Accelerating Reinforcement Learning for Dynamic Spectrum Access in Cognitive Wireless Networks

Nils Morozs

Ph.D.

University of York

Electronics

September 2015

Abstract

This thesis studies the applications of distributed reinforcement learning (RL) based machine intelligence to dynamic spectrum access (DSA) in future cognitive wireless networks. In particular, this work focuses on ways of accelerating distributed RL based DSA algorithms in order to improve their adaptability in terms of the initial and steady-state performance, and the quality of service (QoS) convergence behaviour. The performance of the DSA schemes proposed in this thesis is empirically evaluated using large-scale system-level simulations of a temporary event scenario which involves a cognitive small cell network installed in a densely populated stadium, and in some cases a base station on an aerial platform and a number of local primary LTE base stations, all sharing the same spectrum. Some of the algorithms are also theoretically evaluated using a Bayesian network based probabilistic convergence analysis method proposed by the author.

The thesis presents novel distributed RL based DSA algorithms that employ a Win-or-Learn-Fast (WoLF) variable learning rate and an adaptation of the heuristically accelerated RL (HARL) framework in order to significantly improve the initial performance and the convergence speed of classical RL algorithms and, thus, increase their adaptability in challenging DSA environments. Furthermore, a distributed case-based RL approach to DSA is proposed. It combines RL and case-based reasoning to increase the robustness and adaptability of distributed RL based DSA schemes in dynamically changing wireless environments.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deepest gratitude to my supervisors Tim Clarke and David Grace for providing me with their insightful guidance, yet at the same time giving me so much freedom to explore different research directions.

I would also like to thank my examiners Luiz DaSilva and Paul Mitchell for the motivating in-depth discussions during the viva and the valuable feedback on my work.

Thanks to my colleagues at the Department of Electronics for creating a friendly and intellectually stimulating research environment. In particular, thanks to Andy, Stuart, Tautvydas and Alan for countless discussions over lunch and at the pub that made this work easy and enjoyable.

My sincere thanks go to my family - Ieva, Polina, Vadim, Inessa and Kesha (the dog) - without whom none of this work could have possibly happened.

Finally, I would like to acknowledge that this work has been funded by the ABSOLUTE Project (FP7-ICT-2011-8-318632), which received funding from the 7th Framework Programme of the European Commission.

Declaration

All work presented in this thesis is original to the best knowledge of the author. References to other researchers have been given as appropriate. This work has not previously been presented for an award at this or any other institution. The research presented in this thesis features in a number of the author's publications listed below.

Journal Articles

N. Morozs, T. Clarke, and D. Grace, "Distributed heuristically accelerated Q-learning for robust cognitive spectrum management in LTE cellular systems," in *IEEE Transactions on Mobile Computing*, 2015.

N. Morozs, T. Clarke, and D. Grace, "Distributed heuristically accelerated reinforcement learning for dynamic secondary spectrum sharing," in *IEEE Access*, 2015.

N. Morozs, T. Clarke, and D. Grace, "Cognitive spectrum management in dynamic cellular environments: a case-based Q-learning approach," submitted to *Transactions on Emerging Telecommunications Technologies*, 2015.

Conference Papers

N. Morozs, T. Clarke, and D. Grace, "A novel adaptive call admission control scheme for distributed reinforcement learning based dynamic spectrum access in cellular networks," in *International Symposium on Wireless Communication Systems (ISWCS)*, 2013.

N. Morozs, D. Grace, and T. Clarke, "Case-based reinforcement learning for cognitive spectrum assignment in cellular networks with dynamic topologies," in *Military Communications and Information Systems Conference (MCC)*, 2013.

N. Morozs, T. Clarke, D. Grace, and Qiyang Zhao, "Distributed Q-learning based dynamic spectrum management in cognitive cellular systems: choosing the right learning

rate," in *IEEE International Symposium on Computers and Communications (ISCC)*, 2014.

N. Morozs, D. Grace, and T. Clarke, "Distributed Q-learning based dynamic spectrum access in high capacity density cognitive cellular systems using secondary LTE spectrum sharing," in *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2014.

N. Morozs, T. Clarke, and D. Grace, "Using Bayesian networks for convergence analysis of intelligent dynamic spectrum access algorithms," in *IEEE International Conference on Communications Workshops (ICC Workshops)*, 2015.

N. Morozs, T. Clarke, and D. Grace, "Case-based cognitive cellular systems for temporary events," extended poster abstract in *European Conference on Networks and Communications (EuCNC)*, 2015.

N. Morozs, T. Clarke, and D. Grace, "Intelligent dynamic spectrum access in cellular systems with asymmetric topologies and non-uniform traffic loads," in *IEEE Vehicular Technology Conference (VTC-Fall)*, 2015.

N. Morozs, T. Clarke, and D. Grace, "Intelligent secondary LTE spectrum sharing in high capacity cognitive cellular systems," in *IEEE Vehicular Technology Conference (VTC-Fall)*, 2015.

Chapter 1. Introduction

Contents

1.1 Overview

One of the fundamental tasks of a wireless network is spectrum management, concerned with dividing the available spectrum into a set of resource blocks or channels and assigning them to voice calls and data transmissions in a way that provides a good quality of service (QoS) to the users. Spectrum sharing and flexible dynamic spectrum access (DSA) techniques play a key role in utilising the given spectrum efficiently in the face of an ever increasing demand for mobile data capacity [16][84].

Some of the early work on DSA, then commonly referred to as dynamic channel assignment (DCA), dates back to the early 1970s. For example, Cox and Reudink [24] and Anderson [5] demonstrate through simulation experiments that their proposed DSA algorithms, which give all base stations access to the whole spectrum pool of a cellular network, significantly increase the capacity of mobile cellular systems compared to the classical, fixed channel allocation approach. More recently the idea of DSA and the need for efficient spectrum utilisation has given rise to novel wireless communication systems such as cognitive radio (CR) networks [86] and cognitive cellular systems [34]. Such networks employ intelligent opportunistic DSA techniques that allow them to access licensed spectrum underutilized by the incumbent users.

An emerging state-of-the-art technique for intelligent DSA is reinforcement learning (RL); a machine learning technique aimed at building up solutions to decision problems only through trial-and-error [87]. It has been successfully used for a wide range of DSA problems and scenarios such as CR networks [43][89], small cell networks

[7][26], multi-hop backhaul networks [101], and cognitive wireless mesh networks [18][19]. The chief advantage of the RL approach to DSA is its capability to facilitate full self-organisation in a wireless network. It eliminates the need for the potentially challenging and time-consuming spectrum planning process carried out by human experts, whilst enabling the wireless network to learn flexible and highly efficient spectrum management policies [8]. However, an inherent disadvantage of RL algorithms is their need for the exploration process, which normally involves a large number of trial-and-error iterations, during which the system exhibits poor performance due to its lack of initial knowledge of the environment [87].

The purpose of the work described in this thesis is to increase the adaptability of distributed RL based DSA algorithms by proposing a number of techniques that significantly improve their temporal characteristics such as initial performance, convergence speed and steady-state performance. The ultimate aim of these contributions is to enable reliable opportunistic RL based DSA methods that are a feasible option for implementation in real-world commercial wireless networks.

1.2 Hypothesis

The following hypothesis is used to guide the work presented in this thesis:

"Appropriate use of available heuristic information can accelerate distributed reinforcement learning algorithms to enable highly adaptable dynamic spectrum access in cognitive wireless networks."

The adaptability of cognitive wireless networks is assessed by inspecting the temporal QoS performance of the proposed RL algorithms in a range of large-scale DSA and spectrum sharing simulation scenarios. Specifically, it is essential for the cognitive wireless devices to exhibit sufficiently good performance at the initial stages of learning and to show a high convergence speed in order to be able to adapt to challenging and potentially dynamic radio environments. These aspects of the distributed RL based DSA algorithm performance are the focus of the simulation experiments discussed in this thesis.

1.3 Thesis Outline

The rest of the thesis is organised as follows:

- Chapter 2 first reviews the existing research literature on conventional DSA techniques based on heuristic spectrum awareness information, e.g. geo-location databases, distributed interference measurements and license repositories. It then discusses a number of distributed machine intelligence methods based on RL found in the general artificial intelligence literature. Finally, the last section of Chapter 2 reviews the state-of-the-art in the applications of RL towards intelligent DSA in wireless networks.

- Chapter 3 explains the experimental methodology used for empirical evaluation of the DSA algorithms proposed in this thesis. It presents the details of the cognitive wireless network simulation model, the metrics used to assess the network performance, and two conventional DSA schemes used for baseline comparison.

- Chapter 4 introduces the distributed Q-learning based DSA algorithm used as the basis for the DSA schemes proposed in the further chapters of this thesis. It also introduces the concept of the Win-or-Learn-Fast (WoLF) variable learning rate principle and empirically demonstrates the network performance improvements achieved by applying it to every learning agent in the environment. The last section of Chapter 4 presents the simulation results of using the distributed Q-learning based DSA algorithm in the context of secondary spectrum sharing.

- Chapter 5 presents a novel empirically validated probabilistic model for convergence analysis of distributed RL based DSA algorithms. It is based on a Bayesian network that describes a simple generalised inter-cell interference problem with two base stations and two user equipments.

- Chapter 6 proposes a DSA algorithm designed for Long Term Evolution (LTE) cellular systems - distributed ICIC accelerated Q-learning (DIAQ). It combines distributed RL and standardized inter-cell interference coordination (ICIC) signalling in the LTE downlink, using the framework of heuristically accelerated RL (HARL). Its purpose is to improve the initial performance and the conver-

gence speed of distributed RL based DSA algorithms and, thus, to increase their robustness and adaptability in challenging wireless environments.

- Chapter 7 extends the HARL framework proposed in Chapter 6 and presents a novel mechanism for dynamic secondary spectrum sharing based on it. It utilises a radio environment map (REM) as external information for guiding the learning process of cognitive wireless networks. Furthermore, the novel principle and the general structure of heuristic functions proposed in the context of HARL are applicable to a wide range of self-organisation problems beyond the wireless communications domain.

- Chapter 8 proposes a case-based RL (CBRL) approach that stabilises the performance of intelligent DSA algorithms in dynamic wireless environments. The proposed algorithm is the combination of classical RL and a novel implementation of case-based reasoning (CBR) which aims to facilitate a number of learning processes running in parallel. It is assessed using a number of simulations of a cognitive wireless network with a dynamically changing topology.

- Chapter 9 presents the conclusions of this thesis, summarises its original contributions, and discusses a number of recommendations for further work.

Chapter 2. Literature Review

Contents

2.1 Introduction

Spectrum management is one of the fundamental tasks performed by wireless networks. It is concerned with dividing the available spectrum into a set of resource blocks or channels and assigning them to voice calls and data transmissions in a way that provides a good quality of service (QoS) to the users. Flexible dynamic spectrum access (DSA) and spectrum sharing techniques are often considered the key spectrum management paradigm for utilising the wireless spectrum efficiently in order to accommodate the ever increasing demand for mobile data capacity [16][84]. This motivated

the design of novel wireless communication systems such as CR networks [86] and cognitive cellular systems [34]. Such networks employ opportunistic DSA techniques that allow them to access licensed spectrum underutilized by the incumbent users.

2.2 Dynamic Spectrum Access and Spectrum Sharing

This section first introduces a number of well-established DSA methods designed for cognitive wireless networks that do not involve machine intelligence. It then presents a simplified adaptation of the Mitola's cognition cycle of an intelligent wireless device [61] and discusses how these conventional DSA techniques differ from the intelligent methods that involve all aspects of the Mitola's cognitive cycle.

2.2.1 Spectrum Database Approach

The classical application of DSA in cognitive wireless networks is the use spectrum databases. In particular, the most widely known type of DSA networks that rely on spectrum databases are TV white space (TVWS) based CR networks. Such networks aim to reuse the spectrum allocated to TV broadcasters for other wireless communications, whilst eliminating harmful interference to the incumbent TV receivers, e.g. [30][35].

The coexistence between the primary TV broadcasting networks and the secondary wireless networks is facilitated by geo-location databases that describe in detail the unused TV spectrum bands at given geographical locations, i.e. the white spaces. Such a setup is depicted in Figure 2.1, where the maintenance of the TVWS database is controlled by the national telecoms regulator such as Ofcom in the United Kingdom

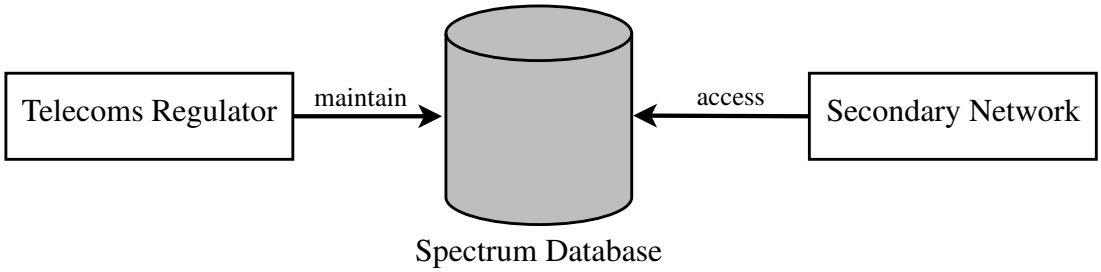


Figure 2.1: Secondary dynamic spectrum access facilitated by a geo-location database

or the Federal Communications Commission (FCC) in the United States [27]. The secondary cognitive networks are then allowed to gain access to the TV spectrum without the need for a license, provided that they do not interfere with the primary licensed users. The key source of information that enables them to satisfy this requirement is the TVWS database.

The most notable example of the practical use of TVWS technologies aided by geolocation databases is the IEEE 802.22 standard for license-exempt wireless regional area networks (WRAN), the first ever world-wide standard for CR and TVWS communications [22]. For example Ishizu et al. [38] have conducted a field experiment where an IEEE 802.22 WRAN system was used to provide broadband communications (4.5 Mbps uplink and 5.2 Mbps downlink) to a remote rural area 12.7 km away. Such wide coverage broadband communications without the need for dedicated spectrum is the main application of IEEE 802.22 WRAN networks since the TVWS occur at appropriately low frequencies to support long distance transmissions, e.g. [52][85].

Such spectrum database DSA methods provide a robust and highly controllable solution for increasing the spectrum utilisation efficiency by allowing secondary cognitive devices access to spectrum bands otherwise unused by the incumbent users. However, there is limited scope for flexibility and adaptability of the secondary wireless devices employing such DSA methods due to the restrictive and relatively static regulatory control of the spectrum databases.

2.2.2 Opportunistic Spectrum Sensing Approach

A more flexible and dynamic approach to DSA which is highly popular in the CR research domain is the use of spectrum sensing for dynamic identification of unused spectrum eligible for secondary access [86][99]. Here, the cognitive wireless devices continuously measure the interference levels on the channels potentially available for secondary reuse and transmit their packets as soon as they detect the unused spectrum due to the interference on a particular channel dropping below a pre-defined threshold, e.g. -107 dBmW for TVWS [23]. A simplified scenario that describes this opportunistic approach to DSA is depicted in Figure 2.2. When the CR device has a new packet

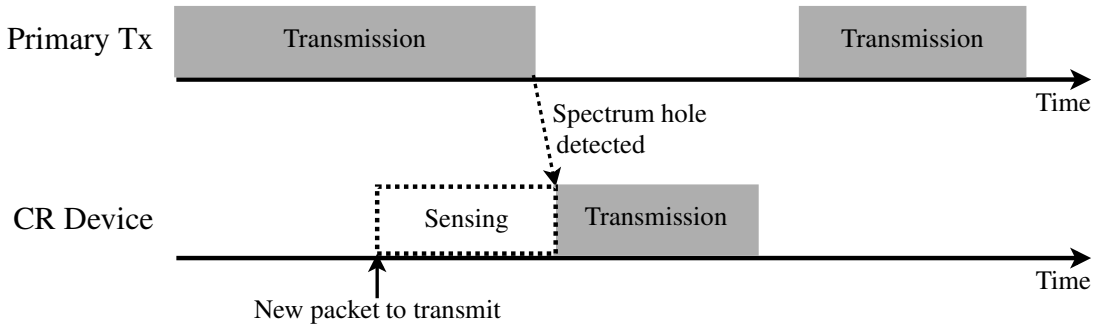


Figure 2.2: Opportunistic spectrum access using spectrum sensing in cognitive radio (CR) networks

to transmit, it starts sensing the interference levels received from the primary system transmitters and in some cases from other CR transmitters. As soon as it detects the lack of primary or secondary transmissions on a particular channel it transmits its own packet using the detected spectrum hole. If another primary user transmission starts before the secondary CR transmission finishes, the latter is interrupted and resumed at the next opportunity.

This "listen-before-talk" principle had already been studied by the wireless communications research community before the concepts of CR and primary-secondary user spectrum sharing were first introduced [47]. A classical example of such interference measurement based DSA algorithms is the scheme proposed by Akerberg and Brouwer [4]. There, the interference level is measured by the base station (BS) on each channel to determine whether it is available for assignment. The authors investigate the effects of varying the interference threshold, which sets the maximum interference level at which a channel can be assigned to a user, for the least interfered channel scheme, i.e. where a channel with the lowest level of interference is always chosen if available.

Since then such interference measurement based DSA methods have become a highly popular approach to spectrum management among CR researchers, as well as among those investigating more traditional wireless networks without the primary-secondary spectrum sharing considerations. For example, Cheng and Chuang [20] show that a simple aggressive least interference DSA algorithm, that always opportunistically assigns a channel with the least aggregate interference without any admission control, achieves the best performance in a classical hexagonal cell network compared to other known interference sensing based approaches. Ramachandran et al. [69]

propose a dynamic interference-aware algorithm that minimises interference between routers in wireless mesh networks, as well as between the mesh networks and the other co-located wireless networks. They also empirically demonstrate the performance improvement gained by their algorithm on a physical IEEE 802.11 testbed. A notable example of research on opportunistic spectrum access in CR networks is the paper by Huang et al. [37], where the authors give a closed form analysis of the primary and secondary CR user performance that provides theoretical insight into the capacity of opportunistic spectrum access under primary user protection constraints. Another well-known example of theoretical work on opportunistic spectrum access in CR networks is a number of cognitive medium access control (MAC) protocols proposed by Zhao et al. [100] based on the partially observable Markov decision process (POMDP) model of the DSA problem they develop. These decentralised protocols are designed to optimise the performance of the secondary users while limiting the interference perceived by the primary users.

Opportunistic spectrum sensing described in this subsection is a significantly more dynamic and adaptable approach to DSA which has the potential to achieve higher spectrum utilisation efficiency, compared with the spectrum database approach. However, there are also some drawbacks associated with it [99]. For example, the hardware required to facilitate precise spectrum sensing is likely to cause a significant increase in the cost and energy consumption of the CR devices. They are also susceptible to the hidden terminal effect, where a CR node is unable to detect an incumbent transmission due to the effects of shadowing and multipath fading, which in turn results in harmful interference for the primary user.

2.2.3 Regulatory Approach

A more recent problem investigated by researchers, mobile network operators (MNOs) and regulators is LTE/LTE-Advanced spectrum sharing facilitated by an emerging framework known as licensed shared access (LSA) [58]. Here, licenses for the use of LTE spectrum are issued upon agreement for a specific geographical area and required time duration. A successful live field trial of implementing LSA-based spectrum sharing has already taken place in Finland [66]. Here, an LSA controller was used to

autonomously configure an existing LTE network based on the incumbent spectrum usage data stored in the LSA Repository, i.e. a process similar to the TVWS database approach described in Subsection 2.2.1 took place but with a higher degree of regulatory control.

This is a static regulatory approach to spectrum sharing that does not involve any intelligence or cognition in wireless devices and does not require any opportunistic spectrum access techniques. The advantage of this approach is its reliability and the same QoS guarantees as those normally provided to the users of conventional LTE networks with their own exclusive spectrum, but with no need for a permanently owned LTE spectrum band. However, the focus of this thesis is to investigate more flexible opportunistic techniques for DSA that have a greater potential in terms of the spectrum utilisation efficiency, since they are not limited by licenses that restrict the number of different spectrum users sharing the same geographical area.

2.2.4 Mitola's Cognition Cycle of a Wireless Device

The three well-established approaches to DSA described in this section so far work based solely on spectrum awareness information either measured by the wireless device itself or obtained from a spectrum database or a license repository. They do not involve all aspects of the cognition cycle of an intelligent wireless device originally introduced by Mitola in his PhD thesis [61] where the term "cognitive radio" was coined. This cognition cycle is shown in Figure 2.3. It identifies six fundamental functions performed by a CR device, specifically by its cognitive engine - observe, orient, learn, plan, decide and act. The CR device *decides* which action it needs to apply to its wireless *environment* and *acts* using the chosen action. It then *observes* the consequences of taking that action, *orients* itself, i.e. processes the observation, and decides upon its next action. In order to decide which action the CR device should take, e.g. which channel it should select for secondary access, it must have a capability to *plan* its own strategy. Utilising different spectrum awareness information sources, e.g. interference measurements, a geo-location database or an LSA license repository, can be viewed as the "planning" function of a CR device. In all three different DSA approaches described in this section the spectrum awareness information is used to

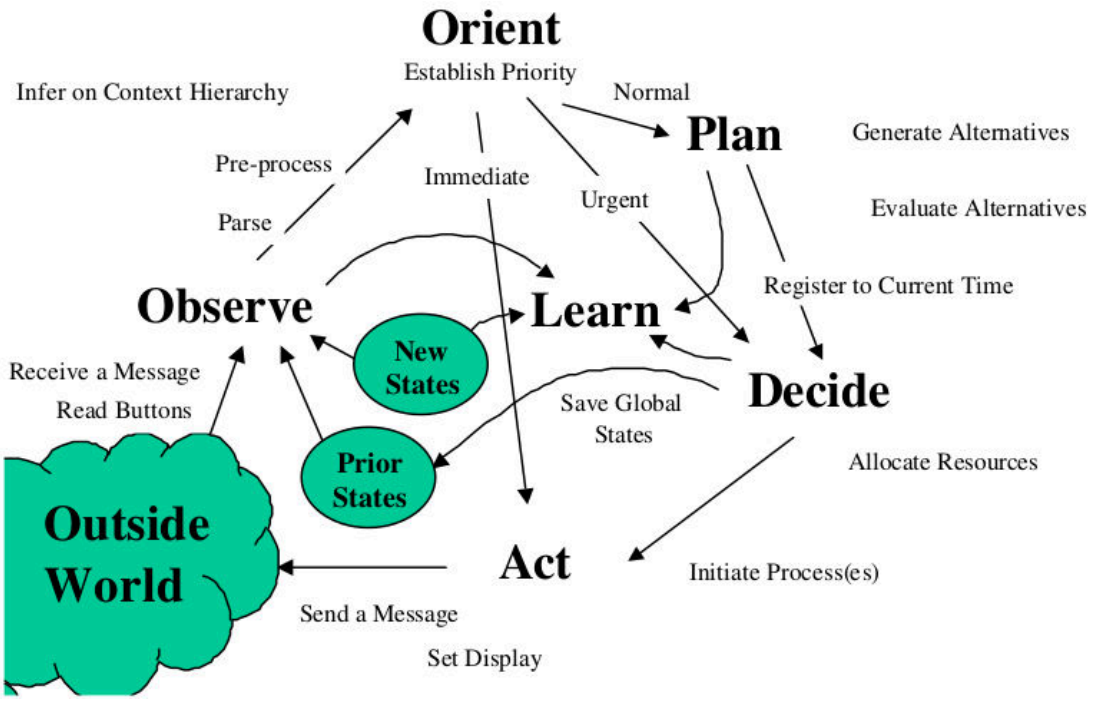


Figure 2.3: Mitola's cognitive radio cycle [61]

identify white spaces or spectrum holes eligible for access by the given wireless device, i.e. this functionality enables the wireless device to *plan* its spectrum assignment strategy. A key function of a CR device as defined by Mitola that is missing from the DSA mechanisms described in this section is the capability to *learn* from its own experience, i.e. to enable the wireless device to gradually build up an internal knowledge base and improve its performance over time.

A close inspection of Mitola's CR cycle shown in Figure 2.3 reveals a fundamental mistake in this diagram - the "learning" function only has incoming arrows and does not output the learnt information to any other function. A way of fixing this issue proposed in this subsection is reversing the arrow between "learning" and "planning", since it makes sense to base one's plans on what has been learnt. It is also consistent with the way humans operate in simple terms - they learn, gain experience and then plan their future actions based on that knowledge.

Secondly, this diagram can be further simplified by removing or modifying several links that are optional for describing a machine's cognition cycle. For example, the arrow from "orient" to "act" is not required, since the machine always needs to *decide* in some way which action to take. Even if it is an immediate random action with no

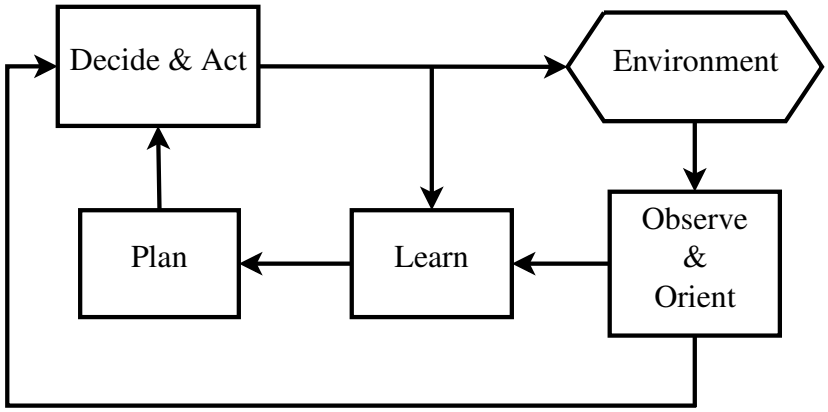


Figure 2.4: Simplified Mitola's cognition cycle of an intelligent wireless device

delay, it can still be classified as a *decision*. The "observe-learn" link can be replaced by the "orient-learn" link, since it is the same information feeding into the learning process, but appropriately processed. The "orient-plan-decide" link can be assumed to flow either through the "learn" process or bypass "plan" and straight to "decide", since the learning updates are the only thing that can possibly change the existing *plan*. Therefore the arrow between "orient" and "plan" can be removed. Finally, the "decide-learn" link is equivalent to the "act-learn" link, since it carries the same information about the chosen action that is fed back to the learning process.

These simplifications result in a considerably more readable and intuitive dual-loop cognition cycle proposed in Figure 2.4. There is an outer *decision loop* which simply looks at the processed outputs of the environment, or its *state*, and chooses an action to be taken in this state. The intelligence is provided by the inner learning and planning loop, where the machine, i.e. an intelligent wireless device, is observing the outputs of the environment caused by its actions and builds up a knowledge base which describes its experience in a way which could be used to derive a plan or a *policy* autonomously. The rest of this chapter reviews the existing methods of achieving such machine intelligence both in general and specifically in the context of DSA.

2.3 Reinforcement Learning

An emerging state-of-the-art technique for intelligent DSA is reinforcement learning (RL); a machine learning technique aimed at building up solutions to decision prob-

lems only through trial-and-error. The fundamental idea behind RL is that learning processes of all organisms are based on interaction with the environment. The aim of RL methods is to imitate this behaviour in artificial systems [87]. The key feature of RL that can potentially enable full self-organisation and high adaptability in cognitive wireless networks is its lack of need for any a priori knowledge of the environment model [93]. This feature is the main reason for the widespread use of RL for DSA in wireless networks, since building an accurate analytical model of an arbitrary wireless communications environment is often unfeasible or even impossible.

The goal of any RL algorithm is to create a function which maps perceived situations or *states* of the environment to *actions* which need to be taken in them. This is known as the *policy function*. It is developed through system experience of trying different actions in each state and noting the result. This trial-and-error approach does not make any assumptions about the environment model, e.g. such as its structure or whether it exhibits the Markov property. Each state or state-action pair receives a numerical *reward* which indicates its desirability. Calculating the reward for each state or state-action pair is handled by the *reward function*. Another important RL term is the *value function*, also referred to as *value table*, *Q-function* or *Q-table*. It maps each state or state-action pair to the total discounted sum of rewards expected to be accumulated over the future, starting from that state. This is equivalent to the reward in the long run, as opposed to an immediate return.

One of the biggest challenges of RL is estimating the value function. The reward function is often relatively easy to design, since it is only concerned with immediate benefits of taking a certain action in a certain state. However, estimating the value function requires predicting the future of the system to some extent, which is a significantly harder task without the knowledge of a system model.

Another challenge of RL is a trade-off between *exploration* and *exploitation*. In each state an RL algorithm always faces two options:

- Choose a previously known action which guarantees the best reward among all other known actions, referred to as the *greedy action*. In this case the system is *exploiting* its current knowledge.

- Choose a previously unknown action which is likely to have a lower reward than the greedy action. However, there is also a low probability of it being better and becoming the new greedy action. In this case the system is *exploring* new possibilities.

2.3.1 Model-Based Reinforcement Learning

One of the fundamental approaches to RL is model-based RL, where a learning agent attempts to build up a model of the environment in a form which would allow it to compute a suitable policy [75][96].

Figure 2.5 shows a flow diagram of the processes involved in model-based RL. There is an outer output-state-action loop, where outputs of the environment are observed and processed to yield the environment state information, and then the best action is chosen for the current state based on the policy of the learning agent. There is also an inner learning loop, whose role is to learn a good policy to be used by the learning agent. It achieves this goal by observing the actions taken by the learning agent and their outcomes and estimating a model of the environment in the form of a transition probability matrix (TPM) and a transition reward matrix (TRM). The role of the TPM is to indicate the probability of being in a certain state, executing a certain action and making a transition to another state. The TRM states the immediate reward received after a certain state-action-state transition. A policy is then computed from the estimated TPM and TRM using a dynamic programming (DP) algorithm and used for

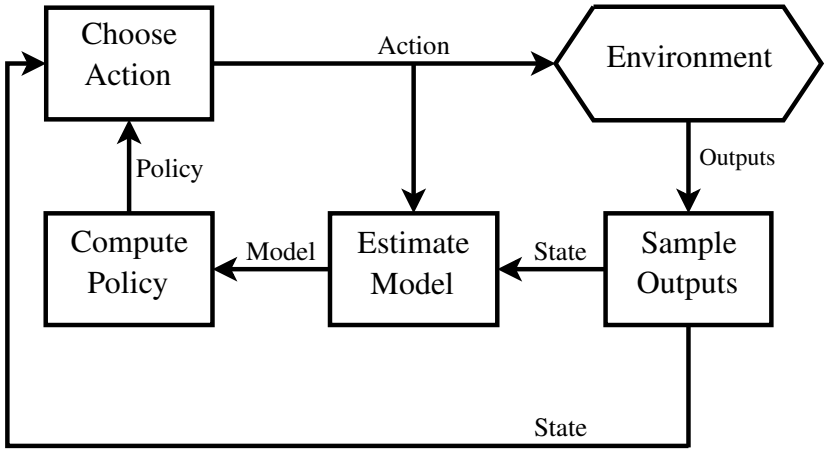


Figure 2.5: Flow diagram of model-based reinforcement learning

choosing an action in the current environment state. It can be seen from the diagram in Figure 2.5 and the description above that model-based RL follows the dual-loop cognition cycle framework described in Subsection 2.2.4, which supports the fact that this is an intelligent learning technique.

Model Estimation

One of the methods of building up TPMs and TRMs is by straightforward counting [75]. It is also referred to as the maximum likelihood model estimation by Wiering [96]. There, for each state-action pair a learning agent counts how many times it made a transition to any other state. It can then normalise the counter values to become a discrete probability distribution. For example, [2, 4, 2] would become [0.25, 0.5, 0.25] which indicates the probabilities of a system to enter one of the three states after executing a certain action in a certain state.

This is a good method for static environments, i.e. where the environment dynamics do not change over time. This approach needs to be modified to be adaptable in dynamically changing environments which are highly relevant in this thesis, where the investigated environment is a wireless network. For example, Wiering [96] proposes a scheme which aims to reset the TPM and TRM counters according to specially derived formulae and then use prioritised sweeping to update the policy. Prioritised sweeping is a method of filtering and analysing only a small portion of the state-action space, to eliminate the need for analysing the full state-action space, most of which might be irrelevant at a given moment in time [63].

Dynamic Programming

Given the TPMs and TRMs built up by the model estimation algorithm, it is then the task of a DP algorithm to derive the best policy from them. There is a large number of DP algorithms for solving the Markov decision processes (MDPs) expressed by TPMs and TRMs. They all have the same goal - solve the recursive Bellman optimality equation, given below [87]:

$$Q^*(s,a) = \sum_{s'} P(s,a,s')[R(s,a,s') + \gamma max_{a'} Q^*(s',a')] \tag{2.1}$$

where $Q^*(s,a)$ is the long-term cumulative reward of taking action a in state s (also referred to as the Q-value), $P(s,a,s')$ is the probability of going to state s' after taking action a in state s (element of TPM), $R(s,a,s')$ is the expected immediate reward when an agent takes action a in state s and goes to state s' (element of TRM), and $\gamma \in [0,1]$ is the discount factor which weights the importance of future long-term rewards with respect to the immediate reward.

It is then straightforward to derive a *greedy* policy from $Q^*(s,a)$, which maximises the Q-value for every state of the environment. The choice of actions would follow the rule given in (2.2), which states that an action with the highest Q-value must be chosen for every state's policy.

$$\pi(s) = argmax_a Q^*(s,a) \tag{2.2}$$

Using the model-based RL approach to DSA could provide valuable insight into the dynamics of cognitive wireless network environments by explicitly building up the knowledge about the transition probabilities between various states of the environment, e.g. discrete probability distributions that describe how particular spectrum assignment decisions in certain situations are likely to affect the QoS of the network. However, for the specific purpose of on-line learning and decision making in an arbitrary wireless environment model-free RL methods described in the next subsection are more flexible and significantly more popular in the research literature.

2.3.2 Model-Free Reinforcement Learning

An alternative to model-based RL methods is model-free RL [45], where the Q-function, also known as the Q-table in discrete state-action space, $Q^*(s,a)$ is estimated directly from received rewards, i.e. without the intermediate step of constructing TPMs and TRMs. This type of RL is more popular, since it is significantly more computationally efficient and does not require the environment to fit the TRM and TPM model

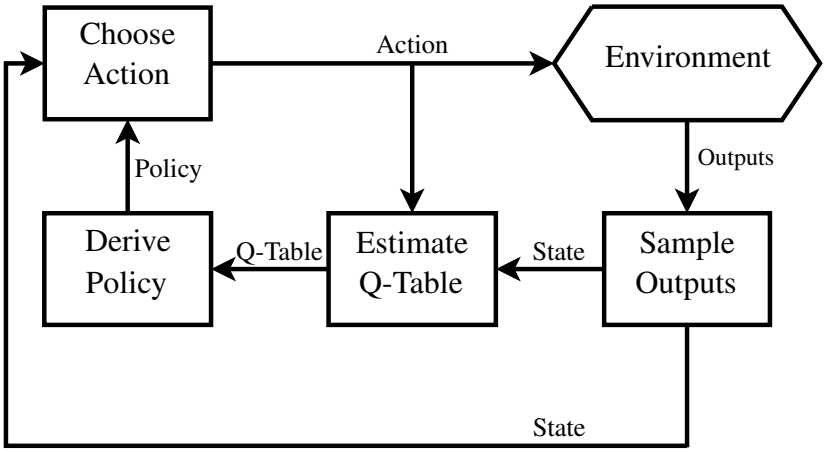


Figure 2.6: Flow diagram of model-free reinforcement learning

template.

Figure 2.6 shows a flow diagram of the processes involved in model-free RL. There is the same outer output-state-action loop as that shown in Figure 2.5, where outputs of the environment are observed and processed to yield the environment state information, and then the best action is chosen for the current state based on the policy of the learning agent. However, the components of the inner learning loop are different from those of model-based RL. Instead of estimating an environment model in the form of TPMs and TRMs, the Q-values of the state-action pairs are directly estimated and stored in the Q-table. The policy derivation step is then much simpler and does not require a full DP algorithm, but is simply the last step of a DP algorithm defined in Equation (2.2). Since model-free RL has an identical dual-loop cognition structure to model-based RL and the modified Mitola's cognition cycle defined in Subsection 2.2.4, it can equivalently be viewed as an intelligent learning technique.

Q-Learning

The most popular RL algorithm is Q-learning introduced by Watkins [94]. It is an off-policy method, i.e. the learning of an optimum policy does not depend on the policy followed by a learning agent. It is updating its policy based on the best possible future scenario, rather than what actually happens after an action is taken. Therefore this approach is not experimentation-sensitive, i.e. the learning is not affected by the amount of exploration performed by an agent.

One of its advantages over other RL algorithms is that it has been mathematically proven, e.g. [39][90], that it is guaranteed to converge on an optimal policy for an MDP in a theoretical case where each state is visited and each action is taken an infinite number of times.

The formula for updating a Q-table entry is given in the equation below [87]:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma max_{a'} Q(s',a') - Q(s,a)) \tag{2.3}$$

where:

- s is the current state of the system,
- a is the action taken in the current state s,
- s' is the next state of the system,
- a' is the action that can be taken in the next state s',
- $Q(s,a)$ is the Q-value of the current state-action pair.
- $\alpha \in [0,1]$ is the learning rate.
- $\gamma \in [0,1]$ is the discount factor.
- $max_{a'} Q(s',a')$ is the maximum Q-value out of all actions in the next state s'.

The key steps of the Watkins' Q-learning algorithm are summarised in Algorithm 1.

Algorithm 1 Watkins' Q-learning algorithm [87]

```
Initialise Q-table arbitrarily
while the learning episode has not finished do
    Detect present state
    while present state is not terminal do
        Choose current action according to action selection policy
        Take this action, observe next state and reward
        Update Q-table entry for current state-action pair using Equation (2.3)
        Store next state as the present state
    end while
end while
```

The simplicity and convergence properties of Q-learning are the key reasons why it is the most widely used RL algorithm and why most multi-agent RL algorithms are

derived from it [14]. It is also the most widely used RL algorithm in the DSA literature reviewed in Section 2.4.

SARSA

The on-policy alternative to Q-learning is the SARSA algorithm [74]. The difference between Q-learning and SARSA is best described by the difference between their update formulae given in Equations (2.3) and (2.4) respectively. Instead of using the Q-value of the best action in the next state - $max_{a'}Q(s', a')$, SARSA uses the action actually chosen in the next state - $Q(s', a')$, giving rise to its name {State, Action, Reward, State, Action}. Therefore, its performance is dependent on the exploration strategy chosen by the learning agent, i.e. it is experimentation-sensitive.

$$Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma Q(s', a') - Q(s, a)) \tag{2.4}$$

Although the fact that SARSA is experimentation-sensitive is often considered as a drawback compared to the classical Q-learning algorithm, there are certain cases where SARSA may exhibit better convergence properties than Q-learning, e.g. a simple grid-world walking problem used by Sutton and Barto [87] throughout their book. Therefore, it is also one of the most widely used RL algorithms in the general RL literature and could prove to be effective for DSA in wireless environments.

Actor-Critic Learning

The actor-critic learning methods belong to another popular type of RL, first investigated by Witten [97]. Their general structure is slightly different from that of Q-learning and SARSA. They explicitly separate the policy and the value table in the learning process as shown in Figure 2.7.

The policy is an *actor* responsible for choosing an action in a given state of the environment, and the value table is a *critic* which observes the outcomes and rewards caused by the chosen actions and critiques them accordingly. If the critique is positive, then the probability of the actor choosing the same action in the same state in future

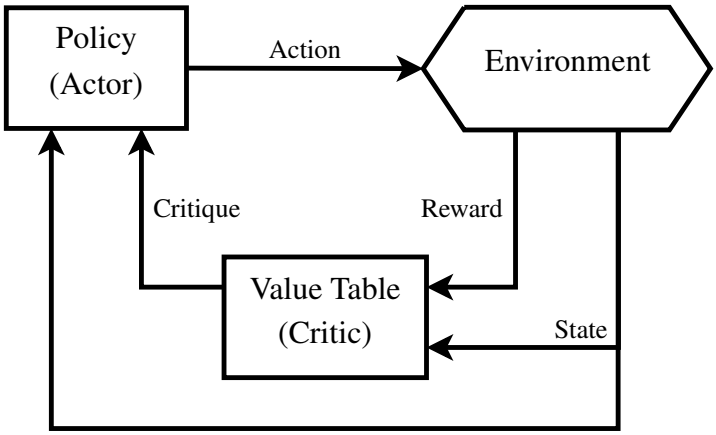


Figure 2.7: Structure of the actor-critic learning methods

should be increased and vice versa [87].

One potential advantage of such methods is their ability to learn an explicitly stochastic policy, i.e. to optimise a probability distribution of selecting various actions in all environment states, e.g. a probability distribution over the potential channels or spectrum holes that could be accessed by a cognitive wireless device. However, a typical disadvantage of the actor-critic methods discussed by Grondman et al. [33] is the lack of adaptability of the critic part of the algorithm in dynamically changing environments. This is a significant issue that limits the applicability of such methods in realistic wireless communications environments which are likely to have a dynamic nature, e.g. in terms of the offered traffic levels and the network topology.

Stateless Q-learning

In some learning problems where an environment does not have to be represented by states, the learning agents are stateless and only the action space and a 1-dimensional Q-table $Q(a)$ can be considered [21][46]. The job of an RL algorithm then becomes simpler, it aims to estimate an expected value of a single reward for each action available to the learning agent:

$$Q(a) = E[r_t] \tag{2.5}$$

where $Q(a)$ is the Q-value of action a and $E[r_t]$ is the immediate reward the learning agent expects to receive after taking action a at time t. An equivalent representation of the classical RL algorithms such as Q-learning and SARSA that consider environments

represented by states is the following:

$$Q(s,a) = E\left[\sum_{t=0}^{T} \gamma^t r_t\right] \tag{2.6}$$

where $Q(s,a)$ is the Q-value of action a in state s, r_t is the numerical reward received t time steps after action a is taken in state s, T is the total number of time steps until the end of the learning process or episode, and γ is a discount factor.

Claus and Boutilier [21] propose the stateless Q-learning algorithm for independent learners in co-operative multi-agent systems, a simplified version of the classical Q-learning algorithm. Its recursive update equation is given below:

$$Q(a) \leftarrow (1-\alpha)Q(a) + \alpha r \tag{2.7}$$

where $Q(a)$ represents the Q-value of the action a, r is the reward associated with the most recent trial and is determined by a reward function, and $\alpha \in [0,1]$ is the learning rate parameter which weights recent experience with respect to previous estimates of the Q-values.

The advantage of formulating learning environments as stateless decision problems and employing the stateless Q-learning algorithm instead of its classical counterpart is the significant reduction in the number of Q-values that need to be estimated by the learning agent, and, therefore, a potentially dramatic reduction in the number of trials needed for it to learn a mature strategy. The latter is also likely to be caused by the fact that the problem of estimating every individual Q-value in stateless Q-learning is significantly simpler as demonstrated by Equation (2.5), as opposed to classical Q-learning or SARSA described by Equation (2.6). Such a significant increase in the speed of the learning process would directly translate into the higher adaptability of RL based cognitive wireless devices, since it would take them less time to learn appropriate DSA policies in a new or dynamically changing wireless environment.

2.3.3 Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) is concerned with cases when there is more than one learning agent in the same environment. MARL has strong links with game theory. An MDP in single-agent RL becomes a stochastic game (SG) in MARL, sometimes also referred to as a multi-agent MDP. A large number of MARL algorithms are based on game theory, since it is one of the most suitable frameworks to model the interactions among several agents in a common environment [55]. This gives rise to the investigation of the applications of MARL to different types of SGs - fully cooperative, fully competitive and mixed games.

Extending RL to the multi-agent case presents several challenges investigated by Busoniu et al. [14]. In many cases a formal definition of a multi-agent learning goal becomes a difficult task. Every learning agent is affected by the actions of the other learning agents. Therefore, the environment is no longer static, it becomes highly dynamic from the viewpoint of each individual agent. This significantly increases the complexity of the learning tasks and invalidates most convergence guarantees of single-agent RL. A popular way to specify a MARL goal is to use a Nash Equilibrium (NE), as used in the game theory context, where none of the agents in the environment has an incentive to deviate from its policy.

Nevertheless, employing the MARL methods also presents a number of benefits [14]. For example, there is scope for experience sharing among the learning agents to improve the initial and steady-state performance of an RL algorithm and, thus, to increase its adaptability. This paradigm lies within the emerging research topic of transfer learning (TL), sometimes also referred to as docitive learning in the wireless communications domain [31]. MARL is also inherently more robust than SARL in that in a certain type of RL problems the faulty agents can be supported or replaced by new ones. Finally, there is a high degree of scalability in MARL, because most MARL algorithms allow easy insertion of new learning agents into the environment.

The rest of this subsection gives examples of several notable MARL algorithms found in the literature.

Nash-Q

The Nash-Q algorithm introduced by Hu and Wellman [36] is an extension of Q-learning to the multi-agent case, where the goal of all agents is to converge to an NE strategy in every state of the environment. The drawback of this algorithm is that every learning agent is supposed to observe the actions taken and rewards received by all other learning agents, and to store all their Q-tables. This is an assumption that may not be valid in many learning problems. It is also inefficient in terms of memory and communication overhead among the agents. However, the advantage of this method, as presented by Hu and Wellman [36], is the proven convergence of this algorithm towards a mixed strategy NE, which is rare in the MARL domain.

Distributed-Q

The Distributed-Q algorithm for fully cooperative SGs is proposed by Lauer and Riedmiller [51]. Here, every learning agent senses the entire environment and performs a single-agent Q-learning algorithm assuming that all other agents will be choosing a certain greedy action at all times. This works extremely well in deterministic environments. However, in the wireless communications domain the real-world learning problems are bound to be highly stochastic instead, due to random environmental effects which cannot be modelled and predicted. It also assumes that every learning agent is able to accurately estimate the greedy actions of the other agents. This may not be possible in a number of distributed multi-agent learning problems.

Conjecture-Based Reinforcement Learning

A more promising variation of multi-agent Q-learning recently proposed by Chen et al. [19] is called conjecture-based RL. It deals with the stochastic nature of the learning process by defining a *conjecture* term which is used in the Q-table update formula. It is effectively a probability of all other learning agents in the environment choosing a particular set of policies, which determines the reward received by the learning agent. It then calculates the expected reward as a weighted sum of possible rewards depending on policies chosen by other agents. Chen et al. [19] successfully use this algorithm

to enable CR devices in a simulated wireless mesh network to learn optimal spectrum and power allocation strategies for improved energy efficiency of the network. However, this approach has only been applied to a relatively small and analytically tractable scenario with six secondary users and five primary users. The scalability of this algorithm has not been tested. For example, it is not clear whether this algorithm would exhibit good performance during the initial exploration stage of the learning process in a significantly larger and more complex wireless environment, and whether it would maintain its property of converging towards optimal strategies.

Independent Single-Agent Reinforcement Learning

The simplest approach to MARL is the "naive" implementation of independent single-agent RL algorithms for each learning agent in the environment, e.g. [77][88]. Despite the fact that the independent learning agents are not even aware of the existence of the other learning agents in the environment, this approach has been successfully applied to various coordination tasks, e.g. [46][77]. For example, an implementation of independent stateless Q-learning agents in a multi-agent environment has also been shown to exhibit remarkably similar convergence performance in a simple coordination task as the "joint action learner", but with significantly less information available to the learning agents [21].

The fundamental advantage of this approach is the lack of assumptions about each learning agent's awareness of the actions performed by the other agents required by the rest of the MARL algorithms described in this subsection so far. It significantly increases the breadth of potential applications of this MARL approach with different information availability constraints, including those in the wireless communications domain.

Heuristically Accelerated Reinforcement Learning

A common disadvantage of RL algorithms is their need for many learning iterations to converge on an acceptable solution. A lot of researchers have been addressing this problem, and one of the more recent promising solutions is the *heuristically ac-*

celerated reinforcement learning (HARL) approach. Its goal is to speed up the RL algorithms, particularly in the multi-agent domain, by guiding the exploration of the state space using additional heuristic information. According to Bianchi et al. [11], a heuristic policy is derived from additional knowledge, either external or internal, which is not included in the learning process. The goal of the heuristic policy is to influence the action choices of a learning agent, i.e. to modify its current policy in a way which would accelerate the learning process. For example, the first evidence of HARL in the literature is the paper by Bianchi et al. [12], where a heuristic function $H(s, a)$ is defined that dictates which actions should be taken in which states to explore the state-space more efficiently. This function can be obtained from additional expert knowledge or "existing clues in the learning process itself" [12]. In [11] the authors prove the convergence of four multi-agent HARL algorithms and demonstrate that they outperform their classical RL counterparts.

This approach is particularly relevant in the DSA environment where various standardised signals with useful spectrum awareness information may be available to the learning agents.

2.4 Intelligent Dynamic Spectrum Access

This section presents recent developments in the field of intelligent DSA. In the context of this thesis intelligent DSA methods are defined as those based on machine intelligence techniques which involve all aspects of the modified Mitola's cognition cycle of wireless devices discussed in Subsection 2.2.4, particularly the learning and planning functionality.

A large amount of research on intelligent DSA in wireless networks focuses on RL techniques, e.g. [7][43][65][89]. The RL algorithms applied to DSA problems can generally be divided into two groups, centralised and distributed. The centralised methods employ one RL agent which controls the operation of the whole network, whereas the distributed methods are multi-agent RL systems which involve significantly less network-level information exchange and primarily use local measurements to make spectrum assignment decisions.

2.4.1 Centralised Reinforcement Learning Approach

Early research work on RL based DSA largely focuses on centralised methods, which use a single control unit for the whole network. It has access to all network information and better suits the original Q-learning algorithm developed by Watkins [94], the most widely used RL algorithm to date. One of the main advantages of the classical single-agent Q-learning approach is that it was proven to converge on an optimal solution in a single-agent MDP context, as explained in Subsection 2.3.2. As soon as other Q-learning agents are introduced into the environment, this convergence is no longer guaranteed.

A classical example of the original Q-learning algorithm applied to a centralised DSA problem is the algorithm proposed by Nie and Haykin [65], where a state of the environment is determined by the index of a cell where a call arrival occurs and the number of channels available for assignment in the given cell. This algorithm is shown to significantly outperform fixed spectrum assignment schemes using a classical cellular network simulation model. It also produces comparable performance to the best known DSA scheme to date - MAXAVAIL, but with a significant reduction in computational complexity. Senouci and Pujolle [78] extend the work of Nie and Haykin [65] and implement a centralised Q-learning algorithm which is capable of learning DSA policies considering call admission control, channel assignment and two classes of traffic, all incorporated in their proposed semi-MDP model of the problem. Singh and Bertsekas [82] model the states of the DSA problem in classical cellular networks using the list of occupied and unoccupied channels at each cell and the event that can cause a state transition, i.e. call arrival, departure or hand-off. They then use a temporal difference RL algorithm [87] to enable the cellular network to learn the best channel reuse patterns depending on the state of the network. These learnt dynamic channel reuse policies are shown to outperform the best analytical methods found in the literature to date.

Although all of these early works have been instrumental in generating great interest in RL based DSA in the wireless communications research community, a crucial part of the evaluation of their proposed RL algorithms is missing, namely the temporal

characteristics of the learning process. Since classical RL algorithms are based solely on the trial-and-error experience of the learning agent, it typically takes a large number of trials for it to build up a mature knowledge base and to learn an acceptable solution to the given decision problem [87]. In RL based DSA this initial exploration process is likely to result in poor QoS provided to the users of the given wireless environment. Therefore, it is essential to analyse the network performance throughout all stages of the learning process and to minimise the deterioration in QoS caused by the exploration process of the RL mechanism in place. The main goal of most contributions of this thesis is to alleviate this problem of poor temporal performance of RL based DSA algorithms, and to make them more robust and adaptable in such challenging real-time decision problems as DSA in wireless networks.

2.4.2 Distributed Reinforcement Learning Approach

Distributed intelligent DSA schemes became significantly more popular than the centralised methods since the introduction of CR networks which normally involve distributed decision making by a number of wireless devices [42]. For example, Jiang et al. [41] apply distributed RL with an explicit random exploration stage to a network of independent CR transmitter-receiver pairs to enable them to learn efficient spectrum sharing patterns. In [43] Jiang et al. further improve the performance of this distributed RL algorithm by combining it with a more efficient weight-driven exploration scheme. In addition to being fully distributed, the fundamental difference between the DSA scheme proposed by Jiang et al. [41][43] and the centralised RL methods discussed in the previous subsection is that the former uses spectrum sensing as the primary source of information to inform spectrum access decisions, while RL is used to enable the CR devices to suggest appropriate channels with potentially low interference on them. Therefore, the poor performance during the exploration stage of the learning process is not a major issue there, since the opportunistic interference sensing approach introduced in Subsection 2.2.2 will always be able to achieve adequate QoS before the knowledge obtained through RL further improves it.

Wu et al. [98] propose a MARL based Q-learning approach where every CR device in the radio environment learns a spectrum and power allocation strategy using the

strategies of other CR transmitters as the state information. One of the disadvantages of this approach is the potentially high communication overhead required to accommodate the assumption that every CR device in the environment is always aware of the current strategy of all other CR devices. Another disadvantage of this approach which is common to all classical RL algorithms is the poor system performance at the initial stage of the learning process. It takes a significant amount of time for the CR devices to learn appropriate DSA strategies that achieve an acceptable probability of successful transmission.

An example of distributed RL based DSA in cellular networks is the implementation of the SARSA algorithm proposed by Lilith and Dogancay [54] which is shown to significantly reduce the call blocking probability in a simulated 49 cell network over a 24-hour period with the typical time-dependent offer traffic pattern, compared to the fixed channel allocation approach. In [53] Lilith and Dogancay also show that their distributed SARSA algorithm with the purposely reduced number of states in the Q-table exhibits comparable performance to that of the centralised RL approach, but with no communication overhead associated with the latter. However, the authors have not compared the performance of their proposed algorithm with any state-of-the-art DSA methods, nor have they compared the temporal variations of the call blocking probability using the distributed RL approach to a non-RL based method, e.g. fixed channel allocation, to verify that a dramatic deterioration in QoS during the traffic peak-times observed in [54] is not caused by the RL exploration process and is common across all considered spectrum management schemes.

Figure 2.8 illustrates how such distributed RL based DSA methods operate in cellular networks. Each BS maintains its own Q-table which, in the case of the stateless Q-learning algorithm described in Subsection 2.3.2, has a Q-value associated with every channel available for assignment. After a sufficient number of trials each BS builds up its own knowledge base which reflects any predictable or unpredictable radio propagation effects from its trial-and-error experience, and uses this table to make the spectrum assignment decisions. A significant advantage of this approach is its scalability. It is not associated with any particular size of the network or its topology. Therefore, if BSs or other cognitive wireless devices are dynamically inserted or removed from the

environment, it will be handled completely autonomously by the learning algorithms implemented in every individual device.

A more modern example of the application of distributed RL based DSA in cellular networks is the algorithm proposed by Bennis et al. [7]. They first present a game theoretic model of interference management in heterogeneous networks that involve a number of small cell BSs underlaying a high power macro-BS considered as the primary spectrum user. They then use this model to design a distributed RL algorithm that enables the small cell BSs to learn appropriate transmitter configurations, i.e. spectrum and power allocation policies, whilst successfully converging towards an equilibrium where the interference received by the primary macro-BS users is below a pre-defined limit. The drawback of their algorithm is the inherent problem encountered in all classical RL algorithms - the poor initial performance due to the lack of prior knowledge of the environment. In the case of the DSA algorithm proposed by Bennis et al. [7], at the start of the learning process performed by the small cell BSs the probability of the primary macro-BS users receiving excessive interference from them is between 0.35 and 0.55 which is unacceptable if strict primary user QoS guarantees have to be adhered to.

Feki et al. [26] propose an RL algorithm based on the cyclic multi-armed bandit formulation of the spectrum sharing problem in LTE cellular networks. Their algorithm autonomously steers each cell in the network towards using the most suitable portions of the available spectrum band, taking into account the spatial offered traffic distri-

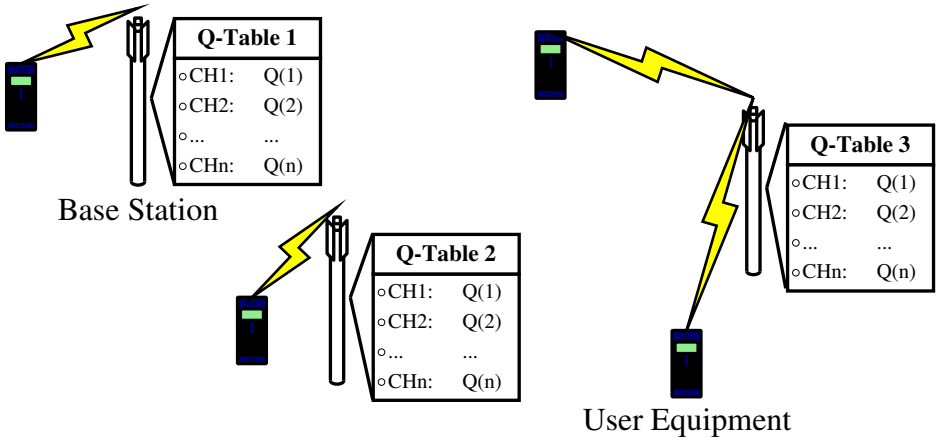


Figure 2.8: Distributed reinforcement learning based dynamic spectrum access in a cellular network

bution. Although the authors focus on the speed of convergence of their proposed distributed RL algorithm, they do not present simulation results that describe the QoS in the network at various stages of the learning process, which is a key aspect of the performance of intelligent DSA algorithms investigated in this thesis.

2.4.3 Transfer Learning Approach

An emerging approach for alleviating the problem of limited information availability and improving the convergence behaviour of distributed RL based DSA algorithms is transfer learning (TL). The fundamental idea behind TL is depicted in Figure 2.9, where, instead of learning DSA strategies completely independently as shown in Figure 2.8, the BSs periodically exchange their acquired knowledge to speed up the learning process of every individual cognitive BS.

For example, Zhao et al. [104] use this methodology to dramatically improve the convergence speed and QoS achieved by a distributed stateless Q-learning algorithm applied to a small cell network covering streets in an urban environment. However, since in this study the BSs are arranged in lines along the streets, the authors force the BSs to use a simple reuse pattern by manipulating the Q-table and inverting the order of preferred spectrum resources of every other BS. Therefore, the transfer of the knowledge acquired by the BSs purely through distributed RL is not a key source of

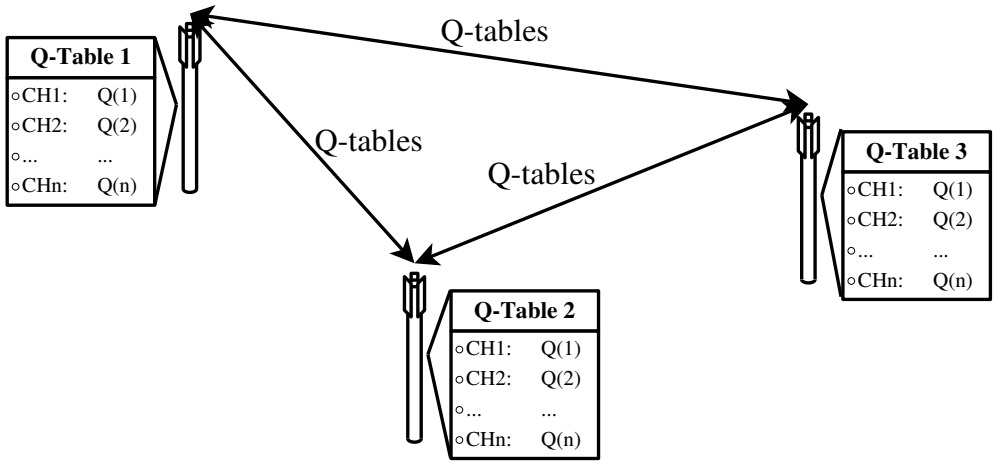


Figure 2.9: Transfer learning based dynamic spectrum access in a cellular network, where the base stations periodically exchange their knowledge to aid the distributed learning process

information used in this particular investigation. In [103] Zhao et al. develop a co-operation management algorithm that dynamically adapts the amount of information exchange overhead required during the TL process depending on the learning stage. It achieves a 90% reduction in the amount of information exchanged among the nodes of a multi-hop backhaul network with no negative effect on the system QoS and throughput. However, despite the significant improvement in the system QoS and convergence speed achieved by TL with relatively little communication overhead compared to fully coordinated DSA schemes, it still suffers from the issue of poor performance at the initial stage of learning when none of the learning agents have had enough time to build up a knowledge base that could be beneficial when transferred to other learning agents.

Cognitive wireless networks that employ TL are also sometimes referred to as docitive networks. For example, Giupponi et al. [31] introduce the concept of docitive networks as an extension to previously proposed cognitive wireless networks, where some opportunistic wireless nodes "teach" other nodes by transferring their knowledge in exactly the same way as in TL. They implement this paradigm in a simulated IEEE 802.22 WRAN coexisting with a primary TV broadcasting network. The docition process is shown to greatly improve the convergence speed of the distributed RL based DSA approach employed by the secondary WRAN BSs. However, the time response plot that compares a number of cognitive and docitive DSA algorithms does not start from zero, but is only given between 300,000 and 400,000 learning iterations. Therefore, it does not show the initial performance of the docitive approach when the are no wireless nodes that could be used as "teachers" for other nodes yet.

Shahid et al. [80] extend the distributed Q-learning based joint resource allocation and power control algorithm to the docitive case, where a number of femto-cell BSs that underlay a macro-cell share their learnt strategies in order to increase the capacity of the secondary femto-cell network whilst adhering to the primary macro-cell user QoS guarantee requirement. Although the docitive approach improves the convergence speed compared to a classical distributed RL approach as expected, the initial performance of both algorithms is extremely poor - a negligibly small capacity of the femto-cell network and a large amount of harmful interference for the primary macro-

cell users is observed. This once again demonstrates that, although TL is a powerful method for speeding up the learning process of RL based DSA algorithms, it still exhibits the fundamental weakness of all trial-and-error based RL algorithms - the poor initial performance due to the lack of prior knowledge available at the start of the learning process. This property of RL based DSA algorithms would also significantly limit their adaptability in dynamic wireless environments, where the learning agents are required to rapidly adapt to changes in their environment, i.e. changes in the parameters of the problem they are trying to solve.

2.5 Conclusion

Spectrum sharing and DSA techniques play a key role in utilising the mobile spectrum efficiently. A large number of classical approaches to DSA are based on spectrum databases, dynamic interference measurements and temporary licenses. However, this thesis focuses on more flexible intelligent DSA techniques that involve the full cognition cycle of wireless devices originally defined by Mitola [61]. The widely investigated state-of-the-art method for intelligent DSA is RL. This chapter gave an overview of a range of single-agent and multi-agent RL algorithms found in the literature both in the general context and those specifically designed for cognitive wireless networks. Although RL presents a promising solution to enable intelligent DSA, the inherent disadvantage of all classical trial-and-error based RL algorithms is the poor system performance at the early stage of the learning process due to the agents' lack of initial knowledge about the environment. This property of RL based DSA algorithms also significantly limits their adaptability in dynamic radio environments.

Chapter 3. Experimental Methodology

Contents

3.1 Cognitive Wireless Network Simulator

This thesis proposes a number of intelligent DSA algorithms designed to be adaptable and robust in realistically challenging wireless environments. In order to empirically demonstrate their adaptability and robustness, a sufficiently complex simulation model is required that would appropriately describe a relevant and realistic DSA and spectrum sharing scenario. Therefore, the simulation scenario chosen for empirical evaluation of the DSA algorithms proposed in this thesis is the stadium temporary event scenario considered in the EU FP7 ABSOLUTE project. It involves a temporary heterogeneous cognitive cellular infrastructure that is deployed in and around a stadium, alongside

a local primary LTE network, to provide extra capacity and coverage to the mobile subscribers and event organizers involved in a temporary event, e.g. a football match or a concert [71]. The details of this scenario and the network architecture are described in the rest of this section.

3.1.1 Scenario and Network Architecture

The scenario is depicted in Figure 3.1. Here, a small cell LTE network is deployed inside the stadium to provide ultra high capacity density to the event attendees, and an eNodeB on an aerial platform (AeNB) is deployed above the stadium to provide wide area coverage. The AeNB is located above the stadium centre point at 300m altitude. The model also includes a local LTE network that consists of 3 primary eNBs (PeNBs) whose coordinates, with respect to the centre point of the stadium, are $(-600, -750)$, $(100, 750)$ and $(750, -800)$ metres.

The stadium small cell network architecture is depicted in Figure 3.2, where the users are located in a circular spectator area 53.7 - 113.7m from the centre of the stadium. The spectator area is covered by 78 eNBs arranged in three rings at 1m height, e.g. with antennas attached to the backs of the seats or to the railings between the different row levels. The seat width is assumed to be 0.5m, and the space between rows - 1.5m, which yields the total capacity of 43,103 seats.

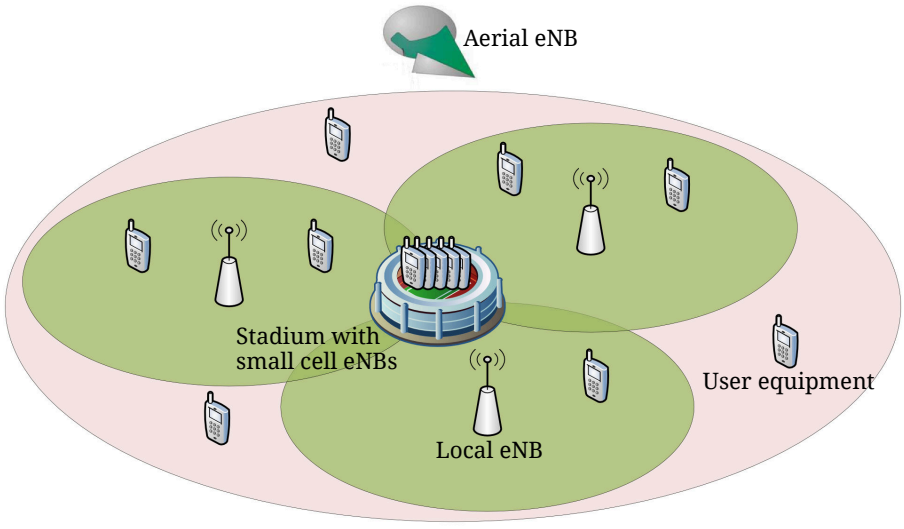


Figure 3.1: Stadium temporary event scenario

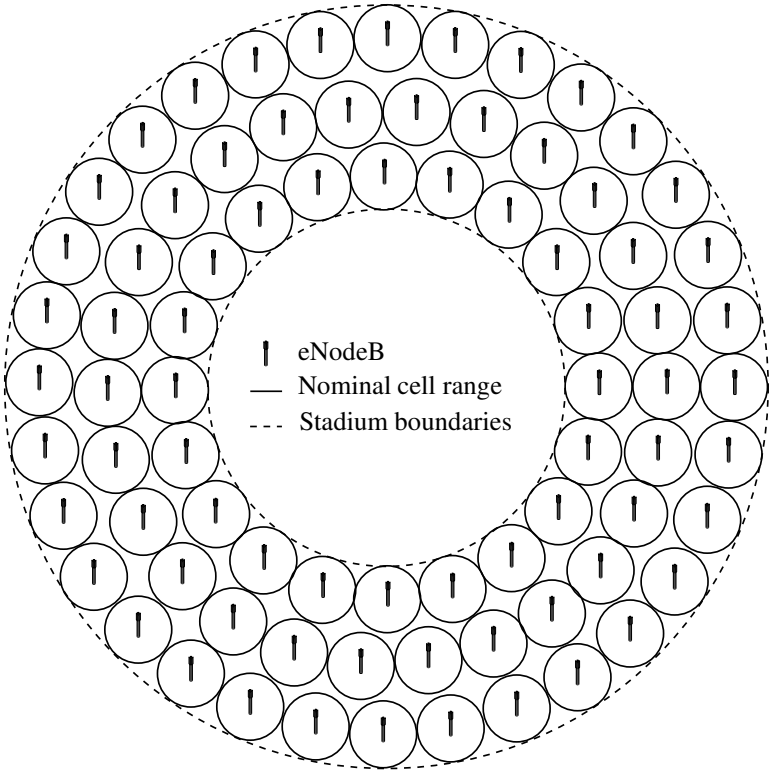


Figure 3.2: Stadium small cell network architecture

500 user equipments (UEs) are randomly distributed outside the stadium, in the circular area from the stadium boundary (5m from the radius of the last row) to 1.5km away from the stadium centre point. 25% of the stadium capacity is filled with randomly distributed wireless subscribers, i.e. $\approx$ 10,776 UEs.

All simulations reported in this thesis use a 20 MHz LTE channel in the 2.6 GHz frequency band. The 20 MHz bandwidth of an LTE channel is divided into 100 virtual resource blocks (VRBs), each having a 180 kHz bandwidth [3]. The spectrum entity that is assigned to any data transmission is referred to as the subchannel and consists of four consecutive VRBs, according to the standardised LTE Type 0 resource allocation [3]. Therefore the transmission bandwidth of a subchannel assigned to any given user is $4 \times 180 = 720$ kHz.

The simulation experiments are divided into two different spectrum management cases:

1. The stadium small cell network has access to its own dedicated 20 MHz LTE channel, e.g. using a temporary LSA license for the use of this spectrum as described in Subsection 2.2.3. In this case its performance is assessed separately, not considering the AeNB and the PeNBs.

2. The cognitive small cells and the AeNB have secondary access to a 20 MHz LTE channel, also used by a local network of three PeNBs. This represents a more challenging secondary spectrum sharing task, where the primary user QoS guarantees are also taken into account.

3.1.2 Radio Propagation

An appropriate model for calculating the propagation loss inside the densely populated stadium is the WINNER II B3 non-line-of-sight model designed for airports, factories, conference halls etc. [50]. It is described by the following equation:

$$PL = 37.8log_{10}(d) + 36.5 + 23log_{10}(\frac{f_c}{5}) + \gamma \tag{3.1}$$

where PL is the path loss in dB, d is the propagation distance in metres, $f_c = 2.6$ is the carrier frequency in GHz and γ is the log-normally distributed shadow fading loss with 0dB mean and 4 dB standard deviation.

The WINNER II C1 line-of-sight suburban macro-cell model is used for propagation between the local PeNBs and the users outside of the stadium [50]:

$$\begin{aligned} PL = \ & 40log_{10}(d) + 11.65 - 16.2log_{10}(h_{BS}) \\ & -16.2log_{10}(h_{UE}) + 3.8log_{10}(\tfrac{f_c}{5}) + \gamma \end{aligned} \tag{3.2}$$

where $h_{BS} = 30m$ is the base station height, $h_{UE} = 1m$ is the height of the UE antenna, and the standard deviation of the shadow fading loss γ is 6 dB.

To calculate the propagation loss between outdoor transmitters and indoor receivers and vice-versa the WINNER C4 II outdoor-to-indoor model is used [50]. However, the C2 term there is replaced by a C1 term, to represent the same suburban outdoor environment described by Equation (3.2) instead of an urban one:

$$PL = PL_{C1}(d) + 17.4 + 0.5d_{in} - 0.8h_{UE} \tag{3.3}$$

where $PL_{C1}(d)$ is the WINNER II C1 path loss described by Equation (3.2) with a 10dB standard deviation of the shadow fading loss, and d_{in} is the indoor part of the

distance between the transmitter and the receiver, i.e. the distance between the stadium shell and the eNB/UE inside the stadium.

The propagation loss between the AeNB and receivers on the ground is calculated using the free space path loss model with 8dB log-normal shadow fading:

$$PL = 20log_{10}(d) + 46.4 + 20log_{10}(\frac{f_c}{5}) + \gamma \tag{3.4}$$

3.1.3 Link Model

A realistic value of the noise floor used for the UE receivers is -124 dBW. It is calculated using the following formula:

$$P_N = 10log_{10}(kTB) + N \tag{3.5}$$

where P_N is the noise power in dBW, $k = 1.38 \times 10^{-23}$ m^2 kg s^{-2}K^{-1} is the Boltzmann constant, $T = 290$ is the noise temperature in K, $B = 2 \times 10^7$ is the bandwidth in Hz and $N = 7$ is the noise figure in dB.

The link quality is determined by the signal-to-interference-plus-noise ratio (SINR), i.e. the ratio between the power of the received signal of interest and the sum of the received powers from interfering transmitters together with the noise power. The SINR at a given receiver on a given subchannel is calculated as follows:

$$SINR = \frac{P_{Tx}^k G_{Tx}^k G_{Rx} PL_k^{-1}}{\sum_{i=1}^{N_I} P_{Tx}^i G_{Tx}^i G_{Rx} PL_i^{-1} + P_N} \tag{3.6}$$

where the signal of interest is received from the transmitter k, G_{Tx}^k is the antenna gain of transmitter k, G_{Rx} is the receiver antenna gain, PL_K is the propagation loss between transmitter k and the receiver, N_I is the number of interfering transmitters, i.e. all other transmitters that are using the same subchannel, and P_N is the receiver noise floor calculated using Equation (3.5) and converted to W. The antenna gains for the eNBs and UEs are 3 dB and 0 dB respectively.

Given the SINR level, the link throughput is calculated using the following 3GPP

truncated Shannon bound model for LTE downlink [2]:

$$Throughput = \begin{cases} 0, & SINR < SINR_{min} \\ \alpha B log_2(1 + SINR), & SINR_{min} \leq SINR < SINR_{max} \\ \alpha B log_2(1 + SINR_{max}), & SINR \geq SINR_{max} \end{cases} \tag{3.7}$$

where $\alpha = 0.6$ is the attenuation factor due to implementation loss, B is the bandwidth of the link, $SINR_{min}$ is the minimum SINR capable of supporting a data transmission, and $SINR_{max} = 22$ dB is the SINR that corresponds to the maximum achievable link throughput. The minimum SINR allowed to support data transmissions to avoid very low quality links is 1.8 dB [44].

3.1.4 Traffic Model

The simulated data traffic is generated using the 3GPP File Transfer Protocol (FTP) model 1 [1]. It is a simple yet realistic model of random bursty traffic that reflects typical behaviour of internet and mobile network users. It uses the negative exponential distribution for the calculation of file inter-arrival times and a fixed file size of 4.2 Mb (≈0.5 MB). The length of each file transmission is calculated by dividing the file size by the link throughput calculated using Equation (3.7).

3.1.5 Power Control and Cell Association

The local PeNBs use the fixed transmit power of 10W. The cognitive base stations, i.e. the stadium small cell eNBs and the AeNB, employ open-loop power control using a constant target received power of -104 dBW, i.e. for a 20 dB signal-to-noise (SNR) ratio. This is a simple power control mechanism that counteracts the effects of shadowing and distance losses and provides a fair signal strength distribution across the whole network [62], e.g. equal received power at the cell centre and the cell edge.

Every UE inside the stadium is associated with a small cell or the AeNB with the minimum estimated downlink path loss, e.g. based on the Reference Signal Received Power (RSRP). The UEs outside of the stadium are associated either with a PeNB or

the AeNB based on the strongest RSRP. The reference signal Tx power of the AeNB is assumed to be 13 dB lower than that of the PeNBs to avoid potential high power interference from the AeNB to the primary users. This is also consistent with a maximum 27 dBmW transmit power of the AeNB defined in the ABSOLUTE project [32].

3.1.6 Inter-Cell Interference Coordination Signalling

The cognitive wireless network scenario described in this section is based on LTE; the current state-of-the-art radio access technology (RAT) for mobile broadband networks. One of the key LTE interference management technologies, that is also an integral part of the DSA algorithms proposed in Chapters 6 and 7 and that features in most of the other simulation experiments discussed in this thesis, is known as inter-cell interference coordination (ICIC). The purpose of ICIC is to reduce interference between adjacent cells by exchanging information between neighbouring eNBs over the dedicated X2 interface [79]. This ICIC signal exchange is depicted in Figure 3.3 using a generic hexagonal cell network architecture. Here, the central eNB is sending an ICIC signal to the eNBs around it to let them know in which parts of the spectrum it is likely to interfere with them.

The format of the messages exchanged between eNBs using ICIC in the LTE downlink is standardized by the 3GPP and referred to as the Relative Narrowband Transmit Power (RNTP) indicator [3]. It contains a bitmap which indicates on which resource blocks an eNB is planning to transmit at high power by setting their corresponding

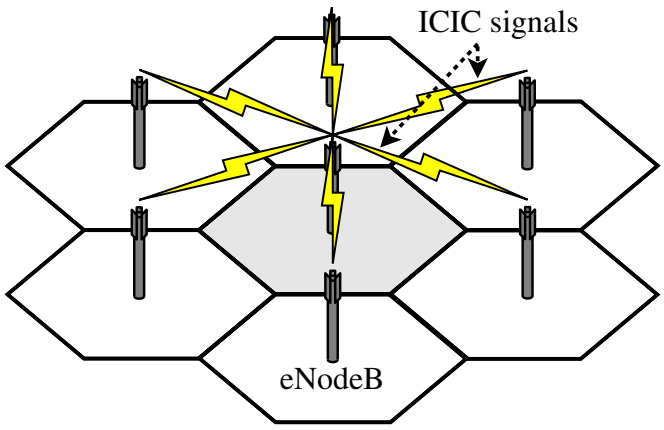


Figure 3.3: Inter-cell interference coordination (ICIC) signalling among neighbouring eNodeBs

bits to *1*, i.e. on which resource blocks it is likely to cause interference in adjacent cells. For example, in a scenario, where a 20 MHz LTE channel consisting of 100 VRBs is allocated to the network, the length of an RNTP message is 100 bits or 25 hexadecimal characters. In this case every subchannel, a minimum entity allocated to a file transmission, consists of 4 adjacent resource blocks [3]. Therefore, every group of 4 bits (i.e. every hexadecimal character) in an RNTP message describes a particular subchannel. For example, if an eNB is planning to use high transmit power on a given subchannel, its corresponding bits in the RNTP message are *1111* or *0xF*, and *0000* or *0x0* in the opposite case.

The threshold used to decide whether a given transmit power is high or low is derived using the average transmit power in the cell and the RNTP threshold, which can take the following set of standardized values [3]:

$$RNTP_{thresh} \in \{\infty, -11, -10, -9, ..., 1, 2, 3\}\ dB \tag{3.8}$$

It is measured in dB relative to the average transmit power in a given cell. To avoid excessive signalling requirements, the minimum allowed time interval between the ICIC message exchanges is 20 ms [79].

3.2 Empirical Evaluation

An important aspect of the empirical evaluation of the quality of service (QoS) and capacity performance of cognitive wireless networks is the appropriate choice of the metrics used to quantify it. The network performance metrics used to analyse the simulation results presented in this thesis are described in the following subsection.

3.2.1 Performance Metrics

The key metrics used to assess the network performance in this thesis are the probability of retransmission $P(re-tx)$, mean and 5% user throughput (UT), and the overall system throughput density.

$P(re-tx)$ is the probability of a file transmission being blocked or interrupted, i.e. the probability of a retransmission being scheduled. It is calculated using the following equation:

$$P(re-tx) = \frac{N_{re-tx}}{N_{re-tx} + N_{successful-tx}} \tag{3.9}$$

where N_{re-tx} and $N_{successful-tx}$ are the number of retransmissions and the number of successfully completed transmissions during one sampling period respectively.

Mean and 5% UT are the metrics that describe the distribution of the average data rates provided to the users. Mean UT is calculated over all UEs in the network, whereas the 5% UT gives the minimum guaranteed UT for 95% of the users. The latter is obtained by calculating the 5th percentile of the UT distribution over all UEs, and is the key metric for ensuring fair QoS distribution across the whole network. The equation for calculating UT for any given UE, as defined in [1], is given below:

$$UT = \frac{\sum_{f=1}^{F} S_f}{\sum_{f=1}^{F} T_f} \tag{3.10}$$

where F is the number of files downloaded by the given UE, S_f is the size of the f'th file, and T_f is the time it took to download it.

System throughput density (STD) of the stadium network is obtained by calculating the average system throughput during the whole simulation and dividing it by the area covered by the eNBs, as shown in the equation below:

$$STD = \frac{Throughput}{\pi R_{outer}^2 - \pi R_{inner}^2} \tag{3.11}$$

where $Throughput$ is the average system throughput measured throughout the whole simulation, R_{outer} is the outer radius of the spectator area - 113.7 m, and R_{inner} is its inner radius - 53.7 m. This performance metric is especially important for small cell scenarios such as the one described in this section, since it demonstrates the spatial efficiency of spectrum reuse achieved by employing such small cell sizes.

In simulations involving the stadium network, the AeNB and the primary system, these metrics are calculated separately for different classes of the users, e.g. based on whether they are inside or outside the stadium or on the type of base station they

are connected to.

3.2.2 Statistical Validation of Results

In order to ensure the validity and statistical significance of the key results presented in this thesis, the following techniques are applied where relevant:

- Data points on the plots of network performance against time or offered traffic sweeps are obtained by averaging over 50 different simulations with different random seeds, UE locations and initial file traffic.
- The offered traffic sweep graphs also include error bars showing the difference between the minimum and the maximum value from 50 different simulations which correspond to a given data point.
- Furthermore, some results are expressed in the form of box plots [59], a compact way of depicting key features of probability distributions such as the median, the 1st and 3rd quartile, and the minimum and maximum data point values.

3.3 Heuristic Schemes for Baseline Comparison

This thesis predominantly uses two heuristic DSA schemes for baseline comparison: a typical approach in standard LTE networks, and an opportunistic approach commonly used in cognitive radio networks. These schemes are described in the following subsections.

3.3.1 Dynamic ICIC

The dynamic ICIC scheme used for baseline comparison in this thesis is a typical approach to interference management in conventional LTE networks [28][79]. It assumes that each eNB always avoids transmitting on the resources used by its neighbours, reported in their ICIC signals explained in Subsection 3.1.6. A given eNB chooses randomly among the subchannels that are not used by any of its neighbours and blocks file

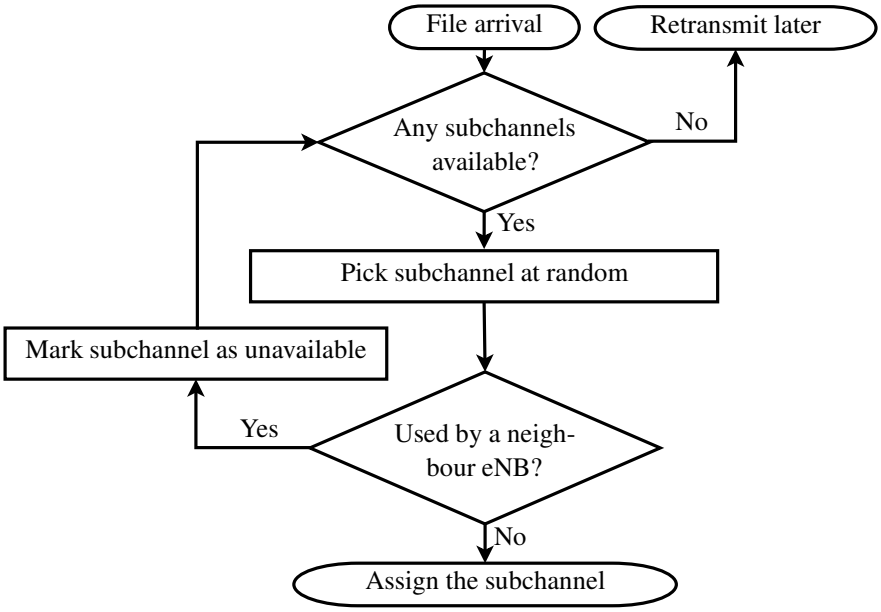


Figure 3.4: Flow diagram of the dynamic ICIC scheme used for baseline comparison

transmissions when no such subchannels are available for assignment. The flowchart of this scheme is shown in Figure 3.4.

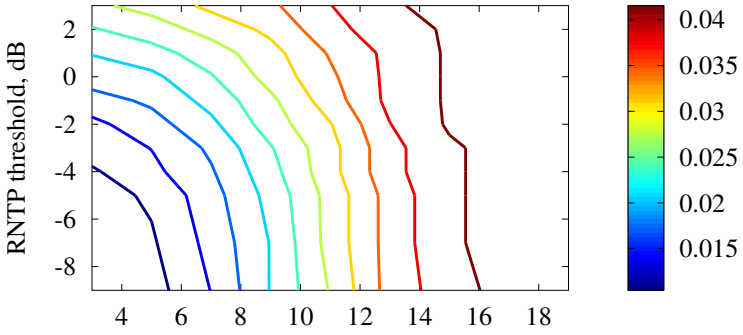
ICIC signal exchanges are assumed to take place highly frequently - every 20 ms [79]. Therefore, the current subchannel usage of a given eNB is always mapped onto its ICIC message, since an eNB is highly likely to continue using the same subchannels for 20 ms until the next ICIC update. All eNBs are assumed to send their ICIC messages at the same time. However, this scheme would work in exactly the same way, if they were not synchronised or if the frequency of the ICIC signals was lower. Every eNB always uses the last received ICIC signal from each of its neighbours, which only affects spectrum assignment decisions for new file arrivals and does not affect current file transmissions.

There are two important parameters in this scheme that have a significant influence on its performance:

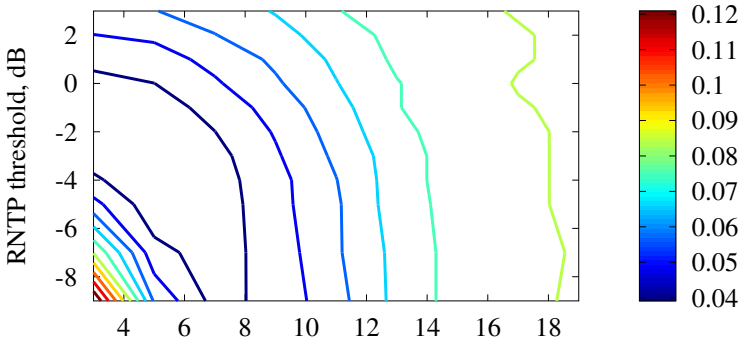
- Minimum neighbour received signal strength (MNRSS) - the minimum proximity of two eNBs in terms of the reference signal strength received from one by another that qualifies them as ICIC signalling neighbours.
- RNTP threshold - the standardised parameter for the LTE downlink used to determine whether a transmit power on a given subchannel is high enough to cause

potential inter-cell interference and whether that subchannel should be reported as busy in the ICIC message the given eNB sends to its neighbours.

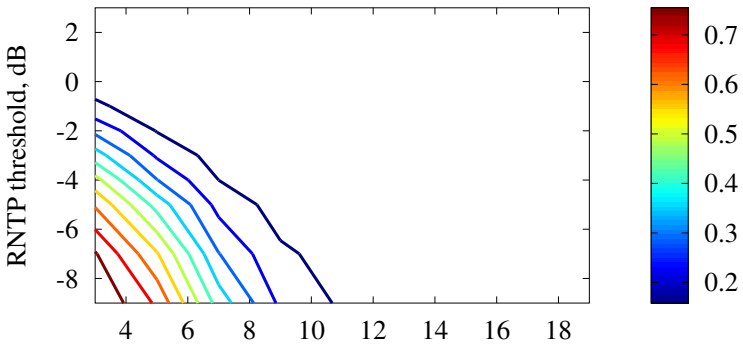
The contour plots in Figure 3.5 show the probability of retransmission at the stadium small cell network with its own dedicated spectrum introduced in Subsection 3.1.1, when it employs the dynamic ICIC scheme depicted in Figure 3.4 with a range of values for the MNRSS and the RNTP threshold. The reference signal power transmitted



(a) 0.6 Gbps offered traffic



(b) 0.9 Gbps offered traffic



(c) 1.2 Gbps offered traffic

Figure 3.5: Probability of retransmission at the stadium network employing dynamic ICIC with a range of RNTP thresholds and MNRSS levels

by every eNB for neighbour detection is assumed to be equal to the average power of data transmissions in its cell. The MNRSS is defined as the power of such signals received by the given eNB in dB relative to the receiver noise floor. If a reference signal received from another eNB is greater or equal to MNRSS, the latter eNB is deemed to be an ICIC signalling neighbour of the former. The RNTP threshold is measured in dB with respect to the average transmit power in the cell.

Figure 3.5 contains three different contour plots representing the system performance at a relatively low (0.6 Gbps), medium (0.9 Gbps) and high (1.2 Gbps) offered traffic level. Figure 3.5a demonstrates that at low traffic loads, low values of the MNRSS and the RNTP threshold achieve a better system QoS. The low values of the MNRSS mean that more eNBs are regarded as each other's neighbours, thus involving more proactive ICIC signalling for inter-cell interference avoidance. Low RNTP thresholds cause more subchannels to be reported in the ICIC signals between neighbouring eNBs, which in turn results in safer and more constrained spectrum assignment policies. However, Figure 3.5c shows a completely opposite pattern at high traffic loads. There, low values of the MNRSS and the RNTP threshold tend to cause a dramatic degradation in the system performance due to an excessive number of subchannels being marked as unavailable resulting in a large number of blocked transmissions. Figure 3.5b shows that at a medium traffic load the optimal choice for these parameters lies in a region between the very high and very low values. Therefore, all three contour plots together demonstrate that the choice of the MNRSS and the RNTP threshold affects the trade-off between the network performance at low and high traffic loads. The simulation experiments presented in the rest of this thesis that involve ICIC signalling in the stadium network use a 5dB MNRSS and the -3 dB RNTP threshold. These values are low enough to perform well at low and medium traffic loads, yet not too low to cause excessive performance degradation at higher traffic loads.

3.3.2 Spectrum Sensing

The opportunistic spectrum sensing scheme described by the flowchart in Figure 3.6 represents a typical cognitive radio approach to DSA, such as those introduced in Subsection 2.2.2. There, a cognitive eNB has the capability of sensing the interference

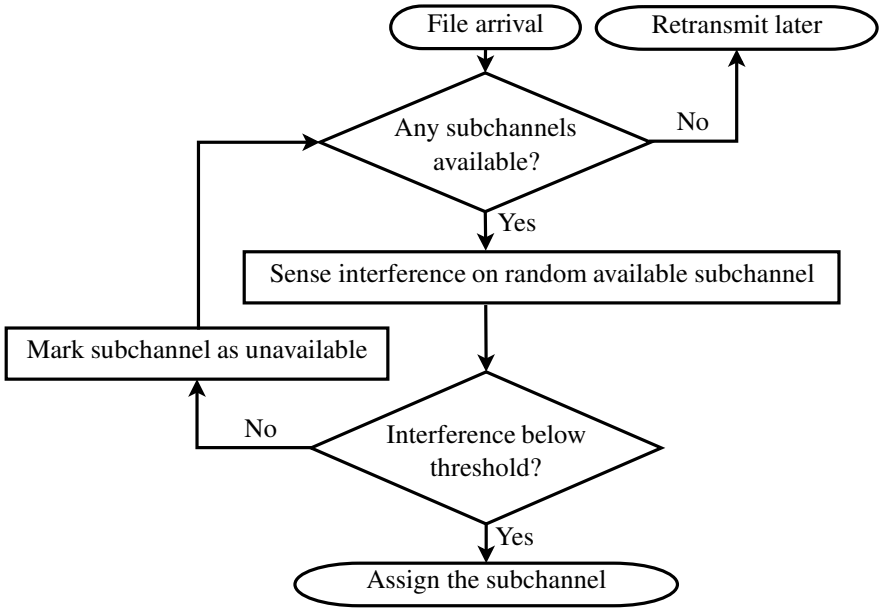


Figure 3.6: Flow diagram of the spectrum sensing based opportunistic spectrum access scheme used for baseline comparison

levels on the subchannels of interest before making spectrum assignment decisions. It chooses a subchannel at random, and senses the interference level on it. If the interference level is below an admission threshold, the subchannel is assigned, otherwise the interference level is sensed on another randomly selected subchannel.

The key parameter in this scheme is the admission threshold, i.e. the maximum amount of interference allowed on the subchannel for it to be deemed safe and eligible for assignment. Figure 3.7 shows how the probability of retransmission in the stadium network varies at different traffic loads and with different values of the interference threshold measured in dB relative to the receiver noise floor. Every data point represents the mean result of 50 simulations using identical parameters but different random seeds, with the error bars showing the minimum and maximum of the corresponding 50 values. Similarly to the dynamic ICIC parameters investigated in Subsection 3.3.1, a trade-off between the system performance at low and high traffic loads has to be achieved. The plot shows that the optimal value for the interference threshold significantly increases, as the offered traffic increases. Similarly to the MNRSS and the RNTP threshold for dynamic ICIC, low interference threshold values in spectrum sensing impose greater restrictions on subchannel selection resulting in better quality links. However, as the traffic load increases it becomes less feasible due to the increase in inter-cell interference levels and the lack of such high quality links. In those cases

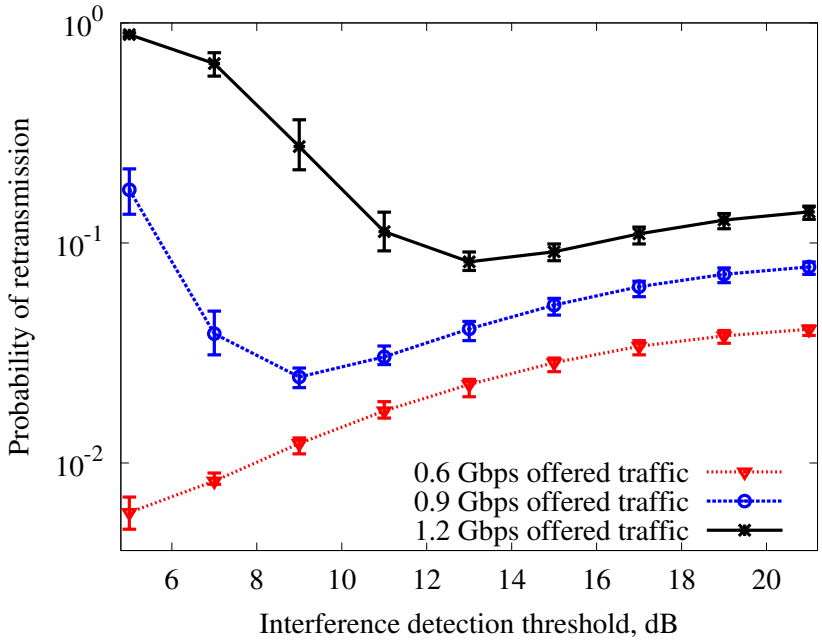


Figure 3.7: Probability of retransmission at the stadium network using the spectrum sensing based DSA scheme with different interference detection thresholds

relaxing the subchannel assignment constraints by raising the interference threshold improves the QoS. All further experiments that employ the spectrum sensing scheme depicted in Figure 3.6 for baseline comparison use an 11 dB interference threshold which is low enough to ensure good QoS at low and medium traffic loads, yet high enough not to cause excessive performance degradation at higher traffic loads.

3.4 Conclusion

This chapter described the methodology used for empirical evaluation of intelligent DSA methods proposed in this thesis. A stadium temporary event scenario, that involves a heterogeneous cognitive cellular system and an incumbent LTE network, is used as the basis for the detailed system-level simulation model of a wireless environment. The key metrics used to assess the performance of the simulated DSA algorithms are the probability of retransmission, mean and 5th percentile user throughput and the overall system throughput density. A standard LTE interference management solution and a spectrum sensing based DSA scheme, typical for CR networks, are used for baseline comparison in the simulation experiments discussed in the rest of this thesis.

Chapter 4. Distributed Q-Learning Based Dynamic Spectrum Access

Contents

4.1 Intelligent Dynamic Spectrum Access

An emerging state-of-the-art technique for intelligent DSA is reinforcement learning (RL); a machine learning technique aimed at building up solutions to decision problems only through trial-and-error, discussed in detail in Section 2.3. It has been successfully applied to a range of DSA problems and scenarios, such as cognitive radio networks [43], small cell networks [7] and cognitive wireless mesh networks [18].

The most widely used RL algorithm in both artificial intelligence and wireless communications domains is Q-learning [94]. Therefore, most of the literature on RL based

DSA focuses on Q-learning and its variations, e.g. [18][102]. Furthermore, this thesis investigates distributed Q-learning based DSA. The distributed Q-learning approach has advantages over centralised methods in that no communication overhead is incurred to achieve the learning objective, and the network operation does not rely on a single computing unit. It also allows for easier insertion and removal of base stations from the network, if necessary. For example, such flexible opportunistic protocols are well suited to disaster relief and temporary event networks, where rapidly deployable architectures with variable topologies are required to supplement any local wireless infrastructure, such as the cognitive wireless network introduced in Section 3.1.

In pure distributed RL based DSA the task of every base station (BS) is to learn to prioritise among the available subchannels only through trial-and-error, with no frequency planning involved, and with no information exchange with other BSs. In this way, frequency reuse patterns emerge autonomously using distributed artificial intelligence with no requirement for any prior knowledge of a given environment. The rest of the section revisits the main principle behind RL and introduces the distributed Q-learning algorithm used as the basis for all work presented in this thesis.

4.1.1 Reinforcement Learning

RL is a model-free type of machine learning which is aimed at establishing the desirability of taking any available action in any state of the environment only through trial-and error [87]. This desirability of an action is represented by a numerical value known as the Q-value - the expected cumulative reward for taking a particular action in a particular state, as shown in the equation below:

$$Q(s,a) = E\left[\sum_{t=0}^{T} \gamma^t r_t\right] \tag{4.1}$$

where $Q(s,a)$ is the Q-value of action a in state s, r_t is the numerical reward received t time steps after action a is taken in state s, T is the total number of time steps until the end of the learning process or episode, and $\gamma \in [0,1]$ is a discount factor.

The task of an RL algorithm is to estimate $Q(s,a)$ for every action in every state, which is then stored in an array known as the Q-table. In some cases where an environment

does not have to be represented by states, only the action space and a 1-dimensional Q-table $Q(a)$ can be considered [21]. The job of an RL algorithm then becomes simpler; it aims to estimate an expected value of a single reward for each action available to the learning agent:

$$Q(a) = E[r_t] \tag{4.2}$$

4.1.2 Distributed Stateless Q-Learning

For this reason the stateless Q-learning algorithm, formulated by Claus and Boutilier in [21], has been chosen as the RL algorithm used for DSA in this thesis. It is a stateless equivalent of the most widely used RL algorithm - Q-learning developed by Watkins in [94]. Expressing the DSA environment as a stateless problem and employing the stateless Q-learning algorithm, as opposed to its classical counterpart, can significantly simplify and speed up the learning process as discussed in Subsection 2.3.2. Figure 4.1 shows a flowchart for one file transmission of how distributed stateless Q-learning can be applied to DSA in cellular systems.

Each BS maintains a Q-table $Q(a)$ such that every subchannel a has a Q-value associated with it. Upon each file arrival, the BS either assigns a subchannel to its transmission or blocks it if all subchannels are occupied. It decides which subchannel to assign based on the current Q-table and the greedy action selection strategy described by the following equation:

$$\hat{a} = \underset{a}{\operatorname{argmax}}(Q(a)) \tag{4.3}$$

where $\hat{a}$ is the subchannel chosen for assignment, and $Q(a)$ is the Q-value of subchannel a.

The values in the Q-tables are initialised to zero, so all BSs start learning with equal choice among all available subchannels. A Q-table is updated by a BS each time it attempts to assign a subchannel to a file transmission in the form of a positive or a negative reinforcement. The recursive update equation for stateless Q-learning, as defined in [21], is given below:

$$Q(a) \leftarrow (1 - \alpha)Q(a) + \alpha r \tag{4.4}$$

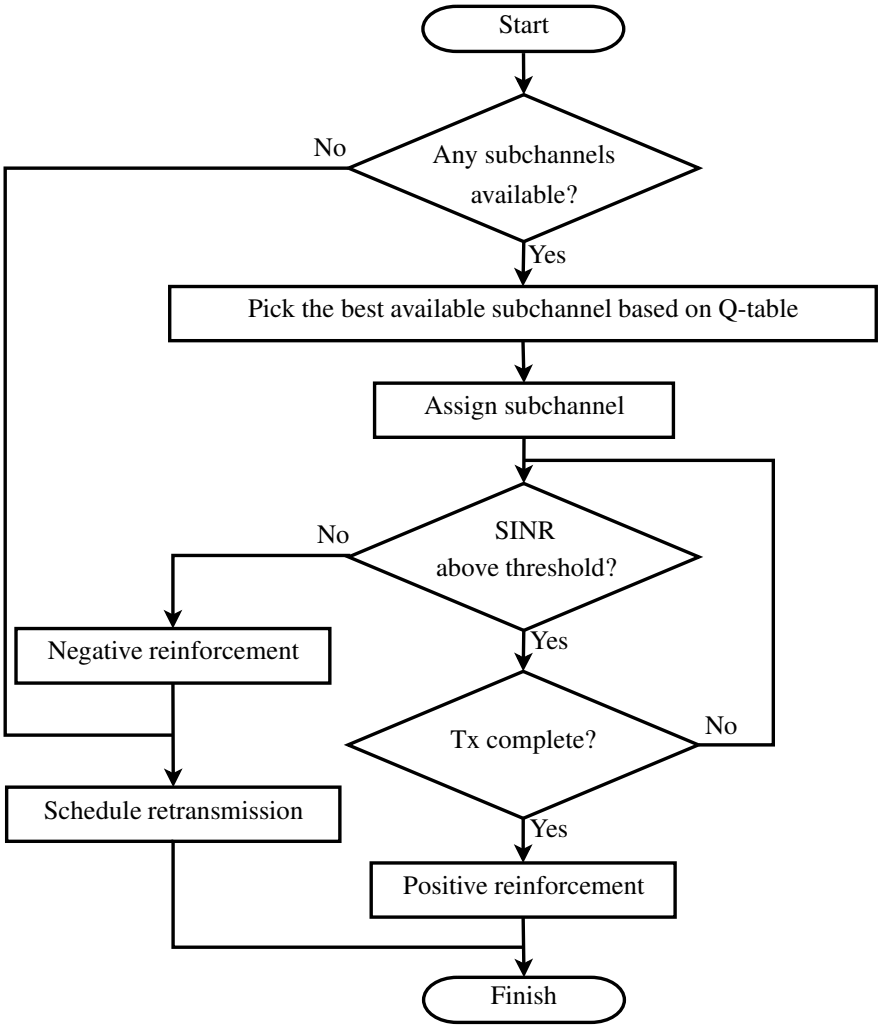


Figure 4.1: Flowchart of the distributed stateless Q-learning based DSA algorithm

where $Q(a)$ represents the Q-value of the subchannel a, r is the reward associated with the most recent trial and is determined by a reward function, and $\alpha \in [0, 1]$ is the learning rate parameter which weights recent experience with respect to previous estimates of the Q-values.

The reward function, which is generally applicable to a wide range of RL problems and which has been successfully applied to DSA problems in the past [43][104], returns two values:

- $r = -1$ (negative reinforcement), if the file transmission fails due to excessive interference on the selected subchannel.
- $r = 1$ (positive reinforcement), if the file transmission is completed using the selected subchannel.

4.2 Choice of the Learning Rate

The learning rate α is a crucial parameter in RL algorithms that can significantly influence the dynamics of the learning process. However, discussing the choice of its value is highly uncommon in the literature on RL based DSA. One of the rare examples where the value of the learning rate is at least specified is [6], where the authors have arbitrarily chosen a value of 0.5, which is simply in the middle of its allowed range of $[0, 1]$. In [49] the authors have swept all possible values of the fixed learning rate to compare different exploration strategies, but do not comment on the difference in performance due to the difference in learning rate values. The majority of other examples in DSA literature do not even specify the learning rate they have chosen, making it impossible to replicate their results.

The purpose of this section is to present the concept of the Win-or-Learn-Fast (WoLF) variable learning rate [13] from the artificial intelligence literature, show how it can be applied in the DSA context and investigate the performance improvements that can be achieved using it in terms of the QoS provided to the network users.

4.2.1 Win-or-Learn-Fast Variable Learning Rate

The WoLF principle proposed by Bowling and Veloso in [13] states that the learning agent should learn faster when it is losing and more slowly when winning. The simple adaptation of the WoLF principle proposed in this section is to split the value of the learning rate α into two cases, α_{win} and α_{lose}, when the subchannel chosen by the BS successfully supports the file transmission and when it fails (blocking or interruption) respectively. If $\alpha_{win} < \alpha_{lose}$, the WoLF principle holds, since the agent is learning slower on successful trials (α_{win}) and faster on the failed ones (α_{lose}).

One of the advantages of using a WoLF variable learning rate is that it encourages thorough exploration in the early stages of learning. Since all values in the Q-tables are initially set to zero and the greedy action selection strategy is followed, if a BS has several successful trials on a particular subchannel, its Q-value will increase and it will continue to be used. If, later on, the interference from other BSs on this subchannel

significantly increases, it will take fewer failed trials for its Q-value to fall below zero than it would if a fixed value of α was used, thus, adapting its policy faster. The rest of this subsection analytically demonstrates these learning dynamics achieved by the WoLF variable learning rate.

First, the recursive Q-table update formula from Equation (4.4) is rewritten using separate terms for the Q-value estimates before ($Q(a)$) and after the update ($Q'(a)$) as follows:

$$Q'(a) = (1 - \alpha)Q(a) + \alpha r \tag{4.5}$$

Second, splitting the learning rate value into two cases, α_{win} and α_{lose}, and substituting the reward values ($r = \pm 1$) into Equation (4.5) yields:

$$Q'(a) = \begin{cases} (1 - \alpha_{win})Q(a) + \alpha_{win}, & r = 1 \\ (1 - \alpha_{lose})Q(a) - \alpha_{lose}, & r = -1 \end{cases} \tag{4.6}$$

Third, rearranging the terms in Equation (4.6) gives the following expression for the change in Q-value $\Delta Q(a) = Q'(a) - Q(a)$:

$$\Delta Q(a) = \begin{cases} -\alpha_{win}Q(a) + \alpha_{win}, & r = 1 \\ -\alpha_{lose}Q(a) - \alpha_{lose}, & r = -1 \end{cases} \tag{4.7}$$

The magnitude of $\Delta Q(a)$ is given by the following equation:

$$|\Delta Q(a)| = \begin{cases} -\alpha_{win}Q(a) + \alpha_{win}, & r = 1 \\ \alpha_{lose}Q(a) + \alpha_{lose}, & r = -1 \end{cases} \tag{4.8}$$

since $\alpha_{win} > 0$, $\alpha_{lose} > 0$ and $Q(a) \in [-1, 1]$.

Figure 4.2 shows a plot of both cases from Equation (4.8), i.e. the linear relationship between the Q-value and the magnitude of its change when a reward of ± 1 is received by the learning agent. It demonstrates that the slope of this relationship is equal to the learning rate α. Therefore, if $\alpha_{win} < \alpha_{lose}$, the slope is higher when the agent "loses" ($r = -1$). This in turn means that most of the time the changes in the Q-values ($|\Delta Q(a)|$) are bigger when the negative rewards are received. The expression

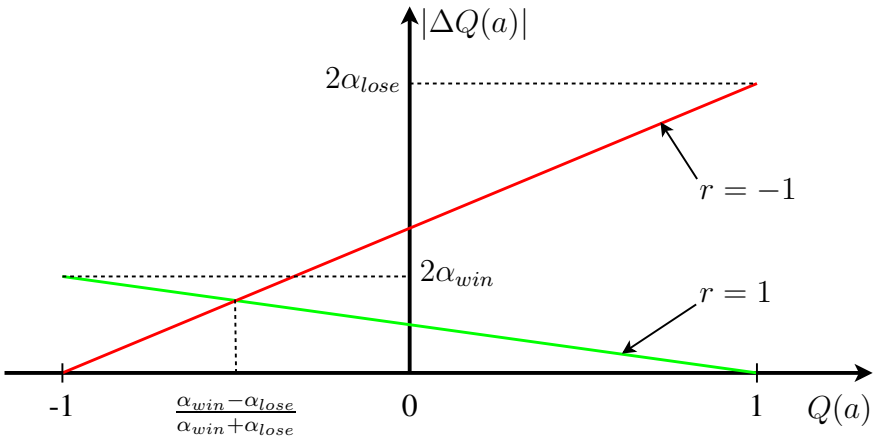


Figure 4.2: The magnitude of the change in the Q-value ($|\Delta Q(a)|$) after a Q-learning update using the WoLF variable learning rate ($\alpha_{win} < \alpha_{lose}$)

for the Q-value at which they are the same for the negative and the positive rewards, i.e. the crossing point between the linear functions plotted in Figure 4.2, can be found by equating the two cases of Equation (4.8) as follows:

$$-\alpha_{win}Q(a) + \alpha_{win} = \alpha_{lose}Q(a) + \alpha_{lose} \tag{4.9}$$

Solving for $Q(a)$ in Equation (4.9) yields the equilibrium value $Q_{eq}(a)$ where $|\Delta Q(a)|$ is the same in both cases:

$$Q_{eq}(a) = \frac{\alpha_{win} - \alpha_{lose}}{\alpha_{win} + \alpha_{lose}} \tag{4.10}$$

Equation (4.10) demonstrates that if $\alpha_{win} < \alpha_{lose}$, $Q_{eq}(a)$ is negative, i.e. the changes in the Q-values are larger when negative rewards are received at the $Q(a) = 0$ point. For example, if a regular learning rate was used instead of WoLF, the slopes of the two linear functions in Figure 4.2 would be the same and the equilibrium point $Q_{eq}(a)$ would be zero. The larger the difference between α_{win} and α_{lose} is, the lower $Q_{eq}(a)$ is and the larger the difference between the Q-value changes for positive and negative rewards is around the $Q(a) = 0$ point. The latter feature of the WoLF variable learning rate is key to avoiding rapid convergence towards local optima at the start of the learning process, since the BSs learn more slowly and "cautiously" from successful trials and faster from the failed trials.

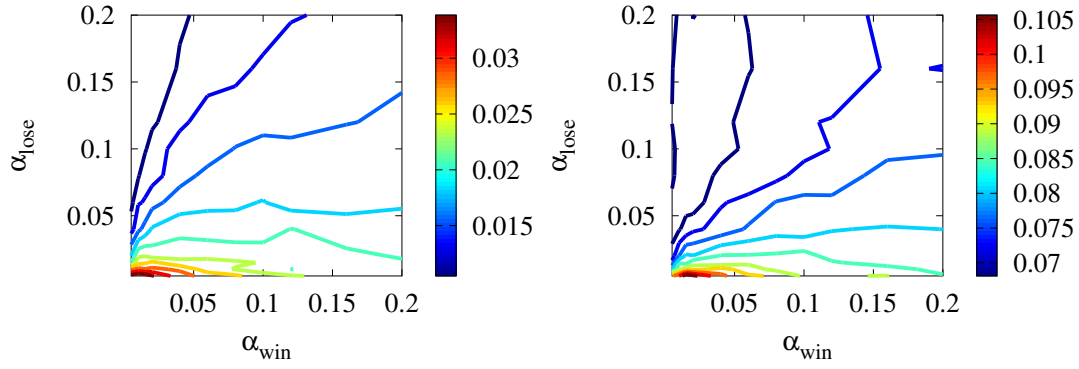
The principle of learning faster when "losing" is also relevant in dynamic learning environments, e.g. when a change in network topology or traffic distribution requires

the BSs to change and adapt their learned policies. In such cases a BS would start exploring other subchannels sooner. Another advantage of the WoLF learning rate is that at any stage of the operation of the network the ratio of successful to failed trials would need to be higher for a subchannel to maintain a high Q-value and keep being assigned, which is consistent with the goal of achieving a low probability of retransmission in a wireless network.

4.2.2 Performance Comparison Using Different Learning Rates

The simulation scenario of a stadium small cell network with its own dedicated spectrum described in Subsection 3.1.1 is used in the rest of this section to test the QoS provided to the UEs, using different combinations of the values of α_{win} and α_{lose}. 25% of the overall stadium capacity is randomly filled with wireless subscribers, i.e. on average 10,776 randomly distributed UEs.

The contour plots in Figure 4.3 show the probability of retransmission results after running the simulations of the distributed Q-learning based DSA algorithm described in Subsection 4.1.2, using different combinations of α_{win} and α_{lose}. The simulations were performed at a relatively low traffic load of 0.7 Gbps and a higher traffic load of 1.2 Gbps. They lasted 1,000,000 transmissions, which constituted 1,000,000 reinforcement learning trials for all eNBs in total. The values of α_{win} and α_{lose} vary within $[0.005, 0.2]$ which covers a range between a very low and a relatively high learning rate.



(a) 0.7 Gbps offered traffic

(b) 1.2 Gbps offered traffic

Figure 4.3: Probability of retransmission using different combinations of learning rates α_{win} and α_{lose}

Both plots demonstrate significant performance improvements when the WoLF variable learning rate principle is used, i.e. the region above and to the left of the 45^o diagonal, where $\alpha_{win} < \alpha_{lose}$. The fixed learning rate values lie on the 45^o diagonal (where $\alpha_{win} = \alpha_{lose}$), and perform noticeably worse than those in the "WoLF region". The plots also show that the further in the WoLF region the combination of α_{win} and α_{lose} are from the 45^o diagonal, the better the system performs within the range of investigated learning rate values. The opposite strategy of setting the learning rate for positive rewards higher than that for the negative rewards, i.e. below and to the right of the 45^o diagonal is shown to perform even poorer than the regular fixed learning rate values. All of this empirical evidence depicted in Figure 4.3 supports the hypothesis that the WoLF strategy for selecting the learning rate values is the best choice for distributed RL based DSA.

4.2.3 Temporal Performance

Figure 4.4 shows the difference in the average QoS time response (i.e. how QoS improves over time) of the distributed Q-learning based DSA algorithm with a typical choice of the fixed learning rate value of 0.1 [102], and the WoLF variable learning rate of $\{0.01, 0.1\}$. Every data point on the graph is the mean of the corresponding data points from 50 different simulations with different random seeds and UE locations. The offered traffic is 1 Gbps.

At the early stages of learning, the WoLF learning rate achieves better QoS due to its increased adaptability to changes in the policies of all eNBs, which are in turn affecting the learning process of every individual eNB. Furthermore, after 1,000,000 transmissions, the QoS achieved using the WoLF learning rate is still significantly better, which suggests that fixed learning rates tend to cause the Q-learning algorithm to converge towards poorer solutions, compared to the WoLF variable learning rates. These results confirm the analytical prediction of the WoLF learning rate achieving superior learning process dynamics discussed in Subsection 4.2.1.

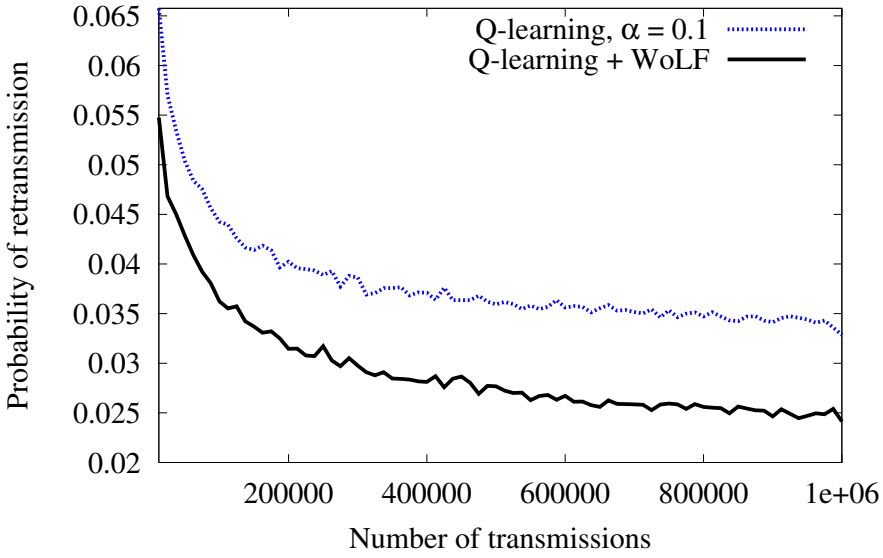


Figure 4.4: Average probability of retransmission temporal response at 1 Gbps offered traffic using the distributed Q-learning based DSA scheme with and without the WoLF variable learning rate

4.2.4 Comparison with Heuristic Schemes

In Figure 4.5 the probabilities of retransmission using the fixed learning rate of 0.1 and the WoLF learning rate $\{0.01, 0.1\}$ are shown across a wide range of traffic loads. It also compares the results with the performance of the following two baseline heuristic schemes:

- a standard LTE dynamic ICIC scheme described in Subsection 3.3.1,
- an opportunistic spectrum sensing based scheme described in Subsection 3.3.2.

The overall simulation length is 1,000,000 file transmissions. Every data point represents the mean result of 50 different simulations at a given traffic load with the error bars showing the minimum and maximum of the corresponding 50 values.

Figure 4.5 shows that the Q-learning based schemes outperform both baseline heuristic schemes at the whole range of traffic loads, demonstrating the effectiveness of the application of RL to DSA in cellular systems. It also shows that by simply changing the fixed learning rate of the Q-learning algorithm ($\alpha = 0.1$) to a WoLF variable learning rate of $\{0.01, 0.1\}$, a 20-41% reduction in the probability of retransmission is achieved at the lower half of the traffic loads (below 1.04 Gbps). There is no notable difference in network performance introduced by the WoLF learning rate at higher

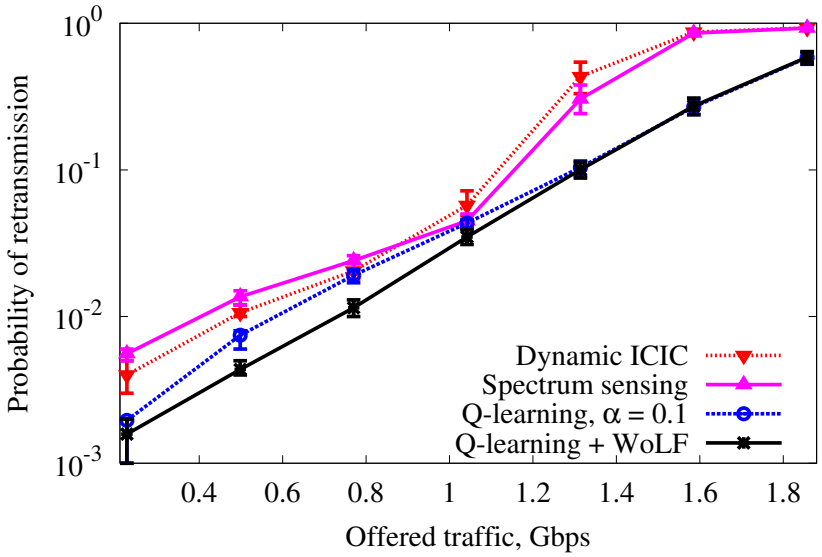


Figure 4.5: Average probability of retransmission at a range of traffic loads using different intelligent and heuristic DSA algorithms

traffic loads, where the probability of retransmission is above ≈10%.

Although no heuristic information obtained through spectrum sensing or ICIC signalling is involved in the "Q-learning + WoLF" approach, it shows a consistent significant improvement in QoS over the purely heuristic schemes. The only disadvantage of the Q-learning approach is the initial learning period, where the QoS starts at a relatively poor level due to the lack of information in the Q-tables and it takes the eNBs time to learn mature DSA policies, as seen in the time responses in Figure 4.4. This issue is addressed in the later chapters of this thesis.

4.3 Q-Learning Based Dynamic Spectrum Sharing

The simulation experiments discussed in this section assess the performance of the distributed stateless Q-learning based DSA algorithm with the WoLF variable learning rate $\alpha \in \{0.01, 0.1\}$ in a dynamic spectrum sharing scenario. There, the additional feature introduced in Subsection 3.1.1 is the presence of a local primary LTE network operating in the suburban area around the stadium. The stadium small cell network has secondary access to the 20 MHz LTE channel used by the primary system. Therefore, the task of the stateless Q-learning based DSA scheme implemented in the secondary system is to learn appropriate spectrum management policies which provide adequate

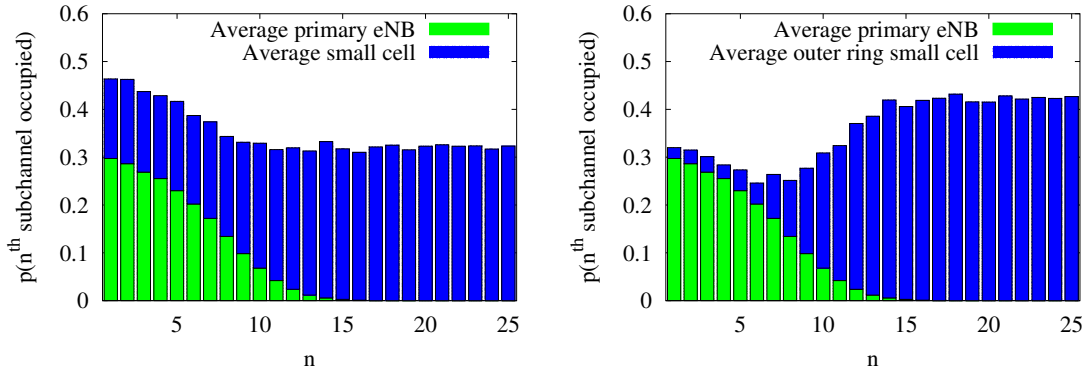
QoS to the secondary users, and which also avoid harmful interference for the primary system.

The primary system is assumed to employ a dynamic ICIC scheme such as that described in Subsection 3.3.1 for the stadium network, where all three eNBs exchange their current spectrum usage as ICIC messages every 20 ms, and exclude the subchannels currently used by the other two eNBs from their available subchannel list. The primary eNBs (PeNBs) always try to assign an available subchannel with the lowest index if any, e.g. they always scan the availability of the subchannels in the same order from the 1st subchannel to the last. In this way, the primary network would make its spectrum usage less random and more appropriate for the cognitive stadium small cell network to share, which is in the interests of both the primary and the secondary system. However, the distributed Q-learning scheme investigated in this chapter does not assume this and would also work regardless of the spectrum management strategy of the primary system.

4.3.1 Spectrum Occupancy Analysis

Figure 4.6 shows the spectrum occupancy patterns that emerge autonomously in the stadium small cell network through distributed machine intelligence afforded by the distributed Q-learning approach, in response to a specific spectrum occupancy pattern used by the local primary LTE network. The simulation lasted a total of 2,000,000 transmissions. The offered traffic in the primary system outside the stadium is 20 Mbps, and 1 Gbps in the stadium small cell network.

Figure 4.6b demonstrates that the outer ring of small cell eNBs depicted in Figure 3.2, which is most vulnerable to interference from the external primary system, has learnt to largely avoid parts of the spectrum most heavily used by the PeNBs. In contrast, most other stadium eNBs have suffered significantly less from the primary system interference on those subchannels, and thus learned to fully reuse them without many negative reinforcements, i.e. blocked/interrupted transmissions. Therefore, the average small cell eNB subchannel occupancy shown in Figure 4.6a is far more evenly distributed. These results demonstrate the efficiency of such an autonomous RL approach, where



(a) Average stadium eNB (b) Average outer ring stadium eNB

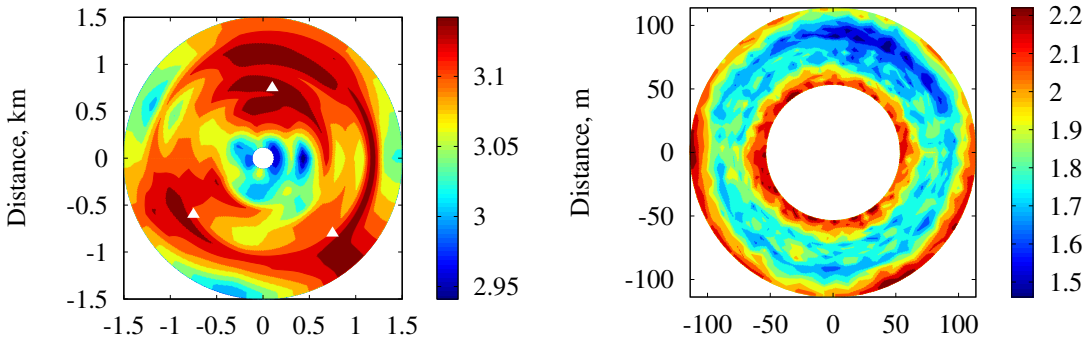
Figure 4.6: Subchannel occupancy of the primary eNBs and the stadium small cells that employ the stateless Q-learning based DSA algorithm

no coordination or spectrum planning is required.

4.3.2 Spatial Distribution of User Throughput

An essential requirement for secondary cognitive cellular systems is to ensure that they do not have a harmful effect on the QoS in the primary system. The contour plots in Figure 4.7 show the spatial distribution of user throughput (UT), i.e. data rates experienced by the primary and the secondary users, achieved by the autonomously emerging spectrum sharing patterns shown in Figure 4.6.

Figure 4.7a shows that the primary user UT varies insignificantly, 2.95-3.15 Mb/s, whilst Figure 4.7b shows that at the same time an adequate QoS (≈1.5-2.2 Mb/s UT)



(a) Primary user UT, Mb/s (b) Secondary user UT, Mb/s

Figure 4.7: Spatial distribution of user throughput (UT) outside (primary users) and inside the stadium (secondary users)

is provided to the ultra-dense population of secondary users. As expected the primary UEs closer to the PeNBs have a higher UT due to the higher quality of the links provided to them in terms of the SINR, whereas the UEs at the cell edge have a slightly lower SINR and UT. Although the primary UEs in the close vicinity of the stadium are most vulnerable to the secondary system interference, the degradation in the UT provided to them is negligible.

The top-right part of the stadium small cell network achieves a poorer QoS than elsewhere due to a higher amount of primary system interference received from the PeNB closest to the stadium. Its location is marked by the white triangle north of the stadium in Figure 4.7a, i.e. north of the (0, 0) coordinate. The stadium network QoS is also visibly better at the edges of the spectator area due to the reduced amount of interference from other small cell eNBs. This is because the users located closer to the middle of the spectator area receive interference from the eNBs in both radial directions, whereas the users located at the edge do not receive inter-cell interference from the areas outside of the doughnut-shaped stadium network. No difference between the QoS at the centre and at the edge of the small cells is observed due to the open-loop power control scheme described in Subsection 3.1.5 that provides the same SNR to both cell-centre and cell-edge UEs.

4.3.3 Primary and Secondary User Quality of Service

The spectrum occupancy and spatial QoS distribution results described in Figures 4.6 and 4.7 show that the secondary stadium small cell network successfully adapts to the spectrum usage of the primary system to minimise the harmful effect of interference from the latter on the former. They also show that the effects of interference from the secondary system on the QoS provided to the primary users are negligible. However, that simulation experiment only considers specific traffic loads outside and inside the stadium, i.e. in the secondary and primary system respectively. The contour plots in Figure 4.8 show the capacity and QoS of the stadium small cell network at a range of primary and secondary system offered traffic values.

Figures 4.8a, 4.8b and 4.8c show that the secondary stadium network is negatively

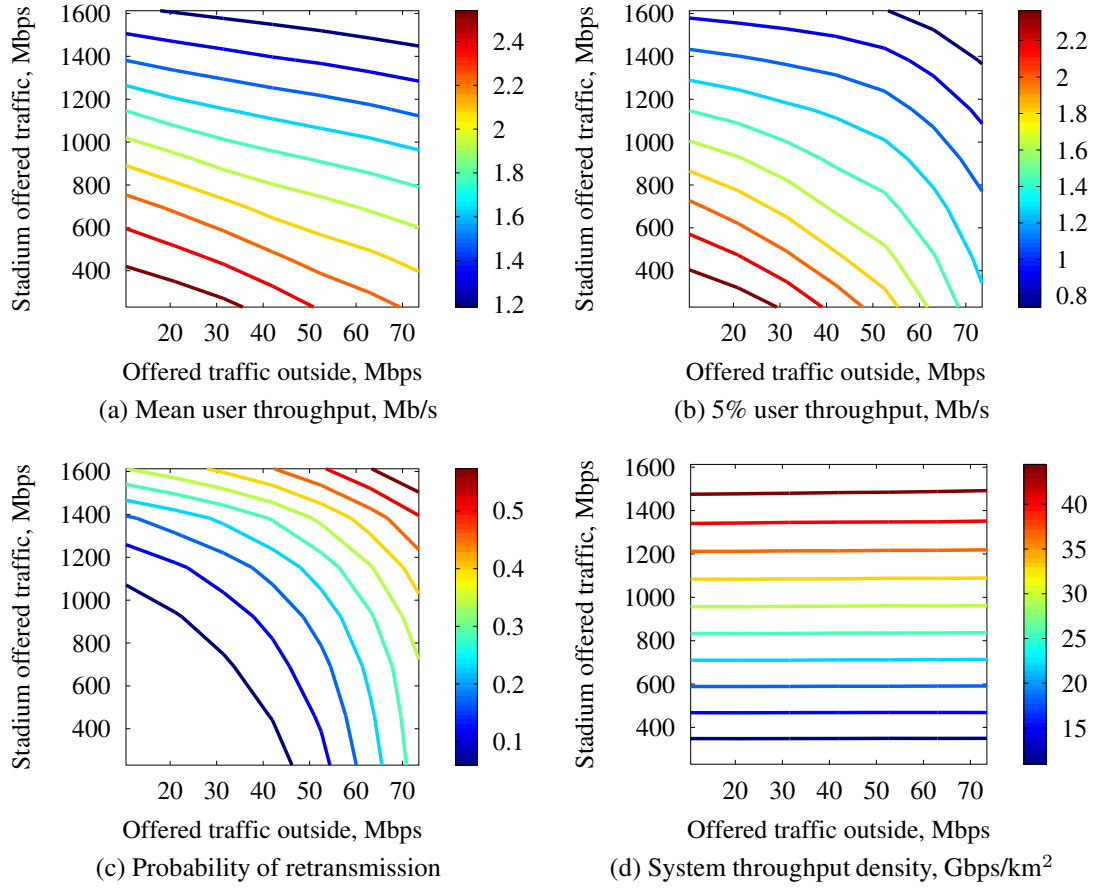


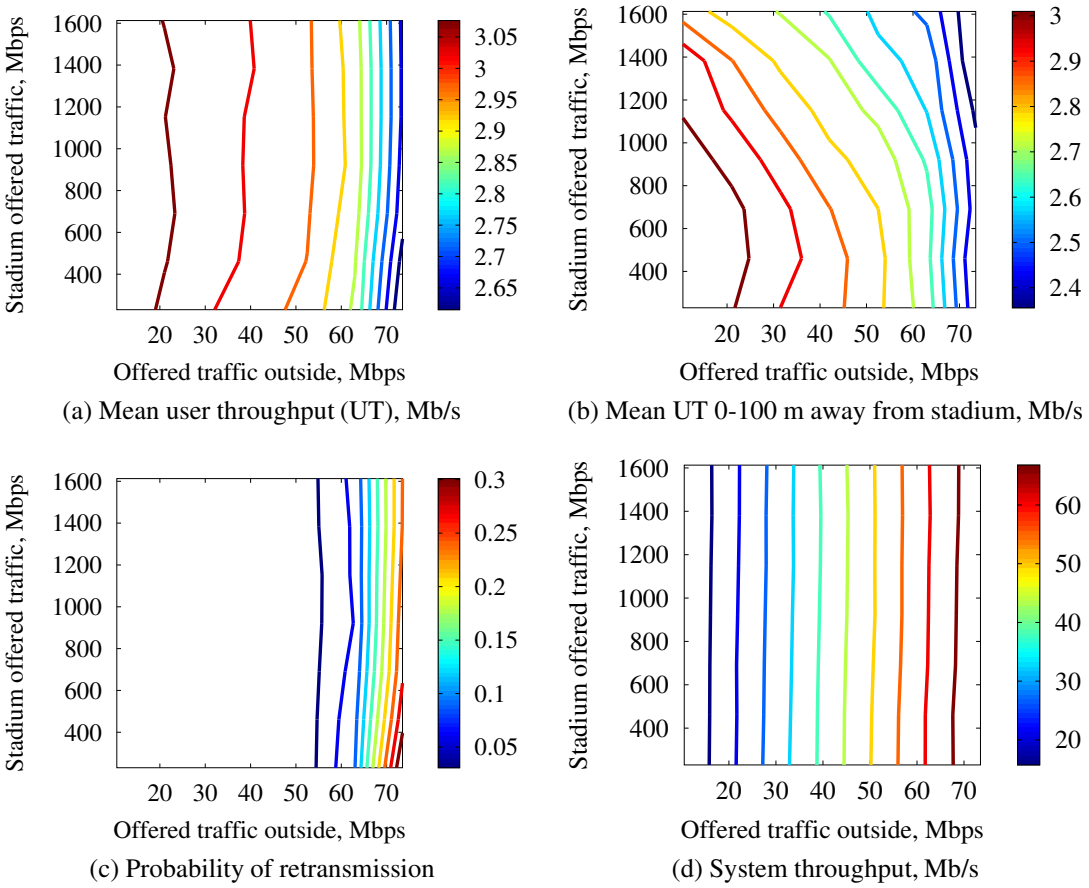
Figure 4.8: Capacity and quality of service in the secondary stadium network at a range of primary and secondary system traffic loads

affected by the interference from the high power local primary system, since its QoS deteriorates both as its own offered traffic increases and as the primary system offered traffic increases (in the horizontal direction on the contour plots). However, Figure 4.8d shows that the capacity of the stadium small cell network is independent of the wide range of primary system traffic load variations investigated in these simulation experiments. This demonstrates that the distributed Q-learning based DSA algorithm investigated in this chapter is able to utilise and reuse the spectrum highly efficiently, even at high primary system traffic loads, where all of it is actively used by the primary system.

A critical requirement for successful coexistence between a primary network and a secondary cognitive cellular system is eliminating harmful effects of the secondary system interference on the primary user QoS. Figure 4.9 contains the same type of 2-dimensional offered traffic sweeps as those in Figure 4.8, but which show the QoS

and system throughput of the primary network outside of the stadium.

Figure 4.9a shows that the overall mean UT of the primary system is independent of the offered traffic variations in the stadium small cell network; therefore, it is not affected by the secondary system interference. Similarly the network-wide probability of retransmission and the overall system throughput shown in Figures 4.9c and 4.9d respectively are unaffected by the secondary system interference. Figure 4.9b shows the mean UT in the area 0-100 m away from the stadium boundary, i.e. the area most vulnerable to interference from the densely populated stadium small cell network as shown in Figure 4.7a. In this case, the contour plot shows that there is indeed a deterioration in the primary user QoS in this area due to an increase in the secondary system traffic load. The maximum decrease in the mean UT of these primary UEs due to a full-scale increase in the secondary system offered traffic from 0.23 to 1.6 Gbps is 8.9%. However, the QoS of the secondary system shown in Figure 4.8 at such a high



(a) Mean user throughput (UT), Mb/s

(b) Mean UT 0-100 m away from stadium, Mb/s

(c) Probability of retransmission

(d) System throughput, Mb/s

Figure 4.9: Capacity and quality of service in the primary network at a range of primary and secondary system traffic loads

traffic load is extremely low, so it would not be feasible for it to support it anyway. At a more feasible stadium offered traffic load of ≈1 Gb/s the maximum deterioration in the mean UT of the primary users in the 100 m vicinity of the stadium is only 2.8% which is significantly more negligible.

4.4 Conclusion

In this chapter the concepts of distributed stateless Q-learning and the Win-or-Learn-Fast (WoLF) variable learning rate principle were introduced. The simulation results empirically demonstrated that it is possible to achieve significant QoS performance improvements and to increase the adaptability of the distributed Q-learning based DSA algorithm simply by choosing an appropriate WoLF learning rate. This machine intelligence based approach was also shown to outperform an opportunistic spectrum sensing scheme and a dynamic ICIC scheme typical for LTE, but with no spectrum sensing or ICIC signalling involved.

In addition, the distributed stateless Q-learning approach to DSA was shown to be effective in a dynamic secondary spectrum sharing scenario, where a stadium small cell network has only secondary access to an LTE channel used by a local primary network. The cognitive stadium network employs the distributed Q-learning algorithm to learn appropriate spectrum management policies that adapt to a specific primary system spectrum usage pattern. It is shown to provide adequate QoS to the secondary users at a wide range of traffic loads up to 1 Gb/s and to support high system throughput densities, whilst having a negligible effect on the primary user QoS with no coordination or spectrum planning involved.

Chapter 5. Bayesian Network Based Convergence Analysis

Contents

5.1 Motivation

An important step in designing RL algorithms not only for DSA applications, but also for any other type of learning problems, is to perform theoretical analysis of their convergence. There is a large amount of previous work on probabilistic analysis of RL algorithms applied to wireless communications problems, where the researchers have stochastically modelled the RL problems to derive their optimal solutions and compare them with the solutions obtained through learning. For example, Pandana and Liu [67] model the problem of average throughput maximisation per total consumed energy in a wireless sensor network as an MDP, derive an optimal solution analytically, and compare it with ones achieved by an RL algorithm. In another example Song and Jamalipour [83] model a vertical hand-off decision problem as a semi-MDP and use Q-learning to solve this model directly. However, none of the stochastic models proposed in the wireless communications domain provide insight into the dynamics of the RL algorithms themselves, as opposed to the learning problems they are applied to.

The purpose of this chapter is to propose a simple Bayesian network model for analysing convergence properties of distributed RL based DSA algorithms such as stateless Q-learning introduced in Chapter 4. This model is based on a minimum complexity 2 base station (BS) 2 user equipment (UE) inter-cell interference problem, and provides a platform for theoretical evaluation of RL algorithms before they are applied to complex real-world DSA problems. In previous work on combining Bayesian networks and RL, the purpose of Bayesian networks was to enhance the performance of RL algorithms by being used as a framework for reasoning under uncertainty, e.g. [48][68]. There appears to be no evidence in the literature of using Bayesian networks as an analysis tool for RL algorithms.

5.2 Simple Inter-Cell Interference Model

In DSA networks all BSs are allowed opportunistic access to the whole spectrum pool available to the network. The main limiting factor for network throughput and QoS performance in DSA networks is inter-cell interference, since all cells are allowed to use the same spectrum. This section presents a simple network model used for theoretical analysis of inter-cell interference.

Figure 5.1 shows a small and analytically tractable DSA network model which can be related to most inter-cell interference problems in general. The aim of this model is to provide a small yet sufficiently complex DSA problem for theoretical analysis of RL algorithms which can then be extrapolated to larger and more realistic scenarios.

The network consists of two BSs and two UEs, each connected to its own BS. If one of the UEs is located within the interference range of the other BS, it suffers from harmful co-channel interference from it. The network is assumed to be allocated 2 subchannels, and the task of both BSs is to learn to use their own subchannel through

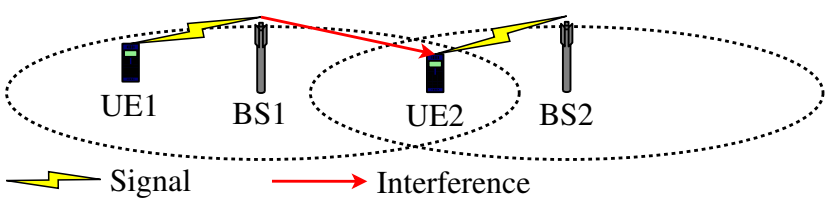


Figure 5.1: 2 base station 2 user equipment network model

distributed machine intelligence.

5.3 Bayesian Network Model

Figure 5.2 presents the Bayesian network which describes the behaviour of the distributed Q-learning algorithm introduced in Subsection 4.1.2 when applied to the simple DSA network model shown in Figure 5.1.

The variables used to denote the Bayesian network nodes are the following:

- $\Pi_n \in \{Same, Diff\}$ - the joint policy of the BSs after n learning iterations. The individual policy of one BS is defined as its preferred subchannel $\pi_x \in \{1, 2\}$ and is derived from the Q-table based on Equation (4.3). The joint policy Π_n takes two values of interest - whether the individual policies of 2 BSs are the same or different ($\Pi_n = Diff$ is the learning objective).
- $I_{UEx} \in \{Yes, No\}$ - whether or not $UE1$ or $UE2$ is located within the interference range of the adjacent BS during the current file arrival.
- $TxOL \in \{Yes, No\}$ - whether file transmissions to $UE1$ and $UE2$ overlap in time during the current iteration.
- $R_{UEx} \in \{S, F\}$ - whether a file transmission to $UE1$ or $UE2$ was successful (S), or whether it failed (F) due to interference. It is conditionally dependent on Π_n, I_{UEx} and $TxOL$.
- $\Pi_{n+1} \in \{Same, Diff\}$ - the joint policy after the Q-learning updates are per-

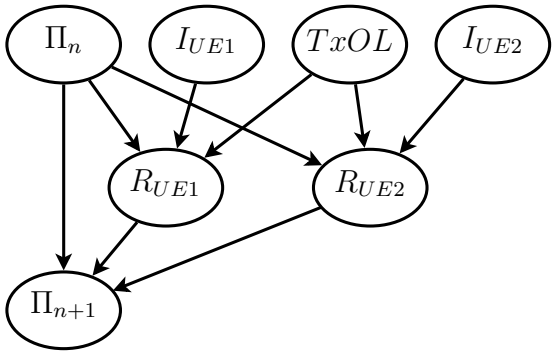


Figure 5.2: Bayesian network describing the behaviour of distributed Q-learning

formed based on Equation (4.4), as a result of the outcome at the current iteration. It is conditionally dependent on Π_n, R_{UE1} and R_{UE2}.

Based on the conditional dependencies described above and depicted in Figure 5.2, the equation for calculating the joint probability distribution over all variables $P_{joint} = P(\Pi_{n+1}, \Pi_n, R_{UE1}, R_{UE2}, I_{UE1}, I_{UE2}, TxOL)$ is the following:

$$\begin{aligned} P_{joint} = \; & P(\Pi_{n+1}|\Pi_n, R_{UE1}, R_{UE2}) \\ & \times P(R_{UE1}|\Pi_n, I_{UE1}, TxOL)\; P(R_{UE2}|\Pi_n, I_{UE2}, TxOL) \\ & \times P(\Pi_n)\; P(I_{UE1})\; P(I_{UE2})\; P(TxOL) \end{aligned} \tag{5.1}$$

which consists of a number of prior probabilities of the form $P(X)$, and conditional probabilities of the form $P(X|Y_1...Y_n)$.

5.3.1 Prior and Conditional Probability Distributions

The prior probability distributions that appropriately describe the given 2 BS 2 UE scenario are defined in Table 5.1. Before any file arrivals at either BS, the Q-tables of both BSs are initialised to zero for both subchannels. Therefore, there is a 50% chance of the BSs choosing the same subchannel, since both of them will choose either subchannel at random, i.e. $P(\Pi_0 = Same) = 0.5$. Furthermore, it is assumed that the interference range overlap of the BSs is such that there is a 40% chance of a UE being located in it, i.e. $P(I_{UEx} = Yes) = 0.4$. Finally, the offered traffic level is assumed to produce a 60% chance of transmissions to both UEs overlapping in time at any given learning iteration, thus potentially resulting in inter-cell interference: $P(TxOL = Yes) = 0.6$. The values chosen for $P(I_{UEx})$ and $P(TxOL)$ only affect the relative difficulty of the DSA problem. They can be changed without the loss of generality of the proposed probabilistic model.

Table 5.1: Prior probability distributions used in the Bayesian network model of distributed stateless Q-learning

$P(\Pi_0)$		$P(I_{UEx})$		$P(TxOL)$	
$Same$	$Diff$	Yes	No	Yes	No
0.5	0.5	0.4	0.6	0.6	0.4

The conditional probability distributions are defined in Table 5.2. The values used for the $P(R_{UEx}|\Pi_n, I_{UEx}, TxOL)$ distribution state that a transmission to $UE1$ or $UE2$ will fail with a probability of 1 ($R_{UEx} = F$) only if the given UE is within the interference range of the other BS ($I_{UEx} = Yes$), transmissions to both UEs overlap in time ($TxOL = Yes$) and both BSs have chosen the same subchannel ($\Pi_n = Same$). Whereas, in any other case, i.e. if $\Pi_n = Diff$, $I_{UEx} = No$ or $TxOL = No$, the transmission will be successful: $R_{UEx} = S$.

The $P(\Pi_{n+1}|\Pi_n, R_{UE1}, R_{UE2})$ table defines how the Q-learning policies of both BSs (Π_{n+1}) are likely to change, given their current joint policy Π_n, and the result of transmissions to both UEs (R_{UE1} and R_{UE2}). Both BSs are running a stateless Q-learning algorithm introduced in Subsection 4.1.2. Firstly, if the transmissions to both UEs are successful ($R_{UE1} = R_{UE2} = S$), then both BSs will reward their respective subchannels and maintain the same policies regardless whether they are the same or different ($\Pi_{n+1} = \Pi_n$). Secondly, if $\Pi_n = Same$ and only a transmission to one of the UEs failed ($\{S, F\}$ or $\{F, S\}$), this UE is more likely to change its policy due to the WoLF learning rate used in its Q-learning algorithm, described in Subsection 4.2.1. Therefore, there is a relatively high probability of the policies being different

Table 5.2: Conditional probability distributions used in the Bayesian network model of distributed stateless Q-learning

$P(R_{UEx}\|\Pi_n, I_{UEx}, TxOL)$								
S	0	1	1	1	1	1	1	1
F	1	0	0	0	0	0	0	0
	$Same$	$Same$	$Same$	$Same$	$Diff$	$Diff$	$Diff$	$Diff$
	Yes	Yes	No	No	Yes	Yes	No	No
	Yes	No	Yes	No	Yes	No	Yes	No
	$\Pi_n, I_{UEx}, TxOL$							

$P(\Pi_{n+1}\|\Pi_n, R_{UE1}, R_{UE2})$					
$Same$	1	Low	Low	$High$	0
$Diff$	0	$High$	$High$	Low	1
	$Same$	$Same$	$Same$	$Same$	$Diff$
	S, S	S, F	F, S	F, F	S, S
	Π_n, R_{UE1}, R_{UE2}				

at the next iteration: $P(\Pi_{n+1} = Diff) = High$. If transmissions to both UEs fail ($\{F, F\}$), both BSs are likely to change their policies to the same other subchannel, thus making $\Pi_{n+1} = Same$ a more likely outcome: $P(\Pi_{n+1} = Same) = High$. The remaining three combinations of Π_n, R_{UE1} and R_{UE2} values are not considered, since they can never occur according to the $P(R_{UEx}|\Pi_n, I_{UEx}, TxOL)$ conditional probability distribution. Regardless of the values used for these combinations in the $P(\Pi_{n+1}|\Pi_n, R_{UE1}, R_{UE2})$ table, they will be multiplied by zero during the calculation of the joint probability distribution defined in Equation (5.1).

5.3.2 Bayesian Network Inference

The aim of the Bayesian network model described above is to establish the marginal likelihood of the joint Q-learning policy at the next iteration $P(\Pi_{n+1})$ by taking a sum over all other variables in P_{joint} as follows:

$$P(\Pi_{n+1}) = \sum_{\Pi_n} \sum_{R_{UE1}} \sum_{R_{UE2}} \sum_{I_{UE1}} \sum_{I_{UE2}} \sum_{TxOL} P_{joint} \tag{5.2}$$

The resulting distribution can then be substituted as the prior for the next learning iteration: $P(\Pi_n) \leftarrow P(\Pi_{n+1})$. This enables iterative evaluation of the Bayesian network model which shows how the probability of transmission failure $P(R_{UEx})$ and the probability of BSs using different subchannels $P(\Pi_n)$ change over time, as the learning process progresses. The individual $P(R_{UEx})$ distribution can be obtained using the same principle of marginalisation as follows:

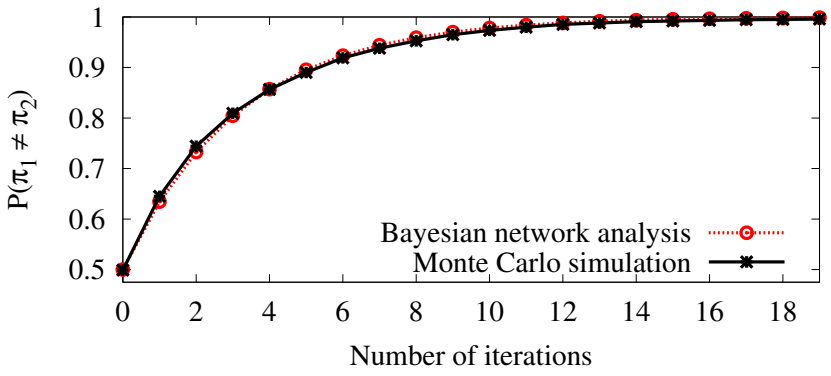
$$P(R_{UE1/2}) = \sum_{\Pi_{n+1}} \sum_{\Pi_n} \sum_{R_{UE2/1}} \sum_{I_{UE1}} \sum_{I_{UE2}} \sum_{TxOL} P_{joint} \tag{5.3}$$

This probabilistic analysis is only valid for the 2 BS 2 UE network model described in Section 5.2, and is not designed to be scalable to larger and more realistic networks. The purpose of this model is to enable theoretical analysis of the relative behaviour of RL algorithms using a simple and tractable problem. An additional, useful approach to evaluating such algorithms used in Chapter 6 is performing realistic large scale simulations and assessing similarities between the simulation results and the theoretical

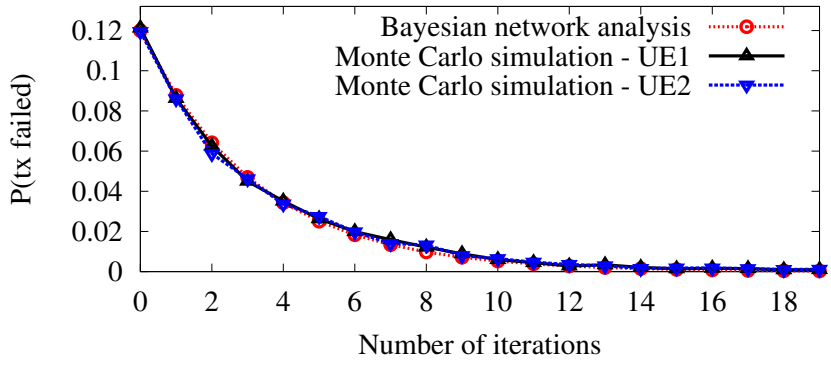
predictions obtained via the method proposed in this chapter.

5.4 Probabilistic Analysis vs Monte Carlo Simulation

Figure 5.3 shows the expected convergence behaviour of distributed Q-learning analytically derived through iterative evaluation of the Bayesian network model developed in this chapter. The values for $High$ and Low in the conditional probability distributions in Table 5.2 are assumed to be $\{0.9, 0.1\}$ without the loss of generality. The analytical results are compared with a Monte Carlo simulation, where the Q-learning algorithm from Subsection 4.1.2 is applied to the 2 BS 2 UE scenario described in Section 5.2. At every transmission arrival the simulation experiment randomly decided whether each UE is within the range of an interfering BS, and whether the transmissions to both UEs overlap in time according to the prior probability distributions defined in Table



(a) Probability of BSs having different policies



(b) Probability of a UE being blocked or interrupted

Figure 5.3: Convergence of distributed Q-learning using Bayesian network analysis and a Monte Carlo simulation

5.1. The probabilities plotted for every learning iteration, i.e. every time step, were obtained by averaging over 10,000 independent runs.

The comparison of the convergence behaviour predicted by the Bayesian network model and that achieved by the Monte Carlo simulation demonstrates remarkable accuracy of the joint policy transition analysis tool proposed in this chapter. Therefore, it is seen to be a valid and effective approach for stochastic modelling of RL based DSA algorithms. It can be used for designing and analysing the convergence and adaptability of more sophisticated RL algorithms by adding nodes and edges to the Bayesian network from Figure 5.2. The added nodes and edges would represent additional functionality and conditional dependencies introduced by the new schemes. This approach would clearly demonstrate in what ways other schemes designed in future using this method extend the basic distributed RL approach depicted in Figure 5.2. For example, this methodology is used for the theoretical analysis of the DSA algorithm proposed in Chapter 6.

5.5 Absorbing Markov Chain Formulation

Figure 5.4 shows an alternative formulation of the convergence properties of distributed Q-learning derived from the Bayesian network model introduced in Section 5.3. It is a Markov chain describing the probabilities of transitions between two different states of the joint policy - $Same$ $(\pi_1 = \pi_2)$ and $Diff$ $(\pi_1 \neq \pi_2)$. The transition probabilities are taken from the $P(\Pi_{n+1}|\Pi_n)$ distribution which, in turn, is calculated using the following definition of conditional probability:

$$P(\Pi_{n+1}|\Pi_n) = \frac{P(\Pi_{n+1}, \Pi_n)}{P(\Pi_n)} \tag{5.4}$$

where $P(\Pi_{n+1}, \Pi_n)$ is obtained by marginalising all other variables from the overall joint distribution as follows:

$$P(\Pi_{n+1}, \Pi_n) = \sum_{R_{UE1}} \sum_{R_{UE2}} \sum_{I_{UE1}} \sum_{I_{UE2}} \sum_{TxOL} P_{joint} \tag{5.5}$$

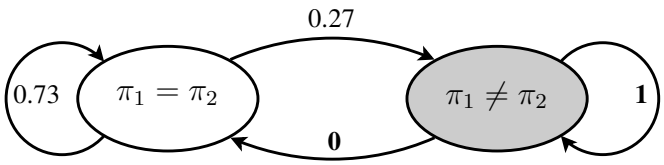


Figure 5.4: An absorbing Markov chain describing the transitions between two states of the joint policy derived from Bayesian network model of the 2 base station 2 user equipment cellular network

Firstly, the Markov chain in Figure 5.4 shows that "$\pi_1 \neq \pi_2$" is an absorbing state, i.e. a state that cannot be left, since the probability of transition from "$\pi_1 \neq \pi_2$" to "$\pi_1 = \pi_2$" is zero. Therefore, this is an absorbing Markov chain which formally demonstrates that the RL algorithm is guaranteed to converge on the desired absorbing state "$\pi_1 \neq \pi_2$". The speed of convergence is controlled by the probability of transition from "$\pi_1 = \pi_2$" to "$\pi_1 \neq \pi_2$", which in this case is 0.27. The objective of future, more advanced RL algorithms, designed using the method proposed in this chapter, is to increase this transition probability to speed up their convergence and, thus, increase their adaptability, whilst preserving the absorbing state "$\pi_1 \neq \pi_2$".

5.6 Conclusion

The Bayesian network based joint policy transition analysis methodology proposed in this chapter is able to provide a simple and accurate probabilistic model of distributed RL algorithms applied to a minimum complexity DSA problem. A Monte Carlo simulation of a distributed Q-learning based DSA algorithm shows that the proposed approach demonstrates remarkably accurate prediction of the convergence behaviour of such algorithms. Furthermore, their behaviour can also be expressed in the form of an absorbing Markov chain, derived from the novel Bayesian network model. This representation enables further theoretical analysis of convergence and adaptability properties of RL based DSA algorithms. Finally, the main benefit of the analysis tool presented in this chapter is that it enables the design and theoretical evaluation of novel RL based DSA algorithms by extending the proposed Bayesian network model, that describes a standard distributed Q-learning scheme.

Chapter 6. Distributed Heuristically Accelerated Q-Learning

Contents

6.1 Motivation

Although RL algorithms such as stateless Q-learning investigated in Chapters 4 and 5 have been shown to be a powerful approach to problem solving, their common disadvantage is the need for many learning iterations before convergence on an acceptable solution, which significantly limits their adaptability in challenging and potentially dynamic multi-agent environments. One of the more recent promising solutions to this issue, proposed in the artificial intelligence domain, is the heuristically accelerated reinforcement learning (HARL) approach. Its goal is to speed up RL algorithms by guiding the exploration process using additional heuristic information [11]. In [10], case-based reasoning is used for heuristic acceleration in a multi-agent RL algorithm

to assess similarity between states of the environment and to make a guess at what action needs to be taken in a given state, based on the experience obtained in other similar states. In [11], Bianchi et al. prove the convergence of four multi-agent HARL algorithms and show how they outperform the regular RL algorithms. There appears to be no evidence in the literature of the HARL approach being applied in the wireless communications domain.

The purpose of this chapter is to alleviate the problem of poor temporal performance of RL based DSA algorithms and, thus, to improve their adaptability, by proposing a cognitive DSA scheme which combines distributed Q-learning and standardised inter-cell interference coordination (ICIC) signalling in LTE networks using a novel adaptation of the HARL framework. Furthermore, it is designed to comply with the current LTE standards and enables robust distributed machine intelligence to be easily implemented in current or future LTE releases.

In previous work on combining ICIC and RL, researchers have only considered applying RL to learning various parameters related to ICIC or radio resource management in Orthogonal Frequency-Division Multiple Access (OFDMA) cellular systems, such as LTE or WiMAX. For example, Simsek et al. [81] use RL to learn optimal cell range bias and power allocation strategies and compare them to static ICIC methods; Dirani and Altman [25] use a fuzzy Q-learning algorithm and ICIC to learn a coordinated power allocation strategy; and Vlacheas et al. [91] use a fuzzy RL principle for automatic tuning of the Relative Narrowband Transmit Power (RNTP) indicator, which is a key ICIC parameter in the LTE downlink. However, no evidence of previous work in the literature was found on using heuristic ICIC methods to enhance the performance of RL based DSA algorithms.

6.2 Heuristically Accelerated Reinforcement Learning

Figure 6.1 shows a novel block diagram representation of the processes involved in HARL. It demonstrates that HARL is an extension of regular RL algorithms. The unfilled blocks and solid lines constitute a block diagram of regular RL depicted in Figure 2.6, whereas the dashed lines and shaded blocks indicate the additional functionality

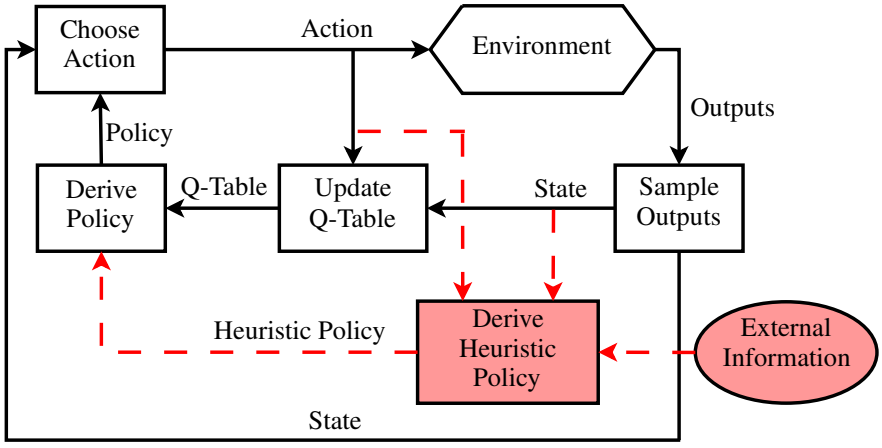


Figure 6.1: Block diagram of heuristically accelerated reinforcement learning

afforded by the heuristic acceleration.

The role of the inner RL loop is to learn a good policy to be used by the learning agent. It achieves this goal by observing the actions taken by the learning agent, sampling the outputs caused by them, and directly estimating (updating) the entries in the Q-table. The role of the policy is to map every state of the environment to the most appropriate action that can be taken in that state. It can be derived from the estimated Q-table and used for decision making. In the context of the DSA problem, the output of interest is whether or not a file transmission is blocked or interrupted, and the action is the piece of resources allocated it.

The key additional element provided by HARL is the derivation of a heuristic policy. According to [11], a heuristic policy is derived from additional knowledge, either external or internal, which is not included in the learning process. Generally, the goal of the heuristic policy $H_t(s, a)$ is to influence the action choices of a learning agent, i.e. to modify its current policy $\pi_t(s)$ in a way which would accelerate the learning process. The format and dimensions of $H_t(s, a)$ should be compliant with the Q-table used by the given learning agent, such that its new combined policy $\pi_t^c(s)$ can be derived using the following equation:

$$\pi_t^c(s) = \underset{a}{\operatorname{argmax}}(Q_t(s, a) + H_t(s, a)) \tag{6.1}$$

where $\pi_t^c(s)$ is the combined policy of the given learning agent for state s at time t based on its Q-table $Q_t(s, a)$ and the heuristic policy $H_t(s, a)$. If $H_t(s, a)$ is always

zero, the algorithm becomes a regular RL algorithm. In the case of the stateless Q-learning algorithm described in Subsection 4.1.2, the heuristic function would not have a state dimension and can be denoted by $H_t(a)$.

6.3 Distributed ICIC Accelerated Q-Learning

This section proposes the distributed ICIC accelerated Q-learning (DIAQ) DSA scheme that combines distributed Q-learning and ICIC using the HARL framework introduced in the previous section to mitigate the issue of poor temporal performance characteristics of Q-learning based DSA algorithms.

As described in Subsection 3.1.6, by using ICIC signalling over the X2 interface, every eNB in an LTE network has the capability of knowing on which virtual resource blocks (VRBs) the neighbouring eNBs are likely to interfere with it, i.e. transmit at a power above the RNTP threshold. In a scenario, where a 20 MHz LTE channel consisting of 100 VRBs is allocated to the network, the length of an RNTP message is 100 bits or 25 hexadecimal characters. There, every subchannel consists of 4 adjacent VRBs, if resource allocation "Type 0" is used [3]. In this case the RNTP messages sent by every eNB to its neighbours contain 25 hexadecimal characters, stating which subchannels they need to reserve to avoid inter-cell interference. $0xF$ denotes that a subchannel is in use, and $0x0$ means it is safe to use by the eNB which receives the RNTP message.

The DIAQ scheme proposed in this section uses these RNTP messages for creating ICIC bitmasks indicating which subchannels are not safe to use for any given eNB, as notified by its neighbours, and using these bitmasks for creating heuristic functions $H_{ICIC}(a)$, which in turn influence the spectrum assignment choices made by the distributed Q-learning based DSA algorithm.

When a request for a new file transmission is received, the eNB starts by aggregating the latest RNTP messages from its neighbours into an ICIC bitmask using a bitwise *OR* operation, as described by the following equation:

$$Mask_{ICIC} = \bigcup_{n=1}^{N} RNTP_n \tag{6.2}$$

where $Mask_{ICIC}$ is a 25 hexadecimal character string representing the subchannels reserved by any of the neighbouring base stations by F, and representing the "safe-to-use" subchannels by 0, $RNTP_n$ is a 25 hexadecimal character RNTP message of the n^{th} neighbouring eNB, and N is the total number of neighbouring eNBs. The RNTP message exchanges can take place as often as every 20 ms [79], and they do not have to be synchronised. Every eNB always uses the latest RNTP message received from a given neighbour.

After creating the ICIC mask, the eNB derives a heuristic function $H_{ICIC}(a)$ as follows:

$$H_{ICIC}(a) = \begin{cases} h_{ICIC} & Mask_{ICIC}(a) = 0xF \\ 0 & Mask_{ICIC}(a) = 0x0 \end{cases} \tag{6.3}$$

where $H_{ICIC}(a)$ is the value of the heuristic function for subchannel a, $Mask_{ICIC}(a)$ is the hexadecimal number in the ICIC bitmask that corresponds to subchannel a, and h_{ICIC} is a fixed negative value with a greater amplitude than the full range of possible $Q(a)$ values. In case of the distributed Q-learning algorithm described in Subsection 4.1.2, $Q(a) \in [-1, 1]$, therefore $h_{ICIC} < -2$. $H_{ICIC}(a)$ can be employed to create a temporary masked Q-table $Q_m(a)$ using the following equation:

$$Q_m(a) = Q(a) + H_{ICIC}(a) \tag{6.4}$$

$Q_m(a)$ is then used for heuristically guided decision making, whilst a normal learning process takes place using $Q(a)$, as defined in the stateless Q-learning update formula given in Equation (4.4).

By using the proposed $Q_m(a)$ and $H_{ICIC}(a)$, the eNB is guaranteed to prioritise the subchannels marked as "safe" by $Mask_{ICIC}$ before the "unsafe" subchannels by shifting the Q-values of the latter to the bottom of the Q-table, whilst still preserving their respective order in terms of the Q-values (due to the fixed value of h_{ICIC}).

The detailed flowchart of the proposed DIAQ scheme is shown in Figure 6.2. The novel ICIC related algorithm steps are red and use dotted outlines. The rest of the flowchart describes a regular distributed Q-learning based DSA process introduced in Subsection 4.1.2. The shaded blue blocks with solid outlines indicate the functions which drive the RL process, i.e. update the Q-table.

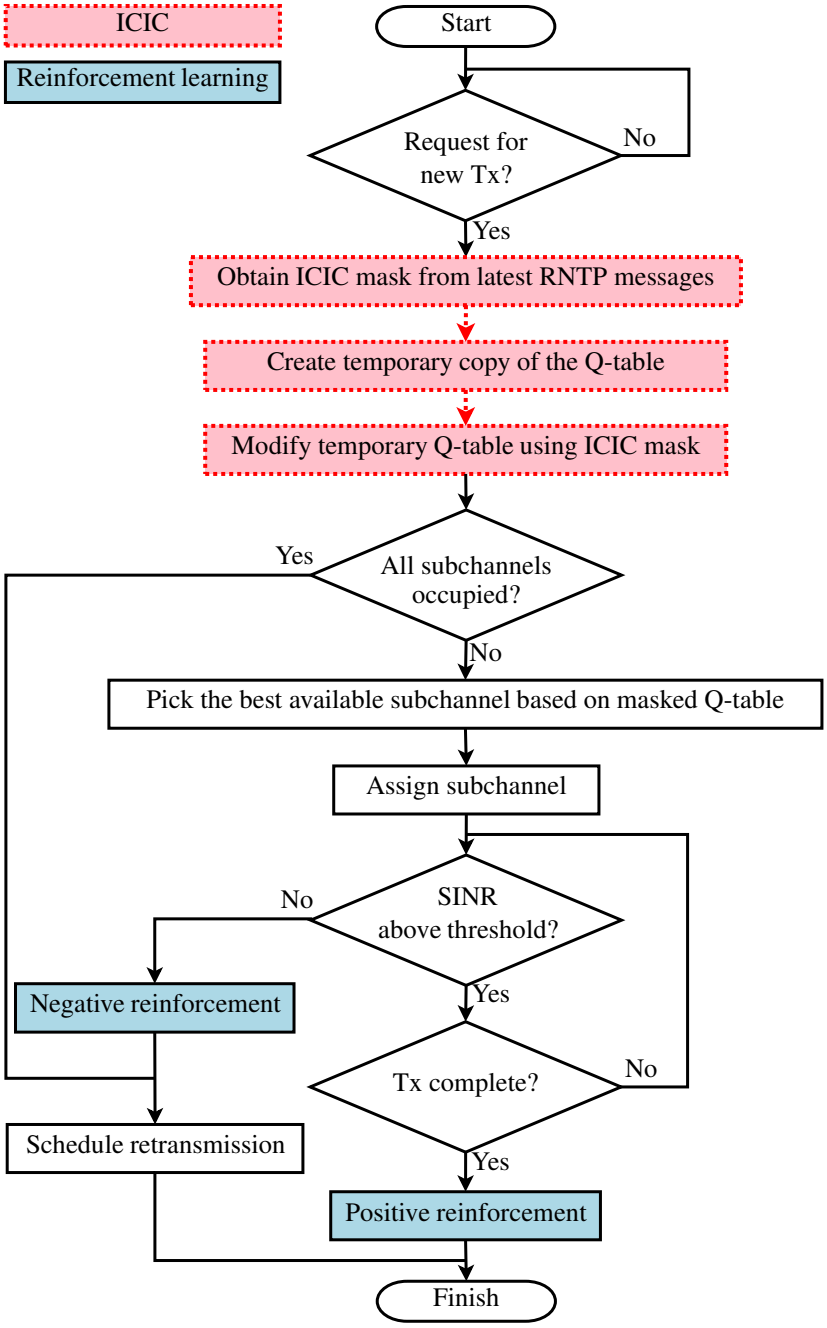


Figure 6.2: Flowchart of the distributed ICIC accelerated Q-learning (DIAQ) scheme

6.4 Theoretical Evaluation

Before testing the developed DIAQ scheme in a realistic cognitive cellular system simulation scenario described in Subsection 3.1.1, its expected performance improvements over regular distributed Q-learning are analytically derived using the simple 2 eNB 2 UE inter-cell interference problem from Section 5.2 and a novel extension to the Bayesian network model proposed in Section 5.3.

6.4.1 Modified Bayesian Network Model

Figure 6.3 presents an adaptation of the Bayesian network model proposed in Section 5.3 which describes the behaviour of DIAQ when applied to the simple 2 eNB 2 UE cellular network described Section 5.2. The shaded nodes and dotted edges show extra dependencies introduced by DIAQ, compared to classical stateless Q-learning described by the Bayesian network shown in Figure 5.2. The variables used to denote the Bayesian network nodes are the following:

- $RNTP \in \{Yes, No\}$ - whether or not, at the latest file arrival time, the corresponding eNB has an up-to-date RNTP message from its neighbour.
- $I_{UEx} \in \{Yes, No\}$ - whether or not $UE1$ or $UE2$ is located within the interference range of the adjacent eNB during the current file arrival.
- $\Pi_n \in \{Same, Diff\}$ - joint policy of the eNBs after n learning iterations. The policy of an eNB is defined as its preferred subchannel (1 or 2). Π_n takes two values of interest - whether the policies of 2 eNBs are the same or different ($\Pi_n = Diff$ is the learning objective).
- $\Pi_n^m \in \{Same, Diff\}$ - joint masked policy, i.e. the combination of Π_n and the heuristic functions of both eNBs defined in Equation (6.3). It is conditionally dependent on Π_n and $RNTP$ (Π_n^m may be different to Π_n, based on the Q-table transformation described by Equation (6.4)).
- $R_{UEx} \in \{S, F\}$ - whether or not a file transmission to $UE1$ or $UE2$ is successful

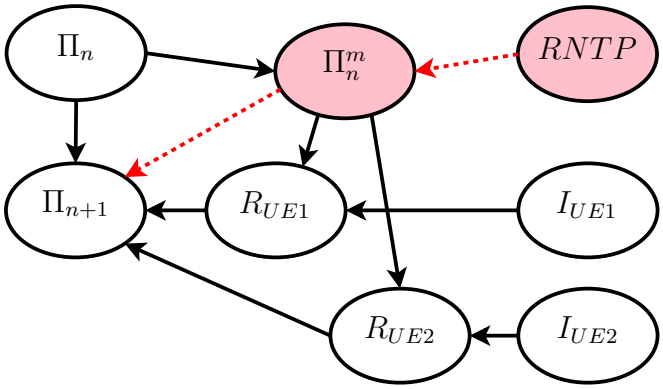


Figure 6.3: Bayesian network describing the behaviour of distributed ICIC accelerated Q-learning applied to the 2 eNB 2 UE dynamic spectrum access network

(S), or whether it failed (F) due to interference. It is conditionally dependent on Π_n^m and I_{UEx}.

- $\Pi_{n+1} \in \{Same, Diff\}$ - the updated joint policy for the next iteration as a result of the outcome at the current iteration. It is conditionally dependent on Π_n, Π_n^m, R_{UE1} and R_{UE2}.

The key difference between this Bayesian network model and the one describing regular stateless Q-learning shown in Figure 5.2 is the addition of the joint masked policy node Π_n^m, which takes into account the $RNTP$ signal from the neighbouring eNB and which is now used for decision making instead of the regular joint Q-learning policy Π_n. The environmental $TxOL$ variable from the original Bayesian network that indicates whether transmissions overlap in time is omitted in the new version for simplicity, since it only affects the convergence speed of the learning process and does not have any effect on the relative performance of DIAQ compared to regular Q-learning. It can be viewed as part of the I_{UE1} and I_{UE2} variables whose only role is determining whether either UE receives harmful inter-cell interference from the adjacent eNB during a pair of transmissions from both eNBs.

Similarly to the Bayesian network model discussed in Section 5.3, based on the conditional dependencies depicted in Figure 6.3, the equation for calculating the joint probability distribution over all variables $P_{joint} = P(\Pi_{n+1}, \Pi_n, \Pi_n^m, R_{UE1}, R_{UE2}, I_{UE1}, I_{UE2}, RNTP)$ is the following:

$$\begin{aligned} P_{joint} = \; & P(\Pi_{n+1}|\Pi_n, \Pi_n^m, R_{UE1}, R_{UE2})\; P(R_{UE1}|\Pi_n^m, I_{UE1})\; P(R_{UE2}|\Pi_n^m, I_{UE2}) \\ & \times P(\Pi_n^m|\Pi_n, RNTP)\; P(\Pi_n)\; P(RNTP)\; P(I_{UE1})\; P(I_{UE2}) \end{aligned} \tag{6.5}$$

which consists of a number of prior probabilities of the form $P(X)$, and conditional probabilities of the form $P(X|Y_1...Y_n)$.

6.4.2 Prior and Conditional Probability Distributions

The prior probability distributions that appropriately describe the 2 eNB 2 UE scenario from Section 5.2 are defined in Table 6.1. The $P(\Pi_0)$ and $P(I_{UEx})$ distributions are identical to those proposed in Section 5.3 for classical stateless Q-learning. The

Table 6.1: Prior probability distributions used in the Bayesian network model of distributed ICIC accelerated Q-learning (DIAQ)

$P(\Pi_0)$		$P(I_{UEx})$		$P(RNTP)$	
$Same$	$Diff$	Yes	No	Yes	No
0.5	0.5	0.4	0.6	$High$	Low

probability of an RNTP message exchange $RNTP$ is a new additional environmental variable that is specific to inter-eNB ICIC signalling used by the DIAQ scheme. $P(RNTP = Yes) = High$ represents a high chance of an RNTP message exchange taking place between current file arrivals at the two eNBs. Since these exchanges can take place as often as every 20 ms, an eNB is highly likely to have an up-to-date RNTP message from its neighbour. If $P(RNTP = Yes)$ is changed to 0, the Bayesian network model will describe the regular stateless Q-learning algorithm introduced in Subsection 4.1.2.

The conditional probability distributions are defined in Table 6.2. The values used for $P(\Pi_n^m|\Pi_n, RNTP)$ state that the masked policies Π_n^m of the eNBs will be the same ($Same$) with a probability of 1, if their Q-learning policies are the same ($\Pi_n = Same$) and there was no RNTP exchange between the file arrivals that could change them ($RNTP = No$). In all other cases, i.e if the Q-learning policies of the eNB are already different ($\Pi_n = Diff$) or if there has been a timely RNTP signal exchange ($RNTP = Yes$) to correct them, the masked policies of the eNBs will always be different ($Diff$). The reasoning behind the $P(R_{UEx}|\Pi_n^m, I_{UEx})$ distribution is to indicate, that a transmission to $UE1$ or $UE2$ will fail with a probability of 1 ($R_{UEx} = F$), if $I_{UEx} = Yes$ and both eNBs have chosen the same subchannel ($\Pi_n^m = Same$). If $\Pi_n^m = Diff$ or $I_{UEx} = No$, then the transmission will be successful: $R_{UEx} = S$.

The $P(\Pi_{n+1}|\Pi_n, \Pi_n^m, R_{UE1}, R_{UE2})$ table defines how the Q-learning policies of both eNBs (Π_{n+1}) are likely to change, given their current Π_n and Π_n^m, and the result of transmissions to both UEs (R_{UE1} and R_{UE2}). Firstly, if both Π_n and Π_n^m are $Same$ or both are $Diff$, and the transmissions to both UEs are successful ($R_{UE1} = R_{UE2} = S$), then both eNBs will reward their respective subchannels and maintain the same policies with a probability of 1 ($\Pi_{n+1} = \Pi_n$). Secondly, if both Π_n and Π_n^m are $Same$ and only a transmission to one of the UEs failed ($\{S, F\}$ or $\{F, S\}$), this UE is

Table 6.2: Conditional probability distributions used in the Bayesian network model of distributed ICIC accelerated Q-learning (DIAQ)

$P(\Pi_n^m \mid \Pi_n, RNTP)$				
$Same$	0	1	0	0
$Diff$	1	0	1	1
	$Same, Yes$	$Same, No$	$Diff, Yes$	$Diff, No$
	$\Pi_n, RNTP$			

$P(R_{UEx} \mid \Pi_n^m, I_{UEx})$				
S	0	1	1	1
F	1	0	0	0
	$Same, Yes$	$Same, No$	$Diff, Yes$	$Diff, No$
	Π_n^m, I_{UEx}			

$P(\Pi_{n+1} \mid \Pi_n, \Pi_n^m, R_{UE1}, R_{UE2})$						
$Same$	1	Low	Low	$High$	$f(n)$	0
$Diff$	0	$High$	$High$	Low	$1 - f(n)$	1
	$Same$	$Same$	$Same$	$Same$	$Same$	$Diff$
	$Same$	$Same$	$Same$	$Same$	$Diff$	$Diff$
	S, S	S, F	F, S	F, F	S, S	S, S
	$\Pi_n, \Pi_n^m, R_{UE1}, R_{UE2}$					

more likely to change its policy due to the WoLF learning rate used in its Q-learning algorithm, described Subsection 4.2.1. Therefore, there is a relatively high probability of the policies being different at the next iteration: $P(\Pi_{n+1} = Diff) = High$. If transmissions to both UEs fail ($\{F, F\}$), both eNBs are likely to change their policies, thus making $\Pi_{n+1} = Same$ a more likely outcome. Lastly, if the Q-learning policies of both eNBs are the same ($\Pi_n = Same$), the masked policies are different ($\Pi_n^m = Diff$), and both transmissions are successful ($R_{UE1} = R_{UE1} = S$), the probability of the $\Pi_{n+1} = Same$ at the next iteration is time-dependent. A realistic approximation of its value at different stages of learning is the following:

$$f(n) = \begin{cases} 0 & n = 0 \\ 0.5 & n = 1 \\ High & n > 1 \end{cases} \tag{6.6}$$

If this is the first learning iteration ($n = 0$), the Q-tables of both eNBs are initialised

to zeros. Therefore, if different subchannels are successfully used ($\Pi_n^m = Diff$), they will be positively reinforced and used at the next iteration with a probability of 1: $P(\Pi_{n+1} = Same|...) = 0$. After one learning iteration, there is about a 50% chance of one of the eNBs changing its Q-learning policy, depending on whether its first trial was a success on its preferred subchannel, or a failure on the other subchannel: $P(\Pi_{n+1} = Same|...) = 0.5$. Afterwards, the eNB, whose Q-learning policy is overriden by the RNTP exchange (since $\Pi_n \neq \Pi_n^m$), is relatively unlikely to change its policy due to the effect of the WoLF learning rates, i.e. the Q-values undergo smaller step changes after successful trials: $P(\Pi_{n+1} = Same|...) = High$.

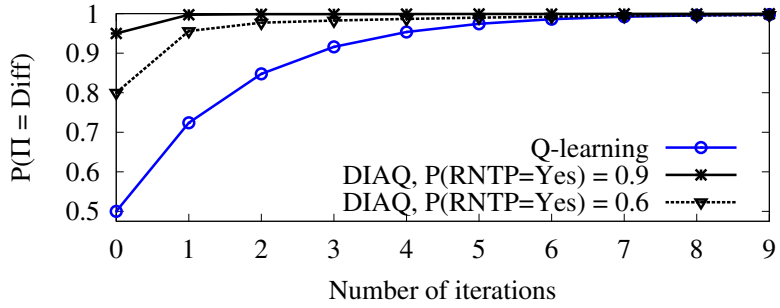
The remaining ten combinations of Π_n, Π_n^m, R_{UE1} and R_{UE2} values are not considered, since they can never occur according to the $P(\Pi_n^m|\Pi_n, RNTP)$ and $P(R_{UEx}|\Pi_n^m, I_{UEx})$ conditional probability distributions. Regardless of the values used for these combinations in the $P(\Pi_{n+1}|\Pi_n, \Pi_n^m, R_{UE1}, R_{UE2})$ table, they will be multiplied by zero during the calculation of the joint probability distribution defined in Equation (6.5).

6.4.3 Convergence Behaviour of DIAQ

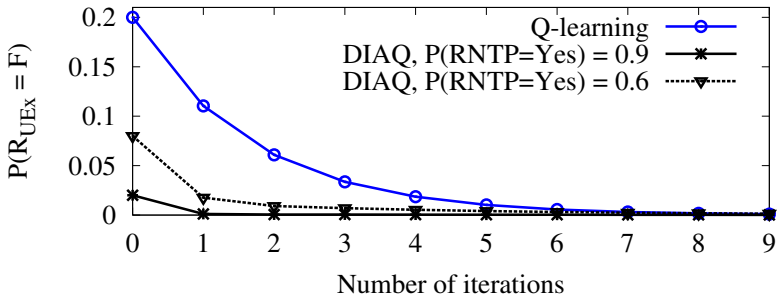
The comparison between the expected convergence behaviour of DIAQ and that of classical stateless Q-learning is obtained using the same approach of iterative evaluation of the Bayesian network model as that proposed in Section 5.3. It shows how the probability of transmission failure $P(R_{UEx})$ and the probability of eNBs using different subchannels $P(\Pi_n^m)$ change over time, as the learning process progresses.

Figure 6.4 shows the results of such iterative evaluation of the Bayesian network from Figure 6.3. It compares the convergence performance of classical stateless Q-learning and DIAQ with $P(RNTP = Yes)$ values of 0.9 and 0.6, respectively the cases where ICIC signalling between the neighbouring eNBs is moderately reliable and relatively unreliable. The values for $High$ and Low in the conditional probability distributions in Table 6.2 are assumed to be $\{0.9, 0.1\}$ without the loss of generality.

Figure 6.4 demonstrates how the presence of RNTP message exchanges in DIAQ, even when they are relatively unreliable ($P(RNTP = Yes) = 0.6$), dramatically speeds up the learning process, especially at its early stages. The eNBs become highly



(a) Probability of eNBs having different policies



(b) Probability of a UE being blocked or interrupted

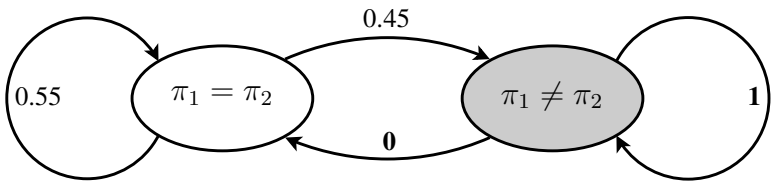
Figure 6.4: Convergence behaviour of Q-learning and DIAQ, using the probabilistic model of the 2 eNodeB 2 user equipment cellular network

likely to converge on the optimal solution ($\Pi = Diff$) significantly faster using DIAQ compared to Q-learning which operates using trial-and-error experience only. Consequently, the temporal performance of the network in terms of the probability of transmission failures shown in Figure 6.4b is also superior using DIAQ.

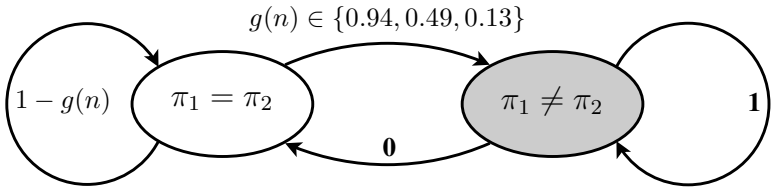
6.4.4 Absorbing Markov Chain Analysis

Figure 6.5 compares the convergence properties of classical stateless Q-learning and DIAQ expressed as absorbing Markov chains that represent the joint policy transition probabilities $P(\Pi_{n+1}|\Pi_n)$. Similarly to the approach described in Section 5.5, this transition probability distribution is calculated by marginalising all variables of the Bayesian network model proposed in Figure 6.3 except Π_{n+1} and Π_n out of P_{joint} and dividing the result by $P(\Pi_n)$ as follows:

$$P(\Pi_{n+1}|\Pi_n) = \frac{\sum_{\Pi_n^c} \sum_{R_{UE1}} \sum_{R_{UE2}} \sum_{I_{UE1}} \sum_{I_{UE2}} \sum_{RNTP} P_{joint}}{P(\Pi_n)} \tag{6.7}$$



(a) Classical stateless Q-learning



(b) Distributed ICIC accelerated Q-learning

Figure 6.5: Absorbing Markov chains describing the transitions between two states of the joint Q-learning policy in the 2 eNodeB 2 user equipment network scenario

The key common feature of both Markov chains in Figure 6.5 is the fact that they are absorbing, i.e. the probability of staying in the desired state "$\pi_1 \neq \pi_2$" is 1. This demonstrates that the additional heuristic functionality of the DIAQ scheme does not break the convergence guarantee of the original stateless Q-learning algorithm introduced Subsection 4.1.2.

Given the fact that "$\pi_1 \neq \pi_2$" is an absorbing state, the probability of transition from "$\pi_1 = \pi_2$" to "$\pi_1 \neq \pi_2$" is directly related to the speed of convergence of the system to the absorbing state, i.e. the higher it is, the faster the system is likely to converge. Figure 6.5b shows that due to a time-variant value in the $P(\Pi_{n+1}|\Pi_n, \Pi_n^m, R_{UE1}, R_{UE2})$ conditional probability distribution from Table 6.2 described by Equation (6.6), the probability of transition from "$\pi_1 = \pi_2$" to "$\pi_1 \neq \pi_2$" also varies with time. In the case where ICIC signalling between the two eNBs is relatively reliable ($P(RNTP = Yes) = 0.9$), it is described by the following equation:

$$P(\Pi_{n+1} = Diff \mid \Pi_n = Same) = \begin{cases} 0.94, & n = 0 \\ 0.49, & n = 1 \\ 0.13, & n > 1 \end{cases} \tag{6.8}$$

This equation demonstrates that DIAQ provides a dramatic improvement in initial performance over the classical distributed RL approach, i.e. due to the high value of the probability of transition from "$\pi_1 = \pi_2$" to "$\pi_1 \neq \pi_2$" at the first learning iteration. Although afterwards this probability rapidly decreases and becomes significantly

lower than that achieved by stateless Q-learning, i.e. 0.13 compared to 0.45, by that time it is highly likely that the system will have already converged. Furthermore, the decision process in DIAQ is based on the masked policies Π_n^m, i.e. the combination of the Q-learning policies and heuristic ICIC information, as opposed to Q-learning policies Π_n only. Therefore, even in the unlikely cases where the eNBs have not learned the optimal strategy after several trials, they are still more likely to employ the correct joint policy "$\pi_1 \neq \pi_2$". This rapid convergence of the joint masked policy achieved by the DIAQ scheme is depicted in Figure 6.4a. In the case where ICIC signalling between the two eNBs is relatively unreliable ($P(RNTP = Yes) = 0.6$), the time-variant values of the probability of transition from "$\pi_1 = \pi_2$" to "$\pi_1 \neq \pi_2$" are $g(n) \in \{0.78, 0.48, 0.24\}$. There, the initial probability of convergence is not as rapid (0.74), yet it is still significantly higher than that achieved by classical stateless Q-learning (0.45).

6.5 Simulation Results

This section presents the results of simulating the proposed DIAQ scheme using the stadium small cell network model with its own dedicated spectrum introduced in Subsection 3.1.1. The performance of this scheme is compared to that of a pure distributed Q-learning algorithm from Subsection 4.1.2 and the typical dynamic ICIC based scheme described in Subsection 3.3.1. The comparison with these two schemes is most appropriate, since they represent two key components of the DIAQ scheme separately - the RL part and the heuristic inter-eNB coordination part. The latter represents a standard approach in LTE [28][79]. Therefore, the results evaluate the importance of both of these components in the proposed DIAQ scheme.

6.5.1 Temporal Performance

Figure 6.6 compares the temporal response of the network in terms of the probability of retransmission at 1 Gbps offered traffic, using dynamic ICIC, pure distributed Q-learning and DIAQ schemes for DSA. The graph shows the average of 50 simulations with different random seeds and UE locations in order to mitigate the noise introduced

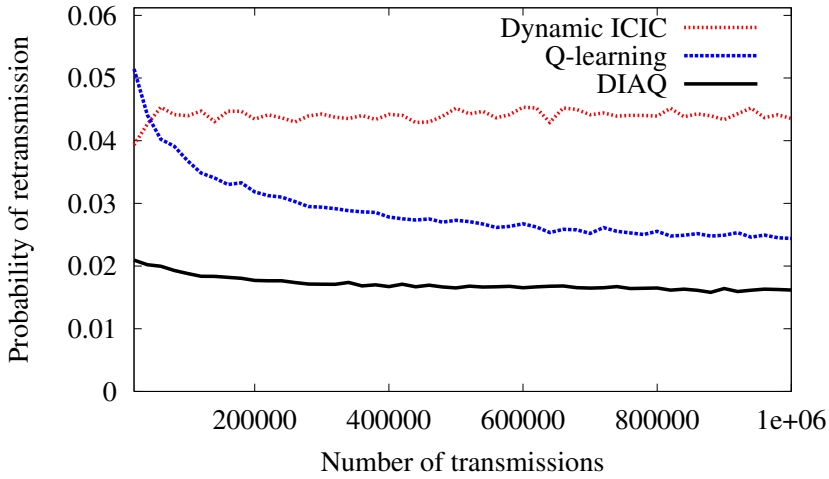


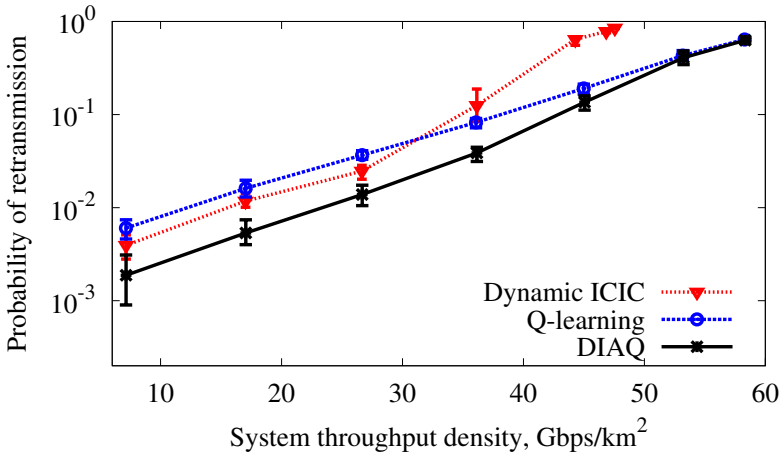
Figure 6.6: Probability of retransmission time response using dynamic ICIC, pure Q-learning and distributed ICIC accelerated Q-learning (DIAQ)

by the bursty nature of the traffic, and to produce a more statistically valid temporal response. Firstly, the graph shows that both Q-learning and DIAQ schemes converge on better DSA policies than the ICIC scheme. Secondly, the DIAQ scheme achieves a significant improvement in the initial performance compared to the classical Q-learning approach. The highly efficient guided exploration process of the DIAQ scheme results in a substantial reduction in initial $P(re - tx)$ by a factor of ≈2.5, compared to pure Q-learning. This improvement is consistent with the theoretically predicted outcome shown in Figure 6.4b.

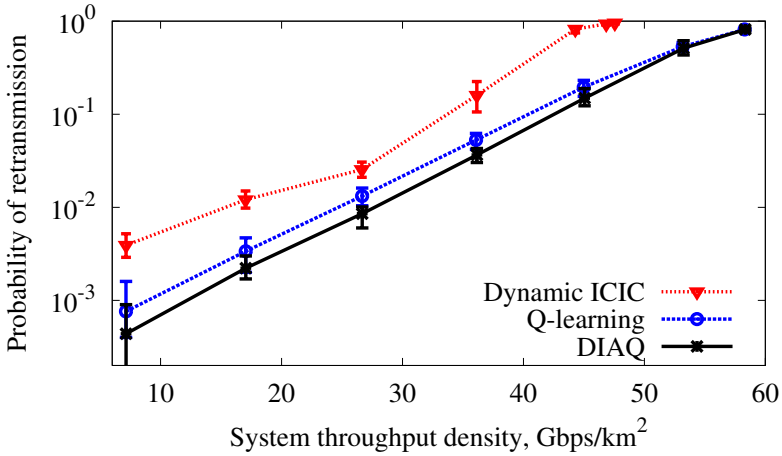
Figure 6.6 also shows that DIAQ still has a significantly lower probability of retransmission compared to both schemes after 1,000,000 trials, when it is approaching its steady state. Therefore, using ICIC to enhance the stateless Q-learning algorithm in this way dramatically speeds up its convergence, and substantially improves both its initial and steady-state performance. Such acceleration of the learning process is crucial in more realistic dynamically changing environments explored in Chapter 8, e.g. with time-varying traffic distributions and topologies. The impact of DIAQ, compared to regular distributed Q-learning, is that it can adapt to new interference environments considerably faster.

6.5.2 Initial and Final Performance

Figure 6.7 shows the difference in initial and final $P(re-tx)$ performance of these schemes at a wide range of traffic loads. It is plotted against the system throughput density to evaluate both the QoS and the system capacity in the same graphs. The initial $P(re-tx)$ in Figure 6.7a is calculated using the first 20,000 transmissions, and the final $P(re-tx)$ in Figure 6.7b is calculated from the last 20,000 file transmissions. The overall simulation length is 1,000,000 file transmissions. Every data point represents the mean result of 50 different simulations at a given traffic load with the error bars showing the minimum and maximum $P(re-tx)$ in those simulations.



(a) Initial probability of retransmission



(b) Final probability of retransmission

Figure 6.7: Initial and final probabilitiy of retransmission using pure ICIC, pure Q-learning and distributed ICIC accelerated Q-learning (DIAQ) at different system throughput densities

Figure 6.7a shows that the dramatic improvement in initial performance using DIAQ instead of the classical Q-learning approach is consistent at most traffic loads. DIAQ introduces a 29-69% reduction in the initial probability of retransmission at system throughput densities below 45 Gbps/km^2. Only at ultra-high system throughput densities does the difference in their performance become negligible. DIAQ also shows a significantly better performance in initial and final probability of retransmission, compared to the dynamic ICIC scheme. Furthermore, the latter only supports system throughput densities of up to 48 Gbps/km^2, whereas DIAQ and Q-learning are significantly more robust at extremely high offered traffic densities. They both manage to support system throughput densities of up to 58 Gbps/km^2. This demonstrates that it is better to take opportunistic spectrum assignment decisions, based on reinforcement learning, instead of blocking transmissions based on ICIC signalling, since the probability of a subchannel not being occupied by any of the neighbouring eNBs tends to zero. In these cases, the heuristic ICIC approach "blindly" blocks most file transmissions, whereas Q-learning is still capable of providing some insight into which subchannels could result in successful transmissions.

6.6 Conclusion

The novel DIAQ scheme proposed in this chapter combines distributed RL and standardized ICIC signalling in the LTE downlink, using the framework of HARL. It is theoretically evaluated using a novel extension of the Bayesian network model proposed in Chapter 5, which explains a predicted improvement in convergence behaviour achieved by DIAQ, compared to classical distributed RL. Large scale simulation experiments of a stadium small cell network show that it provides superior QoS compared to a typical heuristic ICIC approach and a state-of-the-art distributed RL based approach. It achieves a significantly lower probability of retransmission and supports higher system throughput densities of up to 58 Gbps/km^2. A comparison of the probability of retransmission time response characteristics of DIAQ and pure distributed Q-learning reveals a dramatic improvement in performance at the initial stage of learning, a 29% to 69% improvement ranging across all but ultra-high traffic loads, due to the use of heuristics for guiding the exploration process. This result confirms the theoretical pre-

dictions made using the Bayesian network model of the algorithm. DIAQ also exhibits excellent final performance and convergence speed. The dramatic improvements in the initial performance and convergence speed achieved by the heuristic acceleration of the learning process significantly increases the adaptability of the distributed RL based approach to DSA, since the cognitive eNBs are able to adapt to each other's dynamically changing policies considerably faster. Finally, the DIAQ scheme is designed to comply with the current LTE standards. Therefore, it allows easy implementation of robust distributed machine intelligence for full self-organisation in existing commercial networks.

Chapter 7. Robust Intelligent Dynamic Spectrum Sharing

Contents

7.1 Motivation

The key feature of the novel distributed ICIC accelerated Q-learning (DIAQ) scheme proposed in Chapter 6 is the use of heuristic spectrum awareness information for a significant increase in the adaptability and robustness of distributed RL based DSA in terms of the QoS convergence behaviour. The purpose of this chapter is to report on the novel application of the HARL framework to a more complex DSA problem where

the cognitive cellular system shares spectrum with other independent primary and secondary wireless networks. The dynamic spectrum sharing (DSS) scenario described in Subsection 3.1.1 represents a relevant and realistic context for this problem and is used for the development and evaluation of the novel algorithms described in this chapter.

The heuristic acceleration for the RL based DSA algorithms developed in this chapter is provided by a dynamically updated spectrum usage database, also known as the radio environment map (REM), which is a commonly used component in secondary cognitive wireless networks [60]. In previous work on combining RL and dynamic spectrum databases, such as REMs, researchers have considered employing RL algorithms solely for obtaining information that can be stored in these databases, e.g. [17][56]. There appears to be no evidence of previous work in the literature on using REM databases to enhance the performance of RL based DSA and DSS algorithms.

7.2 HARL for Dynamic Spectrum Sharing

The stadium temporary event spectrum sharing scenario described in Subsection 3.1.1 and shown in Figure 7.1 consists of a network of primary eNBs (PeNBs) operating in a suburban area and a secondary cognitive cellular system that itself consists of two separately operating entities - an aerial eNB (AeNB) for wide area coverage and a small cell network for high capacity density inside the stadium.

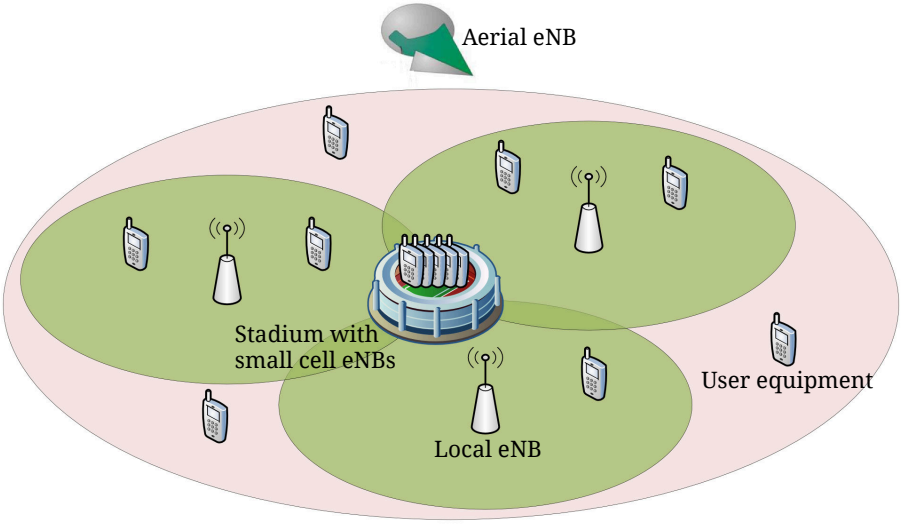


Figure 7.1: Dynamic spectrum sharing scenario designed for stadium temporary events

A study in Section 4.3 has demonstrated that successful dynamic secondary spectrum sharing between a low power stadium small cell system and a relatively high power local PeNB infrastructure can be facilitated using an independent distributed Q-learning algorithm implemented in the former at network-wide traffic loads of up to 1 Gbps. This is largely because the interference between the two systems is attenuated by the stadium shell. However, the scenario investigated in this chapter also involves an AeNB serving line-of-sight (LoS) users both inside and outside the stadium. Therefore, it presents two additional challenges - spectrum sharing between the PeNBs and the AeNB, and spectrum sharing between the AeNB and the stadium small cell network.

The way of achieving these two spectrum sharing tasks proposed in this chapter is to use a REM to continuously monitor and store the information about spectrum usage of the PeNBs and the AeNB. In this way, the AeNB has the means to avoid interfering with the primary system, and the small cell network can avoid interfering with the AeNB. This type of setup is depicted in Figure 7.2, which is a classical way of achieving coexistence between cognitive radio networks and primary spectrum users, especially in the TV white space context as described in Subsection 2.2.1.

The task of the spectrum monitoring system with a REM database is to detect the occupancy of the spectrum resources used by the PeNBs and the AeNB. It is then possible to estimate the probability of spectrum occupancy at every eNB on every individual subchannel that, in turn, can be used to influence the spectrum assignment decisions of the secondary systems.

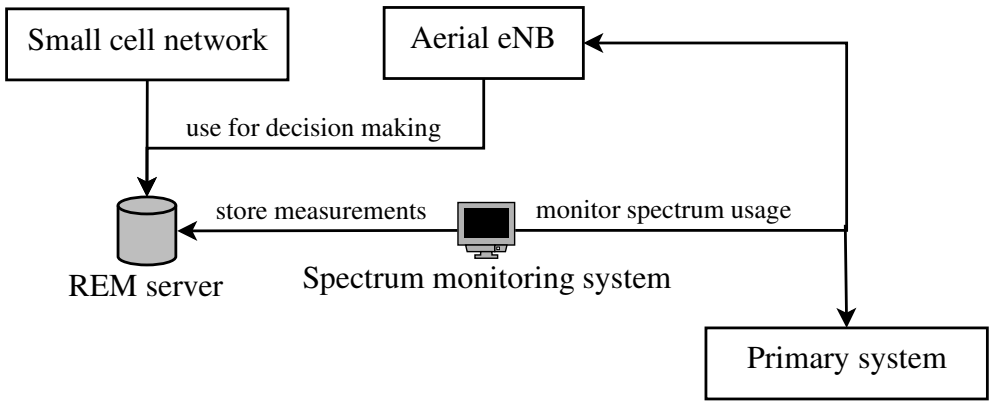


Figure 7.2: Secondary spectrum sharing using a spectrum monitoring system and a radio environment map (REM)

7.2.1 Spectrum Monitoring

One way of implementing reliable spectrum monitoring in such LTE cellular systems is for the primary system to grant the secondary system access to its ICIC signals. The ICIC signals standardized for the LTE downlink are explained in Subsection 3.1.6. In this way, the binary spectrum occupancy information about the PeNBs and the AeNB could be logged at the REM server and used to make predictions about the spectrum availability. Such a protocol is easily implementable, especially if both systems are controlled by the same mobile network operator (MNO). However, if the secondary cognitive network is not controlled by the primary system's MNO, it may not be allowed to access the ICIC signals of the primary system. In such cases, dynamic spectrum monitoring could be achieved by deploying a sensor network around the stadium to detect spectrum usage of every PeNB and AeNB, e.g. using an algorithm for multiple signal classification [76].

Regardless of the detection mechanism, the algorithms proposed in this section assume that the spectrum monitoring system is able to periodically detect whether or not a particular subchannel is being used by a particular PeNB or AeNB. It is designed to return 1 if it is currently occupied, or 0 otherwise.

7.2.2 Spectrum Occupancy Estimation

Given the mechanism for obtaining a stream of binary spectrum occupancy data, it is then important to estimate the probability of subchannel occupancy at every observed eNB, i.e. a probability of a particular subchannel being occupied at a particular eNB based on the previous observations.

A simple and appropriate way of tracking the mean of a data sequence, whilst simultaneously giving more recent observations higher weight compared to older estimates, is the exponentially weighted moving average (EWMA) method [73]. It is described by the following recursive equation:

$$y \leftarrow (1 - \lambda)y + \lambda x \tag{7.1}$$

where y is the mean estimate of the data sequence x, and λ is a factor which controls how quickly the estimated mean adapts to new observations. The role of λ in EWMA estimation is identical to that of the learning rate α in the stateless Q-learning update formula from Equation (4.4). In fact, comparing Equations (4.4) and (7.1) demonstrates that stateless Q-learning is, in fact, an EWMA estimation algorithm of the rewards received by a learning agent.

We propose adapting the EWMA method to estimate the probability of subchannel occupancy $p(occupied)$ in the following way:

$$p(occupied) \leftarrow (1-\lambda)p(occupied) + \lambda b,\ b \in \{0,1\} \tag{7.2}$$

where b is a current binary subchannel occupancy measurement, i.e. $b = 1$ if the given subchannel is occupied, $b = 0$ if it is not. In this way, the EWMA equation is used to estimate the mean of a stream of 1's and 0's, representing $p(occupied) \in [0,1]$.

7.2.3 REM Based Heuristic Function

A threshold P_{min} to determine whether a particular subchannel should be avoided, based on an estimate of $p(occupied)$, can then be defined to obtain the following heuristic function:

$$H_{REM}(a) = \begin{cases} h_{REM} & p_a(occupied) \geq P_{min} \\ 0 & p_a(occupied) < P_{min} \end{cases} \tag{7.3}$$

where $H_{REM}(a)$ is the value of the REM based heuristic function for subchannel a, $p_a(occupied)$ is the EWMA estimate of $p(occupied)$ for subchannel a, h_{REM} is a fixed negative value which shifts the Q-values of the undesirable subchannels down, such that the others are prioritized before them. This heuristic function follows the same principle of shifting Q-values as the one used in DIAQ proposed in Chapter 6.

Such a heuristic function $H_{REM}(a)$ aims to guide the learning process of the cognitive eNBs in a direction desirable for secondary spectrum sharing. The small cell eNBs can coexist with the AeNB by applying the heuristic function from Equation (7.3) to the AeNB subchannel occupancy observations, hereafter referred to as $H_{REM}^{AeNB}(a)$. The

AeNB can coexist with the PeNBs by applying the same principle to PeNB subchannel occupancy observations. In the latter case, since the wide area coverage AeNB is going to interfere with all PeNBs in the area of interest, the probability of subchannel a being occupied by any PeNB is obtained by calculating the sum of $p_a(occupied)$ values of every individual PeNB:

$$p_a^{PeNBs}(occupied) = \sum_{n=1}^{N} p_a^{n^{th}\ PeNB}(occupied) \tag{7.4}$$

where N is the total number of PeNBs. The REM based heuristic function from (7.3) can then be calculated using $p_a^{PeNBs}(occupied)$, hereafter referred to as $H_{REM}^{PeNBs}(a)$.

7.2.4 Superimposed Heuristic Functions

With the introduction of the REM based heuristic function for secondary spectrum sharing, a framework for using several heuristic functions simultaneously is required. For example, in addition to using an ICIC based heuristic function $H_{ICIC}(a)$ introduced in Section 6.3 for internal dynamic spectrum access, the small cell eNBs are now also required to share spectrum with the AeNB using another heuristic function $H_{REM}^{AeNB}(a)$, such that their masked Q-tables $Q_m(a)$ could be constructed using the following principle:

$$Q_m(a) = Q(a) + H_{ICIC}(a) + H_{REM}^{AeNB}(a) \tag{7.5}$$

where $Q(a) \in [-1, 1]$ is an original Q-table of a given eNB maintained using the stateless Q-learning algorithm described in Subsection 4.1.2. There, two heuristic functions $H_{ICIC}(a)$ and $H_{REM}^{AeNB}(a)$ have to be superimposed to modify a learning eNB's policy, such that it incorporates both ICIC and REM information into its learning process.

The author proposes a novel method where every new heuristic function superimposed on the Q-table splits the Q-values into two non-overlapping regions, as shown in Figure 7.3. The normal range of Q-values $Q(a)$ maintained by the stateless Q-learning algorithm from Subsection 4.1.2 is $[-1, 1]$. If the h_{ICIC} parameter of the $H_{ICIC}(a)$ heuristic function is -3, it shifts $Q_m(a)$ values of disapproved subchannels into a non-

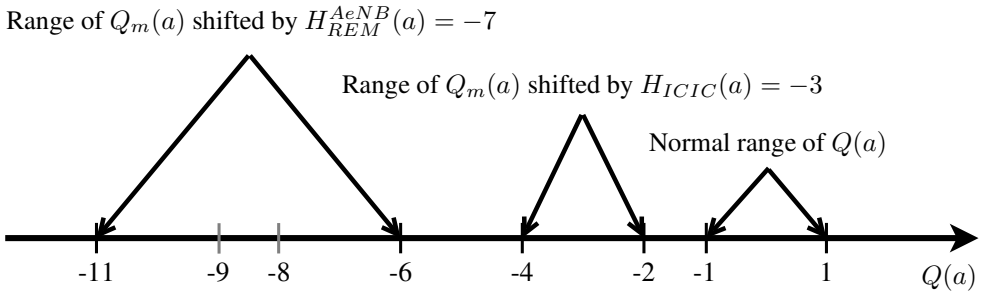


Figure 7.3: The effect of superimposed heuristic functions $H_{ICIC}(a) \in \{0, -3\}$ and $H_{REM}^{AeNB}(a) \in \{0, -7\}$ on the range of masked Q-table values

overlapping region of $(Q(a) - 3) \in [-4, -2]$, thus prioritising them below the subchannels with $Q_m(a) \in [-1, 1]$. If another heuristic function $H_{REM}^{AeNB}(a)$ is used and its h_{REM} constant is -7, it will split $Q_m(a)$ into two regions - $Q_m(a) \in [-4, 1]$ and $(Q_m(a) - 7) \in [-11, -6]$. In this way, the subchannels disapproved by $H_{REM}^{AeNB}(a)$ are guaranteed to be prioritised below any other subchannel. This approach allows an unlimited number of further heuristic functions superimposed on top of each other, as long as their respective importance is known. For example, in this case $H_{REM}^{AeNB}(a)$ responsible for spectrum sharing is prioritised above $H_{ICIC}(a)$ responsible for internal stadium network DSA by setting $h_{REM} < h_{ICIC} - \Delta Q$, where $\Delta Q = 2$ is the difference between the minimum and the maximum possible value in the original Q-table.

7.2.5 Q-Value Based Admission Control

The HARL algorithm required for the AeNB to coexist with the primary system only includes one heuristic function $H_{REM}^{PeNBs}(a)$, since it is a separately controlled entity with no ICIC-compatible neighbouring base stations. Therefore, it uses the following masked Q-table for guiding its learning process:

$$Q_m(a) = Q(a) + H_{REM}^{PeNBs}(a) \tag{7.6}$$

However, another important aspect of secondary spectrum sharing is the primary user protection [29], i.e. making sure the secondary system, in this case the AeNB, does not produce harmful interference for the primary system, in this case the users connected to the PeNBs. A simple technique that could be easily and effectively embedded into

the HARL framework developed in this chapter, i.e. where $H_{REM}^{PeNBs}(a)$ shifts part of the Q-values by a fixed negative number h_{REM}^{PeNBs}, is the introduction of Q-value based admission control (Q-AC) [64]. A Q-value admission threshold q_{AC} can be defined such that all subchannels whose masked Q-values are below it are deemed unavailable for assignment, as follows:

$$A_{allowed} = \{a \mid a \in A' \ \wedge \ Q_m(a) \geq q_{AC}\} \tag{7.7}$$

where A' is the set of currently unoccupied subchannels, i.e. those available for assignment, and $A_{allowed} \subset A'$ is the set of subchannels allowed for assignment based on the admission threshold q_{AC}. In this way, the subchannels with $Q_m(a) < q_{AC}$ are never assigned to data transmissions, which are blocked instead.

The value of q_{AC} can be chosen such that:

$$q_{max} - h_{REM}^{PeNBs} < q_{AC} < q_{min} \tag{7.8}$$

where q_{min} and q_{max} are the minimum and the maximum possible value of the Q-table $Q(a)$ before the transformation respectively. In this way, the subchannels disapproved by the heuristic function $H_{REM}^{PeNBs}(a)$ are always forbidden for assignment at the AeNB, due to their Q-values being shifted below q_{AC}, thus guaranteeing protection of the PeNBs from secondary interference.

7.2.6 HARL Algorithms for Spectrum Sharing

Algorithms 2 and 3 summarize the HARL schemes for dynamic secondary spectrum sharing developed in this section. Algorithm 2 shows the sequence of steps in the distributed REM and ICIC accelerated Q-learning (DRIAQ) scheme, designed for stadium small cells to mitigate interference among themselves and the AeNB, using two superimposed heuristic functions. Algorithm 3 shows the REM accelerated Q-learning algorithm with Q-value based admission control (RAQ-AC), designed for the AeNB to share spectrum and avoid interference with the primary system. Lines {2, 8, 9} of Algorithm 2 and lines {2, 8-12, 14} of Algorithm 3 are specific to the HARL framework developed in this section. If they are removed and $Q_m(a)$ is replaced by $Q(a)$,

the algorithms are simplified down to stateless Q-learning from Subsection 4.1.2.

Algorithm 2 Distributed REM and ICIC accelerated Q-learning (DRIAQ) for stadium small cells

Initialise Q-table to all zeros
Set $h_{ICIC} = -3$ and $h_{REM}^{AeNB} = -7$
while eNB is on **do**
  Wait for a file arrival
  if all subchannels are occupied **then**
    Block transmission
  else
    Update $H_{ICIC}(a)$ and $H_{REM}^{AeNB}(a)$ based on latest ICIC and REM information, using Equations (6.3) and (7.3)
    Combine $Q(a)$ with $H_{ICIC}(a)$ and $H_{REM}^{AeNB}(a)$ into a masked Q-table $Q_m(a)$ using Equation (7.5)
    Assign the best subchannel using $Q_m(a)$ and Equation (4.3)
    Observe the outcome, calculate the reward $r = \pm 1$
    Update $Q(a)$ using Equation (4.4)
  end if
end while

Algorithm 3 REM accelerated Q-learning with Q-value based admission control (RAQ-AC) for the aerial eNB

Initialise Q-table to all zeros
Set $h_{REM}^{PeNBs} = -7$ and $q_{AC} \in (-6, -1)$ as shown in Equation (7.8)
while eNB is on **do**
  Wait for a file arrival
  if all subchannels are occupied **then**
    Block transmission
  else
    Update $H_{REM}^{PeNBs}(a)$ based on latest REM information, using Equation (7.3)
    Combine $Q(a)$ with $H_{REM}^{PeNBs}(a)$ into a masked Q-table $Q_m(a)$ using Equation (7.6)
    if all subchannels with $Q_m(a) \geq q_{AC}$ are occupied **then**
      Block transmission
    else
      Assign the best subchannel using $Q_m(a)$ and Equation (4.3)
    end if
    Observe the outcome, calculate the reward $r = \pm 1$
    Update $Q(a)$ using Equation (4.4)
  end if
end while

7.2.7 Choice of Parameters

The final details required to complete the design of the REM and the REM based heuristic functions are the values of the EWMA algorithm parameter λ from Equation (7.1) and the probability of subchannel occupancy threshold P_{min} for $H_{REM}^{AeNB}(a)$ and $H_{REM}^{PeNBs}(a)$ as used in Equation (7.3). The author proposes using $P_{min} = \lambda$ and $\lambda = 0.008$, while the REM is updated every 200 ms, which is frequent enough to capture the traffic variations of the PeNBs and the AeNB, yet not too frequent to introduce a large overhead of additional REM information that has to be broadcast to all cognitive eNBs. However, other values can be used for these parameters without the loss of generality.

The value $\lambda = 0.008$ is chosen based on the rate of decay of a $p_a(occupied)$ estimate, e.g. the time it would take for a once heavily used subchannel to be assumed unused, if the eNB of interest stopped using it. For example, if $p_a(subchannel) = 0.99$ and afterwards subchannel a is not used for 600 consecutive REM updates, i.e. 2 minutes, the new $p_a(occupied)$ estimate, based on Equation (7.2), is the following:

$$p_a(occupied) = 0.99 \times (1 - \lambda)^{600} = 0.00799 \tag{7.9}$$

which is just below $P_{min} = \lambda = 0.008$. Therefore subchannel a would no longer be undesirable for secondary reuse, based on the heuristic function from Equation (7.3). This value of λ is high enough to be applicable in dynamic environments where the monitored spectrum usage patterns change over time, yet not high enough to dismiss valuable historical spectrum usage information too quickly. This trade-off between the speed and accuracy of the EWMA algorithm, controlled by the λ parameter, is essential and must be carefully considered, e.g. using numerical examples such as the one described in Equation (7.9).

The value $P_{min} = \lambda$ is proposed because it is crucial that, if interference is detected on a previously unused subchannel with $p(occupied) = 0$, the new estimate of $p(occupied)$ is such that this subchannel is recognised as busy straightaway. In this case the $p(occupied)$ estimate will change from 0 to $\lambda = P_{min}$ which is high enough to be flagged by the REM based heuristic function described by Equation (7.3).

7.3 Simulation Results

The spectrum sharing problem described in Subsection 3.1.1 and depicted again in Figure 7.1 involves an AeNB and a network of small cell eNBs that have to share spectrum among themselves and with a primary system of local eNBs operating in the area.

The primary system is assumed to employ the same dynamic ICIC scheme as that used in the simulation experiments in Section 4.3. There, all three PeNBs exchange their current spectrum usage as RNTP messages every 20 ms, and exclude the subchannels currently used by the other two PeNBs from their available subchannel list. However, the DSS schemes developed for the secondary systems in this chapter do not assume this and would also work regardless of the spectrum management strategy of the primary system.

The results of implementing the following three schemes in the secondary cognitive system are discussed in this section:

- "Dynamic ICIC" - all systems use ICIC signalling as described in Subsection 3.3.1. The stadium eNBs receive ICIC messages from the AeNB and from their neighbouring small cells. They only report subchannels used at a Tx power above -3 dB with respect to the average power in the cell, and choose randomly among the subchannels deemed "safe". The AeNB randomly assigns subchannels not used by the primary system, based on the ICIC messages of the latter.
- "DIAQ + Q-learning" - all networks are working independently. The stadium network employs the DIAQ scheme proposed in Chapter 6, and the AeNB is using the stateless Q-learning algorithm from Subsection 4.1.2. This scheme represents a state-of-the-art distributed RL solution to the spectrum sharing problem.
- "DRIAQ + RAQ-AC" - the combination of novel HARL based schemes developed in Section 7.2 and summarized in Algorithms 2 and 3.

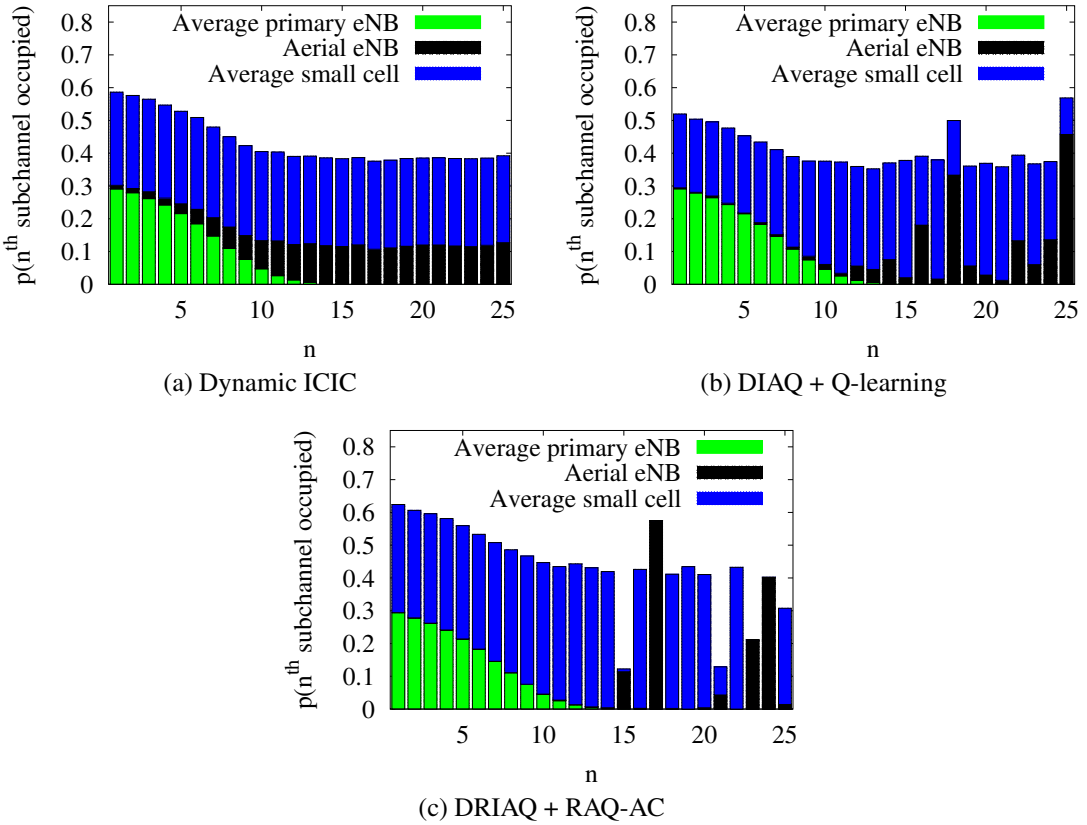
500 UEs are randomly distributed outside the stadium, in the circular area from the stadium boundary (5 m from the radius of the last row) to 1.5 km away from the stadium centre point. 25% of the stadium capacity is filled with randomly distributed

wireless subscribers, i.e. 10,776 UEs on average. The offered traffic is 20 Mb/s outside of the stadium and 1 Gb/s inside. All simulations last 2,000,000 transmissions, most of which take place inside the densely populated stadium. This corresponds to ≈2 hours.

7.3.1 Spectrum Occupancy Analysis

Figure 7.4 shows the subchannel occupancy distribution of the PeNBs, the AeNB, and the small cell eNBs using three different spectrum sharing strategies described in the beginning of this section. The distributions are calculated by measuring the amount of time every eNB spends occupying every subchannel and dividing it by the total simulation time.

Figure 7.4a shows that in the case of "dynamic ICIC" implemented in all systems, the reverse relationship between the spectrum mostly used by the AeNB and that preferred by the primary system is observed, demonstrating the effect of frequent ICIC



(a) Dynamic ICIC

(b) DIAQ + Q-learning

(c) DRIAQ + RAQ-AC

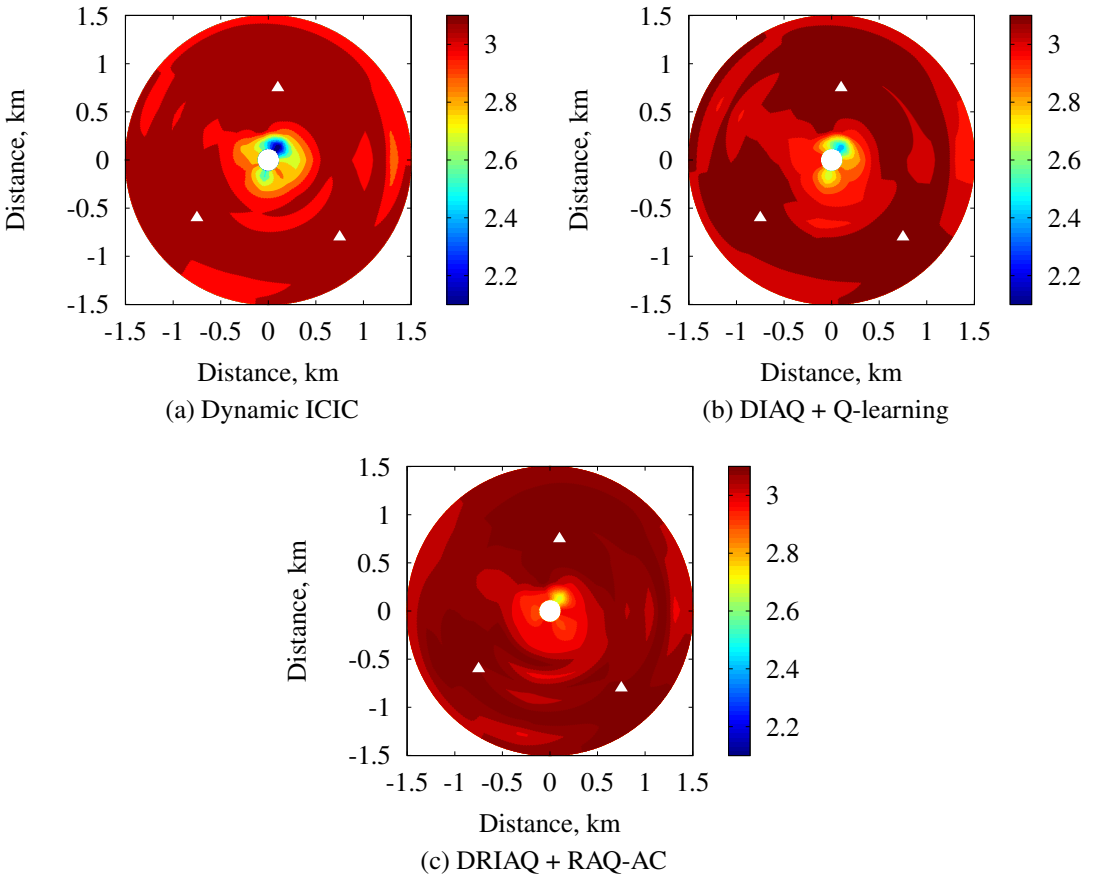
Figure 7.4: Subchannel occupancy of primary eNBs, aerial eNB and small cells using different spectrum sharing schemes

signalling for interference avoidance between the two. It also shows that the small cell network uses the whole spectrum approximately uniformly. Figure 7.4b demonstrates the difference made by introducing distributed Q-learning into the DSS process. The two challenging spectrum sharing relationships associated with this scenario tend to be addressed through distributed machine intelligence. The AeNB learns to avoid using the primary spectrum more than the "dynamic ICIC" approach, whilst the small cell eNBs tend to learn to use the subchannels preferred by the AeNB less than the others, i.e. they learn to avoid interfering with the AeNB, since it often results in blocked and interrupted file transmissions.

Figure 7.4c shows how the novel heuristically accelerated approach further improves the autonomously emerging spectrum sharing pattern by guiding the learning process of the AeNB to avoid interfering with the PeNBs, and discouraging the small cell eNBs from exploring and assigning the subchannels frequently used by the AeNB. Firstly, there is no overlap in the spectrum used by the AeNB and the PeNBs. Secondly, the AeNB uses fewer subchannels (less spectrum), since the small cells successfully adapt their policies to avoid using the AeNB's most preferred subchannels. This in turn positively reinforces the use of the same subchannels by the AeNB through the stateless Q-learning algorithm.

7.3.2 Primary User Quality of Service

Figure 7.5 shows contour plots of the spatial distribution of user throughput (UT) across the area outside of the stadium, covered by the PeNBs and the AeNB. They indicate that the area most susceptible to harmful interference is that in the vicinity of the stadium, where the UEs are connected to the AeNB as well as the PeNBs. There is also interference radiating from the ultra-dense stadium small cell network. Fig. 7.5a shows that the "dynamic ICIC" approach, with a relatively even spectrum occupancy distribution seen in Figure 7.4a, performs poorly and results in a significant decrease in UT in the vicinity of the stadium. Such performance degradation of the UEs located outside of the stadium is unacceptable from the viewpoint of secondary spectrum sharing. A significant improvement in the spatial UT distribution is achieved by using the learning based "DIAQ + Q-learning" approach. The performance is further improved



(a) Dynamic ICIC

(b) DIAQ + Q-learning

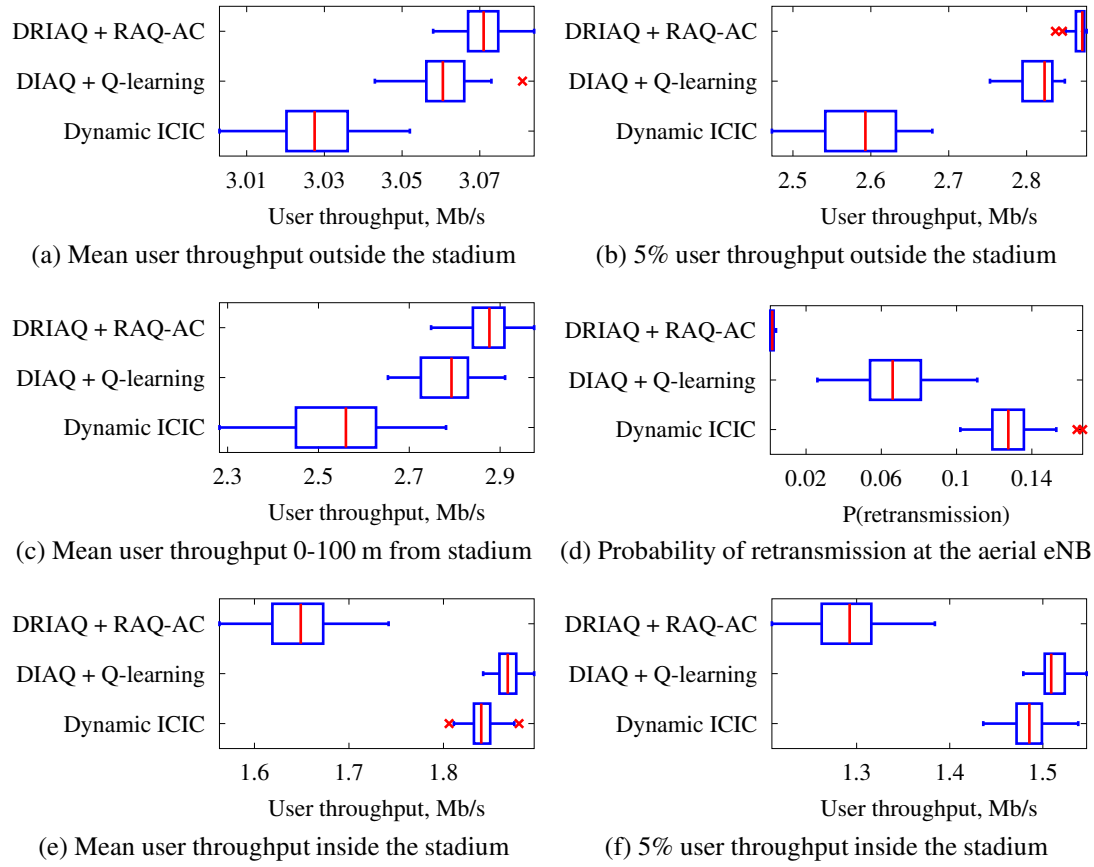
(c) DRIAQ + RAQ-AC

Figure 7.5: Spatial distribution of user throughput (Mb/s) outside of the stadium (the triangles represent the primary eNB locations)

by using the novel "DRIAQ + RAQ-AC" approach proposed in this chapter due to its ability to autonomously achieve the significantly more adaptable spectrum partitioning patterns seen in Figure 7.4c.

7.3.3 Statistical Analysis

The results in Figure 7.6 break down the QoS provided to the primary and secondary system users using the three different DSS strategies. Furthermore, they also verify the statistical significance of performance improvements gained by using the HARL based "DRIAQ + RAQ-AC" scheme proposed in Section 7.2. It shows the results from 50 different simulation setups, i.e. with different random seeds, UE locations and initial traffic, in the form of box plots [59], a compact way of depicting key features of probability distributions. The box boundaries represent the first and third quartile of the distribution, the line between them marks the median result, and the whiskers



(a) Mean user throughput outside the stadium

(b) 5% user throughput outside the stadium

(c) Mean user throughput 0-100 m from stadium

(d) Probability of retransmission at the aerial eNB

(e) Mean user throughput inside the stadium

(f) 5% user throughput inside the stadium

Figure 7.6: Boxplots of the primary and secondary system performance from 50 different simulations

show the minimum and the maximum point within $1.5 \times IQR$ distance from the box boundaries. IQR is the inter-quartile range, the difference between the first and third quartile (the width of the box). Any results further than $1.5 \times IQR$ away from the box are considered the outliers and are plotted as individual data points.

Figure 7.6a shows that the variation in mean UT outside the stadium is negligibly small, when comparing different DSS strategies. However, the box plots of 5% UT outside the stadium in Figure 7.6b reveal a more significant difference in the performance of the simulated DSS schemes. 5% UT for a single simulation is obtained by calculating the 5th percentile of the UT values of 500 users outside the stadium. It is a more important metric than the mean UT, since it represents a minimum QoS guaranteed to 95% of the users, and thus shows how fair the spatial QoS distribution is. Introducing the learning algorithms into the spectrum sharing strategies ("DIAQ + Q-learning") results in an 8.9% increase in median 5% UT outside the stadium compared to "dynamic ICIC", whereas the novel "DRIAQ + RAQ-AC" scheme improves it by

11%. These improvements are statistically significant since there is no overlap between the boxes in the plot. The same improvement pattern is observed in Figure 7.6c which shows the mean UT of the users located in the vicinity of the stadium (0-100m from the boundary), the region most vulnerable to the interference between the small cell network, the AeNB and the PeNBs.

Figure 7.6d demonstrates the most notable performance improvement achieved by "DRIAQ + RAQ-AC". It almost entirely eliminates the retransmissions, i.e. the blocked and interrupted file transmissions, at the AeNB. It results in a 98% decrease in the probability of retransmission compared to "dynamic ICIC" and a 97% decrease compared to a significantly better "DIAQ + Q-learning" scheme. This improvement is achieved due to the high controllability of the exploration process provided by the heuristic functions designed in Section 7.2. They successfully steer the learning process of the AeNB such that it avoids interfering with the PeNBs, whereas the small cell eNBs are continuously discouraged from occupying the resources preferred by the AeNB, as demonstrated by the spectrum occupancy patterns in Figure 7.4c.

Figures 7.6e and 7.6f show that the improvements in QoS, provided by the "DRIAQ + RAQ-AC" scheme to the PeNB and AeNB users, come at the cost of a 10-12% decrease in mean UT and a 13-14% decrease in 5% UT provided to the small cell users, compared with the two baseline schemes. However, this concession made by the stadium small cell network is relatively insignificant and essential in the context of dynamic secondary spectrum sharing. It results in the increased feasibility of secondary LTE spectrum reuse by a temporarily deployed eNB on an aerial platform and an ultra-high capacity density stadium small cell network, that is able to accommodate a vast increase in capacity (1 Gb/s in addition to the primary system's 20 Mb/s offered traffic). Furthermore, the "DRIAQ + RAQ-AC" scheme achieves remarkable reliability of AeNB communications (due to the lack of retransmissions). For example, this could be highly useful in the temporary event scenario for providing a robust dedicated access network to event organizers both inside and outside the stadium.

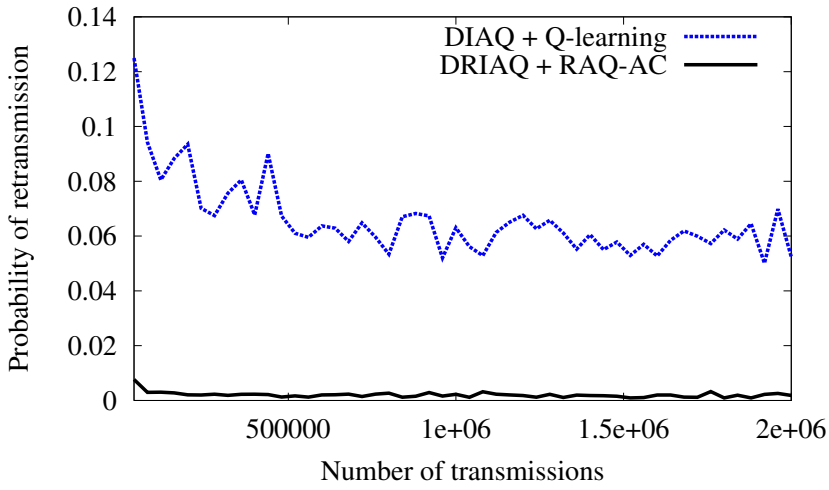


Figure 7.7: Probability of retransmission time response at the aerial eNB

7.3.4 Temporal Performance

Figure 7.7 shows the temporal performance of the two learning based schemes, "DIAQ + Q-learning" and "DRIAQ + RAQ-AC", in terms of the probability of retransmission at the AeNB. All data points are obtained by averaging over 50 different simulations. The time response of "DIAQ + Q-learning" demonstrates that it behaves as a classical RL algorithm, i.e. starts at a relatively poor performance level and gradually improves over time, while the AeNB and the small cell eNBs are learning appropriate spectrum sharing patterns. In contrast, the "DRIAQ + RAQ-AC" time response is a great demonstration of the improvements in the adaptability of cognitive eNBs achieved by introducing heuristic acceleration into the learning process. It starts at a superior probability of retransmission level and maintains it throughout the whole simulation.

7.4 Conclusion

The HARL based framework proposed in this chapter utilises a REM as external information for guiding the learning process of cognitive cellular systems, which are thus able to reuse the LTE spectrum owned by another cellular network. The performance of the DSS and DSA schemes developed in this chapter is assessed using system level simulations of the stadium temporary event scenario described in Subsection 3.1.1. It involves an eNodeB on an aerial platform, a small cell stadium network and a local

primary LTE network, all sharing the same LTE spectrum. Two novel DSS schemes are described in detail - distributed REM and ICIC accelerated Q-learning (DRIAQ) used by the small cell network, and REM accelerated Q-learning with Q-value based admission control (RAQ-AC) used by the aerial eNodeB. These schemes are shown to achieve high controllability of spectrum sharing patterns in a fully autonomous way. They also result in a significant decrease in primary system QoS degradation due to the interference from the secondary cognitive systems, compared to a state-of-the-art RL solution and a purely heuristic typical LTE solution. The spectrum sharing patterns that emerge by using the proposed schemes also result in remarkable reliability of the cognitive aerial eNodeB due to a 97% decrease in the probability of retransmission compared to a classical RL approach.

Furthermore, the novel principle of superimposed heuristic functions proposed in the context of HARL, as well as the general Q-table mask structure of these functions, are not specific to the investigated spectrum sharing scenario, and are generally applicable to a wide range of self-organization problems beyond the wireless communications domain.

Chapter 8. Case-Based Reinforcement Learning for Dynamic Environments

Contents

8.1 Motivation

All DSA algorithms and simulation experiments discussed in this thesis so far only consider static environments, i.e. environments with the same network topologies and the same traffic load levels and distributions. However, the vast majority of real-world

wireless environments are likely to be dynamic, e.g. with variable traffic load distributions and/or network topologies depending on the time of the day. The purpose of this chapter is to assess the performance of distributed RL based DSA algorithms in such dynamic environments, and to propose a way of improving their stability and adaptability. The technique investigated for solving this problem is case-based reinforcement learning (CBRL), a combination of RL and case-based reasoning (CBR). CBR is broadly defined as the process of solving new problems by using the solutions to similar problems solved in the past [95]. In CBRL these solutions are obtained through an RL algorithm.

This combination of RL and CBR has been successfully applied to various decision problems, e.g. dynamic inventory control [40], RoboCup Soccer [15] and control of a simulated mountain car [9]. For example, Jiang and Sheng [40] propose an effective dynamic inventory control algorithm that uses CBR for analysing the similarity between different states of a dynamic multi-agent RL problem. In [9] and [15] the authors develop transfer learning algorithms that transfer knowledge between similar learning tasks whilst using CBR to make this process faster. There appears to be no evidence in the literature of the CBRL approach being applied in the wireless communications domain.

8.2 Dynamic Wireless Environments

A key challenging aspect of the wireless environment considered in this chapter is its dynamic nature due to the variable network topology. The stadium small network introduced in Subsection 3.1.1 adapts its topology to temporal non-uniform variations in the traffic load. In the full secondary spectrum sharing scenario, the dynamic nature of the environment is also caused by periodic deployments of the AeNB. All of these paradigms are explained in more detail in the following subsections.

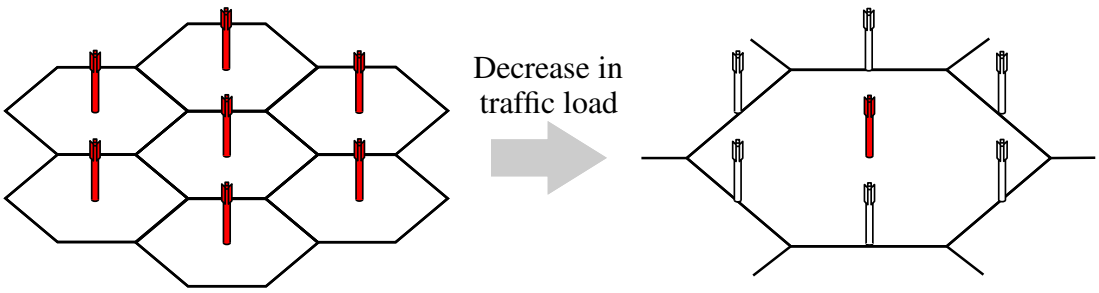


Figure 8.1: A simple topology management case, where a number of eNBs are switched off after a decrease in the overall traffic load

8.2.1 Dynamic Topology Management

Topology management is an increasingly popular area of research, particularly in green communications, where a trade-off between the QoS provided to the users and the energy savings of the network is achieved by dynamically switching various base stations on/off, e.g. [57][70]. A simple illustrative example discussed in [57] is portrayed in Figure 8.1. It involves a classical hexagonal cell layout, where all base stations surrounding the middle one temporarily enter a sleep mode at times when the traffic load is lower, e.g. night time. The users from all seven cells can then be served by the middle base station that would expand its coverage area accordingly. Employing such topology management schemes can result in significant energy savings, since a major part of energy in telecommunications systems is consumed by base stations [57][72].

8.2.2 Dynamic Non-Uniform Traffic Load

Another source of the network's dynamic nature considered in this study is the presence of a dynamically moving traffic hotspot area. For example, a rapid increase in the traffic load in a specific part of the stadium small cell network may be observed if a particular event happens close to the given area, e.g. teams walking out at the opening ceremony of the Olympic Games or a goal at a football match etc. In such cases, the topology management algorithm would cause the network to be fully switched on in the hotspot area (left side of Figure 8.1), and only partially deployed in other areas of lower traffic intensity (right side of Figure 8.1). Furthermore, the experiments described in this chapter assume that the geographical location of this hotspot area varies with time, making the wireless environment asymmetric and dynamic in both

the offered traffic distribution and the network topology.

8.2.3 Rapidly Deployable Aerial Platform

The full secondary spectrum sharing scenario described in Subsection 3.1.1 also involves a local primary LTE network and a cognitive eNB on an aerial platform (AeNB) for wide area coverage, all sharing the same LTE spectrum with the stadium small cell network. The AeNB can be switched on and off several times throughout the duration of the event [71]. For example, it can be switched on for providing the event organizers with a dedicated access network when required, and switched off to have its batteries recharged or to minimise the energy consumption in general. Therefore the additional challenge faced by the cognitive small cell eNBs is to adapt to these sudden changes in their radio environment, while not affecting the QoS in the local primary system.

8.3 Distributed Case-Based Q-Learning

The technique investigated in this chapter for enhancing the stability of RL based DSA algorithms under challenging dynamic conditions of wireless environments is case-based RL (CBRL). Its general principles are introduced in the following subsection.

8.3.1 Case-Based Reinforcement Learning

CBRL is a combination of RL and case-based reasoning (CBR), where the solutions to previously known problems are used to help learning solutions to new problems [95]. Figure 8.2 shows a flow diagram of the processes involved in CBRL. It also demonstrates that it is an extension of classical single-agent RL, i.e. the latter can be viewed as a special case of CBRL.

The unfilled blocks and solid lines in Figure 8.2 constitute a flow diagram of a classical RL algorithm introduced in Figure 2.6. There is an outer output-state-action loop, where outputs of the environment are observed and processed to yield the environment state information, and the best action is chosen for the current state based on the policy

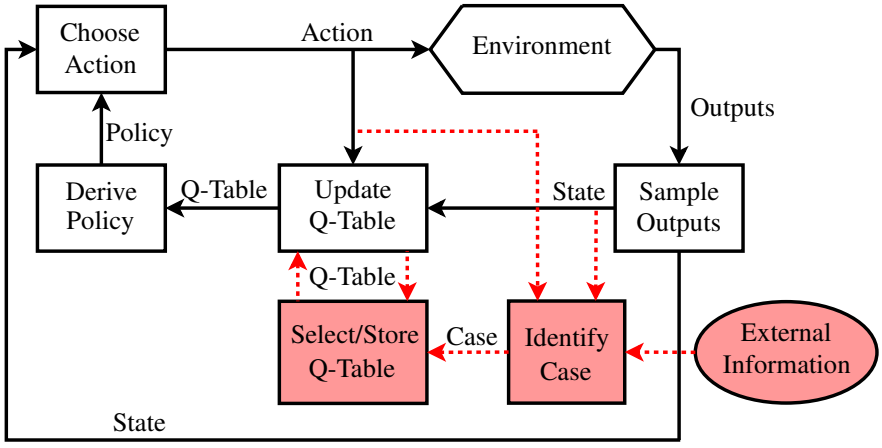


Figure 8.2: Block diagram of case-based reinforcement learning

of the learning agent. In the context of DSA problems, the output of interest is whether or not a given transmission is blocked, interrupted or successfully completed, and the action is a spectrum resource allocated to it. There is also an inner learning loop, whose role is to learn a good policy to be used by the learning agent. It achieves this goal by observing the actions taken by the learning agent and their outcomes, and directly estimating the entries in the Q-table, e.g. using Equation (4.4) in the case of stateless Q-learning. A policy is then derived from the estimated Q-table and used for choosing an action in the current environment state, e.g. as shown in Equation (4.3).

The highlighted blocks and dotted arrows represent additional functionality afforded by CBR to enable the system to learn several solutions to different cases of the environment at once. It introduces another parallel inner loop which continuously observes the input/output relationship of the environment, and identifies its current model or case. It may also have access to other information supplied from elsewhere to aid the identification process. The idea is that for different cases of the environment the estimated models will be sufficiently different to be detected by the identification algorithm, and for every identified model of the environment there will be a stored Q-table associated with it. In this way, a CBRL algorithm always knows what phase the environment is currently in and is always able to use a Q-table most suitable for it.

8.3.2 Case Identification

A crucial part of the CBRL process is an appropriate mechanism for case identification, such that the dynamically changing environment could be described by a finite number of distinct configurations, i.e. cases. All changes in the network environment described in Section 8.2 involve changes in the network topology, e.g. triggered by the temporally and spatially variable traffic load or the periodically deployed eNB on the aerial platform. Therefore, case identification based on network topology is proposed in this section.

In order to limit the potential number of identifiable topology cases and to make this approach scalable and generally applicable to any cellular system, the proposed topology identification process is localised to the second order neighbourhood (2ON) of a given eNB. We define the 2ON of an eNB as the set of its neighbouring eNBs and all their neighbouring eNBs as illustrated in Figure 8.3 for a generic hexagonal cell layout.

The 2ON based topology identification process is localised enough to be scalable and generally applicable in arbitrary cellular networks, yet not too limited to disregard valuable information about the radio environment surrounding a given eNB. To use the example in Figure 8.3, the spectrum management policy of the middle eNB will

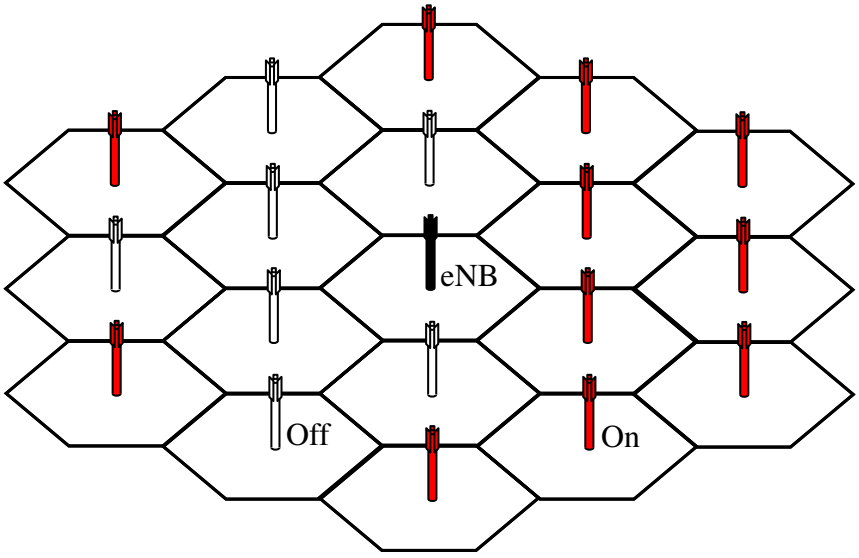


Figure 8.3: Example of a second order neighbourhood used for case identification by the middle eNodeB

be heavily influenced by the on/off configurations of its neighbouring eNBs with their own cognitive spectrum management policies. Equivalently, the latter will be significantly influenced by all of their own neighbouring eNBs, thus potentially having a noticeable impact on the original middle eNB. It is possible to extend this argument to higher orders of neighbouring eNBs, however, their impact on the original eNB in question is likely to be diminishing. In future adaptations of the approach proposed in this chapter further neighbourhoods of eNBs, up to the whole network topology, can also be included in the case identification process without the loss of generality.

The author proposes expressing the on/off configurations of a given eNB's 2ON as a binary string, each bit corresponding to a particular eNB in the 2ON. For example, the following binary string would be used to describe the asymmetric topology case surrounding the middle eNB in Figure 8.3:

$$T_{2ON} = 1010000100111111111_2 \tag{8.1}$$

where T_{2ON} is the binary string describing the network topology surrounding the given eNB. The order of the bits in T_{2ON} corresponds to the sequence of the eNBs in the 2ON depicted in Figure 8.3 counting from the left-hand column of eNBs downwards and excluding the middle eNB itself. Every eNB is assumed to have access to the information about the on/off configuration of its 2ON through a small-scale periodically broadcast radio environment map (REM), which is one of the key features of intelligent cognitive cellular systems [60].

8.3.3 Case Retrieval

Another fundamentally important function that has to be performed by an intelligent CBR agent is case retrieval, i.e. selecting a solution, e.g. a Q-table, that corresponds to the most appropriate stored case to be used at any given moment as shown in Figure 8.2. To facilitate this functionality, a method for comparing a currently identified case with the stored cases and calculating a degree of similarity between them is required. Since every case is expressed in terms of the on/off configuration of the 2ON of a given eNB as shown in the example in Equation (8.1), the similarity measure between any

two cases is defined as the number of eNBs in the 2ON with the same on/off status. In order to calculate it, first, a binary string T_{same} indicating which eNBs in the 2ON are active/idle in both compared cases must be obtained. It is done by performing a bitwise exclusive NOR operation between the binary strings describing the current case $T_{2ON}^{current}$ and one already stored in the case base T_{2ON}^{stored}:

$$T_{same} = \overline{T_{2ON}^{current} \oplus T_{2ON}^{stored}} \tag{8.2}$$

The similarity measure β is then defined as the number of eNBs in the 2ON that have the same active/idle status, thus representing how similar two 2ON topologies are. It is calculated by adding up the bits in T_{same} as follows:

$$\beta = \sum_{n=1}^{N} T_{same}(n) \tag{8.3}$$

where $T_{same}(n)$ is the n^{th} bit of T_{same}, and N is the number of eNBs in the 2ON.

In this way, for any currently identified case the retrieval function will return a stored case using the following principle:

$$\hat{k} = \underset{k}{\operatorname{argmax}}(\beta_k),\ k \in \{1, 2, ..., K\} \tag{8.4}$$

where $\hat{k}$ is the index of the retrieved case, β_k is the similarity measure between the k^{th} stored case and the currently identified case, and K is the total number of stored cases.

8.3.4 Multi-Criteria Case Identification

The case identification and retrieval technique described in this section so far only considers the topology of a homogeneous network, e.g. it can be applied to an isolated stadium small cell network from Figure 3.1. However, if the secondary spectrum sharing scenario from Subsection 3.1.1, which also involves a dynamically deployable AeNB, is considered, then the network environment becomes heterogeneous and an extension to the proposed case identification and retrieval framework is required.

The presence/absence of an entity such as the wide area coverage AeNB in the net-

work environment can be viewed as a separate major criterion for case identification, in addition to localised homogeneous topologies as shown in Figure 8.3. Therefore, the author proposes a bias variable β_{bias} introduced into the case similarity assessment formula given in Equation (8.3), such that the cases with the same AeNB status are recognised as more similar to each other than those with a different AeNB status. The presence/absence of the AeNB is chosen to be a primary criterion for case identification and retrieval, since it represents a significantly more substantial change in the radio environment than changes in the active/idle mode of an eNB's local 2ON from Figure 8.3. Therefore, the extended multi-criteria similarity measure formula is the following:

$$\beta = \sum_{n=1}^{N} T_{same}(n) + \beta_{bias} \tag{8.5}$$

where the bias variable $\beta_{bias} > N$, i.e. a value higher than the maximum possible unbiased similarity measure, when the AeNB status of the two given cases is the same, and $\beta_{bias} = 0$ otherwise.

8.3.5 The Case-Based Q-Learning Algorithm

Algorithm 4 summarises the steps of the proposed case-based Q-learning approach to DSA in dynamic wireless environments. The extra functionality specific to CBR is described by steps 5, 6, 7 and 11, i.e. if these steps are taken out, the algorithm simplifies down to classical stateless Q-learning described in Subsection 4.1.2.

Algorithm 4 Subchannel assignment using case-based Q-learning in dynamic cellular environments

1: Wait for a file arrival
2: **if** all subchannels are occupied **then**
3: Block transmission
4: **else**
5: Identify current case k
6: Find most similar stored case $\hat{k}$ using Equation (8.4)
7: Retrieve Q-table $Q(a)$ associated with $\hat{k}$
8: Assign a subchannel using $Q(a)$ and (4.3)
9: Observe the outcome, calculate the reward $r = \pm 1$
10: Update $Q(a)$ using Equation (4.4)
11: Store $Q(a)$ in case base, associate it with k
12: **end if**

8.4 Simulation Results

The simulation experiments discussed in this section consider both the scenario where the stadium small cell network has exclusive access to a 20 MHz LTE channel, and the full spectrum sharing scenario described in Subsection 3.1.1 which also involves an AeNB and a primary system of local eNBs all sharing the same 20 MHz LTE spectrum.

The primary system is assumed to employ the same dynamic ICIC scheme as that used in the simulation experiments in Sections 4.3 and 7.3. There, all three PeNBs exchange their current spectrum usage as RNTP messages every 20 ms, and exclude the subchannels currently used by the other two PeNBs from their available subchannel list. However, the CBRL scheme proposed in Algorithm 4 does not assume this and would also work regardless of the spectrum management strategy of the primary system.

The results of implementing the following five schemes in the secondary cognitive system are discussed in this section:

- *"Dynamic ICIC"* - all systems use ICIC signalling as described in Subsection 3.3.1 and above for the primary system. The stadium eNBs receive ICIC messages from the AeNB and from their neighbouring small cells. They only report subchannels used at a Tx power above -3 dB with respect to the average power in the cell, and choose randomly among the subchannels deemed "safe". The AeNB randomly assigns subchannels not used by the primary system, based on the ICIC messages of the latter. This approach represents a heuristic baseline DSA scheme, typical for LTE networks [79].
- *"Reinforcement learning (RL)"* - the AeNB and the stadium small cells run the distributed Q-learning algorithm introduced in Subsection 4.1.2.
- *"Case-based reinforcement learning (CBRL)"* - the AeNB is still running classical stateless Q-learning, whereas the stadium small cells run the distributed case-based Q-learning algorithm proposed in this chapter and summarised in Algorithm 4.

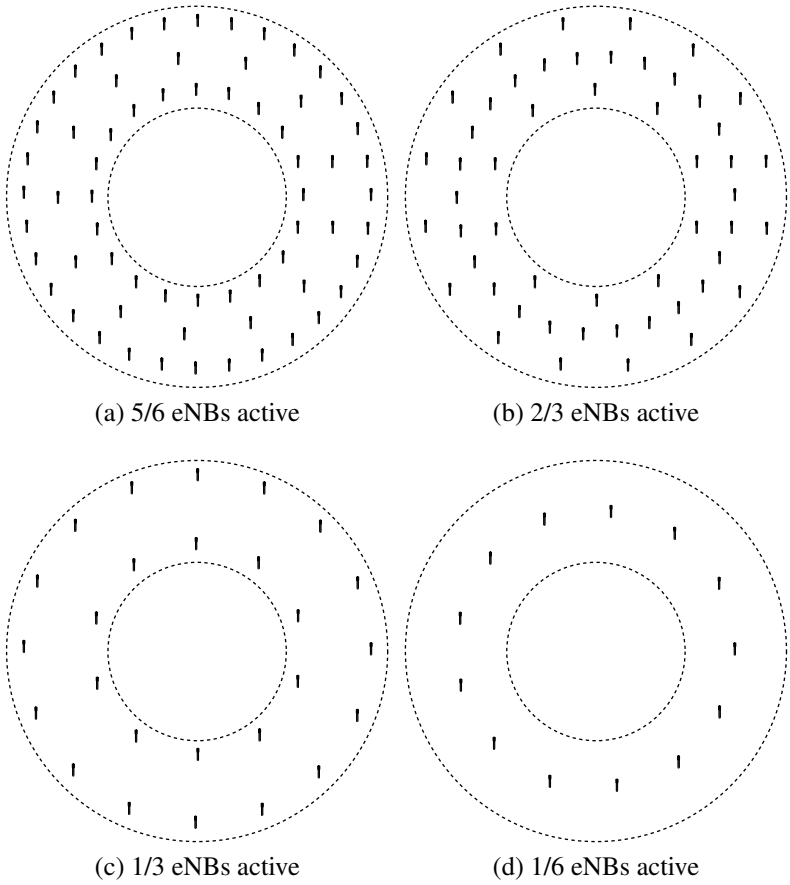
- *"Heuristically accelerated reinforcement learning (HARL)"* - the HARL based schemes proposed in Chapters 6 and 7, i.e. the DIAQ scheme in the scenario with the stadium small cell network only, and "DRIAQ + RAQ-AC" in the full spectrum sharing scenario exactly as used in Chapter 7.
- *"Case-based heuristically accelerated reinforcement learning (CBHARL)"* - The HARL based schemes proposed in Chapters 6 and 7 augmented with the CBR functionality proposed in this chapter. This is achieved by replacing the regular Q-table $Q(a)$, used in step 8 of Algorithm 4 for making a spectrum assignment decision, by the masked Q-table $Q_m(a)$ which takes into account the heuristic ICIC and/or REM information as described in Section 7.2. This approach combines all schemes proposed in this thesis in one and, therefore, represents the entire contribution made by this thesis towards improving the adaptability and robustness of distributed RL based DSA.

25% of the stadium capacity is filled with randomly distributed wireless subscribers, i.e. 10,776 UEs on average. In the full spectrum sharing scenario 500 UEs are randomly distributed outside the stadium in a circular area from the stadium boundary out to 1.5 km from the stadium centre point, producing the total offered traffic of 20 Mb/s.

8.4.1 Topology Management

Figure 8.4 shows how the principle of traffic load dependent dynamic topology management described in Subsection 8.2.1 is adapted to the stadium small cell network used in simulation experiments in this chapter. The following relationship between the network-wide offered traffic density (OTD) and the topology patterns from Figure 8.4 was experimentally found to achieve an appropriate trade-off between the number of eNBs switched off for potential energy savings and the QoS provided to the users:

- all eNBs are active if OTD $>$ 27 Gbps/km^2
- 5/6 eNBs are active if OTD $\in$ (21, 27] Gbps/km^2
- 2/3 eNBs are active if OTD $\in$ (15, 21] Gbps/km^2
- 1/3 eNBs are active if OTD $\in$ (8, 15] Gbps/km^2



(a) 5/6 eNBs active (b) 2/3 eNBs active

(c) 1/3 eNBs active (d) 1/6 eNBs active

Figure 8.4: Traffic load based partial deployments of the stadium small cell network (centralised topology management)

- 1/6 eNBs are active if OTD $\leq$ 8 Gbps/km^2

In this way the stadium network is able to provide adequate QoS to the users across a wide range of traffic loads, whilst achieving significant energy savings when the offered traffic is low by employing these partial small cell network deployments.

8.4.2 Dynamic Traffic Hotspot Area

Another feature of the simulation scenario investigated in this chapter is the presence of a traffic hotspot area within the stadium that changes its geographical location with time. An example of such a hotspot area and its effect on the topology of the stadium network is shown in Figure 8.5. If an increased user activity in the 60 degree sector is observed, while the offered traffic density is lower elsewhere, the topology management algorithm detects the possibility of deploying all available eNBs in the

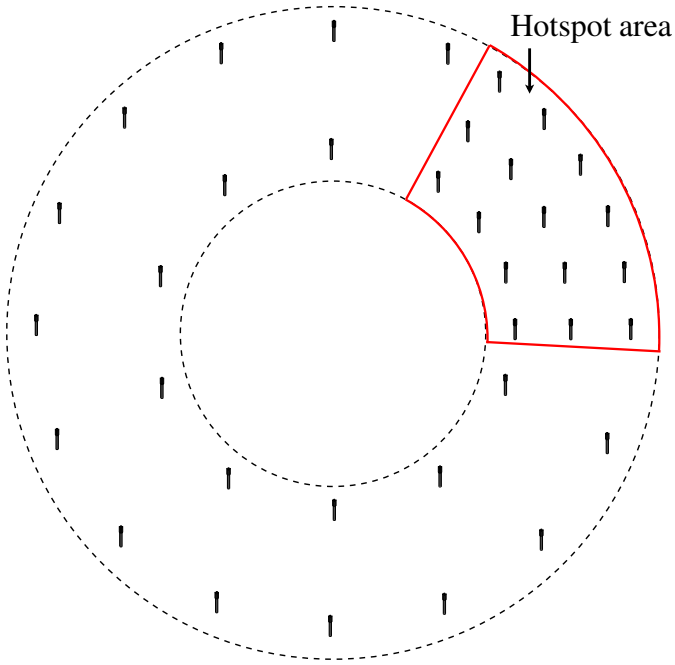
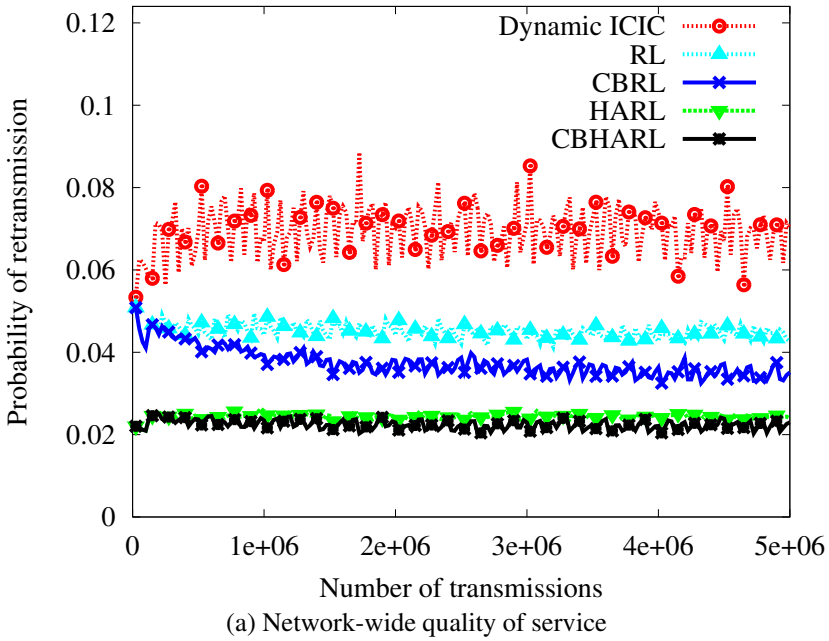


Figure 8.5: Asymmetric network topology due to a local traffic hotspot area

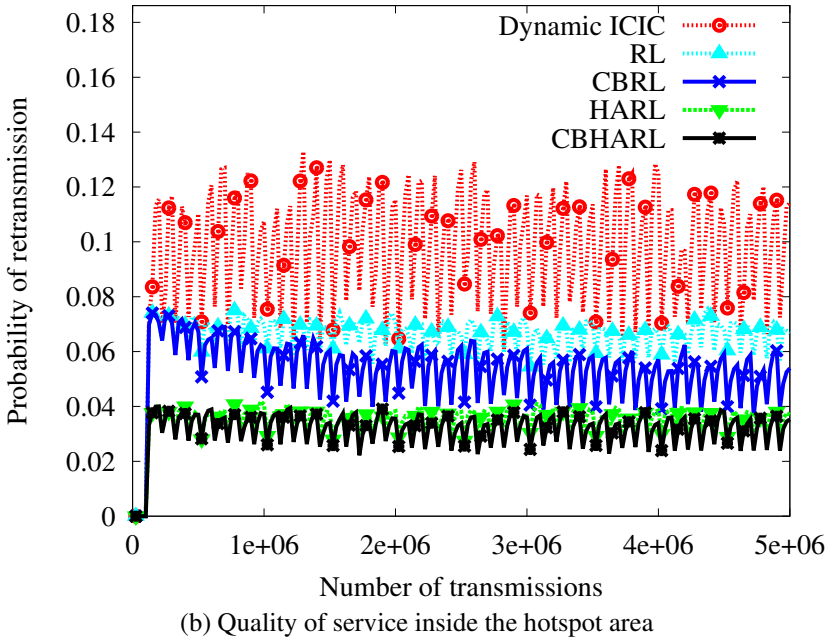
hotspot area and keeping a number of them switched off according to one of the partial deployment patterns from Figure 8.4.

Figure 8.6 shows the probability of retransmission time response in the stadium small cell network inspected individually with its own dedicated spectrum (20 MHz LTE channel). The location of the 60° hotspot area is randomly changed every 100,000 transmissions to one of its six possible locations - {0°, 60°, 120°, 180°, 240°, 300°}. The offered traffic density within the hotspot is 34 Gbps/km^2, and 13 Gbps/km^2 elsewhere. The topology management algorithm is assumed to detect a change in the offered traffic distribution with a delay of 5000 file transmissions. All eNBs within 15° of the edges of the user hotspot area are switched on to make sure it is covered by fully deployed small cells. The plots are obtained by averaging every data point using the results from 50 simulations with different randomly generated UE locations and initial traffic.

Firstly, all RL based schemes significantly outperform the dynamic ICIC approach, demonstrating the effectiveness of applying distributed RL to DSA in cellular networks. Secondly, although the classical RL and CBRL schemes start at the identical QoS level, the latter goes on to gradually improve its performance due to its increased adaptability in the dynamic environment. In contrast, the classical RL process is dis-



(a) Network-wide quality of service



(b) Quality of service inside the hotspot area

Figure 8.6: Probability of retransmission of the small cell stadium network with a dynamically moving traffic hotspot

turbed by the environment changes frequently enough not to show any notable performance improvement over time. As a result, by the end of the simulation the proposed case-based Q-learning scheme shows an ≈22% reduction in the network-wide number of retransmissions shown in Figure 8.6a, compared with the classical Q-learning alternative. However, both plots in Figure 8.6 also shows that, if the ICIC signalling information is available to the cognitive eNBs, employing the HARL based DIAQ scheme proposed in Chapter 6 instead of the CBRL approach results in a far more

significant 50% reduction in the number of retransmissions both network-wide and inside the moving hotspot area only. Augmenting the HARL approach with CBR results only in a marginal further improvement in performance, demonstrating that the heuristic acceleration provided by ICIC signalling in the CBHARL approach plays a significantly more important role.

8.4.3 Temporal Network-Wide Traffic Variations

A further challenge introduced into the simulation experiments hereafter is the variable network-wide traffic load shown in Figure 8.7. These variations in the offered traffic density result in changes in the network topology according to the topology management scheme described in Subsection 8.4.1. Figure 8.8a shows the probability of retransmission time response of the stadium network with such uniform temporal variations in the network-wide traffic load. Due to the uniform nature and a lower number of possible topology cases compared to the dynamic traffic hotspot scenario from the previous subsection, the difference in performance between CBRL and classical RL is larger than that observed in Figure 8.6, especially at times shortly after the network topology transitions. Incorporating CBR into the learning process often results in as much as a two-fold reduction in the probability of retransmission.

Figure 8.8b shows the probability of retransmission time response of the stadium network both with uniform variations in the offered traffic density shown in Figure 8.7 and with the dynamically moving traffic hotspot area shown in Figure 8.5. There,

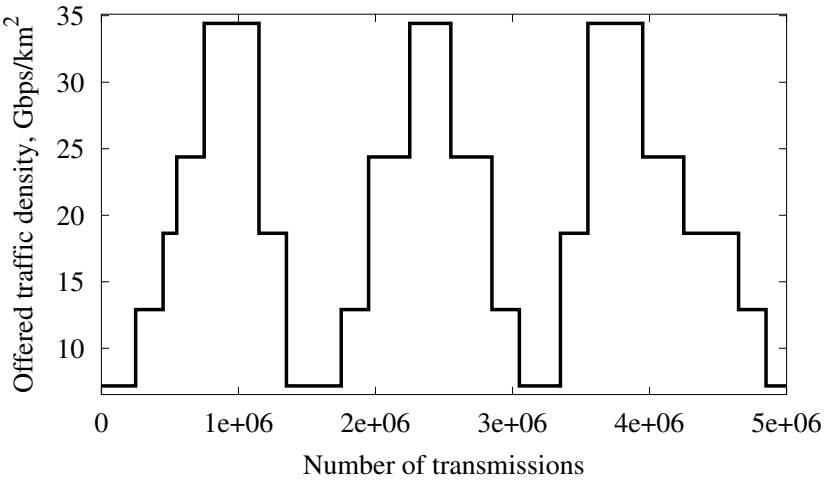
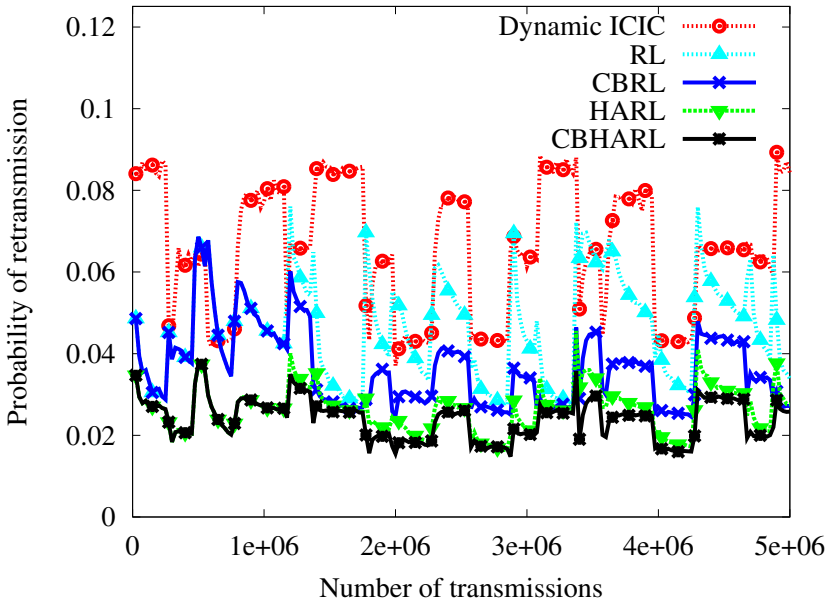
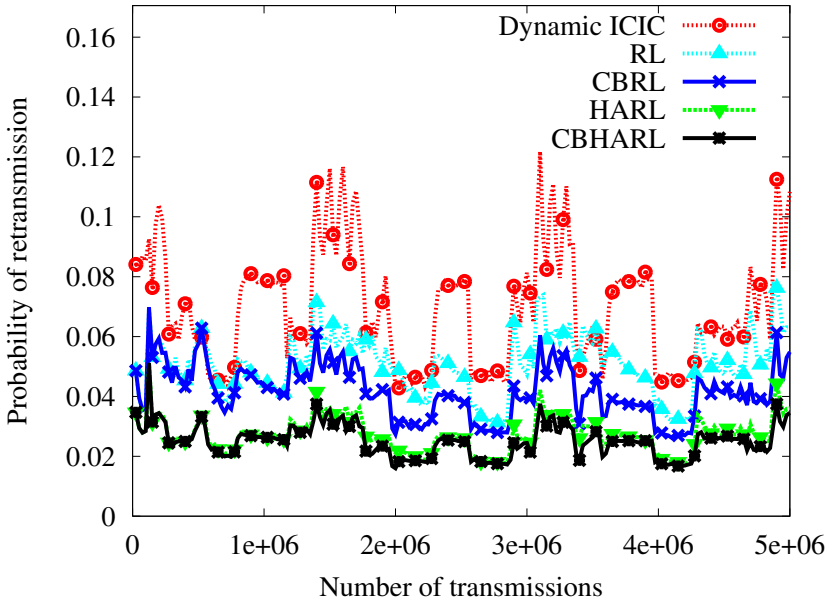


Figure 8.7: Temporal variations in the stadium network-wide offered traffic density



(a) Uniform traffic load



(b) With a dynamic hotspot area

Figure 8.8: Probability of retransmission of the stadium network with temporal variations in the network-wide offered traffic

in contrast to the results in Figure 8.8a, the increase in the complexity of the problem and the number of potential network topology cases reduces the magnitude of the performance improvements gained by CBRL compared to classical RL. Nevertheless, the CBR functionality is still able to provide a consistent noticeable decrease in the number of retransmissions experienced by the UEs in the stadium network.

Both plots in Figure 8.8 once again show that the HARL approach significantly outperforms the CBRL approach due to the availability of additional valuable spectrum

awareness information. However, the difference made by introducing CBR into the heuristically accelerated Q-learning approach is visibly bigger, yet still relatively small, in Figure 8.8a, where previously learned Q-tables are reused to improve the system stability and speed-up the process of adapting the policies of the small cell cognitive eNBs shortly after major changes in the network topology.

8.4.4 Spectrum Sharing with Dynamic Aerial eNB Deployment

The last set of simulation results discussed in this chapter considers the performance of both the primary and the secondary network in the full spectrum sharing scenario described in Subsection 3.1.1. In addition to the dense stadium small cell network, it involves an AeNB and a local network of PeNBs, all sharing the same 20 MHz LTE channel. The stadium small cell network includes both dynamic environment features investigated in the previous subsections:

- a dynamic 34 Gbps/km^2 offered traffic density area depicted in Figure 8.5
- an updated version of the temporal variations in the network-wide traffic load shown in Figure 8.9

The variable network-wide traffic loads are slightly lower than those used in the previous experiments and shown in Figure 8.7, since the 20 MHz LTE channel is no longer fully dedicated to the stadium network, but is shared with the primary system and the

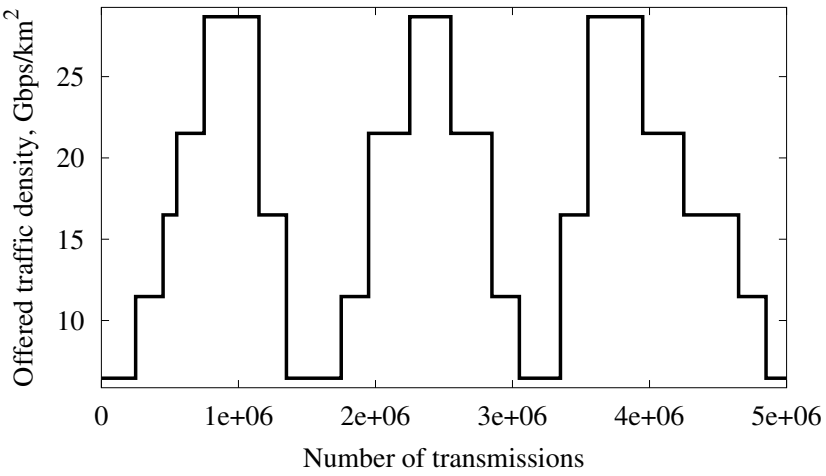
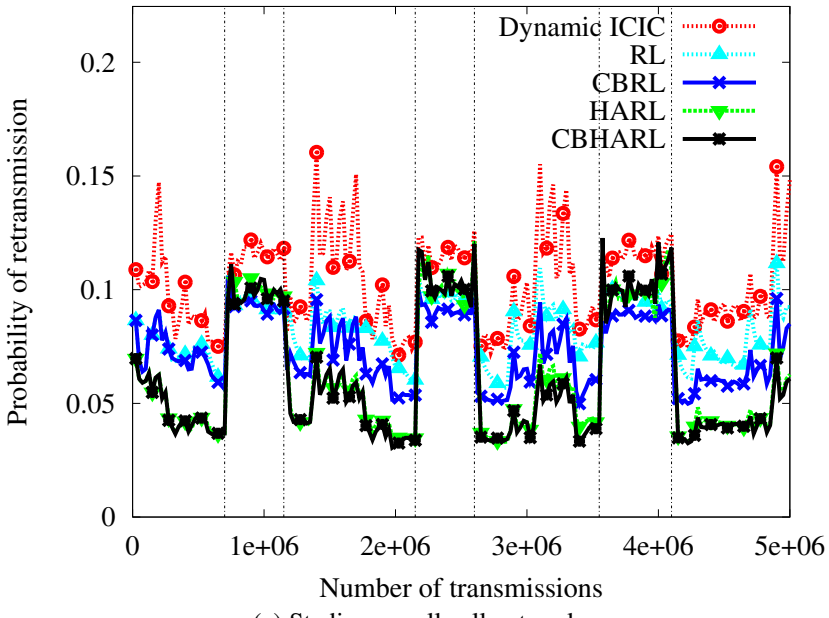


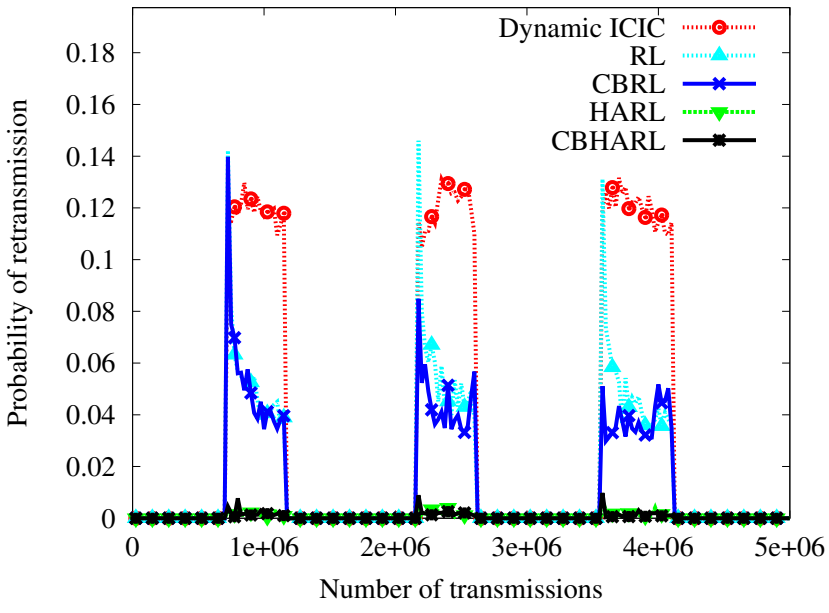
Figure 8.9: Temporal variations in the stadium network-wide offered traffic density in the full spectrum sharing scenario

cognitive AeNB. The latter is running a classical Q-learning algorithm described in Subsection 4.1.2 and is periodically deployed and redeployed into the network.

Figure 8.10 shows how the probability of retransmission changes over time in the two independent secondary systems involved in the spectrum sharing scenario - the stadium small cell network and the AeNB. All simulations start with the AeNB switched off, and the vertical dash-dot lines in Figure 8.10a mark the times when it is switched on and off again. It shows that the performance gap between case-based and classical



(a) Stadium small cell network



(b) Aerial eNB

Figure 8.10: Probability of retransmission of the stadium network and the Aerial eNB in a dynamically changing radio environment

Q-learning in the stadium network is further reduced due to an even more complicated scenario, the presence of an interfering primary network and a higher number of possible network topologies. However, Figure 8.10b shows that employing the CBRL approach in the stadium network dramatically improves the QoS provided by the AeNB shortly after it is switched on for the second and third time, compared to classical RL. This is due to the capability of cognitive small cell eNBs to distinguish between various network topologies, including whether or not the AeNB is switched on. In this way, the stadium small cells are able to revert their Q-learning DSA policies to those most appropriate for the AeNB to share spectrum with them, resulting in the QoS improvement in both of these secondary access networks.

At times when the AeNB is switched off and the small cell network shares the spectrum only with the local PeNBs, the performance of the HARL approach is still consistently better than that of the classical and the case-based RL approach. This confirms that even with the presence of the primary system interference the DIAQ scheme proposed in Chapter 6 produces the best QoS in the stadium small cell network of all schemes investigated in this thesis. However, when the AeNB is switched on, the probability of retransmission at the stadium network achieved by both HARL and CBHARL dramatically increases and becomes slightly higher than that achieved by classical and case-based RL. This is precisely the effect of efficient spectrum sharing patterns shown in Figure 7.4 achieved by the HARL based spectrum sharing schemes proposed in Chapter 7. The QoS decrease at the stadium network is caused by the heuristically accelerated policies of the small cell eNBs that avoid interference with the AeNB, as a result making the latter dramatically more reliable in terms of its own probability of retransmission. This dramatic improvement in the QoS provided to the AeNB users is observed in Figure 8.10b. Introducing CBR into the heuristically accelerated Q-learning approach has a negligible effect on the system performance due to the increased complexity of the learning problem in terms of the number of potential network topology cases. This confirms that the introduction of the heuristic acceleration into the learning process as proposed in Chapters 6 and 7 makes a significantly bigger contribution towards improving the system QoS and its adaptability, than the novel CBR functionality introduced in this chapter.

Table 8.1: Primary user quality of service (QoS) with and without the presence of the secondary network (SN)

QoS metric	**Without SN**	**With SN**
Mean user throughput (UT), Mb/s	3.03	3.07
95th percentile UT, Mb/s	3.16	3.16
5th percentile UT, Mb/s	2.76	2.91
Mean UT 0-100 m from the stadium, Mb/s	2.95	2.96

An essential requirement for cognitive wireless networks is to ensure that they do not have a harmful effect on the QoS in the primary system. Table 8.1 compares the QoS provided to the users outside of the stadium with and without the presence of the stadium users and the secondary network which runs the CBHARL algorithm, i.e. the final product of all contributions of this thesis. It describes the statistical distribution of user throughput (UT) achieved by the primary network.

Table 8.1 shows that the introduction of the secondary stadium network and the AeNB results in no degradation in the overall mean UT, the 5th and 95th percentile UT, and the mean UT provided to the primary users in the 100 m vicinity of the stadium. Interestingly, it even achieves an improvement in the 5th percentile UT, which represents the lowest UT provided to at least 95% of the users and which is an important metric for ensuring fair QoS distribution across the whole network. This is because the AeNB manages to provide higher quality opportunistic links to some primary users than those that could be provided by the local eNBs. The results in Table 8.1 emphatically show that it is possible to develop a temporary heterogeneous cognitive network that is capable of servicing a dramatic increase in the mobile data capacity (544 Mb/s overall throughput compared to 19.8 Mb/s in the primary system only) in a challenging dynamic radio environment, but with no need for additional spectrum and with no degradation in the primary user QoS.

8.5 Conclusion

The CBRL technique proposed in this chapter is an effective and feasible approach to DSA in cognitive cellular systems with dynamic topologies. Large-scale system level

simulations of a stadium small cell network with an asymmetric time-variant topology show that augmenting classical distributed stateless Q-learning with the CBR functionality in this way results in increased adaptability of the cognitive cellular system to changes in its radio environment. For example, it is capable of achieving a two-fold reduction in the number of retransmissions, compared to the classical RL approach, shortly after transitions between different network topologies. However, as the complexity of the dynamic environment and the possible number of network topologies increase, the performance gap between classical and case-based RL decreases. Nevertheless, the proposed distributed case-based Q-learning approach achieves a consistent improvement in the system QoS and its stability in the dynamic cellular environment considered. However, a far more significant contribution towards improving the QoS and its adaptability and robustness in cognitive wireless networks is achieved by the HARL based schemes proposed in Chapters 6 and 7. Therefore, if the heuristic spectrum awareness information used by these schemes is available to the cognitive eNBs, introducing CBR into the learning process has a small effect on system performance.

Simulations of a spectrum sharing scenario, where the stadium small cell network shares the same LTE channel with a cognitive AeNB and a local primary network, show that the CBHARL algorithm, i.e. the final product of all technical contributions proposed in this thesis, achieves a significant improvement in the QoS of the stadium network without the presence of the AeNB, and a dramatic improvement in the reliability of the AeNB at the cost of a small QoS decrease inside the stadium, compared to the classical RL algorithm. Furthermore, these simulations show that the cognitive cellular system that employs the CBHARL DSA scheme with only secondary access to an LTE channel, is able to accommodate a 28-fold increase in the total primary and secondary system throughput, but with no need for additional spectrum and with no degradation in the QoS of the primary users.

Chapter 9. Conclusions and Further Work

Contents

9.1 Conclusions

Cognitive wireless networks that employ flexible DSA methods are considered one of the key technologies for utilising the wireless spectrum efficiently and, thus, accommodating the ever increasing demand for mobile data capacity. Distributed RL is a powerful and widely used approach to DSA due to its capability to facilitate full self-organisation in wireless networks. It eliminates the need for the potentially challenging and time-consuming spectrum planning process carried out by human experts, whilst enabling the wireless networks to learn flexible and highly efficient spectrum management policies. However, an inherent disadvantage of RL algorithms is their need for the exploration process, which normally involves a large number of trial-and-error iterations, during which the system exhibits poor performance due to its lack of initial knowledge of the environment. This property of classical RL algorithms significantly limits their applicability in challenging real-world wireless environments where the primary and/or secondary user QoS guarantees must be accommodated.

The work presented in this thesis has therefore focused on accelerating distributed RL based DSA algorithms and, thus, improving their adaptability in realistic cognitive wireless network environments. First, an adaptation of the Win-or-Learn-Fast (WoLF) variable learning rate principle was proposed to improve the initial and steady-state performance of classical RL based DSA algorithms in cellular networks without the use of any external heuristic information. Next, the heuristically accelerated RL

(HARL) framework developed in Chapter 6 was analytically and empirically shown to achieve significant further improvements in the initial performance and the convergence speed of distributed RL based DSA algorithms. The proposed HARL based DSA scheme utilises the ICIC signals in LTE networks as external heuristic information to accelerate the learning process of every individual base station. After that, the novel HARL framework was extended to the dynamic spectrum sharing scenario which considers both the QoS provided to the users of the cognitive wireless networks as well as their effect on the primary user QoS. The proposed DSA algorithms based on the extended HARL framework make use of the radio environment map (REM) to enable the cognitive base stations to learn efficient DSA patterns, whilst successfully coexisting with the other primary and secondary wireless networks in their environment. Finally, a distributed case-based RL approach to DSA was proposed. It combines RL and case-based reasoning (CBR) to increase the robustness and adaptability of distributed RL based DSA schemes in dynamically changing wireless environments.

A more detailed chapter-by-chapter discussion of the original contributions of this thesis towards the enhanced adaptability of distributed RL based DSA algorithms is given in the following subsection.

9.1.1 Original Contributions

Win-or-Learn-Fast Variable Learning Rate

A novel adaptation of the Win-or-Learn-Fast (WoLF) variable learning rate approach for distributed RL based DSA algorithms is proposed in Chapter 4. It uses two fixed values for the learning rate parameter: a lower value when the learning agent receives positive rewards, i.e. due to successful transmissions, and a higher value for negative rewards which correspond to blocked or interrupted transmissions. In this way every learning agent, i.e. base station, in the wireless environment is learning faster when it is "losing" and more slowly and cautiously when "winning". This simple variable learning rate approach is empirically shown to improve the speed of convergence of a distributed stateless Q-learning based DSA algorithm at the early stage of the learning process. Interestingly, it also tends to converge on better solutions, i.e. those that

result in a better steady-state system QoS, compared to the identical RL algorithm with a traditional, fixed learning rate. This suggests an improvement in adaptability of the distributed learning agents due to larger negative step changes in the Q-tables that enable them to escape local optima.

Bayesian Network Based Convergence Analysis Method

Chapter 5 proposes a Bayesian network based joint policy transition analysis methodology that is able to provide a simple and accurate probabilistic model of distributed RL based DSA algorithms applied to a minimum complexity generalised inter-cell interference problem. A Monte Carlo simulation of the distributed Q-learning based DSA algorithm introduced in Chapter 4 shows that the proposed approach demonstrates remarkably accurate prediction of the convergence behaviour of such algorithms. Furthermore, their behaviour can also be expressed in the form of an absorbing Markov chain, derived from the novel Bayesian network model. This representation enables further theoretical analysis of convergence properties of RL based DSA algorithms. Finally, the main benefit of this analysis tool is that it enables the design and theoretical evaluation of novel RL based DSA algorithms by extending the proposed Bayesian network model, that describes a standard distributed Q-learning scheme.

Distributed ICIC Accelerated Q-Learning

The distributed ICIC accelerated Q-learning (DIAQ) scheme proposed in Chapter 6 combines distributed RL and standardized ICIC signalling in the LTE downlink, using the framework of heuristically accelerated RL (HARL). It is theoretically evaluated using a novel extension of the Bayesian network model proposed in Chapter 5, which explains a predicted improvement in convergence behaviour achieved by DIAQ, compared to classical distributed RL. Large scale simulation experiments of a stadium small cell network show that it provides superior QoS compared to a typical heuristic ICIC approach and the distributed RL based approach introduced in Chapter 4. A comparison of the probability of retransmission time response characteristics of DIAQ and pure distributed Q-learning reveals a dramatic improvement in performance at the initial stage of learning due to the use of heuristics for guiding the exploration

process. This result confirms the theoretical predictions made using the Bayesian network model of the algorithm. DIAQ also exhibits excellent steady-state performance and convergence speed, thus, dramatically increasing the adaptability of the distributed RL approach to DSA. Finally, it is designed to comply with the current LTE standards. Therefore, it allows easy implementation of robust distributed machine intelligence for full self-organisation in existing commercial networks.

HARL for Dynamic Secondary Spectrum Sharing

Chapter 7 extends the HARL framework proposed in Chapter 6 and presents a novel mechanism for dynamic spectrum sharing (DSS) based on it. It utilises a radio environment map (REM) as external information for guiding the learning process of cognitive wireless networks. The DSA and DSS schemes proposed in Chapter 7 are shown to achieve high controllability of spectrum sharing patterns in a fully autonomous way. They also result in a significant decrease in primary system QoS degradation due to the interference from the secondary cognitive systems, compared to a state-of-the-art RL solution and a purely heuristic typical LTE solution. The spectrum sharing patterns that emerge by using the proposed schemes also result in remarkable reliability of the wide coverage cognitive eNodeB on an aerial platform in a scenario where it has secondary access to LTE spectrum shared with a local primary network and a secondary high capacity density small cell network. Furthermore, the novel general structure of heuristic functions proposed in the context of HARL are applicable to a wide range of self-organisation problems beyond the wireless communications domain.

Case-Based RL for Dynamic Wireless Environments

The case-based RL (CBRL) technique proposed in Chapter 8 is an effective and feasible approach to DSA in cognitive cellular systems with dynamic topologies. Large-scale system level simulations of a stadium small cell network with an asymmetric time-variant topology show that augmenting classical distributed stateless Q-learning with the CBR functionality in this way results in increased adaptability of the cognitive cellular system to changes in its radio environment. However, as the complexity of the dynamic environment and the possible number of network topologies increase,

the performance gap between classical and case-based RL decreases. Nevertheless, the proposed distributed case-based Q-learning approach achieves a consistent improvement in the system QoS and its stability in the dynamic cellular environment considered. However, a far more significant contribution towards improving the QoS and its adaptability and robustness in cognitive wireless networks is achieved by the HARL based schemes proposed in Chapters 6 and 7. Therefore, if the heuristic spectrum awareness information used by these schemes is available to the cognitive eNBs, introducing CBR into the learning process has a small effect on system performance.

9.1.2 Hypothesis Revisited

The hypothesis stated at the beginning of this thesis is the following:

"Appropriate use of available heuristic information can accelerate distributed reinforcement learning algorithms to enable highly adaptable dynamic spectrum access in cognitive wireless networks."

The key contributions of this thesis described in Subsection 9.1.1 can be summarised in the context of the above hypothesis as follows:

- The WoLF variable learning rate proposed in Chapter 4 increases the adaptability of the distributed RL based DSA approach by making it more difficult for the distributed learning agents to converge on local optima and, thus, encouraging them to keep looking for better DSA policies.
- The DIAQ scheme proposed in Chapter 6 uses standard ICIC signalling in LTE networks as an additional heuristic information source to dramatically improve the temporal characteristics of the distributed RL based DSA approach, such as its initial and steady-state performance, as well as its convergence speed, all of which contribute towards its significantly increased adaptability in a wireless environment.
- Similarly to the ICIC-aided DIAQ scheme, the HARL based DSS approach proposed in Chapter 7 uses a REM as a heuristic information source to increase the robustness and adaptability of distributed RL based DSA algorithms in scenar-

ios that involve several independent primary and secondary wireless networks sharing the same spectrum.

- The HARL based DSS approach proposed in Chapter 7 also includes a novel framework for utilising an arbitrary number of heuristic information sources together to accelerate the learning process based on several criteria, e.g. incorporating both the REM content and the ICIC signals to increase the adaptability of the small cell eNBs to their neighbouring eNBs as well as to the independently operating aerial eNB.
- The CBRL approach proposed in Chapter 8 uses the network topology information to enable the distributed learning agents to identify different configurations of a dynamic wireless environment, adapt their learning processes to these changes more rapidly and, thus, stabilise their performance.

These contributions are empirically, and in some cases analytically, shown to cause dramatic improvements in the adaptability of distributed RL based DSA methods applied to complex cognitive wireless network environments, thus, proving the hypothesis of this thesis.

9.2 Recommendations for Further Work

This section gives a number of recommendations for further work on the areas explored in this thesis. They predominantly involve extending the applicability of the proposed techniques to a wider range of scenarios beyond the scope of this work.

Effect of WoLF Learning Rate in Different DSA and DSS Scenarios

Although the WoLF principle for varying the learning rate of the distributed stateless Q-learning algorithm is analytically justified, it is empirically evaluated only using the stadium small cell network scenario investigated in this thesis. In order to verify that the WoLF variable learning rate scheme proposed in Chapter 4 is generally applicable to RL based DSA in wireless networks, it has to be tested using a range of different network architectures and DSA/DSS scenarios. Furthermore, since the proposed

WoLF principle is not specific to the stateless Q-learning algorithm used for simulation experiments in this thesis, it would also be interesting to assess the benefits of implementing the WoLF variable learning rate for other RL algorithms applied to wireless communications problems.

Bayesian Network Analysis of Secondary Spectrum Sharing

The Bayesian network based method for theoretical convergence analysis of distributed RL based DSA algorithms proposed in Chapter 5 uses a simple inter-cell interference model with two co-primary spectrum users. It would also be possible to modify the structure of joint policy transition probabilities to adapt this model to a secondary spectrum sharing scenario which involves a primary user with a conventional non-RL based spectrum management policy and a cognitive secondary user learning to avoid interference with the primary user. This method would then have the potential to provide theoretical insight into the effects of the RL based DSS approach on both the primary and the secondary user performance.

Bayesian Network Analysis of CBRL

Another potential application of the Bayesian network based convergence analysis technique proposed in Chapter 5 is the CBRL algorithm for DSA in dynamic wireless environments proposed in Chapter 8. The fundamental difference between the inter-cell interference scenario considered in the Bayesian network model used in Chapters 5 and 6 and one that would describe CBRL is the fact that the application scenario of the latter is dynamic rather than static. Therefore, its main purpose would be to provide theoretical insight into the adaptability of the CBRL-enabled cognitive wireless devices to changes in their radio environment that disrupt their learning process.

Heuristic Acceleration Applied to Different RL Algorithms

The heuristic functions designed to guide the learning process of the HARL algorithms for intelligent DSA and DSS proposed in Chapters 6 and 7 are shown to be highly effective when combined specifically with the stateless Q-learning algorithm introduced

in Chapter 4. It would be beneficial to generalise the mask-based heuristic acceleration principle developed in this thesis to a range of different RL algorithms, such as classical Q-learning, SARSA and actor-critic learning. The classical RL algorithms normally involve a state space in addition to the action space considered by the stateless Q-learning approach. Therefore, the structure of the heuristic functions will likely need to be adapted to their state-action space specifications. However, the author believes that the core principle of creating a temporary Q-table modified by the heuristic function and using it for decision making, as proposed in this thesis, is an approach that would be generally applicable to any centralised or distributed RL algorithm used for DSA in wireless networks.

HARL in Different Spectrum Sharing Scenarios

The stadium temporary event scenario used for the development and simulations of the HARL based DSS methods in Chapter 7 is an appropriately complex and realistic problem. However, the proposed distributed REM and ICIC accelerated Q-learning (DRIAQ) and REM accelerated Q-learning with Q-value based admission control (RAQ-AC) algorithms are specific to that particular scenario. Since the novel HARL framework that forms the basis for these algorithms is generally applicable to arbitrary DSA and DSS problems, a thorough empirical evaluation of different DSS problems solved by similar algorithms based on this framework would significantly widen its impact and applicability.

Different Heuristic Information Sources for HARL

Another possible direction for future work on extending the HARL framework for DSA and DSS proposed in this thesis is to adapt it to other heuristic information sources. For example, the schemes developed in this thesis use standardised ICIC signalling and a specific form of a REM database for the heuristic acceleration of RL algorithms. This framework could be extended to be compatible with other potentially available heuristic information sources, such as interference power measurements, or Q-tables of other distributed learning agents, i.e. combining the HARL framework with transfer learning introduced in Subsection 2.4.3. Adapting the HARL framework

to various heuristic information sources and performing a thorough comparative analysis of their performance would provide new valuable insight into the magnitude of its potential benefits.

CBRL in More Complex Dynamic Environment Scenarios

The work on CBRL based DSA in dynamic cellular environments presented in Chapter 8 only considers scenarios where the wireless network switches between a limited number of possible topologies depending on the traffic load distribution. More complex scenarios with a larger number of potential topologies may require more sophisticated methods for case identification and retrieval. For example, the detected cases may need to be divided into dynamically configured clusters to avoid an excessive number of immature learning processes taking place in parallel, which in turn may have a detrimental effect on the system performance. Therefore, although Chapter 8 shows promising preliminary results, there is scope for developing more advanced CBRL schemes that further enhance the stability and adaptability of cognitive wireless networks in dynamic radio environments.

Exploiting Similarities between Cases in CBRL

One way of increasing the effectiveness of the CBRL approach to DSA not investigated in this thesis is exploiting the similarities between different cases of the environment. For example, the information learnt through the RL process at a particular network topology is likely to be valuable at a different, yet largely identical network topology. A way of updating several Q-tables at once, based on the similarity between the cases they correspond to, may significantly improve the maturity of the information stored in the case base, since each case base entry will take into account a significantly larger amount of the learning agent's trial-and-error experience. A technique that could potentially facilitate this functionality is fuzzy logic [92], since it is based on the fuzzy set theory which allows varying degrees of individual's membership in a particular set. For example, a value between 0 and 1 that represents the degree of membership of a particular network topology in a particular case could be defined such that it could be used to weigh the RL updates performed on the given case.

Glossary

2ON Second Order Neighbourhood

AC Admission Control

AeNB Aerial eNodeB

BS Base Station

CBR Case-Based Reasoning

CBHARL Case-Based Heuristically Accelerated Reinforcement Learning

CBRL Case-Based Reinforcement Learning

CR Cognitive Radio

DIAQ Distributed ICIC Accelerated Q-Learning

DRIAQ Distributed REM and ICIC Accelerated Q-Learning

DCA Dynamic Channel Assignment

DP Dynamic Programming

DSA Dynamic Spectrum Access

DSS Dynamic Spectrum Sharing

eNB Evolved NodeB

EWMA Exponentially Weighted Moving Average

HARL Heuristically Accelerated Reinforcement Learning

ICIC Inter-Cell Interference Coordination

LoS Line-of-Sight

LSA Licensed Shared Access

LTE Long Term Evolution

MDP Markov Decision Process

MNO Mobile Network Operator

MNRSS Minimum Neighbour Received Signal Strength

NE Nash Equilibrium

PeNB Primary eNodeB

POMDP Partially Observable Markov Decision Process

QoS Quality of Service

RAQ-AC REM Accelerated Q-Learning with Q-Value Based Admission Control

RAT Radio Access Technology

REM Radio Environment Map

RL Reinforcement Learning

RNTP Relative Narrowband Transmit Power

RSRP Reference Signal Received Power

SG Stochastic Game

SINR Signal-to-Interference-plus-Noise Ratio

SNR Signal-to-Noise Ratio

TL Transfer Learning

TPM Transition Probability Matrix

TRM Transition Reward Matrix

TVWS TV White Space

UE User Equipment

UT User Throughput

VRB Virtual Resource Block

WoLF Win-or-Learn-Fast

WRAN Wireless Regional Area Network

References

[1] 3GPP. Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA physical layer aspects (3GPP TR 36.814 version 9.0.0 Release 9). December 2010.

[2] 3GPP. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios (3GPP TR 36.952 version 11.0.0 Release 11). December 2012.

[3] 3GPP. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (3GPP TS 36.213 version 11.5.0 Release 11) . December 2013.

[4] D. Akerberg and F. Brouwer. On channel definitions and rules for continuous dynamic channel selection in coexistence etiquettes for radio systems. In *IEEE Vehicular Technology Conference (VTC)*, 1994.

[5] L.G. Anderson. A simulation study of some dynamic channel assignment algorithms in a high capacity mobile telecommunications system. *Vehicular Technology, IEEE Transactions on*, 22:210–217, 1973.

[6] M. Bennis and D. Niyato. A Q-learning based approach to interference avoidance in self-organized femtocell networks. In *2010 IEEE GLOBECOM Workshops (GC Wkshps)*, 2010.

[7] M. Bennis, S.M. Perlaza, P. Blasco, Zhu Han, and H.V. Poor. Self-organization in small cell networks: A reinforcement learning approach. *Wireless Communications, IEEE Transactions on*, 12:3202–3212, 2013.

[8] F. Bernardo, R. Agusti, J. Perez-Romero, and O. Sallent. Distributed spectrum management based on reinforcement learning. In *International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*, 2009.

[9] R.A.C. Bianchi, L.A. Celiberto Jr., P.E. Santos, J.P. Matsuura, and R. Lopez de Mantaras. Transferring knowledge as heuristics in reinforcement learning: A case-based approach. *Artificial Intelligence*, 226:102 – 121, 2015.

[10] R.A.C. Bianchi and R. Lopez de Mantaras. Case-based multiagent reinforcement learning: Cases as heuristics for selection of actions. In *European Conference on Artificial Intelligence (ECAI 2010)*, 2010.

[11] R.A.C. Bianchi, M.F. Martins, C.H.C. Ribeiro, and A.H.R. Costa. Heuristically-accelerated multiagent reinforcement learning. *Cybernetics, IEEE Transactions on*, 44:252–265, 2014.

[12] R.A.C. Bianchi, C.H.C. Ribeiro, and A.H.R. Costa. Heuristically accelerated q–learning: a new approach to speed up reinforcement learning. In *Advances in Artificial Intelligence–SBIA 2004*, pages 245–254. Springer, 2004.

[13] M. Bowling and M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136:215–250, 2002.

[14] L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *Trans. Sys. Man Cyber Part C*, 38(2):156–172, March 2008.

[15] L.A. Celiberto, J.P. Matsuura, R. Lopez de Mantaras, and R.A.C. Bianchi. Reinforcement learning with case-based heuristics for robocup soccer keepaway. In *Robotics Symposium and Latin American Robotics Symposium (SBR-LARS), 2012 Brazilian*, pages 7–13, 2012.

[16] J.M. Chapin and W.H. Lehr. Cognitive radios for dynamic spectrum access - the path to market success for dynamic spectrum access technology. *Communications Magazine, IEEE*, 45(5):96–103, May 2007.

[17] Si Chen, R. Vuyyuru, O. Altintas, and Alexander M. Wyglinski. On optimizing vehicular dynamic spectrum access networks: Automation and learning in mobile wireless environments. In *Vehicular Networking Conference (VNC), 2011 IEEE*, 2011.

[18] Xianfu Chen, Zhifeng Zhao, and Honggang Zhang. Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks. *Mobile Computing, IEEE Transactions on*, 12:2155–2166, 2013.

[19] Xianfu Chen, Zhifeng Zhao, Honggang Zhang, and Tao Chen. Conjectural variations in multi-agent reinforcement learning for energy-efficient cognitive wireless mesh networks. In *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, pages 820–825, 2012.

[20] M.M.-L. Cheng and J.C.-I. Chuang. Performance evaluation of distributed measurement-based dynamic channel assignment in local wireless communications. *Selected Areas in Communications, IEEE Journal on*, 14:698–710, 1996.

[21] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, 1998.

[22] C. Cordeiro, K. Challapali, D. Birru, and N. Sai Shankar. IEEE 802.22: the first worldwide wireless standard based on cognitive radios. In *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, 2005.

[23] C. Cordeiro, M. Ghosh, D. Cavalcanti, and K. Challapali. Spectrum sensing for dynamic spectrum access of TV bands. In *International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, 2007.

[24] D. C. Cox and D. O. Reudink. Dynamic channel assignment in high-capacity mobile communications systems. *Bell System Technical Journal*, 50:1833–1857, 1971.

[25] M. Dirani and Z. Altman. A cooperative reinforcement learning approach for inter-cell interference coordination in OFDMA cellular networks. In *Interna-*

tional Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010.

[26] A. Feki, V. Capdevielle, and E. Sorsy. Self-organized resource allocation for LTE pico cells: A reinforcement learning approach. In *IEEE Vehicular Technology Conference (VTC Spring)*, 2012.

[27] M. Fitch, M. Nekovee, S. Kawade, K. Briggs, and R. MacKenzie. Wireless service provision in TV white space with cognitive radio technology: A telecom operator's perspective and experience. *Communications Magazine, IEEE*, 49:64–73, 2011.

[28] I.G. Fraimis, V.D. Papoutsis, and S.A. Kotsopoulos. A decentralized subchannel allocation scheme with inter-cell interference coordination (ICIC) for multi-cell OFDMA systems. In *IEEE Global Telecommunications Conference (GLOBECOM)*, 2010.

[29] A. Ghasemi and E.S. Sousa. Spectrum sensing in cognitive radio networks: requirements, challenges and design trade-offs. *Communications Magazine, IEEE*, 46:32–39, 2008.

[30] C. Ghosh, S. Roy, and D. Cavalcanti. Coexistence challenges for heterogeneous cognitive wireless networks in TV white spaces. *Wireless Communications, IEEE*, 18:22–31, 2011.

[31] L. Giupponi, A. Galindo-Serrano, P. Blasco, and M. Dohler. Docitive networks: an emerging paradigm for dynamic spectrum management. *Wireless Communications, IEEE*, 17(4):47–54, 2010.

[32] K. Gomez, T. Rasheed, R. Hermenier, Tao Jiang, S. Rehan, D. Grace, L. Reynaud, T. Javornik, I. Ozimek, and L. Le Garrec. FP7-ICT-2011-8-318632-ABSOLUTE/D2.6.1 System-wide Simulations Planning Document. 2013.

[33] I. Grondman, L. Busoniu, G.A.D. Lopes, and R. Babuska. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(6):1291–1307, 2012.

[34] M. Guizani, B. Khalfi, M.B. Ghorbel, and B. Hamdaoui. Large-scale cognitive cellular systems: resource management overview. *Communications Magazine, IEEE*, 53:44–51, 2015.

[35] D. Gurney, G. Buchwald, L. Ecklund, S.L. Kuffner, and J. Grosspietsch. Geolocation database techniques for incumbent protection in the TV white space. In *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2008.

[36] J. Hu and M.P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *In Proceedings of the Fifteenth International Conference on Machine Learning*, pages 242–250. Morgan Kaufmann, 1998.

[37] Senhua Huang, Xin Liu, and Zhi Ding. Opportunistic spectrum access in cognitive radio networks. In *IEEE Conference on Computer Communications (INFOCOM)*, 2008.

[38] K. Ishizu, K. Hasegawa, K. Mizutani, H. Sawada, K. Yanagisawa, T. Keat-Beng, T. Matsumura, S. Sasaki, M. Asano, H. Murakami, and H. Harada. Field experiment of long-distance broadband communications in TV white space using IEEE 802.22 and IEEE 802.11af. In *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2014.

[39] T. Jaakkola, M.I. Jordan, and S.P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Comput.*, 6(6):1185–1201, November 1994.

[40] Chengzhi Jiang and Zhaohan Sheng. Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system. *Expert Syst. Appl.*, 36(3):6520–6526, April 2009.

[41] T. Jiang, D. Grace, and Y. Liu. Two-stage reinforcement-learning-based cognitive radio with exploration control. *Communications, IET*, 5:644–651, 2011.

[42] Tao Jiang. *Reinforcement Learning-based Spectrum Sharing for Cognitive Radio*. PhD thesis, University of York, 2011.

[43] Tao Jiang, D. Grace, and P. D. Mitchell. Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing. *Communications, IET*, 5:1309–1317, 2011.

[44] Tao Jiang, Peng Li, Chunshan Liu, N. Khan, D. Grace, A. Burr, and C. Oestges. EU FP7 INFSO-ICT-248267 BuNGee Deliverable D4.1.2: Simulation Tool(s) and Simulation Results. 2012.

[45] L.S. Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[46] S. Kapetanakis and D. Kudenko. Reinforcement learning of coordination in cooperative multi-agent systems. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI)*, 2002.

[47] I. Katzela and M. Naghshineh. Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey. *Personal Communications, IEEE*, 3(3):10–31, 1996.

[48] M. Kearns and D. Koller. Efficient reinforcement learning in factored MDPs. In *International Joint Conference on Artificial Intelligence (IJCAI) - Volume 2*, 1999.

[49] A.H.R. Ko, R. Sabourin, and F. Gagnon. Performance of distributed multi-agent multi-state reinforcement spectrum management using different exploration schemes. *Expert Systems with Applications*, 40(10):4115 – 4126, 2013.

[50] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzić, M. Milojević, A. Hong, J. Ylitalo, V. Holappa, M. Alatossava, R. Bultitude, Y. de Jong, and T. Rautiainen. IST-4-027756 WINNER II Deliverable D1.1.2: WINNER II channel models. 2008.

[51] M. Lauer and M. Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth*

International Conference on Machine Learning, pages 535–542. Morgan Kaufmann, 2000.

[52] Ying-Chang Liang, Anh Tuan Hoang, and Hsiao-Hwa Chen. Cognitive radio on TV bands: a new approach to provide wireless connectivity for rural areas. *Wireless Communications, IEEE*, 15:16–22, 2008.

[53] N. Lilith and K. Dogancay. Distributed reduced-state sarsa algorithm for dynamic channel allocation in cellular networks featuring traffic mobility. In *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, volume 2, pages 860–865 Vol. 2, 2005.

[54] N. Lilith and K. Dogancay. Distributed dynamic call admission control and channel allocation using sarsa. In *Communications, 2005 Asia-Pacific Conference on*, pages 376–380, Oct.

[55] M.L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163. Morgan Kaufmann, 1994.

[56] J. Lunden, S.R. Kulkarni, V. Koivunen, and H.V. Poor. Multiagent reinforcement learning based spectrum sensing policies for cognitive radio networks. *Selected Topics in Signal Processing, IEEE Journal of*, 7:858–868, 2013.

[57] M.A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo. Optimal energy savings in cellular access networks. In *IEEE International Conference on Communications Workshops (ICC Workshops)*, 2009.

[58] M. Matinmikko, H. Okkonen, M. Palola, S. Yrjola, P. Ahokangas, and M. Mustonen. Spectrum sharing using licensed shared access: the concept and its workflow for LTE-advanced networks. *Wireless Communications, IEEE*, 21:72–79, 2014.

[59] R. McGill, J.W. Tukey, and W.A. Larsen. Variations of box plots. *The American Statistician*, 32:12–16, 1978.

[60] R.K. McLean, M.D. Silvius, K.M. Hopkinson, B.N. Flatley, E.S. Hennessey, C.C. Medve, J.J. Thompson, M.R. Tolson, and C.V. Dalton. An architecture for coexistence with multiple users in frequency hopping cognitive radio networks. *Selected Areas in Communications, IEEE Journal on*, 32:563–571, 2014.

[61] J. Mitola. *Cognitive radio: Model-based competence for software radios*. PhD thesis, KTH, 1999.

[62] A.M. Monk and L.B. Milstein. Open-loop power control error in a land mobile satellite system. *Selected Areas in Communications, IEEE Journal on*, 13:205–212, 1995.

[63] A.W. Moore and C.G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13:103–130, 1993.

[64] N. Morozs, T. Clarke, and D. Grace. A novel adaptive call admission control scheme for distributed reinforcement learning based dynamic spectrum access

in cellular networks. In *International Symposium on Wireless Communication Systems (ISWCS)*, 2013.

[65] Junhong Nie and S. Haykin. A dynamic channel assignment policy through q-learning. *Neural Networks, IEEE Transactions on*, 10(6):1443–1455, 1999.

[66] M. Palola, M. Matinmikko, J. Prokkola, M. Mustonen, M. Heikkila, T. Kippola, S. Yrjola, V. Hartikainen, L. Tudose, A. Kivinen, J. Paavola, and K. Heiska. Live field trial of licensed shared access (LSA) concept using LTE network in 2.3 GHz band. In *IEEE International Symposium on Dynamic Spectrum Access Networks (DYSPAN)*, pages 38–47, April 2014.

[67] C. Pandana and K.J.R. Liu. Near-optimal reinforcement learning framework for energy-aware sensor communications. *Selected Areas in Communications, IEEE Journal on*, 23:788–797, 2005.

[68] P. Poupart, N. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete bayesian reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2006.

[69] K.N. Ramachandran, E.M. Belding-Royer, K.C. Almeroth, and M.M. Buddhikot. Interference-aware channel assignment in multi-radio wireless mesh networks. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2006.

[70] S. Rehan and D. Grace. Combined green resource and topology management for beyond next generation mobile broadband systems. In *Computing, Networking and Communications (ICNC), 2013 International Conference on*, pages 242–246, 2013.

[71] L. Reynaud, S. Allsopp, P. Charpentier, Hanwen Cao, D. Grace, R. Hermenier, A. Hrovat, G. Hughes, C. Ioan, T. Javornik, A. Munari, M.M. Vidal, J. Strother, R. Valcarce, and S. Zaharia. FP7-ICT-2011-8-318632-ABSOLUTE/D2.1 Use cases definition and scenarios description. 2014.

[72] F. Richter, A.J. Fehske, and G.P. Fettweis. Energy efficiency aspects of base station deployment strategies for cellular networks. In *IEEE Vehicular Technology Conference (VTC-Fall)*, 2009.

[73] G.J. Ross, N.M. Adams, D.K. Tasoulis, and D.J. Hand. Exponentially weighted moving average charts for detecting concept drift. *Pattern Recognition Letters*, 33:191 – 198, 2012.

[74] G. A. Rummery and M. Niranjan. On-line q-learning using connectionist systems. Technical report, Cambridge University Engineering Department, 1994.

[75] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003.

[76] R.O. Schmidt. Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on*, 34:276–280, 1986.

[77] S. Sen, I. Sen, M. Sekaran, and J. Hale. Learning to coordinate without sharing information. In *The Twelfth National Conference on Artificial Intelligence (AAAI)*, 1994.

[78] S.-M. Senouci and G. Pujoile. Dynamic channel assignment in cellular networks: a reinforcement learning solution. In *International Conference on Telecommunications (ICT)*, 2003.

[79] S. Sesia, M. Baker, and I. Toufik. *LTE-The UMTS Long Term Evolution: From Theory to Practice*. John Wiley & Sons, 2011.

[80] A. Shahid, S. Aslam, Hyung Seok Kim, and Kyung-Geun Lee. A docitive Q-learning approach towards joint resource allocation and power control in self-organised femtocell networks. *Transactions on Emerging Telecommunications Technologies*, 26:216–230, 2015.

[81] M. Simsek, M. Bennis, and A. Czylwik. Dynamic inter-cell interference coordination in HetNets: A reinforcement learning approach. In *IEEE Global Communications Conference (GLOBECOM)*, 2012.

[82] S. Singh and D. Bertsekas. Reinforcement learning for dynamic channel allocation in cellular telephone systems. In *Advances in Neural Information Processing Systems (NIPS)*, 1997.

[83] Qingyang Song and Abbas Jamalipour. A quality of service negotiation-based vertical handoff decision scheme in heterogeneous wireless systems. *European Journal of Operational Research*, 191:1059 – 1074, 2008.

[84] S. Srinivasa and S.A. Jafar. Cognitive radios for dynamic spectrum access - the throughput potential of cognitive radio: A theoretical perspective. *Communications Magazine, IEEE*, 45:73–79, 2007.

[85] C. Stevenson, G. Chouinard, Zhongding Lei, Wendong Hu, S.J. Shellhammer, and W. Caldwell. IEEE 802.22: The first cognitive radio wireless regional area network standard. *Communications Magazine, IEEE*, 47:130–138, 2009.

[86] Hongjian Sun, A. Nallanathan, Cheng-Xiang Wang, and Yunfei Chen. Wideband spectrum sensing for cognitive radio networks: a survey. *Wireless Communications, IEEE*, 20:74–81, 2013.

[87] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

[88] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *International Conference on Machine Learning (ICML)*, 1993.

[89] Yinglei Teng, Yong Zhang, Fang Niu, Chao Dai, and Mei Song. Reinforcement learning based auction algorithm for dynamic spectrum access in cognitive radio networks. In *Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd*, pages 1–5, 2010.

[90] J.N. Tsitsiklis and R. Sutton. Asynchronous stochastic approximation and q-learning. In *Machine Learning*, pages 185–202, 1994.

[91] P. Vlacheas, E. Thomatos, K. Tsagkaris, and P. Demestichas. Autonomic downlink inter-cell interference coordination in LTE self-organizing networks. In *International Conference on Network and Services Management (CNSM)*, 2011.

[92] L. Wang. *Adaptive Fuzzy Systems and Control, Design and Stability Analysis*. PTR Prentice Hall, 1994.

[93] Wenbo Wang, A. Kwasinski, D. Niyato, and Zhu Han. A survey on applications of model-free strategy learning in cognitive wireless networks. *CoRR*, abs/1504.03976, 2015.

[94] C. Watkins. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, England, 1989.

[95] I. Watson. Case-based reasoning is a methodology not a technology. *Knowledge-Based Systems*, 12(56):303 – 308, 1999.

[96] M.A. Wiering. Reinforcement learning in dynamic environments using instantiated information. In *In Proceedings of the Eighth International Conference on Machine Learning*, 2001.

[97] I.H. Witten. An adaptive optimal controller for discrete-time markov environments. *Information and Control*, 34(4):286 – 295, 1977.

[98] Cheng Wu, K. Chowdhury, M. Di Felice, and W. Meleis. Spectrum management of cognitive radio using multi-agent reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS): Industry Track*, 2010.

[99] T. Yucek and H. Arslan. A survey of spectrum sensing algorithms for cognitive radio applications. *Communications Surveys Tutorials, IEEE*, 11(1):116–130, 2009.

[100] Qing Zhao, Lang Tong, A. Swami, and Yunxia Chen. Decentralized cognitive mac for opportunistic spectrum access in ad hoc networks: A pomdp framework. *Selected Areas in Communications, IEEE Journal on*, 25:589–600, 2007.

[101] Qiyang Zhao and D. Grace. Application of cognition based resource allocation strategies on a multi-hop backhaul network. In *IEEE International Conference on Communication Systems (ICCS)*, pages 423–427, 2012.

[102] Qiyang Zhao and D. Grace. Transfer learning for QoS aware topology management in energy efficient 5G cognitive radio networks. In *International Conference on 5G for Ubiquitous Connectivity (5GU)*, pages 152–157, 2014.

[103] Qiyang Zhao, David Grace, and Tim Clarke. Transfer learning and cooperation management: balancing the quality of service and information exchange overhead in cognitive radio networks. *Transactions on Emerging Telecommunications Technologies*, 26:290–301, 2015.

[104] Qiyang Zhao, Tao Jiang, N. Morozs, D. Grace, and T. Clarke. Transfer learning: A paradigm for dynamic spectrum and topology management in flexible architectures. In *IEEE Vehicular Technology Conference (VTC Fall)*, 2013.