateral deviation from an equilibrium point brings no better results. On the other hand, RL algorithms helps individual agents learn a better strategy based on their own action-reward history. Interestingly, the two tools may be combined; several reinforcement learning techniques have been proved to achieve the equilibrium point.

This thesis aims at investigating the distributed resource management in wireless communications. Specifically, we study the use of reinforcement learning under game-theoretic formulations. The motivation behind is that, while the problem structures can be quite different, we would like to propose a unifying scheme which is suitable for various applications. A general guideline of the proposed scheme is described as follows. First, some components (e.g., base stations or users) in the network are identified as the agents (players) of the game. Second, the utility function is defined in order to reflect the agents'

interests, either individual or common ones. Finally, assuming they are rational and selfish devices, agents act as learning automata to learn their strategies that maximize their individual payoffs. Notice that in addition to the interaction among players, the time-varying external state is also considered in the learning procedure.

Starting with the seminal contributions of Von Neumann, Morgenstern [1] and Nash [2], game theory has been extensively investigated in the previous century. While early works focused on the studies of economy, game theory has become a popular choice for the researchers in wireless networks. Comprehensive surveys on the game-theoretic studies for different wireless network applications can be found in [3,4]. On the other hand, we also see rapid development of RL algorithms over the past few decades. Q-learning [5] is a simple way for agents to learn how to act optimally in controlled Markovian domains. It works by successively improving its evaluations of the quality (Q-value) of particular actions at particular states. Another learning method, referred to as the stochastic learning (SL), is based on the update of probability. Using the techniques in stochastic approximation [6], the SL process tracks the ODE of different dynamics. The resulting state depends on the learning rule adopted. The hybrid learning was discussed [7], where the agents may adopt different learning rules to obtain the strategy. SL has been applied to several areas in wireless networks, for example, precoder selection [8], network selection [9], and cognitive radio [10]. The connection between learning and game has been investigated by Sastry *et al.* [11]. The authors have proposed an SL algorithm and pointed out that NE can be achieved when the algorithm is applied to common-payoff games. In this thesis we will further show that the same algorithm achieves NE for potential games, of which the common-payoff game is a special case.

1.2 Thesis Outline and Contributions

The main content of the thesis is divided into three parts. In Part I (Chapter 2 and 3) we review the fundamental mathematical tools and provide a survey on cognitive radio networks. Part II (Chapter 4 and 5) provide two application examples of fully distributed

learning in distributed resource management. Part III (Chapter 6 and 7) studies the case of distributed learning with partial cooperation. The following is an overview of each chapter.

Chapter 2. This chapter introduces the different concepts that will be used throughout the thesis, together with the fundamental mathematics. The basic ideas in game theory is first reviewed. This problem is formulated as a non-cooperative game. The existence and multiplicity of the Nash equilibrium (NE) solution will be investigated for two different network models. In the second part of this chapter, the stochastic learning algorithm is explained in detail. We give the structure of SLA, and present several update rules. At the end, we show that under certain conditions, the SLA converges to NE.

Chapter 3. The first three examples in this thesis are all related to the spectrum access behaviors of cognitive radio networks (CRNs). Therefore, before entering the examples, we open up one chapter to review the previous works on CRNs. The spectrum access in CRNs is classified as different models according to the way the spectrum is granted to the secondary users. Then the representative works of each model are summarized.

Chapter 4. This chapter presents the first application: the network selection problem in cognitive heterogeneous networks (HetNets) where multiple radio access technologies (RATs) coexist. We formulate the network selection problem as a non-cooperative game where the secondary users (SUs) are the players. In particular, under a cognitive access scenario, the availability of channels for SUs depends on the traffic demands of PUs, and is considered as the time-varying external state. With a reasonably designed utility function, we prove that the game is an OPG. SLA is adopted and each SU's strategy progressively evolves toward the Nash equilibrium (NE) based on its own action-reward history, without the need to know actions in other SUs. The convergence property and the performance in terms of throughput and fairness are again shown through simulations.

Chapter 5. As the second application example of SLA, this chapter studies the spectrum trading in CRNs. Different from the first example, now the licensed spectrum opportunities are sold to multiple unlicensed secondary users by multiple service providers. The spectrum trading is modeled as a multi-leader multi-follower Stackelberg game with two

levels of competition. In the lower-level competition, each secondary user selects a service provider with time-varying channel availability. The service selection is determined by the prices and the quality of service, which depends on the number of residual channels and the behavior of other secondary users. In the upper-level competition, service providers adjust their pricing strategies to maximize their individual revenues. We further propose decentralized, stochastic learning-based algorithms for both levels, where a player's strategy progressively evolves toward the Nash equilibrium (NE) based on its own action-reward history without information of other players' actions. The convergence properties of the proposed algorithms toward NE points are theoretically and numerically verified. The proposed method demonstrates good utility and fairness performances for the secondary users as compared to other service selection schemes.

Chapter 6. The third example considers channel assignment in OFDMA-based two-tier distributed networks. The secondary users are formulated as the players, and the strategy is the channel assignment. There are two major difference from the previous examples. Firstly, unlike the previous examples where a resource unit is granted by the owner to a specific user, here we consider the case that all users access the same spectrum. On top of that, an interference mitigation game is formed. Secondly, each player is allowed to know the action of its neighbors. In this way, a proper utility function can be defined, and the channel assignment problem is formulated as an ordinal potential game which has at least one pure-strategy Nash equilibrium (NE). Then the stochastic learning algorithm discussed in Chapter 2 is applied. The convergence property toward pure strategy NE points is verified through system-level simulations. In addition, performance evaluation is carried out by comparing the proposed algorithm with other methods.

Chapter 7. The last example addresses the joint processing and distributed channel assignment in network MIMO systems. The cooperative frequency reuse among base stations (BSs) can improve the system spectral efficiency by reducing the intercell interference (ICI) through channel selection and precoding. We presents a game-theoretic study of channel selection for realizing network MIMO operation under time-varying wireless channel. We propose a new joint precoding scheme that carries enhanced interference

mitigation and capacity improvement abilities for network MIMO systems. We formulate the channel selection problem as a noncooperative game with BSs as the players, and show that our game is an exact potential game (EPG) given the proposed utility function. A decentralized, stochastic learning-based algorithm is proposed where each BS progressively moves toward the Nash equilibrium (NE) strategy based on its action-reward history and not actions taken by others. The convergence properties of the proposed learning algorithm toward a pure-strategy NE point are theoretically shown and numerically verified for different network topologies. The proposed learning algorithm also demonstrates a fine capacity and fairness performance as compared to other schemes through extensive link-level simulations.

1.3 Publications

The research work conducted during the three years of the thesis has led to several publications.

International Journal Articles

- **L.-C. Tseng**, F.-T. Chien, D. Zhang, R. Y. Chang, W.-H. Chung, and C.-Y. Huang, “Network Selection in Cognitive Heterogeneous Networks Using Stochastic Learning,” to appear in *IEEE Communications Letters*.

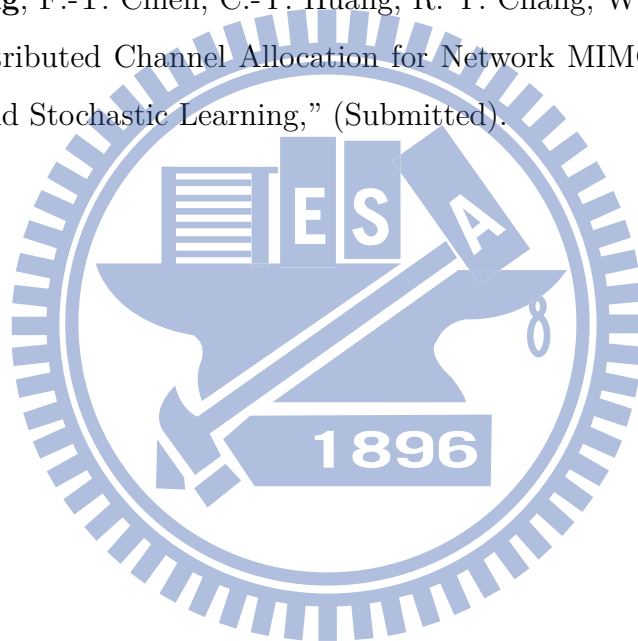
International Conference Proceedings

- C.-H. Lin, **L.-C. Tseng**, C.-Y. Huang “Cognitive Radio Networks: Game Modeling and Self-organization Using Stochastic Learning,” in *Proc. IEEE PIMRC 2013*, Sept. 2013, pp.3006-3010.
- **L.-C. Tseng**, X. Jin, A. Marzouki, and C.-Y. Huang, “Downlink Scheduling in Network MIMO Using Two-Stage Channel State Feedback,” in *Proc. IEEE VTC Fall '12*, Sept. 2012, pp.1-5.

- **L.-C. Tseng**, C.-Y. Huang and A. F. Hanif, “Dynamic resource management for OFDMA-based Femtocells in the Uplink,” in *Proc. IEEE IWCMC '11*, July 2011, pp. 528-533.

Submitted Articles

- **L.-C. Tseng**, F.-T. Chien, C.-Y. Huang, R. Y. Chang, W.-H. Chung and A. Marzouki, “Self-Organized Cognitive Sensor Networks: Distributed Channel Assignment for Pervasive Sensing” (Submitted).
- **L.-C. Tseng**, F.-T. Chien, C.-Y. Huang, R. Y. Chang, W.-H. Chung and A. Marzouki, “Distributed Channel Allocation for Network MIMO: Game-Theoretic Formulation and Stochastic Learning,” (Submitted).





Part I

The Backgrounds

Chapter 2

Stochastic Learning in Games

This chapter aims at introducing the stochastic learning algorithm which is used in game models.

2.1 Introduction

THERE has been much interest in designing learning algorithms toward NE in non-cooperative games. However, the external state (CSI) is unknown and the action is selected by each player simultaneously and independently in each play. Therefore, previous algorithms requiring complete information and implicit ordering of acting players (e.g., those based on better response dynamics (BRD) [12] and fictitious play (FP) [13]) may not be feasible in our self-organized multicell resource allocation problem. In this chapter, we develop a decentralized SL-based algorithm where the BSs move toward the equilibrium strategy based on their individual action-reward history.

2.2 Non-cooperative Game Theoretical Concepts

In this section, we briefly review some game-theoretical concepts which can be seen as the basis throughout the manuscript. We consider the rational and selfish game players

in the sense that a player chooses its best strategy to maximize its own benefit [12].

2.2.1 Game with External State

The four basic components of a non-cooperative game \mathcal{G} with external state are:

- The external state space \mathcal{X} . The state is represented by an independent random variable, and the transitions between the states are independent of the chosen actions.
- The set of players, $\mathcal{N} = \{1, \dots, N\}$, where N is the total number of players
- The action spaces $\mathcal{A} = \{\mathcal{A}_1, \dots, \mathcal{A}_N\}$, where \mathcal{A}_i is the set of actions that player i can take. These nonempty sets can be discrete or continuous, finite or infinite.
- The preference structure of the players. $\{u_i\}_{i \in \mathcal{N}}$ is the utility function of player i that depends on its own action as well as the actions of other players.

The *strategic form* (also called normal-form) of a game \mathcal{G} is represented by a 4-tuple:

$$\mathcal{G} = (\mathcal{X}, \mathcal{N}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \{u_i\}_{i \in \mathcal{N}}) \quad (2.1)$$

For a game with external state, the utility is defined as the expectation of the random reward, i.e.,

$$u_i(a_i, a_{-i}) = \mathbb{E}_{\mathbf{X}}[r_i | (a_i, a_{-i})],$$

where $\mathbb{E}[\cdot]$ denotes the mathematical expectation operator.

In the case of non-cooperative games, in which the players act in a selfish and independent manner, the Nash equilibrium (NE) introduced in [2] provides a solution concept of the game. It represents an operating point which is both predictable and robust to unilateral deviations (which is realistic considering the fact that the players are assumed to be non-cooperative and act in an isolated manner). This means that once the system

is operating in this state, no player has any incentive to deviate because it will lose in terms of its own benefit. The mathematical definition of the NE is as follows:

Definition 2.2.1 (Nash equilibrium). A strategy profile $\mathbf{a}^* = (a_1^*, \dots, a_N^*)$ is a (pure-strategy) Nash equilibrium if

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a'_i, a_{-i}^*), \forall i \in \mathcal{N}, a'_i \in \mathcal{A}_i \quad (2.2)$$

where $a_{-i}^* = (a_1^*, \dots, a_{i-1}^*, a_{i+1}^*, \dots, a_N^*)$ denotes the set of the other players' actions.

2.2.2 Mixed Strategy Extension

We can easily extend the non-cooperative game into a mixed strategy form as in [11]. Let p_{i,s_i} be the probability that player i selects strategy $s_i \in \mathcal{A}_i$, and $\mathbf{p}_i = [p_{i,1}, \dots, p_{i,K}]^T$ be the mixed strategy of player $i, \forall i \in \mathcal{N}$. Let \mathcal{P}_i be the set of probability distribution over the action space of player i , i.e.,

$$\mathcal{P}_i := \left\{ \mathbf{p}_i \mid p_{i,s_i} \in [0, 1], \sum_{s_i \in \mathcal{A}_i} p_{i,s_i} = 1 \right\} \quad (2.3)$$

Then, the mixed extension of utility function $\psi_i : \times_{i \in \mathcal{N}} \mathcal{P}_i \mapsto \mathbb{R}$ is defined as

$$\begin{aligned} \psi_i(\mathbf{p}_i, \mathbf{P}_{-i}) &:= \mathbb{E}_{\mathbf{p}_1, \dots, \mathbf{p}_N} [u_i] \\ &= \sum_{a_1, \dots, a_N} u_i(a_1, \dots, a_N) \left(\prod_{j=1}^N p_{j,a_j} \right). \end{aligned} \quad (2.4)$$

where \mathbf{p}_{-i} is the mixed strategy of players other than i . We have the definition of NE in mixed strategy as follows.

Definition 2.2.2 (mixed-strategy NE). A strategy profile \mathbf{P}^* is a mixed-strategy Nash equilibrium (NE) point of the non-cooperative game \mathcal{G} if and only if

$$\psi_i(\mathbf{p}_i^*, \mathbf{p}_{-i}^*) \geq \psi_i(\mathbf{p}_i, \mathbf{p}_{-i}^*), \quad \forall i \in \mathcal{N}, \forall \mathbf{p}_i \in \mathcal{P}_i \setminus \{\mathbf{p}_i^*\}. \quad (2.5)$$

2.2.3 Potential Games

While the concept of NE describes a possible steady state for a non-cooperative game, NE points do not always exist. An important class of games for which the existence of NE is guaranteed is the potential game introduced in [12]. We first define different kinds of potential games:

Definition 2.2.3. A strategic form game $\mathcal{G} = (\mathcal{N}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \{u_i\}_{i \in \mathcal{N}})$ is an exact potential game (EPG) if there exists a potential function $\Phi : \mathcal{A} \mapsto \mathbb{R}_+$ such that

$$u_i(a'_i, a_{-i}) - u_i(a_i, a_{-i}) = \Phi(a'_i, a_{-i}) - \Phi(a_i, a_{-i}), \forall i \in \mathcal{N}. \quad (2.6)$$

Definition 2.2.4. A strategic form game $\mathcal{G} = (\mathcal{N}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \{u_i\}_{i \in \mathcal{N}})$ is a weighted potential game (WPG) if there exists a potential function $\Phi : \mathcal{A} \mapsto \mathbb{R}_+$ and a weight vector $\mathbf{w} = [w_1, \dots, w_N] \in \mathbb{R}_+$ such that

$$u_i(a'_i, a_{-i}) - u_i(a_i, a_{-i}) = w_i[\Phi(a'_i, a_{-i}) - \Phi(a_i, a_{-i})], \forall i \in \mathcal{N}. \quad (2.7)$$

Definition 2.2.5. A strategic form game $\mathcal{G} = (\mathcal{N}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \{u_i\}_{i \in \mathcal{N}})$ is an ordinal potential game (OPG) if there exists a potential function $\Phi : \mathcal{A} \mapsto \mathbb{R}_+$ such that

$$u_i(a'_i, a_{-i}) \geq u_i(a_i, a_{-i}) \Leftrightarrow \Phi(a'_i, a_{-i}) \geq \Phi(a_i, a_{-i}), \forall i \in \mathcal{N}. \quad (2.8)$$

An important property of potential games is that the objectives of all players align to a *common objective*, that is, the maximization of potential function Φ . Following [12], the local maxima of the potential function are NE points of the game. Thus, every potential game has at least one pure strategy NE.

2.2.4 Achieving NE: Previous Methods

We briefly discuss two previously developed methods to achieve NE.

- **Fictitious Play**

Introduced by G.W. Brown [13], in fictitious play, each player presumes that the opponents are playing stationary (possibly mixed) strategies. At each round, each player thus best responds to the empirical frequency of play of his opponent. Such a method is of course adequate if the opponent indeed uses a stationary strategy, while it is flawed if the opponent's strategy is nonstationary. The opponent's strategy may for example be conditioned on the fictitious player's last move.

- **Best response dynamics**

Each of the players select actions sequentially. In each time slot, a player selects the action that is best response to the action chosen by the other players in the previous time slot. A best response $BR(\cdot)$ is a correspondence (multi-valued mapping) from $\prod \mathcal{A}_i \mapsto 2^{|\mathcal{A}_i|}$:

$$a_i = BR(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_N). \quad (2.9)$$

Furthermore, in finite games, the iterative best-response type algorithms converge to one of the NE states depending on the initial point.

2.3 Evolutionary Game and Replicator Dynamics

Evolutionary game theory studies the behaviors of large populations of agents who repeatedly engage in strategic interactions. Here we review the replicator dynamics, an important part of evolutionary games [14]. When considering the replicator dynamics, it is useful to think of a large population of agents who play a pre-programmed pure strategies and are randomly matched to play against each other. The growth rate of the proportion of players using a certain pure strategy is the difference between the expected

payoff of that pure strategy, given the proportions of players using every pure strategy, and the average expected payoff in that population. The strategy is inherited.

2.3.1 Replicator Dynamics

Consider a population of players. Suppose that there is some evolutionary game (two-player and symmetric) that these critters play with each other. This game has a set of pure strategies \mathcal{S} , and a payoff function $\pi(s, s')$ being the payoff to an agent playing strategy s against another agent playing s' .

Let $\phi_s(t)$ be the measures of the set of players using pure strategy s at time t , and $\theta_s(t) = \frac{\phi_s(t)}{\sum_{s'} \phi_{s'}(t)}$ be the fraction of players. Then the expected payoff to using pure strategy s at time t is $u_s(t) \triangleq \sum_{s'} \theta_{s'}(t) \pi(s, s')$, and the average utility of the whole population is $\bar{u}(t) \triangleq \sum_s \theta_s(t) u_s(t)$. Suppose that each individual is genetically programmed to play some pure strategy, and that this programming is inherited¹. Suppose that the net reproduction rate of each individual is proportional to its score in the stage game, i.e.,

$$\dot{\phi}_s(t) = \phi_s(t) u_s(t). \quad (2.10)$$

Then a continuous time dynamics of the portion can be found as

$$\begin{aligned} \dot{\theta}_s(t) &= \frac{\dot{\phi}_s(t) \sum_{s'} \phi_{s'}(t) - \phi_s(t) \sum_{s'} \dot{\phi}_{s'}(t)}{\left(\sum_{s'} \phi_{s'}(t)\right)^2} \\ &= \theta_s(t) [u_s(t) - \bar{u}(t)]. \end{aligned} \quad (2.11)$$

Equation (2.11) says that strategies with negative scores have negative net growth rates. The population size is varying; if all payoffs are negative, the entire population is shrinking. This is reasonable with the biological interpretation; in economic applications we tend to think of the number of agents playing the game as being constant. But note that even if the rewards are negative, the sum of the population shares is always unity. Note also

¹Indeed, mutation is also considered in the studies of evolutionary game theory, however it is out of the scope of this manuscript.

that if the initial share of strategy s is positive, then its share remains positive: the share can shrink towards zero, but zero is not reached in finite time. Notice that the population share of strategies that are not the best responses to other players current action can grow, as long as these strategies perform better than the population average. This is a key property that distinguishes the replicator dynamic from best-response dynamic and fictitious play.

2.3.2 Stochastic Game

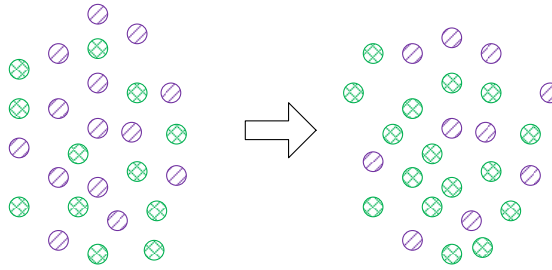
Now we change the *population* concept into a stochastic form of standard game. Formally, we may write the ODE:

$$\frac{dp_{i,s_i}(t)}{dt} = p_{i,s_i}(t) \left[\psi_i(\mathbf{e}_{s_i}, \mathbf{p}_{-i}) - \sum_{s'_i \in \mathcal{A}_i} \psi_i(\mathbf{e}_{s'_i}, \mathbf{P}) p_{i,s'_i}(t) \right]. \quad (2.12)$$

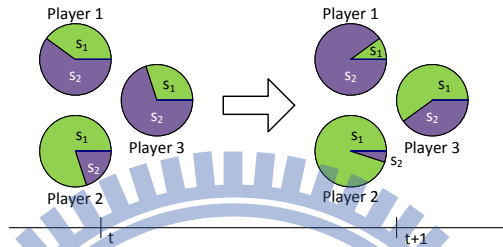
Although the game setting looks quite different, the concepts in replicator dynamics can be applied. The two interpretations are shown in Figure 2.1. Figure 2.1(a) shows an evolutionary game with two types of players. The population of players taking each strategy changes. On the other hand, the strategic game in Figure 2.1(b), the number of players is fixed, and the weighting of each strategy changes. A detailed comparison is given in Table 2.1.

Table 2.1: Comparison

Property	Evolutionary Game	Stochastic Game
Player rationality	not rational	rational
Strategy adopted by each player	same strategy as inherited	mixed-strategy with varying weight
Variables in replicator equations	population share	probability of each strategy
Strategy with higher reward	population growth	higher probability



(a) Evolutionary Game



(b) Stochastic Game

Figure 2.1: Replicator dynamics in evolutionary game and stochastic game.

2.4 Stochastic Learning Algorithm

In this section, we present the structure of the stochastic learning algorithm (SLA), which will be used in later section. When the learning is applied, it has two major advantages over conventional methods. First, the SLA is robust against external states: the learned strategy for each player. Second, Learning under limited information. According to the available information for individual players, the learning is classified as follows.

1. **Fully-distributed learning:** The available information is restricted to action-reward history of each individual player. A player knows nothing about its opponents. Fully distributed learning is usually applied when the payoff is given by an *outsider* which is not a member of the player set.
2. **Distributed learning with partial cooperation:** Sometimes, the reward is calculated by individual player instead of obtained from the environment. In this case, the players may own partial knowledge of other players including their past

actions and the observation on external states. However, each player keeps its own learning process, and the decision making is uncoupled. Notice that the major difference between uncoupled learning and BRD is that the former allows simultaneous strategy updates of players, while the latter requires an implicit ordering of strategy updates.

In this thesis, the examples considered include both cases. Table 2.2 summarizes the information available to the players.

Table 2.2: Information Available to the Players

Information	Fully-distributed Learning	Distributed learning with partial cooperation
Awareness of being in a game	No	Yes
Existence of opponents	No	Partial
Observation of external state	No	Partial
Action spaces of the others	No	No
Joint strategy	No	No
Current action of others	No	No
Last action of the opponents	No	Partial
Last own-action	Yes	Yes
Observation of own reward	Yes	Yes
Own reward function form	No	Yes
Reward function form of the others	No	No

Table 2.3: Summary of Notations in Game-theoretic Formulation

Symbol	Meaning
\mathcal{X}	external state space
\mathbf{X}	random matrix for the external state
\mathcal{N}	set of players
\mathcal{A}_i	set of actions of player i
$s_i \in \mathcal{A}_i$	an element of \mathcal{A}_i
$a_i(n) \in \mathcal{A}_i$	action of player i at iteration n
$a_{-i}(n) \in \mathcal{A}_i$	actions of players except for i at iteration n
$\mathcal{P}_i := \Delta(\mathcal{A}_i)$	set of probability distribution over \mathcal{A}_i
$\mathbf{p}_i(n) \in \mathcal{P}_i$	mixed strategy of player i at slot n
$r_i(n) \in \mathbb{R}$	instantaneous reward of player i at slot n
$\hat{\mathbf{u}}_i(n) \in \mathbb{R}^{ \mathcal{A}_i }$	estimated utility vector of player i at slot n

2.4.1 Generic SLA Structure

Under the SLA, the players can learn their expected payoffs and their optimal strategies by using some simple iterative techniques based on their action-reward history. The actions that give good performance are reinforced and new actions are explored. Therefore, such an approach belongs to the *reinforcement learning*.

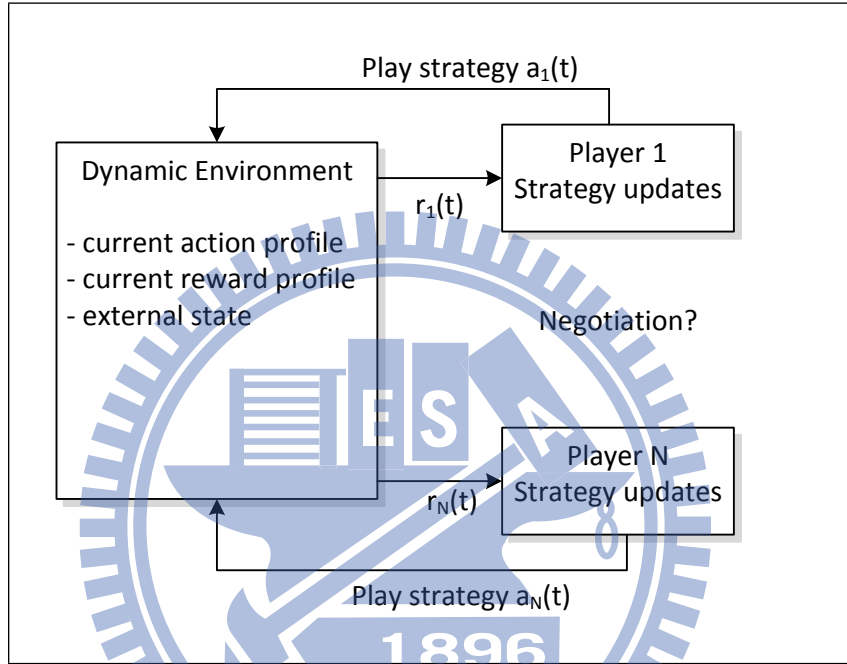


Figure 2.2: Generic SLA structure.

The generic stochastic learning algorithm is described in Algorithm 2.1. In each

Algorithm 2.1 Generic Stochastic Learning

- 1: Initially, set $n = 0$, and the action probability vector as

$$p_{i,s_i}(0) = 1/|\mathcal{A}_i|, \hat{u}_{i,s_i}(-1) = 0, \quad \forall i \in \mathcal{N}, s_i \in \mathcal{A}_i.$$
 - 2: At the beginning of the n th iteration, each player selects an action $a_i(n)$ according to the current action probability vector (i.e., mixed strategy) $\mathbf{p}_i(n)$.
 - 3: At the completion of the n th iteration, each player calculates or receives the instantaneous reward $r_i(n)$.
 - 4: All players update their utility estimation and action probability vector according to the *update rules*.
-

iteration, the action is selected based on the probability distribution over the strategy

set of each player. After each iteration, a player obtains the instantaneous reward from the outsider or through calculation. It updates the action probability vector (i.e., mixed strategy) $\mathbf{p}_i(n)$ as well as the utility estimation vector $\hat{\mathbf{u}}_i(n)$. The utility estimation serves as a reinforcement signal so that higher utility (lower cost) leads to higher probability in the next play. Notably, the proposed learning algorithm is distributed: the strategy selection is based on individual observations instead of the guide from a centralized controller. The update rules for action probability vector and utility estimation are investigated in the next section.

2.5 Update Rules

The general form of an update rule can be expressed as:

$$\begin{cases} \hat{\mathbf{u}}_i(n+1) = f_i(\lambda_i(n), a_i(n), r_i(n), \hat{\mathbf{u}}_i(n), \mathbf{p}_i(n)) \\ \mathbf{p}_i(n+1) = g_i(\nu_i(n), a_i(n), r_i(n), \hat{\mathbf{u}}_i(n), \mathbf{p}_i(n)) \end{cases} \quad (2.13)$$

where $\lambda(n), \nu(n)$ are the *learning rates* for the utility estimation and action probability, respectively. Their values are carefully chosen so that

$$\begin{aligned} \lambda_i(t) \geq 0, \quad \sum_t \lambda_i(t) = +\infty, \quad \sum_t \lambda_i^2(t) < \infty, \\ \nu_i(t) \geq 0, \quad \sum_t \nu_i(t) = +\infty, \quad \sum_t \nu_i^2(t) < \infty. \end{aligned} \quad (2.14)$$

In this section, we introduce two probability update rules, namely, the Bush-Mosteller update rule and the multiplicative-weight update rule. We investigate their ODE approximations.

2.5.1 Bush-Mosteller (BM) Update Rule

With BM update rule, the mixed strategies are updated as follows:

$$\begin{cases} p_{i,s_i}(n+1) = p_{i,s_i}(n) + b\tilde{r}_i(n)(1 - p_{i,s_i}(n)), & s_i = a_i(n) \\ p_{i,s_i}(n+1) = p_{i,s_i}(n) - b\tilde{r}_i(n)p_{i,s_i}(n), & s_i \neq a_i(n) \end{cases} \quad (2.15)$$

where $\tilde{r}_i(n) \in [0, 1]$ is the normalized instantaneous reward, i.e.,

$$\tilde{r}_i(t) = \frac{r_i(t) - r_{min}}{r_{max} - r_{min}}, \quad (2.16)$$

where r_{max} and r_{min} proposition.

Proposition 2.5.1. *With sufficiently small b , the probability matrix sequence $\{\mathbf{P}(n)\}$ converges to \mathbf{P}^* which is the solution of the following ODE:*

$$\frac{dp_{i,s_i}(t)}{dt} = p_{i,s_i}(t) [\psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) - \psi_i(\mathbf{P})] \quad (2.17)$$

The boundary condition is given by $\mathbf{P}(0) = \mathbf{P}_0$, where \mathbf{P}_0 is the initial action probability matrix.

Although the SL-based algorithm with BM rule converges to NE points for potential games, it requires the normalization of the instant reward. This requirement makes the algorithm inapplicable when the extreme values of reward functions are unavailable. Therefore, another update rule is also considered in our works.

2.5.2 Multiplicative-weight Update Rule

The multiplicative-weight update rule consists of the iterative updates for utility estimations and mixed strategies. The rule is described as follows:

$$\begin{cases} \hat{u}_{i,s_i}(n+1) - \hat{u}_{i,s_i}(n) = \eta_i \mathbb{1}_{\{a_i(n)=s_i\}} (r_i(n) - \hat{u}_{i,s_i}(n)) \\ p_{i,s_i}(n+1) = \frac{p_{i,s_i}(n)(1+\epsilon_i)^{\hat{u}_{i,s_i}(n)}}{\sum_{s'_i \in \mathcal{A}_i} p_{i,s'_i}(n)(1+\epsilon_i)^{\hat{u}_{i,s'_i}(n)}} \end{cases} \quad (2.18)$$

where η_i and ϵ_i are the learning rates for utility estimation and action probability, respectively.

Its ODE approximation is discussed in the following proposition. First, by using the ordinary differential equation (ODE) approximation we characterize the long-term behavior of the sequence $\{\mathbf{P}(n)\}$. Second, we establish a sufficient condition for the arrival at NE points for the proposed learning algorithm and prove that the game \mathcal{G} satisfies this condition.

Proposition 2.5.2. *With sufficiently small learning rates η and ϵ :*

1. *The estimated utility converges to*

$$\hat{u}_{i,s_i} \rightarrow \psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i}). \quad (2.19)$$

2. *Asymptotically, the probability matrix sequence $\{\mathbf{P}(k)\}$ can be approximated by the trajectory of the following ODE:*

$$\frac{dp_{i,s_i}(t)}{dt} = p_{i,s_i}(t) [\psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) - \psi_i(\mathbf{P})] \quad (2.20)$$

where $p_{m,s_i}(t)$ is the continuous-time version of $p_{i,s_i}(n)$, and the boundary condition is given by $\mathbf{P}(0) = \mathbf{P}_0$, where \mathbf{P}_0 is the initial mixed strategy matrix.

Proof: For better understanding, we reproduce the proof from [7, Section 4.3]. From the theory of stochastic approximation, the update of the estimated utility in (2.18)

can be given as

$$\hat{u}_{i,s_i} \rightarrow \psi_i(s_i, \mathbf{P}), \text{ if } \eta_i \rightarrow 0, \quad (2.21)$$

and the tracked ODE can be given as [7, Section 4.3], [11, Theorem 3.1]

$$\frac{dp_{i,s_i}(t)}{dt} = \lim_{\epsilon_i \rightarrow 0} \frac{p_{i,s_i}(n+1) - p_{i,s_i}(n)}{\epsilon_i}. \quad (2.22)$$

Next we will show the RHS of the above is exactly that of (2.20).

Let $S = \sum_{s'_i \in \mathcal{A}_i} p_{i,s'_i}(t)(1 - \epsilon_i)^{-\hat{u}_{i,s'_i}}$. Then we have

$$\begin{aligned} & \frac{p_{i,s_i}(t+1) - p_{i,s_i}(t)}{\epsilon_i} \\ &= \frac{p_{i,s_i}(t)}{\epsilon_i} \left[\frac{(1 - \epsilon_i)^{-\hat{u}_{i,s_i}}}{S} - 1 \right] \\ &= \frac{p_{i,s_i}(t)}{S} \left[\frac{(1 - \epsilon_i)^{-\hat{u}_{i,s_i}} - 1 + 1 - S}{\epsilon_i} \right] \\ &= \frac{p_{i,s_i}(t)}{S} \left[\frac{(1 - \epsilon_i)^{-\hat{u}_{i,s_i}} - 1}{\epsilon_i} - \sum_{s'_i \in \mathcal{A}_i} p_{i,s'_i} \left(\frac{(1 - \epsilon_i)^{-\hat{u}_{i,s'_i}} - 1}{\epsilon_i} \right) \right], \end{aligned}$$

where we have employed the update rule for $p_{i,s_i}(t+1)$ in (12) of the manuscript to obtain the first equality.

With the result of (2.21) and $\lim_{\epsilon \rightarrow 0} \frac{(1-\epsilon)^{-u}-1}{\epsilon} = u$, it follows that

$$\begin{aligned} & \lim_{\epsilon_i \rightarrow 0} \frac{p_{i,s_i}(t+1) - p_{i,s_i}(t)}{\epsilon_i} \\ &= p_{i,s_i}(t) \left[\psi_i(\mathbf{e}_{s_i}, \mathbf{P}) - \sum_{s'_i \in \mathcal{A}_i} \psi_i(\mathbf{e}_{s'_i}, \mathbf{P}) p_{i,s'_i}(t) \right] \\ &= p_{i,s_i}(t) [\psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) - \psi_i(\mathbf{P})], \end{aligned} \quad (2.23)$$

where we have used the fact that $\lim_{\epsilon \rightarrow 0} S = 1$. Combining (2.22) and (2.23) above we complete the proof. \blacksquare

Notice that $\psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i})$ is the utility of player m if it employs pure strategy s_m while

other player m' , $\forall m' \in \mathcal{M}, m' \neq m$ employs a mixed strategy $\mathbf{p}_{m'}$, and its value is learned by player m as the estimated utility \hat{u}_{m,s_m} , as shown in (2.19). On the other hand, the ODE for mixed-strategy in (2.20) is the *replicator equation* [14] in which the probability of taking one strategy increases if the current estimated utility of this strategy is larger than the average utility over all strategies and decreases otherwise. Compared to the best response dynamics [12] where a player changes its strategy in the next iteration to the best action according to other players' actions (i.e., the best response), with the replicator dynamics, a player selects an action according to a probability distribution over the strategy set, and adjusts the weighting for each possible action in each iteration based on the utility estimation.

2.6 Convergence of the Proposed Algorithm

Convergence toward pure strategy NE points is an important feature of the proposed learning algorithm. Similar to the discussions in [11] and [10], here we theoretically demonstrate the convergence properties of the proposed SL-based algorithm. First, by using the ordinary differential equation (ODE) approximation we characterize the long-term behavior of the sequence $\{\mathbf{P}(n)\}$. Second, we establish a sufficient condition for the arrival at NE points for the proposed learning algorithm and prove that the game \mathcal{G} satisfies this condition.

Note that the ODE in (2.20) is the *replicator equation* [14] in which the probability of taking one strategy grows if this strategy's current estimated utility is larger than the average utility over all strategies and declines otherwise. Compared to the best response dynamics where a player changes its strategy in the next iteration to the best action according to other players' action, a player adjusts the weighting for each possible action in each iteration with the replicator dynamics.

Proposition 2.6.1 (Folk theorems). *The proposed learning algorithm has the following properties:*

1. All Nash equilibria are stationary points of (2.20);
2. All stationary points of (2.20) are Nash equilibria.

Proposition 2.6.1 is an instance of the Folk theorems in evolutionary game theory [14], and these properties follow directly from the replicator equation in (2.20). For an intuitive explanation, observe that $\psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i})$ is the expected reward function of player i if it employs pure strategy s_i while other player $j, \forall j \in \mathcal{N}, j \neq i$ employs a mixed strategy \mathbf{p}_j . From the definition of Nash equilibrium, the condition

$$\psi_i(\mathbf{e}_{s_i^*}, \mathbf{P}_{-i}^*) = \psi_i(\mathbf{P}^*), \quad \forall i \in \mathcal{N}, s_i \in \mathcal{A}_i \text{ with } p_{i,s_i}^* > 0 \quad (2.24)$$

must hold for an NE strategy profile \mathbf{P}^* . Therefore any Nash equilibrium must lead the right-hand side of (2.20) to zero, and thus constitutes a stationary point of (2.20). It is worth noting that, for a mixed-strategy NE, all survived pure strategies (i.e. s_i with $p_{i,s_i} > 0$) of player i perform equally well when other players follow the mixed strategy \mathbf{P}_{-i}^* .

From the ODE approximation, we find a way to describe the asymptotic behavior of the discrete updates of the mixed strategies for different update rules. In the following, we investigate the convergence property.

2.6.1 Potential Games

We first consider the case that the game is a potential game.

Proposition 2.6.2. *Suppose that there exists a bounded differentiable function $\Psi : \mathbb{R}^{|\mathcal{A}|} \rightarrow \mathbb{R}$ such that*

$$\Psi(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) = \frac{\partial \Psi(\mathbf{P})}{\partial p_{i,s_i}} \quad (2.25)$$

is an increasing function of $\psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i})$. Then, the SL-based algorithm converges to an NE point of a noncooperative game.

Proof. First, we rewrite the ODE in (2.20) as follows:

$$\frac{dp_{i,s_i}(t)}{dt} = p_{i,s_i}(t) \sum_{s'_i \in \mathcal{A}_i} p_{i,s'_i}(t) \left[\psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) - \psi_i(\mathbf{e}_{s'_i}, \mathbf{P}_{-i}) \right]. \quad (2.26)$$

Given that $\Psi(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) = \partial\Psi(\mathbf{P})/\partial p_{i,s_i}$ is an increasing function of $\psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i})$, and let $D_{i,s_i,s'_i} = \psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) - \psi_i(\mathbf{e}_{s'_i}, \mathbf{P}_{-i})$, $E_{i,s_i,s'_i} = \Psi(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) - \Psi(\mathbf{e}_{s'_i}, \mathbf{P}_{-i})$, we may write

$$D_{i,s_i,s'_i} > 0 \Leftrightarrow E_{i,s_i,s'_i} > 0. \quad (2.27)$$

By applying (2.26) and (2.27), the derivation of $\Psi(\mathbf{P})$ with respect to t is given by

$$\begin{aligned} \frac{d\Psi(\mathbf{P})}{dt} &= \sum_{i \in \mathcal{N}} \sum_{s_i \in \mathcal{A}_i} \frac{\partial\Psi(\mathbf{P})}{\partial p_{i,s_i}} \frac{dp_{i,s_i}}{dt} \\ &= \sum_{i \in \mathcal{N}} \sum_{s_i, s'_i \in \mathcal{A}_i} p_{i,s_i} p_{i,s'_i} \Psi(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) \cdot D_{i,s_i,s'_i} \\ &= \frac{1}{2} \sum_{i \in \mathcal{N}} \sum_{\substack{s_i, s'_i \in \mathcal{A}_i \\ s_i < s'_i}} p_{i,s_i} p_{i,s'_i} E_{i,s_i,s'_i} \cdot D_{i,s_i,s'_i} \\ &\geq 0 \end{aligned} \quad (2.28)$$

where the last inequality holds since given the condition in (2.27), D_{i,s_i,s'_i} and E_{i,s_i,s'_i} always have the same sign.

Thus $\Psi(\cdot)$ is non-decreasing along the trajectories of the ODE, and asymptotically all the trajectories will be in the set $\{\mathbf{P} \in \mathcal{P} : \frac{d\Psi(\mathbf{P})}{dt} = 0\}$. From (2.26) and (2.28), the following is known:

$$\begin{aligned} \frac{d\Psi(\mathbf{P})}{dt} &= 0 \\ \Rightarrow p_{i,s_i} p_{i,s'_i} \left[\psi_i(\mathbf{e}_{s_i}, \mathbf{P}_{-i}) - \psi_i(\mathbf{e}_{s'_i}, \mathbf{P}_{-i}) \right]^2 &= 0, \quad \forall i, s_i, s'_i \\ \Rightarrow \frac{dp_{i,s_i}}{dt} &= 0, \quad \forall i, s_i, s'_i \\ \Rightarrow \mathbf{P} &\text{ is a stationary point of the ODE (2.20).} \end{aligned} \quad (2.29)$$

In other words, when starting from an interior point of the simplex of the mixed strategy

space \mathcal{P} , the sequence $\mathbf{P}(n)$ converges to a stationary point of the ODE in (2.26). By Proposition 2.6.2, we complete the proof. \square

Proposition 2.6.2 establishes a sufficient condition that guarantees the convergence toward NE. In what follows, we prove that an ordinal potential game \mathcal{G} satisfies this condition and hence it converges to a pure-strategy NE point by using the SL-based algorithm.

Proposition 2.6.3. *When applied to OPGs, the proposed SLA with both update rules converges to a (possible mixed-strategy) NE point.*

Proof. For OPGs, let $\Psi(\mathbf{P})$ be the mixed extension of the potential function,

$$\Psi(\mathbf{P}) = \sum_{a_l, l \neq i} \Phi(a_1, \dots, a_N) \prod_{j \neq i} p_{j, a_j}. \quad (2.30)$$

By extending the definition of OPG into mixed-strategy, we have that for OPGs

$$\Psi(\mathbf{e}_{s'_i}, \mathbf{P}) - \Psi(s_i, \mathbf{P}) > 0 \Leftrightarrow \psi_i(s'_i, \mathbf{P}) - \psi_i(s_i, \mathbf{P}) > 0 \quad (2.31)$$

$\forall s_i, s'_i \in \mathcal{A}_i, \forall i \in \mathcal{N}$. By Proposition 2.6.2, we complete the proof. \square

Corollary 2.6.1. *When applied to WPGs and EPGs, the proposed SLA with both update rules converges to a (possible mixed-strategy) NE point.*

Note that the learning rates (ϵ_i, η_i) play an important role in the convergence behavior of the proposed SL-based learning algorithm. In particular, smaller learning rates lead to a slower convergence. The choice of learning rates poses a trade-off between accuracy and speed, and may be determined by training in practice.

Remark 2.6.1. Propositions 2.6.2 and 2.6.3 do not guarantee the convergence toward a *pure-strategy* NE. However, our simulation shows that a pure-strategy NE rather than a mixed-strategy NE is usually achieved.

2.6.2 Non-potential Games

While the OPG already relaxes the constraints of problem formulation, there are cases that a potential game cannot be formed. When trying to apply the SLA, we encounter two major questions:

- (1) Does the SLA still converge?
- (2) If the SLA converges, what are the properties of the resulting strategy profile (e.g., is it NE point)?

Similar to MAQL, the convergence is not theoretically guaranteed but usually observed in practical applications. We may also set the limitation of maximum number of rounds to avoid an infinite loop. Furthermore, due to the stochastic approximation to the trajectory of replicator dynamics, the remaining mixed strategy is a kind of *good* strategy against the opponents, though may not be NE point. Therefore, while the proposed SLA possesses some good properties when applied to potential games, we believe that it is still suitable for other problem formulations in which learning is required.

Proposition 2.6.4. *If the proposed algorithm converges to a stationary point of (2.20), the limiting point must be a (possibly mixed-strategy) NE point.*

2.7 Applications: Game Theoretic Modeling

The last section of this chapter is devoted to an overview of how to establish a game theoretic formulation for a radio resource management (RRM) problem in wireless communication systems. A mapping of game theory components to RRM problem is given in Table 2.4.

The players in the game are the mobile users and/or the networks. Players seeking to maximize their payoffs can choose between different strategies, such as: available bandwidth, subscription plan, or available service providers. The payoffs can be estimated

Table 2.4: Summary of Symbols for Game-theoretic Formulation

Game Component	Network Selection Environment Correspondent
Players	The agents who are playing the game: users or/and networks
Strategies	A plan of actions to be taken by the player during the game: available/requested bandwidth, subscription plan, offered prices, available service providers, etc.
Payoffs	The motivation of players represented by profit and estimated using utility functions based on various parameters: monetary cost, quality, network load, QoS, etc.
Resources	The resources for which the players involved in the game are competing: bandwidth, power, etc.
External State	The external state for the game that is not controlled by the players: channel availability, channel coefficients, etc.

using utility functions based on various decision criteria: monetary cost, energy conservation, network load, availability, etc. The games can be formulated so that they can target different objectives, such as maximizing or minimizing different resources - bandwidth, power, etc.

Appendix 2.A Assumptions for Stochastic Approximation

In this appendix, we summarize the basic assumptions for stochastic approximation. Please refer to [6] for more details.

Consider the difference equation $\mathbf{p}(n+1) = \mathbf{p}(n) + \lambda(n)(f(x(n)) + M(n+1))$ in $\mathbb{R}^{|A|}$ and assume that

(A1.) f is Lipschitz.

(A2.) $\lambda(n) \geq 0$, $\sum_{n \geq 0} \lambda(n) = +\infty$, $\sum_{n \geq 0} \lambda^2(n) < +\infty$.

(A3.) $M(n + 1)$ is a martingale difference sequence with respect to the increasing family of sigma-fields $\mathcal{F}(n) = \sigma(x(n'), \hat{u}(n'), M(n'), n' \leq n)$, i.e., $\mathbb{E}[M(n + 1)|\mathcal{F}(n)] = 0$.

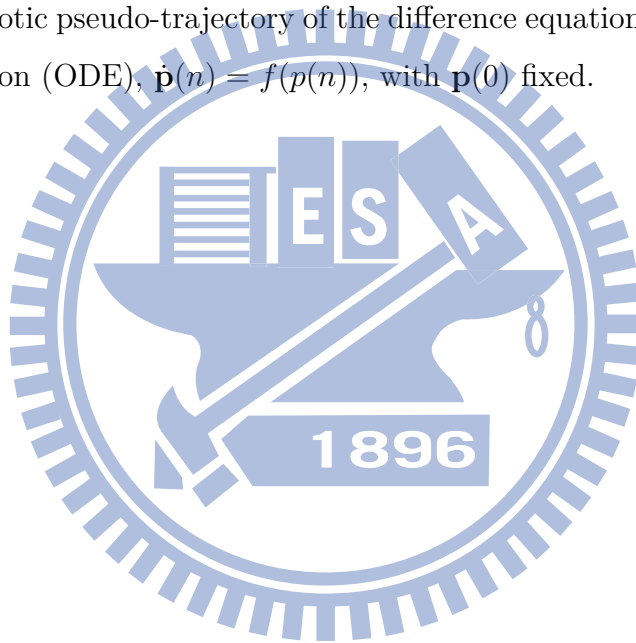
(A4.) $M(n)$ is square integratable and there is a constant $c > 0$ such that

$$\mathbb{E}[M(n + 1)|\mathcal{F}(n)] \leq c(1 + \|\mathbf{x}(n)\|^2) \quad (2.32)$$

almost surely, for all $t \geq 0$.

(A5.) $\sup_n \|\mathbf{p}(n)\| < \infty$ almost surely.

Then, the asymptotic pseudo-trajectory of the difference equation is given by the ordinary differential equation (ODE), $\dot{\mathbf{p}}(n) = f(p(n))$, with $\mathbf{p}(0)$ fixed.



Chapter 3

A Survey on the Spectrum Access of Cognitive Radio Networks

Before showing the application examples, we provide a survey on cognitive radio networks in this chapter. An overview on different access scenarios in cognitive radio networks is given first. Then the examples are given in brief, with the pros and cons.

3.1 Cognitive Spectrum Access

IN cognitive radio networks (CRNs), the cognitive radio (CR) users obtain the spectrum access rights in different ways. By extending the work of Akyildiz *et al.* [15] with investigations afterwards, we categorized the spectrum access scenario of CRNs into four different types. The four spectrum access scenarios of CRNs are depicted in Fig. 3.1 and also briefly introduced as follows.

Opportunistic spectrum access. Nodes in CRNs communicate with each other in an ad-hoc manner on both licensed and unlicensed spectrum bands. Each connection opportunistically access the spectrum with consideration on the co-tier and cross-tier interference.

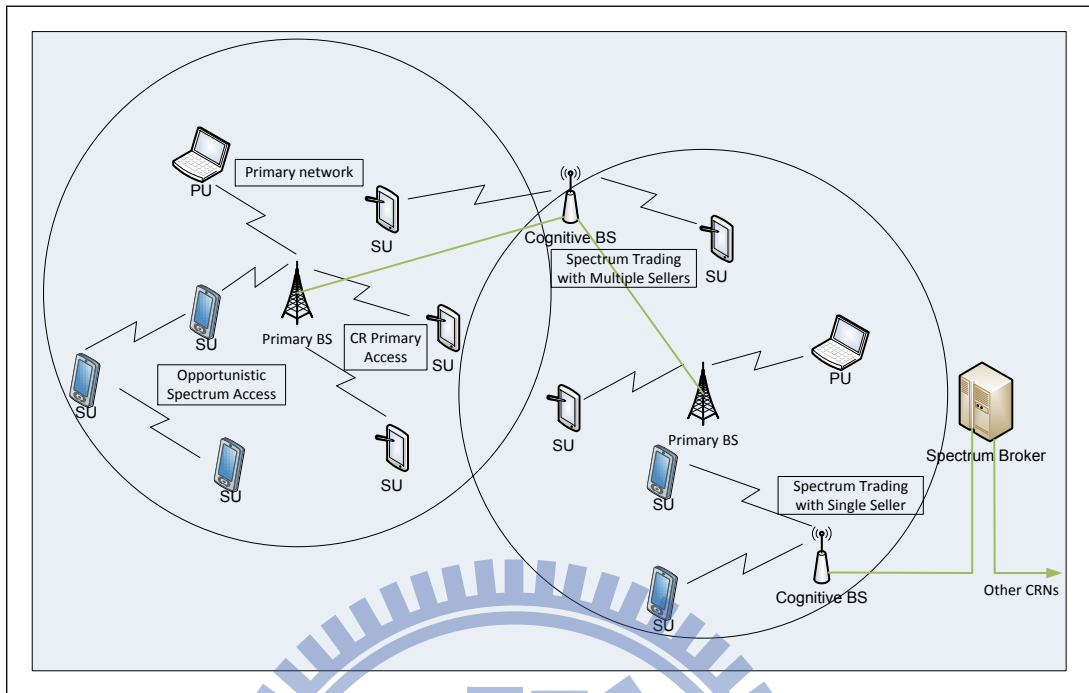


Figure 3.1: Cognitive radio network architecture.

Spectrum trading with single seller. CR users communicate via their own CR base stations (CRBSs), on both licensed and unlicensed spectrum bands. The CRBS determines the amount of resources (i.e., bandwidth) to request from the spectrum seller. When the spectrum is granted by the seller, since all interactions occur inside the CR cluster, the spectrum sharing policy can be independent of that of the primary network.

Spectrum trading with multiple sellers. The CRBS may choose to request the spectrum from different sellers. Therefore, it has to determine the best seller as well as the amount of requested spectrum. Then the CR users access the spectrum through their own CRBSs.

CR primary access. CR users can access the primary base station through the licensed band. CR users require a re-configurable medium access control (MAC) protocol, which enables roaming over multiple primary networks with different access technologies.

3.2 Opportunistic Spectrum Access

Opportunistic spectrum access (OSA) is a promising technique for tackling the spectrum scarcity problem by exploiting the temporally unutilized spectrum bands [16, 17]. In the CR ad hoc access model, the CR links access the spectrum based on sensing and contention. We provide four examples.

Learning-Based OSA with adaptive hopping. Derakhshani and Le-Ngoc [18] presented an adaptive hopping transmission strategy for secondary users (SUs) to access temporarily idle frequency-slots of a licensed frequency band in consideration of the random return of primary users (PUs), aiming to maximize the overall SU throughput.

Learning the hidden Markov model. The work of Choi *et al.* [19] is based on learning consider the hidden Markov model (HMM) and partially observable Markov decision process (POMDP).

Learning under unknown environments. Xu *et al.* [10] considered opportunistic spectrum access in which the CR links contend for the spectrum. The strategy is the channel selection. A CR can access a channel if it wins the contention and the PU is not using this channel. Under the unknown dynamic environment, stochastic learning algorithm is applied to learn the equilibrium of the expected game.

Adaptive channel recommendation. Chen *et al.* [20] proposed a dynamic spectrum access scheme where secondary users cooperatively recommend “good” channels to each other and access accordingly. The spectrum access problem was formulated as an average reward-based Markov decision process (MDP).

OSA for mobile CR. While most existing work focuses on enabling OSA for stationary CRs, Min *et al.* [21] considered mobility of secondary users (SUs). In this work, the channel availability experienced by a mobile SU was modeled as a two-state continuous-time Markov chain (CTMC). To protect PU communications from SU interference, the authors introduce guard distance in the space domain and derive the optimal guard distance that

maximizes the spatio-temporal spectrum opportunities available to mobile CRs. To facilitate efficient spectrum sharing, the secondary network throughput maximization was formulated as a convex optimization problem, and an optimal, distributed channel selection strategy was derived.

3.3 Spectrum Trading with Single Seller

Spectrum trading with single seller has been studied in [22–25]. In the simplest spectrum market, the seller can be interpreted as the only SP or a spectrum broker who collects the residual spectrum from several SPs. The strategy of the seller is the spectrum price.

Cournot game. Niyato *et al.* [22] formulated the bandwidth demand of SUs as a Cournot competition. The unit spectrum price increases with the total demand. Cournot equilibrium is achieved.

Monopolist in spectrum market. Gao *et al.* [23] considered the single SP as the monopolist in a spectrum market.

Auction under imperfect spectrum sensing. Tehrani and Uysal [24] consider the shared used model in cognitive radio networks and design a spectrum trading method to maximize the total satisfaction of the Secondary Users (SUs) and revenue of the Wireless Service Provider (WSP). Specifically, this work considered the risk of imperfect spectrum sensing in spectrum auction.

Random access. Pricing-based spectrum management with random access was considered in [25], where the SUs contend for spectrum access.

3.4 Spectrum Trading with Multiple Seller

When multiple spectrum seller exists, a multi-level framework considering both the behaviors of SPs and SUs is established [26–34]. The spectrum trading mechanism can

be classified according to the way the rights of spectrum access are granted.

We consider a problem of dynamic spectrum leasing in a spectrum secondary market of cognitive radio networks where secondary service providers lease spectrum from spectrum brokers to provide service to secondary users. We first consider the intuitive case in which the spectrum access is determined through iterative negotiations, either between PO and SU or among SUs. In such a spectrum leasing scenario, market mechanism is involved in the spectrum there exists one or more spectrum owner, and the secondary users pay to obtain the right of channel utilization. In the following, we use several example applications of stochastic game theory to cognitive radio networking to illustrate how to formulate a stochastic game for different problems and how to solve the game.

3.4.1 Exclusive Access

In the *exclusive access* model [26–31], the SPs set the per-channel prices, and a channel is exclusively assigned to an SU when it pays the announced price. Under such setting, the game is limited to the competition among SPs on the spectrum price. Xing *et al.* [26] considered discrete price levels and applied stochastic learning algorithm to help the SPs select proper pricing strategies. In [27,28], auction-based spectrum trading was conducted. Bargaining-based approach was considered in [29], which allows the short-term spectrum trading among SUs after the long-term spectrum leasing from SPs. Three-stage trading model was adopted in [30,31], where agents profit from buying the spectrum opportunities from the owners and selling them to the SUs. The major drawback of the exclusive access model is that a negotiation process among sellers and buyers is required before the SUs can start transmission.

Multi-auctioneer Problem (MAP). Gao *et al.* [27] proposed an auction-based mechanism with multiple auctioneers. In the system model, each PO has a number of unoccupied channels for SUs to lease. The POs are considered as auctioneers trying to sell its channels to SUs. On the other hand, the SUs select the preferred PO based on the value to itself and the price announced. The auction is run repeatedly, in which each auctioneer starts

from a lower reserved price, and increase the price if the bid is more than its quota. Equilibrium (defined in the paper) was observed if the the price adjustment step size is small enough. The equilibrium may not be NE, but at least the resulting state is tractable.

Repeated Auction with Bayesian Learning The work of Han *et al.* [35] models the spectrum access in CRN as a repeated auction game subject to sensing costs and the cost of transmission (upon successful bidding for a channel). The formulation is a dynamic game with incomplete information, as the information about other SUs' action is limited. A Bayesian non-parametric belief update scheme is constructed based on the Dirichlet process. In the proposed bidding learning algorithms, SUs can decide whether or not to participate in the bidding according to the belief update.

3.4.2 Shared Access

On the other hand, in the *shared access* model [32–34], the SPs set the subscription prices, and the exact bandwidth assigned depends on the number of SUs sharing the spectrum opportunity of the same SP. Although the bandwidth of each SU is no longer guaranteed, this model avoids the overhead for negotiation. The interaction of SUs is often modeled as a population game [36] due to their distributed nature.

Game-theoretic modeling with multiple sellers and buyers. Niyato *et al.* [32] proposed a game-theoretic framework for the spectrum trading with multiple primary users selling spectrum opportunities to multiple secondary users. The secondary users can adapt the spectrum buying behavior (i.e., evolve) by observing the variations in price and quality of spectrum offered by the different primary users or primary service providers. On the other hand, The primary users adjust their behavior in terms of size of offered spectrum to the secondary users and spectrum price to achieve the highest utility. Evolutionary game theory was considered for the dynamic behavior of secondary users. For the PUs, an iterative algorithm for strategy adaptation was presented.

Dynamic spectrum leasing in secondary market. Zhu *et al.* [33] considered a hierarchical spectrum leasing scenario and developed a two-level dynamic game framework. In

a scenario, secondary service providers lease spectrum from spectrum brokers to provide service to secondary users who are also choosing the service providers. At the lower layer, the dynamic service selection is modeled as an evolutionary game, and the replicator dynamics is applied to model the service selection adaptation and the evolutionary equilibrium is considered to be the solution. At the upper layer, With dynamic service selection, competitive secondary providers dynamically lease spectrum to provide service to secondary users. A spectrum leasing differential game was formulated to model this competition at the upper level. [33] adopted evolutionary equilibrium in the lower-level, while the upper-level competition is modeled as a differential game.

Queueing-based model. Elias *et al.* [34] addressed the joint pricing and network selection problem in cognitive radio networks. This paper studied the steady-state performance of SUs, focusing on delay as the quality of service (QoS) metric.

These methods, however, require the knowledge of the opponents' actions and are difficult to be implemented in distributed systems. Therefore, it is desirable to find self-organized spectrum trading in which the nodes (viz. service providers and secondary users) act independently.

3.5 CR Primary Access

While most approaches involve negotiations, fully-distributed learning algorithm is newly introduced.

Hybrid Learning in 4G heterogeneous networks. Khan *et al.* [9] proposed a fully distributed method, namely, the hybrid learning, for network selection in 4G heterogeneous networks. The users, as learning automata, are embedded with different learning rules. The convergence towards a pure strategy profile was demonstrated without indicating whether or not the achieved strategy profile is an equilibrium point.

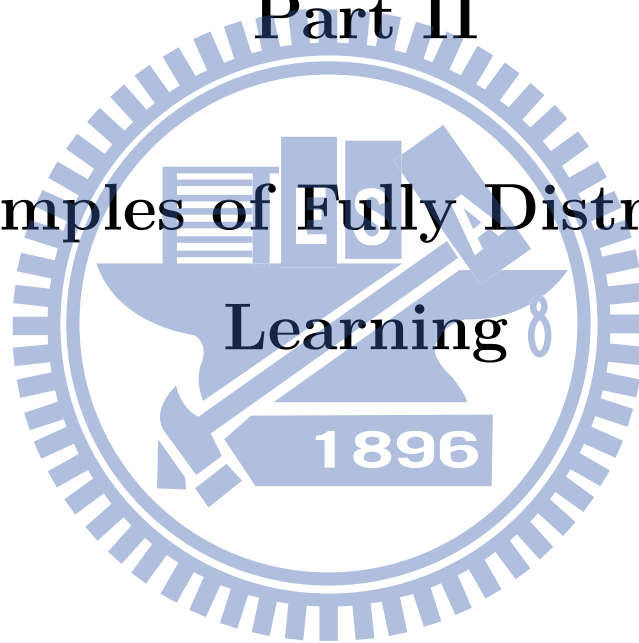
To realize spectrum trading between independent nodes with unknown dynamic spectrum opportunities, our work tackles with fully distributed operations based on stochastic

learning. In addition to the spectrum trading (with price competition only) in [26], stochastic learning has been adopted in different areas of wireless communications. Examples include the precoding strategy in multi-antenna systems [8], network selection in 4G heterogeneous networks [9], and opportunistic spectrum access in CRNs [10]. However, to the best of our knowledge, the fully-distributed operation in self-organized spectrum trading with two-level competitions has yet been extensively investigated, whether based on stochastic learning or not.



Part II

Examples of Fully Distributed Learning



Chapter 4

Network Selection in Cognitive Heterogeneous Networks

COEEXISTENCE of multiple radio access technologies (RATs) is a promising paradigm to improve spectrum efficiency. This chapter presents a game-theoretic study of network selection in a cognitive heterogeneous networking environment with time-varying channels. We formulate the network selection problem as a non-cooperative game with secondary users (SUs) as the players, and show that the game is an ordinal potential game (OPG). A decentralized, stochastic learning-based algorithm is proposed where each SU progressively moves toward the Nash equilibrium (NE) based on its action-reward history and not actions taken by others. The convergence properties of the proposed algorithm toward a pure-strategy NE point are theoretically and numerically verified. The proposed algorithm demonstrates a fine throughput and fairness performance in different network scenarios.

4.1 Introduction

The ever increasing traffic demands have rendered a single-network wireless system insufficient to meet the demands due to inefficient spectrum usage. A heterogeneous

network, where multiple radio access technologies (RATs) coexist, has emerged as a viable alternative solution. In a heterogeneous network, users are allowed to access the spectrum licensed to different spectrum owners, which are called service providers (SPs), and as a result a more efficient spectrum utilization can potentially be achieved. In heterogeneous networks, one significant issue to address is network selection where each user determines which network to associate with.

In this work, we consider the problem of network selection in a heterogeneous network featuring cognitive radio (CR). Specifically, we consider the *primary network access* scenario [15] where both primary users (PUs) and secondary users (SUs) are served by the primary networks. We model the network selection by SUs as a noncooperative game. With our proposed utility function, the game is shown to be an ordinal potential game (OPG) [12]. A stochastic learning algorithm (SLA) is proposed to perform network selection independently at each SU based on its action-reward history and not on other SUs' actions. The convergence property of the algorithm to a pure-strategy Nash equilibrium (NE) point is verified theoretically and numerically. To the best of our knowledge, this work presents the first application of SLA to OPGs in wireless networks. Notably, unlike the consideration of exact potential game (EPG) in [10], our formulation of OPG poses fewer constraints on the design of utility functions and thus facilitates mapping practical resource management problems in distributed networks into proper game-theoretic formulations.

4.1.1 Game-theoretic Problem Mapping

The mapping of network selection problem to game-theoretic formulation is summarized in Table 4.1.

Table 4.1: Mapping to game-theoretic formulation.

Elements in game	Characters in network selection problem
Players	Secondary users
Strategies	Selection of service providers
Reward	Individual user throughput
External state	Number of available channels

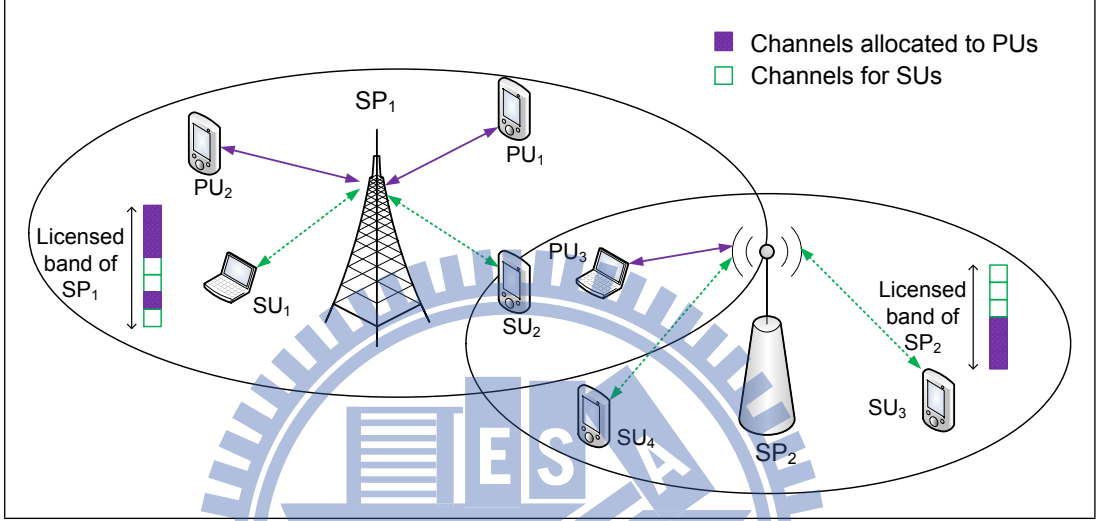


Figure 4.1: An exemplary heterogeneous network with 2 SPs, 3 PUs, and 4 SUs. The filled and blank blocks in the licensed band of each SP denote the busy channels currently used by the PUs and the residual channels available for serving the SUs, respectively.

4.2 System Model

We consider a cognitive heterogeneous network with M SPs and N SUs. The sets of SPs and SUs are denoted by \mathcal{M} and \mathcal{N} , respectively. SP_m owns K_m channels. At time instant j , after resource allocation for PUs, SP_m has $C_m(j)$ residual channels that can be used to serve the SUs. Fig. 1 presents an exemplary heterogeneous network where two RATs coexist.

To reflect a practical wireless heterogeneous network, our system model incorporates the following considerations:

1. Due to hardware and protocol limitations, each SU can subscribe to only one SP at a given time.

2. Each SU selects the SP independently. There is neither central control nor negotiation among SUs.
3. The statistics of the number of residual channels owned by each SP are fixed but unknown to the SUs.
4. The number of SUs in the system, N , is unknown.

Notably, the only information available for decision making is the action-reward history of individual players (SUs).

Let $\mathcal{N}_m(j) = \{i \in \mathcal{N} | a_i(j) = m\}$ be the set of SUs associated with SP_m at time j , where $a_i(j)$ is the action (i.e., network selection) of SU_i at time j . Here, we consider the case where the SUs are of the same priority class, and thus the residual channels are equally divided (can be in both frequency and time domain) to them. Then, if $a_i(j) = m$, the throughput of SU_i at time j is given by

$$r_i(j) = \frac{C_m(j)}{n_m(j)} R_{m,i}, \quad \forall i \in \mathcal{N}_m \quad (4.1)$$

where $n_m(j) \triangleq |\mathcal{N}_m(j)|$ and $R_{m,i}$ is the per-channel throughput of SU_i when SU_i is the only user associated with SP_m . The value of $R_{m,i}$ is determined by the modulation order (e.g., $R_{m,i} = 4$ when 16-QAM is adopted). For notational brevity, we hereafter discard the timing dependence in occasions without ambiguity.

4.3 Self-Organized Network Selection

In this section, we present the game-theoretic formulation of the self-organized network selection problem. The notations used in the formulation are summarized in Table 4.2.

Table 4.2: Summary of Symbols for Game-theoretic Formulation

Symbol	Meaning
\mathcal{N}	the set of SUs
\mathcal{M}	the set of SPs
\mathcal{C}	external state space (channel availability)
$C_m(j)$	number of available channels of SP _{<i>m</i>} at time <i>j</i>
$\mathcal{A}_i \subseteq \mathcal{M}$	the set of actions of player <i>i</i>
$s_i \in \mathcal{A}_i$	an element of \mathcal{A}_i
$a_i(j) \in \mathcal{A}_i$	the action (SP selection) of player <i>i</i> at time <i>j</i>
$a_{-i}(j) \in \mathcal{A}_i$	actions of players except for <i>i</i> at time <i>j</i>
$\mathcal{P}_i := \Delta(\mathcal{A}_i)$	the set of probability distribution over \mathcal{A}_i
$\mathbf{p}_i(j) \in \mathcal{P}_i$	mixed strategy of player <i>i</i> at time <i>j</i>
$r_i(j) \in \mathbb{R}$	instantaneous reward of player <i>i</i> at time <i>j</i>

4.3.1 Game Model

We model the network selection problem as a noncooperative game where the SUs are the players, and the number of residual channels (after the resource allocation of PUs) is considered as the external state. The game is represented as:

$$\mathcal{G} = \left(\mathcal{C}, \mathcal{N}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \{u_i\}_{i \in \mathcal{N}} \right)$$

where \mathcal{C} is the space of external states, \mathcal{N} is the set of players, $\{\mathcal{A}_i\}_{i \in \mathcal{N}}$ is the set of actions (network selection) that player *i* can take, and $\{u_i\}_{i \in \mathcal{N}}$ is the utility function of player *i* that depends on the actions of itself as well as other players.

The SUs are selfish and rational players with the objective of maximizing their individual throughput. Thus, we define the instantaneous reward of player *i* at time *j* as the throughput specified in (4.1). The reward function captures the dynamics of the behavior of PUs as well as the joint behaviors of multiple SUs. Then, we define the utility function

as the expected reward of player i over the channel availability¹, i.e.,

$$u_i(a_i, a_{-i}) \triangleq \mathbb{E}_{C_{a_i}} [r_i | (a_i, a_{-i})] = \frac{R_{a_i, i}}{n_{a_i}} \sum_{k=1}^{K_{a_i}} x_{a_i, k} \cdot k \quad (4.2)$$

where $x_{a_i, k}$ is the probability of $C_{a_i} = k$ with $\sum_{k=1}^{K_{a_i}} x_{a_i, k} = 1$, and n_{a_i} is the number of players taking action a_i , which depends on the action of player i (a_i) as well as other players' actions (a_{-i}). Formally, the game can be expressed as

$$(\mathcal{G}) : \max_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i}), \quad \forall i \in \mathcal{N}. \quad (4.3)$$

4.3.2 Analysis of Nash Equilibrium

With the utility function in (4.2), we show the existence of an NE point for the considered game here.

Proposition 4.3.1. *The game \mathcal{G} is an OPG.*

Proof: Consider the function $\Phi : \times_{i \in \mathcal{N}} \mathcal{A}_i \rightarrow \mathbb{R}_+$:

$$\Phi(a_1, \dots, a_N) = \prod_{m=1}^M \prod_{l=1}^{n_m} \nu_m(l) \cdot \prod_{i=1}^N R_{a_i, i} \quad (4.4)$$

where

$$\nu_m(l) \triangleq \frac{1}{l} \sum_{k=1}^{K_m} x_{m, k} \cdot k \quad (4.5)$$

is the average number of channels allocated by SP_m to each of its SUs when there are l SUs associated with SP_m . Now, consider that player i changes its action unilaterally from a_i to \check{a}_i . Let n_{a_i} and $n_{\check{a}_i}$ be the number of SUs associated with SP_{a_i} and $\text{SP}_{\check{a}_i}$ before the change, respectively. If this change improves the u_i , from the definitions in (4.2) and

¹The same formulation can be applied under fading channels, where the time-varying $R_{m, i}$ is considered as part of the external state and its average value is adopted in u_i . A longer learning period may be required in this case.

(4.5), we have

$$u_i(\check{a}_i, a_{-i}) > u_i(a_i, a_{-i}) \Leftrightarrow \nu_{\check{a}_i}(n_{\check{a}_i} + 1) \cdot R_{\check{a}_i, i} > \nu_{a_i}(n_{a_i}) \cdot R_{a_i, i}. \quad (4.6)$$

Meanwhile, since player i 's change merely affects the resource allocations in SP_{a_i} and $\text{SP}_{\check{a}_i}$, the change in Φ caused by player i 's unilateral deviation is given by

$$\frac{\Phi(\check{a}_i, a_{-i})}{\Phi(a_i, a_{-i})} = \frac{\nu_{\check{a}_i}(n_{\check{a}_i} + 1) \cdot R_{\check{a}_i, i}}{\nu_{a_i}(n_{a_i}) \cdot R_{a_i, i}} > 1. \quad (4.7)$$

From (4.6) and (4.7) we find that the variations in u_i and Φ due to player i 's unilateral deviation have the same sign, i.e.,

$$u_i(\check{a}_i, a_{-i}) - u_i(a_i, a_{-i}) > 0 \Leftrightarrow \Phi(\check{a}_i, a_{-i}) - \Phi(a_i, a_{-i}) > 0. \quad (4.8)$$

Therefore, \mathcal{G} is an OPG with potential function Φ [12]. ■

The existence of a pure-strategy NE is always guaranteed and it coincides with a local maximum of the potential function [12]. Note that an EPG formulation [8] requires

$$u_i(\check{a}_i, a_{-i}) - u_i(a_i, a_{-i}) = \Phi(\check{a}_i, a_{-i}) - \Phi(a_i, a_{-i}). \quad (4.9)$$

Comparing (4.8) and (4.9), it is observed that the constraint on the utility function is relaxed in OPG, which facilitates game-theoretic developments.

4.3.3 Stochastic Learning Procedure

Here, we discuss obtaining the NE via stochastic learning. As the channel availability is time-varying and the action is selected by each player simultaneously and independently in each play, previously developed algorithms requiring complete information (e.g., better response dynamics [12]) may not be applicable. To this end, we propose a decentralized algorithm based on stochastic learning (SL) [11], by which the SUs learn toward the equilibrium strategy profile from their individual action-reward history.

To facilitate the development of the SL-based algorithm, let the mixed strategy $\mathbf{p}_i(j) = [p_{i,1}(j), \dots, p_{i,M}(j)]^T$ be the network selection probability vector for player i , where $p_{i,s_i}(j)$ is the probability that player i selects strategy $s_i \in \mathcal{A}_i$ at time j . The proposed self-organized network selection (SoNS) algorithm is described in Algorithm 4.1.

Algorithm 4.1 Self-organized Network Selection (SoNS)

- 1: Initially, set $j = 0$, and the network selection probability vector as $p_{i,s_i}(j) = 1/|\mathcal{A}_i|, \forall i \in \mathcal{N}, s_i \in \mathcal{A}_i$.
- 2: At every time j , each player selects an action $a_i(j)$ as the outcome of a probabilistic experiment based on $\mathbf{p}_i(j)$.
- 3: The SUs receive the instantaneous reward $r_i(j)$ specified by (4.1) from the SPs.
- 4: Each SU updates its network selection probability vectors according to the following rules:

$$p_{i,s_i}(j+1) = p_{i,s_i}(j) + b \cdot \tilde{r}_i(j) (\mathbb{1}_{\{s_i=a_i(j)\}} - p_{i,s_i}(j)) \quad (4.10)$$

where $0 < b < 1$ is the learning rate, $\mathbb{1}_{\{\cdot\}}$ is the indicator function, and $\tilde{r}_i(j)$ is the normalized reward.

The instantaneous reward (throughput) serves as a reinforcement signal so that a high reward brings a high probability in the next strategy update (Step 4). Also note that network selection based on a probabilistic experiment (Step 2) might result in handover between different networks in the beginning of the learning procedure. However, a stable long-term network selection strategy will be yielded after the learning period (Proposition 2) and the time required for convergence is a small fraction of the total operation time.

Proposition 4.3.2. *The SoNS Algorithm converges to NE when the learning rate b is sufficiently small.*

While the convergence to an NE is guaranteed as $b \rightarrow 0$, a smaller value of b leads to a slower convergence rate. A proper value of b can be numerically determined to strike the desired tradeoff between the accuracy and rate of convergence for practical operations of the algorithm.

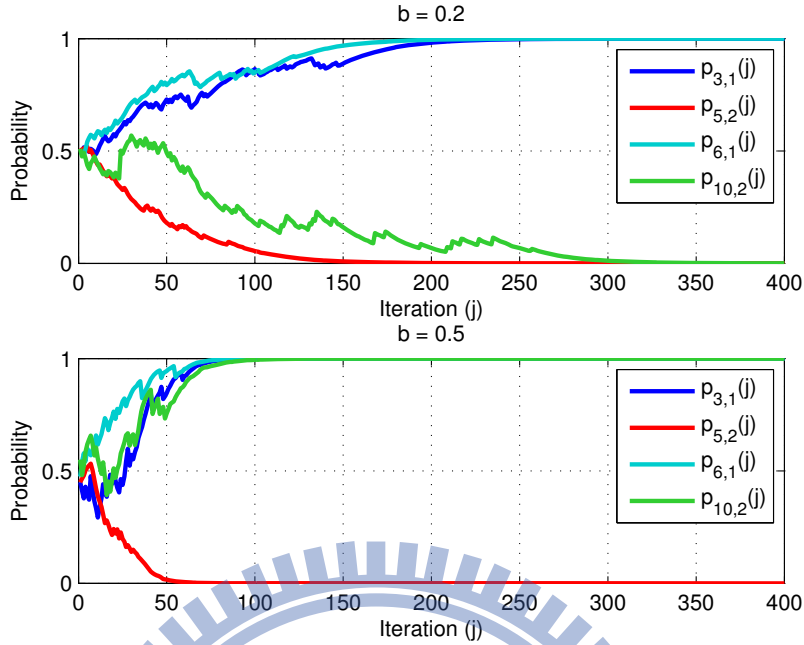


Figure 4.2: Evolution of the mixed strategies (choice probability of actions) of some players, using different learning rates.

4.4 Numerical Results

In order to evaluate the performance of the proposed scheme, we conduct a series of simulations. We consider a heterogeneous network in which there are 2 SPs each owning 3 channels. There are 10 SUs in the network, and the per-channel throughput is set to $R_{m,i} = \{2, 4, 6\}$ to reflect the modulation orders adopted under different RSS conditions. Fig. 4.2 shows the evolution of the choice probabilities of the actions (i.e., mixed strategy) for network selection using the proposed stochastic learning algorithm. With equal initial probabilities, it is observed that the network selection probabilities converge to pure strategies in around 300 and 100 cycles for $b = 0.2$ and $b = 0.5$, respectively. Note that SU #10 takes different strategies in the two cases. In Fig. 3, we test the deviation of the network selection of each of the 10 players. It is shown in Fig. 4.3(a) that when $b = 0.2$, unilateral deviation results in lower throughputs for all players, suggesting an NE point is reached by the learning algorithm. On the other hand, when $b = 0.5$, as shown in Fig. 4.3(b), SU #10 achieves a higher throughput by unilateral deviation, and thus the resulting strategy is not an NE point.

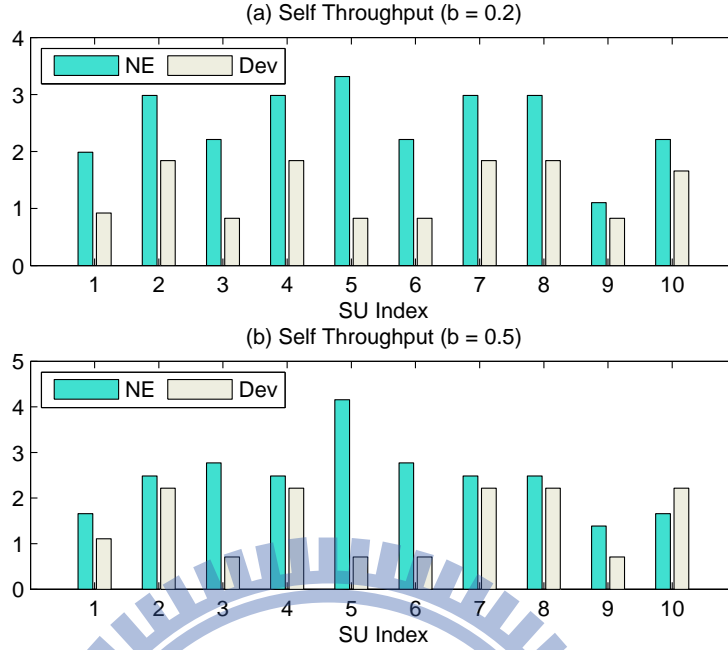


Figure 4.3: Test of unilateral deviation from the resulting strategy profile of each of the 10 players, using different learning rates.

In Table 4.3, we compare the performance of the proposed network selection scheme with two other approaches, namely, best RSS and (centralized) exhaustive search, which are described as follows:

- In the best RSS scheme, each SU chooses the SP with the best per-channel throughput (i.e., $a_i = \arg \max_m R_{m,i}$). If there are more than one best SP, choose arbitrarily.
- In the exhaustive search, the channel availability statistics and the number of SUs are known to a centralized controller, and the action profile is selected so as to maximize the system throughput $u_{sum} = \sum_{i=1}^N u_i$.

The performance of different network selection schemes are evaluated by the system throughput u_{sum} and the fairness among SUs, measured by the Jain's fairness index (JFI), $J = u_{sum}^2 / (N \sum_{i=1}^N u_i^2)$. We consider two scenarios for the simulation. In scenario 1, the SUs are randomly distributed. An SU may have better RSS from SP₁ ($R_{1,i} > R_{2,i}$), from SP₂ ($R_{1,i} < R_{2,i}$), or similar RSS from both SPs ($R_{1,i} = R_{2,i}$). In scenario 2, we set $R_{1,i} = 6$ and $R_{2,i} = \{2, 4\}, \forall i \in \mathcal{N}$. This describes a two-tier network where SP₁ is a

Table 4.3: Comparison of the achievable expected system throughput of three network selection schemes

	Proposed	Best RSS	Exhaustive
Scenario 1, u_{sum}	24.9662	24.4521	27.0621
Scenario 1, JFI	0.8974	0.7759	0.3822
Scenario 2, u_{sum}	25.9379	14.8554	25.9379
Scenario 2, JFI	0.9986	1.0000	0.8894

small-cell serving indoor SUs, while SP_2 is a macro-cell located far apart. We observe that the efficiency of the learned NE strategy (ratio between u_{sum} of the proposed and exhaustive search methods) is above 90% for both scenarios. In addition, the exhaustive search method results in best u_{sum} , but suffers from poor fairness in scenario 1. This is due to the *winners-first* property of exhaustive search: If m can be found so that $R_{m,i} = 6$, SU_i is usually assigned to SP_m ; on the other hand, those SUs with lower $R_{m,i}$ in both networks may be assigned to a less crowded SP instead of their own preference. The best RSS scheme has good system throughput in scenario 1 but not in scenario 2, since in this extreme case, all SUs are crowded in SP_1 and the resources of SP_2 are wasted. In contrast, the proposed method performs well in terms of both throughput and fairness under both scenarios. The results show the advantage of the proposed method: through the learning procedure towards equilibrium, the throughput of each SU is considered and the fairness can be maintained.

4.5 Concluding Remarks

In this chapter, we have studied the problem of self-organized network selection in heterogeneous networks with time-varying channel availability and unknown number of secondary users. We formulated the network selection problem by an ordinal potential game. A decentralized stochastic learning-based algorithm has been proposed. Simulation results have demonstrated the convergence of the algorithm towards a pure strategy Nash equilibrium point. The proposed method outperforms the best RSS scheme in terms of average throughput, while the performance loss compared to the centralized exhaust-

ive search is limited. Moreover, the proposed method achieves good fairness in various network scenarios.



Chapter 5

Spectrum Trading in Multiple-Seller Cognitive Radio Networks

This chapter studies spectrum trading in cognitive radio networks in which multiple service providers (SPs) sell licensed spectrum opportunities to multiple unlicensed secondary users (SUs). Spectrum trading is modeled as a multi-leader multi-follower Stackelberg game with two levels of competition. The SPs as leaders compete in offering spectrum prices first (upper-level subgame) and then the SUs as followers compete in selecting SPs to associate with (lower-level subgame). In the upper-level competition, SPs adjust their pricing strategies to maximize their individual revenues. In the lower-level competition, SUs select SPs based on the offered spectrum prices as well as the number of residual channels and the behavior of other SUs associated with each SP. The lower-level game incorporates the time-varying channel availability as the external state so that the proposed scheme is robust against dynamic channel availability. To achieve self-organized operation, we propose decentralized, stochastic learning-based algorithms for the Stackelberg game. The convergence properties of the proposed algorithms toward the Nash equilibrium (NE) are theoretically and numerically verified. The proposed scheme demonstrates good utility and fairness performances for the SUs as compared to other service selection schemes.

5.1 Introduction

COGNITIVE radio network (CRN) [37] has been considered as a promising solution to the problem of spectrum scarcity. In CRNs, owners of the licensed spectrum are referred to as service providers (SPs). Since the licensed spectrum is not always fully utilized, spectrum holes exist and cause inefficiency. To improve spectrum utilization, the SP may allow secondary users (SUs) to access its licensed spectrum. The SUs pay a fee to compensate for their access to the spectrum owned by an SP, and the payments become the revenue of the SP. When there are multiple SPs, their subscription prices affect the choices of SUs as well as the revenues of SPs. *Spectrum trading* in this multiple-seller spectrum market is the focus of this work.

We study the spectrum trading problem from a game-theoretic perspective. Game theory [38] models the interaction of distributed players and has been an effective tool for studying resource management problems in distributed networks such as CRNs [39] and heterogeneous networks [40]. A game-theoretic approach to spectrum trading was proposed in [32]. The methods in [32, 39, 40] are effective when the resource allocation for PUs is static, but may not be applicable in scenarios where PUs have time-varying behaviors, because of the assumption that either 1) the quality of service (QoS) requirements of PUs are not flexible and the number of residual channels is fixed; or 2) the QoS requirements of PUs are flexible but the preferences of PUs are fixed. When the PUs' traffic demands change, the channel availability may change accordingly and therefore the spectrum trading procedure may need to be executed again, resulting in significant overhead in practical CRN operations.

In this work, spectrum trading is formulated as a two-level Stackelberg game [38]. The SPs as leaders set their spectrum prices first (upper level) and then the SUs as followers select the service based on the offered prices (lower level). Our goal is to find proper service selection (for SUs) and spectrum pricing (for SPs) strategies, with the following considerations. First, players at the same level are unaware of the presence of one another and there is no information exchange among them. This avoids impractical information

exchange requirements. Second, the chosen strategies are robust to the time-varying channel availability, as the demands are unknown in the decision-making stage. Due to the lack of information of the opponents and the randomness of the environment, the subgame perfect Nash equilibrium (SPNE) of the Stackelberg game cannot be achieved through the traditional backward induction method [38]. Therefore, fully distributed methods for strategy selection in both levels are needed.

To achieve the equilibrium in a CRN with unknown and dynamic spectrum opportunities, we propose fully distributed algorithms based on stochastic learning (SL) [7]. With the proposed algorithms, the SPs and SUs learn from their individual action-reward history and adjust their strategies towards the equilibrium. The main contributions of this work are as follows:

- We formulate spectrum trading in a CRN as a two-level Stackelberg game where the upper- and lower-level subgames model the price competition of the SPs and service selection of the SUs, respectively. A unique feature of our considered game is that we formulate an expected game by incorporating time-varying channel availability as the external state. Selection strategies robust against time-varying channel availability can therefore be developed for the proposed game.
- We propose fully distributed SL-based algorithms for the robust Stackelberg game. The algorithms enable self-organized decision making for SPs and SUs and yield strategies that are robust against unknown dynamic channel availability. In our proposed algorithms the only information required for decision making is the action-reward history of individual SUs (SPs) in the lower (upper) level game. The convergence properties of the proposed algorithms toward the equilibrium are theoretically and numerically verified.

The rest of the chapter is organized as follows. In Section 5.2, the system model considered in this work is presented. The game-theoretic formulation of the service selection problem and the SL-based solutions are presented in Section 5.3. The price competition

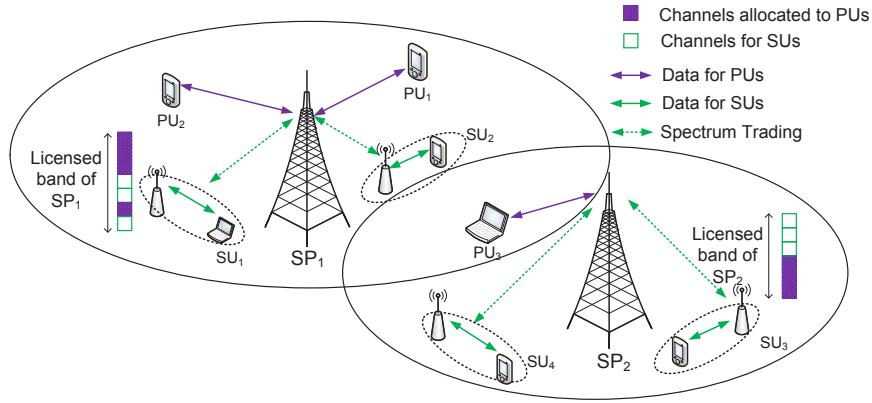


Figure 5.1: An exemplary cognitive radio network with 2 SPs, 3 PUs, and 4 SUs. The filled and blank blocks in the licensed band of each SP denote the busy channels currently used by the PUs and the residual channels available for serving the SUs, respectively.

among SPs is presented in Section 5.4. Numerical results are provided in Section 5.5. Conclusion is drawn in Section 5.6.

5.2 System Model

We consider a cognitive radio network with M SPs and N SUs, where SP_m owns K_m channels in total. The sets of SPs and SUs are denoted by \mathcal{M} and \mathcal{N} , respectively. Fig. 5.1 presents an exemplary cognitive radio network with two coexisting SPs.

5.2.1 Spectrum Trading Mechanism

In cognitive radio networks, after resource allocation to PUs, an SP may possess remaining spectrum (i.e., residual channels) which can be sold to the SUs. We adopt a shared access model in which the SPs set and announce the subscription prices. An SU sends a request message and pays the price if it wants to buy the spectrum opportunities from an SP. SUs have the freedom to dynamically select the SP that will provide the best reward determined by multiple factors (e.g., bandwidth, delay, and price). The procedure is referred to as *service selection*. On the other hand, the SPs may adjust their

pricing strategies iteratively in order to improve their own revenues. The information exchanges including price announcement and spectrum request go through a dedicated control channel. This way, a multiple-seller, multiple-buyer spectrum trading market is formed, where the sellers and buyers correspond to the SPs and SUs, respectively.

When complete information of other SPs is available, the SPs may determine their pricing strategies by anticipating the service selections of SUs (i.e., through backward induction [31, 34]). In our scenario, however, the SPs are independent decision makers who can simply learn the pricing strategy through iterative updates. The reward is collected when the behaviors of SUs converge, as in [32]. Therefore, the strategy update interval of SPs is longer than that of the SUs, so as to wait for the convergence of the choices of the SUs. The number of iterations that the SPs wait for the SUs' strategies to converge is denoted as T_{conv} .

5.2.2 Two-level Competition as a Stackelberg Game

The spectrum trading mechanism described above shows a leader-follower structure in that the SPs (leaders) set the prices first, and then the SUs (followers) perform service selection according to the prices. Such a hierarchical feature suggests the formulation of a Stackelberg game with two levels of competitions. The upper-level competition exists among SPs to sell the residual channels (i.e., spectrum opportunities) to the SUs. For an SP, a higher subscription price means more revenue that will be received from one SU, but also means a possible reduction in the number of subscriptions since other SPs may offer more spectrum opportunities or better prices. An SP must therefore carefully set the price so that its total revenue is maximized. The lower-level competition exists in the service selection of SUs. To buy spectrum opportunities, an SU chooses an SP with larger payoff and lower perceived cost, and its strategy depends on the strategies of other SUs. The elements in lower- and upper-level subgames of a Stackelberg game are summarized in Table 5.1, and the detailed game formulations will be discussed in Section 5.3 and Section 5.4, respectively.

Table 5.1: Elements in a Stackelberg game

Element	Lower-level sub-game	Upper-level sub-game
Player set	The set of SUs, \mathcal{N}	The set of SPs, \mathcal{M}
Strategy set	The set of SPs, \mathcal{M}	The set of candidate price levels, \mathcal{A}_m
Utility	The expected reward (payoff minus price) of an SU	The revenue obtained by selling spectrum opportunities
External state	The number of available channels of SP $_m$, $c_m(j)$	-

To reflect a practical distributed CRN, our system model incorporates the following considerations:

1. Due to hardware and protocol limitations, each SU can buy and access the spectrum opportunity of only one SP at a given time.
2. Service selection is done by each SU independently and simultaneously. There is neither negotiation nor sequential updates among SUs.
3. The statistics of the spectrum opportunities offered by the SPs are fixed but unknown to the SUs.
4. The number of SUs in the system is unknown to any SP and SU; the number of SPs is unknown to any SP.

5.3 Service Selection of Secondary Users

In this section, we present the game-theoretic formulation and self-organized learning procedure for service selection of SUs. As the lower-level subgame of our Stackelberg game formulation, the SUs perform service selection under given spectrum prices announced by the SPs. Our objective is to devise for the SUs a fully-distributed strategy that takes into

Table 5.2: Summary of Notations for Game-theoretic Formulation

Symbol	Meaning
\mathcal{N}	the set of SUs
\mathcal{M}	the set of SPs
\mathcal{C}	the space of external states (channel availability)
$c_m(j)$	number of available channels of SP _{<i>m</i>} at time <i>j</i>
\mathcal{A}_i	the set of actions of player <i>i</i>
$s_i \in \mathcal{A}_i$	the pure strategy of player <i>i</i>
$a_i(j) \in \mathcal{A}_i$	the action (SP selection) of player <i>i</i> at time <i>j</i>
$a_{-i}(j) \in \mathcal{A}_i$	actions of players except for <i>i</i> at time <i>j</i>
$\mathcal{P}_i := \Delta(\mathcal{A}_i)$	the set of probability distribution over \mathcal{A}_i
$\mathbf{p}_i(j) \in \mathcal{P}_i$	mixed strategy of player <i>i</i> at time <i>j</i>
$r_i(j) \in \mathbb{R}$	instantaneous reward of player <i>i</i> at time <i>j</i>

account the effect of congestion, offered spectrum opportunities, and the subscription price. The notations used in the formulation are summarized in Table 5.2.

5.3.1 Game Model

We model the service selection problem as a non-cooperative game where the SUs are the players, and the number of residual channels (after the resource allocation of PUs) is considered as the external state. The game is represented as:

$$\mathcal{G}_1 = \left(\mathcal{C}, \mathcal{N}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \{u_i\}_{i \in \mathcal{N}} \right)$$

where \mathcal{C} is the space of external states, \mathcal{N} is the set of players, $\{\mathcal{A}_i\}_{i \in \mathcal{N}}$ is the set of actions (service selection) that player *i* can take, and $\{u_i\}_{i \in \mathcal{N}}$ is the utility of player *i*. In the service selection game, the utility is defined as the expected reward over the external state.

As a player in the lower-level subgame, an SU receives its instantaneous reward in each play. The reward of an SU is defined as the payoff offered by the SP minus the price paid by the SU. The payoff function is designed to quantify satisfaction levels of SUs, and its value depends on the number of residual channels as well as the number of SUs sharing the same SP. In this work, we assume the SUs are of the same priority class, and thus

the residual channels are equally divided (can be in both frequency and time domains) between them. Let $c_m(j) \in \{0, 1, \dots, K_m\}$ be the number of residual channels of SP_m at time j , and q_m be the spectrum subscription price paid by each SU that is attached to SP_m (i.e., purchasing spectrum opportunities from SP_m). The price q_m is assumed to take possible values on a pre-defined and finite pricing strategy set of SP_m . Let B_m be the channel bandwidth of SP_m , $\mathcal{N}_m(j)$ be the set of SUs attached to SP_m at time j , and $n_m(j) \triangleq |\mathcal{N}_m(j)|$. The bandwidth allocated to an SU attached to SP_m at time j is given by $B_m c_m(j)/n_m(j)$, and the instantaneous reward received by SU_i can be given as

$$r_i(j) = \kappa B_m c_m(j)/n_m(j) - q_m, \quad \forall i \in \mathcal{N}_m \quad (5.1)$$

where the constant κ is interpreted as the monetary value of unit bandwidth seen by an SU. Without loss of generality, we set $\kappa = 1$ in this work. The reward function in (5.1) captures the dynamics of the joint behaviors of multiple SPs as well as SUs. For notational brevity, we hereafter discard the timing dependence (j) in occasions without ambiguity. With the reward function in (5.1), the utility (i.e., expected reward) becomes

$$u_i(a_i, a_{-i}) \triangleq \mathbb{E}_{c_{a_i}} [r_i(a_i, a_{-i})] = B_{a_i} \bar{c}_{a_i} / n_{a_i} - q_{a_i} \quad (5.2)$$

where $\bar{c}_{a_i} \triangleq \mathbb{E}[c_{a_i}]$ is the expected number of residual channels of SP_{a_i} . The utility of player i depends on the action of player i (a_i) and of other players (a_{-i}).

5.3.2 Analysis of Nash Equilibrium

We assume that the SUs are selfish and rational players with the objective of maximizing their individual utility. Formally,

$$(\mathcal{G}_1) : \max_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i}), \quad \forall i \in \mathcal{N}. \quad (5.3)$$

The Nash equilibrium of \mathcal{G}_1 is studied as follows.

Definition 5.3.1 (Nash equilibrium). An action profile $\mathbf{a}^* = (a_1^*, \dots, a_N^*)$ is a pure

strategy Nash equilibrium (NE) point of the non-cooperative game \mathcal{G} if and only if no player can improve its utility by deviating unilaterally, i.e.,

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*), \quad \forall i \in \mathcal{N}, \forall a_i \in \mathcal{A}_i \setminus \{a_i^*\}. \quad (5.4)$$

With the utility function in (5.2), we show the existence of a pure strategy NE point for the lower-level subgame.

Proposition 5.3.1. *The game \mathcal{G}_1 is an exact potential game (EPG).*

Proof: Define the function $\Phi : \times_{i \in \mathcal{N}} \mathcal{A}_i \rightarrow \mathbb{R}_+$ as

$$\Phi(\mathbf{a}) = \sum_{m=1}^M \left(\sum_{l=1}^{n_m} \nu_m(l) - n_m q_m \right) \quad (5.5)$$

where $\nu_m(l) = B_m \bar{c}_m / l$. Now, consider that player i changes its action unilaterally from a_i to \check{a}_i . Let n_{a_i} and $n_{\check{a}_i}$ be the load of (i.e., number of SUs attached to) SP_{a_i} and $\text{SP}_{\check{a}_i}$ before the change, respectively. Note that player i 's change merely affects the SUs subscribing to SP_{a_i} and $\text{SP}_{\check{a}_i}$, and the change in $\Phi(\cdot)$ caused by its unilateral deviation is given by

$$\begin{aligned} & \Phi(\check{a}_i, a_{-i}) - \Phi(a_i, a_{-i}) \\ &= \left[\sum_{l=1}^{n_{\check{a}_i}+1} \nu_{\check{a}_i}(l) - (n_{\check{a}_i}+1)q_{\check{a}_i} \right. \\ & \quad \left. + \sum_{l=1}^{n_{a_i}-1} \nu_{a_i}(l) - (n_{a_i}-1)q_{a_i} \right] \\ & \quad - \left[\sum_{l=1}^{n_{\check{a}_i}} \nu_{\check{a}_i}(l) - n_{\check{a}_i}q_{\check{a}_i} + \sum_{l=1}^{n_{a_i}} \nu_{a_i}(l) - n_{a_i}q_{a_i} \right] \\ &= [\nu_{\check{a}_i}(n_{\check{a}_i}+1) - q_{\check{a}_i}] - [\nu_{a_i}(n_{a_i}) - q_{a_i}] \\ &= u_i(\check{a}_i, a_{-i}) - u_i(a_i, a_{-i}). \end{aligned} \quad (5.6)$$

That is, the changes in $u_i(\cdot)$ and $\Phi(\cdot)$ due to player i 's unilateral deviation are identical. Therefore, \mathcal{G}_1 is an EPG with potential function $\Phi(\cdot)$ [12]. \blacksquare

For EPGs, the existence of a pure-strategy NE is always guaranteed and the NE points coincide with the local maximum of the potential function [12].

5.3.3 Stochastic Learning Procedure for Service Selection

Here, we propose a decentralized algorithm by which the SUs learn toward the NE strategy profile from their individual action-reward history. The algorithm is based on stochastic learning (SL) [11]. To facilitate the development of the SL-based algorithm, let the mixed strategy $\mathbf{p}_i(j) = [p_{i,1}(j), \dots, p_{i,M}(j)]^T$ be the service selection probability vector for player i , where $p_{i,s_i}(j)$ is the probability that player i selects strategy $s_i \in \mathcal{A}_i$ at time j . Let $\mathbf{P}(j) = [\mathbf{p}_1(j), \dots, \mathbf{p}_M(j)]$ be the mixed strategy profile of \mathcal{G}_1 . We denote the mixed extension of utility u_i by $\psi_i(\mathbf{P})$, i.e.,

$$\psi_i(\mathbf{P}) = \sum_{a_1, \dots, a_N} u_i(a_1, \dots, a_N) \prod_{i'=1}^N p_{i', a_{i'}}. \quad (5.7)$$

Letting \mathbf{P}_{-i} be the mixed strategy of players except for player i , we have the definition of NE in mixed strategy as follows.

Definition 5.3.2. A strategy profile \mathbf{P}^* is a mixed-strategy Nash equilibrium (NE) point of the non-cooperative game \mathcal{G} if and only if

$$\psi_i(\mathbf{p}_i^*, \mathbf{P}_{-i}^*) \geq \psi_i(\mathbf{p}_i, \mathbf{P}_{-i}^*), \quad \forall i \in \mathcal{N}, \forall \mathbf{p}_i \in \mathcal{P}_i \setminus \{\mathbf{p}_i^*\}. \quad (5.8)$$

A stochastic learning algorithm is characterized by its rule of updating the mixed strategies (based on the action-reward observation), which is usually referred to as the *learning rule*. Different learning rules may result in different convergence behaviors. For example, the learning rule proposed in [11] which is widely adopted in the literature has been later proved to converge to NE point when applied to potential games [41]. Since the service selection game is a potential game (Proposition 5.3.1), it benefits from

this nice convergence property if the same learning rule is adopted. The learning rule requires the instantaneous reward to be normalized. In our case, the lower bound of the reward is given by $r_{\text{inf}} = -q^{\text{max}}$, where $q^{\text{max}} = \max_m q_m$ is the highest subscription price. On the other hand, the maximum reward obtainable from an SP is achieved when all channels are allocated to one SU and the lowest price is charged. The upper bound of the reward is given by the maximum possible reward from all SPs; in other words, $r_{\text{sup}} = \max_m (\kappa B_m K_m - q_m)$. Then, the normalized reward $\tilde{r}_i(j) \in [0, 1]$ can be obtained as

$$\tilde{r}_i(j) = \frac{r_i(j) - r_{\text{inf}}}{r_{\text{sup}} - r_{\text{inf}}}. \quad (5.9)$$

The proposed self-organized service selection (SoSS) algorithm is described in Algorithm 5.1.

Algorithm 5.1 Self-organized Service Selection (SoSS)

- 1: Initially, set $j = 0$, and the spectrum request probability vector as $p_{i,s_i}(j) = 1/|\mathcal{A}_i|, \forall i \in \mathcal{N}, s_i \in \mathcal{A}_i$.
- 2: At every time instant j , each SU selects an action (i.e., SP) $a_i(j)$ according to $\mathbf{p}_i(j)$.
- 3: The SUs receive the instantaneous reward $r_i(j)$ specified by (5.1).
- 4: Each SU updates its service selection probability vectors according to the following rule:

$$p_{i,s_i}(j+1) = p_{i,s_i}(j) + b \cdot \tilde{r}_i(j) (\mathbb{1}_{\{s_i=a_i(j)\}} - p_{i,s_i}(j)) \quad (5.10)$$

where $0 < b < 1$ is the learning rate, and $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

Notably, the instantaneous reward serves as a reinforcement signal so that a high reward brings a high probability in the next play (Step 4). Moreover, the proposed learning algorithm is fully distributed: the selection of SP is solely based on the individual action-reward history without knowledge of other players' actions.

Proposition 5.3.2. *The SoSS Algorithm converges to a (possibly mixed-strategy) NE point when the learning rate b is sufficiently small.*

Proof: This follows from the fact that the SLA converges to a (possibly mixed-strategy) NE point when applied to an ordinal potential game (OPG) [41], and the fact that EPG belongs to OPG. ■

We emphasize that the proposed game and the convergence proof are not restricted to the model described above with the specific reward function in (5.1), but is applicable to reward functions that take into account other QoS parameters. The logarithm reward function in [32, 33] and the delay-related reward function in [34] are both possible choices.

5.3.4 Social Welfare and Price of Anarchy

While NE marks a steady state of mixed strategies of players, its efficiency needs further justification. The efficiency of equilibrium is typically determined by the price of anarchy (PoA) [42], which is defined as the ratio between the social welfare (SW) of the worst NE and that of the optimal strategy profile. Similar to [34], in our proposed spectrum trading markets, the social welfare is defined as the sum of the utilities of all SUs and SPs. Considering the utility function in (5.2) and after simple manipulations we have

$$SW = \sum_{m:n_m>0} B_m \bar{c}_m. \quad (5.11)$$

Notice that the price q_m does not appear in (5.11), since the price paid by the SUs becomes the utility (i.e., revenue) of the SPs. Considering the case with two SPs, the social welfare is maximized when the spectrum opportunities of both SPs are utilized, i.e.,

$$SW_{\max} = B_1 \bar{c}_1 + B_2 \bar{c}_2.$$

Then, the PoA is given by

$$\text{PoA} = \begin{cases} \frac{B_1 \bar{c}_1}{B_1 \bar{c}_1 + B_2 \bar{c}_2}, & \text{if } B_2 \bar{c}_2 - q_2 \leq \frac{B_1 \bar{c}_1}{N} - q_1; \\ \frac{B_2 \bar{c}_2}{B_1 \bar{c}_1 + B_2 \bar{c}_2}, & \text{if } B_1 \bar{c}_1 - q_1 \leq \frac{B_2 \bar{c}_2}{N} - q_2; \\ 1, & \text{otherwise.} \end{cases} \quad (5.12)$$

As can be seen from (5.12), the PoA is usually good unless the condition of one of the SPs is so poor (e.g., high subscription price or narrow spectrum opportunity) that all SUs

would rather crowd in another SP.

The social welfare is not the only metric for performance evaluation. Specifically, a globally optimal solution that maximizes the social welfare does not guarantee a fair resource allocation in terms of the individual utilities of SUs [43]. The fairness issue will be discussed in Sec. 5.5 through numerical experiments.

5.4 Price Competition among Service Providers

In the upper-level subgame of the proposed Stackelberg game formulation, the SPs compete with each other in determining the subscription price with the objective of maximizing the revenue. In this section, we present the game model and the fully distributed learning in the upper-level price competition game.

5.4.1 Game Model

The upper-level price competition game is modeled as a game played by the SPs. The game is represented as a 3-tuple:

$$\mathcal{G}_2 = \left(\mathcal{M}, \{\mathcal{A}_m\}_{m \in \mathcal{M}}, \{u_m\}_{m \in \mathcal{M}} \right)$$

where \mathcal{M} is the set of players (SPs), $\{\mathcal{A}_m\}_{m \in \mathcal{M}}$ is the set of actions (candidate price levels) that player $m, m \in \mathcal{M}$ can take, and $\{u_m\}_{m \in \mathcal{M}}$ is the utility defined as the revenue of SP_m . The revenue that an SP receives depends on its user load at the equilibrium. Given the pricing vector $\mathbf{q} = [q_1, \dots, q_m, \dots, q_M]$ and letting $n_m^*(\mathbf{q})$ be the number of SUs attached to SP_m under Nash equilibrium and q_{-m} be the price of SPs except for SP_m , the utility of SP_m is given by

$$u_m(q_m, q_{-m}) = q_m \cdot n_m^*(\mathbf{q}). \quad (5.13)$$

As in the lower-level game, we adopt learning algorithm for the SPs to adapt to proper

pricing strategies. In the upper-level competition, as can be seen from (5.13), the revenues received by the SPs depend on the user loads at the convergence point of the lower-level game. If the user loads (and therefore the revenues) vary under a fixed pricing strategy profile, it could be difficult for the SPs to evaluate each strategy and find proper ones.

Fortunately, while there may be multiple NE points, the user loads at NE are unique with a fixed subscription price vector. Following the discussions in [32,34], it is known that for a given subscription price vector, the unique steady-state user loads are characterized by the Wardrop equilibrium [44]. At the equilibrium, the per-SU utilities offered by subscribing to the SPs that have at least one attached SU are equal, and are larger than that experienced by a single SU attached to any unused SP. In other words, $\forall m, m' \in \mathcal{M}$ with $n_m^* > 0$, we have

$$\begin{cases} \nu_m(n_m^*) - q_m = \nu_{m'}(n_{m'}^*) - q_{m'}, & \text{if } n_{m'}^* > 0, \\ \nu_m(n_m^*) - q_m > \nu_{m'}(1) - q_{m'}, & \text{otherwise.} \end{cases} \quad (5.14)$$

Remark. The unique user load profile under Wardrop equilibrium is based on the assumption that the game is non-atomic; that is, the number of SUs is big compared to the number of SPs. When the number of SUs is finite, solving (5.14) may result in non-integer n_m^* 's which do not constitute feasible user loads, and there may be multiple NEs. However, the user load profile under NE is unique when additional constraint is applied.

Definition 5.4.1 (Strict Nash Equilibrium). An action profile $\mathbf{a}^* = (a_1^*, \dots, a_N^*)$ is a pure strategy strict Nash equilibrium point of the non-cooperative game \mathcal{G} if and only if unilateral deviation results in decreased utility for all players, i.e.,

$$u_i(a_i^*, a_{-i}^*) > u_i(a_i, a_{-i}^*), \quad \forall i \in \mathcal{N}, \forall a_i \in \mathcal{A}_i \setminus \{a_i^*\}. \quad (5.15)$$

Proposition 5.4.1. *The user load profile is unique if the lower-level NE is strict.*

Proof: The proof relies on an important property of potential games: an NE always

coincides with a local maximum of the potential function. For a network with M SPs, let $(\hat{n}_1, \dots, \hat{n}_M)$ be a user load profile that maximizes the potential function Φ in (5.5), and denote for a user load n_m the *drift* as $d_m = n_m - \hat{n}_m$. We show that any user load profile (n_1, \dots, n_M) in which there exists $m \in \mathcal{M}$ such that $|d_m| > 0$ cannot be a maximizer of the potential function, and therefore the user load profile is unique.

For better understanding, we consider first the case with $M = 2$ SPs, and a drift in user loads so that $(n_1, n_2) = (\hat{n}_1 + 1, \hat{n}_2 - 1)$. Denote by Φ_{n_1, n_2} the value of the potential function under user loads (n_1, n_2) . Define $\Delta_1 = \Phi_{\hat{n}_1+1, \hat{n}_2-1} - \Phi_{\hat{n}_1, \hat{n}_2}$, we have

$$\Delta_1 = \frac{B_1 \bar{c}_1}{\hat{n}_1 + 1} - q_1 - \frac{B_2 \bar{c}_2}{\hat{n}_2} + q_2 < 0. \quad (5.16)$$

Moving one step further in the same direction of the drift and define $\Delta_2 = \Phi_{\hat{n}_1+2, \hat{n}_2-2} - \Phi_{\hat{n}_1+1, \hat{n}_2-1}$, we have

$$\Delta_2 = \frac{B_1 \bar{c}_1}{\hat{n}_1 + 2} - q_1 - \frac{B_2 \bar{c}_2}{\hat{n}_2 - 1} + q_2 < \Delta_1 < 0. \quad (5.17)$$

By moving further, it can be shown that the potential function is monotonically decreasing and there is no other local maximum of Φ in the same direction. The same observation can be made for the movement in another direction.

When there are M SPs, consider a user load profile $(n_1, \dots, n_M) = (\hat{n}_1 + d_1, \dots, \hat{n}_M + d_M)$, where $d_m \neq 0, \forall m$. Define $\Delta_1 = \Phi_{\hat{n}_1+d_1, \dots, \hat{n}_M+d_M} - \Phi_{\hat{n}_1, \dots, \hat{n}_M}$, we have $\Delta_1 < 0$. For a further drifted user load profile $(n_1, \dots, n_M) = (\hat{n}_1 + d'_1, \dots, \hat{n}_M + d'_M)$ such that $d'_m \geq d_m$ if $d_m > 0$ and $d'_m \leq d_m$ otherwise, define $\Delta_2 = \Phi_{\hat{n}_1+d'_1, \dots, \hat{n}_M+d'_M} - \Phi_{\hat{n}_1+d_1, \dots, \hat{n}_M+d_M}$. Again we have $\Delta_2 < \Delta_1 < 0$ and show that Φ is monotonically decreasing. Therefore the user load profile $(\hat{n}_1, \dots, \hat{n}_M)$ is unique under NE and the proof is completed. ■

5.4.2 Stochastic Learning Procedure for Price Competition

In the upper-level game, learning-based algorithms help the SPs gradually adjust their pricing strategies based on the service selections of the SUs at the equilibrium of the lower-

Algorithm 5.2 Self-organized Pricing (SoP)

1: Initially, set $k = 0$. Set the pricing probability vector and utility estimation as

$$\begin{aligned} p_{m,s_m}(0) &= 1/|\mathcal{A}_m|, \\ \hat{u}_{m,s_m}(-1) &= 0, \quad \forall m \in \mathcal{M}, s_m \in \mathcal{A}_m. \end{aligned}$$

- 2: At the beginning of the k th iteration, each seller selects an action $a_m(k)$ according to the current pricing strategy $\mathbf{p}_m(k)$.
- 3: When the service selection of SUs converges, each seller m receives the utility $u_m(k)$ specified by (5.2) depending on the user load.
- 4: All SPs update their utility estimation and pricing probability vector in iteration k according to the rules:

$$\begin{aligned} &\hat{u}_{m,s_m}(k) - \hat{u}_{m,s_m}(k-1) \\ &= \eta \mathbb{1}_{\{a_m(k)=s_m\}} (u_m(k) - \hat{u}_{m,s_m}(k-1)) \\ p_{m,s_m}(k+1) &= \frac{p_{m,s_m}(k)(1+\epsilon)^{\hat{u}_{m,s_m}(k)}}{\sum_{s'_m \in \mathcal{A}_m} p_{m,s'_m}(k)(1+\epsilon)^{\hat{u}_{m,s'_m}(k)}} \end{aligned} \tag{5.18}$$

where η and ϵ are the learning rates for utility estimation and pricing probability, respectively.

level game. A seller's pricing strategy is defined over a probability space of its candidate price levels.

As in the lower-level game, two main issues are considered when designing the learning algorithm for the upper-level game, namely, the learning rule and the convergence property. First, since the total number of SUs is unknown to the SPs, it is difficult to obtain the upper bound and normalize the revenue. Therefore, a probability update rule different from (5.10) is needed. In this work, we consider the *multiplicative-weight* rule for mixed-strategy update. The learning procedure in the self-organized price (SoP) competition is described in Algorithm 5.2. The multiplicative-weight update rule in (5.18) belongs to the combined fully distributed payoff and strategy reinforcement learning (CODIPAS-RL) [7], in which learning applies to both the expected payoff and the strategies.

The second issue is the convergence behavior when the SL algorithm is applied in the price competition game. Unlike the lower-level game, the upper-level competition with

the utility in (5.14) is a potential game. Thus, we are unable to provide a theoretical proof to guarantee the convergence toward an NE for the price competition game. However, the algorithm still has some nice properties when applied to general strategic games. We investigate first the approximation of (continuous-time) ordinary differential equation (ODE) by the discrete-time mixed strategy update rule, and then the theoretical perspectives of convergence behaviors. The notations of $\mathbf{p}_m, \mathbf{P}_{-m}, \mathbf{P}$, and $\psi_m(\mathbf{P})$ are defined similarly as in Section 5.3, and \mathbf{e}_{s_m} is a unit probability vector (of appropriate dimension) with the s_m -th component being one and all other components being zero.

Proposition 5.4.2. *With sufficiently small learning rates η and ϵ :*

1. *The estimated utility converges to*

$$\hat{u}_{m,s_m} \rightarrow \psi_m(\mathbf{e}_{s_m}, \mathbf{P}_{-m}). \quad (5.19)$$

2. *Asymptotically, the probability matrix sequence $\{\mathbf{P}(k)\}$ can be approximated by the trajectory of the following ODE:*

$$\frac{dp_{m,s_m}(t)}{dt} = p_{m,s_m}(t) [\psi_m(\mathbf{e}_{s_m}, \mathbf{P}_{-m}) - \psi_m(\mathbf{P})] \quad (5.20)$$

where $p_{m,s_m}(t)$ is the continuous-time version of $p_{m,s_m}(k)$, and the boundary condition is given by $\mathbf{P}(0) = \mathbf{P}_0$, where \mathbf{P}_0 is the initial mixed strategy matrix.

Proof: See [7, Section 4.3]. ■

Notice that $\psi_m(\mathbf{e}_{s_m}, \mathbf{P}_{-m})$ is the utility of player m if it employs pure strategy s_m while other player $m', \forall m' \in \mathcal{M}, m' \neq m$ employs a mixed strategy $\mathbf{p}_{m'}$, and its value is learned by player m as the estimated utility \hat{u}_{m,s_m} , as shown in (5.19). On the other hand, the ODE for mixed-strategy in (5.20) is the *replicator equation* [14] in which the probability of taking one strategy increases if the current estimated utility of this strategy is larger than the average utility over all strategies and decreases otherwise. Compared to the best response dynamics [12] where a player changes its strategy in the next iteration

to the best action according to other players' actions (i.e., the best response), with the replicator dynamics, a player selects an action according to a probability distribution over the strategy set, and adjusts the weighting for each possible action in each iteration based on the utility estimation.

Proposition 5.4.3. *The proposed learning algorithm has the following properties:*

1. *All Nash equilibria are stationary points of (5.20);*
2. *All stationary points of (5.20) are Nash equilibria.*

Proof: Proposition 5.4.3 is an instance of the Folk theorems in the evolutionary game theory [14, Chapter 3], and these properties follow directly from the replicator equation in (5.20). Please also refer to [7, Section 4.3]. ■

For an intuitive explanation, observe that for a mixed-strategy NE profile \mathbf{P}^* , all survived pure strategies (i.e., s_m with $p_{m,s_m}^* > 0$) of player m perform equally well when other players follow the mixed strategy \mathbf{P}_{-m}^* . That is, the condition

$$\begin{aligned} \psi_m(\mathbf{e}_{s_m}, \mathbf{P}_{-m}^*) &= \psi_m(\mathbf{P}^*), \\ \forall m \in \mathcal{M}, s_m \in \mathcal{A}_m \text{ with } p_{m,s_m}^* > 0 \end{aligned} \quad (5.21)$$

must hold. Therefore, any NE must lead the right-hand-side of (5.20) to zero and thus constitutes a stationary point of (5.20). In other words, if the proposed algorithm converges to a stationary point of (5.20), the limiting point must be a (possibly mixed-strategy) NE point. Although there is no theoretical proof as in the lower-level game (since \mathcal{G}_2 is not an EPG), the convergence toward NE in the upper-level game is still observed through numerical simulations.

5.5 Numerical Results

In order to evaluate the performance of the proposed scheme and algorithms, we conduct a series of simulations. The distribution of the number of residual channels (i.e., spec-

Table 5.3: Simulation Parameters

Parameter	Value
Number of SPs	$M = 2$
Max. number of channels	$K_m = 3$
Ch. availability of SP ₁	$\mathbf{x}_1 = [0, 0.1, 0.3, 0.6]$
Ch. availability of SP ₂	$\mathbf{x}_2 = [0, 0.4, 0.3, 0.3]$
Pricing strategies	$\mathcal{A}_m = [1, 1.5, 2, 2.5], \forall m$
Learning rate	$(\eta, \epsilon) = (0.1, 0.05)$
Number of SUs	$N = 6$
Learning rate of SUs	$b = 0.3$
Waiting time for obtaining NE	$T_{conv} = 400$

trum opportunities) offered by SP_{*m*} is described by a vector $\mathbf{x}_m = [x_{m,0}, \dots, x_{m,c}, \dots, x_{m,K_m}]$, where $x_{m,c}$ denotes the probability that SP_{*m*} possesses *c* residual channels. The default values of simulation parameters are given in Table 5.3, and these values are adopted in the simulations unless otherwise specified.

We first study the lower-level game under a given price vector $(q_1, q_2) = (1, 1.5)$. The purpose is to observe the convergence behavior and the performance of the proposed algorithm. Then, the upper-level game is involved to observe the price competition.

5.5.1 Convergence Behavior of the Lower-level Game

Fig. 5.2 shows the evolutions of the choice probabilities of the actions (i.e., mixed strategies) for service selection using the proposed SL algorithm. With equal initial probabilities, it is observed that the service selections converge to pure strategies in 350 and 250 iterations for $b = 0.3$ and $b = 0.5$, respectively. We observe that the final user loads of SUs are different: $(n_1, n_2) = (3, 3)$ for $b = 0.3$ and $(n_1, n_2) = (4, 2)$ for $b = 0.5$. As the convergence toward pure strategies is observed in both cases, an intuitive question to ask is whether the learned strategy constitutes an NE point.

To verify the NE property, we test the unilateral deviation from the learned service selection strategies of each of the $N = 6$ players. The comparison in terms of (normalized) utilities is given in Fig. 5.3.

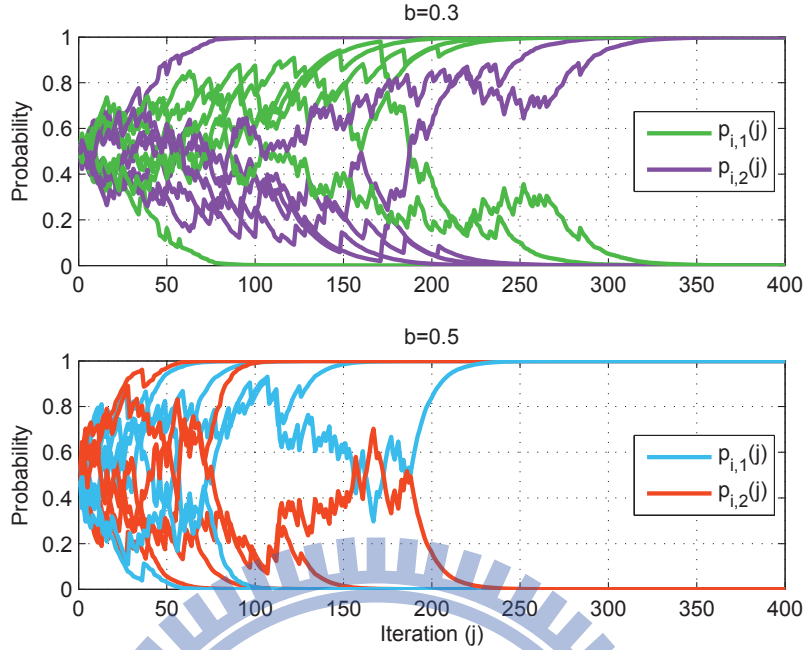


Figure 5.2: Evolution of the mixed strategies (probability of taking different actions) of all players. Each pair of $p_{i,1}(j)$ and $p_{i,2}(j)$ shows the behavior of a player $i \in \mathcal{N}$.

It is shown that when $b = 0.3$, unilateral deviation from the learned strategy results in a lower utility for all players. This confirms that the outcome of the learning algorithm is an NE point. However, when $b = 0.5$, four SUs (#1, #2, #5, and #6) gain higher utility by unilateral deviation, which implies that the resulting strategy is not an NE point. Combined with the results in Fig. 5.2, we observe that the user load under NE is $(n_1, n_2) = (3, 3)$. Since the learning algorithm converges to $(n_1, n_2) = (4, 2)$ when $b = 0.5$, any of the SUs subscribing to SP_1 can improve its utility by unilaterally deviating to SP_2 .

During the learning procedure, the decisions of service selection are made based on probabilistic experiments. When the service selection changes in the next iteration, the switching between different spectrum bands induces some overheads since the SUs need to be re-configured. The evolutions of actions for selected players are shown in Fig. 5.4. While Fig. 5.2 shows that when $b = 0.3$ it takes around 350 iterations for all players to converge to pure strategies, frequent service switching happens only before 250 iterations in the learning procedure. This observation reveals that service switching, if at all happens, usually happens only in the beginning of the entire learning procedure. We

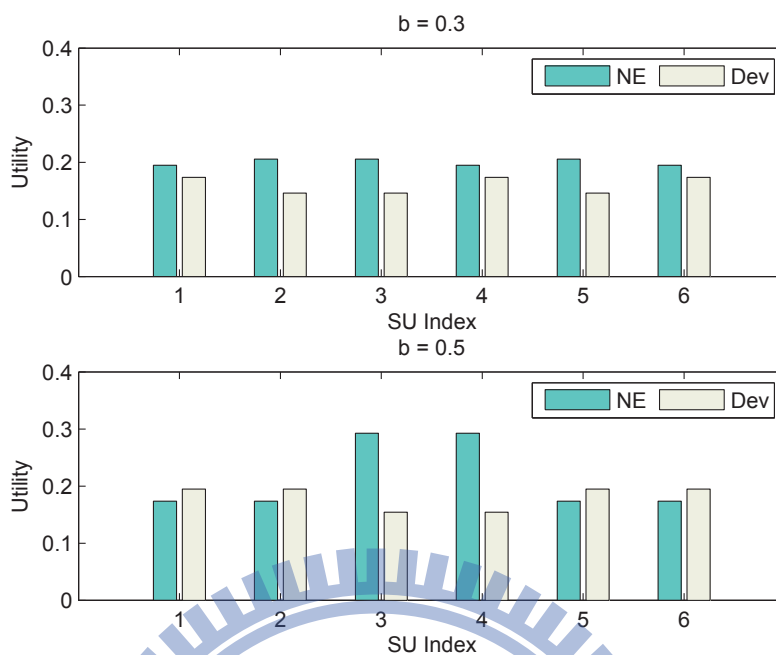


Figure 5.3: Test of unilateral deviation from the learned strategy profile of each of the $N = 6$ players, with learning rates $b = 0.3$ and $b = 0.5$.

also note that our proposed algorithm aims at learning the equilibrium strategy in the long run. The service switching and the incurred reconfiguration are manageable overheads in the early stage of the learning procedure compared to the long operation time.

5.5.2 Performance Comparison in the Lower-level Game

We further compare the performance of the proposed service selection scheme with two other approaches, namely, random selection and exhaustive search, described as follows:

- In the random selection scheme, each SU randomly subscribes to a network in each iteration. Neither learning algorithm nor centralized controller is implemented. Since the SUs possess very little knowledge on the environments, the random selection scheme is an intuitive heuristic leveraging the randomness of the external states.
- In the exhaustive search scheme, it is assumed that there exists a centralized controller which knows all system information including the numbers of SUs and SPs,

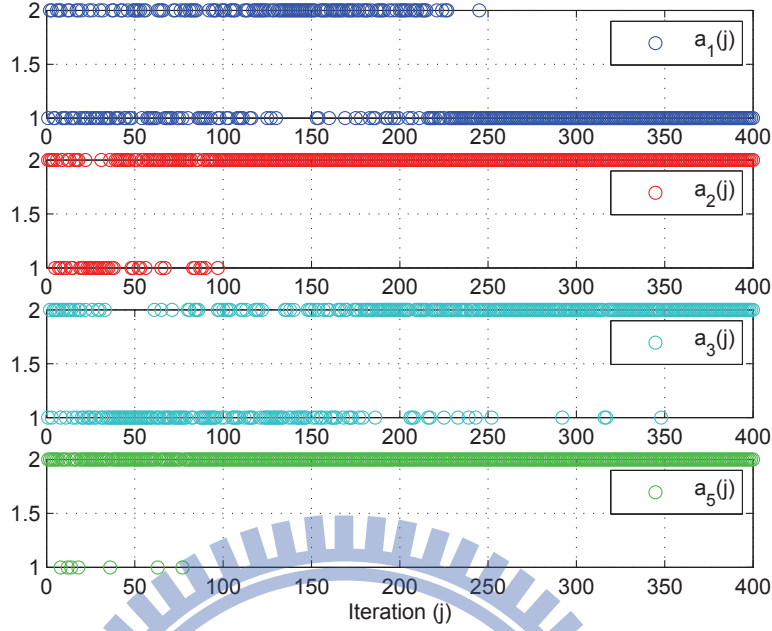


Figure 5.4: Evolution of the actions $a_i(j)$ for selected players.

and the channel availability statistics. The service selection profile is determined by maximizing the expected sum utility, i.e., finding the optimal action profile $\mathbf{a}^{opt} = (a_1, \dots, a_N)$ such that

$$\mathbf{a}^{opt} = \underset{\mathbf{a}}{\operatorname{argmax}} \sum_{i=1}^N u_i(a_i, a_{-i}). \quad (5.22)$$

The performance of different service selection schemes is first evaluated by the average normalized utility per SU. The simulation results are shown in Fig. 5.5. It is shown that the average utility of the learning method is around 87% of that of the exhaustive search. Also, the learning method outperforms the random selection method by about 20%.

As mentioned in Sec. 5.3.4, the second performance metric is the fairness among SUs. Fairness of resource allocation is usually quantified by the Jain's fairness index (JFI) [45], which is defined as

$$J = \frac{(\sum_{i=1}^N u_i)^2}{N \sum_{i=1}^N u_i^2}. \quad (5.23)$$

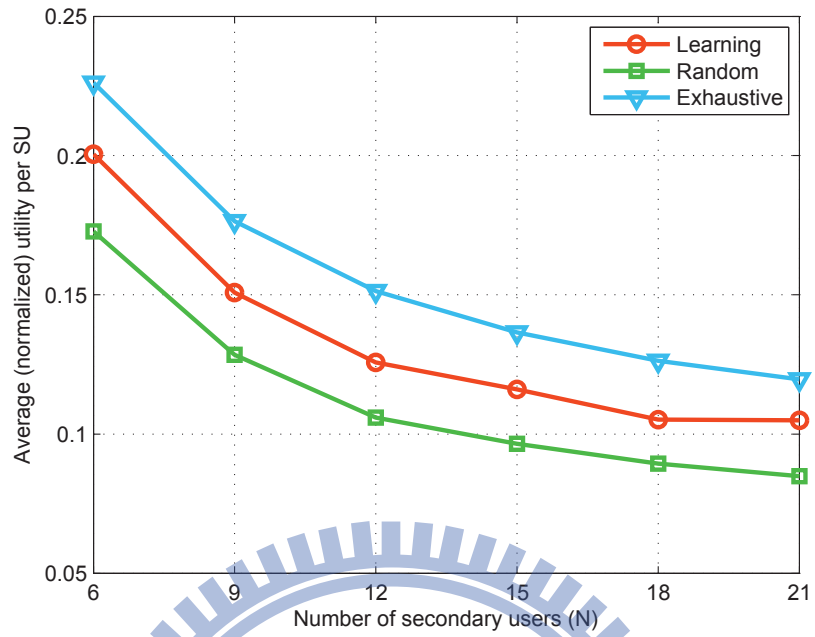


Figure 5.5: Comparison of the average (normalized) utility per SU for different service selection schemes.

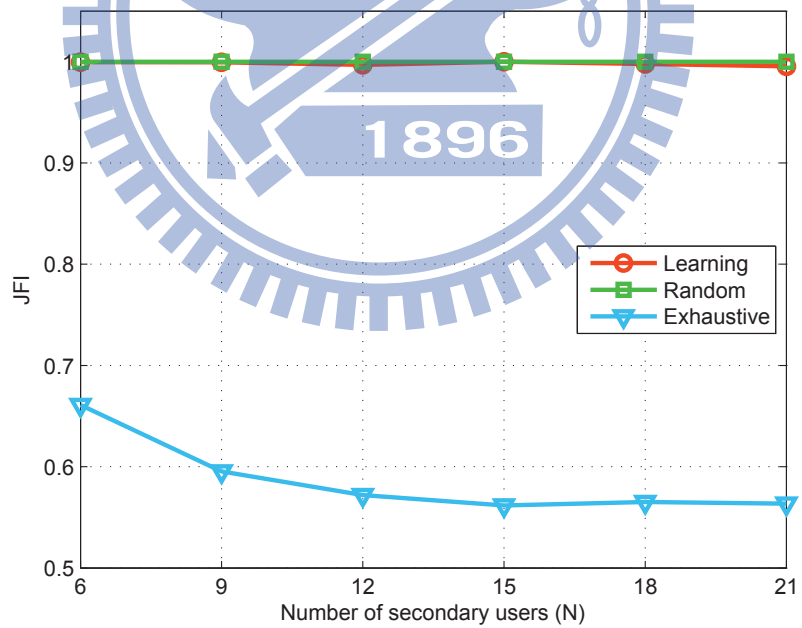


Figure 5.6: Comparison of the JFI using three service selection schemes.

The value of JFI falls in the interval of $[1/N, 1]$, and a higher JFI value indicates better fairness. The JFI of the three service selection schemes are shown in Fig. 5.6. It is

observed that the random selection and the proposed learning algorithm achieves perfect fairness ($J \approx 1$). The exhaustive search approach performs poorly in terms of fairness, as explained as follows. Notice that when $n_1, n_2 > 0$, the summed utility of the lower-level game is given by

$$u_{sum,l} = B_1\bar{c}_1 - n_1q_1 + B_2\bar{c}_2 - n_2q_2. \quad (5.24)$$

Since $q_1 < q_2$ in the current setting, the summed utility is maximized when the user load is $(n_1, n_2) = (N - 1, 1)$. That is, the spectrum opportunity of SP_2 is still utilized while the total payment is minimized. However, the *individual* utility becomes

$$u_i = \begin{cases} B_1\bar{c}_1/(N-1) - q_1, & \text{if } i \in \mathcal{N}_1, \\ B_2\bar{c}_2 - q_2, & \text{otherwise.} \end{cases} \quad (5.25)$$

Apparently, this results in poor JFI. Individual SUs would therefore prefer a more balanced user load.

From the results in Fig. 5.5 and Fig. 5.6, we show that with the proposed game-theoretic formulation, the distributed learning performs well in terms of both the total utility and fairness, under time-varying channel availability.

5.5.3 Convergence Behavior of the Upper-level Game

For the upper-level game, again we first study the convergence behaviors of the pricing strategies. Fig. 5.7 shows the evolution of the choice probabilities of the pricing strategies using Algorithm 5.2. With equal initial probabilities, it is observed that the pricing probabilities converge to pure strategies in around 80 and 120 cycles for seller 1 and seller 2, respectively. When the competition converges to the equilibrium, seller 1 sets subscription price level 2 ($q_1 = 1.5$) and seller 2 sets subscription price level 3 ($q_2 = 2$).

The evolutions of chosen prices and revenues are shown in Fig. 5.8. From the price dynamics in Fig. 5.8(a), it can be observed that while the prices do not change after 110 iterations, the revenues do change several times. This means that, although it rarely

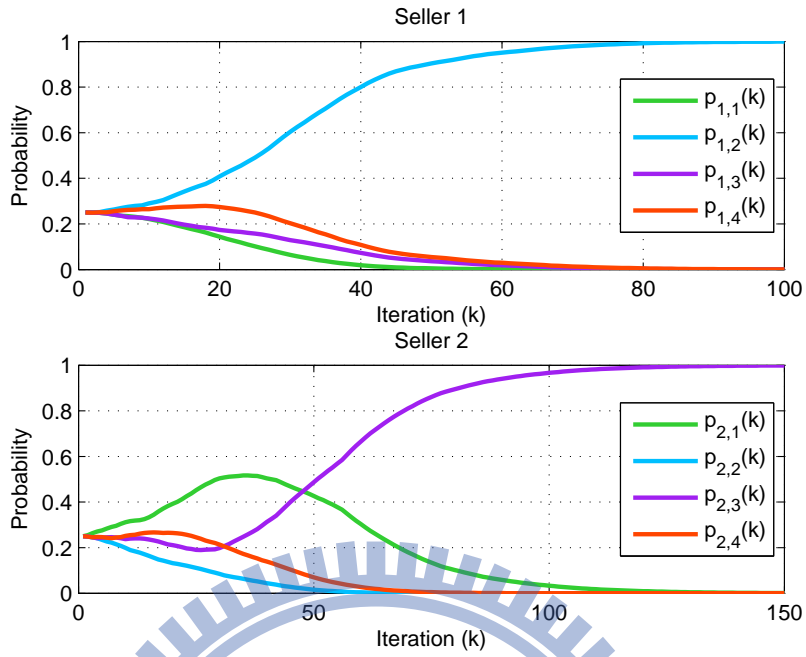


Figure 5.7: Evolution of the mixed strategies (probability of taking different actions) of the $M = 2$ sellers.

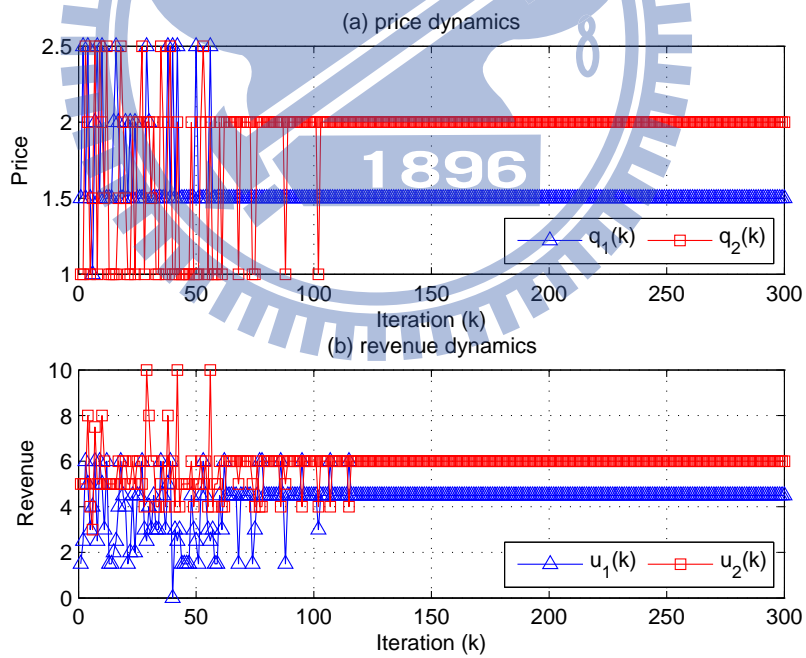


Figure 5.8: Price and revenue dynamics of the $M = 2$ sellers.

happens, the learning in the lower-level game may not converge to the NE distribution. A simple solution is to adopt smaller learning rate b in the lower-level game. However, this

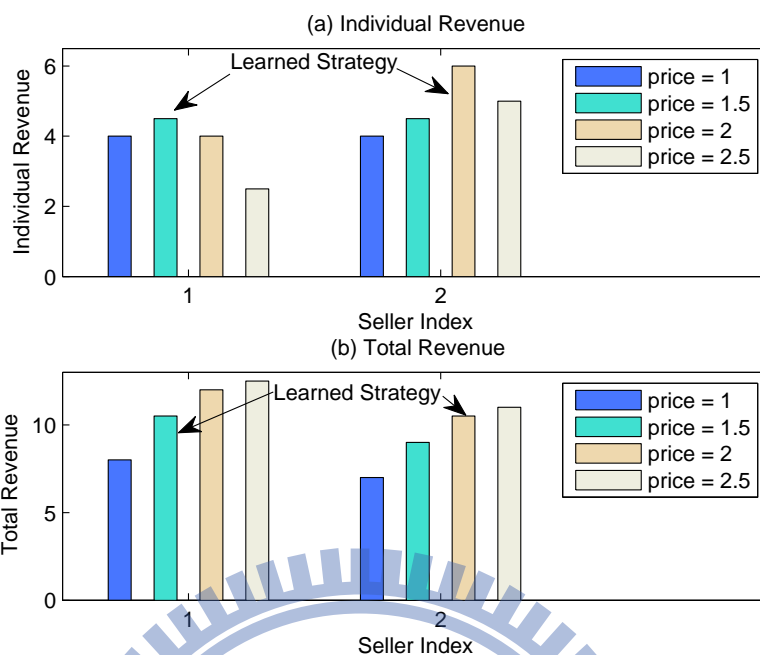


Figure 5.9: Test of different strategies. For each seller, the four bars show its revenues when taking the four different pricing strategies, while its opponent sticks to the learned strategy.

results in slower convergence. Since the occasionally wrong convergence behavior does not affect the learned pricing strategy, a more “aggressive” learning rate is adopted in practice.

The performance of the learned strategy profile is studied in Fig. 5.9. For each of the two sellers, we fix the opponent’s strategy as what it learned through Algorithm 5.2, while the seller itself tests four different pricing strategies and the performances are evaluated. The first performance metric is the individual revenue which can be used for the verification of NE, and the results are provided in Fig. 5.9(a). It is shown that when SP_2 sticks to the NE strategy (i.e. $q_2 = 1.5$), SP_1 gets best utility by also taking the NE strategy ($q_1 = 2$). Similarly, SP_2 must follow its NE strategy when SP_1 does so. On the other hand, as the second performance metric, the results of total revenue (defined as the sum of the revenues of the two SPs) are shown in Fig. 5.9(b). While unilateral deviation from the learned strategy does improve the total revenue in some cases, the seller who benefits (i.e., obtains higher revenue) is the opponent instead of the deviating seller.

Our final note is on the revenue of spectrum sellers under competition. Apparently, the summed revenue is maximized when both sellers set the highest price level (i.e., $q_1 = q_2 = 2.5$). In this case the summed revenue (i.e., the total utility in the upper-level competition) would be $U_{sum,u}^{opt} = 15$. Compared to the summed revenue obtained by using NE pricing strategy, $U_{sum,u}^{NE} = 10.5$, the efficiency of NE is only 70%. From the sellers' point of view, this is interpreted as an *efficiency loss* as a consequence of the game-theoretic formulation and price competition. While the efficiency loss is usually considered as a drawback of NE in traditional game-theoretic studies, the SUs do pay less when attached to either seller. This observation indicates the essence of the spectrum market; that is, the non-cooperative nature prevents the collusion among sellers, and buyers benefit from the price competition of sellers.

5.5.4 Non-unique User Loads

As mentioned before, when the number of SUs is finite, the user load profiles under lower-level NE may be different. To study the influence of nonunique user load profiles, we now consider a manipulated scenario with $N = 3$ SUs and identical parameters for the two sellers. Fig. 5.10 shows the dynamics of prices, utilities, and estimated utilities for both sellers. It is observed that the pricing strategies converge to the third price level (i.e., $q_1 = q_2 = 2$) for both sellers after around 200 iterations. On the other hand, the revenue (utility) oscillates between $u_m(k) = 2$ and $u_m(k) = 4$ for both sellers, which means the user loads oscillate between $(n_1, n_2) = (4, 2)$ and $(n_1, n_2) = (2, 4)$. With simply inspection we know that both user load profiles lead to NE in lower-level game. The estimated revenues (utilities) roughly converge to $\hat{u}_m \approx 3$ for $m = \{1, 2\}$. Assume that the probabilities of two user load profiles are equal (i.e., $\Pr\{(n_1, n_2) = (2, 4)\} = \Pr\{(n_1, n_2) = (4, 2)\} = 0.5$), the expected utilities are $\hat{u}_m = 3$ for $m = \{1, 2\}$. Though there may not be a rigorous proof of the distributions of the user load profiles, such an interpretation complies with the property that the estimated utilities converges toward the expected utilities, as discussed in Proposition 5.4.2.

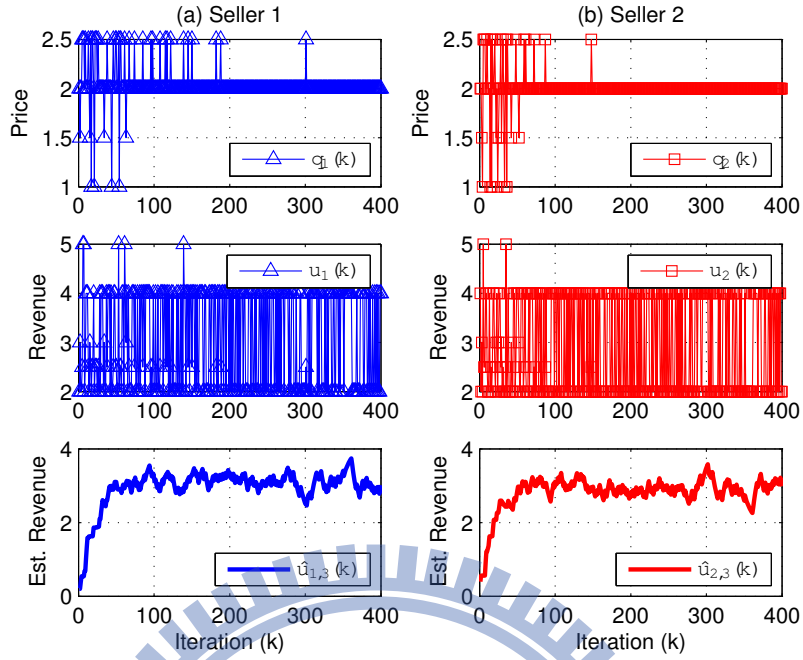


Figure 5.10: Drift in user loads. For each seller, the dynamics of prices, revenues, and estimated revenues are shown.

5.6 Conclusion and Open Issues

In this chapter we studied the problem of spectrum trading with multiple sellers and time-varying spectrum opportunities. We formulated the spectrum trading as a two-level Stackelberg game whose upper and lower level subgames model the service selection of SUs and the price competition of SPs, respectively. Decentralized stochastic learning-based algorithms were proposed for the strategic learning in both levels. Simulation results demonstrated the convergence of the algorithm towards a pure strategy Nash equilibrium point in both levels. In the lower level game, the proposed method outperforms the best random selection scheme in terms of average utility, while the performance loss compared to the centralized exhaustive search is limited. Moreover, the proposed method achieves significantly improved fairness compared to the exhaustive search method. On the other hand, the price competition among spectrum sellers decreases their summed revenue, but brings benefit to the secondary users.

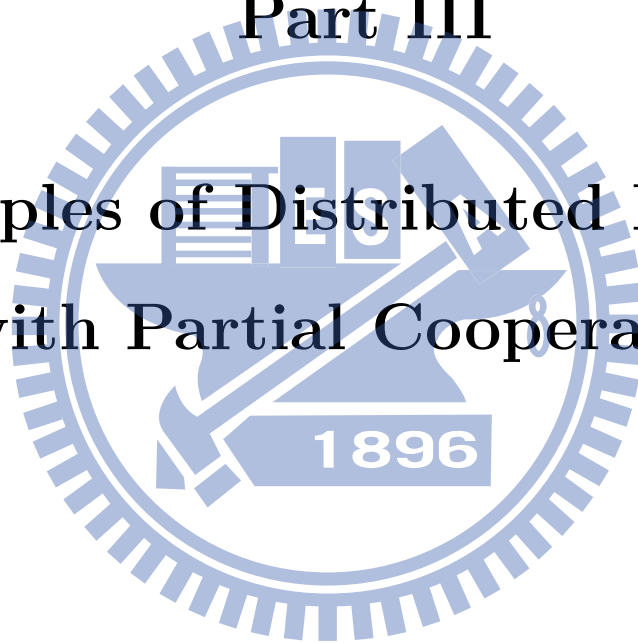
The major direction of extending this work is the consideration on more realistic model

that takes QoS parameters into account.



Part III

Examples of Distributed Learning with Partial Cooperation



Chapter 6

Self-organized Channel Assignment in Two-tier Distributed Networks

In this chapter, we study the channel assignment strategy in orthogonal frequency division multiple access (OFDMA) based two-tier distributed networks where macrocells and distributed cognitive radio networks (CRNs) are overlaid. We formulate the channel selection problem as a potential game which has at least one pure-strategy Nash equilibrium (NE). To achieve the NE we propose a stochastic learning-based algorithm which does not require the information of other players' actions and the time-varying channel. The cognitive radio base stations or cluster heads are considered as players in the game, and act as self-organized learning automata and adjust selection strategies based only on their own action-reward history. The convergence property of the proposed algorithm toward pure strategy NE points is shown theoretically and verified numerically. Simulation results demonstrate that the learning algorithm yields a 26% sensor node capacity improvement as compared to the random selection, and incurs less than 10% capacity loss compared to the exhaustive search.

6.1 Introduction

SPECTRUM utilization can be improved with two-tier networks. Efficient interference mitigation is the key to maintain the performance of two-tier networks. We start our presentation with two examples of two-tier distributed networks.

6.1.1 Examples of Two-tier Distributed Networks

Femtocell Networks

Femtocell technology [46] has been extensively considered in next-generation wireless standards such as 3GPP-LTE [47] as a means to enhance cell coverage and user capacity. In femtocell networks, the low-power and low-cost indoor base stations (referred to as home base stations, HBSs) utilize the wired broadband connection as backhaul and are planned to be easily installed by consumers. By utilizing femtocells, the indoor femtocell user equipment (FUE) is able to attain high data rate due to the short distance from HBS, and operators can reduce the cost in deploying macro base stations (MBSs) with the aid of HBS to serve the FUEs in the coverage holes.

In the absence of a central controller, resource allocation in femtocell networks is implemented in a distributed manner. Resource allocation with interference mitigation can be achieved by assigning different spectrum to adjacent femtocells. These methods can be viewed as variations of frequency planning, and usually require negotiations among HBSs.

Co-channel implementation brings the advantage of efficient spectrum usage. However, it also results in CCI between the femtocell(s) and the macrocell in various ways. In Fig. 6.1, different CCI possibilities are listed according to their sources, their victims, and whether they occur in the DL or the UL. Interference scenarios #1 and #2 involve the CCI between the femtocell user equipment (FUE) and the macrocell network, scenarios #3 and #4 involve the CCI between the macrocell user equipment (MUE) and the femtocell network, while scenarios #5 and #6 involve the CCI scenarios between close-by femtocell

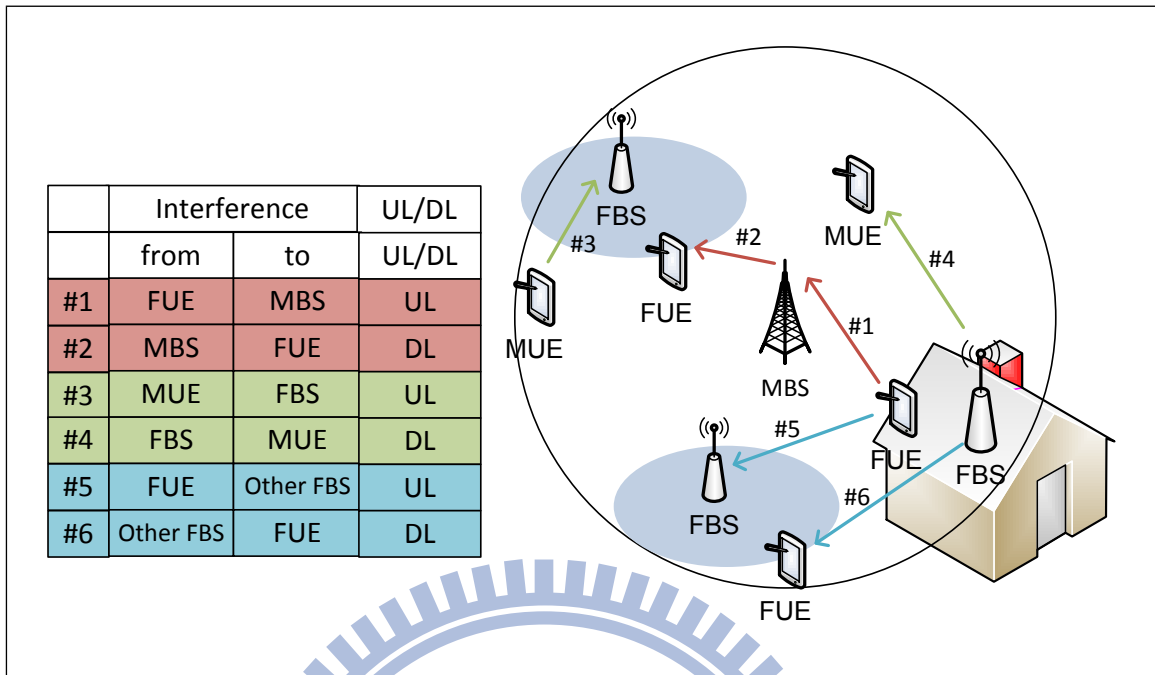


Figure 6.1: Possible interference scenarios related to femtocell communications.

networks. All these interference scenarios can be considered for both time division duplex (TDD) and frequency division duplex (FDD) systems. It should be noted that these scenarios are based on the assumption that femtocell is not allowed to be in DL mode while macrocell is in UL subframe (in TDD systems), or femtocell cannot use the UL frequency band of the macrocell for DL (in FDD systems).

Distributed Sensor Networks

In wireless sensor networks [48], spatially distributed, low-power and low-cost sensor nodes are deployed in a geographical area to monitor the environment. The sensor nodes usually form clusters, and in each cluster there is a energy-rich sensor node acting as the cluster head, while other sensor nodes are referred to as cluster members. A cluster head is a special sensor node with better cognitive radio functionality, and is responsible for the spectrum sensing and the channel assignment among its cluster members.

To enable the various kinds of services [33, 49, 50] provided by a pervasive sensing system, proper radio resource management [51] is important. Due to the spectrum scarcity

and the ad-hoc nature of sensor network deployment, it could be hard to assign licensed bands to sensor networks. Therefore, the cognitive radio [39] technology has been considered as a promising solution to the channel assignment problem of sensor networks. Cognitive radio technology enables dynamic spectrum access (DSA) and allows the unlicensed users by sensing the usage information of the spectrum from the radio environment. Akan *et al.* [52] provided a survey on cognitive radio sensor networks. By utilizing the CR technology, the sensor networks are able to attain high data rate due to the spectrum holes. In addition, dynamic spectrum access helps mitigate the interference incurred by dense deployment of sensor nodes.

Despite the promising features of cognitive sensor networks, the deployment of such heterogeneous networks with sensor clusters underlying the same spectrum as macrocells and in the same geographical area brings new technical challenges. In particular, we are interested in the case of densely populated sensor networks where, due to extensive frequency reuse, the co-channel interference (CCI) among sensor nodes and the cross-tier interference (between the macrocell and sensor networks) affect the system performance.

6.1.2 Contributions

In this chapter, we consider a two-tier distributed network, and address the self-organized channel assignment problem. The main contributions of this work are summarized as follows.

- We model the distributed channel assignment problem as an ordinal potential game (OPG). The game considers time-varying channel availability (as a result of the resource allocation to MUEs) as its external state.
- We propose a fully decentralized channel assignment algorithm in which the channel is selected by each link independently based on its action-reward history. The strategy update of all links are simultaneous, without any coordination. The convergence property of the algorithm to a pure strategy NE point is verified numerically.

Through numerical simulations, we also show that the proposed method performs quite close to the exhaustive search.

6.1.3 Game-theoretic Problem Mapping

The mapping of distributed channel assignment problem to game-theoretic formulation is summarized in Table 6.1.

Table 6.1: Mapping to game-theoretic formulation.

Elements in game	Characters in channel assignment problem
Players	Femtocell BSs or sensor cluster heads
Strategies	Channel assignments
Reward	gSINR (to be defined)
External state	Channel availability

This chapter is organized as follows. In Section 6.2, we review the previous works. In Section 6.3, the system model for two-tier distributed network is presented. Section 6.4 describes the game-theoretic model of the channel selection problem. Section 6.5 presents the stochastic learning procedure carried out by HBSs. Finally, numerical results are given in Section 6.6, and the conclusion is drawn in Section 6.7.

Notations: Normal letters represent scalar quantities; uppercase and lowercase bold-face letters denote matrices and vectors, respectively. Given a finite set \mathcal{A} , $\Delta(\mathcal{A})$ represents the set of all probability distributions over the elements of \mathcal{A} . $\mathbb{1}_{\{cond\}}$ is the indicator function which equals one if the condition *cond* is satisfied, and zero otherwise.

6.2 Related Works

In this section, we present the previous works on the spectrum sharing in distributed networks.

6.2.1 Variations of Frequency Planning

When multi-carrier techniques (e.g. OFDMA) is considered, interference mitigation can be done by allocating different channels to neighboring femtocells, like the cell planning and frequency reuse in traditional cellular systems. However, since we usually assume that there is no centralized controller, this spectrum assignment has to be done in a distributed manner. Examples include distributed random access [46], dynamic frequency planning [53], and clustering (FCRA) [54]. While the FP-like methods guarantees the interference avoidance among nearby femtocells, it does not consider the location of FUEs.

6.2.2 Learning-based Methods

To further improve the spectrum efficiency, machine learning can be implemented. In contrast, self-organized resource allocation in femtocell networks based upon reinforcement learning (RL) mechanisms has been shown effective in the literature. The stochastic learning (SL), in contrast, updates the actions of users based on their individual action-reward history. SL was applied to the spectrum access in cognitive radio networks [10] to achieve the Nash equilibrium (NE) strategy. However, fully distributed SL-based resource allocation in femtocell networks has not been extensively investigated.

We hereby review some representative works on femtocell networks based on learning.

Multi-agent Q-learning (MAQL)

Multi-agent Q-learning (MAQL) could be the most widely applied reinforcement learning method in distributed spectrum access. MAQL was applied to femtocell networks in [55–57]. MAQL involves the actions of other agents as the external state and thus requires the sharing of the knowledge of all agents' actions. The form of Q-learning is usually represented as

$$Q(s, a) \leftarrow Q(s, a) + \alpha \times \left[r(s, a) + \gamma \max_{a'}(Q(s', a')) - Q(s, a) \right], \quad (6.1)$$

where $r(s, a)$ is the immediate reward, α is the learning rate, $0 < \gamma < 1$ is the relative value of delayed versus immediate rewards, s' is the new state after action a . Then the selected action becomes:

$$\pi(a) = \arg \max_a Q(s, a) \quad (6.2)$$

When applied to a system with multiple agents, the external state of one agent involves the actions of other agents. MAQL suffers from the *curse of dimensionality*: When the number of state-action pairs is large or the input variables are continuous, the memory requirements may become infeasible. In addition, the selection of discrete sets for state and action definitions may highly affect the system performance.

In this work, Q-learning is used as the self-organization technique to manage interference in two-tier femtocell networks. While the Q-learning does achieve some kind of final stage, the property of the result is untraceable.

Regret matching and correlated equilibrium

Correlated equilibrium (CE) is a solution concept that is more general than the well known Nash equilibrium.

Definition 6.2.1. a probability distribution $\phi(a)$ on the set S of action is a correlated equilibrium of the game \mathcal{G} if, for every player $i \in N$ and every two actions $s, s' \in S_i$ of i , we have

$$\sum_{a \in \mathcal{A}: a_i = s} \phi(a) (u_i(a_i, a_{-i}) - \phi(s)u_i(s)) \leq 0, \quad (6.3)$$

where $a_{-i} \in \mathcal{A}_{-i}$ denotes the action combination of all players except i (thus $a = (a_i, a_{-i})$). The inequality means that when the recommendation to player i is to choose action s , then choosing s' instead of s cannot yield a higher expected payoff to i . In other words, ϕ is a correlated equilibrium if no player can improve his expected utility via a strategy modification. Hart and Mas-Colell [58] proposed a regret-matching (RM) method and proved the convergence toward CE. With respect to the last action chosen,

the player calculates his or her regret from not having used other actions, when those actions replace the last action each time it was used in the m periods that the player recalls.

Hunag *et al.* [59] considered downlink spectrum allocation in a long term evolution (LTE) system macrocell which contains multiple femtocells. The competition amongst cognitive HBS for spectrum resources was formulated as a non-cooperative game-theoretic learning problem where each agent (HBS) seeks to adapt its strategy in real time. A distributed spectrum access algorithm based on the regret-matching method in [58] was proposed to compute the correlated equilibrium RB allocation policy. However, with the regret-matching method, each FBS needs to evaluate all possible actions. This hinges on two implicit assumptions: (1) each FBS knows the form of its own utility function, and (2) each FBS observes the actions of *all* the other FBSs (players) at each time t . Clearly, these assumptions are unrealistic in practice due to the distributed nature of femtocell networks.

Stochastic learning for Coarse Correlated Equilibrium

To solve the impracticality in [59], Bennis *et al.* [60] further relax the constraints on equilibrium by considering the coarse correlated equilibrium (CCE).

Definition 6.2.2. Consider similar setting to definition 6.2.1, $\phi(a)$ is an ϵ -CCE if

$$\sum_{a \in \mathcal{A}, a_i = s} \phi(a) (u_i(a_i, a_{-i}) - \phi(s)u_i(s)) \geq \epsilon. \quad (6.4)$$

A regret-based SL algorithm was proposed whereby cognitive femtocells mitigate their interference toward the MUEs, on the downlink. Based on these local observations from SINR feedback, FBSs learn the probability distribution of their transmission strategies (power levels and frequency band) by minimizing their regrets for using certain strategies, while adhering to the cross-tier interference constraint. The proposed algorithm is fully decentralized, and is shown to converge to ϵ -CCE point.

Our work in this chapter is also based on SL, but the problem formulation is quite different from [60]. The details will be given in next section.

6.3 System Model

We consider a cognitive two-tier distributed network consisting of one MBS and N clusters under the coverage of the MBS. Our model can be applied to different scenarios: both femtocell network and sensor networks are possible applications. For ease of modeling and explanation, we refer to the considered network as a distributed sensor network. In this case, the method of sensor node clustering and cluster head selection [61] are also interesting topics but are out of the scope of this work. Also we consider only the single-hop transmission and omit the multi-hop routing issue for ad-hoc networks [62]. The sensors are deployed in an apartment block with a dual-stripe room layout, as shown in Fig. 6.2.

In our considered system, the medium access control (MAC) function in a cluster assembles that of cellular systems. The time domain is divided into frames, and a frame is further divided into time slots. In each frame, a cluster head allocates its cluster members (i.e., sensor nodes) in different time slots, following a time division multiple access (TDMA) rule. For simplicity, we assume that in each slot each cluster head allocates one sensor node over one of the available channels. We emphasize that the proposed method can be easily generalized to the cases with multiple sensor nodes per slot. A sensor node is in idle mode unless the current time slot is allocated for it. A cluster head is idle if one of its sensor nodes is transmitting data, and idle otherwise. We therefore introduce an *active ratio*, which is defined as the percentage of active clusters in a time slot. An exemplary time slot allocation is depicted in Fig. 6.3.

The spectrum is divided into C channels, and the channels may be licensed to different macrocells (a.k.a spectrum owners). By utilizing CR, the sensor nodes access the same frequency band as the macrocell does. Since the sensor nodes are in an energy-tight situation and operate with ultra-low power, we assume that the transmission power of a

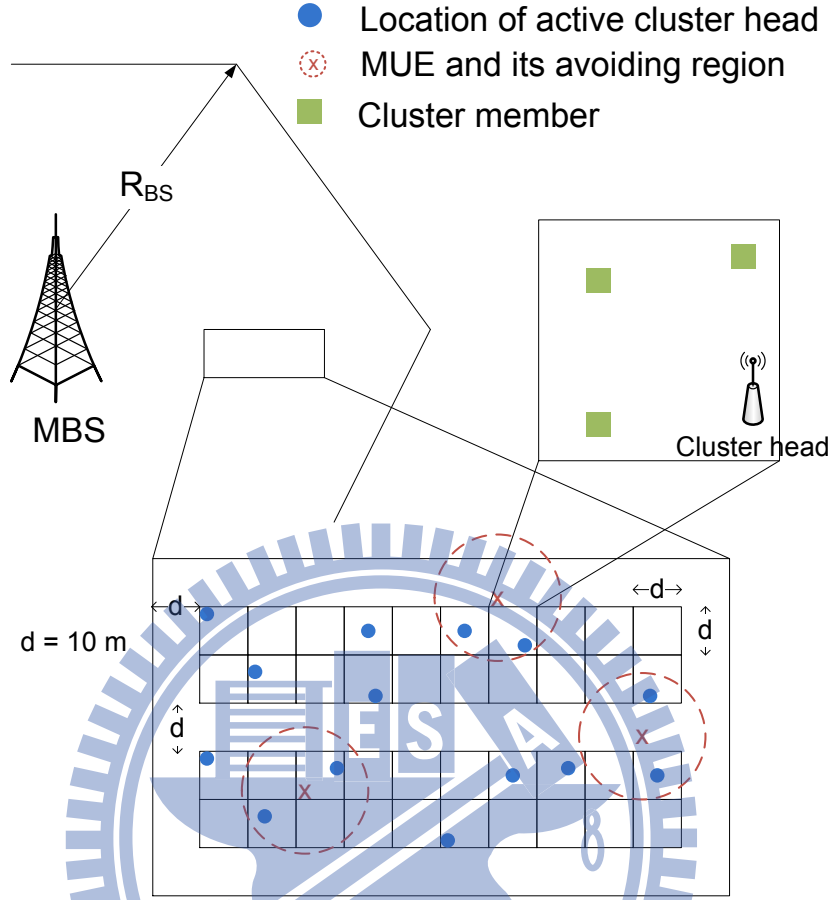


Figure 6.2: Dual-stripe deployment of sensor clusters.

macrocell user equipment (MUE) is much higher than that of the sensors. Thus, the uplink transmission of an MUE will block the nearby sensor nodes using the same spectrum. For cross-tier interference mitigation, we define an *avoiding region* for each MUE. A channel is available to a cluster only if the channel is not assigned to an MUE whose avoiding region covers the cluster head. The channel availability for sensor clusters is expressed as a binary matrix $\mathbf{X} \in \{0, 1\}^{N \times C}$, in which the element $x_{i,c}$ equals one if channel c is available to link i , and zero otherwise. The elements of \mathbf{X} follow the Bernoulli distribution, and can be described by a probability matrix $\boldsymbol{\theta} \in [0, 1]^{N \times C}$, where the element $\theta_{i,c}$ is the probability that $x_{i,c} = 1$.

Assuming perfect synchronization in time and frequency, let P_i denote the power of sensor node i , and $|h_{i,j}|^2$ indicate the link gain between cluster head i and sensor node j .

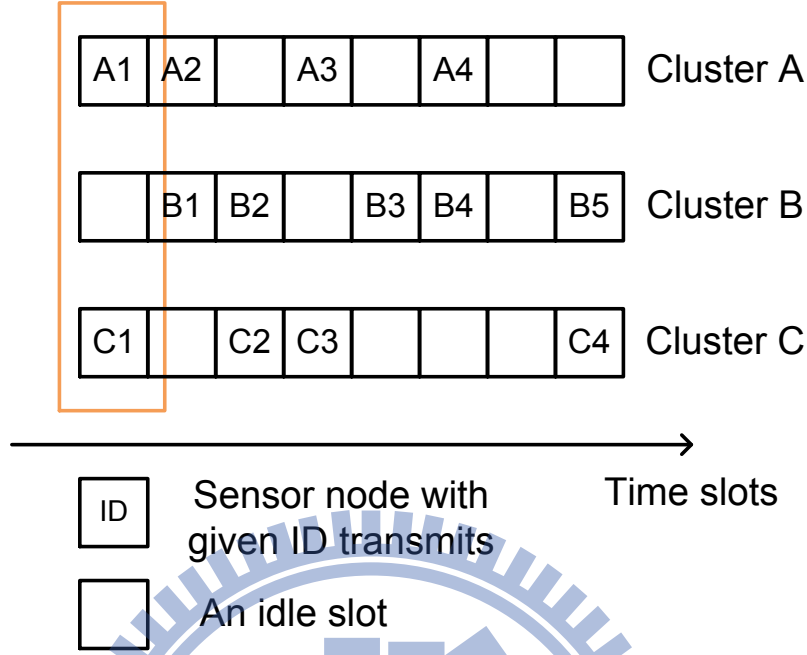


Figure 6.3: Exemplary time slot allocation in a frame. In the first slot, cluster head A and C assign channels for sensor node A1 and C1, respectively.

The interference received by cluster head i from sensor node j is given by

$$I_{j \rightarrow i} = \mathbb{1}_{\{a_i(n)=a_j(n)\}} P_j |h_{i,j}|^2, \quad \forall i, j \in \mathcal{N}, \quad (6.5)$$

where $a_i(n)$ is the action (channel selection) of cluster head i in frame n . For notational brevity, we will hereafter discard the timing dependence of the action $a_i(n)$ in occasions without ambiguity. Then, the signal-to-interference-and-noise ratio (SINR) at cluster head i can be expressed as

$$\gamma_i = \frac{P_i |h_{i,i}|^2}{\sum_{j=1, j \neq i}^N I_{j \rightarrow i} + \sigma^2}. \quad (6.6)$$

Consequently, the expected capacity for link i in bits/s/Hz is given by

$$R_i = \theta_{i,a_i} \log_2(1 + \gamma_i). \quad (6.7)$$

Let $\mathbf{a} = (a_1, \dots, a_N)$ be the channel assignment profile of all active clusters. The global

objective of the system is to find the optimal channel selection profile \mathbf{a}_{opt} that maximizes the sum capacity. Formally,

$$\mathbf{a}_{opt} = \underset{\mathbf{a}}{\operatorname{argmax}} \sum_{i=1}^N \theta_{i,a_i} \log_2(1 + \gamma_i). \quad (6.8)$$

To reflect a practical distributed network, our system model incorporates the following considerations:

1. The uplink resource allocation for MUEs is time-varying during the learning period, and the channel availability statistics (i.e., θ_{i,a_i}) is fixed but unknown to any secondary users.
2. There is no centralized controller and the channel selection is performed independently by each cluster head.
3. The number of clusters in the system, N , is unknown.

With these considerations, solving (6.8) is a challenging task, since the only available information for decision making at each individual player is its own action-reward history. Thus, a fully distributed channel selection scheme is proposed.

6.4 Game-theoretic Model

In this section, we present the game-theoretic formulation of the self-organized channel selection problem. Our objective is to devise for each cluster head a distributed channel assignment strategy that takes into account the effect of both the second-tier and cross-tier interference. We summarize our notations related to the game formulation in Table 6.2.

Table 6.2: Summary of Notations in Game-theoretic Formulation

Symbol	Meaning
\mathcal{X}	external state (channel availability)
\mathbf{X}	a realization of external state (channel availability)
\mathcal{N}	set of players
\mathcal{A}_i	set of actions of player i
$s_i \in \mathcal{A}_i$	an element of \mathcal{A}_i
$a_i(n) \in \mathcal{A}_i$	action (channel selection) of player i at slot n
$a_{-i}(n) \in \mathcal{A}_i$	actions of players except for i at slot n
$\mathcal{P}_i := \Delta(\mathcal{A}_i)$	set of probability distribution over \mathcal{A}_i
$\mathbf{p}_i(n) \in \mathcal{P}_i$	mixed strategy of player i at slot n
$r_i(n) \in \mathbb{R}$	observed utility of player i at slot n
$\hat{\mathbf{u}}_i(n) \in \mathbb{R}^{ \mathcal{A}_i }$	estimated utility vector of player i at slot n
(ϵ_i, λ_i)	learning rates of player i

6.4.1 Problem Formulation and Game Model

The channel selection problem described in the previous section can be modeled by a normal-form game with external state, expressed as a 4-tuple:

$$\mathcal{G} = (\mathcal{X}, \mathcal{N}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \{u_i\}_{i \in \mathcal{N}})$$

where \mathcal{X} is the external state (channel availability) space, \mathcal{N} is the set of players (cluster heads), \mathcal{A}_i is the set of actions (selections of channels) that player i can take, and $\{u_i\}_{i \in \mathcal{N}}$ is the utility function of player i that depends on his own action as well as the actions of other players.

Inspired by [63], the reward function is designed to consider the interference received (inward) and generated (outward) by each link. In this way, the cluster heads implicitly cooperate to reduce the interference generated toward other secondary users. We define the *generalized SINR* (gSINR) for player i as

$$\tilde{\gamma}_i = \frac{P_i |h_{i,i}|^2}{\sum_{j=1, j \neq i}^N (I_{j \rightarrow i} + I_{i \rightarrow j}) + \sigma^2}. \quad (6.9)$$

Then the instantaneous reward function of cluster head i is designed as

$$r_i = \begin{cases} \log_2(1 + \tilde{\gamma}_i), & \text{if } x_{i,a_i} = 1; \\ 0, & \text{otherwise.} \end{cases} \quad (6.10)$$

By the definition in (6.10), when the channel is available, the reward is given by Shannon's capacity formula where both inward and outward interference are accounted for. When the channel is not available, the reward is zero. Notice that the calculation of the reward function in (6.10) relies on the knowledge of other players' action. This leads to overhead due to the required information. The implementation is possible, and discussion on such protocol design can be found in [63]. The *self-organization* claimed in this work is based on the fact that the action in each time instant is selected by each player independently and simultaneously.

For systems with the channel availability as the external state, the utility function is defined as the expected reward of player i over the external state (i.e., channel availability \mathbf{X}), i.e.,

$$u_i(a_i, a_{-i}) = \theta_{i,a_i} \log_2(1 + \tilde{\gamma}_i). \quad (6.11)$$

Furthermore, if the cluster heads are assumed to be selfish and rational players, they will compete to maximize their own individual utility. In fact, a selfish cluster head will not only maximize the capacity of its own user but also reduce the interference. Formally, the game \mathcal{G} is expressed as:

$$(\mathcal{G}) : \max_{a_i \in \mathcal{A}_i} u_i(a_i, a_{-i}), \quad \forall i \in \mathcal{N}. \quad (6.12)$$

Notice that the calculation of gSINR requires the knowledge of the interference each HBS causes to FUEs served by other HBSs, which can be obtained via proper protocol design in distributed systems [63]. While additional signaling is brings higher complexity, the utility function design induces self-acting coordination of HBSs. When competing to maximize the individual utility, a selfish and rational HBS will not only maximize its own capacity but also reduce the interference toward the FUEs served by other HBSs.

Moreover, the formulation can be easily generalized to the cases with multiple FUEs per slot (e.g., following the way that [64] generalizes [63]).

6.4.2 Analysis of Nash Equilibrium

With the utility function defined in (6.11), we show the existence of an NE point for the proposed game in the following proposition.

Proposition 6.4.1. *The game \mathcal{G} is an ordinal potential game (OPG) which possesses at least one pure strategy NE.*

Proof: Consider the function $\Phi : \times_{i \in \mathcal{N}} \mathcal{A}_i \rightarrow \mathbb{R}_+$:

$$\Phi(\mathbf{a}) = \log_2 \left(1 + \frac{\sum_{k=1}^N P_k |h_{k,k}|^2}{\sum_{k=1}^N \sum_{j=1, j \neq k}^N I_{k \rightarrow j}} \right). \quad (6.13)$$

Now consider an improvement step made by cluster head i that changes its action unilaterally from a_i to \check{a}_i , so that $u_i(\check{a}_i, a_{-i}) > u_i(a_i, a_{-i})$. Defining $I_{i \rightarrow j} \triangleq \mathbb{1}_{\{\check{a}_i = a_j\}} P_i |h_{j,i}|^2$, and $I_{j \rightarrow \check{i}} \triangleq \mathbb{1}_{\{\check{a}_i = a_j\}} P_j |h_{i,j}|^2$, we have

$$\begin{aligned} & u_i(\check{a}_i, a_{-i}) > u_i(a_i, a_{-i}) \\ \Leftrightarrow & \sum_{j=1, j \neq i}^N [I_{i \rightarrow j}^\check{+} + I_{j \rightarrow \check{i}}] < \sum_{j=1, j \neq i}^N [I_{i \rightarrow j} + I_{j \rightarrow i}] \\ \Leftrightarrow & \sum_{j=1, j \neq i}^N [I_{i \rightarrow j}^\check{+} + I_{j \rightarrow \check{i}}] + \sum_{j=1, j \neq i}^N \sum_{k=1, k \neq i, j}^N I_{j \rightarrow k} \\ < & \sum_{j=1, j \neq i}^N [I_{i \rightarrow j} + I_{j \rightarrow i}] + \sum_{j=1, j \neq i}^N \sum_{k=1, k \neq i, j}^N I_{j \rightarrow k}. \end{aligned} \quad (6.14)$$

Here we have used the fact that when cluster head i changes its action, the effects are only on the interference that it receives ($I_{j \rightarrow i}$) and generates ($I_{i \rightarrow j}$). From (6.13) and (6.14), we obtain

$$u_i(\check{a}_i, a_{-i}) - u_i(a_i, a_{-i}) > 0 \Leftrightarrow \Phi(\check{a}_i, a_{-i}) - \Phi(a_i, a_{-i}) > 0. \quad (6.15)$$

Therefore, \mathcal{G} is an OPG with potential function Φ , and the existence of a pure strategy NE is always guaranteed [12] since it coincides with the local maxima of the potential function. This completes the proof. ■

Notice that the term $\sum_{k=1}^N \sum_{j=1, j \neq k}^N I_{k \rightarrow j}$ in the potential function Φ denotes the summation of all mutual interference in the sensor network. Therefore, every NE point is the strategy profile that is a local maximum of the summed interference.

6.5 Stochastic Learning Procedure

Here, we discuss obtaining the NE via stochastic learning. As the channel state is time-varying and the action is selected by each player simultaneously and independently in each play, previous algorithms that require complete information (e.g., better response dynamics [12]) may not be applicable here. Thus, we propose a decentralized stochastic learning (SL)-based algorithm by which the BSs learn toward the equilibrium strategy profile from their individual action-reward history.

The proposed distributed channel assignment (DCA) algorithm for cognitive sensor networks is described in Algorithm 6.1.

In each play, the channel selection is based on a probability distribution over the set of channels. After each play, cluster head i obtains the instantaneous reward and updates the mixed strategy (i.e. channel selection vector) $\mathbf{p}_i(n)$ and utility estimation $\hat{\mathbf{u}}_i(n)$. Notably, the utility estimation serves as a reinforcement signal so that higher utility induces higher probability in the next play. Furthermore, the proposed learning algorithm is fully distributed, and the channel selection is solely based on individual action-reward experience without a centralized controller. In fact, the proposed algorithm belongs to the combined fully-distributed payoff strategy reinforcement learning (CODIPAS-RL) [9]. The evolution of the mixed strategies is described as follows.

Proposition 6.5.1. *The DCA Algorithm converges to a pure strategy NE for OPGs if the learning rates are sufficiently small.*

Algorithm 6.1 Distributed Channel Assignment (DCA)

- 1: Initially, set $n = 0$. Set the channel assignment probability vector and utility estimation as

$$p_{i,s_i}(0) = 1/|\mathcal{A}_i|, \hat{u}_{i,s_i}(-1) = 0, \forall i \in \mathcal{N}, s_i \in \mathcal{A}_i.$$

- 2: At the beginning of the n th slot, each player selects an action $a_i(n)$ according to the current channel assignment probability $\mathbf{p}_i(n)$.
- 3: In each slot, each BS transmits data. At the end of each slot, each BS receives the instantaneous reward $r_i(n)$ specified by (15) depending on the precoding scheme.
- 4: All players update their channel assignment probability vector and utility estimation according to the rules:

$$\begin{cases} \hat{u}_{i,s_i}(n) - \hat{u}_{i,s_i}(n-1) \\ \quad = \eta_i \mathbb{1}_{\{a_i(n)=s_i\}} (r_i(n) - \hat{u}_{i,s_i}(n-1)) \\ p_{i,s_i}(n+1) = \frac{p_{i,s_i}(n)(1+\epsilon_i)^{\hat{u}_{i,s_i}(n)}}{\sum_{s'_i \in \mathcal{A}_i} p_{i,s'_i}(n)(1+\epsilon_i)^{\hat{u}_{i,s'_i}(n)}} \end{cases} \quad (6.16)$$

where ϵ_i and η_i are the learning rates for action probability and utility estimation, respectively.

6.6 Numerical Results

For system-level simulations, we consider a cognitive sensor network deployed within the coverage of a cellular network. As in Fig. 1, the simulation environment includes one macrocell covering one dual-stripe apartment block. The apartment block contains 40 single-floor apartments. There is one sensor cluster in each apartment. When a sensor cluster is active, its cluster head assigns one channel to cluster members randomly located in the same apartment. Without loss of generality, we consider the channel assignment in the first slot of each frame, in which for each active cluster there is one cluster member. The simulation parameters are listed in Table 6.3.

6.6.1 Convergence of the proposed SL-based learning algorithm

We first study the time-evolving behaviors of the proposed stochastic learning method.

Table 6.3: Simulation Parameters

Parameter	Value
Min. distance between nodes	3 m
Carrier Frequency	2 GHz
Number of Channels	2
Transmission Bandwidth of Each Channel	180 kHz
Path Loss and Shadowing	Table A.2.1.1.2-8 [47]
Penetration loss	Table A.2.1.1.2-8 [47]
Sensor Transmission Power	1mW
Thermal Noise	-174 dBm/Hz
Learning Rates (default)	$(\lambda_i, \epsilon_i) = (0.1, 0.1)$

Evolution of mixed-strategies

Fig. 6.4 shows the evolutions of the channel assignment probabilities (i.e., mixed strategy) using the proposed SL-based algorithm. We consider different learning rates and study the convergence behaviors. It is observed that, with equal initial probability, the channel assignment probability converges to a pure strategy (i.e., the probability of choosing one strategy approaches one) in around 80 and 20 iterations for $\epsilon = 0.1$ and $\epsilon = 0.5$, respectively. As expected, larger learning rate results in faster convergence.

Verification of NE

As shown in Fig. 6.4, the convergence toward pure strategy is observed for both $\epsilon = 0.1$ and $\epsilon = 0.5$. An intuitive question to ask is: Does the resulting strategy profile achieve the Nash equilibrium? In Fig. 6.5, we verify the NE property by testing the unilateral deviation with a 25% active ratio and different learning rates. As can be seen from Fig. 6.5(a), when $\epsilon = 0.1$, a unilateral deviation results in lower utility for all players. In other words, the outcome of the learning algorithm is an NE point. On the other hand, when $\epsilon = 0.5$, as shown in Fig. 6.5(b), link #4 and #8 both achieve higher throughput by unilateral deviation, and thus the resulting strategy is no longer an NE point. These results reflect the trade-off between accuracy and convergence speed we mentioned before.

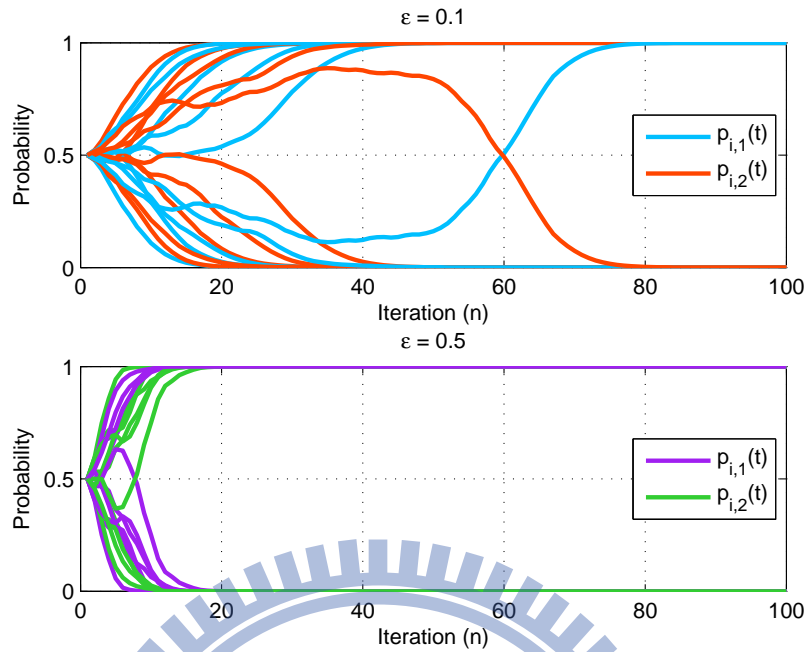


Figure 6.4: Evolution of the mixed strategies (probability of taking different actions) of all players. Each pair of $p_{i,1}(t)$ and $p_{i,2}(t)$ shows the behavior of player i .

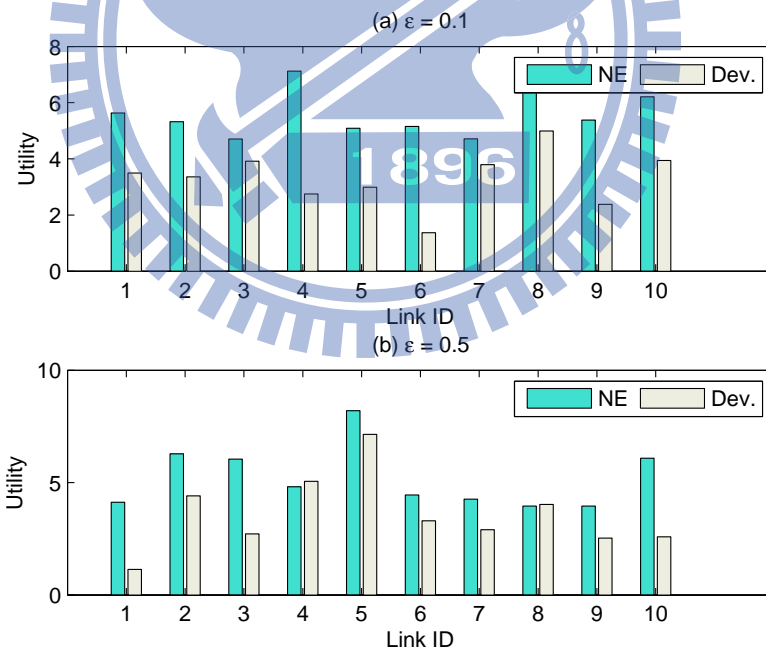


Figure 6.5: Test of unilateral deviation from the resulting strategy profile of each of the 10 players.

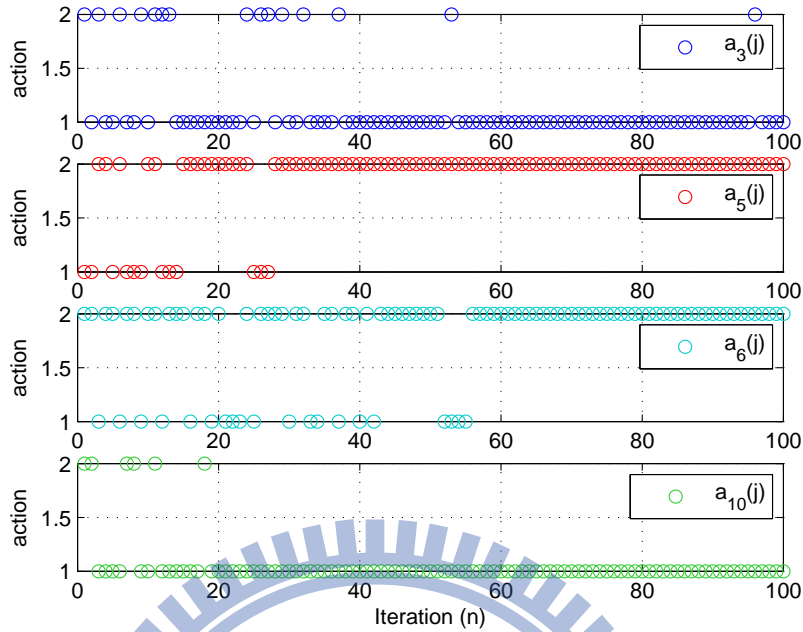


Figure 6.6: Evolution of the actions $a_i(j)$ for some players.

Evolution of Actions

During the learning procedure, the channel assignment is based on probabilistic experiments. When the channel assignment changes in the next frame, the switching between different channels brings overhead since the sensor node needs to be re-configured. The evolution of actions for selected players are shown in Fig. 6.6. As can be seen, while Fig. 6.4(a) reveals that it takes around 80 iterations for all players to converge to pure strategies, the actions seldom change after about 60 iterations in the learning procedure. This suggests that channel switching, if at all happens, usually happens only in the beginning of the entire learning procedure. Actually, our proposed learning algorithm aims at learning the equilibrium strategy in the long run. The channel switching and the incurred sensor node reconfiguration are manageable overheads compared to the long operation time.

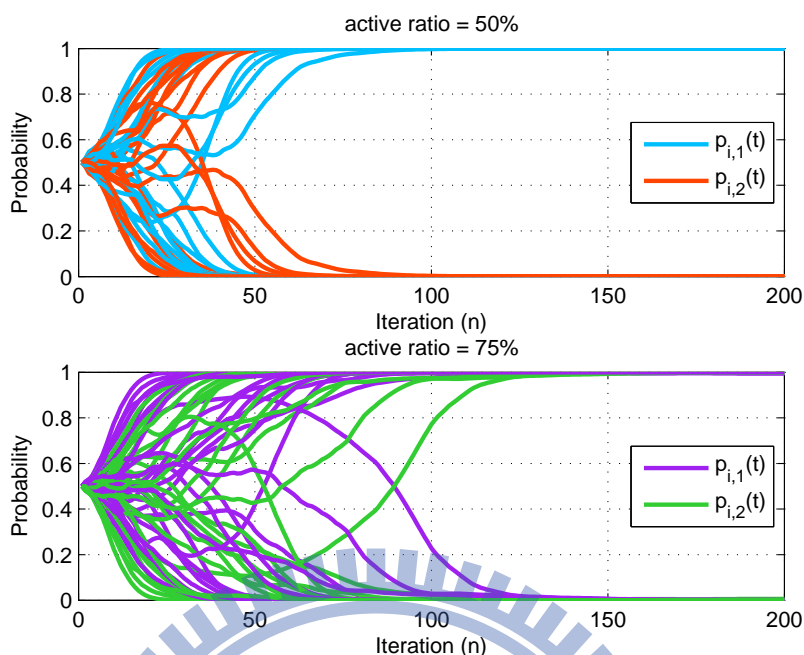


Figure 6.7: Evolution of the mixed strategies (probability of taking different actions) of all players with active ratios of 50% and 75%. Each pair of $p_{i,1}(t)$ and $p_{i,2}(t)$ shows the behavior of a player $i \in \mathcal{N}$.

Different active ratios

We further consider different active ratios, and investigate the convergence behaviors under different levels of mutual interference. The results for active ratio of 50% and 75% are shown in Fig. 6.7. We observe that the convergence toward pure strategy takes around 100 and 150 iterations for active ratio of 50% and 75%, respectively. Comparing the case of 25% active ratio in Fig. 6.4(a), we see that it takes fewer iterations for densely active networks to converge than for sparsely active sensor networks.

6.6.2 Capacity performance

Capacity under unilateral deviation

In Fig. 6.5 we have shown that unilateral deviation leads to decreased utility. While the altruistic utility function design reduces the mutual interference, we are also interest

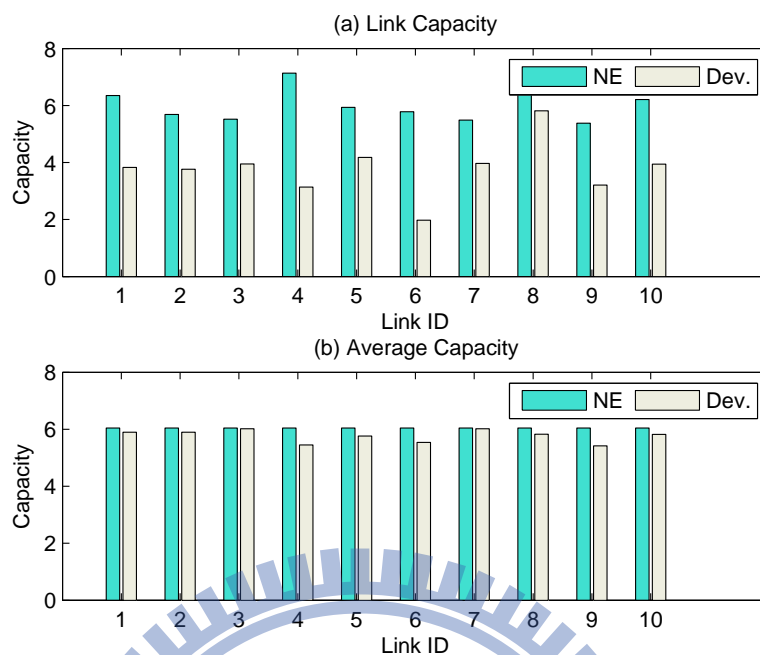


Figure 6.8: Test of unilateral deviation from the resulting strategy profile of each of the 10 players.

in the performance of Nash equilibrium strategy in terms of the throughput of each cluster as well as the whole system. Therefore, in Fig. 6.8 we test the change on capacity under unilateral deviation from the NE strategy for all players. As depicted in Fig. 6.8(a), there is no significant change on the average capacity per sensor link when only one player unilaterally deviates from its NE strategy. From Fig. 6.8(b) we observe that for all players, deviation from NE strategy decreases their own capacity.

Comparison with Other Methods

We further compare the performance of the proposed channel selection scheme with two other approaches, namely, random allocation and exhaustive search, described as follows:

- In the random allocation scheme, each cluster head randomly selects a channel for its sensor node in each frame. Neither learning algorithm nor centralized controller is implemented.

Table 6.4: Comparison of the capacity and fairness for different channel assignment schemes

Number of SUs	Proposed	Exhaustive	Random
active ratio = 25%, R_{avg}	6.0426	6.2433	4.7912
active ratio = 25%, J	0.9370	0.8512	0.9516
active ratio = 50%, R_{avg}	4.8375	4.9454	4.0955
active ratio = 50%, J	0.8855	0.8235	0.9056

- In the exhaustive search scheme, it is assumed that there exists a centralized controller which knows all system information including the channel gains, the channel availability statistics, and the number of clusters. The channel assignment profile is determined by maximizing the expected sum capacity (i.e., solving (6.8)).

The performance of different channel selection schemes are evaluated by the average capacity per sensor node, $R_{avg} = \frac{1}{N} \sum_{i=1}^N R_i$ and the fairness among sensor nodes. In the literature, fairness of resource allocation is usually quantified by the Jain's fairness index (JFI) [45], which is defined as

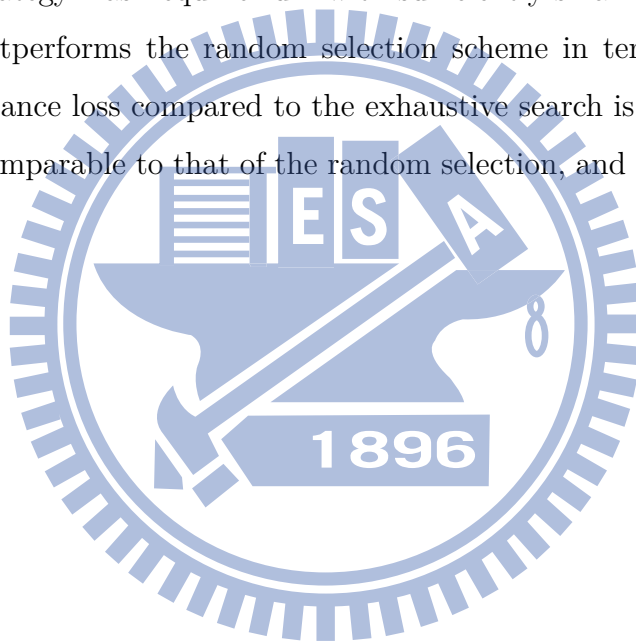
$$J = \frac{(\sum_{i=1}^N R_i)^2}{N \sum_{i=1}^N R_i^2}. \quad (6.17)$$

The value of JFI falls in the interval of $[1/N, 1]$, and a higher JFI value indicates better fairness.

The simulation results of average capacity and JFI for different active ratios are summarized in Table 6.4. We observe that the exhaustive search method results in the best average capacity with worst fairness. The random selection scheme, in contrast, has the lowest average capacity but good fairness due to its randomness nature. The proposed method performs well balanced in terms of both average capacity and fairness. The results show the advantages of the proposed method: through the learning procedure toward equilibrium, the capacity of each player is considered and fewer players are sacrificed. If we examine the final channel selection profile, it is observed that in the progress of convergence toward the NE point, the proposed learning algorithm allocates the mutually interfered users on different channels.

6.7 Concluding Remarks

In this work, we studied the problem of self-organized channel assignment in distributed two-tier networks with unknown channel and unknown number of clusters. We presented a game-theoretic approach to distributively manage interference and enable the coexistence of sensor and macrocell operations in a scenario where secondary nodes operate in the same spectrum as a cellular system. We modeled channel assignment problem by means of an ordinal potential game. A decentralized stochastic learning algorithm has been proposed. Simulation results have demonstrated the convergence of the algorithm toward a pure strategy Nash equilibrium with sufficiently small learning rates. The proposed method outperforms the random selection scheme in terms of average capacity, while the performance loss compared to the exhaustive search is limited. In addition, its fairness level is comparable to that of the random selection, and surpasses the exhaustive search scheme.



Chapter 7

Distributed Channel Allocation in Network MIMO

The cooperative frequency reuse among base stations (BSs) can improve the system spectral efficiency by reducing the intercell interference (ICI) through channel selection and precoding. This chapter presents a game-theoretic study of channel selection for realizing network multiple-input multiple-output (MIMO) operation under time-varying wireless channel. We propose a new joint precoding scheme that carries enhanced interference mitigation and capacity improvement abilities for network MIMO systems. We formulate the channel selection problem as a non-cooperative game with BSs as the players, and show that our game is an exact potential game (EPG) given the proposed utility function. A decentralized, stochastic learning-based algorithm is proposed where each BS progressively moves toward the Nash equilibrium (NE) strategy based on its action-reward history and not actions taken by others. The convergence properties of the proposed learning algorithm toward a pure-strategy NE point are theoretically shown and numerically verified for different network topologies. The proposed learning algorithm also demonstrates a fine capacity and fairness performance as compared to other schemes through extensive link-level simulations.

7.1 Introduction

UNIVERSAL frequency reuse is a key technique to improve the throughput of broadband wireless networks. However, frequency reuse among neighboring cells inevitably results in intercell interference (ICI) and degrades the achievable throughput performance. To overcome this problem, ICI management techniques such as ICI coordination (ICIC) and base-station cooperation (BSC) have been proposed [65,66]. BSC, also known as network multiple-input multiple-output (MIMO), is a multi-antenna signal processing technique that enables several nearby BSs to jointly serve multiple mobile stations (MSs). The implementation of network MIMO may require a partial or full sharing of channel state information (CSI) and data among the BSs.

Much of the research on network MIMO and multicell cooperation has focused on signal processing techniques in an orthogonal frequency-division multiple access (OFDMA) system. The channel assignment for each MS is generally assumed determined or treated separately from the network MIMO mechanism. Efficient channel allocation (particularly in a distributed manner) for network MIMO in a multi-antenna multicell environment has not yet been extensively studied. The aim of this work is therefore to study the distributed channel allocation problem in network MIMO systems. We adopt a game-theoretic approach and incorporate reinforcement learning procedures into the proposed channel selection game where each player (i.e., the BS) can act (i.e., perform channel selection) without explicitly knowing other players' actions and the forms of utility functions. The main contributions of this work are as follows:

- We propose a novel joint processing scheme where an MS is jointly served by a set of selected BSs. The capacity advantages of the proposed scheme over conventional precoding methods are numerically demonstrated.
- We formulate the channel allocation problem as a non-cooperative game and show the existence of Nash equilibrium (NE). A stochastic learning (SL)-based algorithm is developed to achieve self-organized channel allocation. The convergence beha-

vivors of the proposed algorithm toward an NE point are theoretically proven and numerically verified for different network topologies.

The rest of the chapter is organized as follows. In Section 7.2, we review related works on precoding in multicell multi-antenna networks as well as those on distributed resource allocation. In Section 7.3, the system model and the proposed joint processing are described. The game-theoretic formulation of the channel allocation problem is presented in Section 7.4 and the SL-based solutions are presented in Section 7.5. Numerical results are provided in Section 7.6. Conclusion is given in Section 7.7.

7.2 Related Works

7.2.1 Precoding with BS Cooperation

Under the concept of network MIMO, the actual implementation may vary depending on the degrees of CSI and data sharing that are needed.

Static Clustering

In the static clustering scheme, a fixed set of nearby BSs cooperate in jointly serving the users where precoding techniques for single-cell multiuser MIMO systems (e.g., block-diagonalization (BD) [67]) are applied to mitigate the multiuser interference. One disadvantage of static clustering is its requirement of a full sharing of data and CSI within a cluster, which creates a significant overhead on the system operation [68]. [69] considered the BD scheme. The overhead will be even greater if inter-cluster interference is considered [70].

Partial Data

To reduce the information exchange overhead, partial cooperation has been proposed which does not require a full sharing of CSI and/or data. Kaviani *et al.* [71] proposed a

precoding scheme according to the minimum mean square error (MMSE) criterion and Kerret and Gesbert [72] developed a sparse precoding method which determines the most efficient data sharing patterns, both assuming partial data sharing among the BSs.

partial CSI

Distributed MIMO precoding was introduced by Kerret and Gesbert [73] assuming partial CSI sharing but full data sharing. Zakhour *et al.* [74, 75] proposed a distributed precoding scheme by maximizing the virtual signal-to-interference-and-noise ratio (VSINR) with local CSI. Bjornson *et al.* [76] developed a network MIMO scheme for large cellular networks, where the precoding vectors are computed in centralized (by a central controller) or fully distributed (by each BS independently) fashion with partial CSI and data.

7.2.2 Spectrum Sharing

The realization of network MIMO in an OFDMA system involves an important issue: channel allocation. Traditionally, frequency planning with spatial reuse was considered [46, 53, 54] to mitigate the ICI among adjacent cells. Dynamic channel allocation schemes were proposed for cognitive radio networks (CRNs) [77] and network MIMO [76], which however requires the presence of a central station or negotiations among BSs. The development of self-organized, fully-distributed resource allocation schemes can be facilitated by the application of game theory. Self-organized resource allocation in wireless networks based on reinforcement learning (RL) has been studied [8, 10, 11, 56, 57, 63, 78]. Within the RL framework, multiagent Q-learning (MAQL) was applied to CRNs [56] and femtocell networks [57]. MAQL involves the actions of other agents as the external state and thus requires the sharing of the knowledge of all agents' actions. The stochastic learning (SL), in contrast, updates the actions of users based on their individual action-reward history. SL has been applied to the game-theoretic study of dynamic spectrum access in

CRNs [10, 63] and precoder selection in multiple access channels [8] to learn the equilibrium strategy profile [11]. An application of SL on both strategy and payoff, referred to as the combined fully-distributed payoff strategy reinforcement learning (CODIPAS-RL), was found for MIMO power loading [78]. Hybrid CODIPAS-RL was applied to heterogeneous 4G networks and the convergence of users' network selection was observed [9]. While SL algorithms have shown promise for wireless applications in the literature, their applications in fully distributed resource allocation for multiantenna multicell networks as well as distributed networks of random geometry have not been well studied.

Notations: Normal letters represent scalar quantities; upper-case and lower-case boldface letters denote matrices and vectors, respectively. $(\cdot)^T$ and $(\cdot)^H$ stands for the transpose and the conjugate transpose, respectively. \mathbf{I} and $\mathbf{0}$ represent the identity matrix and zero vector with proper size, respectively. $\mathbb{1}_{\{cond\}}$ is the indicator function which equals one if the condition *cond* is satisfied, and zero otherwise.

7.3 System Model

7.3.1 The Network MIMO System

We consider the downlink of an N -cell network MIMO OFDMA system where in each cell there is one BS equipped with N_t antennas serving several single antenna MSs. The set of BSs is denoted as \mathcal{N} . The time domain is divided into slots, while the licensed spectrum is divided into K available orthogonal subchannels, each having the same bandwidth. In each time slot, each BS serves one MS over one of the available subchannels following the time division multiple access (TDMA) policy. A subchannel may be reused by multiple BSs.

In the considered multicell network, an MS may be served cooperatively by multiple BSs in a network MIMO setting. Since BSs and MSs distant apart cause negligible interference to each other, we consider joint transmission only among nearby BSs to

reduce the overhead of data sharing and CSI exchange on the backhaul. For ease of exposition, we make the following definitions:

- Each MS_i estimates and feedbacks the CSI to a set of BSs in its coordination set, which is defined as

$$\mathcal{C}_i = \{b \in \mathcal{N} \mid \rho_{ib}^2 \geq \alpha_{th}\rho_{ii}^2\} \quad (7.1)$$

where ρ_{ib}^2 is the large-scale channel gain between BS_b and MS_i , which can be obtained by averaging over the estimated channel gain at the receiver, and the threshold $0 < \alpha_{th} \leq 1$ is a system design parameter.

- Each MS_i receives the data from its service set, which is defined as

$$\mathcal{D}_i = \{b \in \mathcal{N} \mid \rho_{ib}^2 \geq \beta_{th}\rho_{ii}^2\} \quad (7.2)$$

where $\beta_{th} \geq \alpha_{th}$ and $\mathcal{D}_i \subseteq \mathcal{C}_i$.

In the network MIMO system, a BS_b ($b \in \mathcal{C}_i$) can mitigate the interference to the MSs in the coverage area of the other BSs in \mathcal{C}_i through proper precoder designs. An illustrative example of the network MIMO system with joint processing is given in Fig. 7.1.

To reflect a practical wireless network, our system model incorporates the following considerations:

1. The channel state is time-varying. Its statistical characteristics (e.g., Rayleigh fading) are fixed but unknown at the BSs during the learning period. Only several adjacent subchannels in the OFDMA system are used for operation in network MIMO model. Their total bandwidth is wider than the coherence bandwidth and the MSs move slowly so that the slot length is longer than the coherence time. Thus, the channel undergoes frequency-flat block fading.
2. The number of cells, N , is unknown.
3. Each BS selects the channel independently without having to consider the actions taken by other BSs.

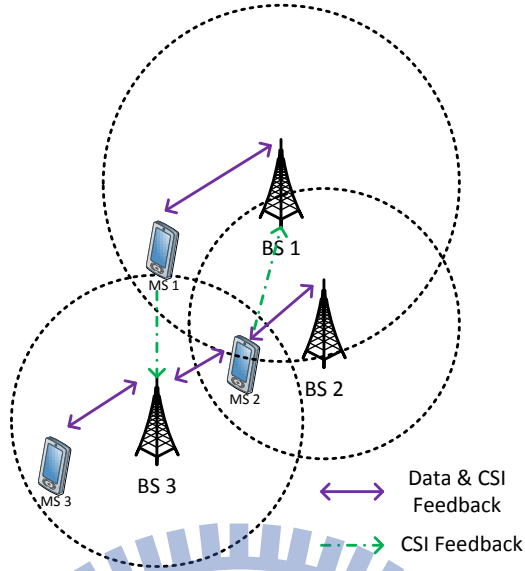


Figure 7.1: Illustration of distributed channel selection with joint precoding in multicell networks. For MS_1 , $\mathcal{C}_1 = \{1, 3\}$ and $\mathcal{D}_1 = \{1\}$, where BS_1 and BS_3 both receive CSI feedback from MS_1 and perform interference mitigation but only BS_1 serves MS_1 . For MS_2 , $\mathcal{C}_2 = \{1, 2, 3\}$ and $\mathcal{D}_2 = \{2, 3\}$, where BS_2 and BS_3 jointly serve MS_2 while all three BSs perform interference mitigation. For MS_3 , $\mathcal{C}_3 = \{3\}$ and $\mathcal{D}_3 = \{3\}$, where only BS_3 serves MS_3 .

Notably, the only available information for the proposed channel selection game is the history of each individual player's channel selection strategies and rewards.

7.3.2 Transmitter Precoding

In the network MIMO system considered in [75], only data are shared among the BSs and the precoding vector is calculated separately at each BS. Here, we propose a joint processing method in which the BSs in the serving set of each user exchange their knowledge of CSI and determine the precoding vector jointly. Similar to [75], a power splitting procedure is considered which allows each BS to split its transmission power among the MSs that it needs to serve. Let P_b be the transmission power of BS_b on one subchannel. We adopt a simple equal-power splitting method so that the power allocated

to MS_{*i*} by BS_{*b*} is given by

$$P_{ib} = \frac{P_b}{\sum_{i=1}^N \mathbb{1}_{\{\mathcal{D}_i \ni b\}}}, \quad \forall i \text{ s.t. } \mathcal{D}_i \ni b. \quad (7.3)$$

Signal transmission in the multicell network MIMO system is modeled as follows. Let $\mathbf{h}_{ib} \in \mathbb{C}^{N_t \times 1}$ represent the channel from BS_{*b*} to MS_{*i*}. The symbol x_i denotes the data intended for MS_{*i*}, where $\mathbb{E}[|x_i|^2] = 1$ and $\mathbb{E}[x_i^* x_j] = 0, \forall i \neq j$. The data symbol x_i is precoded by precoders $\mathbf{w}_{ib} \in \mathbb{C}^{N_t \times 1}, \forall b \in \mathcal{D}_i$. Let $D_j = |\mathcal{D}_j|$ be the cardinality of the serving set of MS_{*j*}. Then, the channels from the BSs in \mathcal{D}_j to MS_{*i*} can be expressed as

$$\mathbf{h}_{i, \mathcal{D}_j} = \left[\sqrt{P_{jb_1}} \mathbf{h}_{ib_1}^T, \dots, \sqrt{P_{jb_{D_j}}} \mathbf{h}_{ib_{D_j}}^T \right]^T \quad (7.4)$$

and the collective precoding vector for MS_{*i*} as

$$\mathbf{w}_i = \left[\mathbf{w}_{ib_1}^T, \dots, \mathbf{w}_{ib_{D_i}}^T \right]^T \quad (7.5)$$

Let $a_i(n)$ be the selected channel for MS_{*i*} (i.e., the action taken by BS_{*i*}) at slot n . For notational brevity, we will hereafter discard the timing dependence of the action $a_i(n)$ in occasions without ambiguity. The discrete-time baseband signal received by MS_{*i*} is given by

$$y_i = \mathbf{h}_{i, \mathcal{D}_i}^T \mathbf{w}_i x_i + \sum_{j=1, j \neq i}^N \mathbb{1}_{\{a_i = a_j\}} \mathbf{h}_{i, \mathcal{D}_j}^T \mathbf{w}_j x_j + z_i \quad (7.6)$$

where the first term is the desired signal, the second term represents the ICI, and z_i is additive complex Gaussian noise with variance σ^2 . Therefore, the signal-to-interference-and-noise ratio (SINR) at MS_{*i*} can be formulated as

$$\gamma_i = \frac{\|\mathbf{h}_{i, \mathcal{D}_i}^T \mathbf{w}_i\|^2}{\sum_{j=1, j \neq i}^N \mathbb{1}_{\{a_i = a_j\}} \|\mathbf{h}_{i, \mathcal{D}_j}^T \mathbf{w}_j\|^2 + \sigma^2}. \quad (7.7)$$

The achievable capacity for MS_{*i*} in bits/s/Hz is given by

$$R_i = \log_2 \left(1 + \frac{\gamma_i}{\Gamma} \right) \quad (7.8)$$

where $\Gamma = -\ln(5\text{BER})/1.5$ is a function of the required bit error rate (BER), often known as the SINR gap [79].

We denote the precoding vector \mathbf{w}_i for MS_{*i*} by $\mathbf{w}_i = \mu_i \hat{\mathbf{w}}_i$, where μ_i is an adjustment factor to maintain the per-BS power constraint and $\hat{\mathbf{w}}_i$ is the unit-norm vector that maximizes the *modified signal-to-leakage-and-noise ratio* (mSLNR). Different from the SLNR in [80], we consider the mSLNR to reflect a practical network MIMO operation, which is defined in terms of the signal power received by MS_{*i*} and the *available* information about the interference caused to other MSs (produced by the signals from \mathcal{D}_i intended for MS_{*i*}) plus the noise power. The distinction on the interference part is made to reflect the fact that not all CSI can be acquired by the BSs in \mathcal{D}_i and thus the interference powers imposed on other users may not be available. Specifically, in our consideration a BS in \mathcal{D}_i can only acquire the CSI to MS_{*j*} ($i \neq j$) if this BS is also in \mathcal{C}_j . Mathematically, $\hat{\mathbf{w}}_i$ is given by

$$\hat{\mathbf{w}}_i = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \frac{\|\mathbf{h}_{i,\mathcal{D}_i}^T \mathbf{w}\|^2}{\underbrace{\sigma^2 + \sum_{j=1, j \neq i}^N \mathbb{1}_{\{a_i=a_j\}} \|\tilde{\mathbf{h}}_{j,\mathcal{D}_i}^T \mathbf{w}\|^2}_{\text{mSLNR of MS}_i}} \quad (7.9)$$

where

$$\tilde{\mathbf{h}}_{j,\mathcal{D}_i} = \left[\sqrt{P_{ib_1}} \check{\mathbf{h}}_{jb_1}^T, \dots, \sqrt{P_{ib_{D_i}}} \check{\mathbf{h}}_{jb_{D_i}}^T \right]^T \quad (7.10)$$

with

$$\check{\mathbf{h}}_{jb} = \begin{cases} \mathbf{h}_{jb}, & \text{if } b \in \mathcal{C}_i \cap \mathcal{C}_j \\ \mathbf{0}, & \text{otherwise} \end{cases} \quad (7.11)$$

The vector $\check{\mathbf{h}}_{jb}$ reflects our mSLNR consideration; that is, it is equal to the CSI when this information can be collected (via feedbacks or backhaul communications), and zero otherwise.

The solution to (7.9) is given by

$$\hat{\mathbf{w}}_i = \frac{\mathbf{K}_i^{-1} \mathbf{h}_{i, \mathcal{D}_i}}{\|\mathbf{K}_i^{-1} \mathbf{h}_{i, \mathcal{D}_i}\|} \quad (7.12)$$

where $\mathbf{K}_i = \sigma^2 \mathbf{I} + \sum_{j \neq i} \mathbb{1}_{\{a_i = a_j\}} \tilde{\mathbf{h}}_{j, \mathcal{D}_i} \tilde{\mathbf{h}}_{j, \mathcal{D}_i}^H$. We then employ a heuristic approach similar to [70] to obtain the adjustment factor μ_i as

$$\mu_i = \frac{1}{\max\{\|\mathbf{w}_{ib_1}\|, \|\mathbf{w}_{ib_2}\|, \dots, \|\mathbf{w}_{ib_{D_i}}\|\}}. \quad (7.13)$$

Note that the multicell precoding scenario considered in [74] is a special case of our proposed method. In this local precoding scheme, each BS's knowledge of CSI is limited to the channel between itself and the MSs under its coverage. Each BS's CSI is obtained through a feedback mechanism and maintained locally. By setting $\beta_{th} > 1$ in our system, the serving set of each MS will consist of its home BS only and thus the system reduces to local precoding. The performance of local precoding may be limited since the neighboring BSs of an MS act only as a source of interference without providing any useful data streams. The performance comparison of local precoding and joint processing is presented in Section 7.6.

7.4 Channel Selection for Network MIMO

In this section, we present the game-theoretic formulation of the self-organized channel selection to realize the network MIMO scheme described in Sec. 7.3. Our objective is to devise a distributed channel selection strategy that takes into account the effect of ICI. We summarize our notations related to the game formulation in Table I.

Table 7.1: Summary of Notations in Game-theoretic Formulation

Symbol	Meaning
\mathcal{H}	external state (channel state) space
\mathbf{H}	random matrix for the channel state
\mathcal{N}	set of players
\mathcal{A}_i	set of actions of player i
$s_i \in \mathcal{A}_i$	an element of \mathcal{A}_i
$a_i(n) \in \mathcal{A}_i$	action (channel selection) of player i at slot n
$a_{-i}(n) \in \mathcal{A}_i$	actions of players except for i at slot n
$\mathcal{P}_i := \Delta(\mathcal{A}_i)$	set of probability distribution over \mathcal{A}_i
$\mathbf{p}_i(n) \in \mathcal{P}_i$	mixed strategy of player i at slot n
$r_i(n) \in \mathbb{R}$	instantaneous reward of player i at slot n
$\hat{\mathbf{u}}_i(n) \in \mathbb{R}^{ \mathcal{A}_i }$	estimated utility vector of player i at slot n
(ϵ_i, η_i)	learning rates of player i

7.4.1 Game-Theoretic Formulation

We model the channel selection as a non-cooperative game with external state, expressed as a 4-tuple:

$$\mathcal{G} = (\mathcal{H}, \mathcal{N}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \{u_i\}_{i \in \mathcal{N}})$$

where \mathcal{H} is the external state (channel state) space, $\mathcal{N} = \{1, \dots, N\}$ is the set of players (BSs), $\mathcal{A}_i = \{1, \dots, K\}$ is the set of actions (selections of channels) that player i can take, and u_i is the ergodic utility function of player i defined as the expected reward over the time-varying channel state, i.e.,

$$u_i(a_i, a_{-i}) \triangleq \mathbb{E}_{\mathbf{H}} [r_i(a_i, a_{-i}; \mathbf{H})] \quad (7.14)$$

where a_{-i} represents the actions of other players except for i , and $r_i : \times_{i \in \mathcal{N}} \mathcal{A}_i \mapsto \mathbb{R}$ represents the instantaneous reward function for player i under a given channel state \mathbf{H} . By intuition, the achievable capacity in (7.8) may be considered as the reward function. However, we notice that in [63] the interference terms related to the action of player i are treated as the cost of player i , and the negation of summed cost is defined as the reward. The advantage of this reward function design lies in that, during the learning procedure, in addition to maximizing its own rate, a player now also tends to minimize the

interference generated to other players due to its action. Therefore, implicit *coordination* can be achieved even with a noncooperative game formulation. In this work, with the joint processing scenario and inspiration by [63], we propose to design the reward function as

$$r_i(a_i, a_{-i}; \mathbf{H}) \triangleq - \left[\sum_{j=1, j \neq i}^N I_{j \rightarrow i} + \sum_{j=1, \mathcal{D}_j \cap \mathcal{C}_i \neq \emptyset}^N \sum_{m=1, m \neq i, j}^N I_{j \rightarrow m} \right] \quad (7.15)$$

where

$$I_{j \rightarrow i} \triangleq \mathbb{1}_{\{a_i = a_j\}} \frac{\left\| \tilde{\mathbf{h}}_{i, \mathcal{D}_j}^T \mathbf{w}_j \right\|^2}{\left\| \tilde{\mathbf{h}}_{j, \mathcal{D}_j}^T \mathbf{w}_j \right\|^2} \quad (7.16)$$

is the interference caused at MS_{*i*} by the signal intended for MS_{*j*} normalized by the received signal power of MS_{*j*}. The considered reward function is composed of the *I*-values that may vary when player *i* changes its action. The first term in (7.15) accounts for the total interference caused at MS_{*i*} as a result of external BSs. This *selfish* part reflects similar interest to (7.8): lower interference means higher achievable capacity. On the other hand, the second term in (7.15) is the *altruistic* part of the reward function, which accounts for the interference imposed on other MSs by the signal intended for MS_{*j*} when $\mathcal{D}_j \cap \mathcal{C}_i \neq \emptyset$. Effectively, the reward function in (7.15) considers both the suppression of the interference that each BS causes to out-of-cell MSs and the optimization of the desired received signal power in each cell.

7.4.2 Existence of Nash Equilibrium

We assume that the players (i.e., the BSs) in the proposed game are selfish and rational. In other words, they will compete to maximize their individual utilities, i.e., maximizing their own throughput while reducing the interference generated to others.

Definition 7.4.1. An action profile $\mathbf{a}^* = (a_1^*, \dots, a_N^*)$ is a pure strategy Nash equilibrium (NE) point of the noncooperative game \mathcal{G} if and only if no player can improve its utility

by deviating unilaterally, i.e.,

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*), \quad \forall i \in \mathcal{N}, \forall a_i \in \mathcal{A}_i \setminus \{a_i^*\}. \quad (7.17)$$

With the reward function defined in (7.15), we show the existence of an NE point for the proposed game in the following proposition.

Proposition 7.4.1. *The proposed channel selection game \mathcal{G} is an exact potential game (EPG) with at least one pure strategy NE point.*

Proof: For a channel selection profile (a_i, a_{-i}) , consider the following function $\Phi : \times_{i \in \mathcal{N}} \mathcal{A}_i \mapsto \mathbb{R}$ for the game \mathcal{G} :

$$\Phi(a_i, a_{-i}) = \mathbb{E}_{\mathbf{H}} \left[- \sum_{j=1}^N \sum_{m=1, m \neq j}^N I_{j \rightarrow m} \right]. \quad (7.18)$$

Observing that player i 's change does not affect the precoder of MS $_j$ if $\mathcal{D}_j \cap \mathcal{C}_i = \emptyset$, we define

$$r_{-i}(a_{-i}; \mathbf{H}) \triangleq - \sum_{\substack{j=1, \\ \mathcal{D}_j \cap \mathcal{C}_i = \emptyset}}^N \sum_{\substack{m=1, \\ m \neq i, j}}^N I_{j \rightarrow m}. \quad (7.19)$$

Considering a unilateral strategy for player i that changes its action unilaterally from a_i to \check{a}_i , we have

$$\begin{aligned} & u_i(\check{a}_i, a_{-i}) - u_i(a_i, a_{-i}) \\ &= \mathbb{E}_{\mathbf{H}}[r_i(\check{a}_i, a_{-i}; \mathbf{H})] - \mathbb{E}_{\mathbf{H}}[r_i(a_i, a_{-i}; \mathbf{H})] \\ &= \mathbb{E}_{\mathbf{H}}[r_i(\check{a}_i, a_{-i}; \mathbf{H}) + r_{-i}(a_{-i}; \mathbf{H})] - \mathbb{E}_{\mathbf{H}}[r_i(a_i, a_{-i}; \mathbf{H}) + r_{-i}(a_{-i}; \mathbf{H})] \\ &= \Phi(\check{a}_i, a_{-i}) - \Phi(a_i, a_{-i}). \end{aligned} \quad (7.20)$$

According to the definition in [12], \mathcal{G} is an EPG with Φ as its potential function, and the existence of a pure strategy NE point is guaranteed. This completes the proof. \blacksquare

One important property of a potential game is that the interests of players align to a global objective: maximization of the potential function. For example, with (7.18), the

players in \mathcal{G} actually minimize the total cost in the system. This property suggests the possibility of distributed learning toward the equilibrium.

7.4.3 Acquisition of the Interference Information

Obtaining the exact interference information for the reward function in (7.15) may be difficult in a practical protocol design. However, if we combine the two terms in (7.15) and approximate

$$I_{j \rightarrow i} \approx 0, \forall j \in \mathcal{N} \text{ s.t. } \mathcal{D}_j \cap \mathcal{C}_i = \emptyset, \quad (7.21)$$

the instantaneous reward function in (7.15) can be approximated by

$$r_i \approx - \sum_{j=1, \mathcal{D}_j \cap \mathcal{C}_i \neq \emptyset}^N I_j^{out} \quad (7.22)$$

where

$$I_j^{out} = \sum_{j=1, j \neq m}^N I_{j \rightarrow m} \quad (7.23)$$

which defines the outward interference of player j . In other words, the reward function of player i takes into account the players whose interference set overlaps with the serving set of player i . A two-step protocol can therefore be established:

1. Each player calculates its own I_j^{out} based on the CSI feedback, and
2. Each player exchanges the information with other players.

7.5 Stochastic Learning-based Channel Selection Algorithm

There has been much interest in designing learning algorithms toward NE in non-cooperative games. However, the external state (CSI) is unknown and the action is

selected by each player simultaneously and independently in each play. Therefore, previous algorithms requiring complete information and implicit ordering of acting players (e.g., those based on better response dynamics (BRD) [12] and fictitious play (FP) [13]) may not be feasible in our self-organized multicell resource allocation problem. In this section, we develop a decentralized SL-based algorithm where the BSs move toward the equilibrium strategy based on their individual action-reward history.

7.5.1 Algorithm Description

The proposed SL-based channel selection algorithm is described in Algorithm 7.1. In each play, the channel is selected based on the probability distribution over the set of channels. After each play, a player obtains the instantaneous reward and updates the channel selection probability vector as well as the utility estimation vector $\hat{\mathbf{u}}_i(n)$. The utility estimation serves as a reinforcement signal so that higher utility (lower cost) leads to higher probability in the next play. Notably, the proposed learning algorithm is fully distributed: the channel selection is solely based on individual action-reward experience without a centralized controller. Moreover, although the SL-based algorithm proposed in [11] also converges to NE points for potential games, its probability update rule requires the normalization of the instant reward such that its value will lie in $[0, 1]$. This requirement of normalization makes the algorithm inapplicable when the extreme values of reward functions are unavailable. This restriction however does not apply to the proposed algorithm due to a different probability update rule.

7.5.2 Convergence Properties of the Proposed Algorithm

Convergence toward pure strategy NE points is an important feature of the proposed learning algorithm. Similar to the discussions in [11] and [10], here we theoretically demonstrate the convergence properties of the proposed SL-based algorithm. First, by using the ordinary differential equation (ODE) approximation we characterize the long-term behavior of the sequence $\{\mathbf{P}(n)\}$. Second, we establish a sufficient condition for

Algorithm 7.1 Stochastic Learning toward NE

- 1: Initially, set $n = 0$, and the channel selection probability vector as $p_{i,s_i}(n) = 1/|\mathcal{A}_i|, \forall i \in \mathcal{N}, s_i \in \mathcal{A}_i$.
- 2: At the beginning of the n th slot, each player selects an action $a_i(n)$ according to the current channel selection probability $\mathbf{p}_i(n)$.
- 3: In each slot, each BS transmits data. At the end of each slot, each BS receives the instantaneous reward $r_i(n)$ specified by (7.16) depending on the precoding scheme.
- 4: All BSs update their channel selection probability vector and utility estimation according to the rules:

$$\begin{cases} p_{i,s_i}(n+1) = \frac{p_{i,s_i}(n)(1-\epsilon_i)^{-\hat{u}_{i,s_i}(n)}}{\sum_{s'_i \in \mathcal{A}_i} p_{i,s'_i}(n)(1-\epsilon_i)^{-\hat{u}_{i,s'_i}(n)}} \\ \hat{u}_{i,s_i}(n+1) - \hat{u}_{i,s_i}(n) = \eta_i \mathbb{1}_{\{a_i(n)=s_i\}} (r_i(n) - \hat{u}_{i,s_i}(n)) \end{cases} \quad (7.24)$$

where ϵ_i and η_i are the learning rates for action probability and utility estimation, respectively.

the arrival at NE points for the proposed learning algorithm and prove that the game \mathcal{G} satisfies this condition.

Proposition 7.5.1. *With sufficiently small ϵ_i and η_i , the probability matrix sequence $\{\mathbf{P}(n)\}$ converges to \mathbf{P}^* which is the solution of the following ODE:*

$$\frac{dp_{i,s_i}(t)}{dt} = p_{i,s_i}(t) \left[\psi_i(s_i, \mathbf{P}) - \sum_{s'_i \in \mathcal{A}_i} \psi_i(s'_i, \mathbf{P}) p_{i,s'_i}(t) \right]. \quad (7.25)$$

The boundary condition is given by $\mathbf{P}(0) = \mathbf{P}_0$, where \mathbf{P}_0 is the initial channel selection probability matrix. The estimated utility converges to

$$\hat{u}_{i,s_i}(n) \rightarrow \psi_i(s_i, \mathbf{P}). \quad (7.26)$$

Proof: See [7, Section 4.3]. ■

Note that the ODE in (7.25) is the *replicator equation* [14] in which the probability of taking one strategy grows if this strategy's current estimated utility is larger than the average utility over all strategies and declines otherwise. Compared to the best response dynamics where a player changes its strategy in the next iteration to the best action

according to other players' action, a player adjusts the weighting for each possible action in each iteration with the replicator dynamics.

Proposition 7.5.2. *The proposed learning algorithm has the following properties:*

1. *All Nash equilibria are stationary points;*
2. *All stationary points that are not Nash equilibria are unstable.*

These properties follow directly from the replicator equation in (7.25). For an intuitive explanation, we first define

$$\bar{\psi}_i \triangleq \sum_{s'_i \in \mathcal{A}_i} \psi_i(s'_i, \mathbf{P}) p_{i,s'_i}(t) \quad (7.27)$$

which can be interpreted as the expected utility over current action probabilities. Then, by definition, achieving NE implies

$$\psi_i(s_i, \mathbf{P}) = \bar{\psi}_i, \quad \forall s_i \in \mathcal{A}_i. \quad (7.28)$$

This also constitutes a stationary point of the ODE in (7.25).

Proposition 7.5.3. *Suppose that there exists a nonnegative function $\Psi : \mathbb{R}^{|\mathcal{A}|} \rightarrow \mathbb{R}$ such that*

$$\psi_i(s_i, \mathbf{P}) = \frac{\partial \Psi(\mathbf{P})}{\partial p_{i,s_i}}. \quad (7.29)$$

Then, the SL-based algorithm converges to a pure strategy NE point of a noncooperative game.

Proposition 7.5.3 establishes a sufficient condition that guarantees the convergence toward NE. In what follows, we prove that the proposed channel selection game \mathcal{G} satisfies this condition and hence it converges to a pure-strategy NE point by using the SL-based channel selection algorithm.

Proposition 7.5.4. *When applied to EPGs, the proposed SL-based channel selection algorithm converges to an NE point.*

Note that the learning rates (ϵ_i, η_i) play an important role in the convergence behavior of the proposed SL-based learning algorithm. In particular, smaller learning rates lead to a slower convergence. The choice of learning rates poses a trade-off between accuracy and speed, and may be determined by training in practice.

7.6 Numerical Results and Discussions

In this section, our theoretical developments are numerically verified in hexagonal cellular networks as well as distributed networks of random geometry. Universal frequency reuse is adopted in our link-level simulations. The simulation setup follows the 3GPP model [81] and is summarized in Table 7.2.

7.6.1 Convergence Behaviors of the Proposed Learning Algorithm

We plot the evolution of the channel selection probability (i.e., the mixed strategies) of the proposed stochastic learning algorithm for four arbitrarily selected players in Fig. 7.2. It is observed that, with equal initial probabilities, the channel selection probabilities converge to a pure strategy in around 700 cycles. For other players in the game which are not shown, a similar convergence result is also observed.

Fig. 7.3 shows the evolution of the estimated cost vector (i.e., $-\hat{\mathbf{u}}_i$) of two selected players. As can be seen, the BSs tend to select the channel with lower estimated cost (solid lines). Fig. 7.2 and Fig. 7.3 demonstrate that, with high probability, mutually interfering cells can coordinate their transmissions on different channels even without negotiations.

We verify the (mean-field) NE property by testing the deviation of the channel selection of each of the 19 players. The results shown in Fig. 7.4 are time-averaged values starting from the slot where the pure strategy can be identified until the end of simulation. It is shown in Fig. 7.4(a) that for all players a unilateral deviation produces higher (time-averaged) cost; in other words, the learning algorithm converges to an NE point. In addition, we test the change of (time-averaged) capacities under unilateral deviation. As

Table 7.2: The Simulation Setup

Cellular Parameters	
Number of Cells, N	19 (wrap-around)
Cell Radius, R_{BS}	500 m
Min. MS to home BS distance	$0.7R_{BS}$
Number of Tx Antennas, N_t	2
OFDMA Parameters	
FFT Size	128
Carrier Frequency	2 GHz
Subcarrier Spacing	15 kHz
Number of Subchannels	6
Number of Subcarriers per Subch.	12
Subch. for network MIMO mode	subch. 1 & 2 ($K = 2$)
Channel Model Parameters	
PathLoss (dB)	$34.5 + 35 \log_{10} d$ (d in m)
Shadowing Std. Dev.	8 dB
Speed of MSs	3 km/h
Fast Fading	Ray-based model (Sec. 5 of [81])
Power Control Parameters	
Trans. Power	46 dBm
Thermal Noise Power	-174 dBm/Hz
Other Parameters	
Thresholds for Coordination	$\alpha_{th} = 0.1, \beta_{th} = 0.3$ (default)
Learning Rates	$\epsilon_i = \eta_i = 0.1, \forall i \in \mathcal{N}$

can be seen from Fig. 7.4(b), for most MSs a deviation from the NE strategy reduces their own capacity. Finally, as depicted in Fig. 7.4(c), there is no significant change on the average capacity when only one player unilaterally deviates from the NE strategy.

7.6.2 Capacity Performance for Different Channel Selection Strategies

Here, we compare the capacity performance of the proposed channel selection strategy with two other methods, namely, the random allocation and exhaustive search, which are described as follows:

- In the random allocation scheme, each BS randomly selects a channel for its MS in

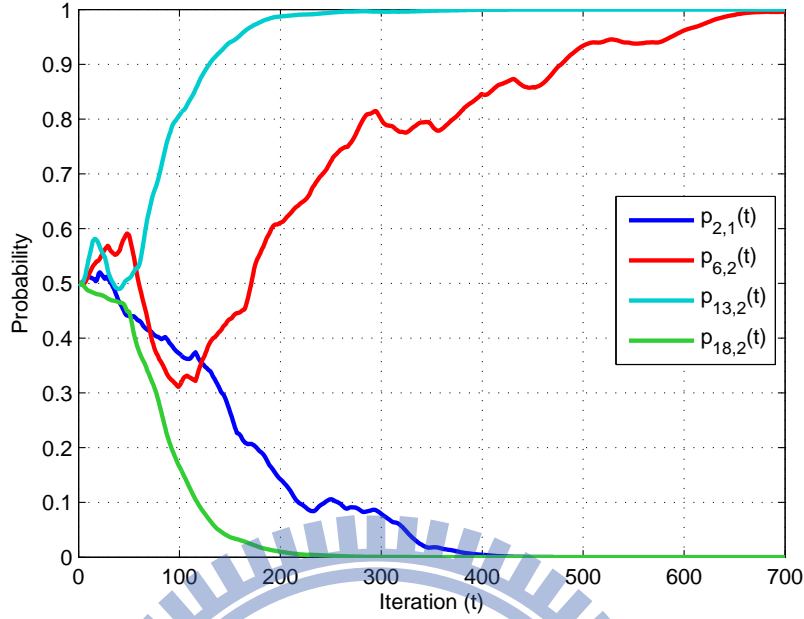


Figure 7.2: Evolution of the mixed strategies (probability of taking different actions) of four selected players when joint processing is adopted.

each frame. No learning algorithm is implemented.

- In the exhaustive search scheme, it is assumed that there exists a centralized controller which knows all system information including the channel gains, the channel availability statistics, and the number of BSs. The channel selection profile is determined by minimizing the total number of mutually interfering links, i.e.,

$$\mathbf{a}_{exh} = \operatorname{argmin}_{\mathbf{a} \in \mathcal{A}} \sum_{i=1}^N \sum_{j \in \mathcal{C}_i, j \neq i}^N \mathbb{1}_{\{a_i = a_j\}}. \quad (7.30)$$

Fig. 7.5 compares the cumulative distribution function (CDF) of the average cell capacity in each time slot for different channel selection strategies.

As can be seen, the proposed learning algorithm significantly outperforms the random selection approach and performs close to the exhaustive search approach. This demonstrates the proposed learning algorithm's ability to allocate mutually interfered players on different channels in its convergence toward the NE point.

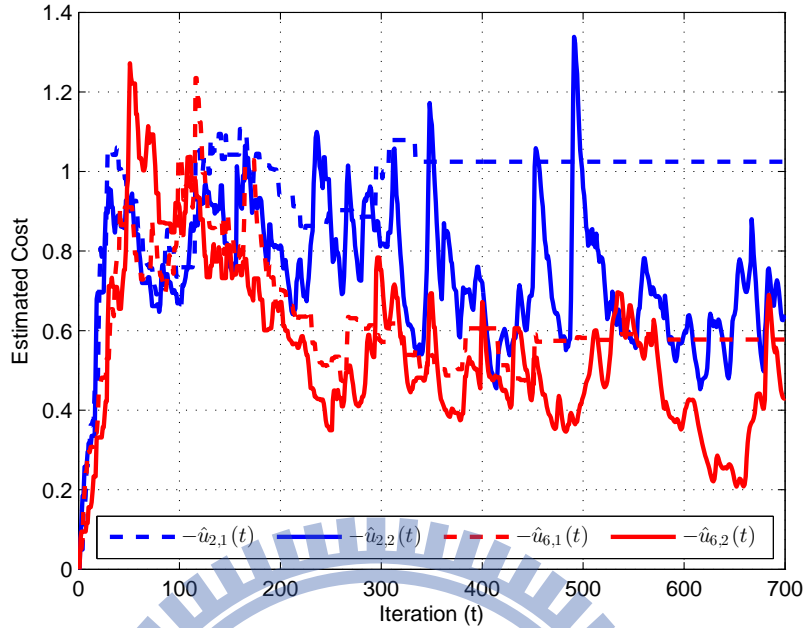


Figure 7.3: Evolution of the estimated cost of taking different actions for two selected players (marked by blue and red colors, respectively) when joint processing is adopted.

7.6.3 Capacity Performance and Fairness for Different Precoding Schemes

As mentioned in Section 7.3.2, local precoding is a special case of joint processing. Here, we investigate the impact of different precoding schemes on the performance of the proposed learning algorithm. The average per-MS capacities for different combinations of channel selection and precoding schemes are summarized in Table 7.3. For the proposed learning algorithm, it is shown that joint processing yields 10%–30% improvement over local precoding across different channel selection strategies. The results also suggest that a lower threshold β_{th} will lead to a higher average cell capacity, since when joint processing is adopted nearby cells serve the MS instead of simply mitigating its interference. Besides, we observe an increased capacity gap between the random selection and the exhaustive search when joint processing is applied. This is because in joint processing a neighboring BS becomes a serving BS, and when adjacent cells are using the same subchannel the signal for another MS becomes a strong interference source.

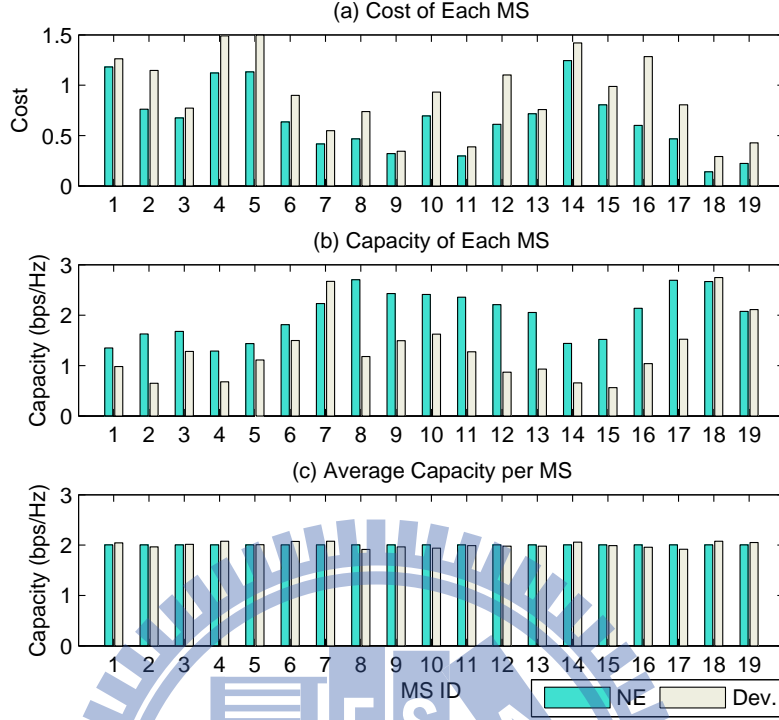


Figure 7.4: Cost and capacity for each player for the NE strategy and unilateral deviation from the NE strategy.

Table 7.3: Capacity per MS (bps/Hz) for Different Combinations of Channel Selection and Precoding Schemes

Precoding	Learning	Random	Exhaustive
Local Precoding	1.6476	1.5246	1.7006
Joint Processing, $\beta_{th} = 0.5$	1.8406	1.6924	1.8993
Joint Processing, $\beta_{th} = 0.3$	2.1052	1.8835	2.1811

In addition to the average per-MS capacity, the fairness among players is examined. Fairness of resource allocation is usually measured by the Jain's fairness index (JFI) [45] which is defined as

$$J = \frac{\left(\sum_{i=1}^N \bar{R}_i \right)^2}{N \sum_{i=1}^N \bar{R}_i^2} \quad (7.31)$$

where \bar{R}_i is the time-averaged capacity of player i over the whole simulation. The value of JFI falls in $[1/N, 1]$, and a higher JFI value represents better fairness. The JFI of the three channel selection strategies are summarized in Table 7.4. As can be seen, the

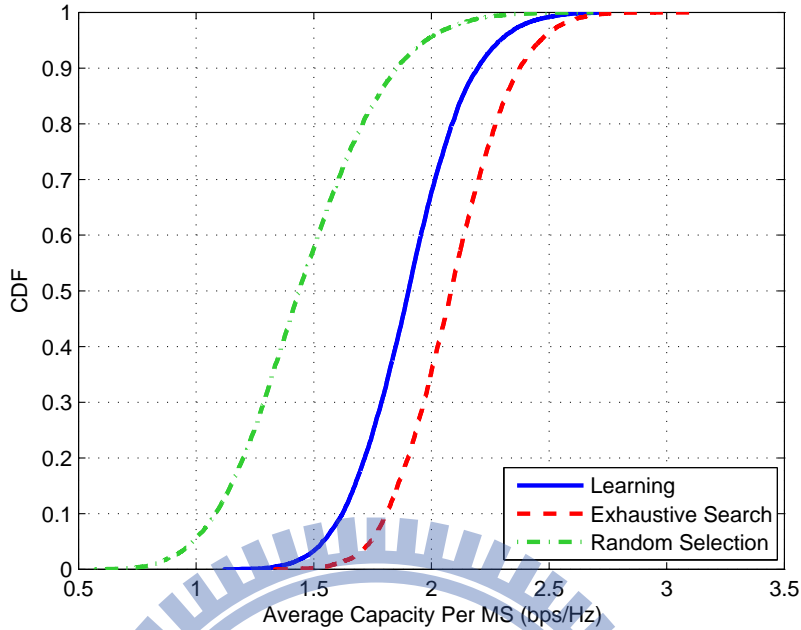


Figure 7.5: Comparison of the achievable capacity for three channel selection strategies when joint processing is adopted.

Table 7.4: JFI (7.31) for Different Combinations of Channel Selection and Precoding Schemes

Precoding	Learning	Random	Exhaustive
Local Precoding	0.8507	0.9034	0.8530
Joint Processing, $\beta_{th} = 0.5$	0.8847	0.9280	0.8809
Joint Processing, $\beta_{th} = 0.3$	0.8903	0.9371	0.9034

random selection scheme, due to its fully randomized nature, achieves the best fairness in terms of the time-averaged cell capacity while the other two channel selection strategies are also reasonably fair.

7.6.4 Performance Results for Distributed Networks with Random Geometry

The proposed learning algorithm can be implemented in any network with universal frequency reuse. Here, we consider the scenario where the transmission links are randomly placed, which reflects the typical network topology of distributed networks (e.g., cognitive

radio and femtocell networks). We generate a topology of 10 links, with the transmitters randomly distributed inside a 1 km by 1 km square area and each receiver located at a distance of 120–150 m away from its transmitter. The transmission power is set to $P_0 = 23$ dBm, with pathloss and shadowing given by the line-of-sight (LOS) urban-micro model [81]. Other simulation parameters follow those in Table 7.2. A snapshot of the network topology is shown in Fig. 7.6. Only local precoding is considered in this scenario, since joint processing requires backhaul communications among transmitters, making its implementation difficult in distributed networks.

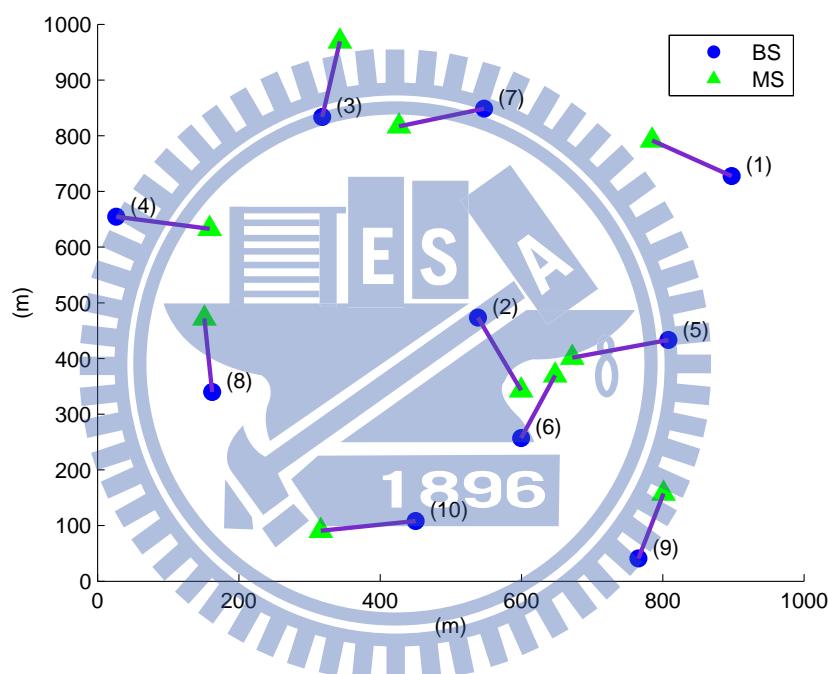


Figure 7.6: A snapshot of the nodes' positions and network topology. The link ID is shown in parenthesis next to the link.

The evolution of the mixed strategies is depicted in Fig. 7.7. The convergence toward the pure strategy is clearly observed. In addition, a comparison of different players shows that the convergence behavior is highly related to the interference condition of individual links. For relatively isolated players (e.g., link 9), it takes longer time to converge. In contrast, for players in crowded regions (e.g., links 2, 5, and 6), the convergence is generally faster but with large variation. This can be explained through the proposed reward function. Observe that in the definition in (7.15), higher interference means higher cost.

Thus, the difference between the cost of choosing channels is smaller for isolated links than for links in crowded region. The multiplicative-weights update rule makes a larger probability adjustment in each step in the latter case, resulting in a faster convergence.

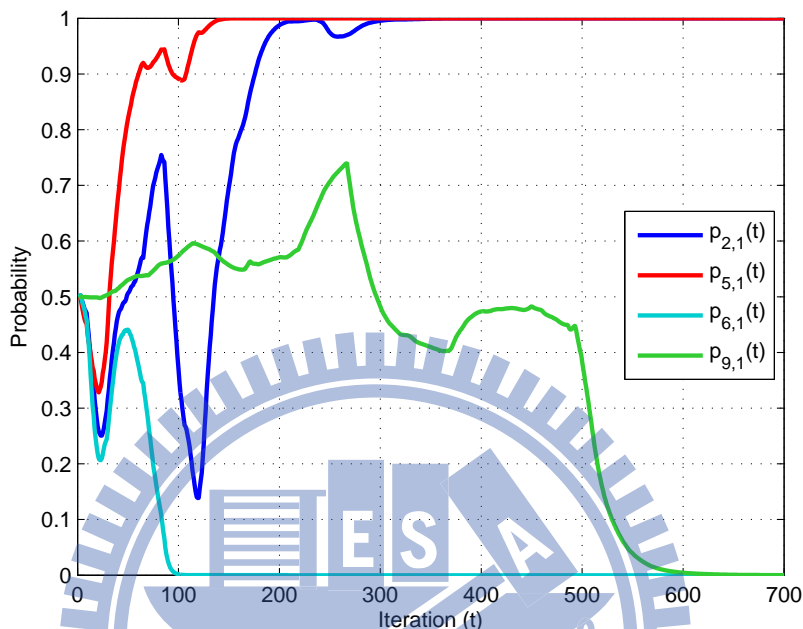


Figure 7.7: Evolution of the mixed strategies of four selected players when local precoding is applied to the distributed network.

The performance of the proposed learning algorithm is shown in Fig. 7.8. Fig. 7.8(a) compares different channel selection strategies and shows that the learning algorithm outperforms the random selection. Specifically, for highly interfered users, the proposed algorithm significantly improves the capacity compared to the random selection. The test of deviation from the NE property is conducted and the NE property is again verified in Fig. 7.8(b). The increase of cost due to unilateral deviation from NE is significant for highly interfered (crowded) players and slight for isolated players. These observations show that the proposed learning algorithm is effective in networks with random geometry for all kinds of interference conditions.

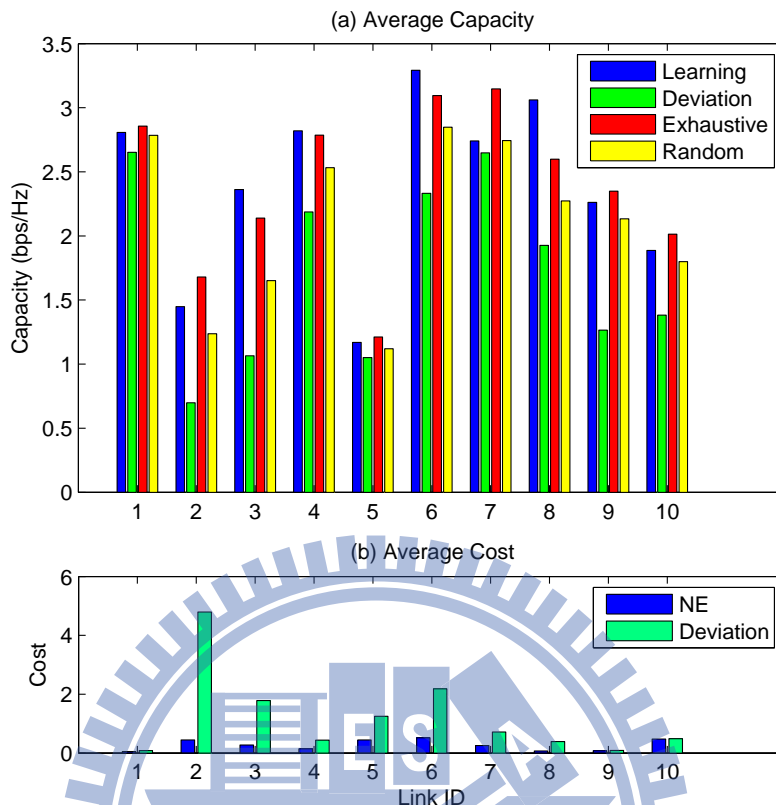


Figure 7.8: (a) Achievable capacity for different channel selection strategies. (b) Cost for each player for the NE strategy and unilateral deviation from the NE strategy.

7.7 Concluding Remarks and Open Issues

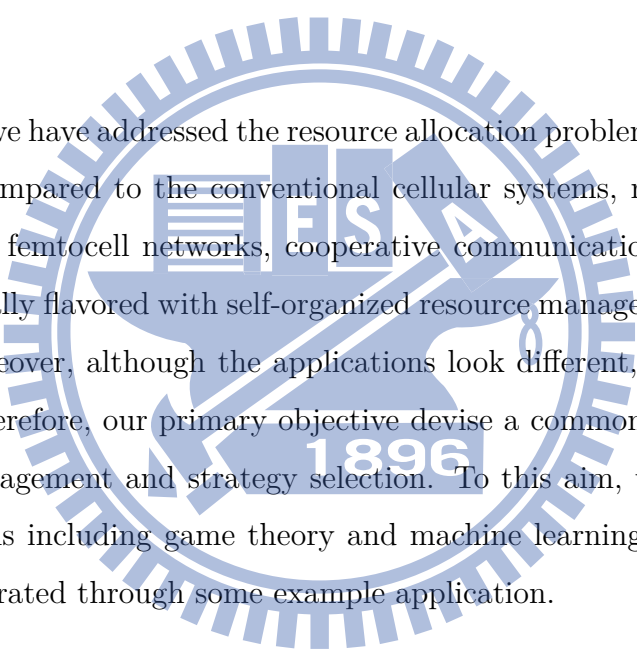
In this chapter, we have studied the problem of distributed channel selection in multicell network MIMO systems with time-varying channel and unknown number of BSs through a game-theoretic approach. We have proposed a practical joint processing scheme where each MS is jointly served by a set of nearby BSs. We have formulated the channel selection problem as a noncooperative game where the reward function was properly defined so that the game was shown to be an exact potential game. To achieve the Nash equilibrium strategy, we have proposed a stochastic learning-based decentralized algorithm by which each cell adjusts its channel selection strategy according to its individual action-reward history without knowing the actions taken by other BSs. The convergence property of the proposed algorithm in achieving a pure-strategy NE point

was theoretically proven and numerically verified for different network scenarios. The performance of the proposed algorithm in terms of the achievable capacity and fairness was also examined.



Chapter 8

Conclusion and Perspectives



In this thesis, we have addressed the resource allocation problems in decentralized wireless networks. Compared to the conventional cellular systems, recent topics in wireless networks, such as femtocell networks, cooperative communications, and cognitive radio networks, are usually flavored with self-organized resource management due to the distributed nature. Moreover, although the applications look different, they indeed bear some resemblances. Therefore, our primary objective is to devise a common guideline to decentralized resource management and strategy selection. To this aim, we have studied related mathematical tools including game theory and machine learning. Then the application has been demonstrated through some example application.

From the previously literature, we have seen that game theory, which studies the interaction of multiple players, was a commonly adopted mathematical tool. Based on game theory, each player selects the best strategy and pursuit equilibrium. However, the problem lies in the method to achieve equilibrium strategy profile. Among the proposed methods, best-response dynamics (BRD) is the most well-known one as it promises to achieve Nash equilibrium through the so-called finite improvement path (FIP). BRD suffers from an implementation complexity issue as it actually requires a coordinator to control the order of sequential strategy updates. Reinforcement learning (RL), on the other hand, is a self-organization technique by which each player treats the behaviors of

other players as the nature, and learns toward a final strategy through its own action-reward history. Coordinator is not required in RL, and most important of all, the update is simultaneous.

Before the discussion on practical applications, we presented the details of the mathematical tools. The general ideas and important definitions of game theory were reviewed first. Based on game-theoretic formulation, a math tool, named stochastic learning, has been introduced. We have proposed to use stochastic learning as the basis of self-organized resource management. Under this structure, the players in the game are assumed to be autonomous and capable of learning proper strategies based on the action-reward history. A non-cooperative game theoretical framework was used to investigate the solution to this problem, though defined quite differently in different applications. Iterative algorithms based on the best-response functions were implemented to compute the Nash equilibrium solutions. Three different applications were considered in the chapters afterwards.

The first application that we introduced in this thesis is the network selection in cognitive heterogeneous networks. In wireless networks where different RATs coexist, finding a proper network to subscribe to for each secondary user turned out to be a challenging issue, as the strategies of all users affect each other. We formulated the network selection problem as an ordinal potential game, and fully-distributed decision making is performed by individual users.

The second application considered the spectrum trading in cognitive radio networks. A two-level Stackelberg game is formulated where the service providers act as leaders and set the spectrum price first, and the secondary users, as followers, perform service selections accordingly. Learning algorithm can be implemented in both levels, with which the service providers and secondary users find their proper pricing and service selection strategies. An important extension of our work would be the study considering heterogeneous users (i.e., the users have different priority and QoS requirements) using either SLA or other mathematical tools.

As the third application, we considered the channel assignment in two-tier distributed networks. In such a scenario, even secondary users accessing the same spectrum interfere

with each other, the spectrum efficiency can be maintained as long as the interference is manipulated properly. The interference mitigation includes the cross-tier and co-tier problem. Stochastic learning was used to learn the Nash equilibrium point. An interesting point of our design is the consideration of implicit cooperation. Instead of maximizing its own throughput, our setting implicitly force each player to consider the interference toward other users. This work can be extended from two aspects: (i) multiple user, (ii) the detailed protocol.

When the mutually interfering scenario is extended to multi-antenna systems, more sophisticated designs, on both the physical and MAC layers should be involved. In the last example of the thesis, we have addressed the channel assignment problem of designing multi-user MIMO processing techniques. From the physical layer perspective, interference mitigation can be achieved by utilizing the spatial diversity. Furthermore, channel allocation was approach from a game-theoretic perspective. There are some interesting open issues, for example, we may consider the case with multiple receiving antenna.

System level simulations are applied to all applications. We have seen that as expected, NE is achieved through the proposed algorithm. Also, the SLA usually retains fairness as compared to the globally optimal solution. The biggest problem could be the acquisition of instantaneous reward. A straightforward way to maintain the distributed property in real-world implementation is to design proper protocols.

Finally, we may conclude that the proposed distributed learning method is capable of achieving NE and thus will be an important tool for the self-organized wireless systems. However, the most significant barrier is the game formulation: the convergence property is not guaranteed if the game is a potential function cannot be found. We have discovered the OPG. Although there are still steps to be taken in order to make our studies relevant from a real-world point of view, the importance of our work lies in the fact that they represent the limits of performance that can be achieved in practice.

Bibliography

- [1] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior*. Princeton university press, 1947.
- [2] J. Nash, “Non-cooperative games,” *The Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, 1951.
- [3] Z. Han, D. Niyato, W. Saad, and A. Hjørungnes, *Game theory in wireless and communication networks: theory, models, and applications*. Cambridge University Press, 2011.
- [4] K. R. Liu and B. Wang, *Cognitive radio networking and security: A game-theoretic view*. Cambridge University Press, 2010.
- [5] C. J. Watkins and P. Dayan, “Q-learning,” *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.
- [6] V. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Cambridge University Press Cambridge, 2008.
- [7] H. Tembine, *Distributed Strategic Learning for Wireless Engineers*. CRC Press, 2012.
- [8] W. Zhong, Y. Xu, M. Tao, and Y. Cai, “Game theoretic multimode precoding strategy selection for MIMO multiple access channels,” *IEEE Signal Processing Lett.*, vol. 17, no. 6, pp. 563 –566, Jun. 2010.
- [9] M. Khan, H. Tembine, and A. Vasilakos, “Game dynamics and cost of learning in heterogeneous 4G networks,” *IEEE J. Select. Areas Commun.*, vol. 30, no. 1, pp. 198 –213, Jan. 2012.
- [10] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, “Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution,” *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1380 –1391, Apr. 2012.
- [11] P. Sastry, V. Phansalkar, and M. Thathachar, “Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information,” *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 5, pp. 769 –777, May 1994.

- [12] D. Monderer and L. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, pp. 124–143, 1996.
- [13] G. Brown, "Iterative solution of games by fictitious play," *Activity analysis of production and allocation*, vol. 13, no. 1, pp. 374–376, 1951.
- [14] D. Fudenberg and D. Levine, *The Theory of Learning in Games*. MIT press, 1998, vol. 2.
- [15] I. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Commun. Mag.*, vol. 46, no. 4, pp. 40–48, Apr. 2008.
- [16] S. Huang, X. Liu, and Z. Ding, "Opportunistic spectrum access in cognitive radio networks," in *Proc. IEEE INFOCOM '08*. IEEE, 2008, pp. 1427–1435.
- [17] I. F. Akyildiz, W.-Y. Lee, and K. R. Chowdhury, "Crahn: Cognitive radio ad hoc networks," *Ad Hoc Networks*, vol. 7, no. 5, pp. 810–836, 2009.
- [18] M. Derakhshani and T. Le-Ngoc, "Learning-based opportunistic spectrum access with adaptive hopping transmission strategy," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 3957–3967, 2012.
- [19] K. W. Choi and E. Hossain, "Opportunistic access to spectrum holes between packet bursts: A learning-based approach," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 8, pp. 2497–2509, 2011.
- [20] X. Chen, J. Huang, and H. Li, "Adaptive channel recommendation for opportunistic spectrum access," vol. 12, no. 9, pp. 1788–1800, 2013.
- [21] A. Min, K.-H. Kim, J. Singh, and K. Shin, "Opportunistic spectrum access for mobile cognitive radios," in *Proc. IEEE INFOCOM, '11*, 2011, pp. 2993–3001.
- [22] D. Niyato and E. Hossain, "Competitive pricing for spectrum sharing in cognitive radio networks: Dynamic game, inefficiency of nash equilibrium, and collusion," *Selected Areas in Communications, IEEE Journal on*, vol. 26, no. 1, pp. 192–202, jan. 2008.
- [23] L. Gao, X. Wang, Y. Xu, and Q. Zhang, "Spectrum trading in cognitive radio networks: A contract-theoretic modeling approach," *IEEE J. Select. Areas Commun.*, vol. 29, no. 4, pp. 843–855, 2011.
- [24] M. N. Tehrani and M. Uysal, "Auction based spectrum trading for cognitive radio networks," *IEEE Commun. Lett.*, vol. 17, no. 6, pp. 1168–1171, 2013.
- [25] L. Yang, H. Kim, J. Zhang, M. Chiang, and C. W. Tan, "Pricing-based decentralized spectrum access control in cognitive radio networks," *IEEE/ACM Trans. Networking*, vol. 21, no. 2, pp. 522–535, 2013.

- [26] Y. Xing, R. Chandramouli, and C. Cordeiro, "Price dynamics in competitive agile spectrum access markets," *IEEE J. Select. Areas Commun.*, vol. 25, no. 3, pp. 613–621, 2007.
- [27] L. Gao, Y. Xu, and X. Wang, "MAP: Multiauctioneer progressive auction for dynamic spectrum access," vol. 10, no. 8, pp. 1144–1161, Aug. 2010.
- [28] S. H. Chun and R. La, *IEEE/ACM Trans. Networking*, no. 1, pp. 176–189.
- [29] D. Xu, X. Liu, and Z. Han, "Decentralized bargain: A two-tier market for efficient and flexible dynamic spectrum access," vol. 12, no. 9, pp. 1697–1711, Sep. 2013.
- [30] L. Qian, F. Ye, L. Gao, X. Gan, T. Chu, X. Tian, X. Wang, and M. Guizani, "Spectrum trading in cognitive radio networks: An agent-based model under demand uncertainty," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3192–3203, 2011.
- [31] L. Duan, J. Huang, and B. Shou, "Duopoly competition in dynamic spectrum leasing and pricing," vol. 11, no. 11, pp. 1706–1719, 2012.
- [32] D. Niyato, E. Hossain, and Z. Han, "Dynamics of multiple-seller and multiple-buyer spectrum trading in cognitive radio networks: A game-theoretic modeling approach," vol. 8, no. 8, pp. 1009–1022, 2009.
- [33] K. Zhu, D. Niyato, P. Wang, and Z. Han, "Dynamic spectrum leasing and service selection in spectrum secondary market of cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1136–1145, Mar. 2012.
- [34] J. Elias, F. Martignon, L. Chen, and E. Altman, "Joint operator pricing and network selection game in cognitive radio networks: Equilibrium, system dynamics and price of anarchy," *IEEE Trans. Veh. Technol.*, vol. PP, no. 99, pp. 1–1, 2013.
- [35] Z. Han, R. Zheng, and H. Poor, "Repeated auctions with bayesian nonparametric learning for spectrum access in cognitive radio networks," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 3, pp. 890–900, 2011.
- [36] W. H. Sandholm, *Population Games and Evolutionary Dynamics*. MIT press Cambridge, 2010, vol. 88.
- [37] J. Mitola III and G. Q. Maguire Jr, "Cognitive radio: making software radios more personal," *IEEE Personal Commun. Mag.*, vol. 6, no. 4, pp. 13–18, 1999.
- [38] D. Fudenberg and J. Tirole, *Game Theory*. MIT Press, 1991.
- [39] B. Wang, Y. Wu, and K. Liu, "Game theory for cognitive radio networks: An overview," *Computer Netw.*, vol. 54, no. 14, pp. 2537–2561, 2010.
- [40] R. Trestian, O. Ormond, and G.-M. Muntean, "Game theory-based network selection: Solutions and challenges," *IEEE Trans. Contr. Syst. Technol.*, vol. 14, no. 4, pp. 1212–1231, 2012.

- [41] L.-C. Tseng, F.-T. Chien, D. Zhang, R. Y. Chang, W.-H. Chung, and C.-Y. Huang, "Network selection in cognitive heterogeneous networks using stochastic learning."
- [42] T. Roughgarden and E. Tardos, "Introduction to the inefficiency of equilibria," *Algorithmic Game Theory*.
- [43] J.-Y. Le Boudec, "Rate adaptation, congestion control and fairness: A tutorial," Nov. 2012.
- [44] J. G. Wardrop, "Some theoretical aspects of road traffic research." in *ICE Proceedings: Engineering Divisions*, vol. 1, no. 3. Thomas Telford, 1952, pp. 325–362.
- [45] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," *DEC Research Report TR-301*, 1984.
- [46] K. Sundaresan and S. Rangarajan, "Efficient resource management in OFDMA femtocells," in *Proc. ACM Mobihoc '09*. ACM, 2009, pp. 33–42.
- [47] 3GPP, "E-UTRA: Further advancements for E-UTRA physical layer aspects," 3GPP Technical report (TR 36.814) v9.0.0, March, 2010.
- [48] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Commun. Mag.*, vol. 40, no. 8, pp. 102–114, 2002.
- [49] J. Ng, "Ubiquitous healthcare: healthcare systems and application enabled by mobile and wireless technologies," *Journal of Convergence*, vol. 3, no. 2, pp. 31–36, 2012.
- [50] S. Silas, K. Ezra, and E. Blessing Rajsingh, "A novel fault tolerant service selection framework for pervasive computing," *Human-centric Computing and Information Sciences*, vol. 2, no. 1, pp. 1–14, 2012. [Online]. Available: <http://dx.doi.org/10.1186/2192-1962-2-5>
- [51] G. H. Carvalho, A. Anpalagan, I. Woungang, and S. K. Dhurandher, "Energy-efficient radio resource management scheme for heterogeneous wireless networks: a queueing theory perspective," *Energy*, vol. 3, no. 4, 2012.
- [52] O. Akan, O. Karli, and O. Ergul, "Cognitive radio sensor networks," *Network, IEEE*, vol. 23, no. 4, pp. 34–40, 2009.
- [53] D. Lopez-Perez, A. Valcarce, G. de la Roche, and J. Zhang, "OFDMA femtocells: A roadmap on interference avoidance," *IEEE Commun. Mag.*, vol. 47, no. 9, pp. 41–48, september 2009.
- [54] A. Hatoum, N. Aitsaadi, R. Langar, R. Boutaba, and G. Pujolle, "FCRA: Femto-cell cluster-based resource allocation scheme for OFDMA networks," in *Proc. IEEE ICC'11*, Jun. 2011, pp. 1–6.

- [55] M. Bennis and D. Niyato, "A Q-learning based approach to interference avoidance in self-organized femtocell networks," in *Proc. IEEE GLOBECOM Workshops '10*. IEEE, 2010, pp. 706–710.
- [56] H. Li, "Multiagent Q-learning for aloha-like spectrum access in cognitive radio systems," *EURASIP J. Wireless Commun. Netw.*, 2010.
- [57] A. Galindo-Serrano and L. Giupponi, "Femtocell systems with self organization capabilities," in *Proc. NetGCooP '11*, Oct. 2011, pp. 1–7.
- [58] S. Hart and A. Mas-Colell, "A reinforcement procedure leading to correlated equilibrium," *Economic Essays*, pp. 181–200, 2001.
- [59] J. Huang and V. Krishnamurthy, "Cognitive base stations in lte/3gpp femtocells: A correlated equilibrium game-theoretic approach," *Communications, IEEE Transactions on*, vol. 59, no. 12, pp. 3485–3493, december 2011.
- [60] M. Bennis, S. M. Perlaza, and M. Debbah, "Learning coarse correlated equilibria in two-tier wireless networks," in *Proc. IEEE ICC '12*, Jun. 2012, pp. 1592–1596.
- [61] B. Singh and D. Lobiyal, "A novel energy-aware cluster head selection based on particle swarm optimization for wireless sensor networks," *Human-centric Computing and Information Sciences*, vol. 2, no. 1, pp. 1–18, 2012. [Online]. Available: <http://dx.doi.org/10.1186/2192-1962-2-13>
- [62] X. Li, N. Mitton, A. Nayak, and I. Stojmenovic, "Achieving load awareness in position-based wireless ad hoc routing," *Journal of Convergence*, vol. 3, no. 3, 2012.
- [63] N. Nie and C. Comaniciu, "Adaptive channel allocation spectrum etiquette for cognitive radio networks," in *Proc. IEEE DySPAN '05*, Nov. 2005, pp. 269–278.
- [64] Q. D. La, Y. H. Chew, and B. H. Soong, "Performance analysis of downlink multi-cell OFDMA systems based on potential game," vol. 11, no. 9, pp. 3358–3367, 2012.
- [65] H. Zhang, H. Dai, and Q. Zhou, "Base station cooperation for multiuser MIMO: Joint transmission and BS selection," in *Proc. IEEE CISS '04*, 2004.
- [66] R. Y. Chang, Z. Tao, J. Zhang, and C.-C. J. Kuo, "Multicell OFDMA downlink resource allocation using a graphic framework," *IEEE Trans. Veh. Technol.*, vol. 58, no. 7, pp. 3494–3507, Sep. 2009.
- [67] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.
- [68] G. Caire, S. A. Ramprashad, and H. C. Papadopoulos, "Rethinking network MIMO: Cost of CSIT, performance analysis, and architecture comparisons," in *Proc. ITA '10*, 2010, pp. 1–10.

- [69] Y. Hadisusanto, L. Thiele, and V. Jungnickel, “Distributed base station cooperation via block-diagonalization and dual-decomposition,” in *Proc. IEEE GLOBECOM '08*, 2008, pp. 1–5.
- [70] J. Zhang, R. Chen, J. Andrews, A. Ghosh, and R. Heath, “Networked MIMO with clustered linear precoding,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, 2009.
- [71] S. Kaviani, O. Simeone, W. Krzymien, and S. Shamai, “Linear MMSE precoding and equalization for network MIMO with partial cooperation,” in *Proc. IEEE GLOBECOM '11*, Dec. 2011, pp. 1–6.
- [72] P. de Kerret and D. Gesbert, “Sparse precoding in multicell MIMO systems,” in *Proc. IEEE WCNC '12*, Apr. 2012, pp. 958–962.
- [73] —, “The multiplexing gain of a two-cell MIMO channel with unequal CSI,” in *Proc. IEEE ISIT '11*, Aug. 2011, pp. 558–562.
- [74] R. Zakhour and D. Gesbert, “A two-stage approach to feedback design in multi-user MIMO channels with limited channel state information,” in *Proc. IEEE PIMRC '07*, pp. 1–5.
- [75] R. Zakhour, Z. Ho, and D. Gesbert, “Distributed beamforming coordination in multicell MIMO channels,” in *Proc. IEEE VTC Spring '09*, Apr. 2009, pp. 1–5.
- [76] E. Bjornson, N. Jalden, M. Bengtsson, and B. Ottersten, “Optimality properties, distributed strategies, and measurement-based evaluation of coordinated multicell OFDMA transmission,” *IEEE Trans. Signal Processing*, vol. 59, no. 12, pp. 6086–6101, Dec. 2011.
- [77] M. Bloem, T. Alpcan, and T. Başar, “A stackelberg game for power control and channel allocation in cognitive radio networks,” in *Proc. ICST VALUETOOLS '07*, 2007, p. 4.
- [78] H. Tembine, “Dynamic robust games in MIMO systems,” *IEEE Trans. Syst., Man, Cybern. B*, vol. 41, no. 4, pp. 990–1002, Aug. 2011.
- [79] A. Goldsmith and S.-G. Chua, “Variable-rate variable-power MQAM for fading channels,” *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.
- [80] M. Sadek, A. Tarighat, and A. Sayed, “A leakage-based precoding scheme for downlink multi-user MIMO channels,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1711–1721, May 2007.
- [81] 3GPP, “Spatial channel model for multiple input multiple output (MIMO) simulations (release 10),” 3GPP Technical report (TR 25.996) v10.0.0, Mar. 2011.