# Reinforcement Learning-based Spectrum Sharing for Cognitive Radio

Tao Jiang

Doctor of Philosophy (Ph.D.)

University of York

Department of Electronics

September 2011

# Abstract

This thesis investigates how distributed reinforcement learning-based resource assignment algorithms can be used to improve the performance of a cognitive radio system. Decision making in most wireless systems today, including most cognitive radio systems in development, depends purely on instantaneous measurement. The purpose of this work is to exploit the historical information the cognitive radio device has learned through the interactions with the unknown environment. Two system architectures have been investigated in this thesis. A point-to-point architecture is examined first in an open spectrum scenario. Then, for the first time distributed reinforcement learning-based algorithms are developed and examined in a novel two-hop architecture for Beyond Next Generation Mobile Network.

The traditional reinforcement learning model is modified in order to be applied to a fully distributed cognitive radio scenario. The inherent exploration versus exploitation trade-off seen in reinforcement learning is examined in the context of cognitive radio. A two-stage algorithm is proposed to effectively control the exploration phase of the learning process. This is because cognitive radio users will cause a higher level of disturbance in the exploration phase. Efficient exploration algorithms like pre-partitioning and weight-driven exploration are proposed to enable more efficient learning process. The learning efficiency in a cognitive radio scenario is defined and the learning efficiency of the proposed schemes is investigated. Results show that the performance of the cognitive radio system can be significantly enhanced by utilizing distributed reinforcement learning since the cognitive devices are able to identify the appropriate resources more efficiently.

The reinforcement learning-based ‘green’ cognitive radio approach is discussed. Techniques presented show how it is possible to largely eliminate the need for spectrum sensing, along with the associated energy consumption, by using reinforcement learning to develop a preferred channel set in each device.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

I would like to express my sincere gratitude to my first supervisor Dr David Grace, for his constant encouragement and support. Without his inspiration and guidance, this work could not have been possible.

I am also grateful for the enormous support and valuable suggestions from my second supervisor, Dr Paul Mitchell.

During the past few years, I have worked with a group of nice colleagues. Many thanks to all the members of the Communications Research Group for creating a friendly and helpful environment. Especially to Dr Yiming Liu who has helped me a lot when I firstly started working on my PhD project.

Last but not least, I wish to express my deepest gratitude to my parents and my wife, for their selfless love and support. Their dedication provided the foundation of this work.

# Declaration

Some of the research presented in this thesis have resulted in publications in journals, conference proceedings and EU research project deliverables. A list of the publications is provided below and at the end of the thesis.

All contributions presented in this thesis as original are as such to be the best knowledge of the author. References and acknowledges to other researchers have been given as appropriate.

List of Publications:

*Book Chapter*

T. Jiang, D. Grace,: 'Reinforcement Learning-based Cognitive Radio for Open Spectrum Access', ***Cognitive Communications: Distributed Artificial Intelligence (DAI), Regulatory Policy & Economics, Implementation****,* Wiley, 2011. (accepted)

*Journal*

T. Jiang, D. Grace, and P.D. Mitchell,: 'Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing', ***IET Communications.,*** Volume 5, Issue 10, pp.1309-1317, July 2011, DOI:10.1049/iet-com.2010.0258

T. Jiang, D. Grace, and Y. Liu,: 'Two-stage reinforcement-learning-based cognitive radio with exploration control', ***IET Communications.,*** Volume 5, Issue 5, pp.644-651, March 2011, DOI:10.1049/iet-com.2009.0803

X. Chen, Z. Zhao, T. Jiang, D. Grace, and H. Zhang,: 'Inter-cluster connection in cognitive wireless mesh networks based on intelligent network coding' ***EURASIP Journal on Advances in Signal Processing - Special issue on dynamic spectrum access for wireless networking***, March 2009, DOI=10.1155/2009/141097

*Conference*

T. Jiang, D. Grace, and P.D. Mitchell,: ‘Improvement of Pre-partitioning on Reinforcement Learning Based Spectrum Sharing’, ***IET International Communication Conference on Wireless Mobile and Computing (CCWMC)***, pp.299-302, 2009 (**Best paper award**)

D. Grace, J. Chen, T. Jiang, and P.D. Mitchell,: ‘Using Cognitive Radio to Deliver 'Green' Communications’, ***4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)***, pp.1-6, 2009

X. Chen, Z. Zhao, H. Zhang, T. Jiang, and D. Grace,: ‘Inter-Cluster Connection in Cognitive Wireless Mesh Networks Based on Intelligent Network Coding’, ***IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)***, pp.1251-1256, 2009

T. Jiang, D. Grace, and Y. Liu,: ‘Cognitive radio spectrum sharing schemes with reduced spectrum sensing requirements’, ***IET Seminar on Cognitive Radio and Software Defined Radios: Technologies and Techniques***, September, 2008, London

T. Jiang, D. Grace, and Y. Liu,: ‘Performance of Cognitive Radio Reinforcement Spectrum Sharing Using Different Weighting Factors’, ***Third International Conference on Communications and Networking in China (ChinaCom)***, pp.1195-1199, 2008

*FP7 BuNGee Project Deliverables*

**ICT-BuNGee D3.1**, Y. Leiba, et al., ‘Baseline RRM & JointAccess/Self-Backhaul Protocols’, May 2011, Available: http://www.ict-bungee.eu/

**ICT-BuNGee D4.1.1**, T. Jiang, A. Papadogiannis, D. Grace, A. Burr,: 'Interim Simulation', February 2011, Available: http://www.ict-bungee.eu/

**ICT-BuNGee D1.2**, M. Goldhamer, et al., 'Baseline BuNGee Architecture', November 2010, Available: http://www.ict-bungee.eu/

# Chapter 1. Introduction

**Contents**

## *1.1 Overview*

Efficient utilization of the physical radio spectrum is a fundamental issue of wireless communications. The conventional licensed frequency allocations are overly inflexible, resulting in both spatially and temporally inefficient usage of radio spectrum. According to Federal Communications Commission (FCC), 15% to 85% of the allocated spectrum is utilized with large temporal and geographical variations [1-2]. Meanwhile, the demands for wireless communication have increased significantly in both the number of users and the required quality of wireless transmission [3]. The conflict between the inefficient usage of spectrum and the rapid growth of wireless services calls for a more flexible and intelligent solution to manage such an important natural resource. Cognitive Radio (CR), a new paradigm of wireless communication, has been considered as a potential way to accomplish such an important task [2, 4-7]. By combining the abilities of spectrum awareness, intelligence and radio flexibility, a cognitive radio will be able to adapt itself to the changes in the local environment [8]. It is foreseen that a large amount of underutilized spectrum will be efficiently used by applying cognitive radio techniques.

Early stage channel assignment schemes are largely based on Fixed Channel Assignment (FCA) [9-11]. FCA requires frequency plans to take place in order to limit the interference. The service area is divided into a number of cells and a subset of channels are assigned to each cell. Depending on the requirements of the quality of service of the network, frequency reuse patterns are developed such that the same subset of channels are reused at a safe distance [10]. FCA schemes are more efficient in terms of handling uniform traffic. However, FCA schemes are not traffic adaptive. When it comes to fluctuating traffic, even though there are channels available in the neighbouring cells, calls are blocked due to insufficient channels in the current cells [9].

Dynamic Channel Assignment (DCA) was developed as a better solution to serve fluctuating traffic. In a DCA scheme, instead of having a fixed frequency plan, all channels are placed into a channel pool and potentially available to all the local users [12-13]. The channels are then assigned on a call-by-call basis. Research shows that DCA schemes achieve better performance when handling uneven traffic which varies both spatially and temporally [9]. DCA schemes can be divided into two categories: Centralized Dynamic Channel Assignment (CDCA) and Distributed Dynamic Channel Assignment (DDCA).

In the CDCA schemes, a centralized controller assigns channels from the channel pool to the calls. Extensive information needs to be exchanged between the base station and the central controller, resulting in a large number of control overhead [14-17]. DDCA schemes utilize localized information to select the suitable channel without any communication with other base station (user) [18-19]. DDCA schemes normally rely on interference or Carrier-to-Interference Ratio measurements. DDCA schemes remove the control overhead required by the CDCA schemes, but the behaviour of DDCA schemes is likely to be more selfish than other approaches so that new activations may introduce excessive interference to the existing transition links [20]. Cognitive radio, a DDCA based technique, is likely to avoid causing such problem by utilizing more advanced spectrum sensing techniques and new functions like learning. One of the main differences between cognitive radio and conventional DDCA techniques is that unlike previous DDCA techniques which were designed for all users in a dedicated band, cognitive radio is proposed to use the spectrum licensed to other systems (the primary network) [4].

## *1.2 Purpose*

The purpose of this thesis is to explore how the historical information of the wireless system can be utilized through reinforcement learning to improve the system performance. Previous DDCA schemes depended only on instantaneous measurement with the past experience being wasted. By exploiting learning techniques, such information can be used to facilitate the transmissions between entities.

Although learning has normally been considered as an essential part of cognitive radio, no clear understanding on when, where and how learning could be applied to a cognitive radio system has been reached, especially in a fully distributed scenario [4-6, 21]. This work concentrates on how to apply reinforcement learning-based techniques to the channel assignment of cognitive radios and tackles the problems found in the process of applying reinforcement learning to cognitive radio.

## *1.3 Communication Architectures*

This thesis examines distributed cognitive radio spectrum sharing techniques for two architectures: a point-to-point architecture and a dual-hop beyond next generation mobile network architecture [22]. An open spectrum scenario is considered where all users are given equal priority to use the spectrum – a cognitive only band where the users are purely cognitive radios [23-26]. It is worth investigating the system performance in a cognitive-only band since it is likely in the future that devices in such (‘unlicensed’) bands will become increasingly cognitive, enabling them to deal with interference and reconfiguration, allowing new more efficient techniques and solutions to be developed.

- *Point-to-point*

A basic transmitter-receiver pair communication system is used as illustrated in figure 1.1 because we try to focus on the complex and autonomous behaviour of cognitive radio users who constantly change their action policy according to the experience gained through learning. We believe the technique is widely applicable to other system models. A certain number of transmitter-receiver pairs are randomly distributed in a service area and the locations of pairs are fixed. The transmission range and *SINR* exclusion area [27] are all shown in this figure. Omnidirectional antennas are applied at all transmitters and receivers. The pairs are fully distributed, meaning that no information is directly exchanged with other pairs.

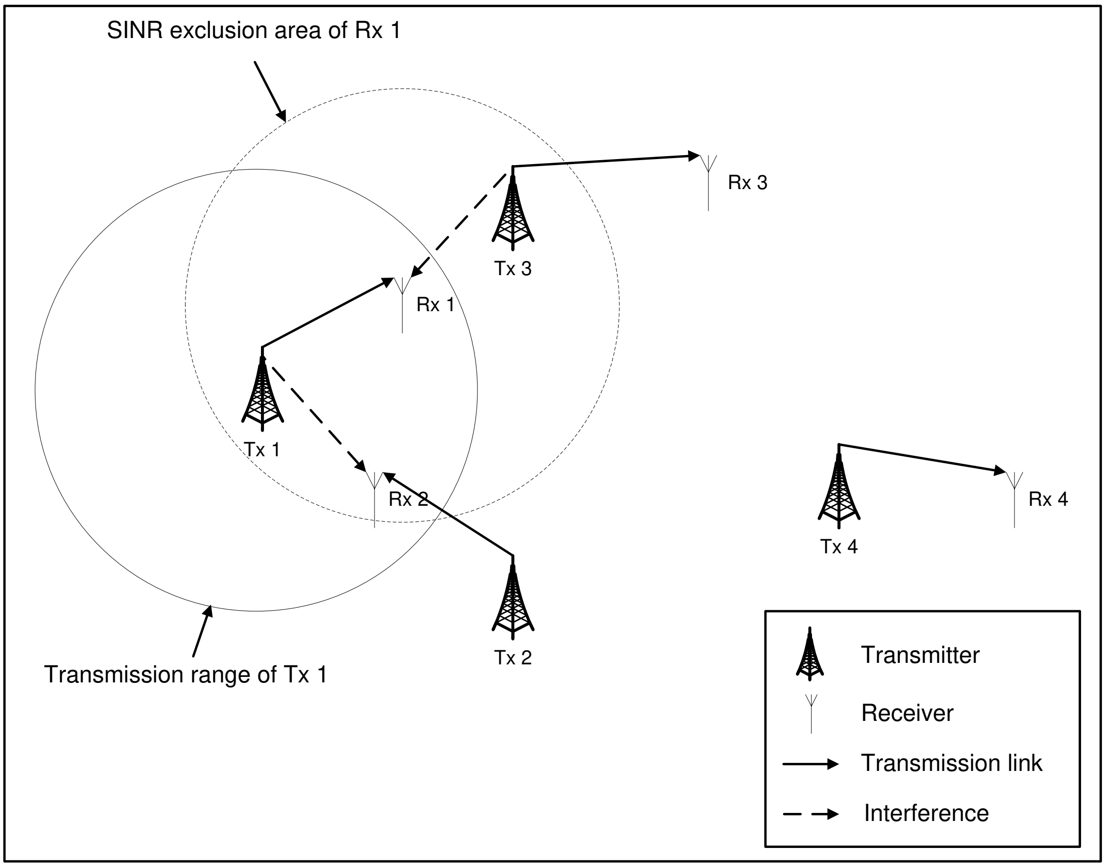


*Figure 1.1 Point-to-Point Architecture*

- *Dual-hop beyond next generation mobile network architecture*

In order to provide sufficient capacity density to the dense city centre area, a dual-hop architecture has been proposed by the FP7 Beyond Next Generation Mobile Broadband Project [22]. This novel dual-hop architecture is used later in this thesis. Reinforcement learning-based cognitive radio approaches are developed for the beyond next generation mobile network. The dual-hop architecture is shown in figure 1.2 [22].

The first tier of the system is the self-backhaul network, where the Hub Base Station (HBS) is mounted over roof-tops. Access Base Stations (ABS) are connected with HBS via a Hub Subscriber Station (HSS) antenna. HBS serves the data streams wirelessly from or to a large number of low-cost ABSs. The second tier of the system is the access network, where the ABSs are placed under roof-top along the streets and these ABSs provide access to the Mobile Subscribes (MS).

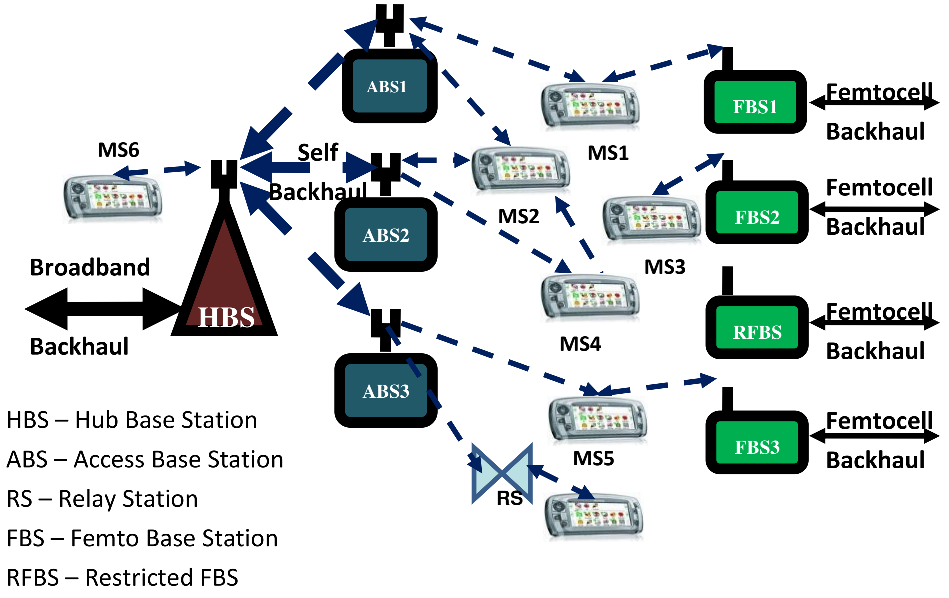


*Figure 1.2 Beyond Next Generation Mobile Network Architecture (directly reproduced from [22])*

The point-to-point architecture is applied in chapter 4, chapter 5, chapter 6, and chapter 7 in order to gain a deep understanding on the behaviour of learning-based users. The much more complex dual-hop architecture is then used in chapter 8 to investigate further the impact of reinforcement learning on beyond next generation mobile systems. The research work obtained by applying these two types of architectures will demonstrate the applicability of the learning-based techniques developed in this work.

## *1.4 Thesis Outline*

The rest of this thesis is outlined as follows:

Chapter 2 provides the background information of this work. A literature review that summarises papers on channel assignment techniques is given first and then comprehensive information of a cognitive radio system is given. After that the information related to reinforcement learning is provided. The intelligent channel assignment techniques are reviewed finally to introduce the state-of-art approaches relevant to this work.

The system modelling methodology, simulation techniques, the key measurements for evaluating the system performance and the verification strategy are introduced in chapter 3. Simulation is used extensively in this work since the behaviour of the learning-based users is too complex to be fully analysed mathematically.

Chapter 4 introduces the generic reinforcement learning model and the value function we developed for cognitive radio. The learning model and the value function are the basis of this work. The performance of the learning based approaches is discussed and the influence of the weighting factors is also discussed in this chapter.

The trade-off between exploration and exploitation seen in reinforcement learning is investigated in the context of cognitive radio in chapter 5. Discussions on how this trade-off could practically influence the learning-based cognitive radios are given. A two-stage algorithm is developed to control the exploration phase, and how this two-stage algorithm is able to improve the system performance is also examined.

The two-stage algorithm introduced in chapter 5 is then used as the basis to develop efficient exploration techniques in chapter 6. The exploration phase can only be limited rather than completely eliminated from the learning process. Thus, it is desirable that more efficient exploration techniques to be applied in line with the exploration control algorithm developed in chapter 5. Two efficient exploration techniques are introduced: Pre-partitioning and Weight-driven exploration. The pre-partitioning scheme randomly reserves a certain amount of spectrum resources for each user. The available action space which the cognitive radio needs to explore is then significantly reduced, which in turn shortens the exploration stage significantly. In the weight-driven exploration scheme, the exploitation phase is gradually moved into exploration by applying a weight-driven probability distribution to influence action selection during exploration. The performance of these two approaches is compared with a commonly used uniform random approach in this chapter.

Chapter 7 explores the ‘green’ aspect of the proposed learning based schemes, concentrating on the power consumption reduction achieved by learning. This is done to reduce the requirement for spectrum sharing through reinforcement learning. The energy consumption of the schemes introduced in chapters 4 - 6 are compared and discussed.

Chapter 8 explores the possibility of applying reinforcement-based cognitive radio techniques to the novel dual-hop beyond next generation mobile network architecture. The system model and the propagation environment are very complex since the system is designed for dense city centre areas where a large number of building blocks can be found, and several types of directional antenna are used along with advanced MIMO techniques.

A very detailed simulator is developed in this chapter to model the wireless system and its surrounding environment. Distributed reinforcement learning-based channel assignment techniques are developed for the first time for such system. A single learning engine is designed to process the information for both hops of the wireless link simultaneously. Performance of the developed schemes is discussed in also chapter 8.

In chapter 9 the ideas of taking the research work in this thesis forward are discussed. The main conclusions and the novel contributions of this work are summarized finally in chapter 10.

# Chapter 2. Literature Review

**Contents**

## 2.1 Introduction

The purpose of this chapter is to provide the essential concepts and the background information related to this thesis. The concepts of Cognitive Radio and Reinforcement Learning are introduced first in section 2.2 and 2.3 respectively. A brief review on Distributed Dynamic Channel Assignment Techniques is provided in section 2.4. After that, a comprehensive literature review on the state-of-art ‘intelligent’ channel assignment techniques is given in section 2.5. The cognitive radio related projects around the world are also reviewed in section 2.6. Finally, conclusions are given in section 2.7. The information provided in this chapter is essential in terms of understanding the techniques introduced later in this thesis.

## 2.2 Cognitive Radio

The assignment of spectrum to transmissions and to users is a fundamental issue of wireless communications. Numerous channel assignment methods have been proposed for sharing the limited physical resource. The traditional licensed spectrum allocation strategies employed by radio regulatory bodies is very restrictive and extremely inflexible, resulting in highly underutilized spectrum usage. Figure 2.1 is an example of the spectrum usage in a few places in the UK

([28]). The temporal and geographical variations of the spectrum usage can be clearly seen.

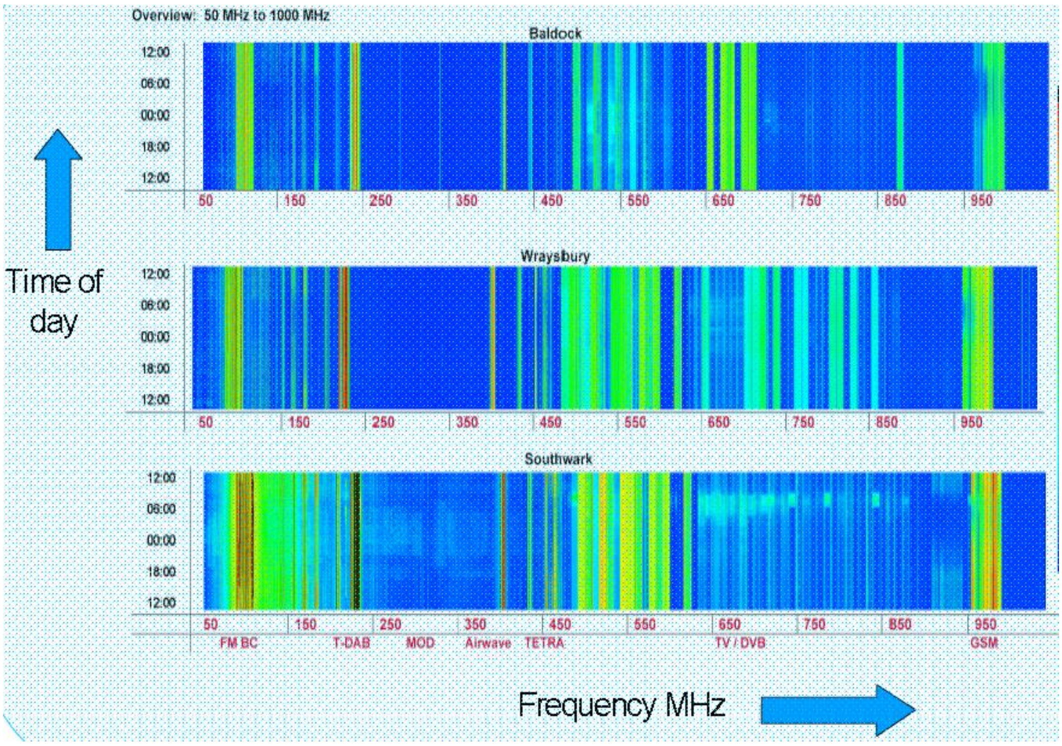


*Figure 2.1 Spectrum Occupancy Measurements in a Rural Area (top), near Heathrow Airport (middle) and in Central London (bottom) (directly reproduced from [28])*

The colours in figure 2.1 represent the level of channel usage, from blue (unused frequency) to red (heavily used frequency). It can be seen that even in central London, the amount of the heavily used frequency bands is still relatively small. The frequency bands in figure 2.1 are largely unoccupied regardless time and location.

A fully dynamic spectrum access technique called Cognitive Radio which was first introduced in [4, 7], has been considered as a potential way to improve the inefficient spectrum utilization. The inefficient usage of the existing spectrum can be improved through opportunistic access to the licensed bands without interfering with the existing users. The definition of cognitive radio suggested by ITU-R [29] is: 'a *radio system employing a technology, which makes it possible to obtain knowledge of its operational environment, policies and internal state, to dynamically adjust its parameters and protocols according to the knowledge* *obtained and to learn from the results obtained*'. The fundamental objective of cognitive radio is to enable an efficient utilization of the wireless spectrum through a highly reliable approach.

An important concept is the definition of spectrum hole. It is defined as: '*a spectrum hole is a band of frequencies assigned to a primary user (licensed user), but, at a particular time and specific geographic location, the band is not being utilized by that user*' [2]. The efficient use of the spectrum will be promoted by exploiting the spectrum holes. If the spectrum hole is requested by primary user, the cognitive user will move to another spectrum hole or stay in the same band, changing its transmit parameters to avoid interference. This process is illustrated in figure 2.1 (reproduced from [2]).

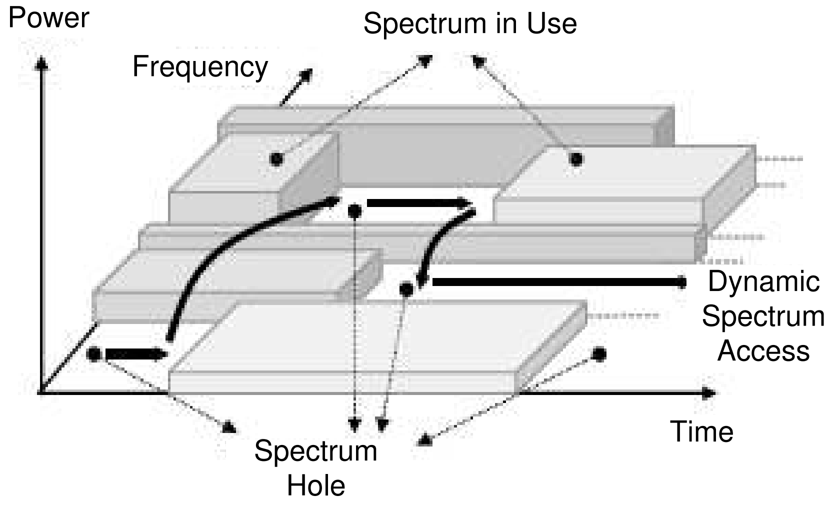


*Figure 2.2 Example of the Utilization of Spectrum holes (directly reproduced from [2])*

Based on the definition of cognitive radio, two main elements can be outlined: the cognition part and the reconfigurability. By combining these two functions together, cognitive radios are able to access the spectrum in a fully dynamic way.

### Cognition

Cognitive capability is the most distinguishing feature of cognitive radio when compared with DDCA, although many schemes like IEEE 802.22 lack any form of intelligence, so could be alternatively considered as performing DDCA [30]. The cognition aspect helps capture the variations of the radio environment over a period of time or space [5]. Spectrum awareness provides the opportunity to fundamentally change the way we manage the radio spectrum. Through this capability, the spectrum holes will be identified and therefore the available spectrum and the appropriate transmitting parameters can be selected. The cognitive cycle which is also the task required for cognitive operation is shown in figure 2.3 (reproduced from [2]). We can see three main elements in this figure: spectrum sensing, spectrum analysis and spectrum decision. These functions are the basis of the on-line interaction between cognitive radio and the unpredictable environment. The details of the functions are as follows [2, 6]:

- *Spectrum sensing: Cognitive radio scans the available spectrum, estimating the interference level of it.*

- *Spectrum analysis: Based on the information provided by spectrum sensing, cognitive radio will estimate the channel state and the channel capacity.*

- *Spectrum decision: The decision-making part is the main research area in this thesis. According to the previous information provided by spectrum sensing and spectrum analysis, cognitive radio needs to determine not only which available channel to use but also the transmission parameters, e.g. the transmission mode, the data rate and transmission power etc [5].*

After the above 3 steps, cognitive radio will have enough information to adjust its operating parameters to perform the communication. The cognition part is the intelligence intensive part of cognitive radio where different intelligent techniques are applied, including reasoning and learning. The decisions made by individual users will change the environment and other users will adapt themselves to these changes by going through the 3 steps repeatedly.

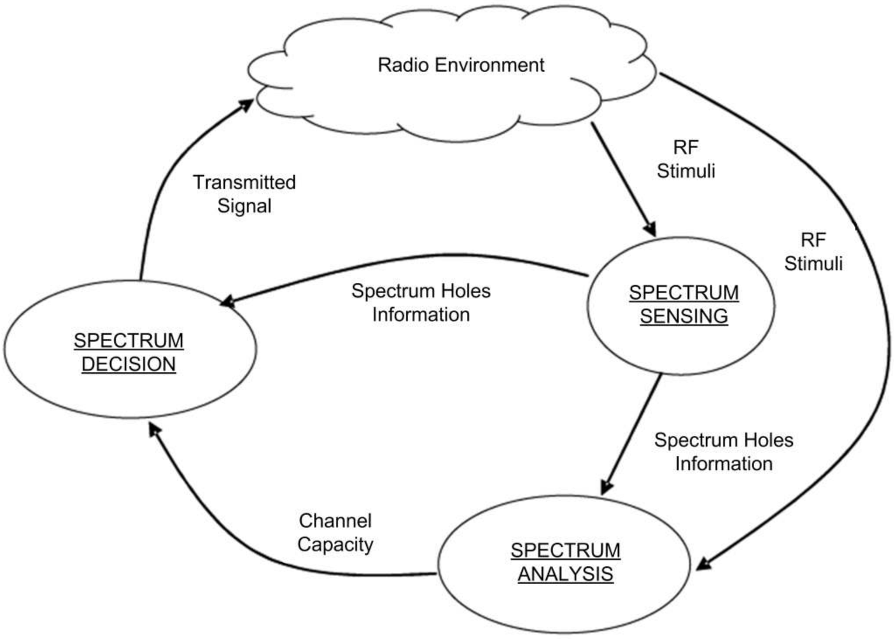


*Figure 2.3 Basic Cognitive Cycle (directly reproduced from [2])*

### Reconfiguration

Another important feature of cognitive radio is the capability of adaption [4, 8]. Cognitive radio will adapt its internal states to the variations of the wireless environment by adjusting certain operating parameters. There are a few basic operating parameters that can be reconfigured by cognitive radio:

- *Carrier frequency: The capability of adjusting the carrier frequency is a fundamental function of cognitive radio. If the current spectrum hole in use is no longer suitable, the cognitive radio needs to move to the most appropriate frequency band according to the spectrum decision made by it.*

- *Transmission power: Dynamic transmission power control can also be performed in cognitive radio scenario. The appropriate transmission* *power level will be applied to decrease the interference and allow more users sharing the same spectrum.*

- *Modulation: The modulation scheme is also reconfigurable. By realizing the characteristics of the targeting spectrum and the environment, cognitive radio is able to select the most suitable modulation to perform the communication.*

Cognitive radio will operate in a very complex heterogeneous scenario. The online adaptation of the operating parameters provides the basis for cognitive radio to dynamically interact with the environment. By dynamically exploiting the spectrum holes, cognitive radio is able to use spectrum efficiently.

## *2.3 Reinforcement Learning*

### 2.3.1 Machine Learning

Machine learning is a field that is concerned with the design and development of algorithms and techniques that allow agents automatically improve with experience [31-32]. It is a multidisciplinary field that draws on results from artificial intelligence, probability theory and statistics, computational complexity theory, control theory, information theory, etc. It is largely applied to the area of natural language processing, syntactic pattern recognition, search engines, medical diagnosis, bioinformatics, brain-machine interfaces and cheminformatics, speech and handwriting recognition, object recognition in computer vision, game playing and robot locomotion.

A general definition of a learning problem can be given as [31]: '*A agent is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.* A well defined learning problem will have 3 essential elements: a class of tasks, the measure of the improving performance and the source of experience.

We consider $x_1,x_2,x_3...x_t...$, is the sequence of input received by the learning agent, where $x_t$ is the input at time $t$, then three main kinds of machine learning can be distinguished [31-32]:

- *Supervised learning: Besides input, a sequence of desired outputs $y_1$, $y_2$, ...$y_t$... is also given to the agent. The objective of learning is to derive a function that maps inputs to desired outputs. In other words, the agent needs to predict the correct output given a new input.*

- *Unsupervised learning [33]: Agents receive the inputs $x_1,x_2,x_3...x_t...$, but without any supervised target outputs or any rewards. It is closely related to the problem of density estimation in statistics. Unsupervised learning is designed to find patterns in the data without any kind of feedback.*

- *Reinforcement learning [34]: Agents interact with an unpredictable environment by selecting different actions and receiving rewards accordingly. The goal is to maximize the long-term rewards which it will receive in the future.*

### 2.3.2 Reinforcement Learning

Reinforcement learning, a sub-area of machine learning, uses a mathematical way to evaluate the success level of actions [34-35]. Its emphasis on individual learning from the direct interactions with the environment makes it perfectly suited to distributed cognitive radio scenarios. Reinforcement learning has been considered as the most suitable learning approach for cognitive radio systems in this work. There are mainly two reasons:

1. Reinforcement learning is an individual learning approach where the learning agent learns only on local observations. This is perfectly suited to cognitive radios who also work on a fully distributed fashion.
2. Reinforcement learning learns on a trial-and-error basis that no environment model is required. This is also perfectly suited to cognitive

radio systems which constantly interact with an ‘unknown’ radio environment on a trial-and-error basis.

The original reinforcement learning model [35] where agents are interacting with the environment as illustrated in figure 2.4 consists of:

1. a set of possible states, represented by S;
2. a set of actions, A;
3. a set of numerical rewards R;

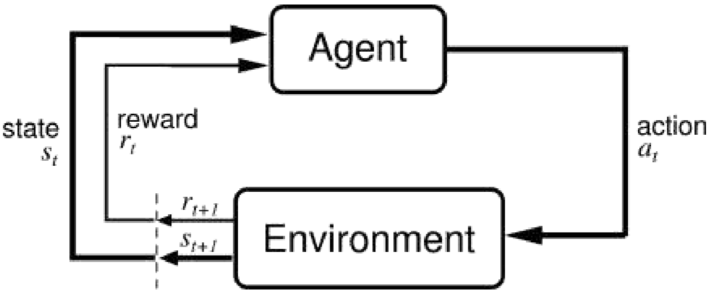


*Figure 2.4 Standard Reinforcement Learning Model (directly reproduced from [34])*

The learner is called the agent. The outside world which it interacts with is called the environment. At each time step *t,* the agent perceives the state of its surrounding environment, $s_t \in S$. Based on $s_t$, the agent chooses an action, $a_t \in A(s_t)$, where $A(s_t)$ is the set of available actions at time *t*. At the next time step *t+1*, the environment makes a transition to a new state $s_{t+1}$ and the agent receives a reward $r_t$. The objective is to develop an optimal policy π: $S \rightarrow A$ that can maximize the reward at state *S*. Given a state *s* and a policy π, the selection of a specific action is denoted as $a = \pi(s)$.

In the standard reinforcement learning algorithm, the value of the current state *s* under a policy π which is denoted by $V^{\pi}(s)$ is the basis to choose the action *A(s)*. An optimal policy is supposed to maximize $V^{\pi}(s)$ at each trial. $V^{\pi}(s)$ is formally defined as [34-35]:

$$V^{\pi}(s) = E\{\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) \big| s_t = s\} \qquad (2\text{-}1)$$

Where *E* is the expectation operator, γ is a discount factor (0 < γ < 1). *r(s, π(s))* is the immediate reward if the agent chooses action *a* = *π(s)* given a state *s*. Equation (2-1) can also be written as:

$$V^{\pi}(s) = R(s,\pi(s)) + \gamma \sum\nolimits_{s'} P(s'|s,\pi(s)) V^{\pi}(s') \qquad (2\text{-}2)$$

Where *R(s, π(s))=E{r(s, π(s))}* is the mean value of *r(s, π(s)). s'* stands for the goal states which *s* will transit to by taking the action *π(s)*. Given that there may be multiple successor states *s',* the probability *P(s'|s,π(s))* defines the probability of making a transition from state *s* to different successor states.

The optimal value function $V^{\pi^*}(s)$ under the optimal policy $\pi^*$ can be defined as:

$$V^{\pi^*}(s) = \max_{a\in A}\left(R(s,\pi(s)) + \gamma \sum\nolimits_{s'} P(s'|s,\pi(s)) V^{\pi^*}(s')\right) \qquad (2\text{-}3)$$

Based on the optimal value function $V^{\pi^*}(s)$, the optimal policy $\pi^*$ is specified as:

$$\pi^*(s) = \arg\max_{a\in A}\left(R(s,\pi(s)) + \gamma \sum\nolimits_{s'} P(s'|s,\pi(s)) V^{\pi^*}(s')\right) \qquad (2\text{-}4)$$

*R(s, π(s))* is effectively the cumulative reward in the state of *s*. The other part of the equation is the expected feedback of its successor states *s'*.

## *2.4 Traditional Distributed Dynamic Channel Assignment Techniques*

This work primarily investigates the application of distributed reinforcement learning to cognitive radio spectrum sharing. It can be considered as an extension of previous Distributed Dynamic Channel Assignment (DDCA) schemes since cognitive radio itself is a DDCA based technique. Thus, previous DDCA schemes are briefly reviewed in this section. It is worth mentioning that most of these DDCA schemes do apply a listen-before-talk style strategy which is quite similar to the spectrum awareness function of cognitive radio. Interference and *SINR* measurements are measured at entities that the channels are assigned based on these measurements.

Early stage research on DDCA can be traced back to the 1980s. In 1989, Åkerberg proposed to apply an interference threshold to determine whether a channel is available [36]. A Least Interference Channel (LIC) assignment scheme is investigated in this paper, and the Grade of Service (GoS) performance has been compared with Non-LIC schemes under different interference threshold settings. The results in this paper show that a tighter interference threshold always guarantees a better GoS performance when using different propagation models. This is because call dropping has been considered much more important than call blocking here. In fact, the GoS in this paper has been defined as the blocking probability plus 10 times of the dropping probability. Thus, a tighter interference threshold achieves a better call dropping-dominated GoS in all cases. However, if a different way of defining GoS is applied, for example a call blocking-dominated GoS, the conclusion could be very different. The results in [37] prove this argument by investigating a similar algorithm with a different performance measurement. The performance of the proposed algorithm has been compared with the MAXAVAIL DCA scheme [12]. It shows that LIC with no interference threshold performs the best.

In [38], the authors proposed a Local Autonomous Dynamic Channel Allocation (LADCA) with power control. Call blocking and call dropping are considered equally important when assessing the performance. The results in this paper show that distributed channel assignment and distributed power control can be combined, providing improved system performance. It can also be seen that nearly all unsuccessful calls are dropped calls that the capacity is limited largely by call dropping rather than call blocking.

Two CIR-based DDCA schemes are examined in [39], First Available (FA) and Best Quality (BQ). Instead of measuring interference level on the channels, these two schemes directly use CIR measurements as the basis to assign channels. The FA scheme chooses the first channel in a pre-defined list that satisfies the CIR requirement. The BQ scheme takes the channel with the highest CIR. The paper shows that the FA scheme is able to achieve a near-optimum performance by allowing call reassignment when the CIR of existing calls fall below a CIR threshold. Law extended the work in [40] by comparing the performance of the

FA scheme with a LIC scheme in the context of Digital European Cordless Telecommunications (DECT) environment. The outage probability for a desired Quality of Service (QoS) is defined in this paper as a measurement of system performance. The results show that the LIC scheme performs better when applying different interference thresholds.

The widely cited paper by Chuang [41] on the subject of DDCA proposed individually scheduled frequency-updating events at fixed facilities (base station/ radio ports). A port receiver turns off its transmitter and scans all available channels after receiving a call request. Then the least interference channel is assigned to the call. This perhaps is one of the early models of spectrum sensing. As a result, a self-organizing frequency assignment is achieved that optimal frequency reuse patterns could be reached after the system converges to a stable point despite unknown factors like random port location and shadowing. The performance of the approach has been compared with a random assignment scheme and a pre-planned assignment scheme. It shows that the least interference algorithm performs significantly better than the random assignment. More importantly the proposed scheme achieves a performance similar to the optimal pre-planned assignment approach.

In 1993, Chuang published another paper that investigates DDCA in the context of TDMA portable radio system [42]. A DDCA scheme has been proposed which aims to achieve a balanced uplink and downlink CIR by considering the best available action for both sides of the transmission link. A sophisticated process has been defined in order to identify such channels. The results show that the proposed scheme outperforms the approaches which only concern either the base station or the portable device.

The upper and lower bounds of the capacity of DDCA schemes are been investigated using analytical models [15]. In [43], interference based DDCA schemes have been studied. The upper bound and lower bound of the probability of unsuccessful calls are derived. It shows that the performance of the DDCA schemes could be better than FCA schemes. In [44], Whitehead provided the estimations of the capacity gain of DDCA algorithms. Geometric analysis of interference adaptation has been taking into account when analysing the trucking-efficiency. The results show that DDCA with integral interference-balancing power control can almost eliminate the variance of CIR, providing a capacity gain of 3. The advantages of DDCA schemes are clear that it provides similar performance comparing to FCA, however it removes the requirements of frequency planning. A pictorial model has been introduced in [20] to explain how call dropping happens when the devices are in a vulnerable region. Different factors that could affect DDCA schemes are also discussed in this paper. The proposed scheme is able to reduce call dropping significantly by pairing calls and measuring interference levels at both sides of the transmission link.

A large number of DDCA schemes have been developed in the last few decades. Most of these schemes are interference or CIR based. The schemes we have reviewed in this section are some of the most classic work in the area which should provide a clear idea on how DDCA works. A more comprehensive review on the more relevant intelligence-based channel assignment schemes are provided later.

## *2.5 Intelligent Channel Assignment Techniques*

### 2.5.1 Reinforcement Learning-based Schemes

Reinforcement learning-based channel assignment can be generally categorized into centralised algorithms where channels are assigned at a centralized server, and distributed algorithms where spectrum decisions are made by individual users. Research work in the field largely focused on centralised scenarios prior to the introduction of cognitive radio. Distributed learning-based algorithms draw more attention after cognitive radio has been introduced since cognitive radio works in a distributed fashion [21, 45-46]. However, it is more difficult to define the learning model in this scenario since entities are fully distributed and decisions are made only according to the local measurements. It is unlikely for a cognitive radio to obtain the information at the network level. Thus, the state of the system is more difficult to be defined and the state transition is not directly derivable. More details are given in the following sections [47].

#### *2.5.1.1 Reinforcement Learning-based Schemes prior to Cognitive Radio*

Q-learning [31], a reinforcement learning approach, has been frequently studied in centralized scenarios where the system information is available at the network level. Centralized learning-based channel assignment has been well studied for cellular networks in the last decade. The states *s* and *s'* are easier to be defined in this case because system level information is widely available in the centralized system.

**Applications**

The centralized Q-learning based dynamic channel assignment proposed by Junhong Nie and Simon Haykin [48] is the most widely cited work in this area. Q-learning has been applied to a cellular system that the channels are assigned on a call by call basis by utilizing the information gained through learning. Instead of fixed frequency planning, the learning based system is able to obtain an optimal channel assignment policy through the interaction with the wireless environment.

The system states are defined based on the channel availability information in different cells of the system. An action (assign a channel) will be chosen at different system states based on the Q-values of actions. The Q-values will be updated when the reward/cost is available. Extensive simulation results are provided in the paper that the Q-learning based approach has been compared with a Fixed Channel Assignment (FCA) scheme and one of the best performance Dynamic Channel Assignment (DCA) schemes MAXAVAIL [12] in a 49 cell mobile communication system scenario. It shown that the Q-learning based DCA algorithm performs better than the FCA in different traffic conditions, including spatially uniform and non-uniform traffic and time varying traffic. The Q-learning based DCA also achieves a similar performance with MAXAVAIL, however the computational complexity has been greatly reduced by using Q-learning.

In [49], Senouci and Pujoile extended the work of Nie and Haykin that they consider not only the channel assignment but also the call admission control for mobile netwok. Two classes of traffic are assumed in their work. The main contribution of this paper is that the traffic condition (number of calls each cell) has been considered when defining system state along with the channel availability information. Call rejection has also been considered when updating the Q-values of channels that previous research only concerns channel assignment problem. The results show that the Q-learning based approach is able to achieve an optimal policy in a real-time system. Compared with other DCA schemes, the Q-learning based scheme performs better when dealing with significant variations of the environment. The self-adaptive feature of learning-based algorithms is seen as one of the most important advantages of such algorithm.

### *2.5.1.2 Reinforcement Learning-Based Cognitive Radio Schemes*

There are typically three main steps in the learning-based cognitive radio schemes as illustrated in figure 2.5: Frequency Awareness, Frequency Resource Management, and Action [50]. An intelligent frequency decision making process is enabled through learning and reasoning. The reinforcement learning-based learning engine enhances the ability of cognitive radio device to select the appropriate spectral resources by exploiting the historical information kept in the knowledge base.

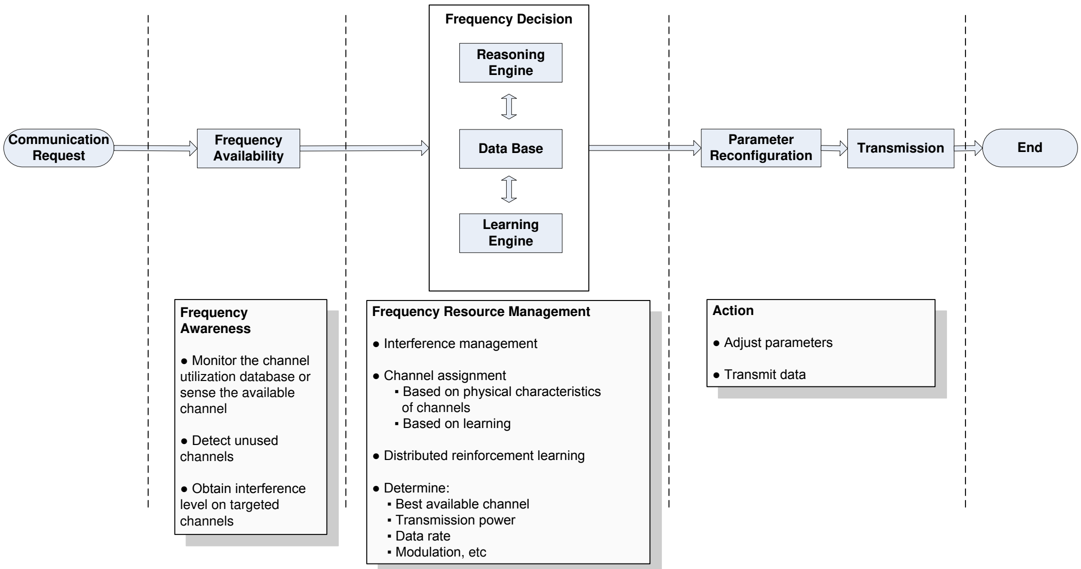


*Figure 2.5 Cognitive Radio based Radio Resource Management*

The learning engine processes external observations, e.g. interference and spectrum holes at first, and then combines such information with the historical information of successful or unsuccessful channel usage. This updated knowledge base will be used by the reasoning engine to make a decision on which resource to use in order to maximise the probability of success. After that, the operating cognitive radio user will adjust its transmission parameters according to the decision made by the frequency resource management function, and the transmission is carried out eventually.

### Applications

The authors of [51] considers a Q-learning based approach which gives rewards to the cognitive radio users after each data transmission. The Primary User's (PU) channel usage is assumed to be uniformly distributed on available channels. The state in the learning model has been defined by the number of neighbouring nodes. Data packet transmission is successful when an acknowledgment has been received, otherwise the transmission is unsuccessful. For each successful data packet transmission, a positive constant value of RW is awarded, otherwise a negative value CT is assigned. In practice, the value of the RW and CT are based on the amount of revenue and cost that a network operator earns or incurs for each successful or unsuccessful data packet transmission. It shows that the Q-learning based approach is able to increase the throughput typically by 2.84 times. However, it is worth mentioning that only one single user reinforcement learning based secondary user (SU) is assumed in this paper, meaning that other entities are using non-learning based channel assignment scheme. The system model and the learning model have been significantly simplified in this case. Later in [52-53], Yau and other authors extended the study to a multi-agent reinforcement learning scenario where more learning-based entities are operating. A Carrier Sense Multiple Access (CSMA) based system is assumed that the Q-values are updated after every packet transmission. The learning model in this paper requires the location information of entities at the system level in order to define the states of the system. It is shown that by enabling multi-agent Q-learning, the performance can be further enhanced. However, it is not clear how many system overheads and computational tasks will occur for updating and exchanging the required user location information.

A theoretical study has been carried out by Husheng Li in [54]. The author has made many assumptions in the paper in order to carry out the theoretical analysis. Multi-agent Q-learning has been assumed in a simple 2 SU x 2 Channel case. No PU is assumed. It is assumed that all of these two channels are available all the time, meaning that the only task for the two SUs is to avoid interfering with each other. In addition, spectrum sensing is ignored in the paper that the SU will not sense the channel before transmission. Although the results in the paper show that the algorithm converges to an equilibrium and the SUs learn to avoid collision quickly, the overly simplified system model makes the theoretical analysis similar to the cases which are well studied in Computer Science research. The wireless communication aspect of the research has almost been ignored.

Multi-agent reinforcement learning for cognitive radio has been studied in a more realistic scenario in [55]. A Q-learning based joint channel and power allocation scheme has been proposed. The state of the system has been defined by using the transmit power level and the channel utilization information of all users. Again this requires system level information that a large amount of system overhead information could be generated. There is no discussion in the paper on the costs of introducing such Q-learning based approach. It shows that the Q-learning based approach performs better than a random assignment approach.

A secondary cognitive radio system model based on IEEE 802.22 standard is considered in [56], where Q-learning based techniques is applied to learn how to control the transmit power in order to reduce the aggregated interference at PUs receivers. The learning model requires local information as well as network information to define the states of the system. It shows that the Q-learning approach is able to learn an optimal action policy to maintain the aggregated interference in the primary user network under a desired value.

### 2.5.2 Game Theory-based Cognitive Channel Assignment

Game Theory has originally been proposed by the mathematician John von Neumann to study human behaviour [57]. It is an interdisciplinary research area where mathematics and social and behavioural sciences are brought together. Game theory has been considered as an analytical tool which is widely used in different areas including finance, computer science and engineering, etc [58].

**Applications**

Game Theory is one of the tools that more and more researchers in this field try to apply to radio resource management. More flexible, efficient, and fair spectrum usage are been achieved by game theoretical dynamic channel assignment techniques where the behaviours of network users can be analyzed by Game Theory [59-60].

Significant work has been done by James Neel in applying Game Theory to the radio resource management of cognitive radio systems [61-64]. The convergence of the proposed approaches towards the Nash Equilibrium (NE) of the games has been studied extensively in their work.

In [65-66], Mangold formulated the game by using the information of Quality of Service (QoS) and data rate. The coexistence of Wireless LAN (IEEE 802.11) access points has been investigated in the context of game theory. The improvement gained by applying Game Theory can be clearly seen in his work.

Nie and Comaniciou have also proposed a game theoretic framework for distributed adaptive channel allocation of cognitive radio [67]. Two objective functions are proposed for the spectrum sharing games, capturing the behaviour of both selfish users and cooperative users. It shows that the non-cooperative games have the advantage of a low overhead requirement for information exchange. The cooperative spectrum sharing etiquette improves the overall system performance with a higher level of requirement for information exchange.

### 2.5.3 Genetic Algorithms

A Genetic Algorithm is an optimization technique that uses a number of bio-inspired evolutionary concepts, like inheritance, selection, mutation and crossover. The aim is to find a solution to an optimization problem [68]. A random solution is usually generated at the beginning of optimization. Then at each generation, the fitness of the solution is evaluated by a predefined fitness function, and the solution will be modified accordingly. The algorithm is terminated when a satisfactory fitness level is achieved or a maximum number of generations are reached.

**Applications**

A Genetic Algorithm can also be applied to distributed optimization problems. A genetic algorithm based frequency allocation approach for distributed cognitive radio networks is proposed by Si Chen and Alexander M. Wyglinski in [69]. A fitness function is developed to intelligently allocate frequency bands for subcarriers in a Non-Contiguous Orthogonal Frequency Division Multiplexing (NC-OFDM) system. Four system parameters are selected to be optimized by the proposed approach, normalized transmission power, modulation index, center frequency and bandwidth. Simulation shows that the proposed Genetic Algorithm is able to simultaneously minimize the bit error rate and the out-of-band interference, while maximizing the overall throughput.

## *2.6 Cognitive Radio-related Projects and Research*

Cognitive radio has been an emerging research area recently. There has been a rapid growth on the cognitive radio related activities worldwide in the last few years. A large number of cognitive radio-related research projects have been funded globally [70]. Europe, North America and East Asia are the most active areas in terms of cognitive radio research. Many aspects of cognitive radio communications have been investigated. Energy efficiency, spectrum efficiency, QoS and the self-organising features of cognitive radio systems are the most popular topics in the field. However, the main limitation of the current state-of-art research is the hardware implementation. Most of the research outcomes are demonstrated theoretically and only a few of the projects have successfully demonstrated their achievements through real hardware implementations.

### 2.6.1 European Projects

1. **Beyond Next Generation Mobile Broadband**

   The Beyond Next Generation Mobile Broadband (BuNGee) Project [71] started in January 2010 and will finish in June 2012. It aims to increase the mobile network capacity density to well beyond what the current next-generation techniques promise, employing a novel two-hop wireless system with new square or cross shaped cells. A multi-beam directional antenna, network MIMO, and cognitive radio based radio resource management techniques have all been investigated in order to deliver the demanded 1 Gbps/km$^2$ capacity density. The project involves 9 partners including University of York. It is funded by the European Union as part of the European Union's Framework 7. The work in this thesis has directly contributed to the BuNGee project.

2. **End-to-End Reconfigurability (E$^2$R,E$^2$RII), and End-to-End Efficiency (E$^3$)**

   The E$^2$R and E$^2$RII Project [72] aimed to develop prototype of reconfigurable devices, offering extensive options to regulators, operators, and users in the context of heterogeneous system.

   The E$^3$ Project [73] has investigated the integration of cognitive wireless systems with Beyond 3G systems, ensuring interoperability, flexibility and scalability between existing legacy and future wireless systems.

   These projects started in January 2004 and finished in December 2009. E$^2$R was funded under the 6th Framework Programme (FP6). E$^2$RII and E$^3$ were funded under the 7th Framework Programme (FP7). The coordinators were Motorala and Alcatel-Lucent.

3. **Quality of Service and Mobility Driven Cognitive Radio Systems (QoSMOS)**

   The ongoing 3-year FP7 QoSMOS project [74] started in January 2010. The project aims to enable '*the utilization of licensed and unlicensed bands for mobile broadband systems by integrating a cognitive radio framework*'. Opportunistic use of under-utilized spectrum bands is investigated with managed Quality of Service (QoS) and seamless mobility.

4. **Cognitive Radio Systems for Efficient Sharing of TV White Spaces in European Context (COGEU)**

   Different from many technology-oriented research projects in the area, the FP7 COGEU project [75] carries out research in policy, business and technical domains. The main objective is to enable an efficient utilization of unused TV white space for mobile networks by introducing secondary spectrum trading and new spectrum commons regime. This project will also define new methodologies for TV White Space equipment certification and compliance addressing coexistence with the DVB-T/H European standard. The technical aspect of this project aims to develop cognitive radio based techniques to support mobile applications in TV White Space.

5. **Cognitive Radio and Cooperation Strategies for Power Saving in Multi-Standard Wireless Device (C2POWER)**

   The power consumption of mobile devices is ever-increasing. Under the coordination of Instituto de Telecomunivacoes, the 3-year FP7 C2POWER project [76] aims to develop energy saving technologies for mobile systems by investigating the combination of cognitive radio and cooperative strategies. The desired energy saving will be achieved without compromising any of the existing performance requirements, i.e. data rate and QoS requirements. This project primarily concerns with two topics '*Cooperative power saving strategies between neighbouring nodes using low power short range communications*' and '*Cognitive handover* *mechanisms to select the Radio Access Technology which has the lowest energy demand in heterogeneous environments*'.

6. **Spectrum and Energy Efficiency through Multi-band Cognitive Radio (SACRA)**

The FP7 SACRA project's main objective is to develop a multi-band cognitive radio technology, achieving significant spectrum and energy efficiency improvement [77]. There are few aspects the project will investigate:

- Spectrum efficiency achieved by multi-band cognitive radio
- The minimization of electronic component number in wireless systems
- The energy optimization achieved by optimizing architecture and algorithms implementation
- The minimization of environmental interference by better assignment of frequency band

The project aims to, by the end of 2012, develop a proof-of –concept able to '*communicate jointly and cognitively in two separate frequency bands*'. The targeted two separate frequency bands are 790-862 MHz and 2.6 GHz.

7. **Quantitative Assessment of Secondary Spectrum Access (QUAZAR)**

Cognitive radio aims to significantly improve the spectrum efficiency by exploiting the under-utilized spectrum even if the cognitive radio devices are not authorised to do so. However, very limited research has been carried out to demonstrate that the large improvements to spectrum efficiency could be achieved. The FP7 QUAZAR project [78] aims to fill the gap by investigating the practical benefits of secondary access to primary spectrum. The impact of secondary users on primary users will be evaluated. The project will provide a roadmap and guidelines on new business models and initiate proposals to go beyond the current regulatory framework by the end of 2012.

8. **Opportunistic Networks and Cognitive Management Systems for Efficient Application Provision in the Future Internet (OneFIT)**

   The FP7 OneFIT project [79] aims to develop and validate opportunistic networks that are managed by advanced cognitive radio systems, enabling improved service for future internet. Opportunistic networks, cognitive management systems are two key aspects that the OneFIT project will primarily look into. The project started in 01/2010, and it finishes in 12/2011. There are 12 partners in total. The University of Piraeus Research Center is the project coordinator.

9. **Sensor Network for Dynamic and Cognitive Radio Access (SENDORA)**

   The main achievement of the FP7 SENDORA project [80] (from 01/2008 to 12/2010) is novel sensor network based techniques that support the coexistence of both licensed and unlicensed wireless users in a local area. The project investigated a wide range of topics including the identification and analysis of the business models of the wireless sensor network aided cognitive radio techniques, wireless sensor network aided dynamic resource allocation for cognitive radio and the design of a flexible and reconfigurable architecture. The achievement of the project is expected to contribute to future research in the related area and future network standards.

10. **Physical Layer for Dynamic Spectrum Access and Cognitive Radio (PHYDYAS)**

    The main objective of the FP7 PHYDYAS project [81] (01/2008 – 12/2010) was to develop advanced physical layer techniques that suitable for dynamic spectrum management and cognitive radio. A filter bank-based multicarrier technique has been investigated by PHYDYAS in the context of cognitive radio. The outcomes of the project show that the performance and the flexibility of systems are enhanced by exploring the spectral efficiency of filter banks and the independence of sub-channels.

11. **Flexible and Spectrum-Aware Radio Access through Measurements and Modelling in Cognitive Radio Systems (FARAMIR)**

    The on-going FP7 FARAMIR project [82] aims to develop advanced environmental and spectral awareness techniques for future wireless system. A knowledge base of radio environment will be built that cognitive radio is able to store and access information from it. Advanced spectrum sensing techniques and algorithms are also proposed. Extensive spectrum utilization measurements will be collected in Europe to gain a better knowledge on the spectrum usage at the same time.

12. **Advanced Coexistence Technologies for Radio Optimisation and Unlicensed Spectrum (ACROPOLIS)**

    The FP7 ACROPOLIS project [83] (10/2010 – 09/2010) aimed to tackle the medium and long term, interdisciplinary and fundamental research problems found in cooperative and cognitive communications. The fast development of wireless communication requires the integration of interdisciplinary knowledge. Experts in cooperation and coexistence, comprising research area such as cognitive radio, cognitive networking and flexible networking, are brought together to strength European knowledge and leadership in the relevant area.

### 2.6.2 North American Projects

1. **Defence Advanced Research Projects Agency (DARPA) Next Generation (XG) Project**

   The DARPA XG project [70] aimed to '*develop both the enabling technologies and system concepts to dynamically redistribute allocated spectrum along with novel waveforms in order to provide dramatic improvements in assured military communications in support of a full range of worldwide deployments'*. A set of advanced Dynamic Spectrum Access techniques have been developed by the project that the achievements could be the basis for the further development of cognitive radio. There were 3 phases in the project: the first phase was technical investments, from 2002 to 2003; then the system and protocol design phases lasted from 2003 to 2005; the third phase system development and demo started in 2005 and finished in 2008.

2. **Defence Advanced Research Projects Agency (DARPA) Wireless Network After Next (WNaN) Project**

   The key objective of the DARPA WNaN [70] is to develop an advanced high density wireless ad-hoc network of cognitive radios. Low-cost wireless nodes are assumed to be the basic elements of the network. However such low-cost nodes have certain physical layer limitations. By introducing technologies and system concepts which are able reduce the demands on physical and link layers through better node configurations and topology management, the wireless network is able to deliver the demanded capacity, enabling reliable communications at low system cost. The project also aims to develop a prototype handheld wireless node which could be deployed to form a high-density wireless network by the end of this project.

3. **National Science Foundation (NFS) Security Provisioning for Cognitive Radio Networks**

   The National Science Foundation has been one of the biggest funding bodies in the US. A number of cognitive radio related projects have been funded by NFS. Security Provisioning for Cognitive Radio Networks project [70] is one of these projects. This project aims to develop a comprehensive security system that serves as a secure backbone for cognitive radio networks that coexist with primary network in different system architectures and coexistence scenarios. The techniques developed are designed to be embedded into the whole network, enabling secured and reliable spectrum access for the cognitive radio networks.

4. **National Science Foundation (NFS) Cognitive Antennas for Wireless Ad Hoc Networks**

   The NFS Cognitive Antennas for Wireless Ad Hoc Network project [70] investigates a network consisting of cognitive radios with reconfigurable antennas. Reconfigurable antennas provide a new dimension of opportunities to reduce interference and increase link robustness. The project aims to show how the extra beam-domain opportunities enabled by reconfigurable antennas are able to support more aggressive frequency reuse and in turn increase the system capacity. Multi-sensor data fusion, distributed control techniques are all proposed.

5. **National Science Foundation (NFS) Human Behaviour Inspired Cognitive Radio Network Design**

   With the ability of environment awareness, learning, reasoning and self-adaptation, it is possible that cognitive radios to take irrational actions like human. This is a risk for future cognitive radio based networks. The NFS Human Behaviour Inspired Cognitive Radio Network Design project [70] investigates the possible behaviour of cognitive radios that is similar to human behaviour and social interactions.

6. **National Science Foundation (NFS) Cognitive Femtocells: Breaking the Spatial Reuse Limits of Cellular Systems**

   This NFS project aims to apply cognitive femtocells to indoor environment like residential buildings and offices, providing significantly improved service and coverage [70]. The cognitive femtocells are designed to share limited radio resource (the cellular frequency band) on an opportunistic fashion. No frequency planning is assumed for the femtocells. The cognitive femtocells need to explore the spectrum opportunities in the available bands.

7. **National Science Foundation (NFS) Beyond Listen-Before-Talk: Advanced Cognitive Radio Access Control in Distributed Multi-User Networks**

   This project aims to achieve improved spectral efficiency by applying advanced cognitive radio access and power control algorithms [70]. This is achieved by exploiting different levels of primary user's Data Link Control (DLC) signalling and feedback information, rather than purely rely on spectrum sensing.

### 2.6.3 Asian Research

Extensive research in the relevant area is being performed in Far East, especially in China and Japanese. In China, cognitive radio and cognitive radio networks related research are very active and largely funded under the Chinese 863 and 973 programs, and the National Natural Science Foundation of China. The topics cover most aspects of cognitive communications, including spectrum sensing, resource management, security, etc. A large number of highly cited publications are been produced by these projects. Japanese research in this field is leading by the National Institute of Information and Communications Technology (NICT). Japanese research projects run by NICT aims to research and develop the technologies for new generation user-centric wireless networking which will be highly reliable and robust in various environments. The research topics in Japan include spectrum sensing, radio resource acquisition/management, software defined radio and Universal RF, wideband mixers/antennas for multimode / multi-band communications. Extensive research activities in Japan also contribute to the standardization of next generation wireless communication systems.

## *2.7 Conclusions*

This chapter provides the background information related to this work. The information of cognitive radio system is given first where the definition of cognitive radio is given. The main features and the cognition cycle of cognitive radio systems have also been discussed. Reinforcement learning is introduced in section 2.3 where the details of the learning model are available.

The traditional non-learning based DDCA schemes are briefly reviewed in section 2.4. It is worth to summarize the research work carried out in this area since cognitive radio itself is a DDCA based technique. Similarities between cognitive radio and DDCA schemes have also been discussed in this section that the listen-before-talk style strategy has been proposed long before the introduction of cognitive radio.

The more relevant ‘intelligent’ channel assignment techniques are reviewed in section 2.5. Centralized learning-based schemes and distributed learning-based schemes are discussed with details. Most of the learning-based algorithms require system level information in order to define the states of the system in the learning model. However, such information is not readily accessible in a fully distributed cognitive radio system. Game Theory and Genetic Algorithm based schemes are also discussed in this section. Thus, the state-of-art techniques relevant to this work are properly introduced.

Furthermore, the cognitive radio related research projects have been reviewed in section 2.6. The European research projects, American research projects and the research carried out in Far East have all been discussed.

# Chapter 3. System Modelling and Performance Evaluation Methodology

**Contents**

## *3.1 Introduction*

This chapter presents the research methodologies, system modelling techniques and the key measurements used in this thesis. System modelling by using professional simulation software is one of the most widely used approaches to conduct research nowadays, especially in Engineering. This is because firstly computing power has grown significantly in the last 10 years. With the rapid development of professional computing and programming software, the computing power is sufficient and convenient to modelling real-life communication systems in a very detailed fashion. Secondly, the cost for conducting the system modelling work is low and the reconfigurability of such models is significant. Thus simulation has been considered as a time-cost effective approach to verify different techniques.

Considerable efforts have been made on developing the simulator in this work. The multi-agent reinforcement learning scenario investigated in this work is extremely difficult to be evaluated mathematically. Thus, extensive system modelling tasks have been carried out and the results are used to evaluate the system capacity in this thesis. Significant effort has also been spent on investigating and discussing such results in order to explain the user behaviour under different situation.

The system modelling techniques are introduced in the next section. Then the key measurements we used to evaluate the system capacity are described in section 3.3. The information of modelling verification and the conclusion are given in section 3.4 and section 3.5 respectively.

## *3.2 System Modelling Techniques*

A number of programming tools are available to conduct system modelling tasks, from the most basic C language to well designed OPNET and Matlab. All of those tools could be the platforms of carrying out the system modelling tasks and they all have their own advantages and disadvantages.

Matlab, one of the most popular professional numerical computing environment and programming language, is used as the main system modelling tool in this thesis. It has similarities to C, but offers a much more user friendly environment. A great deal of work has been done in facilitating Matlab with matrix computing that the matrix computing tasks could be performed very easily. A large number of functions and toolboxes, which are very helpful in terms of developing the code for system modelling, are also available in Matlab that the flexibility and the extensibility of the program done in Matlab are incredible. Thus, it normally requires less time when developing or modifying simulators using Matlab than using C language. The disadvantage of Matlab is that the code runs much slower compared to C. Again with the rapid growth of the computing power, execution time of codes is a far less important factor in terms of developing a simulator. Instead easier development and modification of the code are more important. Thus Matlab is considered to be the most appropriate system modelling tool to in this work.

Monte Carlo simulation has been used in this work since the desired results are infeasible to be computed with a deterministic algorithm. Monte Carlo simulation, a statistical simulation method, relies on repeated random sampling to generate statistical results [84]. A relatively large number of trials are needed in this case to reduce the effect of the random fluctuation. The results become more statistically accurate if more trials are taken into account. Event-based strategies are used extensively in the simulation where only discrete events are captured. In other words, measurements are taken when new events happened in the system. The simulation execution time of such event-based strategy is significantly less than a time-continuous simulation approach. The general process of such event-based simulation is illustrated in figure 3.1. The simulator will firstly generate the location of the entities, the propagation environment, and the arrival and departure time of users (the time of events) based on a set of predetermined parameters. Then the simulator goes through every event and the measurements are taken in each event. After a large number of events have been sampled, the statistics of the large number of samples are obtained to illustrate the behaviour of the systems.

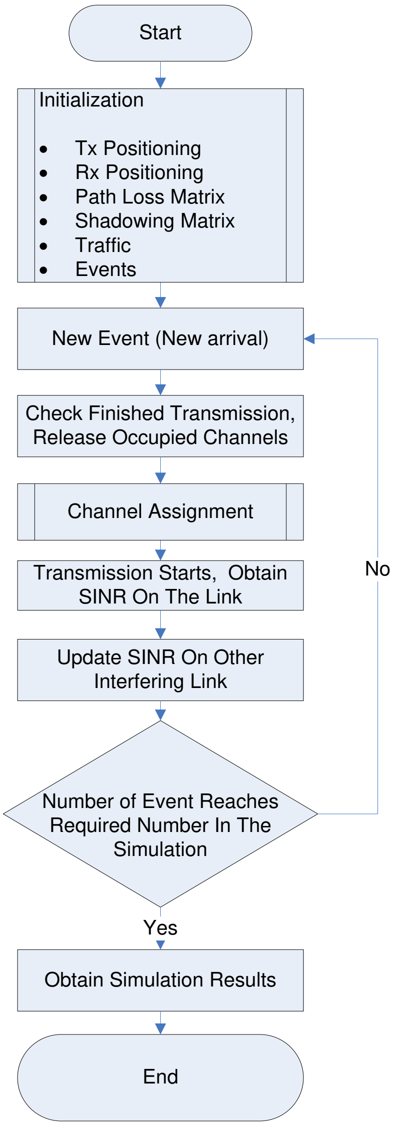

Start
Initialization
Tx Positioning
Rx Positioning
Path Loss Matrix
Shadowing Matrix
Traffic
Events
New Event (New arrival)
Check Finished Transmission, Release Occupied Channels
Channel Assignment
Transmission Starts, Obtain SINR On The Link
Update SINR On Other Interfering Link
Number of Event Reaches Required Number In The Simulation
No
Yes
Obtain Simulation Results
End

*Figure 3.1 Typical Event-based Simulation Process*

Performance measurements are normally plotted against traffic load in this thesis because it is the best way to show the behaviour of the system under different traffic conditions. For example, blocking probability versus offered traffic and throughput versus offered traffic. Plots of different schemes at a same offered traffic load are also provided to show more details of such schemes.

## *3.3 Performance Measures*

A few performance parameters are selected to evaluate the system capacity in this work. Signal-to-Interference-plus-Noise-Ratio (SINR) is used to evaluate link quality, i.e. to determine whether the current user will lose its current service, or to determine the data rate depending on the adaptive modulation applied to the system. Blocking probability and dropping probability are normally used to evaluate link based wireless system, e.g. speech-oriented wireless service. These two parameters are used extensively in this thesis to describe system capacity. Throughput, commonly used to determine the capacity of data-oriented wireless services, is used in this thesis to evaluate the capacity of beyond next generation mobile networks. The Cumulative Distribution Function (CDF) is used to process the initial data and to deliver the statistical behaviour of the results.

### 3.3.1 Signal-to-Interference-plus-Noise-Ratio (SINR)

Signal-to-Interference-and-Noise Ratio (SINR) [85], also known as Carrier-to-Interference-and-Noise Ratio (CINR), is one of the fundamental parameters to measure the link quality of users in wireless communication. It is defined by the quotient of the average received signal power (*S* or *C*) and the average received co-channel interference power (*I*) plus the noise power from other sources (*N*). Two types of architectures are investigated in this work and the SINR has been derived separately:

- Point-to-Point

If we consider a network with $M$ transmitter-receiver pairs and $Q$ channels, the SINR measured at the $n^{th}$ receiver on channel $q$ can be obtained as:

$$\gamma_n = \frac{p_n g_{n,q}}{\sum_{i=1,i\neq n}^{M} p_i g_{i,q} + \sigma^2} \tag{3-1}$$

where $p_n$ is the transmit power of the $n^{th}$ transmitter, $g_{n,q}$ is the gain of the wireless link on channel $q$, $\sigma^2$ is the noise power.

- Dual-Hop Beyond Next Generation Mobile Network

If we consider a wireless network with $M$ HBSs, each HBS has $L$ beams, serving $N$ ABSs. Each ABS provides service to $K$ MSs. There are $Q$ Channels available in total, each channel is divided into $R$ Subchannels. If we assume $P^H = \begin{pmatrix} p_1^{H,1} & . & . & . & p_1^{H,L} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ p_M^{H,1} & . & . & . & p_M^{H,L} \end{pmatrix}$ is the HBS transmission power matrix, and $p_m^{H,l}$ denotes the transmission power of the beam $l$ of HBS $m$. $P^A = \begin{pmatrix} p_1^{A,1} & . & . & . & p_N^{A,1} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ p_1^{A,M} & . & . & . & p_N^{A,M} \end{pmatrix}$ is the ABS transmission power matrix, then $p_n^{A,m}$ denotes the transmission power of ABS $n$, which is associated with HBS $m$.

A frequency separation of backhaul and access is assumed so that the backhaul network and the access network do not interfere with each other. Then for the backhaul network, SINR measured at ABS $n$ (signal from HBS $m$ in channel $q$ and subchannel $r$) can be derived as:

$$\gamma_{n,q,r}^{m,l} = \frac{p_m^{H,l} g_{q,r}^{B,m,l}}{\sum_{i=1,i\neq m}^{M} \sum_{j=1}^{L} p_i^{H,j} g_{q,r}^{B,i,j} + \sum_{i=1,i\neq l}^{L} p_m^{H,i} g_{q,r}^{B,m,i} + \sigma^2} \tag{3-2}$$

where $g_{q,r}^{B,m,l}$ is the gain of the wireless link from the $l$th beam of HBS $m$ to ABS $n$. $\sum_{i=1,i\neq m}^{M} \sum_{j=1}^{L} p_i^{H,j} g_{q,r}^{B,i,j}$ is the interference from other HBSs to ABS $n$.

$\sum_{i=1,i\neq l}^{L} p_m^{H,i} g_{q,r}^{B,m,i}$ is the interference comes from other beams of HBS $m$, using the same channel $q$ and subchannel $r$. $\sigma^2$ is the noise power.

Similarly for the access network, the SINR received at MS $k$ (signal from ABS $n$ (associated with HBS m) in channel $q$ and subchannel $r$) is:

$$\gamma_{k,q,r}^{n,m} = \frac{p_n^{A,m} g_{q,r}^{A,n,k}}{\sum_{i=1,i\neq m}^{M}\sum_{j=1}^{N}\sum_{u=1}^{K} p_j^{A,i} g_{q,r}^{A,j,u} + \sum_{i=1,i\neq n}^{N}\sum_{j=1}^{K} p_i^{A,m} g_{q,r}^{A,i,j} + \sigma^2} \tag{3-3}$$

where $g_{q,r}^{A,n,k}$ is the link gain between ABS $n$ and MS $k$. $\sum_{i=1,i\neq m}^{M}\sum_{j=1}^{N}\sum_{u=1}^{K} p_j^{A,j} g_{q,r}^{A,j,u}$ is the interference from all the ABSs in other cells that are using the same frequency. $\sum_{i=1,i\neq n}^{N}\sum_{j=1}^{K} p_i^{A,m} g_{q,r}^{A,i,j}$ is the interference from other ABSs in the same cell, and $\sigma^2$ is the noise power.

### 3.3.2 Blocking Probability and Dropping Probability

Blocking probability and dropping probability [3] are the measurements we use to evaluate the grade of service in this thesis. The blocking probability at time $t$ can be defined as:

$$P_B(t) = \frac{N_b(t)}{N_a(t)} \tag{3-4}$$

where $P_B(t)$ is the blocking probability at time $t$. $N_b(t)$ is the total number of blocked activations of the system by time $t$ and $N_a(t)$ is the total number of activations of the system by time $t$. Similarly, the dropping probability is defined as follows:

$$P_D(t) = \frac{N_d(t)}{N_{sa}(t)} \tag{3-5}$$

where $P_D(t)$ is the dropping probability by time $t$. $N_d(t)$ is the total number of dropped transmissions by time $t$ and $N_{sa}(t)$ is the total number of accepted activations by time $t$.

Note that for the dual-hop links when calculating $P_D(t)$ for the end-to-end link, the $N_{sa}(t)$ only takes into account the successful transmission over the end-to-end link.

### 3.3.3 Throughput and Throughput Density

System throughput is the major measurement we used in this thesis to describe the system performance of the beyond next generation mobile network. Throughput density is also defined to show the performance since beyond the next generation mobile network is designed primarily to deliver a high throughput density. Adaptive modulation is assumed that all entities are transmitting at the highest data rate that the wireless links can support (best effort basis) based on the *SINR* levels of these links. Therefore, the system throughput can be defined as:

$$Thr_s = \sum_{i=1}^{N_u} \sum_{k=1}^{n_i} \sum_{t=0}^{T_k} Thr_{MIMO-TSB}(t) \cdot BW_c \cdot P_{TDD} \tag{3-6}$$

$Thr_{MIMO\text{-}TSB}(t)$ is the data rate of a MIMO link obtained at time $t$, and it is updated constantly in the simulation. The Truncated Shannon Bound is proposed in [86] and we use it to determine the data rate of links. If we assume an adaptive modulation and coding (AMC) codeset, then the data rate of a link can be obtained by:

$$Throught, Thr, bps/Hz = \left| \begin{array}{lll} Thr = 0 & for & SINR < SINR_{MIN} \\ Thr = a \cdot S(SINR) & for & SINR_{MIN} < SINR < SINR_{MAX} \\ Thr = Thr_{MAX} & for & SINR > SINR_{MAX} \end{array} \right.$$

$$S(SINR) = \log_2(1 + SINR)$$

Where $S(SINR)$ is the Shannon Bound. $a$ is the anttenuation factor. $SINR_{MIN}$ is the minimum *SINR* threshold of the codeset, and $SINR_{MAX}$ is the *SINR* when max throughput is reached. $Thr_{MAX}$ is the maximum throughput of the codeset. For the work carried out in chapter 8, the parameters are defined as: $a$ = 0.65, $SINR_{min}$ = 1.8 dB, $SINR_{max}$ = 21 dB and $Thr_{max}$ = 4.5 bps/Hz.

$T_k$ is the transmission time of the $k^{th}$ transmission of an entity, and $n_i$ is the total number of transmissions that have been finished by the $i^{th}$ entity in the simulation.

$N_u$ is the total number of users in the system. $n_i$ is determined by the offered traffic level and the probability of a transmission been successful:

$$n_i = OT \cdot P_s^i(t) \tag{3-7}$$

$P_s^i(t)$ is the transmission successful probability of entity $i$ at time $t$, and it can be defined as:

$$P_s^i(t) = (1 - P_B^i(t)) \cdot (1 - P_D^i(t)) \tag{3-8}$$

$P_B^i(t)$ and $P_D^i(t)$ are the blocking probability and the dropping probability of entity $i$ at time $t$ respectively.

$OT$ in function 3-7 is the system offered traffic. The offered traffic level of a user $OT_u$ can be defined as:

$$OT_u = \frac{T_{ser}}{T_{ser} + T_{\text{int}}} \tag{3-9}$$

where $T_{ser}$ is the mean transmission service time of a user and $T_{int}$ is the mean transmission interarrival time of a user. $OT_u$ effectively shows the percentage of transmission time in the simulation, e.g. $OT_u = 0.5$ means that this user is transmitting data in 50% of the simulation time. $OT_u = 1$ means that this user is constantly transmitting in the simulation.

$OT$ then can be defined as:

$$OT = OT_u \cdot N_u \tag{3-10}$$

$OT$ effectively shows the average number of active users at any time in the simulation, e.g. if $OT_u = 1$, $N_u = 100$, then $OT = 100$, it means that 100 users are constantly transmitting data in the system. If $OT_u = 0.5$, $N_u = 100$, then $OT = 50$, it means that averagely 50 users are transmitting at any moment in the simulation.

$BW_c$ is the subchannel bandwidth. $P_{TDD}$ is the percentage of time slots have been allocated to the downlink or uplink.

Throughput density can then be defined by:

$$Thr_D = Thr_s / A_s \tag{3-11}$$

where $A_s$ is the service area.

### 3.3.4 Cumulative Distribution Function (CDF)

As we mentioned before, in order to obtain statistically accurate results we need to apply Monte Carlo simulation. However, a very large amount of unprocessed data can be expected by conducting Monte Carlo simulation. Appropriate mathematical analysis in this case is required to show the statistical behaviour of the results.

The cumulative distribution function is the main statistical method applied in this report. The CDF of *x* is defined as [84]:

$$CDF \equiv F(x) = \int_{-\infty}^{x} f(t)dt \tag{3-12}$$

Where *f(x)* is the probability density function of *x*. The results of our simulation like blocking probability and dropping probability are mainly measured at regular points in the service area. The CDF of these results will clearly show the probability of a valued random variable with a given distribution.

## 3.4 Verification

Queueing Theory [87] is commonly used to analyse the behaviour of call-oriented communication system. Well defined analytic models, like the Erlang B formula, have been developed using Queueing Theory to describe different types of queueing systems. Normally performance measurements like blocking probability could be analysed based on Queueing Theory.

However there is no analytic model available for multi-agent reinforcement learning scenario. The autonomous behaviour of the completely distributed users is extremely difficult to be fully described mathematically. Moreover, spectrum sensing also changes the user behaviour significantly since it constantly tries to guide the transmissions moving from heavily interfered channels to less interfered channels. Thus, like the way by which most research is carried out in such multi-agent learning scenario in Computer Science [31-32], this thesis mainly uses Monte Carlo simulation to evaluate the performance of different schemes.

Nevertheless, detailed analysis is given in each chapter to discuss the behaviour of users in specified cases. Basic mathematical analyses for the influence of weighting factors and different settings of the preferred channel set are available in chapters 4 and 5. A more comprehensive analytical model is developed in chapter 6 where the total number of the trials cognitive radios are required to discover before an optimal channel is found in a multi-agent scenario is derived mathematically based on the value function and Probability Theory. The theoretical and experimental learning costs of different schemes are compared also in chapter 6 which verifies both the analytical model and the simulator.

## *3.5 Conclusions*

This chapter describes the ways in which the system modelling tasks are conducted in this thesis. The simulation tool and system modelling techniques are discussed first. Matlab is used as the main simulation tool in this work and Monte Carlo approach is applied to generate statistically meaningful results. The key measurements are then defined in order to evaluate the system capacity. Finally a brief discussion on the verification of the simulation approach is given.

# Chapter 4. Reinforcement Learning-based Spectrum Sharing

**Contents**

## *4.1 Introduction*

The fundamental objective of cognitive radio is to enable an efficient utilization of the wireless spectrum through a highly reliable approach. Although a cognitive radio may be able to analyze the physical environment before it sets up a communication link, the best system performance is unlikely to be achieved by either a random spectrum sensing strategy or a fixed spectrum sensing policy [88]. The system performance is expected to be improved by utilizing the historical information of the wireless environment gained through learning-based techniques.

This chapter introduces the reinforcement learning-based distributed spectrum sharing scheme which enables efficient usage of spectrum by exploiting users' past experience. In our spectrum sharing scheme, a reward value is assigned to a used resource based on the reward function. Cognitive radio users select spectrum resources to use based on the weight values assigned to the spectral resources - resources with higher weights are considered higher priority. Furthermore we investigate and compare the system performance of different sets of reward values which effectively are the weighting factors in the reward function. In fact, we will show how different weighting factor values have significant impact on the system performance, and that inappropriate weighting factor setting may cause some specific problems.

The cognitive radio based reinforcement learning model will be presented in section 4.2 first. The purpose of our work is not only to develop a number of algorithms for cognitive radio for certain scenarios, but more importantly to propose a generic learning model which is widely applicable to cognitive communications.

A few essential concepts like the value function and weighting factors will then be discussed in section 4.3 – 4.4. In section 4.5, a reinforcement learning-based algorithm is described. The schemes presented in this chapter were developed in the early stage of this work. The analysis of the results in section 4.6 focuses on the channel partitioning by reinforcement learning-based algorithms, the impact of weighting factors and the improvement of system performance in terms of blocking probability and dropping probability.

## *4.2 Cognitive Radio based Reinforcement Learning Model*

The reinforcement learning model developed for the cognitive radio scenario is illustrated in Figure 4.1. The cognitive radio is the learning agent. The wireless spectrum is effectively the environment. The way we implement reinforcement learning in the CR scenario is slightly different from the original model presented previously in Chapter 2. This is caused by a few built-in features of cognitive radio. In the original reinforcement learning system, the value of the current state *s* under a policy π which is denoted by $V^{\pi}(s)$ is the basis to choose the action A(s). An optimal policy is supposed to maximize $V^{\pi}(s)$ at each trial. $V^{\pi}(s)$ is formally defined as [34]:

$$V^{\pi}(s) = E\{\sum_{t=0}^{\infty} \gamma^{t} r(s_t, \pi(s_t)) \big| s_t = s\} \tag{4-1}$$

where $E$ is the expectation operator, $\gamma$ is a discount factor ($0 < \gamma < 1$). $r(s, \pi(s))$ is the immediate reward if the agent chooses action $a = \pi(s)$ given a state $s$. Equation (4-1) can also be written as:

$$V^{\pi}(s) = R(s,\pi(s)) + \gamma \sum_{s'} P(s'|s,\pi(s)) V^{\pi}(s') \tag{4-2}$$

where $R(s, \pi(s))=E\{r(s, \pi(s))\}$ is the mean value of $r(s, \pi(s))$. $s'$ stands for the goal states which $s$ will transit to by taking the action $\pi(s)$. Given that there may be multiple successor states $s'$, $P(s'|s,\pi(s))$ defines the probability of making a transition from state $s$ to different successor states.

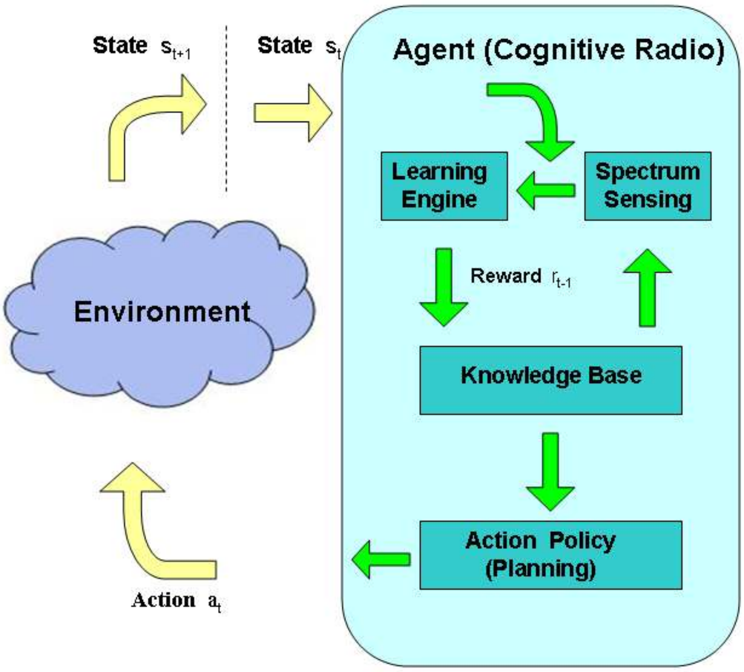


*Figure 4.1 The Reinforcement Learning Model in Cognitive Radio Scenario* [89]

The optimal value function $V^{\pi*}(s)$ under the optimal policy $\pi^*$ can be defined as:

$$V^{\pi*}(s) = \max_{a \in A}\left(R(s,\pi(s)) + \gamma \sum_{s'} P(s'|s,\pi(s)) V^{\pi*}(s')\right) \tag{4-3}$$

Based on the optimal value function $V^{\pi*}(s)$, the optimal policy $\pi^*$ is specified as:

$$\pi^*(s) = \arg\max_{a \in A}\left(R(s,\pi(s)) + \gamma \sum\nolimits_{s'} P(s'|s,\pi(s))V^{\pi^*}(s')\right) \tag{4-4}$$

*R(s, π(s))* is effectively the cumulative reward in the state of *s*. The other part of the equation is the expected feedback of its successor states *s'*.

It can be clearly seen from equation (4-1) to equation (4-4) that in order to obtain the optimal policy $\pi^*$, the information of *s'* is vital. Information like the number of potential successor states and the estimated value of each of the states *s'* are essential. Earlier work which tried to apply reinforcement learning into communications is largely focused on cell-based and centralized scenarios, where they are able to obtain the information of *s* and *s'* because such information is fully available within the communication system. For instance, in [67], Nie and Haykin define state $s_t$ at time *t* as:

$$s_t = (i, A(i))_t \tag{4-5}$$

where $i \in \{1,2,\cdots,N\}$ is the cell index, *i* indicates there is an event (call arrival or departure) occurring in cell *i*. $A(i) \in \{1,2,\cdots,M\}$ stands for the available channels in cell *i* at time *t*. By utilizing the information *i* and *A(i)*, the system can obtain the information to calculate $V^{\pi}(s)$ and to discover its optimal policy $\pi^*$ eventually.

Our strategy is to develop a policy π that maps memory (weight values) to action π: $W \rightarrow A$ instead of the original approach which maps the state of environment to action π: $S \rightarrow A$ [90]. On one hand, the agents are fully distributed in our scenario so that decisions are made only according to the local measurements. It is unlikely for a CR to obtain the information at the network level. On the other hand, it is worth considering whether it is necessary to obtain such information in a cognitive radio scenario even if it is possible. Cognitive radio is able to sense the target spectrum before activation and it is not supposed to transmit data until unoccupied spectrum has been found. With the ability of spectrum sensing, the information of available resources or occupied resources is not necessary if the objective is to find appropriate spectrum for the user. The only matter is how to discover the appropriate spectrum efficiently. Choosing the most successful spectrum by reinforcement learning combined with spectrum sensing is the suggested method in this work. A few amendments have been made to the learning model. The reinforcement learning model which we used consists of [34]:

1. A set of memories, *W*. *W* is a set of weights of the performed actions which are stored in the knowledge base;
2. A set of actions, *A*;
3. A set of numerical rewards *R*;

A CR will access the communication resource according to the memory of reinforcement learning. The success level of a particular action, which is whether the target spectrum is suitable for the considered communication request, is assessed by the learning engine. Based on the assessment, a reward is assigned in order to reinforce the weight of the performed action in the knowledge base. Since the actions are all strongly connected to the target resources, the weight is practically a number which is attached to a used resource and this number reflects the successful level of the resource. Our goal is to develop an optimal policy mapping weight to action $\pi$: $W \rightarrow A$ that can maximize the value of the current memory $V^{\pi}(w)$. Given a set of available weights of used resources and a policy $\pi$, the selection of a specific action is denoted as $a = \pi(w)$. Then $V^{\pi}(w)$ is defined as:

$$V^{\pi}(w) = \sum\nolimits_{w'} P(w'|w, \pi(w)) \cdot w' \tag{4-6}$$

Where $w$ is the weight of used resources of an agent at time $t$, $w'$ is the expected values of weights after agent takes an action $\pi(w)$. $P(w'|w,\pi(w))$ is the probability of selecting an action after taking the action $\pi(w)$. Accordingly, the optimal value function under the optimal policy $\pi^*$ can be defined as:

$$V^{\pi^*}(w) = \max_{a \in A}\left(\sum\nolimits_{w'} P(w'|w, \pi(w)) \cdot w'\right) \tag{4-7}$$

The optimal policy in our work therefore can be specified as:

$$\pi^*(w) = \arg\max_{a \in A}\left(\sum\nolimits_{w'} P(w'|w, \pi(w)) \cdot w'\right) \tag{4-8}$$

At each communication request the agent chooses a resource which can maximize $V^{\pi}(w)$ according to its current memory. Based on the result, the learning engine updates the knowledge base by a reward *r*. The inner loop within cognitive radio in figure 4.1 will proceed constantly to update the knowledge base. Global information is not necessary in this case. From this point of view, the complexity of the communication system is reduced.

## *4.3 Value Function*

Reinforcement learning is a computational approach to learn how to map situations to actions, and it is well suited to problems which include a long-term versus short-term reward trade-off. A key element of reinforcement learning is the value function [91]. A CR user updates its knowledge based on the feedback of the value function. In other words, the CR user adjusts its operation according to the function. From this point of view, the value function in reinforcement learning is also the objective function of cognitive radio in our scenario. The following linear function is used as the objective function to update the spectrum sharing strategy in this work [59, 88]:

$$W_t = f_1 W_{t-1} + f_2 \tag{4-9}$$

where $W_{t-1}$ is the weight of a channel at time *t-1*, and $W_t$ is the weight at time *t* according to previous weight $W_{t-1}$ and the updated feedback from system. $f_1$ and $f_2$ are the weighting factors at time *t* that will take on different values depending on the localized judgment of current system states and the environment. In order to update the weights in the knowledge base, either a reward value or a punishment value is assigned to *f* based on the evaluation of the success level of CR users' action.

## *4.4 Weighting Factors*

Weighting factors have great influence on the system performance, they reflect the degree of responses of a learning agent towards the changes of environment, i.e. a high reward or punishment value means that the learning node will adjust its actions swiftly according to the changes of the wireless environment, and a mild reward or punishment means that the learning node is adapting itself gradually based on the interactions with the environment [88].

The values of weighting factors are shown in table 4.1. Based on the degree of success, either a reward or a punishment is assigned to the weight of the used spectrum.

*Table 4.1 Weighting Factor Values*

| **SCHEMES** | $f_1$ | | $f_2$ | |
|---|---|---|---|---|
| | **Reward** | **Punishment** | **Reward** | **Punishment** |
| Mild Punishment | 1 | 1 | 1 | -1 |
| Harsh Punishment | 1 | 0 | 1 | 0 |
| Discounted Punishment | 1 | 0.5 | 1 | 0 |

The reward value of 1 is used in all of the three schemes in Table 4.1. The main difference between these schemes is the values assigned to punishment factors. In the first scheme, the absolute values of the reward value and the punishment value are equal. In other words the weight is increased or decreased by the same step size. This scheme is also named the '*mild punishment scheme*'. In the second scheme, if the attempt for communication fails, the weight is directly reduced to zero. Therefore we call it the '*harsh punishment scheme*'. Practically, the second scheme is a low complexity learning scheme where the CR users only remember the last successful spectrum and keep using it at new activation until the request for that resource is declined. Then the user picks up a channel randomly and keeps using it as long as the quality of communication in that channel is above the requirement. Weights are reduced by a certain percentage in the third scheme, and a percentage of 50% is used to reduce the weight of an unsuccessful channel. We can refer to the scheme as the '*discounted scheme*'.

## *4.5 Reinforcement Learning based-Distributed Spectrum Sharing Algorithms*

The basic Reinforcement Learning-based distributed CR spectrum sharing algorithm is illustrated in Fig. 4.2 [88]. We consider the CR users are a set of transmitting-receiving pairs of nodes, denoted as $U$, uniformly distributed in a square area and all the pairs $U_i \in U$ are spatially fixed. There are 3 main steps in the process:

**Step 1: Spectrum selection**. At the beginning of each activation, $U_i$ chooses a channel according to the weights of the available resources. It starts with the channel with the highest weight, or picks up a channel randomly if all resources have same priority. The selected channel is denoted as $C_k$ where $C_k \in C$ and $C$ is the available channel set.

**Step 2: Spectrum sensing.** $U_i$ senses the interference level on $C_k$. If the interference level $I$ of $C_k$ is below the interference threshold $I_{\mathrm{thr}}$, $U_i$ is activated. Otherwise if $I > I_{\mathrm{thr}}$ , the weight of $C_k$ for $U_i$ is decreased by a punishment weighting factor and $U_i$ returns back to step 1.

**Step 3: *SINR* measuring.** After step 2, the existing users within the same channel can measure the Signal-to-Interference-plus-Noise Ratio (*SINR*) at their receivers. The purpose of measuring *SINR* is to maintain the communication quality of the channels. We set up a *SINR* threshold $SINR_{thr}$. If the *SINR* of the activated pair $U_i$ is greater than the threshold ($SINR_i > SINR_{thr}$), $U_i$ successfully uses the spectrum and the weight of the channel will be increased by a reward. If $SINR_i < SINR_{thr}$, $U_i$ is blocked by the channel and the weight is updated with a punishment.

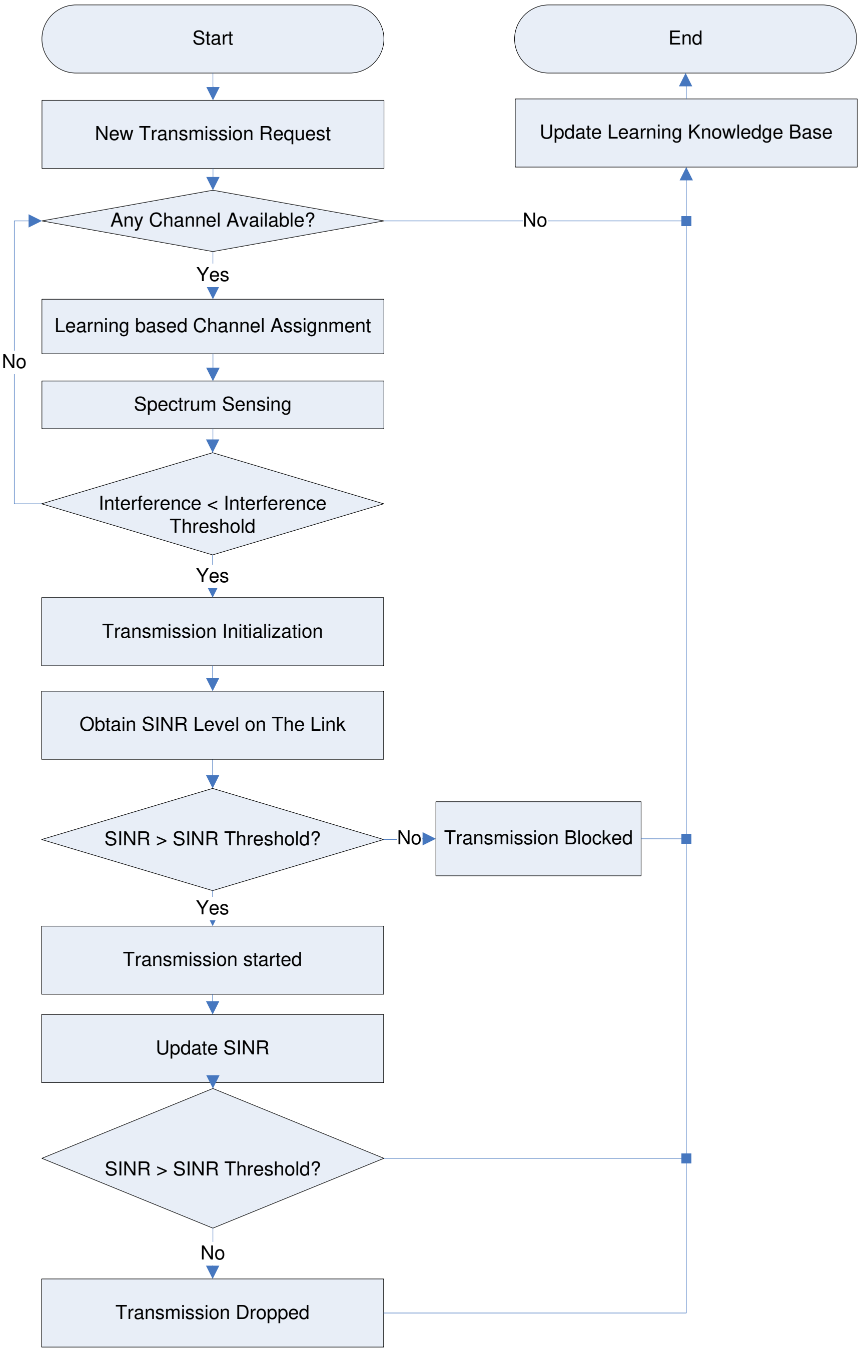


*Figure 4.2 Reinforcement Learning-based Spectrum Sharing Algorithm*

The CR users follow the above steps in every transmission process. One condition applies to the system that $N(U_i)<N_{max}$, $N(U_i)$ denotes the number of sensed channels of $U_i$ in each activation and $N_{max}$ is the maximum number of channels which a CR user is allowed to scan in a single activation. If $N(U_i)>N_{max}$, and $U_i$ is still searching for an unoccupied resource, it is blocked and waits for the next activation. It is unrealistic to allow users to keep sensing and searching for a better resource without a time limit, because sensing is a power-intensive and time-consuming process.

## 4.6 Simulation Scenarios

Fig 4.3 is an example layout of the CR nodes. We use a basic transmitter-receiver pair communication system model because we try to focus on the behavior of CR users and consequently achieve a deep understanding of such complex behavior. We believe the technique is widely applicable for other system models. The Okumura-Hata propagation model [85] is used along with log-normal shadowing with a standard deviation of 8 dB. CR pairs are uniformly distributed on a square service area. An event-based scenario is employed in our work, and at each event a random subset of pairs are activated, system parameters used in this work is shown in table 4.2.

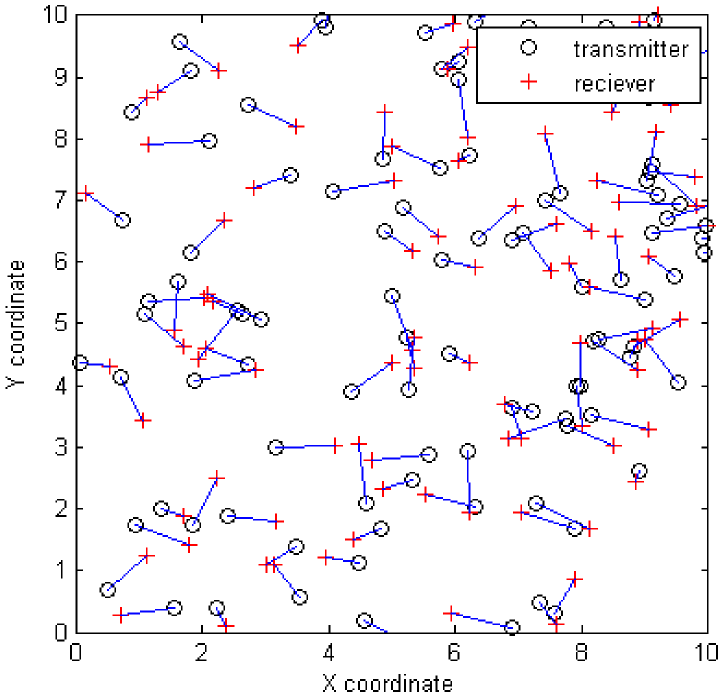


*Figure 4.3 Sample of Spatial Layout of Cognitive Radio Pairs for Simulation*

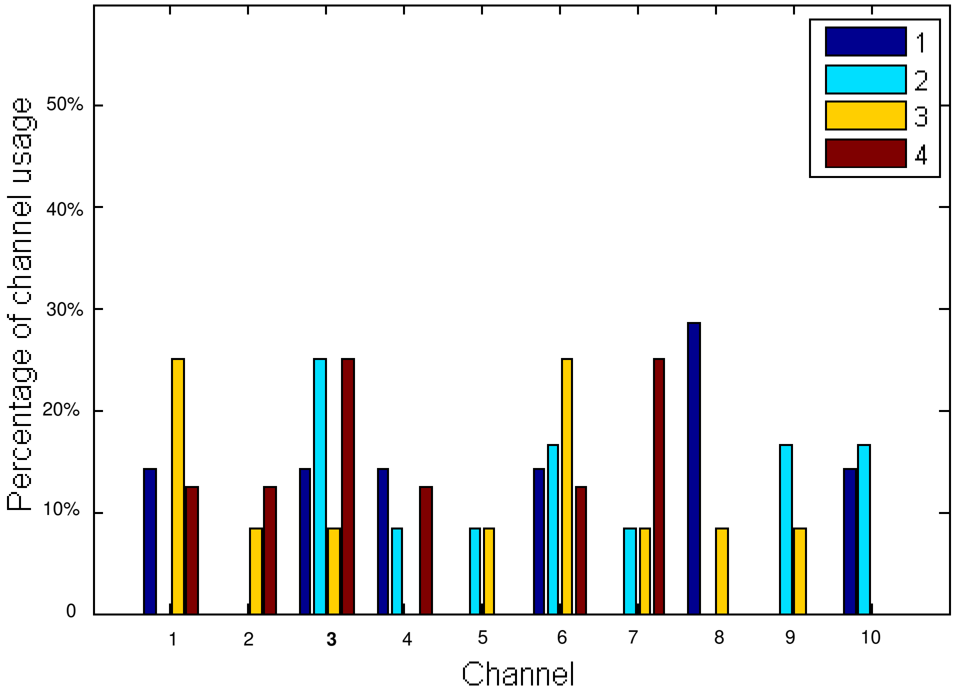
*Table 4.2 Simulation Parameters*

| Parameter | Value |
|---|---|
| Service Area | 1000km$^2$ |
| Number of pairs | 1000 |
| Maximum number of activated users | 400 |
| Link Length | 200m – 1500m |
| Transmitter Antenna gain | 0 dBi |
| Interference threshold | -40 dBm |
| *SINR* threshold | 10 dB |
| Noise floor | -137dBm |

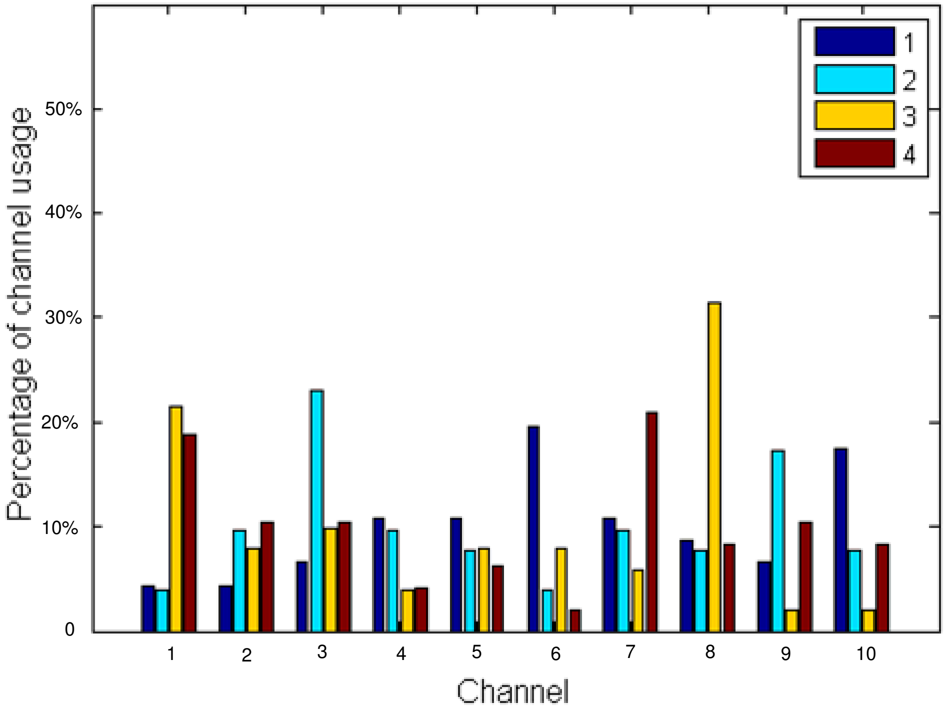
## *4.7 Results and Analysis*

### 4.7.1 Channel Partitioning by Reinforcement Learning

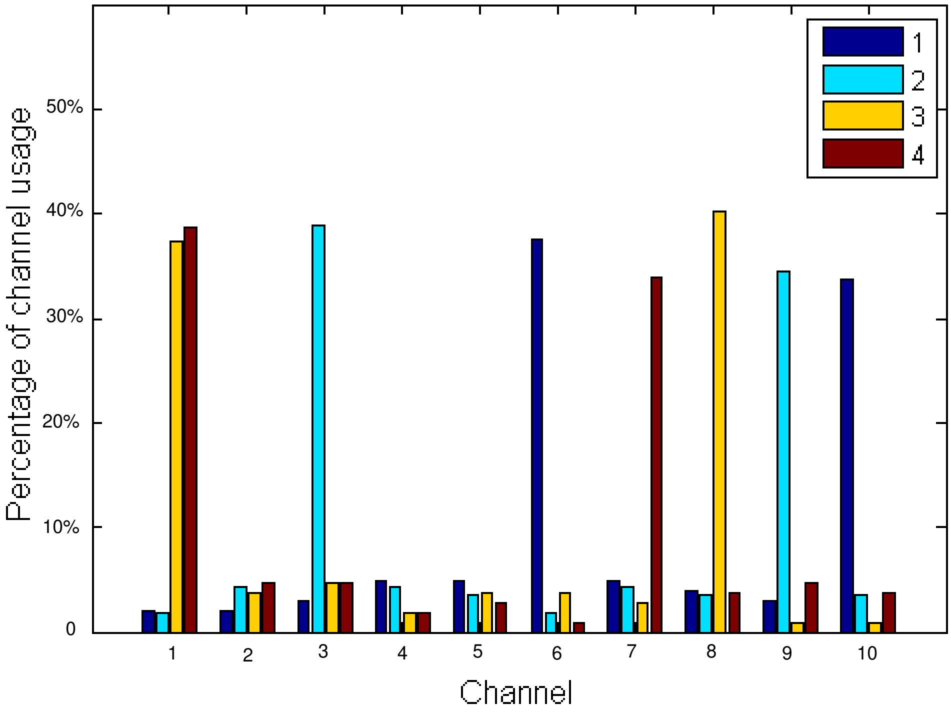
A quick and efficient channel partitioning is the most desirable result in our work since it will promote more efficient and reliable communications. The available spectrum will be partitioned autonomously by individual reinforcement learning and therefore CR users are able to avoid improper spectrum. Figure 4.4 (1)-(4) represent how the channel partitioning emerges during the simulation. A small number of 10 is used in this simulation to define the number of available channels and the number of users, because in this way the channel partitioning can be illustrated directly. We randomly choose 4 users out of 10 and number them 1-4 at the beginning of the simulation. By recording the channel usage of those 4 pairs, the channel usage of those pairs during the simulation can be obtained.



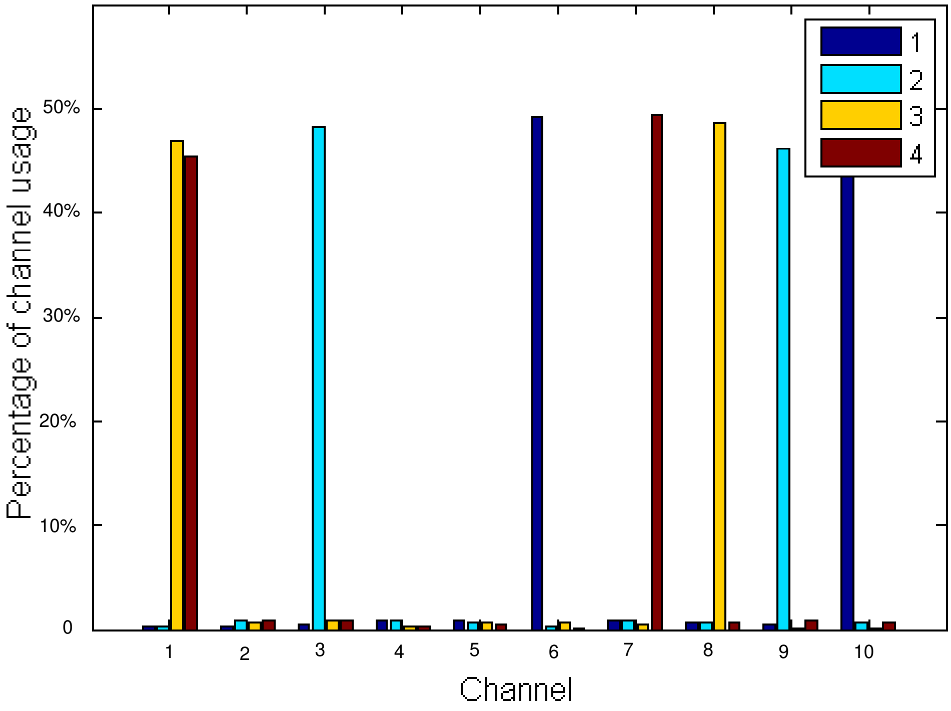
*(1) Event 50*



*(2) Event 100*



*(3) Event 500*



*(4) Event 100*

*Figure 4.4 Channel Usage at (1) Event 50, (2) Event 100, (3) Event 500, (4) Event 1000*

At the beginning of the simulation (Figure 4.4 (1)), CR users use almost all resources equally. After a certain simulation time, at event 100 (Figure 4.4 (2)) a few channels already show their priority to certain users, like user 3 prefers channel 8 and user 2 prefers channel 3. However, the channel usage of user 1 is still fairly equal at this stage. The channel usages at event 500 and event 1000 are shown in Figure 4.4 (3), Figure 4.4 (4) respectively. It can be seen that a spectrum sharing equilibrium is established and therefore the channel usage converged to few preferred channels. The CR users are able to avoid collisions by utilizing their experience from learning consequently.

The behaviour of user 1 in this case clearly illustrates how the learning-based autonomous channel partitioning works. At the beginning, user 1 preferred to use channel 8 where 30% of the activations of user 1 succeeded in this channel. Between event 50 and event 100, communication failures happen on channel 8. User 1 remembered that and tried to avoid this channel thereafter. The channel usage then converged to channel 6 and 10 where user 1 had a better opportunity to successfully transmit data.

### 4.7.2 System Performance

Fig 4.5 – Fig 4.6 illustrate the performance of schemes which we discussed above. Blocking probability is measured at regular points in the service area and a Cumulative Distribution Function (CDF) of system blocking probability at these points is derived. In order to analyse the level of system interruption, a CDF of dropping probability is calculated at the same time. All CR users' parameters are exactly the same for each scheme evaluation, with different system performance being caused only by different weighting factor values.

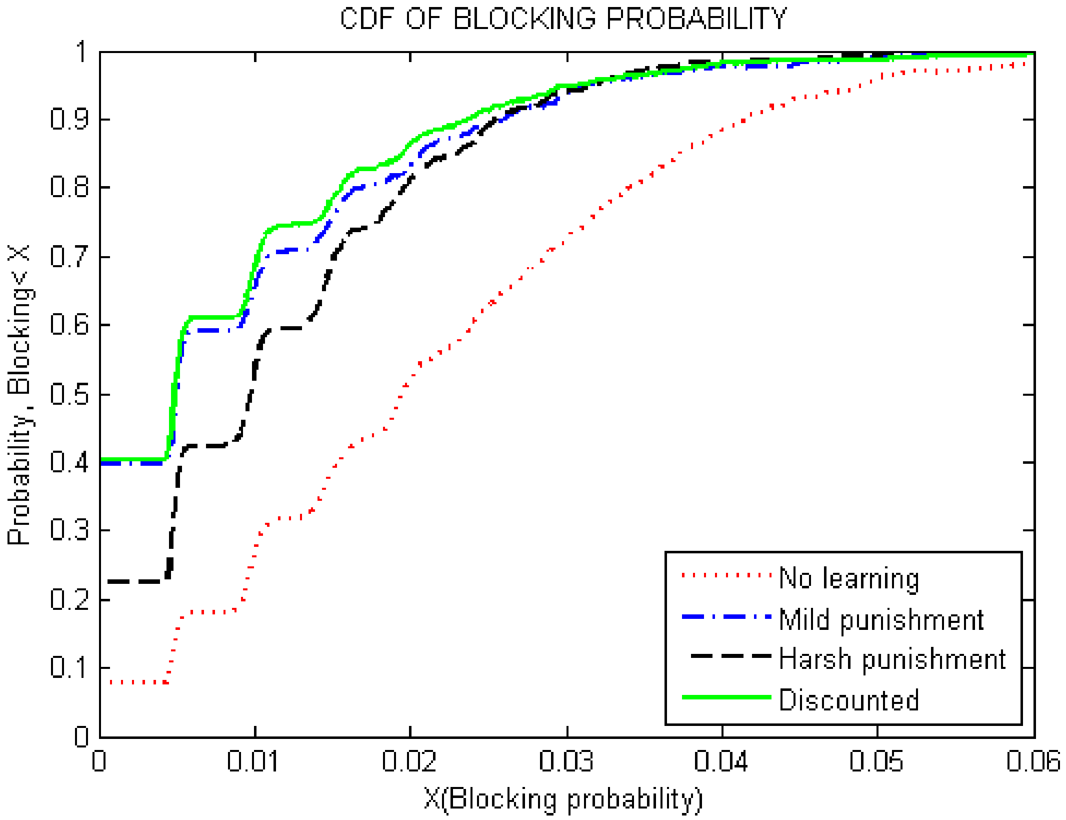


*Figure 4.5 Cumulative Distribution Function of System Blocking Probability at Discrete Points over the Service Area*

Fig 4.5 shows the CDF of system blocking probability of the three learning schemes along with a lower bound performance of random spectrum sharing without reinforcement learning. Comparing with the red dotted line which is the CDF of the no learning scheme, the blocking probability of our reinforcement learning spectrum sharing schemes are much lower than the scheme without learning. About 90% of users blocking probability in the discounted scheme are below 0.02, but in the no learning scheme only 50% users are able to meet this requirement. By using a reinforcement learning way to share spectrum, the blocking probability can be significantly reduced. It can be seen that the discounted scheme has the best performance in Fig 4.5. The overall blocking probability of the discounted scheme is about 40% of that of the no learning scheme. The blocking probability of the mild punishment scheme is slightly higher than the discounted scheme. This is because of the setting of punishment value. We believe that the value of weighting factor reflects the degree of the reaction of a user to a specific action. The higher the value is, the higher the degree is. In the discounted scheme, the weight of an unsuccessful channel is reduced by a certain percentage at each time. According to equation (4-10) if the request for a channel has been refused *n* times, the weight of that channel is:

$$W_t = f_1^{\,n} \cdot W_{t-\mathrm{n}} \tag{4-10}$$

If a user in the mild punishment scheme is in the same situation, the weight of the unsuccessful channel will be:

$$W_t = W_{t-n} - n \tag{4-11}$$

Take n = 3, $W_{t-n} = 100$ for example, we assume that 100 is the highest weight of all available spectrum for a CR user. After the best channel has failed to communicate three times, the weight of that channel $W_t$ in the discounted scheme is 12.5, the channel probably no longer at the top of the priority list for the CR user. But in the mild punishment scheme the weight $W_t$ is 97, it is still high enough to maintain its position as a good channel for the user. Since the reaction of the discounted scheme towards a communication failure is stronger and quicker than that of the mild punishment scheme, the performance of the discounted scheme is better.

Nevertheless the punishment factor is not the higher the better. The black dashed line is the CDF of the harsh punishment scheme. In this scheme the weight of the unsuccessful spectrum is directly decreased to zero but the system blocking probability is still higher than the discounted scheme. This is because the 'over-reactive' behaviour of the harsh punishment scheme. If a spectrum sharing scheme sets a punishment factor overly severe, the results of learning could be significantly changed by a rare occurrence. In the results of simulation, the best performance is achieved by the discounted scheme.

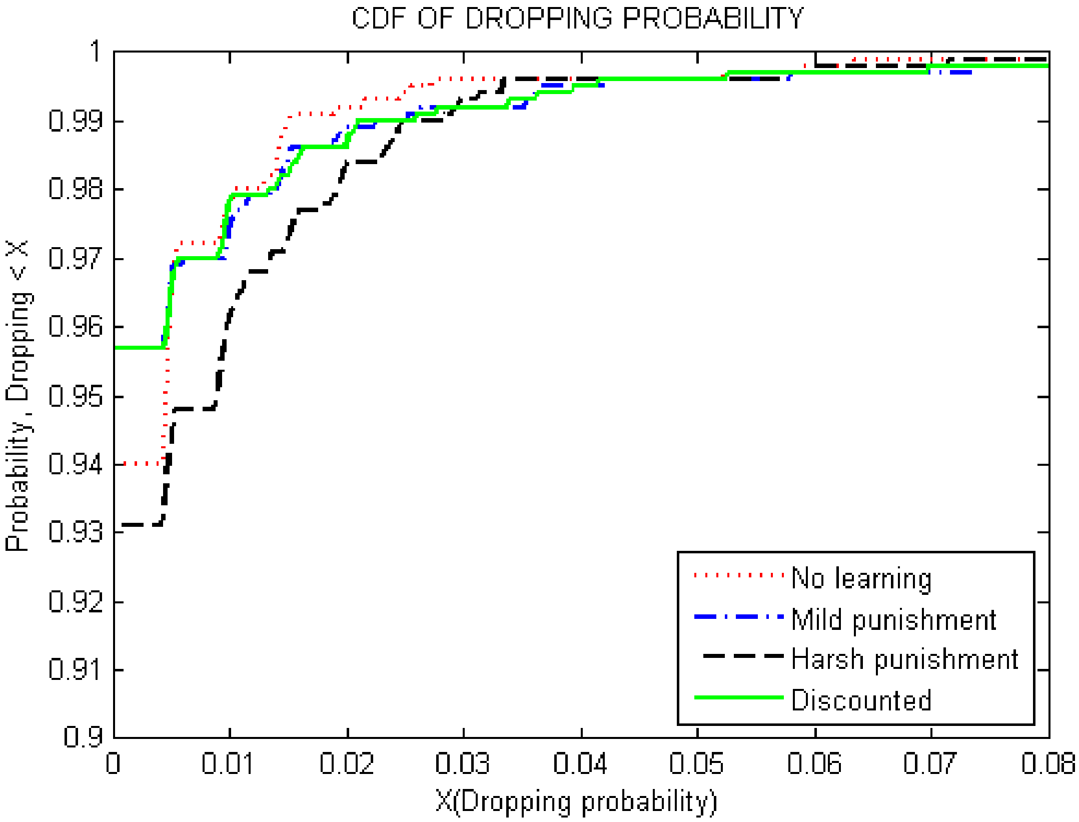


*Figure 4.6 Cumulative Distribution Function of System Dropping Probability at Discrete Points over the Service Area*

It can be seen that in every reinforcement learning scheme there are about 5% of users whose blocking probability is above 0.03. The performance of blocking probability of these users is difficult to improve no matter how the system defines the weighting factors, because these users are located at a high user density area and the opportunity for these users to successful set up a communication link is limited.

Fig 4.6 illustrates the CDF of dropping probability which demonstrates the level of system interruption. It shows that about 93% users are never dropped by system throughout the simulation. Since our schemes only take advantage of the information of system blocking to update the weights of spectrum, the performance of reinforcement learning-based scheme is no longer better than the no learning scheme. On the contrary, the dropping probability of the no learning scheme is lower than learning schemes. This is because a few CR users regard the channels with high dropping probability as their preferred resources and keep using these channels as long as their blocking probability is low. Using the information of system dropping along with blocking to adjust weights may be a potential method to achieve a better system performance. Further work needs to be done to examine this argument.

Fig 4.7 and Fig 4.8 show the spatial plots of the no learning and discounted schemes respectively. Since the users in our scenario are spatially fixed, the blocking probability is strongly connected to the user density in a certain area. From Fig 4.7 and Fig 4.8 we can clearly see the improvement of system performance by applying the reinforcement learning. Not only the 'high blocking' area of no learning scheme is significantly reduced by the discounted scheme, but also the blocking probability of some 'red hotspot' regions are also decreased.

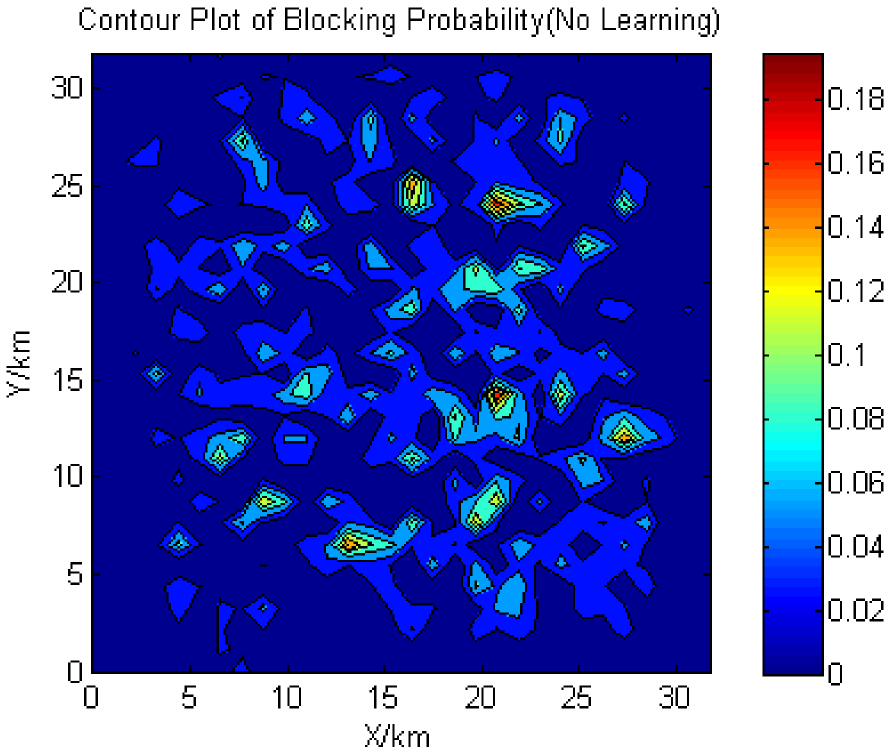


*Figure 4.7 Contour Plot of Blocking Probability of No Learning Scheme*

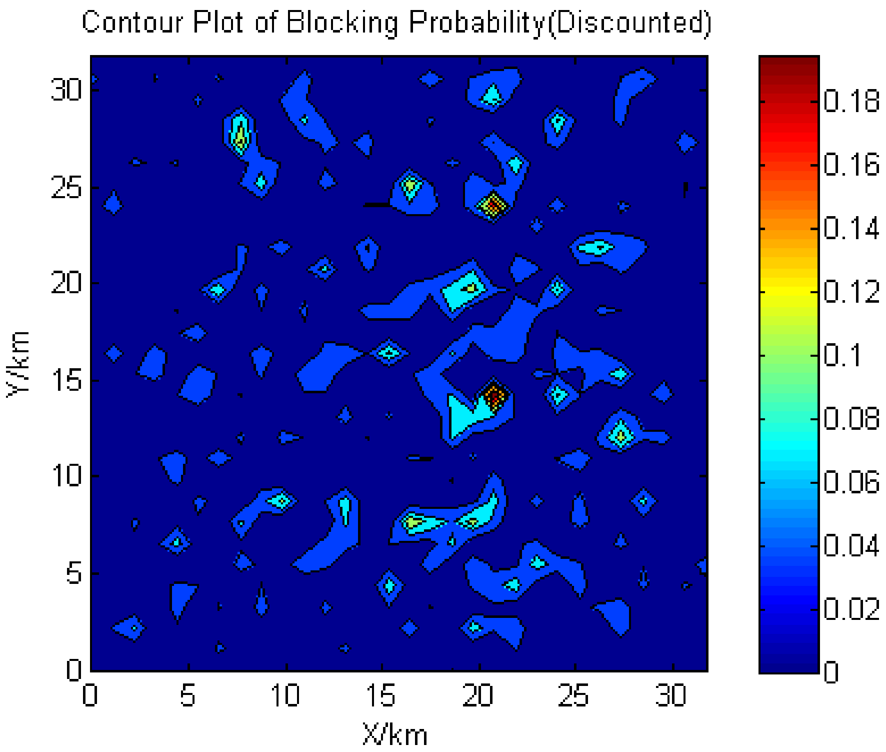


*Figure 4.8 Contour Plot of Blocking Probability of Discounted Punishment Scheme*

## *4.8 Conclusions*

In this Chapter, we introduced a reinforcement learning model for cognitive radio and a few basic reinforcement learning-based spectrum sharing schemes. By utilizing the ability of learning, cognitive agents can remember their preferred communication resources and enable an efficient approach to spectrum sensing and sharing accordingly.

Simulation results show that reinforcement learning-based spectrum sharing algorithms achieve a better system performance compared to non-learning algorithms. The weighting factors have a significant impact on the performance of the communication system. How to set the reward value is one of the key issues in the reinforcement learning scheme. Three different strategies on defining reward values have been investigated: the discounted scheme, the mild punishment scheme and the harsh punishment scheme. The results show that the discounted scheme achieves the best performance. Weighting factor values reflect the degree of the reaction of a user towards an action. The higher the reward/punishment value is, the stronger the reaction is. In our case, neither a mild reaction nor a harsh reaction achieves the best results. The system achieves better performance only if the reward value is assigned appropriately. From the measurements of system blocking and dropping probability, the performance improvements of applying our reinforcement learning scheme can be clearly seen. About 90% of users have a blocking probability below 0.02 in the discounted scheme, compared with a situation of 50% with the no learning scheme. The overall blocking probability of the discounted scheme is 60% lower than that of the no learning scheme. In addition, we have compared the system performance of different sets of reward values. About 90% users perform better in the discounted scheme than in the harsh punishment scheme. In this case, the scheme with a discounted punishment factor achieves the best performance.

# Chapter 5. Exploration Control for Reinforcement Learning-based Cognitive Radio

**Contents**

## *5.1 Introduction*

No matter when an agent learns to interact with an environment, two different tasks need to be carried out. The agent must firstly explore the action space, and then the actions discovered need to be exploited to gain enough experience. Neither of the two tasks can be performed exclusively in the learning process [92]. These two opposing tasks need to be combined in the learning process. The trade-off between exploration and exploitation needs to be more carefully controlled for an agent in order to efficiently learn from the interactions with a dynamic environment.

The trade-off between exploration and exploitation is seen as one of the fundamental challenges of reinforcement learning [34]. However, very few of the existing reinforcement learning algorithms for cognitive radio tackle this challenge. A learning CR needs to explore the wireless environment to find available resources. Meanwhile, the CR also has to exploit the resources discovered in exploration to obtain enough experience to distinguish between good and bad options. The trade-off between exploration and exploitation needs to be balanced in order to improve the performance of the CR system [89].

Thus, a two stage reinforcement learning-based algorithm is described in this chapter for a fully distributed scenario to balance the trade-off between exploration and exploitation [89]. A ‘warm up’ stage is proposed where distributed CR users search for optimum resources and learn from the experience of searching. Once users have obtained a set of preferred resources, they will only sense the spectrum with higher priority prior to establishing communications. The warm up stage is effectively the period of exploration for a CR user to discover new resources in our case. Therefore, the exploration phase becomes controllable by applying different warm up strategies. We will show how the balance between exploration and exploitation is not only theoretically important but also crucial to a CR system in practice.

More details of the exploration-exploitation trade-off are provided in section 5.2. Then the exploration control techniques are introduced in section 5.3, and the two stage algorithm is presented in section 5.4. After that, the simulation scenario and the results are discussed in section 5.5, 5.6 respectively. Finally, the conclusions are given in section 5.7.

## *5.2 Exploration Control Techniques for Cognitive Radio*

### 5.2.1 Warm-Up Stage and Preferred Resource Set

The trade-off between exploration and exploitation is one of the fundamental challenges of reinforcement learning. Exploring users are more likely to cause more disturbance, as their transmissions are more likely to interfere with hidden terminals. Thus, the exploration of the learning-based cognitive radios needs to be carefully controlled.

Our idea to solve this problem is to define a 'warm up' stage and a preferred resource set. 'Warm up' is a stage where distributed CR users search for available actions and learn from the experience of searching. In the warm up stage, agents explore the available spectrum pool by accessing all physical resources with equal probability. The weights of the used resources will be modified after every action. In other words, in the warm up stage an agent updates the knowledge base constantly but uses a random action policy in figure 4.1.

We define a specific threshold such that if the weight of a used resource is above the threshold, the action of taking this resource is considered as a preferred action and the resource is selected into the preferred resource set. By playing the game repeatedly, an agent will obtain a full set of preferred resources. This period is practically the process of exploring for agents. Once a CR user finds a set of ideal resources, the exploration stage will be suspended and the user starts to exploit the set of preferred resources. By constantly taking the preferred actions according to the optimal action policy $\pi^*$, users obtain continuing feedback to verify whether the selected resources are the appropriate targets for themselves. Meanwhile, users who have already obtained their preferred resource set will move back to the warm up stage again when the weight of any preferred resource has decreased under the preferred channel weight threshold.

By adjusting the size of the preferred resource set and the value of the preferred resource weight threshold, the stage of exploration becomes controllable which means it is possible to balance the exploration versus exploitation trade-off in our scenario. Our simulation results will show clearly that the appropriate settings of warm-up and the preferred resource set can be crucial.

### 5.2.2 Two-Stage Reinforcement Learning-based Spectrum Sharing Algorithm

The two stage reinforcement learning based spectrum sharing algorithm with exploration control is illustrated in Fig.5.1. The cognitive radio users will firstly enter the warm-up stage to randomly explore the spectrum space. After a certain number of optimal resources have been discovered, the user will then exploit these optimal resources only. The different spectrum sharing strategies applied in the warm up and exploitation stages are highlighted in the flowchart to help readers to gain a thorough understanding. Note that no modification has been made to the reinforcement learning model and the value function. The learning part of the algorithm remains the way as it was introduced in section 4.2 and 4.3.

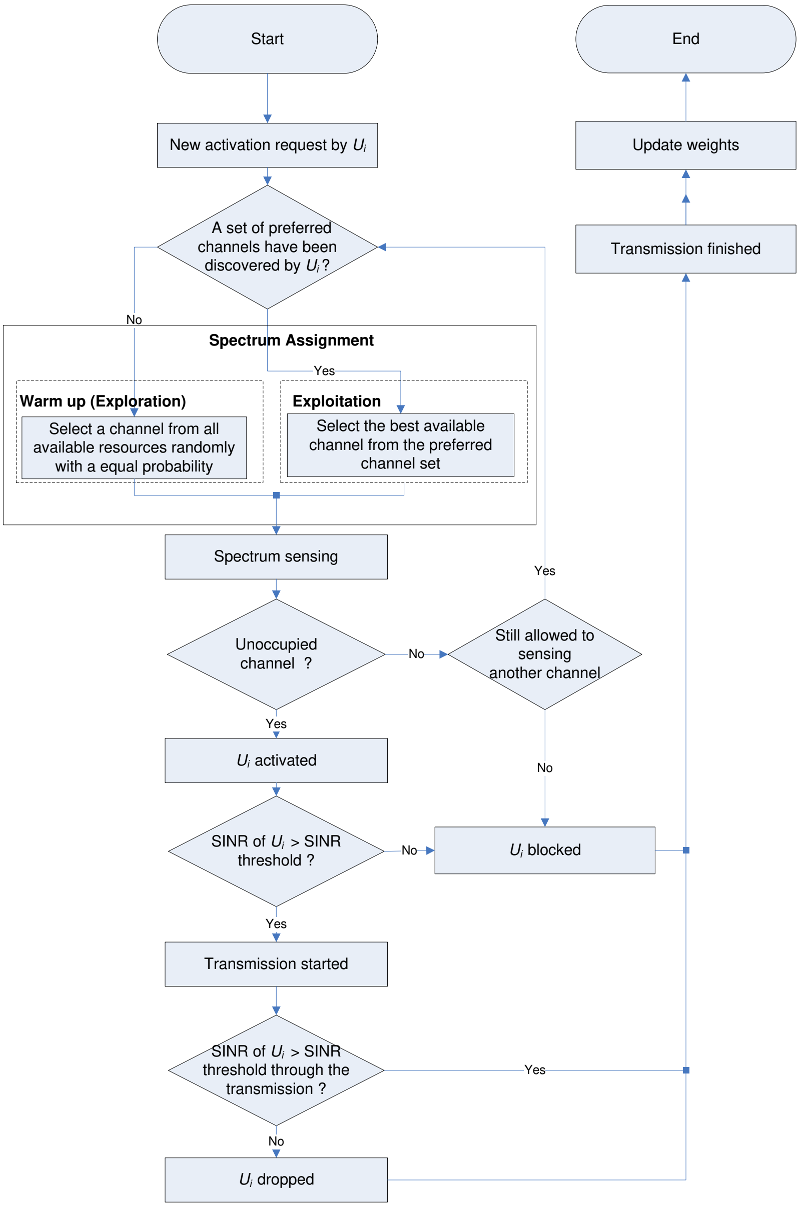

Start
End
New activation request by $U_i$
Update weights
A set of preferred channels have been discovered by $U_i$ ?
Transmission finished
No
Spectrum Assignment
Yes
Warm up (Exploration)
Select a channel from all available resources randomly with a equal probability
Exploitation
Select the best available channel from the preferred channel set
Spectrum sensing
Yes
Unoccupied channel ?
No
Still allowed to sensing another channel
Yes
No
$U_i$ activated
SINR of $U_i$ > SINR threshold ?
No
$U_i$ blocked
Yes
Transmission started
SINR of $U_i$ > SINR threshold through the transmission ?
Yes
No
$U_i$ dropped

*Figure 5.1 Algorithm Flowchart*

CR user is denoted as $U_i$, $U_i \in U$ and $U$ is the CR user set. The steps of our algorithm are given as follows.

**Step 1: State evaluation.** In this step, $U_i$ evaluates its own local system state (warm-up stage or exploitation stage). In this case, it is whether $U_i$ has found its preferred resource set. A preferred resource weight threshold ($W_{thr}$) has been defined and $U_i$ compares the weight of the used channel with $W_{thr}$ at every communication request. If the weight is above $W_{thr}$ , $U_i$ considers the resource as a preferred channel and this channel is selected to the preferred resource set. If the preferred channel set of $U_i$ has been filled with suitable channels, $U_i$ will be considered in the exploration. Otherwise it remains in warm-up stage.

**Step 2: Spectrum Selection.** Depending on the result of the evaluation in step 1, there are different rules in this step:

If $U_i$ is still in the warm-up stage, it chooses a channel randomly from the spectrum pool. $U_i$ senses the interference level on that channel. If the interference level $I$ of the channel is below the interference threshold $I_{\mathrm{thr}}$, $U_i$ is activated. Otherwise the weight of the spectrum is decreased and $U_i$ starts with a new channel again. If $U_i$ is in the main stage. $U_i$ senses the spectrum in their preferred resource set according to the action policy $\pi$.

**Step 3: *SINR* measuring.** After step 2, the existing users within the same channel can measure the *SINR* at their receivers. The purpose of measuring *SINR* is to maintain the communication quality of the channels. We set up a *SINR* threshold $SINR_{thr}$. If the *SINR* of the activated pair $U_i$ is greater than the threshold ($SINR_i > SINR_{thr}$), $U_i$ successfully uses the spectrum and the weight of the channel will be increased by a reward. If $SINR_i < SINR_{thr}$, $U_i$ is blocked by the channel and the weight is updated with a punishment. In addition, according to the measurement of *SINR* of the existing users, the existing users whose *SINR* is decreased below the *SINR* threshold are dropped and the channel weight for these users are also decreased accordingly.

## 5.3 Simulation Scenarios

In this chapter, we keep using the basic transmitted-receiver pair communication system model introduced in section 4.6 because we try to focus on the autonomous behavior of the learning based CR users and consequently achieve a deep understanding of such behavior. Again we believe this technique is widely applicable to other system models. In fact, the techniques developed in this work have been successfully applied to multicasting wireless communication systems and CSMA based multiple access schemes. The fundamental issues of applying RL based techniques to CR discovered in this work have also emerged in the research work on similar topics in this field, e.g. exploration versus exploitation trade-off.

We are also interested here in an open spectrum scenario where the entire spectrum is fully shared, where radio regulations are sufficiently light-touch to give all services equal opportunity to use the spectrum. Such a scenario is seen today to a limited extent in the unlicensed bands.

The IEEE 802.22 standard is considered as a suitable basis to select parameters since it is the first wireless standard based on CR techniques [30]. Therefore, we link our modeling scenarios with IEEE 802.22 to the highest extent. The most commonly utilized Okumura-Hata propagation model is used along with log-normal shadowing with a standard deviation of 8 dB. The values of the parameters are shown in table 5.1. Most of the values are commonly used in this type of system. The carrier frequency is defined as 700 MHz to utilize the TV white space. The transmitter antenna height of 30 m is used to comply with both the requirements of proposed WRAN system (transmitter up to 30 m) and the Okumura-Hata model (which requires transmitters to be in the range 30-200 m) [93].

Moreover, one of the important topics of applying reinforcement learning to cognitive radio is to investigate how the users are able to avoid hidden terminals purely through reinforcement learning. Hidden terminals are the main cause of dropped calls and it is difficult to tackle the hidden terminal problems only by sensing in a fully distributed system. Therefore, a relatively high interference detection threshold of -30 dBm is applied in the simulation, meaning that we implement the communication system in an environment where there is intentionally higher dropping probability.

*Table 5.1 Simulation Parameters*

| **Parameter** | **Value** |
|---|---|
| Service area | 1000km$^2$ |
| Number of users | 1000 |
| Link length | 1km-2km |
| Number of channels | 100 |
| Carrier Frequency | 700MHz |
| Transmitter antenna height | 30m |
| Transmit power | 30dBm |
| Transmitter antenna gain | 0dBi |
| Receiver antenna gain | 0dBi |
| Bandwidth | 1MHz |
| Noise floor | -114dBm |
| Interference threshold | -30dBm |
| SINR threshold | 10dB |

Weighting factor values are shown in table 5.2. Based on the degree of success, either a reward or a punishment is assigned to the weight of the used spectrum. It can be seen in table 5.2 that the absolute values of the reward value and the punishment value are equal. The weight is increased or decreased by the same step size. Moreover, the size of preferred channel set is set to 5, which is 5% of the available resources and the preferred channel weight threshold is set to 5.

*Table 5.2 Weighting Factor Values*

| $f_1$ | | $f_2$ | |
|---|---|---|---|
| Reward | Punishment | Reward | Punishment |
| 1 | 1 | 1 | -1 |

## *5.4 Results and Analysis*

Figure 5.2 illustrates the Cumulative Distribution Function (CDF) of system blocking probability of the reinforcement learning scheme along with a lower bound performance of random spectrum sharing without reinforcement learning. The blocking probability has been measured individually at each transmitter pair in the service area and a CDF of system blocking probability of these pairs is derived. The individual blocking probability values at each user is thus able to show the spatial performance of users. Such information is important since the user density in a particular area could significantly influence the way that a local user behaves. On the contrary, the blocking probability at the system level does not deliver such information.

The X axis in figure 5.2 is the value of the blocking probability obtained from the model. The Y axis is the probability that $P\{x<P_B\}$ where $P_B$ is a given value of blocking probability. The overall blocking probability, the blocking probability in the warm up stage, and the blocking probability of the exploitation stage are plotted separately. It can be seen that the blocking probability in the warm up stage is approximately equal to that of the no learning scheme. This is because in the warm up stage, users pick channels in a random way which is the same as the no learning scheme. As soon as agents obtain their preferred resource set by learning and move to the exploitation stage, the performance is significantly improved. The black dashed line, which represents the blocking probability of the exploitation stage, achieves the best performance, when compared with the other lines in figure 5.2, and the overall performance of the reinforcement learning scheme is enhanced consequently.

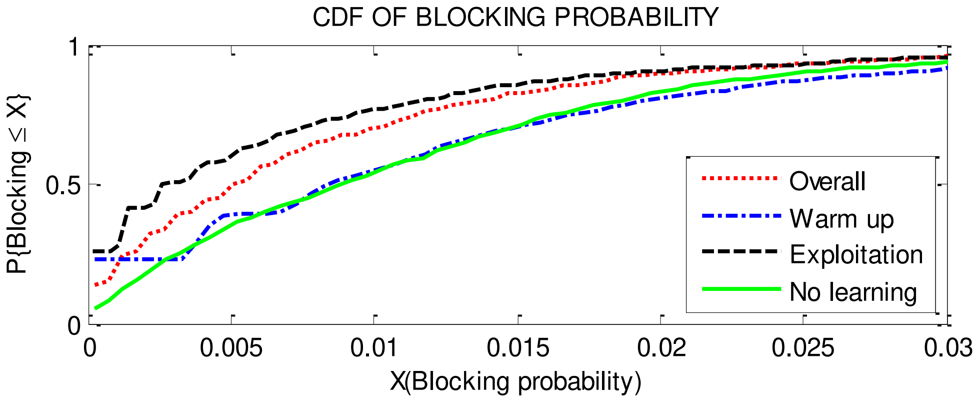


*Figure 5.2 Cumulative Distribution Function of System Blocking Probability of Transmitter and Receiver Pairs*

Since we use the information of local dropping and blocking to update the memory of the CR user, the system performance of dropping probability is also improved. The performance of the two stage algorithm in terms of dropping probability is almost identical to that of blocking. Figure 5.3 shows the CDF of dropping probability which illustrates the level of system interruption. The relative performance is similar to figure 5.2. The dropping probability in the warm-up stage and the dropping probability of the no learning scheme are on the same level and the overall dropping is greatly improved by reinforcement learning in the exploitation stage.

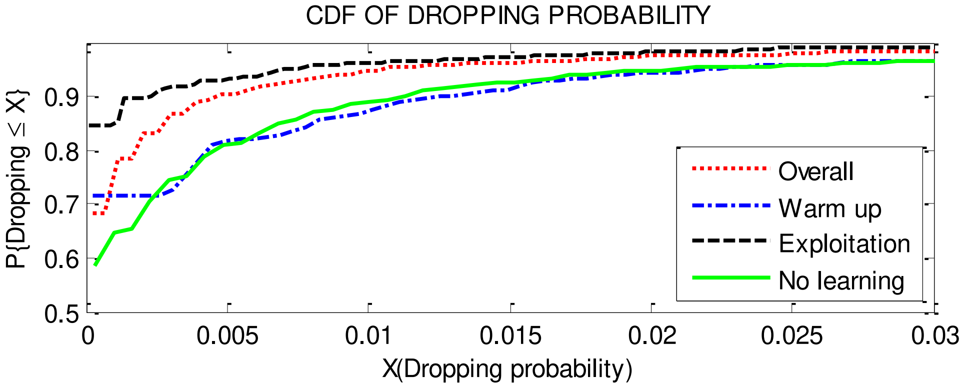


*Figure 5.3 Cumulative Distribution Function of System Dropping Probability of Transmitter and Receiver Pairs*

The problem of the trade-off between exploration and exploitation can be clearly seen in the results. The learning users cause a higher level of disturbance to the environment when they are exploring the spectrum space. In order to reduce transmission failures, users are required to utilize the results of exploration as early as possible, subject to finding suitably good channels. The channel usage of the $k^{th}$ user in channel $l$ at time $t$ is defined by the following equation:

$$u_{k,l}^{t} = \frac{N_{k,l}^{t}}{\sum_{j=1}^{N_c} N_{k,j}^{t}} \tag{5-1}$$

where $u_{k,l}^{t} \in U_k^t$ ($1 \leq l \leq N_c$, $N_c$ is the total number of available channels), $U_k^t$ is the channel usage measurements vector of user $k$ ($k \in \{1,2,\ldots,N_u\}$. $N_u$ is the total number of user) at time $t$. $N_{k,l}^{t}$ is the total number of activations of user $k$ in channel $l$ by time $t$. The measurement $u_{k,l}^{t}$ does not take into account the activation duration in this case.

$U_k^t$ is sorted from the highest to the lowest, showing the channel utilization of user $k$ in a descending manner. The sorted vector is represented by $U_{std}$. $U_{std}$ of $N_u$ users at different time $t$ are then shown in Figure 5.4. Only the channel usage in the top ten utilized channels are shown since it is sufficient to illustrate the users' behaviour. The figure effectively shows the distribution of channel usage of users from the most frequently used channel to the least used channel, which in turn shows how the activations of users are converging to their preferred channels. It is quite obvious that at the beginning of the simulation, the channel usage is almost equal which means users are trying different channels in the warm up stage. After CR users have found their preferred resources, gradually the usage converges to a few highly successful channels. About 33% of activations succeeded in the first tested channels after 2000 events. This figure becomes about 50% after 3000 events, which means half the communication requests of the whole system are successfully accepted in the best available channels for the users.

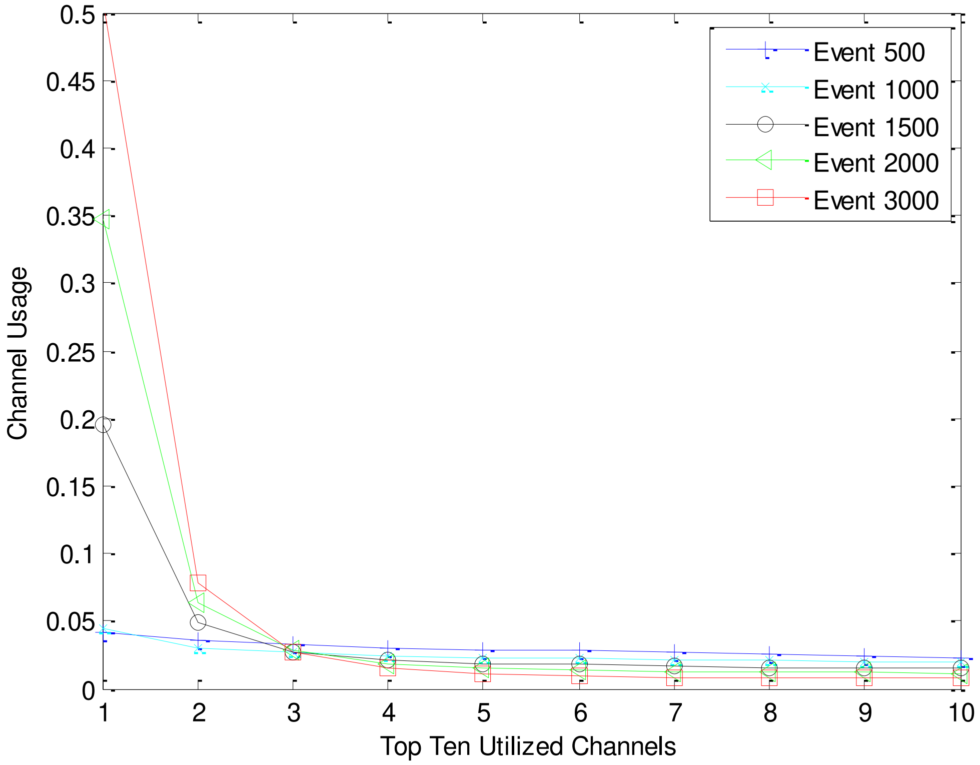


*Figure 5.4 Average Values of $U_{std}$ through Thousands of Events*

Figure 5.4 also shows the convergence behavior of our learning scheme. Like other learning algorithms for dynamic channel assignment, our scheme needs a sufficiently high number of trials to converge to its optimal state. From the start of the simulation to about event 1500, CR users found their preferred resources in warm up gradually, with our learning scheme converging to its ideal spectrum sharing strategy. The learning scheme will arrive at its spectrum sharing equilibrium after the majority of CR users have obtained their ideal channel set. The available spectrum has been autonomously partitioned by individual reinforcement learning consequently, and the users are able to avoid unsuitable channels by using their prior experience. How to obtain a quick and efficient convergence is crucial in this case. If we suppose all the activations will succeed in the first tested channel and purely consider the number of actions for a user to get a set of preferred resources, the number of actions which can be denoted by $N_{at}$ will be in a closed interval: $N_{at} \in [N_{at\min}, N_{at\max}]$ where $N_{atmin}$ is the minimum number of actions which a user is required to implement in order to obtain a full set of preferred spectrum and $N_{atmax}$ is the maximum number accordingly. $N_{atmin}$ can be defined as:

$$N_{at\min} = S_p W_{thr} \tag{5-2}$$

$S_p$ is the size of the preferred channel set and $W_{thr}$ stands for the preferred channel weight threshold. $N_{atmax}$ is defined as:

$$N_{at\max} = S_c (W_{thr} - 1) + S_p \tag{5-3}$$

where $S_c$ is the size of the available channel set. If a quick and efficient spectrum sharing equilibrium is desired, $N_{atmin}$ and $N_{atmax}$ need to be reduced appropriately. The methods which are investigated in this chapter are to adjust the settings of $S_p$ and $W_{thr}$. For instance, if we use $S_p$ = 5, $W_{thr}$ = 5 and $S_c$ = 100, therefore $N_{atmin}$ = 25 and $N_{atmax}$ =405 can be calculated by equations (5-2) and (5-3). A user will need at least 25 activations to obtain a set of preferred spectrum and a maximum of 405 activations from this point of view. If a smaller value 1 is used to define $W_{thr}$, $N_{atmin}$ will be 5 and $N_{atmax}$ will be 5 as well. The upper bound of the interval has been decreased by 94%. The users in this case will need only 5 activations to end the stage of exploration.

The warm up stage can be controlled by adjusting the size of the preferred channel set and the value of the preferred channel weight threshold. Figure 5.5 and figure 5.6 show the blocking probability and the dropping probability versus preferred channel weight threshold respectively. The size of the preferred channel set is fixed at 5 in the simulation. It can be seen that the blocking probability and dropping probability of the warm up stage remain at a high level due to the random action policy. The best overall performance is achieved by the lowest value of the threshold, indicating the invasive nature of the channel assignment selection and the unsuccessful utilization of channels particularly during exploration. The overall blocking probability is about equal to the blocking probability of the exploitation stage if the threshold is 1. The reason is quite obvious: the available spectrum pool has been partitioned immediately if a low threshold has been applied. Figure 5.7 illustrates the percentage of activations in the warm up stage and the exploitation stage versus preferred channel weight threshold. It can be seen that about 99% of activations are activated in the exploitation stage when the threshold is 1. A quick channel partitioning enables efficient spectrum sharing in this case.  Again, the behaviour of dropping probability of users is exactly the same as blocking and it can be explained the same way as above.

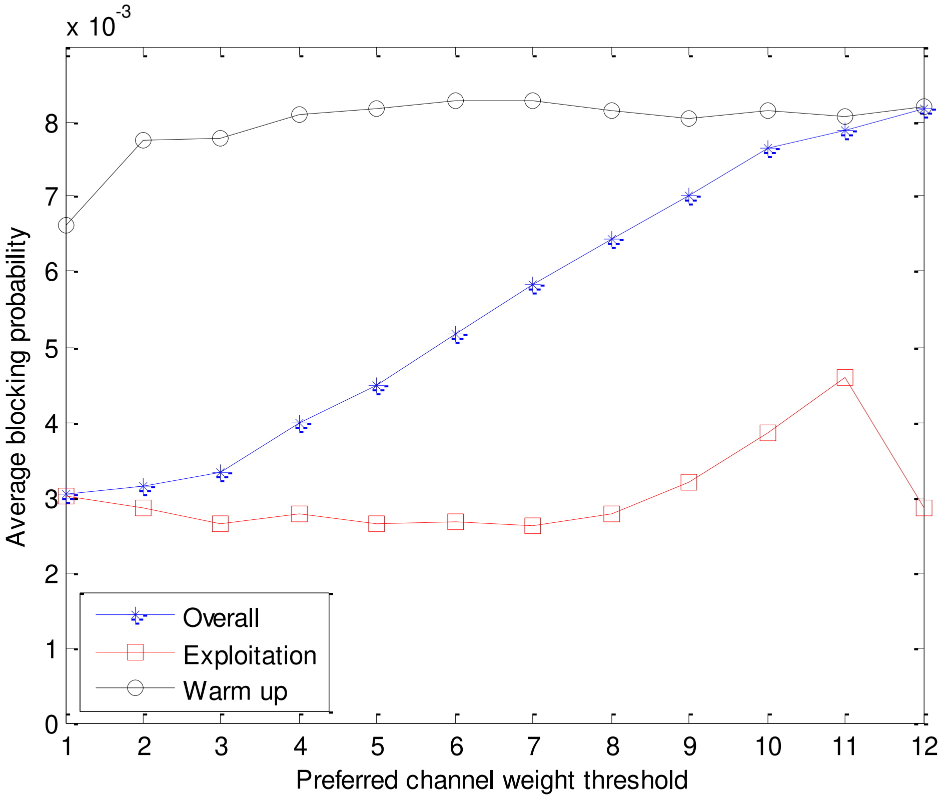


*Figure 5.5 Average Blocking Probability with Different Preferred Channel Weight Threshold*

The overall performance keeps rising if we increase the preferred channel weight threshold. This is because fewer and fewer users are able to obtain a set of preferred resources. It can be seen in Figure 5.7 that after the threshold of 12, the activations in the exploitation stage are very close to 0 which means users can hardly move into the exploitation stage. Therefore the overall performance in figure 5.5 and 5.6 are gradually equivalent to the performance of the warm up stage.

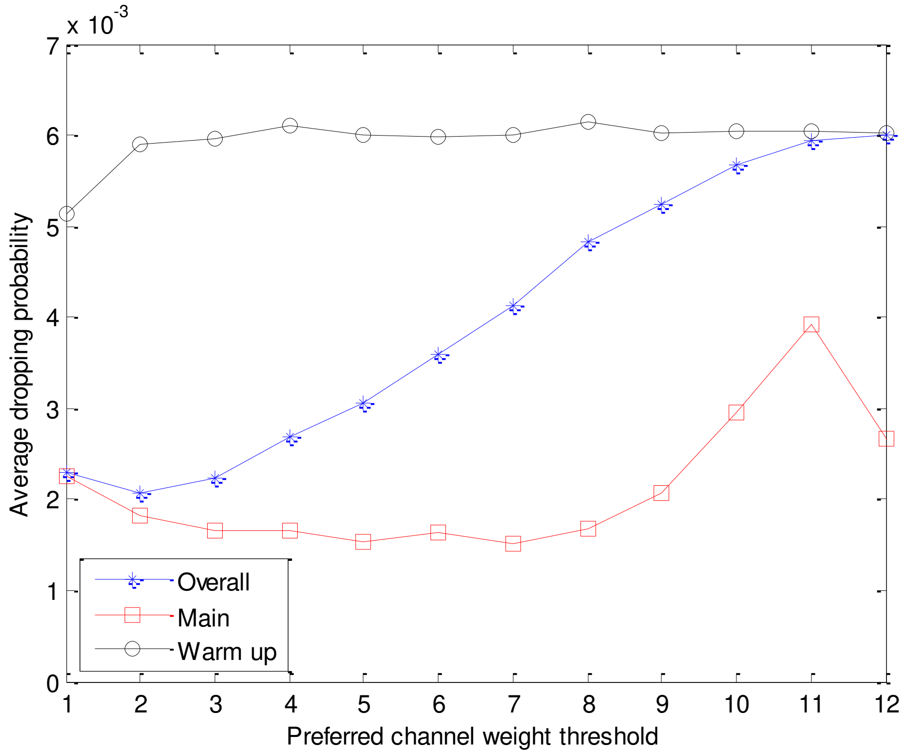


*Figure 5.6 Average Dropping Probability with Different Preferred Channel Weight Threshold*

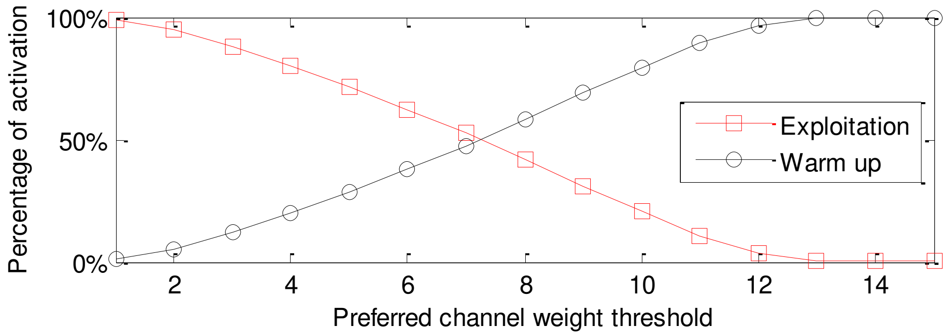


*Figure 5.7 Percentage of Activation with Different Preferred Channel Weight Thresholds*

The behaviour of the blocking probability in the exploitation stage changes accordingly. The red line rises when the weight threshold is increased to 11 since there are an increasing number of users exploring, and the users in the exploitation stage receive an increasing number of interruptions from those who are still searching for ideal spectrum. However, the blocking probability is decreased after the weight threshold is above 11. This is because only a very small number of users who are in very good locations where they receive much less disruption will be able to be activated in exploitation in this case, and the blocking probability of these well-located users is lower. It can be seen from figure 5.5 that when the threshold is above 12, the users are seldom activated in the exploitation stage due to an overly high weight threshold. Thus, the events that happened in exploitation are not statistically sufficient to show the users' behavior correctly for these higher weight thresholds in this scenario. The behavior of the dropping probability in the exploitation stage in figure 5.6 can also be explained the same way above.

Figure 5.8 shows the blocking probability when applying different sizes of the preferred channel set. The preferred channel weight threshold is fixed at 5 in the simulation. The blocking probability in the warm up stage remains at about 0.013 regardless the value of the size. The blocking probability of the exploitation stage, and the overall performance, is much greater than it is in warm up if the size is below 5. This is a result of the preferred channel set being relatively small, meaning that the alternatives for users are not sufficient. Therefore the ability of spectrum sensing is too constrained. Even though users are able to obtain a set of preferred resources fairly quickly by applying a small set size, the probability for them to stay in the exploitation stage is still very low. After the size of 5, the performance is relatively stable. With the capability of spectrum sensing and a sufficient set of ideal resources, the blocking probability can be significantly reduced.

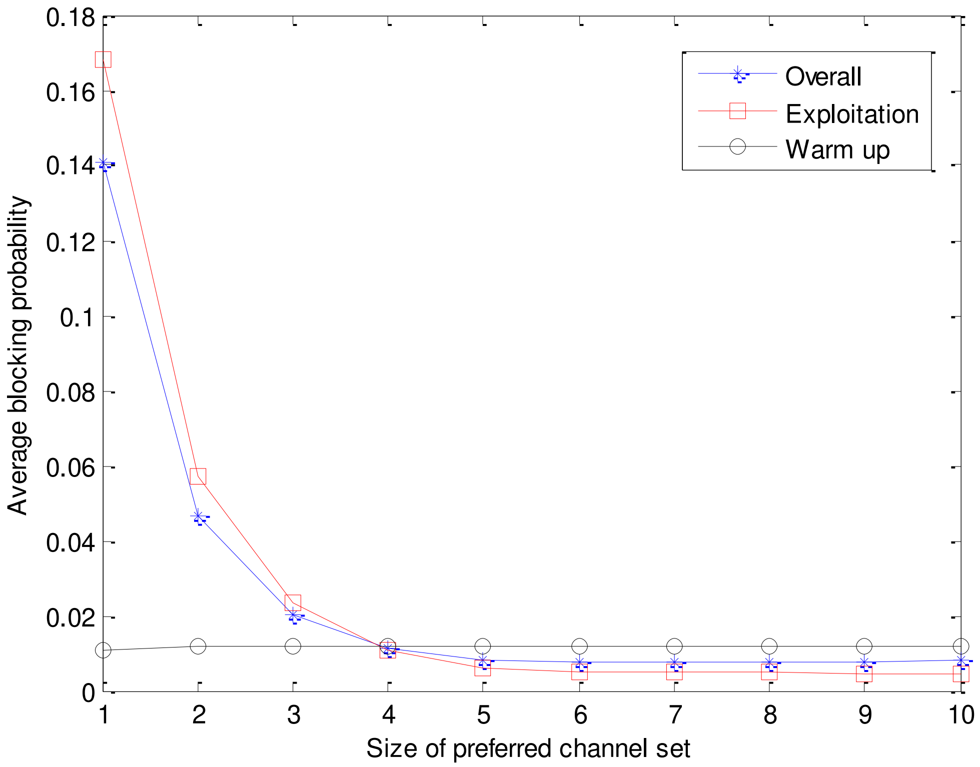


*Figure 5.8 Average Blocking Probability With Different Size of Preferred Channel Set*

The behavior of the dropping probability is shown in figure 5.9. The dropping probability of the exploitation stage is higher at the beginning because users in exploitation stage are experiencing a high level of interruption caused by users who are searching for ideal spectrum. Since the channels in the preferred channel set are insufficient, users are moved back to warm up frequently in this case. It can be seen that after a bigger size has been applied, the dropping probability in the exploitation stage will maintain at a low level which means users are able to avoid bad spectrum by using the prior experience. However, the overall dropping probability keeps rising and will be asymptotically equivalent to the dropping probability of the warm up stage because of the reduction of activations in the exploitation stage which are caused by the increase in the size of the preferred resource set.

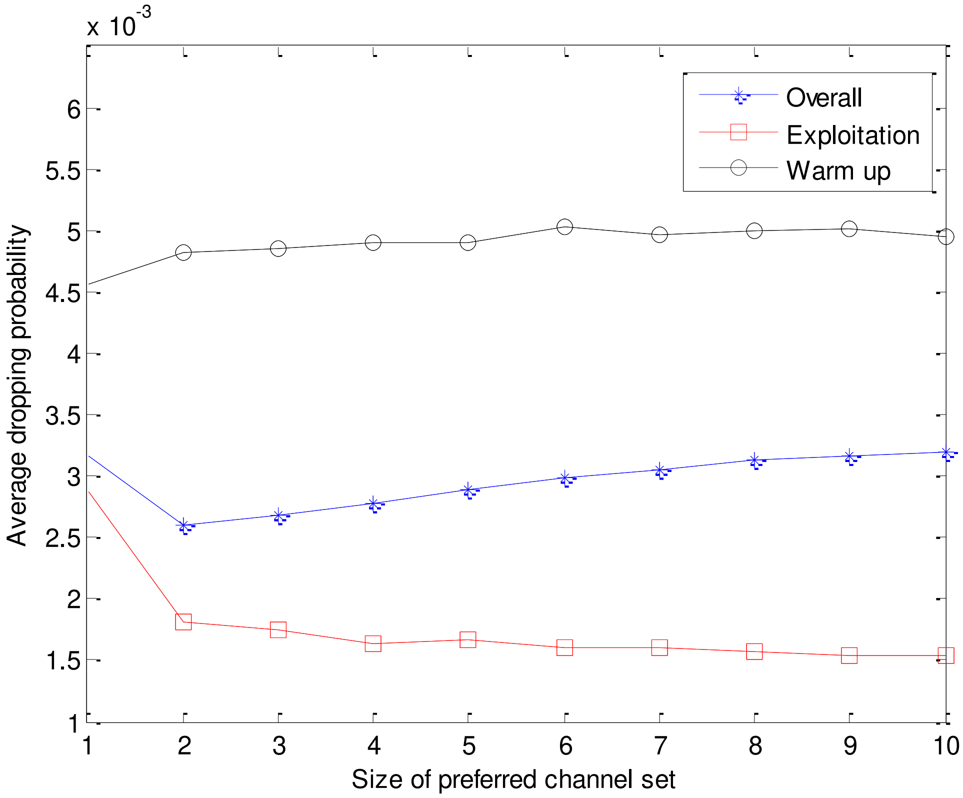


*Figure 5.9 Average Dropping Probability with Different Size of Preferred Channel Set*

## 5.5 Conclusions

A two stage spectrum sharing scheme for cognitive radio is introduced in this chapter. The described algorithm is able to practically control the exploration phase of the learning process. It shown that the trade-off of exploration and exploitation seen in reinforcement learning has a significant influence on RL-based communications system. Due to the hidden terminal effect exploring cognitive radios are likely to cause more disturbance compared with cognitive radios in the exploitation stage, which have successfully learned a preferred set of channels, enabling them to avoid excessive mutual interference.

A warm-up stage has been defined to practically control the phase of exploration in reinforcement learning process. By adjusting the settings of the preferred resource set, the trade-off between exploration and exploitation can be successfully balanced from the system performance perspective. It can be seen from the simulation results that a quick and efficient channel partitioning can be obtained by using a small preferred channel weight threshold. Moreover, either an overly small size of preferred resource set or an overly big size will cause more system interruptions rather than sharing spectrum peacefully, and an optimal spectrum sharing policy will not be discovered consequently. If a size of 1 is applied, the blocking probability is about 16 times higher than that of size 5. The dropping probability of the exploitation stage is also higher when an overly small size is applied. Both blocking probability and dropping probability decrease to a low level when a bigger size of preferred resource set is applied. By mapping memory to the action space and keeping a set of preferred channels, the channel usage of a user will converge to a few highly successful channels. About 17% of transmissions have been carried out in the best 5 channels after 1000 event. After event 3000, this figure is 70%, resulting in improved system performance.

# Chapter 6. Efficient Exploration for Cognitive Radio

**Contents**

## *6.1 Introduction*

Reinforcement learning is a learning approach which emphasizes individual learning from direct interactions with a dynamic environment. This distinct feature of reinforcement learning makes it perfectly suited to a distributed cognitive radio scenario. Very few of the existing reinforcement learning algorithms for cognitive radio address the issue of learning efficiency of the communication system. An agent will firstly explore the action space allowing the actions to be discovered which then need to be exploited to gain enough experience [92]. Practically, exploration is the process where the cognitive radio examines unused channels in the available spectrum pool. Cognitive radios will only use the channels discovered by exploration in the exploitation phase. The tradeoff between exploration and exploitation needs to be more carefully controlled for an agent in order to efficiently learn from the interactions with a dynamic environment. Previous research work in chapter 5 showed how the exploration versus exploitation tradeoff has a significant influence on system performance and how it is possible to practically control the exploration phase. Cognitive radio users will receive a higher level of interference when the users are exploring their available spectrum space since it is often necessary for a user to transmit on a channel in order to completely verify that its transmission can be received at a receiver. This exploration and potential interference does give rise to significantly better system performance in the exploitation phase since the behavior of users is more stable in this stage.

A basic two stage reinforcement learning-based spectrum sharing scheme is proposed in chapter 5. Cognitive radio users search for preferred resources and learn from the experience of searching in the exploration stage. Once users have obtained a set of preferred resources, the exploration stage is finished. Cognitive radio users will then move to the exploitation stage and only use the spectrum assigned a higher usage priority. This two stage algorithm is able to practically separate the exploration phase and the exploitation phase in the learning process, meaning that the exploration versus exploitation tradeoff is controllable. However, the fundamental exploration strategy we applied is still the most inefficient one – the uniform random exploration. Thus the efficiency of the exploration phase has the potential to be improved by applying more efficient exploration strategies.

This chapter introduces efficient exploration techniques for reinforcement learning-based cognitive radio. Two novel approaches are presented, pre-partitioning and weight-driven exploration, to enable efficient exploration in the context of cognitive radio. More importantly, the learning efficiency of a learning-based cognitive radio is defined and investigated. In the pre-partitioning scheme, users will randomly reserve a certain amount of spectrum resources before their transmissions start. The available action space which a cognitive radio needs to explore is then significantly reduced, which in turn shortens the exploration stage significantly. In the weight-driven exploration scheme, a certain level exploitation has been carried out in the exploration stage by applying a weight-driven probability distribution to influence action selection during exploration. Thus, exploration will be more efficient and the overall performance of the cognitive radio system can be improved.

In section 6.2 the general efficient exploration problem is described in the context of reinforcement learning. The efficient exploration techniques developed in this work, pre-partitioning and weight-driven exploration, are then introduced in section 6.3 as means of tackling this problem. The spectrum sharing algorithm is also introduced and the learning efficiency of the proposed approaches is investigated. In section 6.4 we examine the performance of the efficient exploration algorithms in more detail. The conclusions are drawn in section 6.5.

## *6.2 Efficient Exploration Techniques for Cognitive Radio*

The problem faced by all reinforcement learning-based cognitive radio systems is clearly illustrated in figure 6.1 and figure 6.2. These two figures show the system blocking probability and the system dropping probability achieved with our uniform random exploration algorithm. The system performance is worse in the exploration phase because the exploring users will cause more interference to the environment. A lower number of system interruptions are achieved in the exploitation stage since the channel usage of users converges to their preferred resources and the collisions are avoided. Therefore, an efficient exploration is highly desirable in order to reduce the exploration stage.

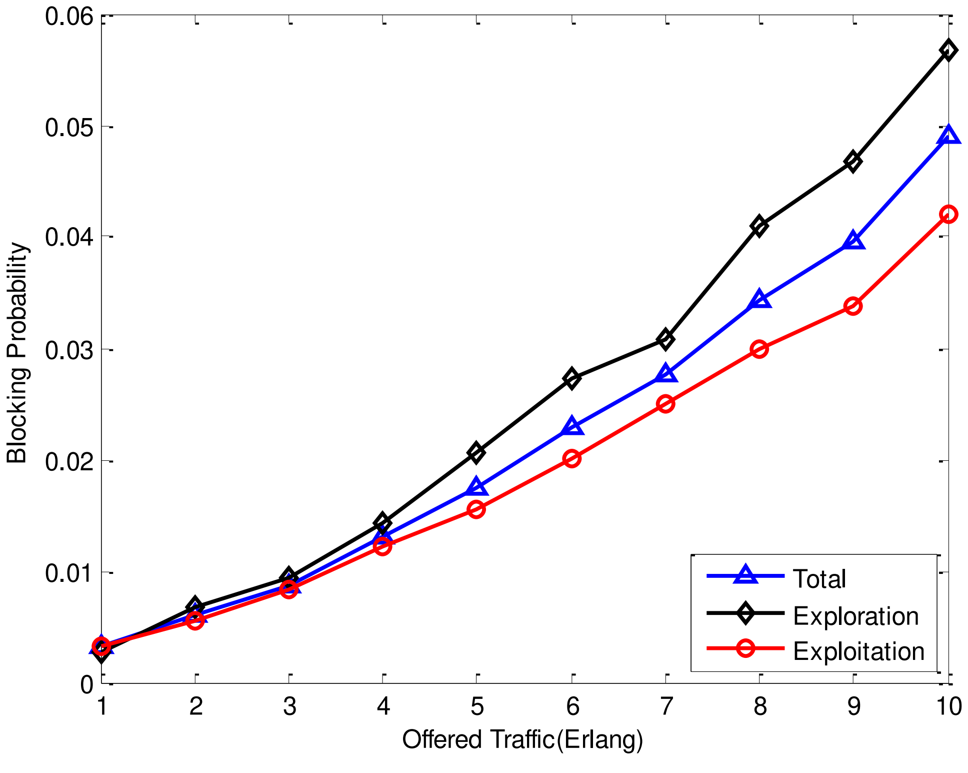


*Figure 6.1 System Blocking Probability of Uniform Random Exploration at Different Offered Traffic Levels*

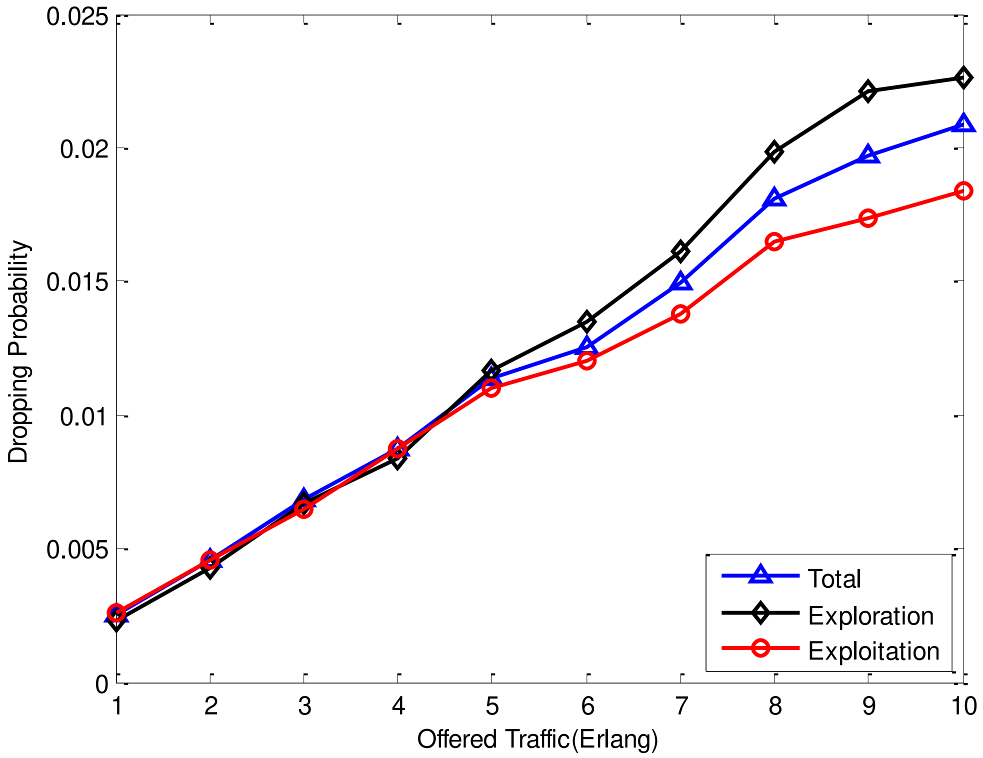


*Figure 6.2 System Dropping Probability of Uniform Random Exploration at Different Offered Traffic Levels*

It is crucial that CRs identify their preferred resources efficiently from the interactions with a dynamic radio environment. Thus, two efficient exploration techniques are proposed in order to accelerate exploration phase. In the pre-partitioning scheme [94], the potential action space of cognitive radios is reduced by initially randomly partitioning the spectrum in each cognitive radio. Cognitive radios are able to finish their exploration stage faster than more basic reinforcement learning-based schemes. In the weight-driven exploration scheme [95], exploitation is merged into exploration by taking into account the knowledge gained in exploration to influence action selection, thereby achieving a more efficient exploration phase. Furthermore, the learning efficiency in a cognitive radio scenario is defined and the learning efficiency of the proposed schemes is investigated.

### 6.2.1 Pre-Partitioning

Pre-Partitioning approach randomly partitions the available spectrum pool for all the cognitive radios. Each individual cognitive radio user reserves a certain number of channels and then select appropriate spectrum to transmit. Based on the level of success, the weight of the used resource is modified and then stored in the knowledge base, and this information will be utilized as guidance in selection of resource for future transmission.

The opportunity for a channel to be exploited by a cognitive radio is increased by pre-partitioning since the action space is reduced. Thus, cognitive radios are able to discover a number of preferred channels and move to the exploitation stage quicker, which in turn improves the system performance. The analytical results of this approach are given in section 6.2.4 along with the analysis of other approaches proposed in this chapter.

### 6.2.2 Weight-Driven Exploration

Most of the existing reinforcement learning-based algorithms for cognitive radio including our previous algorithms apply uniform random exploration strategy with uniform probability. Like a ‘uniform random walk’, conventional cognitive radio explores the available spectrum pool by accessing all resources with equal probability, regardless of the information gained by exploration. Research shows that the uniform random exploration is the most inefficient approach to achieve a goal [92].

As a result, a weight-driven probability distribution is proposed for the exploration process in this work to influence the action strategy by utilizing current weight information in exploration. Weights are values attached to a used resource and the values reflect the successful level of usage of this resource historically. Therefore, weights of used resources correspond to the historical information learned by cognitive radio users. Weight-driven exploration is a variation of Boltzmann exploration [92]. Boltzmann exploration uses a parameter called temperature (*T)* to control the probability of executing exploration. The difference is that in weight-driven exploration the temperature *T* is constantly changing. The weight-driven probability is defined as:

$$P(c) = \frac{w_c}{\sum_{c' \in C} w_{c'}} \tag{6-1}$$

where *P(c)* is the probability of a channel being selected. $w_c$ is the weight value of the channel at current state. *C* is the whole available resource space, *c'* is the channel in the available resource space. $w_{c'}$ is the weight of *c'* at the current state.

All the weights of resources will start with an equal value. Therefore, weight-driven exploration will start with a uniform random exploration at the first trial. After that, the exploration strategy is constantly modified by the weight-driven probability distribution. The higher the weight of the resource, the more likely the resource will be selected. On the one hand, the weight-driven probability distribution ensures exploration by bringing randomness into resource selection. On the other hand, the weight-driven probability distribution also utilizes the information gained in exploration to guide the exploration process itself. By applying exploitative information, exploration will be much more efficient.

### 6.2.3 Efficient Exploration based Cognitive Radio Spectrum Sharing Algorithm

Fig.6.3 shows the algorithm for fully distributed cognitive radio spectrum sharing, operating on individual transmitter-receiver pairs. The basic two-stage reinforcement learning based algorithm for cognitive radio proposed previously is used as a basis here. A preferred resource set is used to separate and control the exploration phase. A used resource is considered as a preferred resource when the weight of the resource is above a specific weight threshold, and it will be placed into a preferred resource set.

A user will firstly evaluate its own state at the beginning of a new activation. If the preferred resource set is fully occupied by good resources, the user will stay in exploration. Otherwise it will move to the exploitation stage. The spectrum assignment part in Fig.6.3 is highlighted in the flowchart to clearly show the different action strategies proposed in this thesis. Pre-partitioning and the weight-driven exploration are the approaches that can be applied to the spectrum assignment part in the exploration stage. In the exploration stage, a cognitive radio searches for available resources and learns from the experience of searching. After the preferred resource set is fully occupied by good resources, exploration will be suspended and the user will exploit resources in the preferred resource set only. A user will start to explore again if any of its preferred resources are no longer suitable for transmission (if the weight of a channel is decreased below the weight threshold).

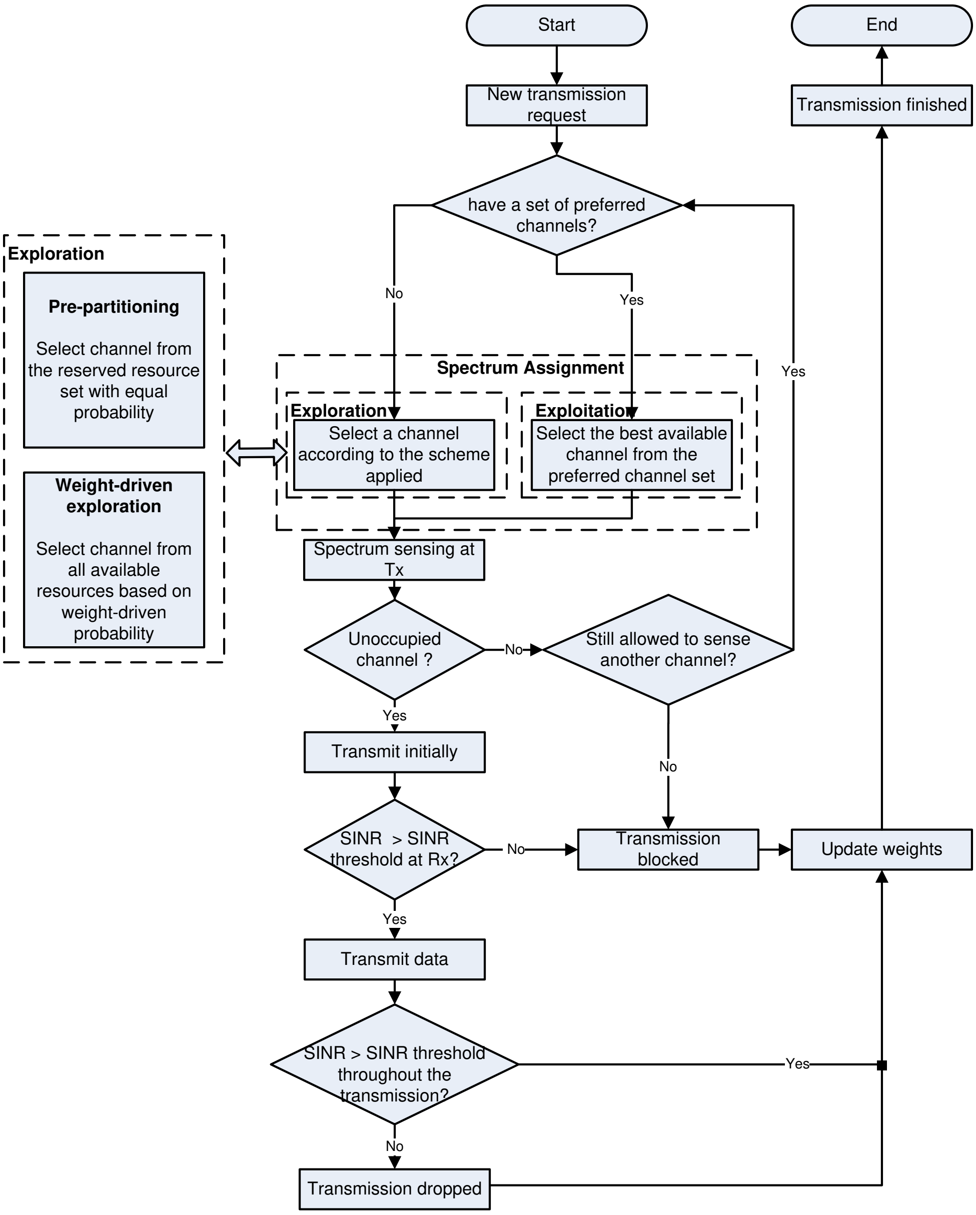

Start
End
New transmission request
Transmission finished
have a set of preferred channels?
No
Yes
Yes
Exploration
Pre-partitioning
Select channel from the reserved resource set with equal probability
Weight-driven exploration
Select channel from all available resources based on weight-driven probability
Spectrum Assignment
Exploration
Select a channel according to the scheme applied
Exploitation
Select the best available channel from the preferred channel set
Spectrum sensing at Tx
Unoccupied channel ?
No
Still allowed to sense another channel?
Yes
Transmit initially
No
SINR > SINR threshold at Rx?
No
Transmission blocked
Update weights
Yes
Transmit data
SINR > SINR threshold throughout the transmission?
Yes
No
Transmission dropped

*Figure 6.3 Flowchart of Efficient Exploration Algorithm as Simulated for Individual Transmitter-Receiver Pairs*

According to the channel assignment strategy applied under current state, the user will make a decision on channel selection. After that, spectrum sensing will be performed on the selected channel. If it is an unoccupied channel, the transmission will start, subject to SINR measuring. The user will try to find another channel if the selected channel failed to meet the threshold in spectrum sensing. The weight of the channel will be updated according to the results of the user's activation in this channel.

### 6.2.4 Learning Efficiency

The dynamic nature of cognitive radio calls for an efficient learning process, maximizing the useful information gained by learning while minimizing the costs of learning. To provide a measure of how efficient the learning process is, we can define the learning efficiency as:

$$Learning\ \ Efficiency = \frac{Useful\ \ \ Learning\ \ \ Cost}{Total\ \ \ Learning\ \ \ Cost} \tag{6-2}$$

Where the total learning cost is the time consumed by a learning agent to finish a task, and the useful learning cost is the time consumed to exploit the optimal strategy only. In the cognitive radio spectrum sharing case, the total learning cost is the number of trials the cognitive radio uses to find the optimal channel, and the useful learning cost is the number of trials the user uses to exploit this optimal channel. Thus, the learning efficiency for cognitive radio spectrum sharing can be defined as:

$$Learning\ \ Efficiency = \frac{Useful\ \ \ Trials}{Total\ \ \ Number\ \ \ of\ \ \ Trials} \tag{6-3}$$

The number of useful trials can be obtained by the equation as follows:

$$V_T = \sum_{n=1}^{N_E} r_n \tag{6-4}$$

where $V_T$ is the targeted weight value, and a channel is considered as an optimal resource when the weight of a channel $w_c$ is equal to $V_T$. $n$ is the number of trials and $N_E$ is the useful trials used to exploit the optimal resource. $r_n$ is the reward received at each trial.

Equation (4-9) is used to update the weight values in this thesis and the accumulated weight value after $n$ trials is:

$$W_n = f_1^n W_0 + f_2 \sum_{m=1}^{n} f_1^{m-1} \tag{6-5}$$

$W_0$ is the initial weight of a channel. $f_1$ and $f_2$ are weighting factors introduced in section 2.5. Therefore equation (6-5) can be rewritten as:

$$V_T = f_1^{N_E} W_0 + f_2 \sum_{n=1}^{N_E} f_1^{n-1} \tag{6-6}$$

If $E[P(c)]$ is the expected value of the probability for a channel to be selected in each trial and the total number of trials that a cognitive radio uses to find the optimal resource is $N_T$ , $N_E$ can be obtained as:

$$N_E = N_T E[P(c)] \tag{6-7}$$

Equation (6-6) then can be written as:

$$V_T = f_1^{N_T E[P(c)]} W_0 + f_2 \sum_{n=1}^{N_T E[P(c)]} f_1^{n-1} \tag{6-8}$$

The total number of trials a cognitive radio requires to find an optimal channel (when the weight of the channel $W_c$ equals to $V_T$) can then be obtained from equation (6-8). The targeted weight value is effectively the preferred channel weight threshold in our algorithm. The influence of $V_T$ and how to define $V_T$ in a cognitive radio system have been investigated in our previous work [89].

It is possible to perform a basic analysis of learning cost of each scheme. To simplify the environment faced by the learning-enabled cognitive radio, we assume optimistically that all selected actions will succeed and the weight of the successful action will increase by 1 in each trial. Thus, $f_1$ and $f_2$ always equal to 1. We also assume that $W_0$ equals 0. Therefore, equation (6-8) can be written as:

$$
\begin{aligned}
V_T &= \sum_{n=1}^{N_T E[P(c)]} f_1^{n-1} \\
&= N_T E[P(c)]
\end{aligned}
\tag{6-9}
$$

$N_T$ then can be obtained as:

$$
N_T = \frac{V_T}{E[P(c)]} \tag{6-10}
$$

It is also very important to notice that by giving a fixed $V_T$, the higher the *E[P(c)]*, and the lower the $N_T$ . In other words, in order to find an optimal channel quickly, the expected value of the probability for a channel to be selected in each trial needs to be increased. The purpose of the proposed efficient exploration in this paper is to increase *E[P(c)]*.

In the uniform random scheme, the user accesses available channels with equal probability, and the probability for a channel to be selected in each activation can be calculated by:

$$
P_u(c) = \frac{1}{N_c} \tag{6-11}
$$

where $N_c$ is the total number of available channels. The probability for a channel to be selected in the pre-partitioning scheme can also be obtained:

$$
P_p(c) = \frac{1}{N_r} \tag{6-12}
$$

where $N_r$ is the number of channels in the reserved channel set. The decrease of the learning cost by pre-partitioning can be illustrated if we compare the learning cost of the uniform random exploration scheme and pre-partitioning scheme by the following equation:

$$
\frac{N_{T-prepartitioning}}{N_{T-uniform}} = \frac{N_r}{N_c} \tag{6-13}
$$

Thus, a small reserved channel set reduces the learning cost theoretically. However, an overly small value of $N_r$ may not enable good channels to be discovered from a radio system perspective, so the system performance may not improve by as much as we expect theoretically. This tradeoff affecting system performance of reducing the level of $N_r$ and obtaining good channels from the radio system perspective is discussed in more detail later.

Figure 6.4 compares the learning cost of two proposed schemes with the uniform random exploration scheme for a single cognitive radio user. The learning cost is effectively the number of trials taken in training. Obtaining an analytical expression for $E[P_w(c)]$ in the weight-driven exploration scheme is complex and beyond the scope of this thesis, since the probability of selecting a channel changes in every trial. Moreover, the probability distribution also changes according to equation (6-1). Therefore, figure 6.4 only includes results obtained by simulation. The theoretical results are calculated by the equations above. $W_0$=0, $N_c$=100 and $N_r$=30, with the same values used in the simulation. The number of trials the agent used to find the best available channel can also be obtained at different targeted weight values. The reduction in the learning cost as a result of pre-partitioning and weight-driven exploration can be clearly seen from this figure. Thus, the proposed exploration techniques are significantly more efficient than the uniform random exploration.

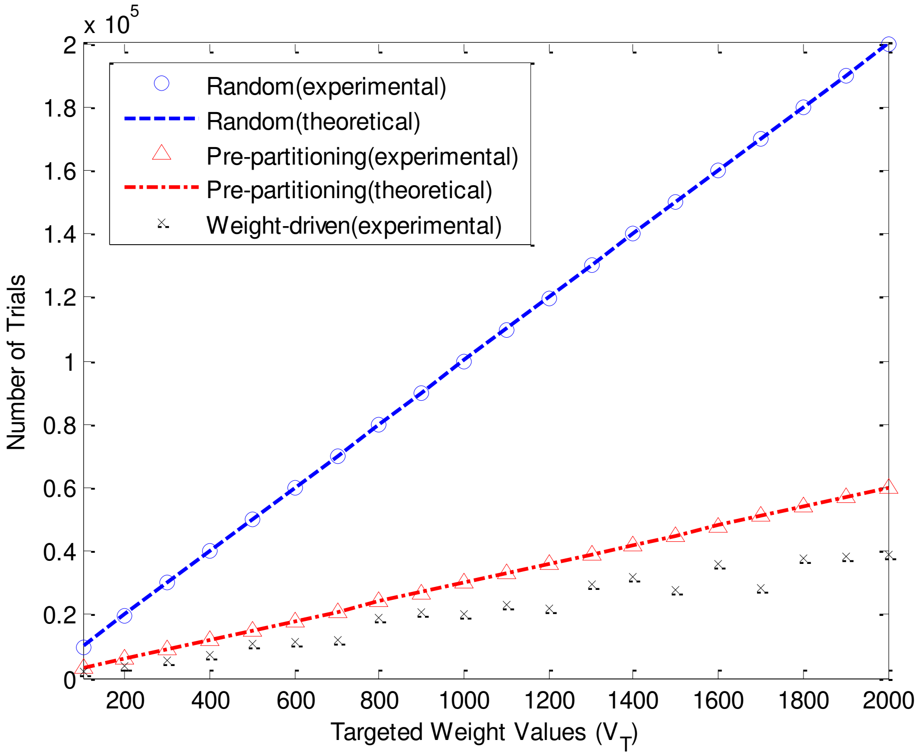


*Figure 6.4 Exploration Costs (Number of Trials Required per Task) for A Learning Agent*

## *6.3 Results and Analysis*

We keep using IEEE 802.22 as the basis to select parameters in this chapter. The important parameters used in this simulation are shown in table 6.1. The propagation model applied is Okumura-Hata propagation model with 8dB log-normal shadowing [85]. No further power control policy is applied. The value function 4-9 is also used in this chapter and the values of weighting factors are shown in table 6.2.

*Table 6.1 Simulation Parameters*

| **Parameter** | **Value** |
|---|---|
| Service area | 100km$^2$ |
| Number of users | 100 |
| Link length | 0.2km-1.5km |
| Number of channels | 20 |
| Carrier Frequency | 700MHz |
| Transmitter antenna height | 30m |
| Transmit power | 30dBm |
| Transmitter antenna gain | 0dBi |
| Receiver antenna gain | 0dBi |
| Bandwidth | 1MHz |
| Noise floor | -114dBm |
| Interference threshold | -20dBm |
| SINR threshold | 10dB |
| Size of preferred channel set | 3 |
| Preferred channel weight threshold | 3 |
| Size of reserved resource set | 20 |

*Table 6.2 Weighting Factor Values*

| $f_1$ | | $f_2$ | |
|---|---|---|---|
| Reward | Punishment | Reward | Punishment |
| 1 | 1 | 1 | -1 |

We also keep using the open spectrum scenario and the transmitter-receiver pair system model where all users are given equal priority to use the spectrum – a cognitive only band where the users are purely cognitive radios. It is worth investigating the system performance in a cognitive only band since it is likely in the future that devices in such ('unlicensed') bands will become increasingly cognitive, enabling them to deal with interference and reconfigure, allowing new more efficient techniques and solutions to be developed. Our approach is different from pure opportunistic scheduling since we understand a cognitive radio to have distinct features of spectrum cognition, intelligence and reconfigurability. It is these three features that have the potential to significantly enhance the capability of future communication systems.

Figure 6.5 shows the significant improvement achieved by applying pre-partitioning and weight-driven exploration in terms of overall blocking probability, compared with a no learning scheme and the uniform random exploration scheme.

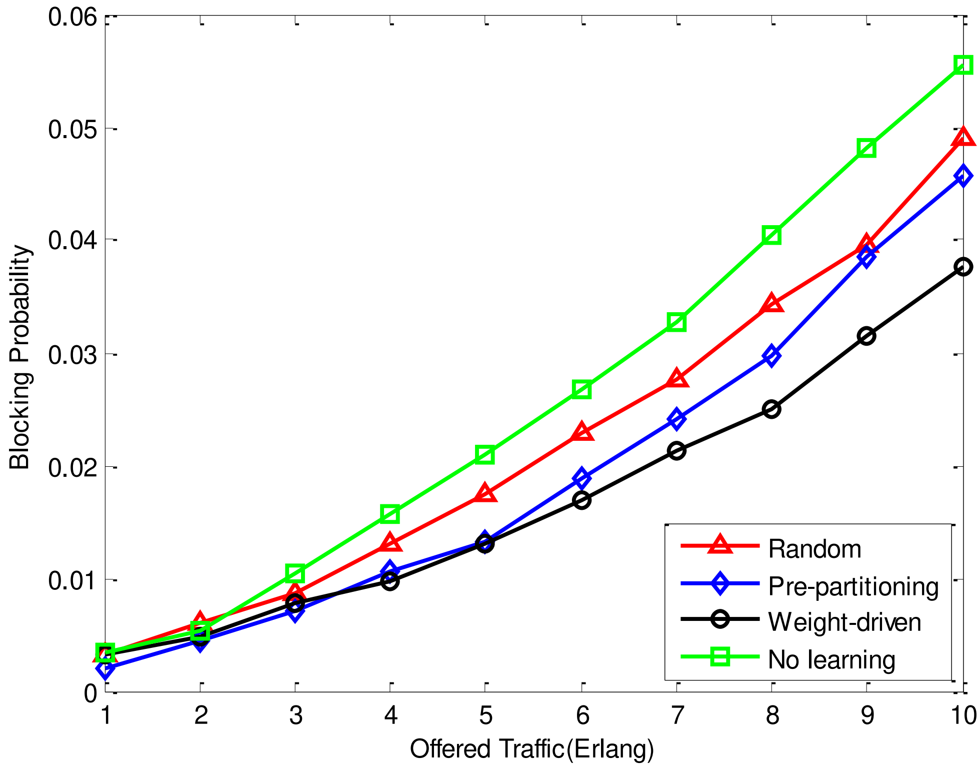


*Figure 6.5 System Blocking Probability at Different Offered Traffic Levels*

It can be seen that the performance of the basic uniform random exploration algorithm has been improved by random spectrum pre-partitioning. This is because random pre-partitioning will significantly reduce the size of the available spectrum pool of each user. Therefore, the requirements for the learning part of the agent to explore the action space are reduced. In other words, the initial exploration stage of a cognitive radio user is accelerated by pre-partitioning. The blocking probability of the weight-driven exploration scheme is also significantly lower than the uniform random exploration scheme. This is because weight-driven exploration is much more efficient and the users will find their optimal resources faster than the uniformly exploring users. Users cause less interference to others since they are spending less time exploring.

The weight-driven exploration scheme also performs better than the pre-partitioning scheme in general. It is shown that the blocking probability of the weight-driven exploration scheme is higher than the pre-partitioning scheme when the offered traffic is lower than 4 Erlangs. In this case, more direct spectrum partitioning is more efficient. The interfering pairs are quickly constrained in their reserved spectrum set and are no longer a source of interference. However, the blocking probability of the weight-driven exploration scheme is lower when the offered traffic is above 4 Erlangs. The users in the pre-partitioning scheme suffer from a higher level of blocking probability since they only have access to a random subset of the entire spectrum pool (20% of the spectrum pool in this simulation), meaning that they have fewer alternatives if the level of transmission requests is high and the reserved channels are not suitable for communication. The blocking probability of the pre-partitioning scheme will increase quickly if we increase the offered traffic. The drawback of pre-partitioning is clear that some users may be constrained to a set of channels which have a high level of interference. Consequently these users may find it difficult to find unoccupied spectrum to use for communication. It is clear that a small preferred channel set is more suitable for a low offered traffic scenario. Pre-partitioning will lose its advantage when the offered traffic is high. Therefore, the advantage of the weight-driven exploration scheme on system blocking is clear.

Transmission dropping is mainly caused by hidden terminals. Therefore, the improvement on system dropping probability is purely achieved by reinforcement learning. Fig.6.6 compares the dropping probabilities of the 3 reinforcement learning based schemes in the same way as Fig.6.5.

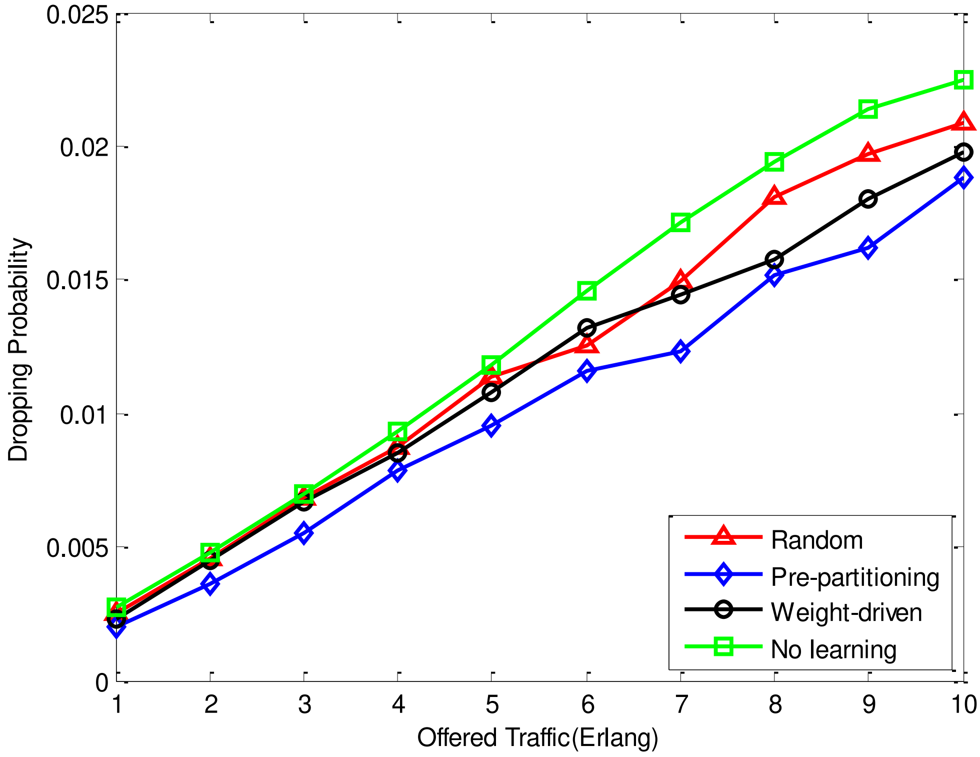


*Figure 6.6 System Dropping Probability at Different Offered Traffic Levels*

It shows that the users in the weight-driven exploration scheme will not achieve a significant improvement when the offered traffic is low. The weight-driven exploration scheme and the uniform random exploration scheme are about the same from 1 Erlang to 5 Erlangs. The information users obtained through dropped transmissions is not sufficient to learn to avoid a transmission from being dropped. In other words, users learn slower than the rate at which the environment changes because the number of dropped transmissions is too small to aid learning. However, if we keep increasing offered traffic to 6 Erlangs or more, the dropping probability will be improved by reinforcement learning and a further reduction in dropping probability can be achieved by applying weight-driven exploration. On the one hand, there will be more connections dropped by the system and users will obtain more information to learn. On the other hand, the mean interarrival time is shorter which means the information learned by users will be used more efficiently to avoid dropping.

Unlike the weight-driven exploration scheme, the reduction in system dropping probability by pre-partitioning is significant. The dropping probability of the pre-partitioning scheme is the lowest of the 3 schemes. This is because the spectrum pool is quickly partitioned and the users are constrained in their channel sets. The probability that transmissions which are dropped by hidden terminal effect is reduced. The trade-off between blocking probability and dropping probability of the communication system is clear in this case. The improvement in terms of dropping probability by pre-partitioning is obtained at the expense of a higher level of transmission blocking.

Fig.6.7 shows that the probability of a user being activated in the exploitation stage is increased by applying the proposed approaches. The improvement of the efficient exploration scheme can be clearly seen. The number of activations in the exploitation phase of the weight-driven exploration scheme is about 40% higher on average than the uniform random exploration scheme. The figure is about 25% for the pre-partitioning scheme. Moreover, this figure drops more slowly in the weight-driven exploration scheme if we increase the traffic load which means that the users not only converge to exploitation faster by weight-driven exploration, but the probability of remaining in the exploitation stage is also higher. The percentage of activations in exploitation will only decrease by about 2.5% in the weight-driven exploration scheme if we increase the offered traffic from 1 Erlang to 10 Erlangs. However, this figure is 19% in the uniform exploration scheme. The line corresponding to pre-partitioning scheme in Fig.6.7 drops even faster than the uniform random scheme when the offered traffic is above 6 Erlangs. Here the users struggle to find a suitable channel since the available resources are very limited in the pre-partitioning scheme.

Fig.6.8 shows the blocking probability of the weight-driven approach in more detail. It can be seen that by applying weight-driven exploration, not only is the blocking probability of the exploration stage significantly lower compared to the previous scheme, but the blocking probability in the exploitation stage is also improved. The available spectrum is quickly partitioned through reinforcement learning in this case and the system is more stable.

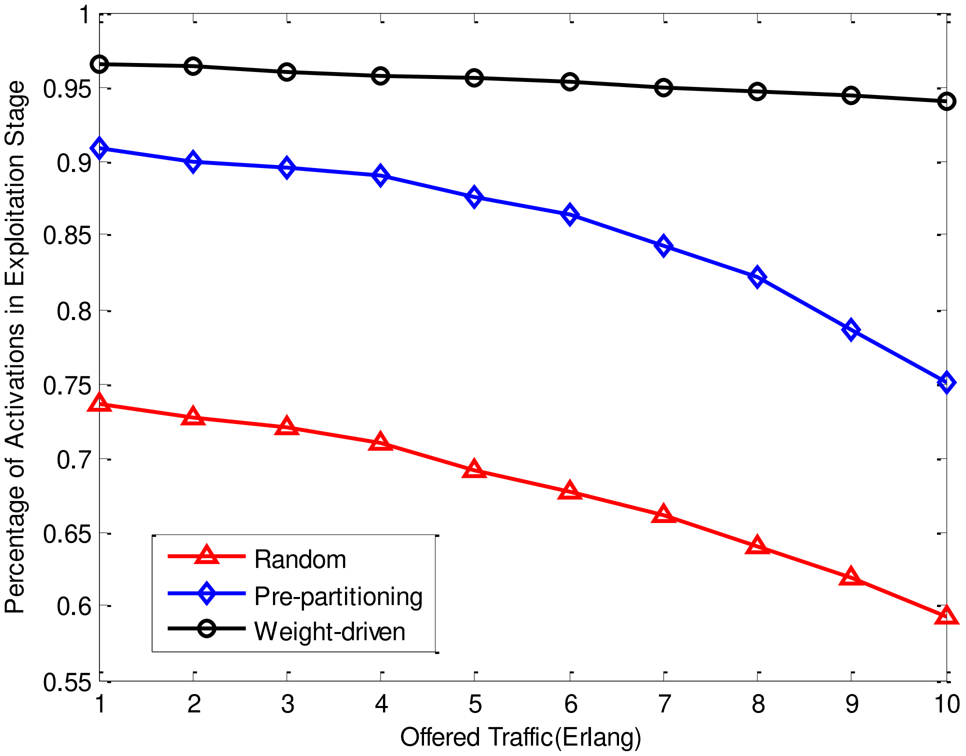


*Figure 6.7 Percentage of Activation in Exploitation at Different Offered Traffic Levels*

The blocking probability of the weight-driven scheme in exploration is even lower than the exploitation stage when the traffic load is lower than 7 Erlangs. It can be explained based on the results shown in Fig.6.7. Fig.6.7 shows the percentage of activations in the exploitation stage. It can be seen that only about 5% of the activations take place in the exploration phase. The rest of the activations are accepted in exploitation. The probability of a user being activated to explore the action space is significantly reduced by weight-driven exploration. Thus, the users can hardly be blocked in exploration when the offered traffic level is low. The spectrum pool is quickly partitioned in a local area and users will almost directly obtain a set of preferred resources. The transmission blocking mainly occurs in the exploitation stage in this case. The blocking probability of exploration will be higher than exploitation if we increase the offered traffic because more and more users will stay in the exploration stage. Additional blocking will take place in exploration accordingly.

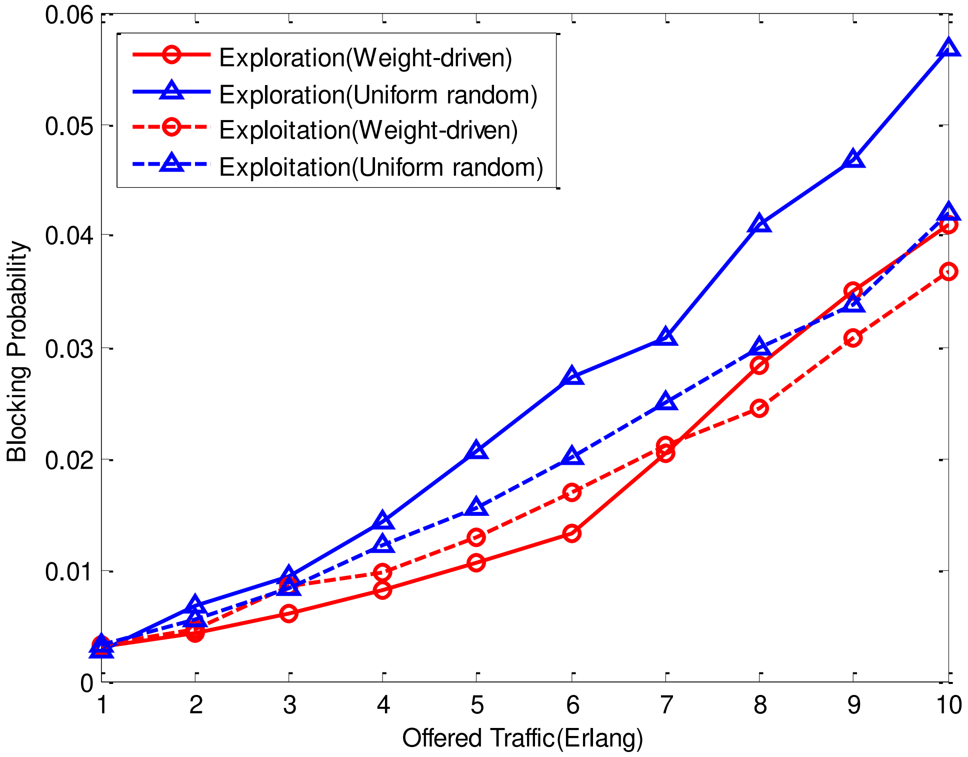


*Figure 6.8 System Blocking Probability in Different Stages at Different Offered Traffic Levels*

Fig.6.9 shows the dropping probability in different stages of the weight-driven exploration scheme and uniform random exploration scheme. The dropping probabilities of the exploitation stage in both schemes are also on the same level. No further improvement can be obtained in exploitation. The further reduction of the overall dropping probability by weight-driven probability in Fig.6.9 is obtained only because the dropping probability of the exploration stage in uniform random exploitation scheme is significantly increased when the offered traffic is above 6 Erlangs and the overall dropping probability is increased accordingly. The dropping probability of the weight-driven scheme in exploration is also lower than the exploitation stage when the traffic load is lower than 8 Erlangs. The users are hardly dropped in exploration since the exploration phase is too short. The dropping probability in weight driven exploration will be higher when more users stay in exploration stage.

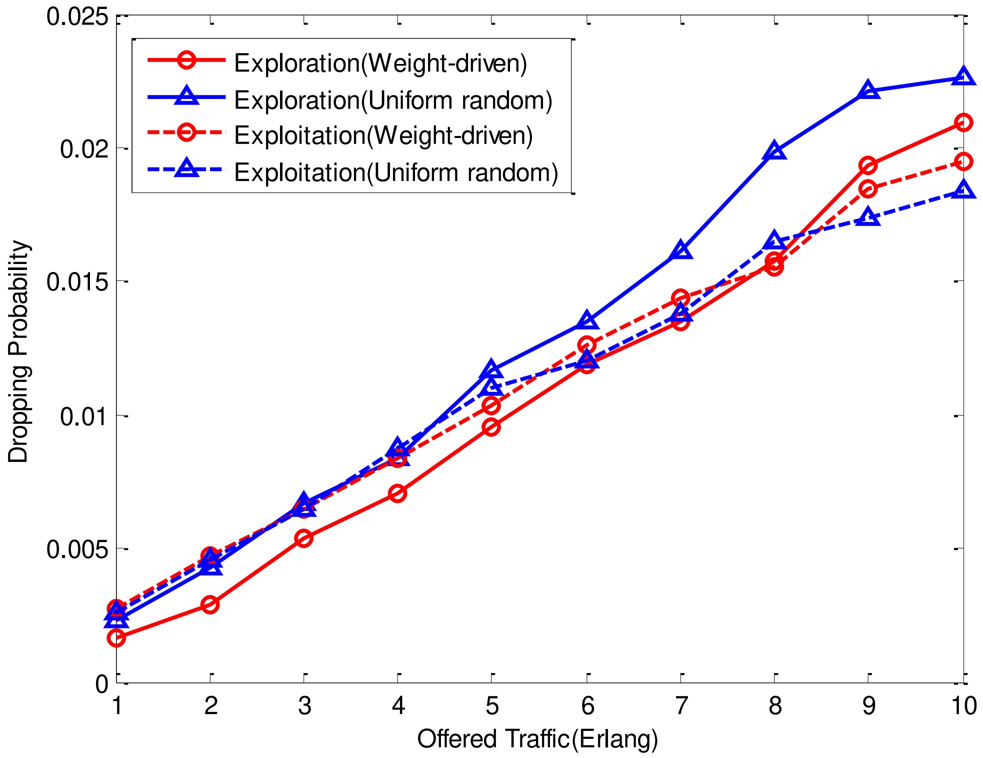


*Figure 6.9 System Dropping Probability in Different Stages at Different Offered Traffic Levels*

## 6.4 Conclusions

In this chapter, the exploration efficiency and learning efficiency have been investigated in the context of cognitive radio spectrum sharing. Two novel efficient exploration algorithms for reinforcement learning-based cognitive radio have been introduced, which are able to significantly improve the exploration efficiency of a learning cognitive radio.

By randomly reserving a subset of available spectrum, the spectrum pool is fully partitioned before transmissions start. Simulation results show how the overall performance of a pre-partitioning scheme is better than the traditional uniform random exploration scheme. By pre-partitioning the spectrum pool, approximately 25% more activations are accepted in the exploitation phase than with the uniform random exploration scheme. Moreover, pre-partitioning has the potential to reduce the hidden terminal problem since the transmissions of different users will quickly converge to different channel sets in this case.

Therefore, users in a local area have a better opportunity to avoid each other, and the probability that transmissions are interfered by hidden terminals is reduced

Weight-driven exploration utilizes the information gained in exploration to guide the exploration process itself, and weight-driven exploration also ensures exploration by merging randomness into action selection. Simulation results show how the number of activations in the exploitation stage is 40% higher on average by applying weight-driven exploration. The system is more stable and the overall performance is significantly better than the uniform random exploration scheme.

# Chapter 7. Learning-based Green Cognitive Radio

**Contents**

## *7.1 Introduction*

According to Moore's Law [96], the processing power of CPUs and the mass storage capacity of devices double approximately every two years. The rapid growth of computing power significantly promotes the capacity of wireless communication devices. This in turn attracts massive data flow between the wireless devices. Thus, bandwidth efficiency has been the primary research topic in the field of wireless communications over the past few decades, resulting in many highly complex techniques and devices that are capable of delivering significantly high bandwidth efficiency. Figure 7.1 (directly reproduced from [97]) shows the significant development of data rates in wireless systems in recent 20 years. For example, the data rates of mobile networks has increased from merely a few kbps to hundreds of Mbps in about 15 years.

Nevertheless, the increase of power consumption associated with achieving high bandwidth efficiency is also significant. It is estimated that a 16-20% increase each year can be expected on the power consumption of mobile networks [98]. In other words, the power consumption of mobile network doubles every 5 years. The global energy usage of mobile networks is about 123.98 Billion kWh per year, and is already contributing to about 1% of the total world energy consumption [99]. Thus, how to reduce the energy consumption of wireless communications system and enable 'green' communication has recently become an increasingly important topic.

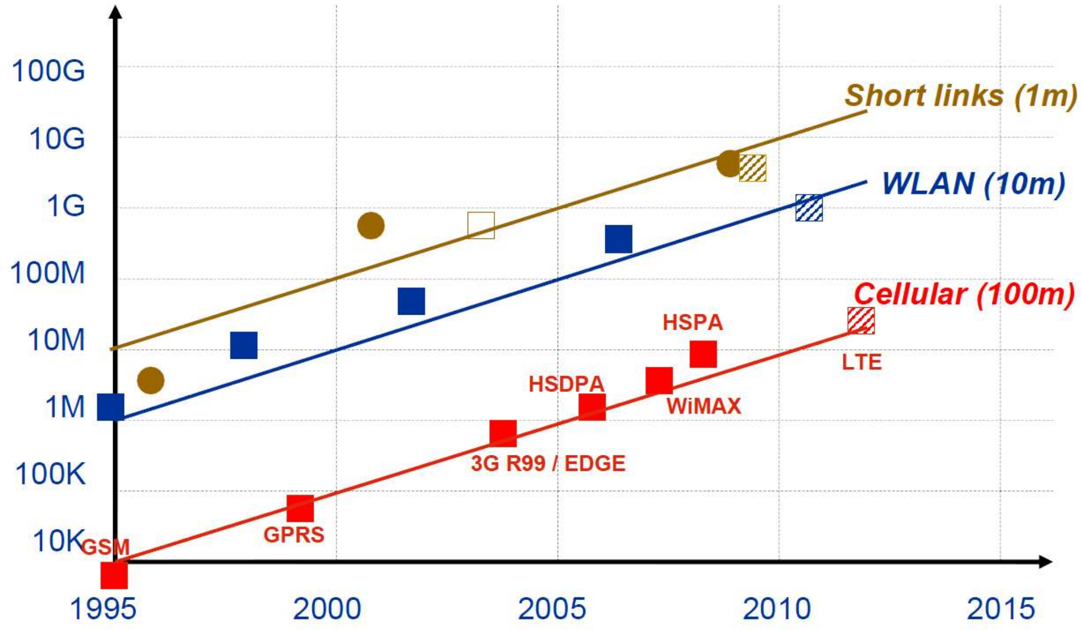


*Figure 7.1 Wireless System Development Roadmap (directly reproduced from [97])*

Cognitive radio systems are becoming increasingly complex, largely due to the increasing flexibility required by different services. This complexity in turn leads to power intensive devices. The purpose of this chapter is to introduce distributed reinforcement learning-based strategies that are able to reduce the energy consumption and the complexity of a cognitive radio system.

## 7.2 Learning-based Green Cognitive Radio

Spectrum awareness is a vital element of cognitive radio. Cognitive radio needs to either periodically or continuously sense the spectrum to obtain the information of the environment. The process of spectrum sensing, as a means of monitoring radio activity in a given bandwidth, is a power-intensive and time-consuming process. Take the 'Cognitive, Radio-Aware, Low-Cost (CORAL) Research Platform' for example [100], the cognitive radio platform that developed by Communications Research Centre Canada takes about 20 seconds to scan a 83 MHz wide frequency band. The time consumption is significant. The energy consumption of spectrum sensing is also likely to be high. Since there are no figures available to directly show the power consumption of spectrum sensing for cognitive radio, we take an example from the field of wireless sensor networks. The energy consumption of a sensor node in transmitting and receiving mode is

24.75 mW and 13.5 mW respectively, and there is no power consumption difference between listening and receiving [101]. The energy consumption of listening and receiving is about 50% of transmitting. Thus it is foreseen that spectrum sensing will be one of the main sources of energy consumption within a CR device.

Two main approaches of spectrum sensing have been proposed for cognitive radio: energy detection and feature detection [2]. The disadvantage of the energy detection technique is that the energy detector is unable to differentiate between modulated signals, interference and noise. However, energy detection has the advantage of lower energy and time consumption compared with feature detection. Feature detection needs significantly longer observation time to recognize different types of signals and it is computationally complex, which calls for a more sophisticated device both in software and hardware.

It has been shown how distributed reinforcement learning is perfectly suited to cognitive radio spectrum sharing scenarios in the previous chapters. Reinforcement learning-based techniques have the potential to efficiently exploit the available spectrum resource. Instead of sensing the entire available spectrum arbitrarily, this approach is able to share the spectrum based on an optimum spectrum sharing strategy, discovered by agents from their interaction with the wireless communication environment. Therefore energy efficiency of the wireless communication device can be improved. Moreover, spectrum decisions can be made only by the results of learning instead of spectrum sensing after an adequate channel partitioning has been established. In other words, spectrum sensing can be replaced by intelligence if the available spectrum can be partitioned autonomously by individual learning. A cognitive radio is able to avoid unsuitable spectrum thereafter. The energy efficiency for spectrum sensing can be maximized accordingly [102].

The purpose of this chapter is to explore the green aspects of the reinforcement learning-based algorithms introduced in previous chapters. This is achieved by limiting the requirement by using the experience gained by the learning users. Once users are mature enough to choose a suitable channel purely by learning, they are allowed to set up wireless links without sensing the target resource beforehand.

There are two ways in which energy efficiency can be improved. The first one is to use distributed reinforcement learning to limit or even replace sensing techniques whenever it is possible, achieving good performance in most of the communication tasks. The other way is to enable a very efficient utilization of very basic spectrum sensing techniques by distributed reinforcement learning. By combining appropriate learning techniques with a few essential elements of cognitive radio, low complexity ‘intelligent’ strategies will be able to significantly reduce the energy consumption of cognitive radio [102].

Thus, two spectrum reduction schemes are compared with a full sensing scheme where Cognitive Radio users scan the target spectrum at the beginning of every activation: (1) a restricted sensing scheme that users only sense the spectrum in their ideal resource set; and (2) a minimum sensing scheme where users directly use their preferred resources to communicate without sensing. The time and power consumption of these schemes is also shown to illustrate the benefits of our scheme. The further spectrum sensing reduction achieved by efficient exploration techniques we introduced in chapter 6 is also investigated in this chapter in section 7.2.2.

### 7.2.1 Learning-based Spectrum Sensing Reduction Algorithms

The spectrum sensing reduction algorithms is illustrated in figure 7.2

The two-stage algorithm introduced in chapter 5 is the basis of our spectrum sensing reduction schemes. $U_i$ will evaluate its own local system state first. In this case, it is whether $U_i$ has found its preferred resource set.

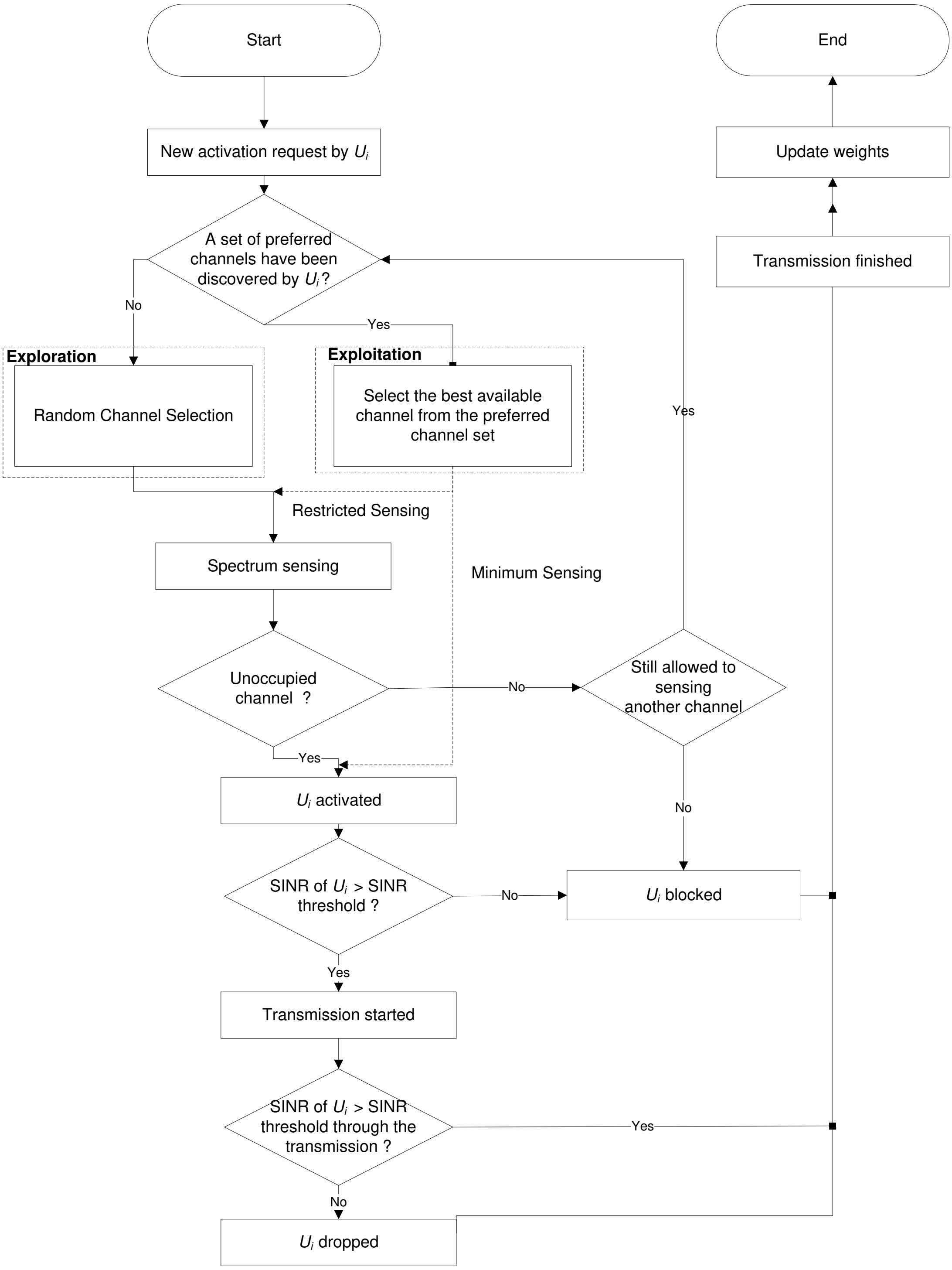

Start
End
New activation request by $U_i$
Update weights
A set of preferred channels have been discovered by $U_i$?
Transmission finished
No
Yes
Exploration
Exploitation
Random Channel Selection
Select the best available channel from the preferred channel set
Yes
Restricted Sensing
Spectrum sensing
Minimum Sensing
Unoccupied channel ?
No
Still allowed to sensing another channel
Yes
$U_i$ activated
No
SINR of $U_i$ > SINR threshold ?
No
$U_i$ blocked
Yes
Transmission started
SINR of $U_i$ > SINR threshold through the transmission ?
Yes
No
$U_i$ dropped

*Figure 7.2 Algorithm Flowchart*

If $U_i$ is still in the exploration stage, it chooses a channel randomly from the available spectrum set, and then $U_i$ senses the interference level on that channel. If $U_i$ is in the exploitation stage then:

- Restricted sensing scheme: $U_i$ senses the spectrum in their ideal resource set randomly.
- Minimum sensing scheme: $U_i$ directly accesses the spectrum in the preferred channel set without sensing.

Instead of sensing the entire available spectrum arbitrarily, the scheme is designed to share the spectrum based on a spectrum sharing strategy discovered by the agents from their interaction with the wireless communication environment. Thus, by reducing the requirement of spectrum sensing or even bypass spectrum sensing in the exploitation stage, the overall energy consumption is expected to be decreased.

The improvement in the energy efficiency of spectrum sensing can be assessed in an indirect way by examining the number of sensed channels. This has been assessed through simulation. 1000 cognitive radio transmitter-receiver pairs are assumed to be uniformly distributed throughout a square service area of 1000 km$^2$. 100 channels are available for communication. The Okumura-Hata propagation model is used along with log-normal shadowing with a standard deviation of 8 dB. The wireless link length is uniformly distributed between 1 km and 2k m. A carrier frequency of 300 MHz is used and the transmitter antenna height is set to 30 m. The transmit power is fixed at 30 dBm and no further power control policy is applied. The gains of the transmit and receive antennas are both fixed at 0 dBi.

An event-based scenario is considered. At each event a random subset of pairs are activated, up to a maximum of 400. Only energy detection sensing is applied to cognitive radio. A fixed interference threshold of -40 dBm is used. The *SINR* threshold is set to 10 dB. The improvement on energy efficiency associated with spectrum sensing by reinforcement learning is shown in figure 7.3.

The advantages of the intelligence-based strategies can be clearly seen from figure 7.3. The number of sensed channels effectively represents the time and energy consumption of spectrum sensing. The crossed line maintains its position around 1.15 because cognitive radios sense every target channel on a random basis. The triangle line converges to 1 since the probability to successfully set up a transmission link in the first tested channel is increased by reinforcement learning. The restricted sensing scheme where cognitive radio can avoid use of an improper resource by learning also outperforms in terms of bandwidth efficiency. The total energy consumption of the minimum sensing scheme after event 2000 is only about 1.72% of the full sensing scheme, assuming that energy consumption increases with the number of channels sensed. The line with the star symbol which represents the energy consumption of the minimum sensing falls towards zero which means spectrum sensing can be stopped if the available spectrum is fully partitioned by learning.

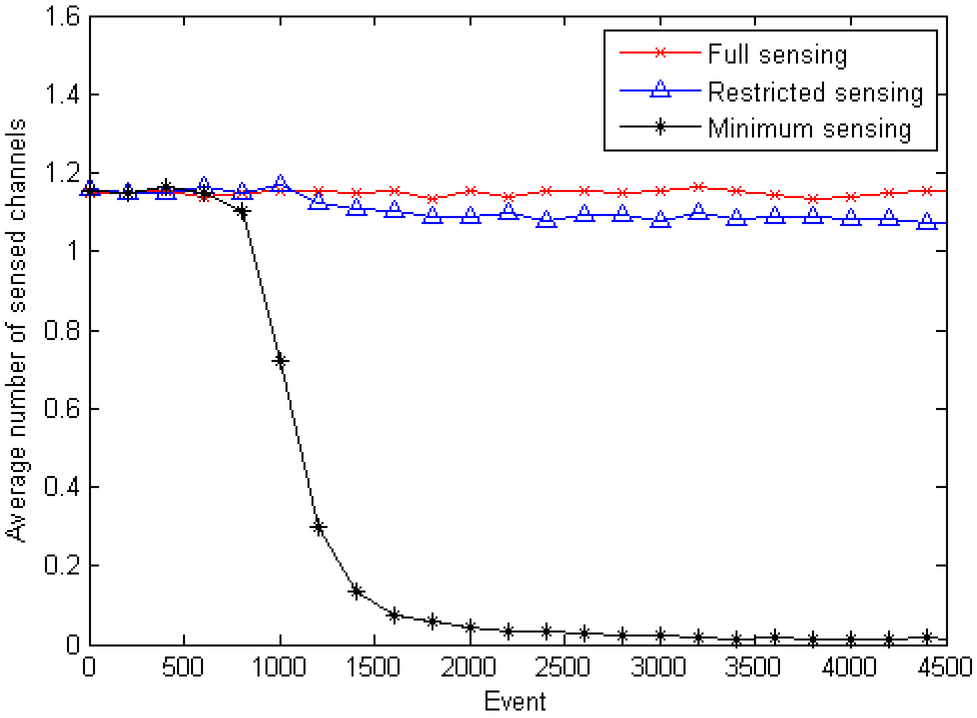


*Figure 7.3 Average Number of Sensed Channels*

Fig. 7.3 also shows the convergence behaviour of our learning schemes. Like other learning algorithms for dynamic channel assignment [56], our scheme needs a sufficiently high number of stages to converge to its optimal state. From the start of the simulation to event 2000, our learning scheme has converged to its ideal spectrum sharing strategy. CR users found their preferred resource set gradually. After event 2000, the learning scheme finally arrived at its spectrum sharing equilibrium which practically means CR users' preferred resource sets are fully occupied by good channels. The user is able to avoid improper channels by using its prior experience. Though the node is designed to move back to the pre-play stage if only one of its preferred channels is no longer good to communicate, the state of the learning scheme is extremely stable. Obviously, the CR users in our scheme have the potential to share spectrum in a ‘polite’ way even if they do not sense beforehand.

In order to illustrate the system performance in more detail, we record the number of sensed channels in each activation and plot the CDF of it in figure 7.4. It can be seen that about 77% of the transmission activations in the minimum sensing scheme succeed without sensing the target spectrum. The restricted sensing scheme performs slightly better than the full sensing scheme. About 90% of the communication requests in the restricted sensing scheme succeed before the user tests the third channel, but in the full sensing scheme only 85% users are able to meet this requirement. Figure 7.4 also shows that about 99% requests are accomplished before sensing four channels.

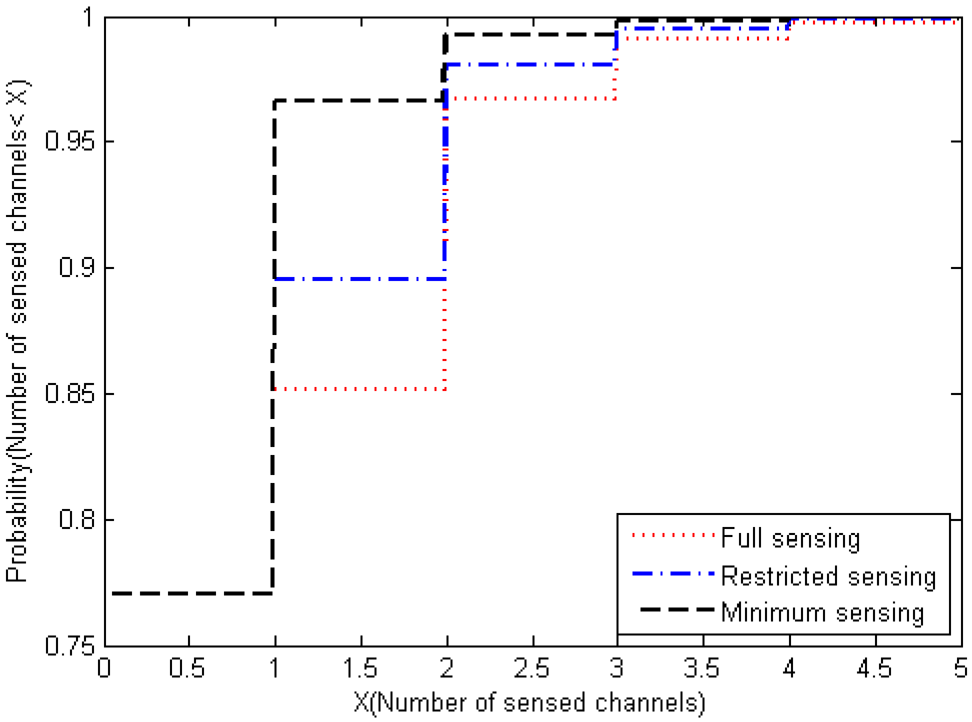


*Figure 7.4 Cumulative Distribution Function of the Number of Sensed Channels in Each Communication Activation*

Fig 7.5 illustrates the CDF of system blocking probability of the three schemes which we discussed before. About 70% users’ blocking probability in the minimum sensing scheme are below 0.04. But in the full sensing scheme and the restricted sensing scheme, it is about 87% and 95% respectively. Comparing with the red dotted line which is the CDF of the full sensing scheme, the blocking probability of the minimum sensing scheme is higher. It is reasonable that a scheme which always chooses a free channel to operate performs better than a scheme occasionally picks a channel without sensing. It is not expected that the minimum sensing scheme can show its advantages from this point of view. On the contrary, the restricted sensing scheme achieves a better performance compared to the full sensing scheme. This is because the user in the restricted sensing scheme is able to sense the channels which have higher probability to success according to prior experience. This is particularly important because communication can still be dropped.

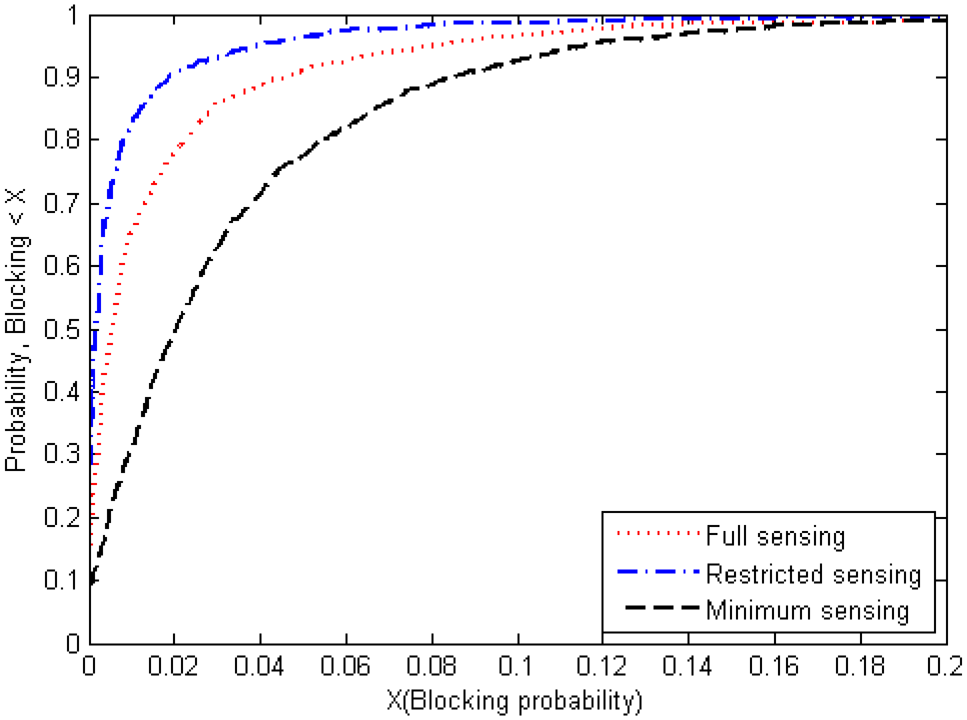


*Figure 7.5 Cumulative Distribution Function of System Blocking Probability at Discrete points over the Service Area*

It can be seen that in every scheme there are about 2% of users whose blocking probability is above 0.2. The blocking probability of these users is difficult to improve no matter which scheme is applied. This is because these users are located either at an extremely high user density area or at a place suffering significant shadowing. The opportunity for these users to successfully set up a communication link is limited.

Figure 7.6 shows the CDF of dropping probability which illustrates the level of system interruption. Since the information of system dropping is also used to update the spectrum sharing strategy, the performance of the restricted sensing scheme is better than the scheme without learning. But just like the performance of blocking probability, the dropping probability of the minimum sensing is also higher than the full sensing scheme. A scheme which stops sensing to some extent cannot perform better than the full sensing scheme in the aspect of communication quality. However, it can be seen that the overall performance of the minimum scheme is acceptable given that the gap between the minimum sensing scheme and others is not large. The genuine benefit of the limited sensing schemes is discussed in the following paragraphs.

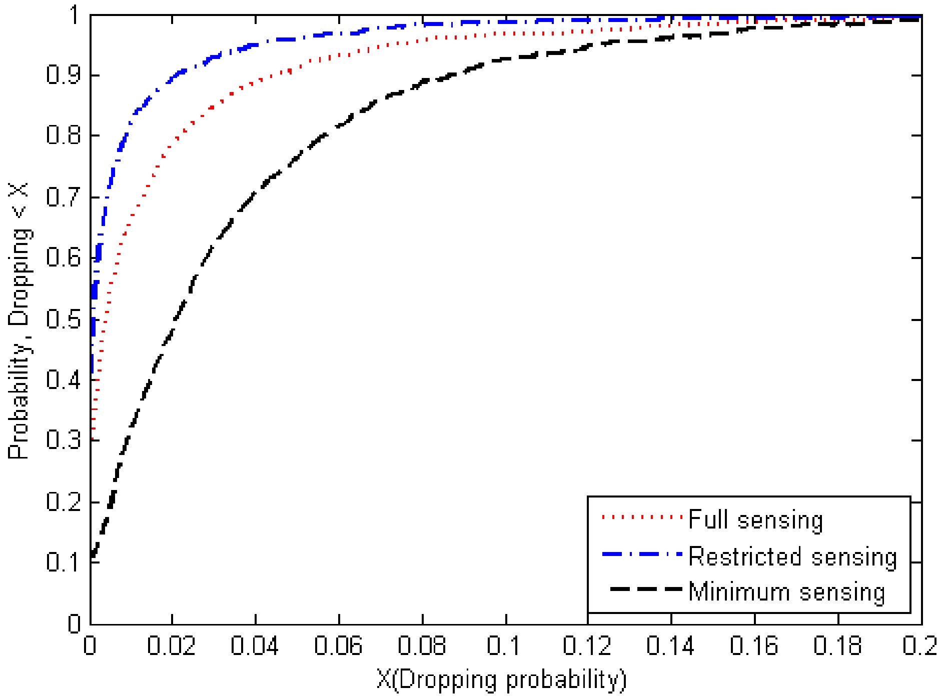


*Figure 7.6 Cumulative Distribution Function of System Dropping Probability at Discrete points over the Service area*

By utilizing reinforcement learning, the need for spectrum sensing is significantly reduced. The overall time and energy consumption of spectrum sensing in the minimum sensing scheme is about 23% of the full sensing scheme. After the minimum sensing scheme converged to its spectrum sharing equilibrium, this figure is only 1.72%. The restricted sensing scheme improves the system performance in two aspects: the sensing consumption is 5% lower than the full sensing scheme. Furthermore, the blocking and dropping probability is also the lowest of the three schemes. Since time and power efficiency are critical issues in real time communication, the advantages of our learning scheme is definite.

It is possible to improve the energy efficiency even further if more complex sensing approaches are applied [103]. Take IEEE 802.22 for example [30], which employs two stages of sensing: fast sensing and fine sensing. Fast sensing estimates interference power in different channels over a very short time period and returns limited information for fine sensing. Fine sensing will then sense the target channels for a significantly longer period of time. Here energy consumption can be further reduced by applying artificial intelligence techniques described previously.

### 7.2.2 Spectrum Sensing Reduction by Efficient Exploration

The efficiency of spectrum sensing can be improved further by applying efficient exploration techniques described in chapter 6. Again the energy consumption of spectrum sensing can be assessed indirectly by examining the number of sensed channels per activation, and this is shown in Fig 7.7.

The pre-partitioning scheme and weight-driven exploration scheme are compared with a no learning scheme and a uniform random exploration-based learning scheme. It can be seen that the efficient exploration-based schemes will consume less power than the other two schemes since the efficient exploration techniques will significantly improve the learning efficiency of the users, meaning that the uses are able to discover their preferred resource more efficiently.

Thus, the energy consumption of the basic uniform random exploration approach is further improved by the proposed efficient exploration techniques. The biggest reduction in exploration is achieved by weight-driven exploration. Compared to the pre-partitioning scheme, users in the weight-driven scheme will not exclude any channels in exploration. Therefore, the performance of the weight-driven exploration scheme will not be affected by the absence of initial choice.

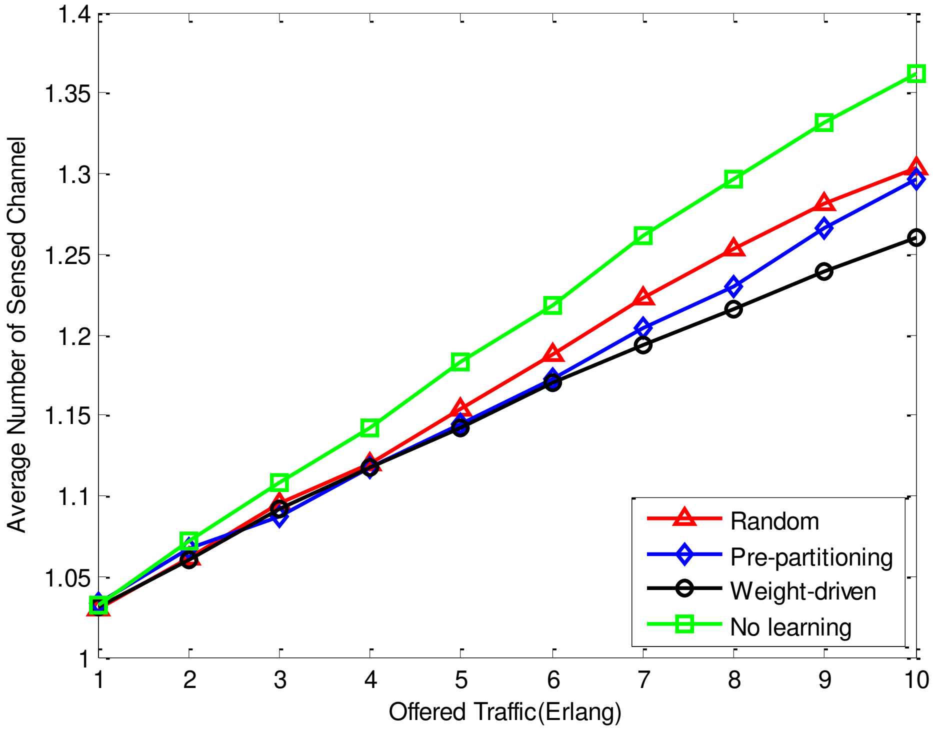


*Figure 7.7 Average Number of Sensed Channels per Activation*

## *7.3 Conclusions*

This chapter has addressed the issue of Green cognitive radio. We have shown how by exploiting reinforcement based learning, it is possible to virtually partition a set of channels to reduce the need to sense the radio spectrum as frequently, resulting in fewer channels to be scanned overall, enabling energy savings.

It is shown that by acquiring a subset of preferred resources, the restricted sensing scheme and the minimum sensing scheme are able to significantly reduce the need for spectrum sensing. The overall energy consumption of spectrum sensing in the minimum sensing scheme is about 23% of the full sensing scheme. After the minimum sensing scheme converged to its spectrum sharing equilibrium, this figure is only 1.72%. The restricted sensing scheme improves the system performance in two aspects: the sensing consumption is 5% lower than the full sensing scheme. Moreover, the blocking and dropping probability is also the lowest of the three schemes. The advantages of our learning scheme is obvious since time and power efficiency are critical issues in real time communication

Efficient learning-based algorithms are able to reduce the requirement of spectrum sensing further. In the pre-partitioning scheme, the spectrum pool is fully partitioned before transmissions start. Weight-driven exploration utilizes the information gained in exploration to guide the exploration process itself, and weight-driven exploration also ensures exploration by merging randomness into action selection. Thus more activations are accepted in the exploitation phase that the requirements for spectrum sensing are reduced. Simulation results show that the weight-driven exploration scheme achieves the highest spectrum sensing reduction in all the proposed schemes.

# Chapter 8. Reinforcement Learning-based Cognitive Channel Assignment for Dual-Hop Beyond Next Generation Mobile Networks

**Contents**

## *8.1 Introduction*

One of the requirements for the next generation wireless communication system is to provide high throughput per user. However, it is also crucial that future systems are able to provide the required average data rate to all active users simultaneously within a geographical area. This is particularly true in the highly populated urban city centre areas where the demand for wireless broadband service is the highest. The current 4G technologies LTE and WiMAX are designed to provide a throughput density of about 100 Mbps/km$^2$ in traditional cellular deployments [104]. The capacity density of current next generation techniques may be adequate in less populated areas, but in an area where the user density is much higher, the capacity of the system becomes inadequate. The highest population density city Mumbai has a population density of 29,650 people/km$^2$. The typical population density in the commercial areas of a European city is about 8,000 people/km$^2$ [104]. If we assume only 10% of the 8000 people are subscribed to the wireless broadband service and only 20% of those subscribers require simultaneous wireless broadband service access, and if we also assume the data rate of each subscriber is 5 Mbps, then the required overall capacity density will be:

$$8{,}000 \times 10\% \times 20\% \times 5Mbps = 800Mbps \qquad (8\text{-}1)$$

It can been seen that the required capacity density in a near future is an order of magnitude higher than the current next generation systems.

Thus the primary goal of the Dual-Hop Beyond Next Generation Mobile Network is to improve the overall capacity density of the mobile network to 1 Gbit/s/km$^2$ anywhere in the service area [22]. In order to meet this ambitious goal, a number of novel advanced techniques have been introduced, including the application of a two-tier system architecture (shown later in figure 8.1), multi-beam directional antenna, multi-beam assisted MIMO and collaborative MIMO, and cognitive radio techniques.

The complex dual-hop architecture not only leads to a greater level of interference, but also requires a higher level of complexity in radio resource management. Spectrum decisions therefore better to be made in a distributed fashion in order to achieve the aggressive target of 1 Gbps/km$^2$ throughput density. The cognitive radio based technique is a feasible approach for the resource management of the access and self-backhaul. Comparing with conventional dynamic radio resource management (RRM) approaches, cognitive radio based techniques have the potential to improve spectrum efficiency, reduce overall complexity, and improve link reliability. The work in this chapter is developed from the techniques we have introduced in previous chapters. The advantages of our reinforcement learning-based schemes are clear that not only they outperform non-learning schemes, but also the self-organizing features enabled by learning are particularly desirable for the Beyond Next Generation Mobile System's complex resource management tasks.

## *8.2 Beyond Next Generation Mobile Network: System Model*

A novel dual-hop architecture shown in figure 8.1 has been proposed for the beyond next generation mobile network [22]. In order to provide a cost-efficient high capacity density anywhere in the cell, the system is composed of an access network and a self-backhaul network. The key elements of the novel architecture shown in the figure are:

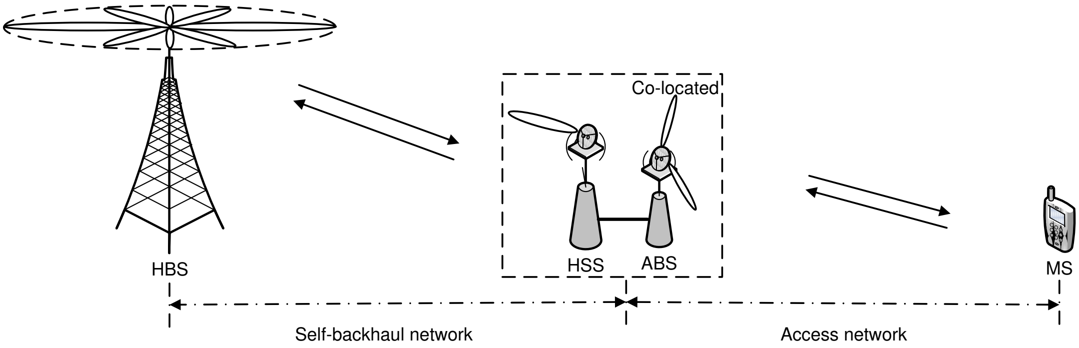


*Figure 8.1 Beyond Next Generation Mobile Network System Architecture*

- Hub Base Station (HBS): an entity that connected to the operater's backhaul network. Multi-beam directional antenna is deployed over roof-top at HBSs, providing high capacity self-backhaul link to the access network entities.
- Access Base Station (ABS): a low-cost entity that provides the access to the mobile subscribes. A large number of ABSs will be mounted below roof-top on electricity poles, traffic lights, traffic signs, etc. ABS has wired connections to the Hub Subscriber Station. ABSs also have two single-beam directional antennas pointed to two opposing directions, providing the desired capacity to the mobiles.
- Hub Subscriber Station (HSS): an entity that connected to ABS by a wired link and to the HBS by a single-beam directional antenna.
- Mobile Subscribers (MS): connected to different ABSs depending on their location.

A large number of ABSs are deployed along the streets, providing sufficient coverage to MSs. The aggregated traffic at ABSs are then transmitted to the associated HBSs. By applying directional antenna and advanced radio resource management techniques, the ambitious goal of 1 Gbps/km$^2$ is possible to be delivered.

## *8.3 Reinforcement Learning-based Channel Assignment for Beyond Next Generation Mobile Network*

Reinforcement learning-based cognitive radio channel assignment techniques have the potential to be applied to beyond next generation mobile network systems, including base stations and mobile users. Cognitive techniques, such as spectrum sensing and machine learning, are able to enable a very aggressive frequency reuse while reducing the radio resource management complexity. Entities in the system will have the access to all the available frequency bands, and the channel decision will be made individually by the entities in a distributed fashion. Available frequency bands are periodically monitored, through channel utilization data base or spectrum sensing. Learning techniques will then utilize the information gained in sensing. In this case a single learning engine processes the information for both hops of the wireless link simultaneously as shown in figure 8.2. The goal is to enable an efficient autonomous spectrum sharing between entities through a reliable dynamic channel assignment algorithm.

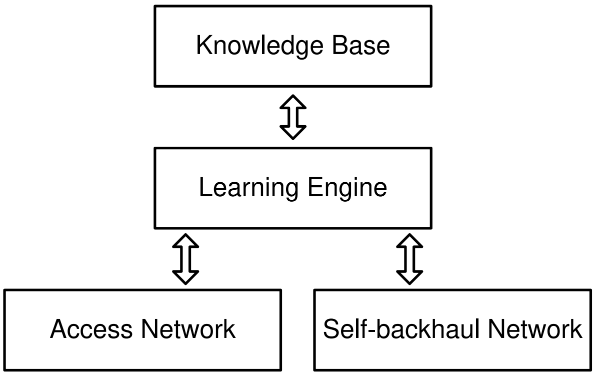


*Figure 8.2 Learning Engine for Beyond Next Generation Mobile Network*

Channel decisions are made individually at each entity. By using reinforcement-based learning, entities will assess the success level of a particular action. Entities in the access networks and the self-backhaul networks will select channels based on the weights assigned to the spectral resources - resources with higher weights are considered higher priority. The following linear function we used in previous chapters has also been applied to the dual-hop system:

$$W' = f_1 W + f_2 \tag{8-2}$$

where $W$ is the weight of a channel at time *t-1*, and $W'$ is the weight at time $t$ according to previous weight $W$ and the updated feedback from system. $f_1$ and $f_2$ are the weighting factors at time $t$ that will take on different values depending on the localized judgment of current system states and the environment.

The algorithm is shown in figure 8.3. We consider $E_i$ is entity $i$ in the dual-hop system. $E_i \in E$ and $E$ is the entity set that contains all reinforcement learning based BS and MS. By randomly choosing channels, the operating entity $E_i$ will explore the spectrum space first. We define a specific threshold such that if the weight of a used resource is above the threshold, the action of taking this resource is considered as a preferred action and the resource is regarded as a preferred resource.

It is assumed that the spectrum sensing is carried out at the receiver end of the wireless link. Beyond next generation mobile network is designed primarily for the dense city centre area where the propagation environment is very complex. The utilization of directional antenna in such area makes it possible that the received signal power and the interference power vary significantly in a few meters range. Thus, spectrum sensing at the transmitter is not accurate enough to identify the interference level on the targeted channel. Spectrum sensing at the receiver end therefore is more desirable in this case.

It is realized that the protocol needs to be properly defined to support this CR function. A certain amount of control information needs to be exchanged between the transmitter and the receiver. The control information overhead incurred in the CR dynamic channel assignment process is not expected to be high. A downlink transmission request to the receiver and an uplink response back to the transmitter with the subchannel discovered in sensing are the information that needs to be exchanged. In order to support this feature, the information required to be exchanged between entities is defined in figure 8.4 and figure 8.5:

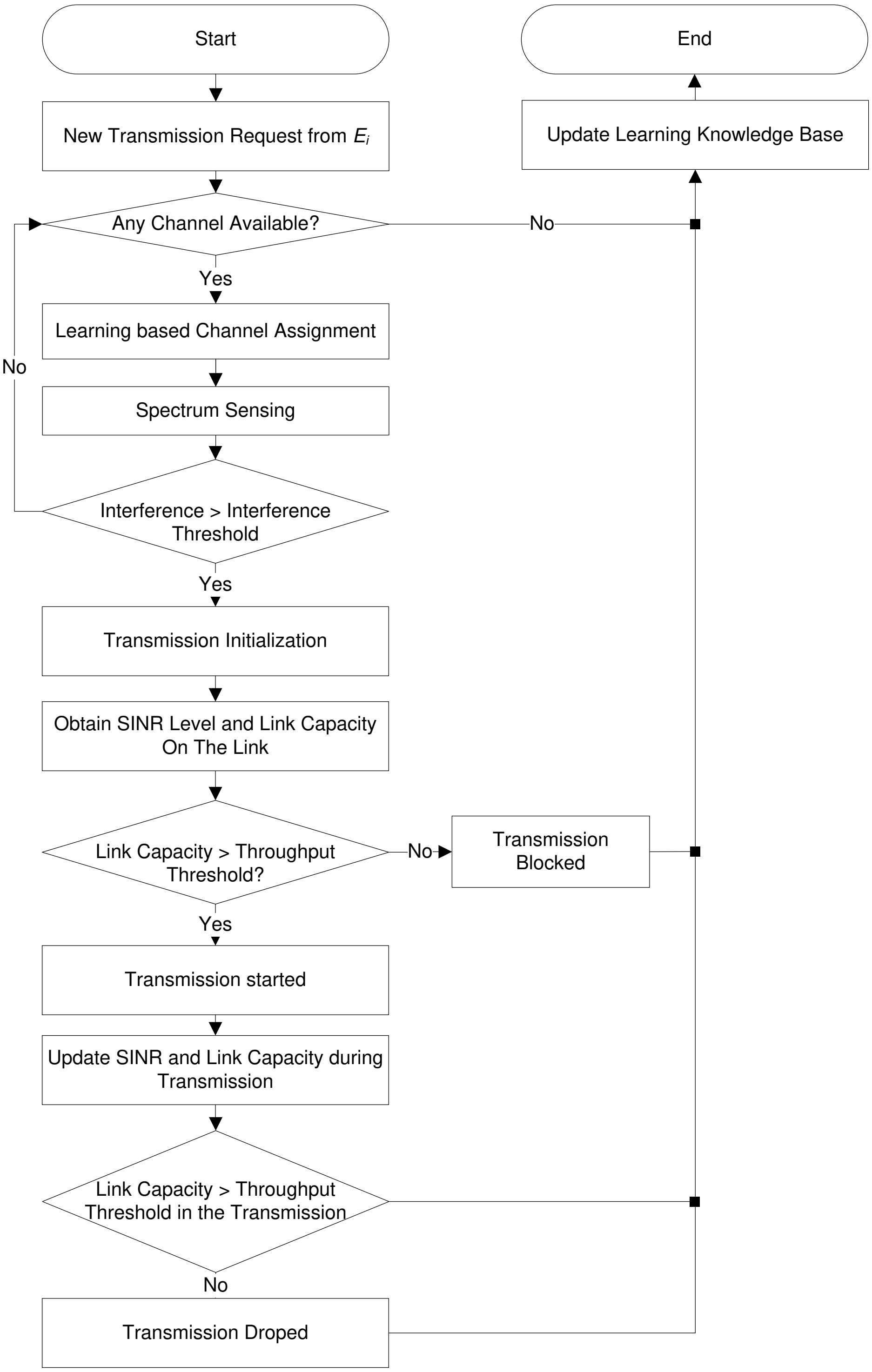

Start
New Transmission Request from $E_i$
Any Channel Available?
Yes
No
Learning based Channel Assignment
Spectrum Sensing
Interference > Interference Threshold
No
Yes
Transmission Initialization
Obtain SINR Level and Link Capacity On The Link
Link Capacity > Throughput Threshold?
No
Transmission Blocked
Yes
Transmission started
Update SINR and Link Capacity during Transmission
Link Capacity > Throughput Threshold in the Transmission
No
Transmission Droped
End
Update Learning Knowledge Base

*Figure 8.3 Learning-based Channel Assignment Algorithm*

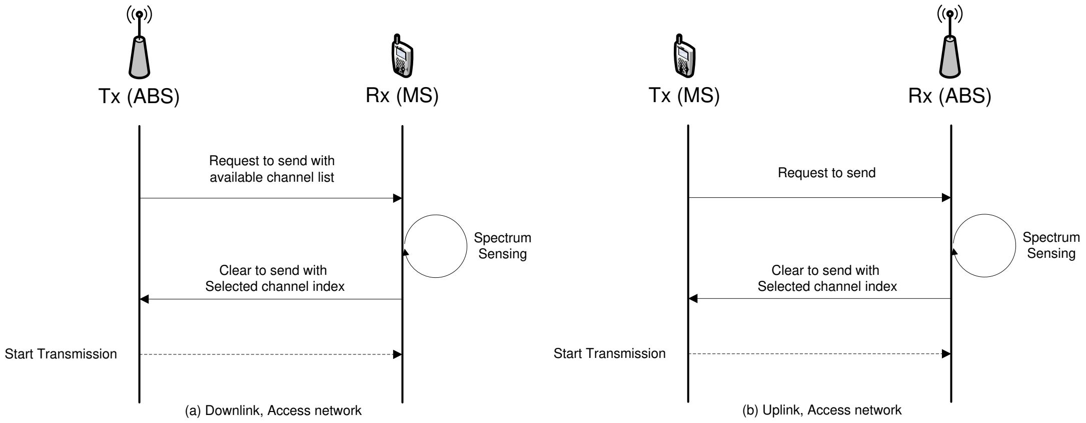

Tx (ABS)
Rx (MS)
Request to send with available channel list
Spectrum Sensing
Clear to send with Selected channel index
Start Transmission
(a) Downlink, Access network
Tx (MS)
Rx (ABS)
Request to send
Spectrum Sensing
Clear to send with Selected channel index
Start Transmission
(b) Uplink, Access network

*Figure 8.4 Information Flow between Entities in Access Network*

It is assumed that the channel availability information is kept in HBSs for the self-backhaul network and in ABSs for the access network. Figure 8.4 (a) shows the information flow between entities in the access network for a downlink transmission request. The available channel list with the request to send will be sent to the mobile subscriber first by the associated ABS and then the MS carries out spectrum sensing on the available channels. After that, the MS will send the clear to send message with the selected channel index back to the ABS and the ABS can then proceed to transmit data on the selected channel.

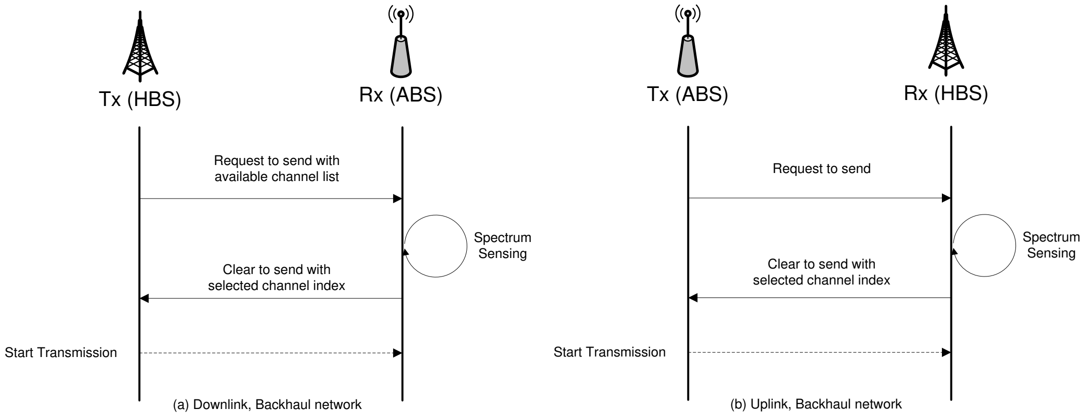

Tx (HBS)
Rx (ABS)
Request to send with available channel list
Spectrum Sensing
Clear to send with selected channel index
Start Transmission
(a) Downlink, Backhaul network
Tx (ABS)
Rx (HBS)
Request to send
Spectrum Sensing
Clear to send with selected channel index
Start Transmission
(b) Uplink, Backhaul network

Figure 8.5 Information Flow Between Entities in Self-Backhaul Network

In the case of uplink transmissions, as shown in figure 8.4 (b), the MS only send the request to send to the ABS prior to the transmission and the ABS will sense all available channels. Then the ABS will send the selected channel index with the clear to send message back to the MS.

The information exchanged between entities in the self-backhaul network is illustrated in figure 8.5. The information exchanged and the process itself are similar to the cases of the access network in figure 8.4. For downlink transmission requests, HBS need to send the available channel information to the receiving ABS first. However in the case of uplink transmission, such information is not required since spectrum sensing is carried out at HBSs.

It can be seen from figure 8.4 and figure 8.5 that the information exchanged between entities is mainly channel index and this approach only requires a single information exchange prior to the transmission. Thus, the overhead information generated in the wireless system is limited to the minimum level.

## *8.4 System Modelling Scenario*

### 8.4.1 Deployment Scenario and System Parameters

A Manhattan-grid environment is used in this simulation, and the square topology is applied as in figure 8.6 [22, 105]. There are 11 streets both East-West (E-W) and North-South (N-S), and that forms a 10×10 block area. The HBS antennas are placed above rooftop in the centre of their cell and 2 HBS form a 4 cell square deployment environment in this case. The HBS beams are indexed clockwise from 1 to 11. HSS and ABS are indexed as shown in figure 8.9. It is assumed that HSS6 (ABS6) of the HBS1 cell is co-located with HSS17 (ABS17) of the HBS 2 cell. ABS 6 and ABS 17 in this scenario is serving the E-W street and the N-S street respectively.

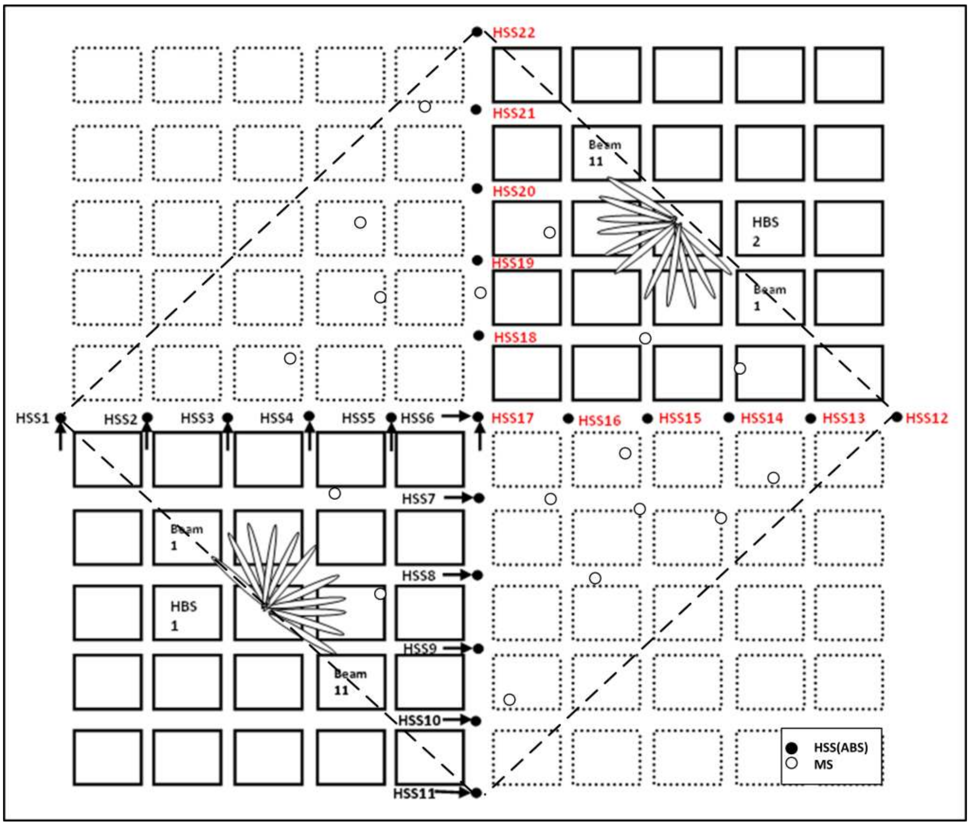


*Figure 8.6 Modelling Scenario*

MSs are randomly distributed within the area surrounded by the dashed line since only two sectors (12 beams) of each HBS directional antenna (facing towards to centre area) have been considered.  MS is associated with ABS based on a set of rules based on their location:

- MS on N-S street: connect to the nearest ABS on the same vertical street
- MS on E-W street: connect to the nearest ABS on the same horizontal street
- MS in a building block: connect to the nearest ABS
- MS on a street cross: connect to the nearest ABS either vertically or horizontally

Any MSs outside the highlighted service area will not be served by any of the ABSs in figure 8.6, and therefore are not considered. The following table 8.1 shows the system parameters.

*Table 8.1 System Parameters [105]*

| Parameter | Value |
|---|---|
| Deployment area dimension | 900m*900m |
| Number of streets in one dimension | 11 |
| Street width | 15 m |
| Building block size | 75m*75m |
| Number of building blocks per cell | 25 |
| Service area size | 0.405 km$^2$ (636.4m*636.4m) |
| Number of HBS | 2 |
| Number of ABS | 22 |
| HBS antenna pattern | 19 dBi -21 dBi |
| HSS antenna pattern | 13 dBi |
| ABS antenna pattern | 17 dBi |
| HBS antenna height | 25m |
| HSS, ABS antenna height | 5m |
| MS antenna height | 1.5m |
| MS antenna | Omnidirectional antenna |
| HBS transmission power | 37dBm |
| HSS transmission power | 27dBm |
| ABS transmission power | 37dBm |
| MS transmission power | 23dBm |
| Carrier frequency | 3.5 GHz |
| Number of Channels | 8 |
| Throughput threshold | 0.86 bps/Hz |
| Lognomal shadowing | 6dB |
| Noise floor | -114 dbm/MHz |
| HBS - HSS propagation model | Ray-tracing based channel model |
| ABS - MS propagation model | WINNER II B1, WINNER II B4 |
| MIMO | TSB + MIMO(MMSE) |
| Traffic | Poisson traffic model |
| Duplexing | TDD (50%-50% split for Downlink and Uplink) |

Note that 8 10 MHz channels in total are assumed, 4 channels for the backhaul network and 4 channels for the access network. These channels can be assumed if there are 40 MHz licensed frequency bands and 40 MHz unlicensed frequency bands are used, or we can assume that the interference between the backhaul network and the access network can be limited under a certain level that the 40 MHz licensed band can be shared by them. TDD duplexing is used and a 50%-50% split is assumed for Downlink and Uplink.

30 OFDMA format subchannels are assumed within a 10 MHz channel. No advanced scheduling techniques are applied at the interim stage. Only one subchannel will be randomly allocated to a user at one time subject to availability. If an entity's transmission has failed to be carried out due to interference, the transmission will be terminated at the first selected channel: no retransmission is allowed.

The antenna data are obtained by CASMA [105]. The 24-beam HBS directional antenna has gains between 19 dBi and 21 dBi for different beams with a beamwidth of approximately 15×10 degree in azimuth and elevation. A 13 dBi directional antenna is applied at the HSS and pointed towards the largest power ray direction as it suggested in [22]. The HSS antenna has a beamwidth of approximately 40×40 degree in azimuth and elevation. Two 17 dBi antennas with approximately 25×25 degree beamwidth are assumed at each ABS, pointing in the two opposite directions either vertically or horizontally along the street and in parallel to the ground. The azimuth angle and the elevation angle are calculated between MS and ABS beams, and a 3D antenna pattern is used for obtaining the appropriate ABS antenna gain when the azimuth angle and the elevation angle are available. The 3D antenna pattern is illustrated as figure 8.7.

Note that a throughput threshold has been used to check the quality of the wireless link instead of a *SINR* threshold as used in previous chapters. This is because by applying MIMO, the achievable data rate at a certain *SINR* level is no longer a deterministic value. Instead, it is a variable that follows a distribution, meaning that even at a *SINR* as low as -5 dB, it may still be possible to transmit data. Thus, a fixed *SINR* threshold which is traditionally used in such a scenario is no longer accurate enough. A throughput threshold of 0.86 bps/Hz is used in this case instead of the *SINR* threshold. The throughput value of 0.86 bps/Hz is calculated by using the Truncated Shannon Bound [86] when *SINR* value is assumed at the lowest level of 1.8 dB.

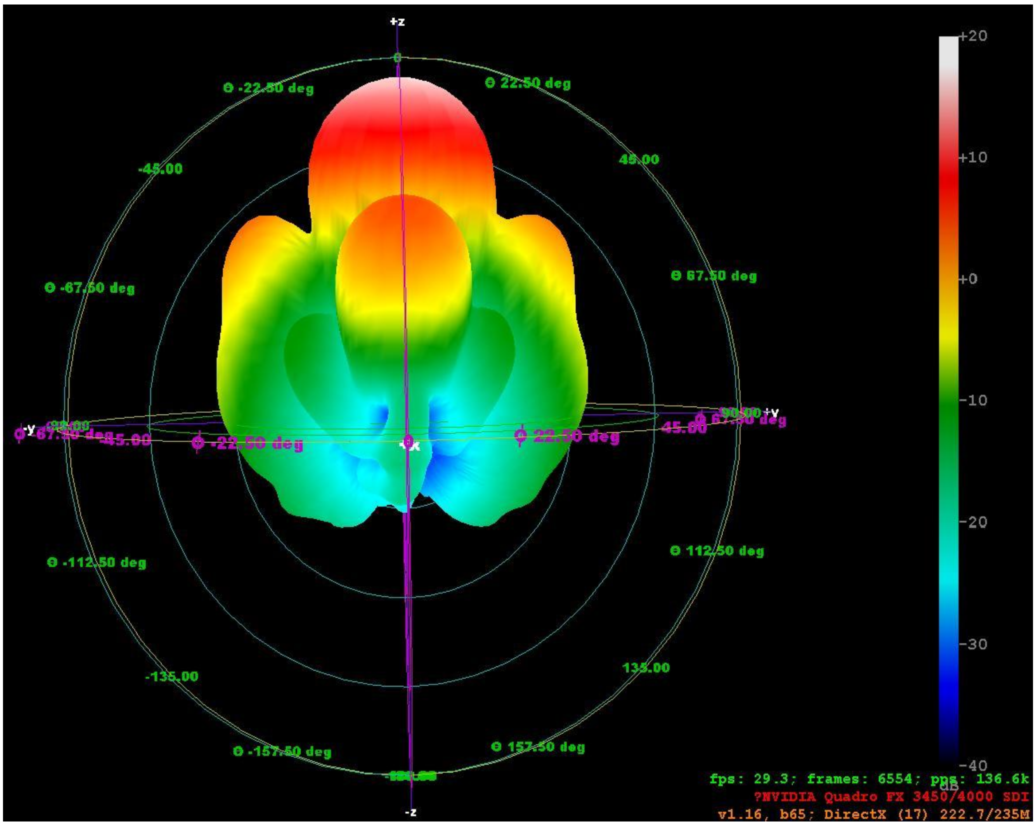


*Figure 8.7 3D ABS Antenna Pattern (directly reproduced from [105])*

### 8.4.2 MIMO

Advanced MIMO techniques have the potential to significantly improve the link capacity. The achievable data rates of both the backhaul network and the access network of the dual-hop network can be significantly enhanced by using multiple antennas and suitable signal processing techniques at the hub base stations (HBSs), access base stations (ABSs) and mobile stations (MSs). HBSs and ABSs are dual polarized antennas which can act as an antenna array and MSs are assumed to be equipped with at least two antennas each. Therefore we assume that the backhaul and access links (HBS-ABS and ABS-MS) are 2x2 MIMO channels.

In SISO channels the achievable throughput is a deterministic value based on the path loss, shadowing and small-scale fading; and this can be evaluated easily. However in MIMO channels, throughput is no longer a deterministic value. Instead the achievable rate follows a distribution.

A method has been developed jointly that maps each value of average link SINR to a statistical distribution of achievable rates [105]. From -5 to 40 dB and for 1 dB step size, we generate 3000 samples offline which fully capture the statistical features of the 2x2 MIMO link. MMSE linear detection is assumed. Then at the system level:

1. For each wireless link we calculate the SINR taking into account the path loss and shadowing of the useful and the interfering links.
2. For the obtained SINR we choose the closest value of SINR for which we have an available MIMO throughput distribution (a vector of throughput values).
3. For the chosen SINR value we select at random a throughput value from the corresponding vector.

Thus, 2x2 MIMO can be taking into account when evaluating system performance by using the module we have developed.

### 8.4.3 Propagation Models

#### *8.4.3.1 Self-backhaul network*

A number of channel models have been used to calculate path loss between the entities in the dual-hop network. A ray-tracing based channel model developed at Université catholique de Louvain [105] is used to estimate the path loss between the HBS and HSS. The path loss values of the backhaul network obtained by the ray-tracing tool is shown in table 8.2 [105].

*Table 8.2 Backhaul Network Ray-Tracing Results*

| | Beam 1 | Beam 2 | Beam 3 | Beam 4 | Beam 5 | Beam 6 | Beam 7 | Beam 8 | Beam 9 | Beam 10 | Beam 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HSS 1 | -116.00 | -133.60 | -138.60 | -140.40 | -142.30 | -139.60 | -151.30 | -138.50 | -146.80 | -147.40 | -149.70 |
| HSS 2 | -106.70 | -110.30 | -129.40 | -130.50 | -128.60 | -130.60 | -139.30 | -150.60 | -151.30 | -150.20 | -148.50 |
| HSS 3 | -109.00 | -101.10 | -120.70 | -115.60 | -120.80 | -120.80 | -133.10 | -134.00 | -138.20 | -139.80 | -131.50 |
| HSS 4 | -120.80 | -117.40 | -106.60 | -101.50 | -106.60 | -109.50 | -114.50 | -123.00 | -132.80 | -133.30 | -126.30 |
| HSS 5 | -141.80 | -147.10 | -114.60 | -126.30 | -112.20 | -112.80 | -135.70 | -114.60 | -140.60 | -130.00 | -138.30 |
| HSS 6 | -140.00 | -138.00 | -141.00 | -135.90 | -126.10 | -106.60 | -121.60 | -130.20 | -129.80 | -130.00 | -132.30 |
| HSS 7 | -129.60 | -145.60 | -137.60 | -137.10 | -127.60 | -119.30 | -113.40 | -123.20 | -131.90 | -130.80 | -130.00 |
| HSS 8 | -123.50 | -132.10 | -120.80 | -123.50 | -111.80 | -101.00 | -99.20 | -99.50 | -112.50 | -121.60 | -119.30 |
| HSS 9 | -144.60 | -126.60 | -147.10 | -133.20 | -138.30 | -128.80 | -120.40 | -115.50 | -101.30 | -102.00 | -107.30 |
| HSS 10 | -151.00 | -131.80 | -135.90 | -148.40 | -147.30 | -137.00 | -147.30 | -137.20 | -126.20 | -112.10 | -112.90 |
| HSS 11 | -158.70 | -148.90 | -145.20 | -143.80 | -145.10 | -134.10 | -148.30 | -143.80 | -139.40 | -127.30 | -111.10 |

### *8.4.3.2 Access Network*

WINNER II [106] provides a comprehensive set of channel models that are capable of covering the propagation environment of the access network. In this work, WINNER II B1 and WINNER II B4 have been used to model the wireless environment of Manhattan-grid city centre area because they are the most advanced models available for the modelling of such environment. WINNER II B1 is used to calculate the pass loss between ABS and MS that is located outside of a building block. The path loss between ABS and MS inside of a building block is estimated by using WINNER II B4.

*8.4.3.2.1 WINNER II B1 – Urban Micro-Cell*

The propagation environment investigated by WINNER II in an urban micro-cell scenario is quite similar to the access networks' propagation environment [106]. A Manhattan-grid layout is considered and all BS and MS antennas are assumed well below the rooftops of the surrounding buildings. All ABS and MS are assumed to be outdoor as illustrated in figure 8.8. Both Line-of-Sight (LOS) and Non-Line-of-Sight (NLOS) cases have been considered, allowing for temporary blockage of the LOS, for example by large vehicles. The LOS and NLOS path loss are calculated as follows:

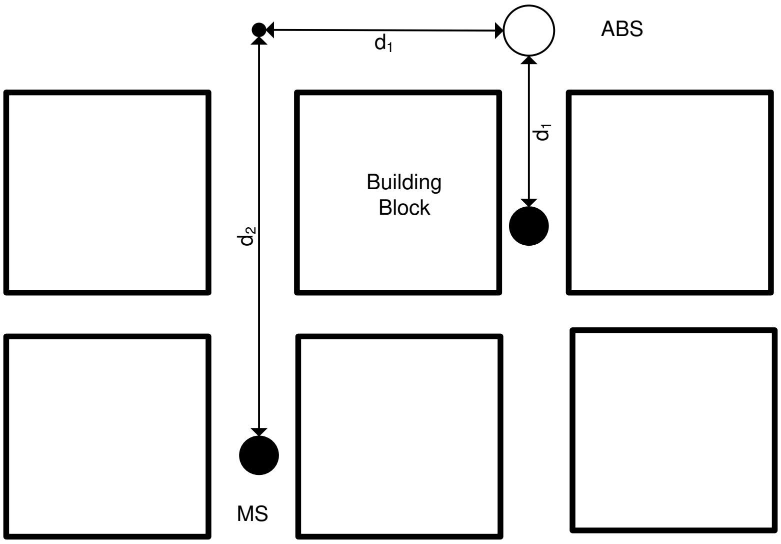


*Figure 8.8 WINNER II B1 Path Loss Calculation*

***LOS***

If the MS and ABS are on a same street, then the path loss can be calculated by:

$$PL = 40.0\log_{10}(d_1) + 9.45 - 17.3\log_{10}(h'_{BS}) - 17.3\log_{10}(h'_{MS}) + 2.7\log_{10}(f_c/5.0) \qquad (8\text{-}3)$$

Where

$$h'_{BS} = h_{BS} - 1 \qquad (8\text{-}4)$$

and

$$h'_{MS} = h_{MS} - 1 \tag{8-5}$$

$d_1$ is the distance between ABS and the LOS MS, $h_{BS}$ is the ABS antenna height and $h_{MS}$ is the MS antenna height.

***NLOS***

If the MS and ABS are not on the same street, then the path loss can be calculated by:

$$PL = \min(PL(d_1, d_2), PL(d_2, d_1)) \tag{8-6}$$

where

$$PL(d_k, d_l) = PL_{LOS}(d_k) + 20 - 12.5n_j + 10n_j \log_{10}(d_l) + 3\log_{10}(f_c/5.0) \tag{8-7}$$

and

$$n_j = \max(2.8 - 0.0024d_k, 1.84), \tag{8-8}$$

$PL_{LOS}$ is the path loss of B1 LOS and $k, l \in \{1,2\}$, $d_1$ and $d_2$ are distance between the entities along the street as it is shown in figure 2-3.

### *8.4.3.2.2 WINNER II B4 – Outdoor to Indoor*

The layout that is considered in WINNER II B4 [106] is also urban micro-cell. The only difference is WINNER II B4 only considers the path loss between on street BSs and in building MSs. Therefore, in this simulation the path loss between ABS and in building MS is obtained by using WINNER II B4 propagation model. The scenario is shown in figure 8.9.

The path loss in this case can be calculated by:

$$
\begin{aligned}
&PL = PL_b + PL_{tw} + PL_{in} \\
&\begin{cases} PL_b = PL_{B1}(d_{out} + d_{in}) \\ PL_{tw} = 14 + 15(1-\cos(\theta))^2 \\ PL_{in} = 0.5 d_{in} \end{cases}
\end{aligned}
\tag{8-9}
$$

Where $PL_{B1}$ is the B1 path loss, $d_{out}$ is the distance from the ABS to the penetration point on the wall, and $d_{in}$ is the distance from that point to the mobile terminal. $\theta$ is the angle between the wall and the wireless link.

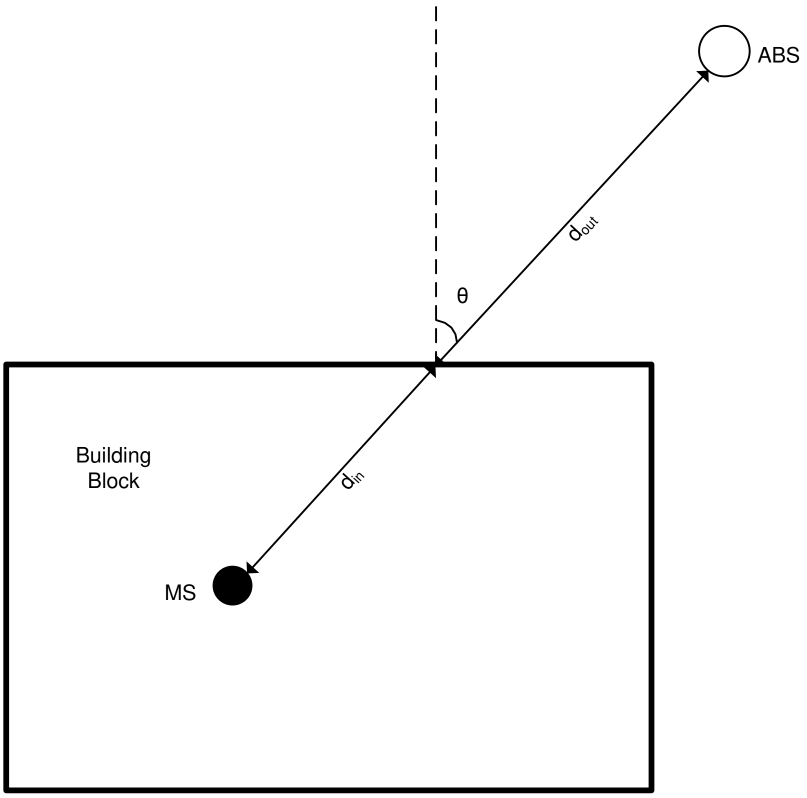


*Figure 8.9 WINNER II B4 Path Loss Calculation*

### 8.4.4 Traffic Model

The basic Poisson traffic model is used in the simulation to generate the traffic for both downlink and uplink [87]. The interarrival and service time of transmissions follow the negative exponential distribution. Note that this link level traffic model only generates user arrival time and user departure time based on the interarrival and service time, not the number of packets that need to be transmitted over a wireless link. Therefore, after a wireless link has been established, it is assumed that an entity will transmit data on a best effort basis.

### 8.4.5 Radio Resource Management

Two radio resource management strategies have been modelled, the fixed frequency planning proposed by Alvarion [22] and a cognitive radio based dynamic sub-channel assignment approach.

#### *8.4.5.1 Frequency Planning*

The details of the frequency plan are shown in figure 8.10. At the HBS side, 4 different channels are used for each group of 4 neighbouring beams in the order channel 1 to channel 4. ABSs located at the top and bottom of the cell are designed to serve N-S streets, and ABSs on the left and right serve the E-W streets. The two ABS beams pointing in opposite directions should use two different channels. ABSs that serve N-S streets use two different channels from those that serve E-W streets.

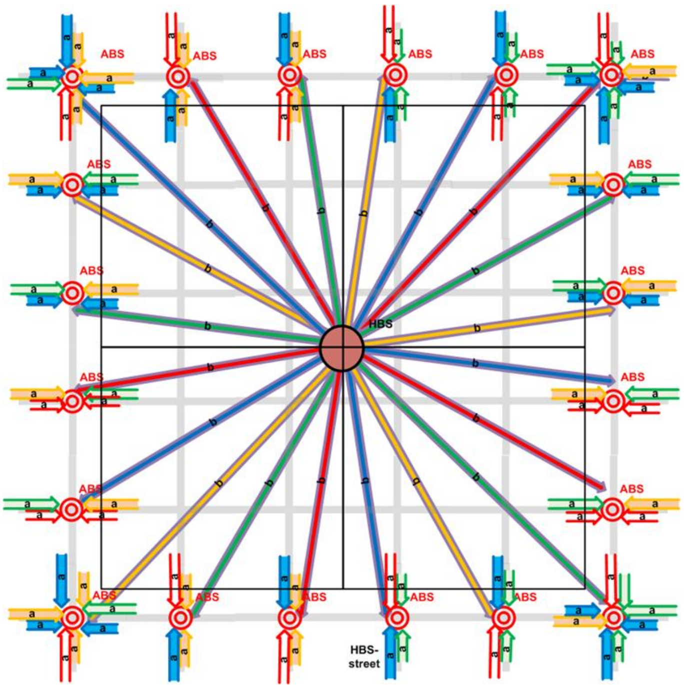


*Figure 8.10 Frequency Planning (directly reproduced from [22])*

Four 10 MHz channels for the backhaul network and four 10 MHz channels for the access network are utilized in the simulation. TDD duplexing is used and a 50%-50% split is assumed for the downlink and uplink.

#### *8.4.5.2 Cognitive Radio*

Two cognitive radio RRM approaches have been modelled in this work [107]. The first one is the most basic spectrum sensing only approach within the scope of the frequency plan. The only difference between this approach and the fixed frequency planning is that spectrum sensing has been assumed at the receiver side of a wireless link prior to the actual transmission. An interference threshold is used to examine the randomly selected subchannel. Available subchannels will be sensed one after another until the first available subchannel has been discovered.

The second approach is the reinforcement learning-based CR approach. The algorithm we introduced in section 8.3 is implemented. By exploring the historical information gained through the interaction with the environment, the entities are able to identify their preferred resources more efficiently.

## *8.5 Results*

This section presents the results obtained from the simulation. Three different RRM approaches we described in section 8.4.5 are implemented: 1.Pure frequency planning. 2. Frequency planning + spectrum sensing. 3. Spectrum sensing + learning. Results are given in forms of performance measures presented in the previous section. The results of the end-to-end links are shown in each figure to give the details of the system. Note that in the figures to show the system throughput and the throughput density, a green line has been added to show the maximum throughput that can be obtained theoretically at different offered traffic levels. It has been calculated by using equation 3-6 by assuming that $Thr_{MIMO-TSB}$ = 9 bps/Hz (the highest achievable data rate through MIMO) and $P_{TDD}$ = 0.5.

We have simulated a scenario where MSs are uniformly distributed over the entire service area. Note that MSs are assumed to be at street level only and no windows are assumed in the buildings (clearly a pessimistic case). The building layouts in most cities will in practice also mean that most users are closer to the streets than assumed here. All the indoor MSs are covered by the ABSs on streets, and no other approaches are assumed to provide indoor coverage, e.g. femto-cells. The propagation environment is very harsh for the indoor MS in this case. The majority of the MSs are placed indoors when a uniform MS density distribution is used over the service area. Thus, the results provided in this section will show the worst case system performance of Beyond Next Generation Mobile Network.

### 8.5.1 Frequency Planning

Figure 8.11 and figure 8.12 show the system throughput measurements of the downlink and the uplink respectively. It can be seen that at relatively low traffic levels when $OT_s$ is below 20, the system is able to transmit data at its maximum capacity. However when the offered traffic is increased further, the throughput of the access network degrades at both downlink and uplink due to interference, and the access network becomes the bottleneck of the system in terms of the end-to-end throughput.

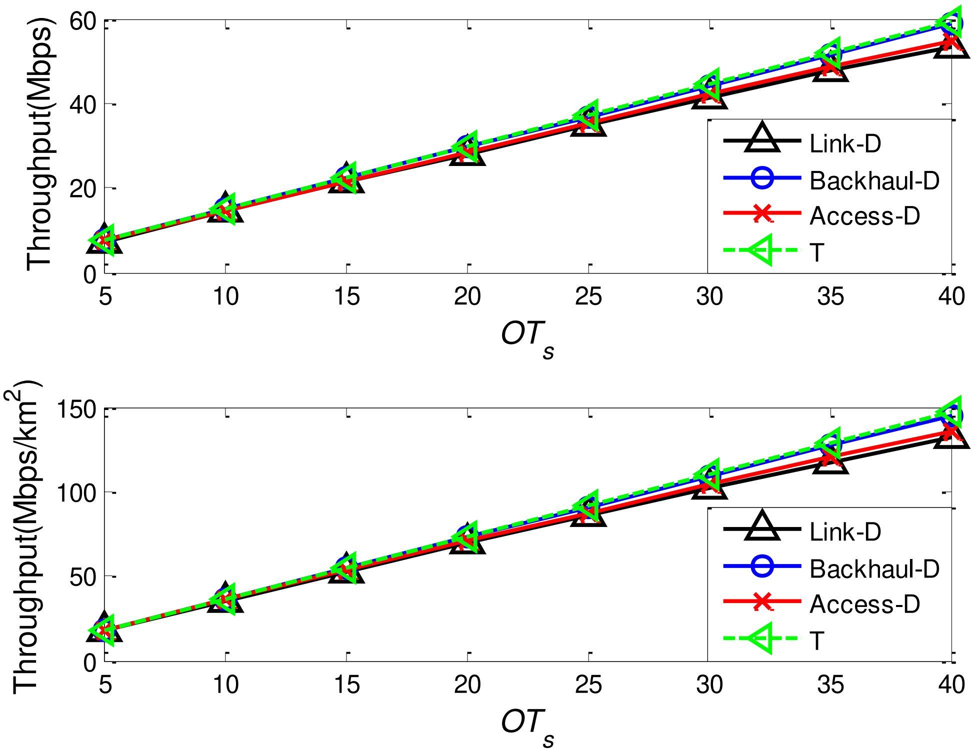


*Figure 8.11 System Throughput and Throughput Density-Downlink - 40 Users*

We can also see that the backhaul network is able to transmit data almost at the maximum level throughout the simulation. The degradation of the backhaul throughput at downlink and uplink is very limited. This means that by using frequency planning at the HBS, the interference between beams has been greatly reduced. The highest end-to-end link throughput is about 53 Mbps for the downlink and about 52 Mbps for the uplink when $OT_s$ is 40. That gives a throughput density of about 132 Mbps/km$^2$ downlink and 127 Mbps/km$^2$ uplink. The total throughput density is then about 0.26 Gbps/km$^2$. The theoretical maximum throughput is about 0.293 Gbps/km$^2$ in this case. Therefore, the degradation of the link throughput is not significant and the system can still obtain a throughput close to the maximum level. The throughput density is well below the targeted 1 Gbps/km$^2$ due to the insufficient coverage of in building MSs.

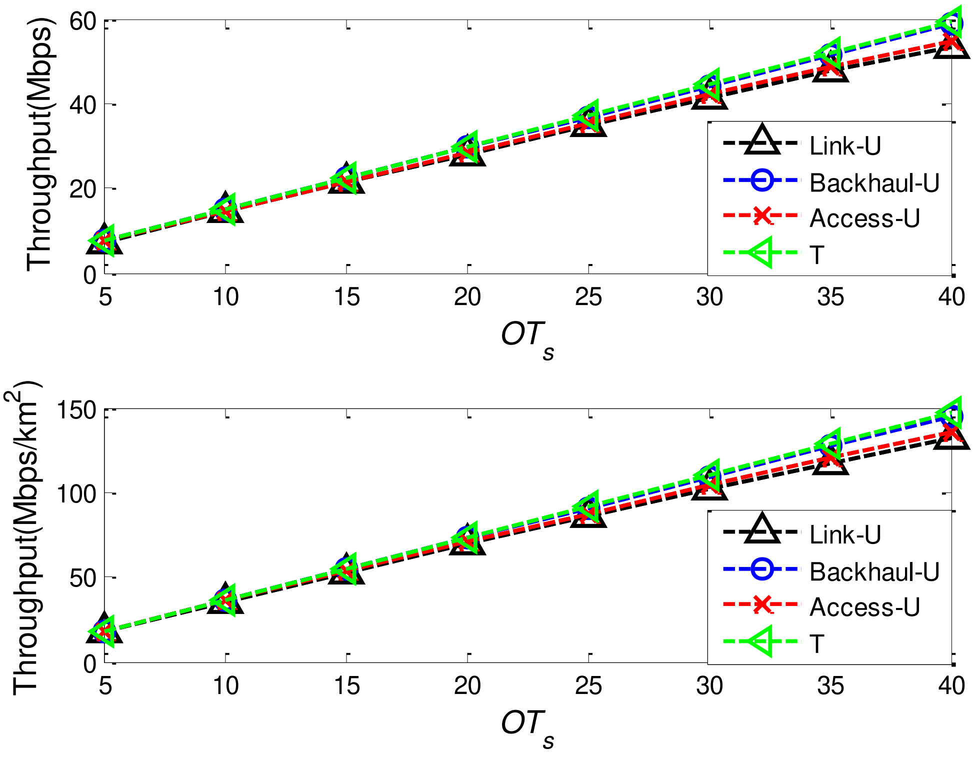


*Figure 8.12 System Throughput and Throughput Density- Uplink - 40 Users*

Figure 8.13 and figure 8.14 show the grade of service of the system, and these two figures explain why the simulation has been stopped at $OT_s$=40. Normally the blocking probability threshold is 5% and the dropping probability threshold is 0.5% for a wireless system [36]. The grade of service is considered to be poor if blocking probability and dropping probability are above these thresholds. Users are either struggling to find an appropriate channel to use or getting a large number of interruptions from others. Especially in a high dropping probability scenario where users are able to transmit data at a fairly high data rate, the transmission is very likely to be interfered by others. The blocking probability is about 2% both downlink and uplink when $OT_s$=40, and the dropping probability is also on the same level. These figures are too high, meaning that the service quality is actually very poor.

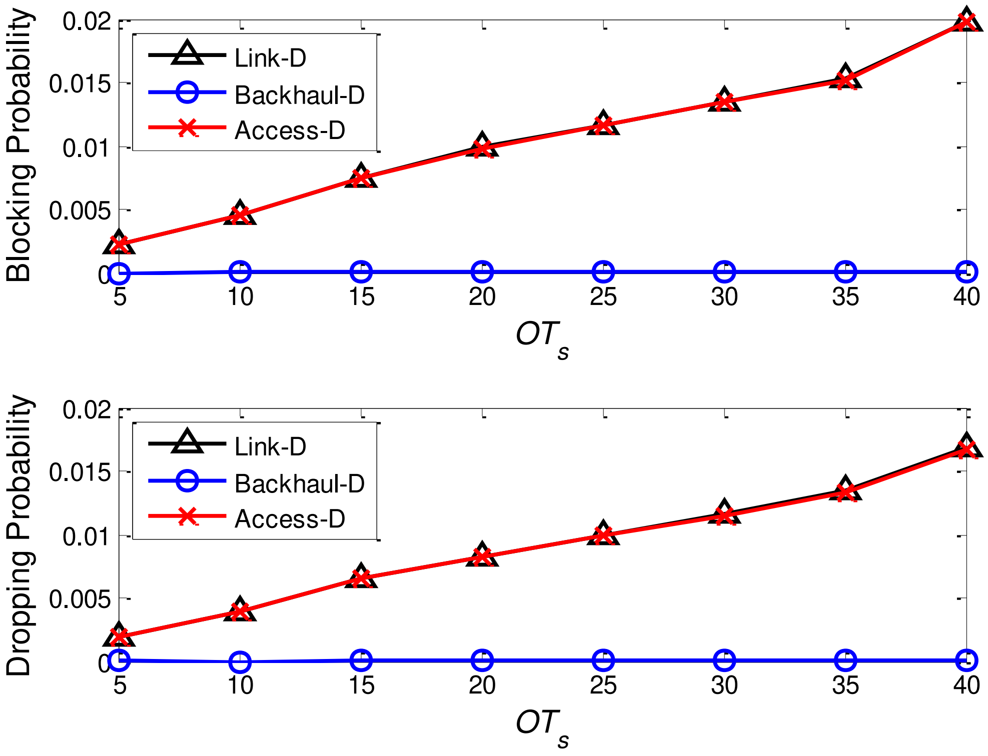


*Figure 8.13 Blocking Probability and Dropping Probability - Downlink - 40 Users*

The backhaul network is receiving significantly less interruption compared with the access network, i.e. entities in the backhaul network receive significantly less interference than the entities in the access network. Please note that the blocking probability and the dropping probability of the backhaul network are not zero in the figures. They are two orders of magnitude smaller than the access network, so the lines of the backhaul network are very close to the X axis. The blocking and dropping of the system are mainly caused by the access network in this case.

Thus, if we consider a scenario where the interference of the wireless system depend purely on the frequency plan and no further means of indoor coverage is assumed, the coverage provided by the base stations is hardly sufficient. More advanced RRM techniques are desirable that the frequency resource can be shared more efficiently. Cognitive radio based techniques have the potential to solve this problem.

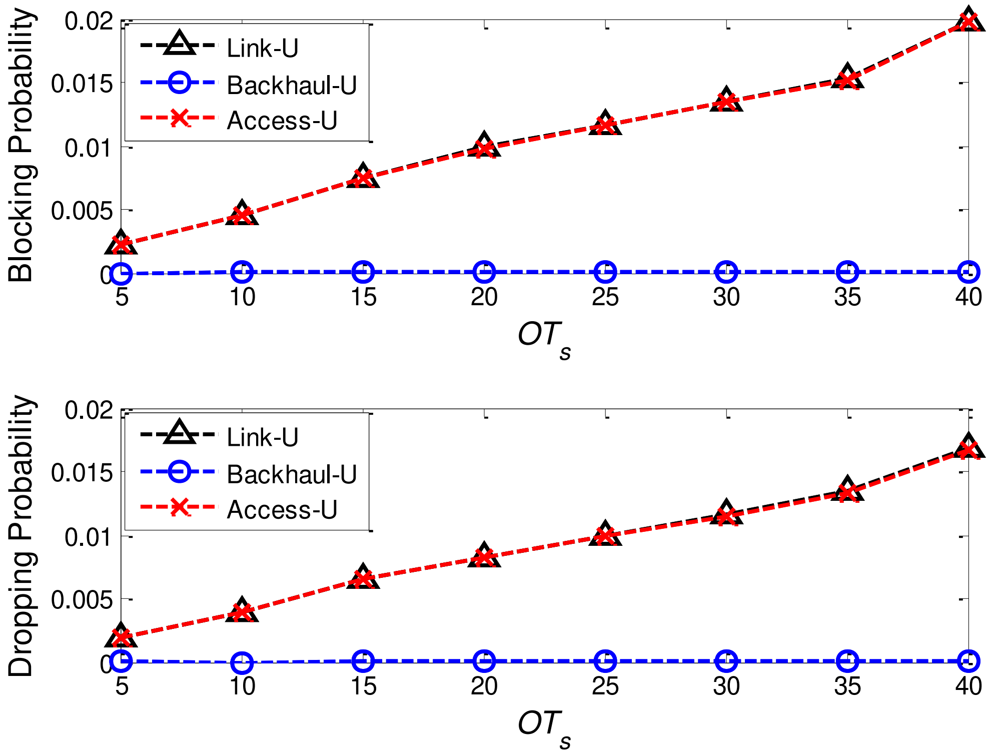


*Figure 8.14 Blocking Probability and Dropping Probability - Uplink - 40 Users*

### 8.5.2 Cognitive Radio Approaches

A very basic spectrum sensing approach has been modelled within the scope of the frequency plan given in this section to initially indicate the achievable throughput density of the cognitive approaches. Only the downlink has been simulated but the uplink performance is expected to be broadly the same as the downlink because a 50%-50% TDD is assumed and the downlink and the uplink will not interfere with each other. The performance of cognitive radio approaches are expected to be better than the pure frequency planning approach, hence a total number of 220 users has been used.

The reinforcement learning-based cognitive radio algorithm we introduced in section 8.3 is also implemented without any frequency plan. All frequency channels are equally available to all entities. Thus the complexity of radio resource management is significantly reduced. System performance of the learning-based approach has been compared with the pure frequency planning approach and the frequency planning + spectrum sensing approach.

Figure 8.15 illustrates the system blocking probability and dropping probability of three different schemes. By applying an interference threshold of -120 dBm to check the targeted channel prior to the transmission at the receivers and reinforcement learning-based channel assignment techniques, the dropping probability and the blocking probability of the cognitive approaches are significantly reduced.

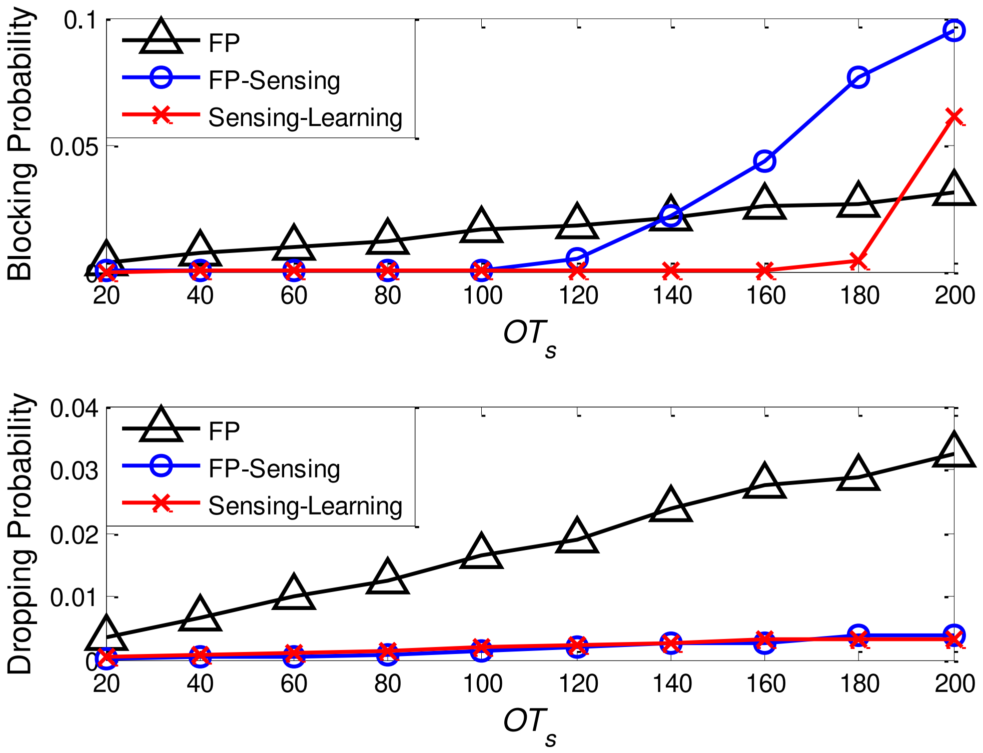


*Figure 8.15 System Blocking Probability and Dropping probability - Downlink - 220 Users*

The dropping probability of the two cognitive radio approaches remain at a very low level below 0.5% all through the simulation. The blocking probability of the frequency planning + spectrum sensing approach is below 5% when $OT_s$ is lower than 160. However, it increases significantly when $OT_s$ is above 120. By applying spectrum sensing, cognitive devices are able to avoid interfered channels that the dropping probability is greatly reduced. However, the blocking probability of the system is increased due to spectrum sensing, especially when the traffic load is high.

It can be seen that distributed reinforcement learning techniques are able to significantly reduce the blocking probability further while maintaining a very low dropping probability. The available spectrum pool is partitioned by distributed reinforcement learning in this case that the entities are able to discover their preferred resources more efficiently. The spectrum sensing + reinforcement learning approach is able to support a traffic level as high as $OT_s = 200$. In other words, the distributed reinforcement learning-based approaches are able to deliver a higher capacity under the same grade of service requirements.

Figure 8.16 shows the results of the downlink system throughput and the downlink throughput density. By applying a 5% blocking probability threshold and a 0.5% dropping threshold, the highest throughput the pure frequency planning approach can support is around 53 Mbps. The frequency planning + spectrum sensing approach is able to increase this figure to about 223 Mbps. The throughput density of the downlink is about 550 Mbps/km$^2$ in this case. The spectrum sensing + reinforcement learning approach performs the best that a throughput of 275 Mbps can be achieved. Comparing to the frequency planning + spectrum sensing approach, a 24% throughput increase is achieved by applying distributed reinforcement learning. This gives a downlink throughput density of approximately 680 Mbps/km$^2$. Again, a similar uplink performance can be expected so that an overall throughput density beyond 1.2 Gbps/km$^2$ is likely to be achieved.

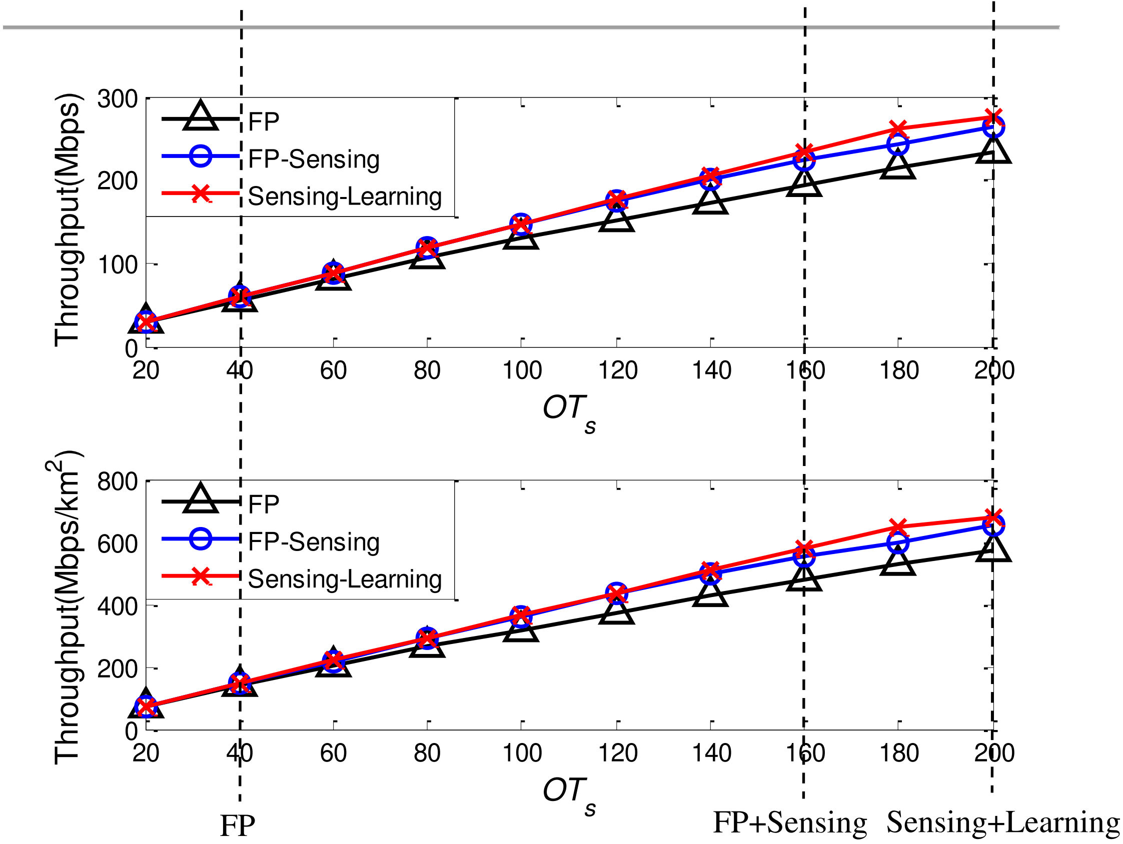


*Figure 8.16 System Throughput and Throughput Density- Downlink - 220 Users*

## *8.6 Conclusions*

Based on the techniques developed in previous chapters in this thesis, distributed reinforcement learning-based channel assignment techniques are developed and examined in the novel two-hop architecture for future Beyond Next Generation Mobile Network Systems. Performance of the learning-based approaches are compared with no learning approaches.

Three RRM approaches are modelled in this section: 1. frequency planning, 2. frequency planning with cognitive radio spectrum sensing, 3. Spectrum sensing with distributed reinforcement learning. In the pure frequency planning approach, the throughput density is about 132 Mbps/km$^2$ downlink and 127 Mbps/km$^2$ uplink. The total throughput density is about 0.26 Gbps/km$^2$. However, the grade of service is relatively poor, with blocking probability about 2% both downlink and uplink when $OT_s$=40, and the dropping probability is also on the same level. This is because a random subchannel assignment strategy is assumed without any further interference protection to the existing transmissions.

The frequency planning + spectrum sensing approach has also been modelled that the most basic spectrum sensing function is assumed at the receiver. The throughput density has been increased to 223 Mbps and the downlink throughput density is about 0.55 Gbps/km$^2$.

The blocking probability and dropping probability are much lower by introducing distributed learning-based techniques to further protect users from interference along with spectrum sensing.  It is shown that the spectrum sensing + reinforcement learning approach performs the best that a downlink throughput of 275 Mbps can be delivered under the same grade of service requirements. This gives a downlink throughput density of about 0.68 Gbps/km$^2$ and a overall throughput density beyond 1.2 Gbps/km$^2$ can be expected. The distributed reinforcement learning enables a spectrum partitioning in the service area that the entities are able to discover their preferred resources more efficiently.

# Chapter 9. Further Work

**Contents**

## *9.1 Dynamic Learning Techniques for Cognitive Radio System*

By combining the abilities of spectrum awareness, intelligence and radio flexibility, a cognitive radio is able to adapt itself to the changes in the local environment. However, unlike most cases studied in Computer Science, the surrounding environment of the learning-based wireless system is constantly changing, e.g. user location, user density, traffic load, etc. Thus, a fixed learning strategy is not likely to deliver the best performance. Dynamic learning algorithms that are able to adapt themselves to the changes of the environment are desirable.

The learning-based algorithms we developed in this work have the potential to be improved further if the learning algorithms are optimized based on the 'live' information observed by the system. In other words, the learning algorithms are optimized from time to time according to the changes of the environment.

A few aspects of the learning algorithm have the potential to be optimized. Firstly, the learning parameters can be optimized according to the changes of environment. In chapter 4 we studied the influence of weighting factors in a general way. By linking the learning parameters with the environment, the system performance has the potential to be improved further. It is also possible to better balance the exploration-exploitation trade-off by dynamically adjusting the setting of the preferred resource set according to the environment which in turn achieves an improved performance.

## 9.2 Low Complexity Cognitive Radio System

With the rapid development of wireless communication techniques, the communication system is increasingly complex. A Cognitive radio approach is designed to incorporate software defined radio, environment awareness, self awareness and dynamic decision making which are all extremely challenging research issues. It is foreseen that the ‘full’ Mitola cognitive radio [7] is still decades from implementation due to such complexity.

The development of low complexity cognitive radio which will deliver much of the functionality of the full cognitive radio can be expected. By combining appropriate intelligence with a few essential elements of cognitive radio, the system might be able to perform well in most of the communication tasks. The research work introduced in chapter 7 showed that a near optimal system performance can be expected by applying low complexity learning strategies. Thus, the information provided in chapter 7 is useful for future research on this topic.

A top level examination of the existing cognitive radio approach is required, to understand the complexity of different components of cognitive radio and whether such complexity is appropriate. Based on the previous research outcomes we have obtained, low complexity intelligent strategies could be developed to enable a rapid and wide implementation of cognitive radio.

## 9.3 Learning-based Channel and Power Joint Allocation

This thesis primarily studies the learning-based spectrum sharing algorithms. The learning-based devices explore time-space-frequency 3-dimensional opportunities by utilizing learning. It is assumed that the transmission power level is fixed. Better performance could be expected if the learning-based adaptive power control could also be carried out along with the learning-based channel assignment. The learning algorithm we developed in this work has the potential to be used jointly for both channel assignment and transmit power control.

It is also possible to develop a two-phase learning approach for joint allocation of channel and transmit power. Different phases of the learning algorithm are concerned with the allocation of channels and transmit power separately. The information learned in different phases could be shared or completely isolated depending on the system performance.

## *9.4 Multi-Learning Cognitive Radio*

Reinforcement learning has been the only intelligent approach we have considered so far due to its natural fitness with the distributed cognitive radio scenario. However they are many other learning approaches, like supervised learning and unsupervised learning, could be used along with reinforcement learning to deliver a better performance.

Different 'intelligent' techniques could be applied to different aspects of cognitive radio system in line with reinforcement learning to achieve a better performance. Take Game Theory for example, Game Theory studies how an individual can make choices depending on the choices of others. It is a strong candidate to be used as the decision making techniques for cognitive radio. In other words, game theory is used to obtain the best choices of resources (channel, transmission power, etc) for every transmission task based on the information learned by the device and the instantaneous measurements of the system. Then the reinforcement learning algorithm learns from the decisions and updates the knowledge base accordingly. The Game Theory part ensures the user select the best action when it needs to perform a task and the reinforcement learning part makes sure that the selection of actions are learnt by the user. In this scenario, game theory concerns the short term performance where reinforcement learning is responsible for the long term performance.

Although this thesis concerns only the distributed reinforcement learning approach, the research work in this thesis could be the starting point to further develop multi-learning algorithms for cognitive radio. Multi-leaning users have the potential to interact with others more smoothly. Different learning techniques deal with different tasks that the benefits of applying learning are maximized.

## *9.5 'Docitive' Approach for Learning-based Cognitive Radio*

The learning-based cognitive radio systems we introduced in this work take a relatively long time to converge to the optimal or sub-optimal point [88, 102]. A number of techniques, like exploration control and efficient exploration, have been developed to improve the learning efficiency [89, 95]. However, the slow convergence is still one of the most difficult challenges seen in such multi-agent learning system. This is especially true in a cognitive radio scenario where the devices work in a fully distributed fashion and the environment they are interacting with is constantly changing.

Thus, it is desirable to allow learning information to be exchanged between neighbouring devices, improving the convergence speed of the users. It is also important to ensure that the information exchanged between devices is kept at a minimum level. 'Docition' is a word which effectively means teaching [108]. It is a process where users are teaching others the experience they gained through learning. Therefore they are able to learn much more efficiently. By exchanging limited information between learning devices, the convergence performance is likely to be significantly improved and a better overall performance can be expected.

# Chapter 10. Summary and Conclusions

**Contents**

## *10.1 Summary and Conclusions*

This thesis has investigated the fundamental issues in applying reinforcement learning to cognitive radio spectrum sharing. Firstly, a generic reinforcement learning model has been proposed to cognitive radio along with a linear value function. Then, a two-stage algorithm has been introduced to properly balance the exploration versus exploitation trade-off seen in reinforcement learning-based systems. Two efficient exploration techniques: pre-partitioning and weight-driven exploration, have been proposed to improve the system performance further. This work has attracted lots of interests and contributed to a number of research works on the relevant topics collaboratively with Zhejiang University, CTTC, and the University of Sydney [109-110]. This work has also contributed to other research works within the Communications Research Group on the topics including CSMA based cognitive radio systems [111-112] and multicasting cognitive radio systems [113]. This work has also directly contributed to the FP7 BuNGee project. A brief summary and conclusions for the thesis are given below:

The first chapter provides a brief introduction to the thesis and the purpose of this work. Chapter 2 presents a comprehensive literature review on the research related to this work. This focuses mainly on intelligent radio resource management schemes, which take advantage of the artificial intelligence algorithms developed mainly in the area of Computer Science.

Chapter 3 describes the research methodologies, simulation techniques and the key measurements used to evaluate the performance. The Monte Carlo simulation approach has been used extensively in this work. Matlab is used as a tool to carry out the simulation tasks. Blocking probability and dropping probability are the main performance measurements we used in this thesis in the early stage. Throughput and throughput density are used later along with blocking probability and dropping probability to evaluate the system performance of the dual-hop beyond next generation mobile network.

In Chapter 4 we firstly introduced the reinforcement learning model for cognitive radio system. The learning model proposed in section 4.2 is the basis of this work since the goal of our work is not only to develop intelligent spectrum sharing algorithms for cognitive radio but more importantly to build a generic reinforcement learning model eventually. After the introduction of the learning model, the value function and the weighting factors were defined.

It is shown that reinforcement learning-based approaches perform better than the non-learning cognitive radio algorithms. By utilizing the ability of learning, cognitive devices exploit their preferred resources with a higher priority. This enables an autonomous partition of the available spectrum. By autonomously partitioning the local available channel pool, the channel usage of different users converges to different channels. Thus, interference can be reduced. The value function and the weighting factors are the basis to assess the success level of the performed actions. The results also show that the settings of weighting factor values have significant influence on the system performance. Weighting factor values need to be properly defined based on the characteristics of the wireless system in order to achieve better performance.

One of the fundamental challenges seen in reinforcement, the trade-off between exploration and exploitation, has been examined in the context of cognitive radio in chapter 5. A learning cognitive radio needs to explore the wireless environment to find available resources. Meanwhile, the cognitive radio also has to exploit the resources discovered in exploration to obtain enough experience to distinguish between good and bad options. The trade-off between exploration and exploitation needs to be balanced in order to improve the performance of the cognitive radio system.

A two stage reinforcement learning-based algorithm has been proposed in this chapter to control the trade-off between exploration and exploitation. A ‘warm up’ stage is proposed where distributed cognitive radio users search for optimum resources and learn from the experience of searching. Once users have obtained a set of preferred resources, they exploit the preferred resources with higher priority and stop searching for new channels. It is shown in this chapter how the balance between exploration and exploitation is not only theoretically important but also crucial to a cognitive radio system in practice.

It can be seen from the simulation results that a quick and efficient channel partitioning can be obtained by using a small preferred channel weight threshold. Moreover, either an overly small size of preferred resource set or an overly big size will cause more system interruptions rather than sharing spectrum peacefully, and an optimal spectrum sharing policy will not be discovered consequently.

The purpose of Chapter 6 is to introduce efficient exploration techniques which are able to reduce the exploration phase of the learning users even further. Cognitive radio users will receive a higher level of interference when the majority of the users are exploring their available spectrum space. The two-stage algorithm proposed in the previous chapter is used as a basis.

Two novel approaches are presented, pre-partitioning and weight-driven exploration, to enable efficient exploration in the context of cognitive radio. The learning efficiency of a learning-based cognitive radio has been defined and investigated. The pre-partitioning scheme randomly reserves a certain amount of spectrum resources for each user. The available action space which the cognitive radio needs to explore is then significantly reduced, which in turn shortens the exploration stage significantly. In the weight-driven exploration scheme, the exploitation phase is gradually moved into exploration by applying a weight-driven probability distribution to influence action selection during exploration. The exploration is more efficient and the overall performance of the cognitive radio system has been improved. Results show that efficient exploration techniques improve the system performance significantly compared with the commonly used uniform random exploration approach and the weight-driven exploration scheme achieves the best performance.

Chapter 7 explores the ‘green’ aspect of the proposed learning-based schemes, concentrating on the power consumption reduction achieved by learning. This is done by reducing the requirement for spectrum sharing through reinforcement learning. Cognitive radio needs to either periodically or continuously sense the spectrum to obtain the information of the environment. It is foreseen that spectrum sensing will be one of the main sources of energy consumption within a cognitive radio device. By utilizing reinforcement learning, cognitive users are able to identify the appropriate channel quicker since they start with their preferred channels. Thus, the time and power consumed by spectrum sensing are reduced.

It is shown that by acquiring a subset of preferred resources, the restricted sensing scheme and the minimum sensing scheme are able to significantly reduce the need for spectrum sensing. The efficient exploration based algorithms are able to reduce the requirement of spectrum sensing further. Weight-driven exploration scheme achieves the highest spectrum sensing reduction in all the proposed schemes.

In chapter 8, the learning-based techniques are implemented in a novel two-hop architecture for beyond next generation mobile network. The system model and the propagation environment are very complex since firstly the system is designed for the dense city center area where the a large number of building blocks could be found, and secondly several types of directional antenna are used along with advanced MIMO techniques. Thus, a very detailed simulator is developed to model the beyond next generation mobile network and its surrounding environment.

Distributed reinforcement learning-based channel assignment techniques are developed for the dual-hop system. A single learning engine processes the information for both hops of the wireless link simultaneously in this case. Three radio resource management approaches are modelled in this section: 1. frequency planning, 2. frequency planning with cognitive radio spectrum sensing, 3. Spectrum sensing with distributed reinforcement learning. Results show that the basic spectrum sensing + frequency planning approach outperforms the pure frequency planning approach. The learning-based cognitive radio approach not only achieves the best performance but also removes the need for frequency planning completely. Thus, the proposed approach provides the targeted capacity to the users while significantly reducing the resource management complexity.

## *10.2 Summary of Novel Contributions*

This thesis concentrates on different aspects of the application of learning-based techniques to cognitive radio system. Very limited work had been carried out on this topic by Bublin [59] before the starting point of this work. The novel contributions of this thesis are highlighted in this section. Most of the work has been published in a number of journal and conference papers. Some of the work has contributed to the EU FP7 BuNGee project. A publication list is also provided in this thesis.

### 10.2.1 Distributed Reinforcement Learning-based Channel Assignment for Open spectrum Cognitive Radio

- The concept of applying distributed reinforcement learning techniques to cognitive radio system is perhaps the most significant novel contribution of this work. Although learning had been considered an essential part of cognitive radio before we started this work, it was not clear where and how we could apply machine learning techniques to the cognitive radio system. Research work which directly shows the benefits of applying machine learning techniques to cognitive radio system was very limited back to the time when we started this work.

- The similarity between the behaviour of reinforcement learning nodes and cognitive radios has been identified in the early stage of this work: they all work in a distributed fashion and they are all interacting with an ‘unknown’ environment. Thus, reinforcement learning has been considered a perfect tool for cognitive radio.

- In chapter 4, a novel generic reinforcement learning model has been proposed to cognitive radios. The original reinforcement learning model has been modified in order to fit with the cognitive radio scenario. A linear function has been proposed as the value function in this chapter to update the knowledge base.

- Another contribution is that we examine the performance of learning-based cognitive radios in an open spectrum scenario where the entire spectrum is fully shared, where radio regulations are sufficiently light-touch to give all services equal opportunity to use the spectrum. Such a scenario is seen today to a limited extent in the unlicensed bands.

These contributions have been published by *IET Communications* [89]. The learning model and the value function proposed in this work have also been used by a number of other papers on different topics [111-114].

### 10.2.2 Impact of Weighting Factors

The impact of weighting factors has been investigated in chapter 4. Weighting factors have great influence on the system performance, it reflects the degree of responses of a learning agent towards the changes of environment, i.e. a high reward or punishment value means that the learning node will adjust its actions swiftly according to the changes of the wireless environment, and a mild reward or punishment means that the learning node is adapting itself gradually based on the interactions with the environment. It is crucial that weighting factor values are defined properly. Three different strategies have been discuss in chapter 4 and the contributions has been published in [88] in *3rd International Conference on Communications and Networking in China (ChinaCom).*

### 10.2.3 Exploration-Exploitation Trade-off Control for Learning-based Cognitive Radio

The exploration-exploitation trade-off seen in reinforcement learning has been tackled in the context of cognitive radio for the first time in chapter 5. A novel two-stage algorithm has been proposed in this chapter. A ‘warm up’ stage is suggested where distributed cognitive radio users search for optimal resources and learn from the experience of searching. Once users have obtained a set of preferred resources, they will only sense the spectrum with higher priority prior to establishing communications. The performance of the exploration and the exploitation phases are investigated. Results show that the system performance in the exploration phase is worse since exploring users cause a higher level of disturbance to the environment. Results in chapter 5 also show that the novel two-stage spectrum sharing algorithm is able to practically control the exploration phase by adjusting the setting of preferred resource set. These contributions have been published in [89] by *IET Communications*.

### 10.2.4 Efficient Exploration Techniques for Cognitive Radio

Chapter 6 develops novel efficient exploration techniques for reinforcement learning-based cognitive radio. Previous published works apply the most basic uniform random exploration techniques, which is not likely to be the best exploration strategy. Two novel techniques have been proposed: Pre-partitioning and Weight-driven exploration: by randomly reserving a subset of available spectrum, the spectrum pool is fully partitioned before transmissions start; weight-driven exploration utilizes the information gained in exploration to guide the exploration process itself, and weight-driven exploration also ensures exploration by merging randomness into action selection. Results in this chapter show that the proposed techniques are able to significantly reduce the exploration phase, which in turn delivers a better system performance.

The contributions in this chapter have been partly published in [95] by *IET Communications*, and partly in [94] in *IET International Communication Conference on Wireless Mobile and Computing* (This paper has also received the Best Paper Award).

### 10.2.5 Reduction of Spectrum Sensing Power Consumption

One of the most important novel contributions of the proposed algorithms is the reduction of the requirement for spectrum sensing. The ‘Green’ aspect of this work is discussed in chapter 7. Cognitive radio needs to either periodically or continuously sense the spectrum to obtain the information of the environment. It is expected that spectrum sensing will be one of the biggest power-consuming sources in a cognitive device. The algorithms we developed in this work make spectrum sensing more efficient, meaning that cognitive radio can sense fewer channel by exploiting the experience gained through learning. It is even possible to directly assign channel based on learning without sensing when the local spectrum is autonomously partitioned. Results in chapter 7 show that by acquiring a subset of preferred resources, the restricted sensing scheme and the minimum sensing scheme are able to significantly reduce the need for spectrum sensing. The efficient exploration based algorithms reduces the requirement for sensing even further.

The ideas and the novel contributions in this chapter have been published in [102] in *IET Seminar on Cognitive Radio and Software Defined Radios: Technologies and Techniques* and in [103] in *4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM).*

### 10.2.6 Reinforcement Learning-based Channel Assignment for Dual-Hop Beyond Next Generation Mobile Network

In chapter 8, for the first time reinforcement learning-based radio resource management techniques have been developed for the dual-hop beyond next generation mobile networks in a highly populated city centre area. Chapter 8 provides the information on how the reinforcement learning-based techniques could be used along with other advanced techniques like multi-beam directional antenna and MIMO. A novel cognitive radio approach has been proposed where a single learning engine processes the information for both hops of the wireless link simultaneously. It shows that learning-based channel assignment schemes not only achieves the highest throughput density but also significantly reduces the complexity of the radio resource management.

The work in this chapter has directly contributed to the *EU FP7 Beyond Next Generation Mobile Broadband Project* and has been published in the project deliverable [105, 107]. The work in this chapter has also contributed to ETSI BRAN Work Item of TR 101 534 (DTR/BRAN-0040008).

### 10.2.7 System Level Modelling of MIMO Techniques

In order to sufficiently model the beyond next generation mobile network systems and evaluate the system performance, a novel approach of modelling MIMO techniques at the system level has been developed collaboratively within the FP7 Beyond Next Generation Mobile Network Project.

In SISO channels the achievable throughput is a deterministic value based on the path loss, shadowing and small-scale fading; and this can be evaluated easily. However in MIMO channels, throughput is no longer a deterministic value. Instead the achievable rate follows a distribution.

A method has been developed jointly that maps each value of average link SINR to a statistical distribution of achievable rates. For a useful range of average SINR values, we obtained offline the empirical distributions of the achievable data rates. In the system level simulation the simulator randomly take a value from the empirical throughput distribution based on the SINR value of the transmission link.

The related work in chapter 8 has directly contributed to the *EU FP7 Beyond Next Generation Mobile Broadband Project* and has been published in the project deliverable [105].

### 10.2.8 Dual-Hop Wireless System Simulator

A novel two-layer modularized simulator has been developed in this work to accurately capture the features of the dual-hop Beyond Next Generation Mobile system architecture. The first layer models the self-backhaul link between HBS and HSS. The second layer models the access link between ABS and MS. The modelling of the dual-hop system is a very complex task that needs to address a range of issues across the Physical, MAC and Network layers. A novel modularized structure is used to maximize the flexibility of the simulator and its compatibility with future developments. A number of modules are developed to model different aspects of the system, including a location module, traffic module, propagation module, MIMO module and RRM module. Thus, the advanced features of the system, from the PHY layer to the network layer, have all been captured by the simulator.

The related work in chapter 8 has directly contributed to the *EU FP7 Beyond Next Generation Mobile Broadband Project* and has been published in the project deliverable [105].

# Publications

*Book Chapter*

T. Jiang, D. Grace,: ‘Reinforcement Learning-based Cognitive Radio for Open Spectrum Access’, ***Cognitive Communications: Distributed Artificial Intelligence (DAI), Regulatory Policy & Economics, Implementation****,* Wiley, 2011. (accepted)

*Journal*

T. Jiang, D. Grace, and P.D. Mitchell,: ‘Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing’, ***IET Communications*.,** Volume 5, Issue 10, pp.1309-1317, July 2011, DOI:10.1049/iet-com.2010.0258

T. Jiang, D. Grace, and Y. Liu,: ‘Two-stage reinforcement-learning-based cognitive radio with exploration control’, ***IET Communications*.,** Volume 5, Issue 5, pp.644-651, March 2011, DOI:10.1049/iet-com.2009.0803

X. Chen, Z. Zhao, T. Jiang, D. Grace, and H. Zhang,: ‘Inter-cluster connection in cognitive wireless mesh networks based on intelligent network coding’ ***EURASIP Journal on Advances in Signal Processing - Special issue on dynamic spectrum access for wireless networking***, March 2009, DOI=10.1155/2009/141097

*Conference*

T. Jiang, D. Grace, and P.D. Mitchell,: ‘Improvement of Pre-partitioning on Reinforcement Learning Based Spectrum Sharing’, ***IET International Communication Conference on Wireless Mobile and Computing (CCWMC)***, pp.299-302, 2009 (**Best paper award**)

D. Grace, J. Chen, T. Jiang, and P.D. Mitchell,: ‘Using Cognitive Radio to Deliver 'Green' Communications’, ***4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)***, pp.1-6, 2009

X. Chen, Z. Zhao, H. Zhang, T. Jiang, and D. Grace,: ‘Inter-Cluster Connection in Cognitive Wireless Mesh Networks Based on Intelligent Network Coding’, ***IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)***, pp.1251-1256, 2009

T. Jiang, D. Grace, and Y. Liu,: ‘Cognitive radio spectrum sharing schemes with reduced spectrum sensing requirements’, ***IET Seminar on Cognitive Radio and Software Defined Radios: Technologies and Techniques***, September, 2008, London

T. Jiang, D. Grace, and Y. Liu,: ‘Performance of Cognitive Radio Reinforcement Spectrum Sharing Using Different Weighting Factors’, ***Third International Conference on Communications and Networking in China (ChinaCom)***, pp.1195-1199, 2008

*FP7 BuNGee Project Deliverables*

**ICT-BuNGee D3.1**, Y. Leiba, et al., ‘Baseline RRM & JointAccess/Self-Backhaul Protocols’, May 2011, Available: http://www.ict-bungee.eu/

**ICT-BuNGee D4.1.1**, T. Jiang, A. Papadogiannis, D. Grace, A. Burr,: ‘Interim Simulation’, February 2011, Available: http://www.ict-bungee.eu/

**ICT-BuNGee D1.2**, M. Goldhamer, et al., ‘Baseline BuNGee Architecture’, November 2010, Available: http://www.ict-bungee.eu/

# Bibliography

[1] FCC, "Notice of proposed rule making and order," ET Docket No 03-222, Dec, 2003.

[2] I. F. Akyildiz*, et al.*, "Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer Networks,* vol. 50, pp. 2127-2159, Sep, 2006.

[3] J. D. Gibson, *The Mobile Communications Handbook*, 1st ed.: IEEE Press, 1996.

[4] J. Mitola and G. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Communication,* vol. 6, pp. 13-18, Aug, 1999.

[5] S. Haykin, "Cognitive Radio: Brain-Empowered Wireless Communications," *IEEE Journal on selected areas in communications,* vol. 23, pp. 201-220, Feb, 2005

[6] B. Fette, *Cognitive Radio Technology*: Newnes, 2006.

[7] J. Mitola, "Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio," Ph.D., Teleinformatics, Royal Institute of Technology (KTH), May, 2000.

[8] J. Mitola, *Cognitive Radio Architecture: The Engineering Foundations of Radio XML*: Wiley, 2006.

[9] I. Katzela and M. Naghshineh, "Channel Assignment Schemes for Cellular Mobile Telecommunication Systems : A Comprehensive Survey," *IEEE Personal Communication,* vol. 3, pp. 10 - 31, Jun, 1996.

[10] S. Tekinay and B. Jabbari, "Handover and channel assignment in mobile cellular networks," *Communications Magazine, IEEE,* vol. 29, pp. 42-46, 1991.

[11] T. Kahwa and N. Georganas, "A Hybrid Channel Assignment Scheme in Large-Scale, Cellular-Structured Mobile Communication Systems," *IEEE TRANSACTIONS ON Communications,* vol. 26, pp. 432-438, April 1978.

[12] K. N. Sivarajan*, et al.*, "Dynamic channel assignment in cellular radio," in *Vehicular Technology Conference, 1990 IEEE 40th*, 1990, pp. 631-637.

[13] K. Okada and F. Kubota, "On dynamic channel assignment in cellular mobile radio systems," in *Circuits and Systems, 1991., IEEE International Sympoisum on*, 1991, pp. 938-941 vol.2.

[14] R. Beck and H. Panzer, "Strategies for handover and dynamic channel allocation in micro-cellular mobile radio systems," in *Vehicular Technology Conference, 1989, IEEE 39th*, 1989, pp. 178-185 vol.1.

[15] A. Gamst, "Some lower bounds for a class of frequency assignment problems," *Vehicular Technology, IEEE Transactions on,* vol. 35, pp. 8-14, 1986.

[16] K. Sallberg*, et al.*, "Hybrid channel assignment and reuse partitioning in a cellular mobile telephone system," in *Vehicular Technology Conference, 1987. 37th IEEE*, 1987, pp. 405-411.

[17] R. W. Nettleton and G. R. Schloemer, "A high capacity assignment method for cellular mobile telephone systems," in *Vehicular Technology Conference, 1989, IEEE 39th*, 1989, pp. 359-367 vol.1.

[18] M. Serizawa and D. J. Goodman, "Instability and deadlock of distributed dynamic channel allocation," in *Vehicular Technology Conference, 1993 IEEE 43rd*, 1993, pp. 528-531.

[19] J. Zander, "Radio resource management in future wireless networks: requirements and limitations," *Communications Magazine, IEEE,* vol. 35, pp. 30-36, 1997.

[20] D. Grace*, et al.*, "Reducing call dropping in distributed dynamic channel assignment algorithms by incorporating power control in wireless ad hoc networks," *Selected Areas in Communications, IEEE Journal on,* vol. 18, pp. 2417-2428, 2000.

[21] D. Čabrić*, et al.*, "A Cognitive Radio Approach for Usage of Virtual Unlicensed Spectrum," presented at the 14th IST Mobile Wireless Communications Summit, Dresden, Germany, June 2005.

[22] M. Goldhamer*, et al.* BuNGee D1.2 Baseline RRM & JointAccess/Self-Backhaul Protocols. 2011. Available: http://www.ict-bungee.eu/

[23] R. J. Berger, "Open Spectrum: A Path to Ubiquitous Connectivity," *Queue,* vol. 1, pp. 60-68, 2003.

[24] O. Yu, *et al.*, "Dynamic Control of Open Spectrum Management," in *Wireless Communications and Networking Conference, 2007.WCNC 2007. IEEE*, 2007, pp. 127-132.

[25] K. V. Katsaros, *et al.*, "Design challenges of open spectrum access," in *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium on*, 2008, pp. 1-5.

[26] X. Yiping, *et al.*, "Dynamic spectrum access in open spectrum wireless networks," *Selected Areas in Communications, IEEE Journal on,* vol. 24, pp. 626-637, 2006.

[27] D. Grace, *et al.*, "The Effects of Interference Threshold and SINR Hysteresis on Distributed Channel Assignment Algorithms for UFDMA," presented at the IEEE International Conference on Universal Personal Communications (ICUPC), 1997.

[28] "Cognitive Radio Technology:A Study for Ofcom," Summary Report, QinetiQ Ltd, Feb, 2007.

[29] ITU-R. WRC-12 Agenda Item 1.19: Software-Defined Radio (SDR) and Cognitive Radio Systems (CRS). 2010. Available: http://www.itu.int/ITU-R/information/promotion/e-flash/4/article5.html

[30] C. Cordeiro, *et al.*, "IEEE 802.22: the first worldwide wireless standard based on cognitive radios," presented at the Dynamic Spectrum Access Networks (DySPAN) 2005.

[31] T. M. Mitchell, *Machine Learning*: McGraw-Hill, 1997.

[32] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*: Pearson Education, 2003.

[33] Z. Ghahramani, "Unsupervised Learning," ed: University College London, September 2004.

[34] R. S. Sutton and A. G. Barto, *Reinforcement learning : An Introduction*: The MIT Press, 1998.

[35] L. P. Kaelbling, *et al.*, "Reinforcement Learning: A Survey," *Journal of artificial intelligence Research,* vol. 4, pp. 237-285, May. 1996.

[36] D. Akerberg, "On Channel Definitions and Rules for Continuous Dynamic Channel Selection in Coexistence Etiquettes for Radio Systems," in *IEEE Vehicular Technology Conference*, Stockholm, June 1994, pp. 809-813.

[37] M. M.-L. Cheng and J. C.-I. Chuang, "Performance Evaluation of Distributed Measurement-Based Dynamic Channel Assignment in Local Wireless Communications," *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS,* vol. 14, pp. 698-710, May,1996.

[38] G. J. Foschini and Z. Miljanic, "Distributed Autonomous Wireless Channel Assignment Algorithm with Power Control," *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY,* vol. 44, pp. 420-429, August 1995.

[39] S. R. Saunders and F. R. Bonar, "Prediction of mobile radio wave propagation over buildings of irregular heights and spacings," *Antennas and Propagation, IEEE Transactions on,* vol. 42, pp. 137-144, 1994.

[40] A. Law*, et al.*, "Comparison of performance of FA and LIC DCA call assignment within DECT," in *Networking Aspects of Radio Communication Systems , IEE Colloquium on*, 1996, pp. 4/1-4/6.

[41] J. C.-I. Chuang, "Autonomous Adaptive Frequency Assignment for TDMA Portable Radio Systems," *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY,* vol. 40, pp. 627-635, August 1991.

[42] J. C.-I. Chuang, "Performance Issues and Algorithms for Dynamic Channel Assignment," *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS,* vol. 11, pp. 955-963, August 1993.

[43] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," *Vehicular Technology, IEEE Transactions on,* vol. 41, pp. 57-62, 1992.

[44] J. F. Whitehead, "Performance and capacity of distributed dynamic channel assignment and power control in shadow fading," in *Communications, 1993. ICC 93. Geneva. Technical Program, Conference Record, IEEE International Conference on*, 1993, pp. 910-914 vol.2.

[45] R. W. Thomas*, et al.*, "Cognitive Networks: Adaptation and Learning to Achieve End-to-End Performance Objectives," *IEEE Communications Magazine,* vol. 44, pp. 51-57, Dec 2006

[46] A. N. Mody*, et al.*, "Recent Advances in Cognitive Communications," *IEEE Communications Magazine,* vol. 45, pp. 54 - 61 October 2007.

[47] C. Clancy*, et al.*, "Applications of Machine Learning to Cognitive Radio Networks," *IEEE Wireless Communications,* vol. 14, pp. 1536-1284, August 2007.

[48] J. Nie and S. Haykin, "A Dynamic Channel Assignment Policy Through Q-Learning," *IEEE Transactions on Neural Networks,* vol. 10, pp. 1443-1455, NOV. 1999.

[49] S.-M. Senouci and G. Pujolle, "Dynamic channel assignment in cellular networks: a reinforcement learning solution," presented at the International Conference on Telecommunications, Feb. 2003.

[50] L. Dasilva and A. Mackenzie, "Cognitive Networks: Tutorial," presented at the CrownCom Orlando, Florida, US,, July, 2007.

[51] K.-L. A. Yau*, et al.*, "A context-aware and Intelligent Dynamic Channel Selection scheme for cognitive radio networks," in *4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications. CROWNCOM '09.*, Hannover, 2009, pp. 1-6.

[52] K. L. A. Yau*, et al.*, "Applications of Reinforcement Learning to Cognitive Radio Networks," in *Communications Workshops (ICC), 2010 IEEE International Conference on*, 2010, pp. 1-6.

[53] K. L. A. Yau*, et al.*, "Enhancing network performance in Distributed Cognitive Radio Networks using single-agent and multi-agent Reinforcement Learning," in *Local Computer Networks (LCN), 2010 IEEE 35th Conference on*, 2010, pp. 152-159.

[54] H. Li, "Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: A two by two case," in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, 2009, pp. 1893-1898.

[55] C. Wu*, et al.*, "Spectrum management of cognitive radio using multi-agent reinforcement learning," presented at the Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Industry track, Toronto, Canada, 2010.

[56] A. Galindo-Serrano and L. Giupponi, "Aggregated interference control for cognitive radio networks based on multi-agent learning," in *4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications. CROWNCOM '09.*, Hannover, 2009.

[57] M. J. Osborne, *An Introduction to Game Theory*: Oxford University Press, 2004.

[58] R. Gibbons, "An Introduction to Applicable Game Theory," *The Journal of Economic Perspectives,* vol. 11, pp. 127-149, 1997.

[59] M. Bublin*, et al.*, "Distributed spectrum sharing by reinforcement and game theory," presented at the 5th Karlsruhe workshop on software radio, Karlsruhe, Germany, March. 2008.

[60] C. K. Tan*, et al.*, "Game theoretic approach for channel assignment and power control with no-internal-regret learning in wireless ad hoc networks," *IEE Communications,* vol. 2, pp. 1159-1169, 2008.

[61] J. O. Neel*, et al.*, "The Role of Game Theory in the Analysis of Software Radio Networks," presented at the SDR Forum Technical Conference, 2002.

[62] J. O. Neel and J. Reed, "Game models for cognitive radio algorithm analysis," presented at the Software Define Radio Forum Technical Conference 2004.

[63] J. O. Neel*, et al.*, "Game theoretic analysis of a network of cognitive radios," presented at the The 45th Midwest Symposium on Circuits and Systems Aug. 2002.

[64] J. O. Neel, "Game Theory in the Analysis and Design of Cognitive Radio Networks," presented at the DySPAN, 2007.

[65] S. Mangold*, et al.*, "Equilibrium Analysis of Coexisiting IEEE 802.11e Wireless LANs," in *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications. PIMRC 2003.* , 2003, pp. 321-325.

[66] S. Mangold and K. challapali, "Coexistence of Wireless Networks in Unlicensed Frequency Bands," presented at the Wireless World Research Forum, Zurich, Switzerland, 2003.

[67] N. Nie and C. Comaniciu, "Adaptive channel allocation spectrum etiquette for cognitive radio networks," *Mobile Networks and Applications,* vol. 11, pp. 779-797, 16 December, 2006.

[68] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, 1 ed.: Addison-Wesley Professional, 1989.

[69] S. Chen and A. M. Wyglinski, "Efficient spectrum utilization via cross-layer optimization in distributed cognitive radio networks," *Computer Communications,* vol. 32, pp. 1931-1943, December 2009.

[70] O. Holland, "Cognitive Communications Projects around the World," York-Zhejiang Summer School on Cognitive Communications2010.

[71] *Beyond Next Generation Mobile Broadband Project Website*. Available: http://www.ict-bungee.eu/

[72] *End-to-End Reconfigurability (E2R,E2RII) Project Website*. Available: http://www.e2r.motlabs.com

[73] *End-to-End Efficiency (E3) Project Website*. Available: https://ict-e3.eu/

[74] *Quality of Service and Mobility Driven Cognitive Radio Systems (QoSMOS) Project Website*. Available: http://www.ict-qosmos.eu/

[75] *Cognitive Radio Systems for Efficient Sharing of TV White Spaces in European Context (COGEU) Project Website*. Available: http://www.ict-cogeu.eu/

[76] *Cognitive Radio and Cooperation Strategies for Power Saving in Multi-Standard Wireless Device (C2POWER) Project Website*. Available: http://www.ict-c2power.eu/

[77] *Spectrum and Energy Efficiency through Multi-band Cognitive Radio (SACRA) Project Website*. Available: http://www.ict-sacra.eu/

[78] *Quantitative Assessment of Secondary Spectrum Access (QUAZAR) Project Website*. Available: http://www.quasarspectrum.eu/

[79] *Opportunistic Networks and Cognitive Management Systems for Efficient Application Provision in the Future Internet (OneFIT) Project Website* Available: http://www.ict-onefit.eu/

[80] *Sensor Network for Dynamic and Cognitive Radio Access (SENDORA) Project Website*. Available: http://www.sendora.eu/

[81] *Physical Layer for Dynamic Spectrum Access and Cognitive Radio (PHYDYAS) Project Website*. Available: http://www.ict-phydyas.org/

[82] *Flexible and Spectrum-Aware Radio Access through Measurements and Modelling in Cognitive Radio Systems (FARAMIR) Project Website*. Available: http://www.ict-faramir.eu/

[83] *Advanced Coexistence Technologies for Radio Optimisation and Unlicensed Spectrum (ACROPOLIS) Project Website*. Available: http://www.ict-acropolis.eu/

[84] N. Drakos, "Introduction to Monte Carlo Methods," Computer Based Learning Unit, University of Leeds, Aug 1994.

[85] S. Saunders, *Antennas and propagation for wireless communication systems*: Wiley, 1999.

[86] A. Papadogiannis and A. G. Burr, "Multi-beam Assisted MIMO - A Novel Approach to Fixed Beamforming," presented at the Future Network and Mobile Summit, Warsaw, Poland, 2011.

[87] L. Kleinrock, *Queueing Systems Volume I: Theory*: John Wiley & Sons, 1975.

[88] T. Jiang*, et al.*, "Performance of Cognitive Radio Reinforcement Spectrum Sharing Using Different Weighting Factors," presented at the International Workshop on Cognitive Networks and Communications (COGCOM) in conjunction with CHINACOM'08, , Hangzhou, China, August, 2008.

[89] T. Jiang*, et al.*, "Two-stage reinforcement-learning-based cognitive radio with exploration control," *Communications, IET,* vol. 5, pp. 644-651, 2011.

[90] T. Jiang*, et al.*, "Two Stage Reinforcement Learning Based Cognitive Radio with Exploration Control," *accepted by IET Communications,* 2009.

[91] S. Kapetanakis and D. Kudenko, "Reinforcement learning of coordination in cooperative multi-agent systems," presented at the Eighteenth national conference on Artificial intelligence, Edmonton, Alberta, Canada, 2002.

[92] S. B. Thrun, "Efficient Exploration in Reinforcement Learning ", Technical Report: CS-92-102, School of Computer Science, Carnegie-Mellon University, USA, 1992.

[93] G. Chouinard, "FCC R&O 08-260 Proposed text for Antenna Height," CRC, January, 2009.

[94] T. Jiang*, et al.*, "Improvement of Pre-partitioning on Reinforcement Learning Based Spectrum Sharing," presented at the IET International Communication Conference on Wireless Mobile& Computing (Best Paper Award) Shanghai, China, 2009.

[95] T. Jiang*, et al.*, "Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing," *Communications, IET,* vol. 5, pp. 1309-1317, 2011.

[96] G. E. Moore, "Cramming More Components Onto Integrated Circuits," *Proceedings of the IEEE,* vol. 86, pp. 82-85, 1998.

[97] G. Fettweis, "The wireless roadmap," presented at the Panel session of IEEE VTC 2007-fall.

[98] W. Van Heddeghem*, et al.*, "Energy in ICT - Trends and research directions," in *Advanced Networks and Telecommunication Systems (ANTS), 2009 IEEE 3rd International Symposium on*, 2009, pp. 1-3.

[99] M. Pickavet*, et al.*, "Worldwide energy needs for ICT: The rise of power-aware networking," in *Advanced Networks and Telecommunication Systems, 2008. ANTS '08. 2nd International Symposium on*, 2008, pp. 1-3.

[100] J. Sydor*, et al.*, "Cognitive, Radio-Aware, Low-Cost (CORAL) Research Platform," presented at the IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN), Singapore, 2010.

[101] W. Ye*, et al.*, "Medium Access Control With Coordinated Adaptive Sleeping for Wireless Sensor Networks," *IEEE TRANSACTIONS ON NETWORKING,* vol. 12, pp. 493-506, 2004.

[102] T. Jiang*, et al.*, "Cognitive Radio Spectrum Sharing Schemes with Reduced Spectrum Sensing Requirements," presented at the The IET Seminar on Cognitive Radio and Software Defined Radios: Technologies and Techniques London, September, 2008.

[103] D. Grace*, et al.*, "Using Cognitive Radio to Deliver 'Green' Communications," in *4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications*, Hannover, 2009.

[104] "Description of Work," BuNGee Project Document, 2009.

[105] T. Jiang*, et al.* BuNGee D4.1.1 Interim Simulation. 2011. Available: http://www.ict-bungee.eu/

[106] P. Kyosti*, et al.*, "IST-4-027756 WINNER II D1.1.2," Available: https://www.ist-winner.org/WINNER2-Deliverables/D1.1.2v1.1.pdf 2007.

[107] Y. Leiba*, et al.* BuNGee D3.1 Baseline RRM & JointAccess/Self-Backhaul Protocols. 2011.

[108] L. Giupponi*, et al.*, "Docitive networks: an emerging paradigm for dynamic spectrum management [Dynamic Spectrum Management]," *Wireless Communications, IEEE,* vol. 17, pp. 47-54, 2010.

[109] X. Chen*, et al.*, "Inter-Cluster Connection in Cognitive Wireless Mesh Networks Based on Intelligent Network Coding," *EURASIP Journal on Applied Signal Processing special issue,* 2009.

[110] X. Chen*, et al.*, "Inter-Cluster Connection in Cognitive Wireless Mesh Networks Based on Intelligent Network Coding," presented at the The 20th IEEE International Symposium On Personal, Indoor and Mobile Radio Communications (PIMRC), 2009.

[111] H. Li*, et al.*, "Collision reduction in cognitive radio using multichannel 1-persistent CSMA combined with reinforcement learning," in *Cognitive Radio Oriented Wireless Networks & Communications (CROWNCOM), 2010 Proceedings of the Fifth International Conference on*, 2010, pp. 1-5.

[112] H. Li*, et al.*, "Cognitive radio multiple access control for unlicensed and open spectrum with reduced spectrum sensing requirements," in *Wireless Communication Systems (ISWCS), 2010 7th International Symposium on*, 2010, pp. 1046-1050.

[113] M. Yang and D. Grace, "Cognitive Radio with Reinforcement Learning Applied to Multicast Downlink Transmission with Power Adjustment," *Wireless Personal Communications,* vol. 57, pp. 73-87, 2011.

[114] M. Yang and D. Grace, "Cognitive radio with reinforcement learning applied to heterogeneous multicast terrestrial communication systems," presented at the 4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM) 2009.