AN ABSTRACT OF THE THESIS OF

Pavithra Venkatraman for the degree of Master of Science in
Electrical and Computer Engineering presented on November 04, 2010.
Title: Opportunistic Bandwidth Sharing Through Reinforcement Learning

Abstract approved: _____

Bechir Hamdaoui

The enormous success of wireless technology has recently led to an explosive demand for, and hence a shortage of, bandwidth resources. This expected shortage problem is reported to be primarily due to the inefficient, static nature of current spectrum allocation methods. As an initial step towards solving this shortage problem, Federal Communications Commission (FCC) opens up for the so-called *opportunistic spectrum access* (OSA), which allows unlicensed users to exploit unused licensed spectrum, but in a manner that limits interference to licensed users. Fortunately, technological advances enabled cognitive radios, which are viewed as *intelligent* communication systems that can *self-learn* from their surrounding environment, and *auto-adapt* their internal operating parameters in real-time to improve spectrum *efficiency*. Cognitive radios have recently been recognized as the key enabling technology for realizing OSA. In this work, we propose a machine learning-based scheme that exploits the cognitive radios' capabilities to enable effective OSA, thus improving the efficiency of spectrum utilization. Specifically, we formulate the OSA problem as a finite Markov Decision Process (MDP), and use reinforcement learning (RL) to locate and exploit bandwidth opportunities effectively. Simulation

results show that our scheme achieves high throughput performance without requiring any prior knowledge of the environment's characteristics and dynamics.

Opportunistic Bandwidth Sharing Through Reinforcement Learning

by

Pavithra Venkatraman

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented November 04, 2010
Commencement June 2011

Master of Science thesis of Pavithra Venkatraman presented on November 04, 2010

APPROVED:

Major Professor, representing Electrical and Computer Engineering

---

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Pavithra Venkatraman, Author

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my major advisor Dr. Bechir Hamdaoui for giving me an opportunity to work in his research group. My deepest thanks to him for his valuable inputs and intuitive suggestions that helped me all through the research work. His encouragement motivated me in achieving my goals.

I would like to thank Dr. Huaping Liu and Dr. Thinh Nguyen for serving on my committee and reviewing my manuscript. Special thanks to both of them for their wonderful lectures in class. I want to thank the Graduate Council Representative, Dr. Yun-Shik Lee for being a part of my committee. I would like to thank Dr. Bella Bose for his constant words of encouragement and support throughout my Masters study. My sincere thanks to Ferne Simendinger for helping me on the administrative side.

I thank my Networking group members Samina Ehsan, Akhil Sivanantha, Nessrine Chakchouk, Omar Alsaleh, Megha Maiya, and Majid Alkaee Taleghan for their knowledge sharing and insightful technical discussions.

Thanks to my husband Karthik Jayaraman for his invaluable support and for constantly motivating me to get the work done ahead of schedule. I express my heartfelt thanks to my brother, parents and grandparents for being understanding and patient in letting me choose my own career path and guiding me through times of stress. Their kindness and affection remains unparalleled and I am lucky to have them for life. I thank the almighty for showering blessings upon me.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

1. INTRODUCTION

There is a huge demand for radio spectrum due to the rapid growth in wireless technology. Unfortunately the spectrum supply has not catered to this growing demand. The shortage in spectrum supply has primarily been due to the inefficient, inflexible, static nature of the existing spectrum allocation methods and definitely not due to the scarcity of available spectrum [1]. This fact is well supported by measurement-based studies [2, 3] which have shown that the average occupancy of spectrum over all frequencies is a paltry 5.2% and that the occupancy of some bands in the 30-300 MHz range is less than 1%. This measurement data confirms the availability of many spectrum opportunities along time, frequency, and space that wireless devices and networks can potentially utilize. Therefore, it is imperative to develop mechanisms that enable effective and efficient exploitation of these spectrum opportunities.

FCC's long-term vision for solving the spectrum shortage problem is to evolve towards more liberal, flexible spectrum allocation policies and usage rights, where spectrum will be managed and controlled dynamically by network entities and end-user devices themselves with little to no involvement of any centralized regulatory bodies. As an initial step towards this liberal paradigm, FCC promotes the so-called *opportunistic spectrum access* (OSA), which improves spectrum *efficiency* by allowing unlicensed, secondary users (SUs) to exploit unused licensed spectrum, but in a manner that limits interference to licensed, primary users (PUs). Indeed, OSA is becoming a practical reality nowadays: As of November 4th, 2008, FCC [4] established rules to allow unlicensed users to operate in TV-band spectrum on a

secondary basis at locations where that spectrum is open. These TV-band spectrum opportunities will be used by unlicensed fixed, portable, and mobile users to support applications like wireless home networking and video services.

Now that we have the approval of regulatory bodies like FCC to promote OSA, the question that comes naturally is whether *we have the technology and the techniques necessary to enable it or not?* Fortunately, technological advances enabled cognitive radios (CRs), built on software-defined radios [5], which have recently been recognized as one of the key emerging technologies [6] that can potentially make OSA a reality. CRs are viewed as intelligent wireless communication systems that are capable of *self-learning* from their surrounding environment, and *auto-adapting* their internal operating parameters in real-time to improve spectrum *efficiency* with no intervention [7].

The apparent promise of OSA has indeed created significant research interests, resulting in much research, ranging from protocol design [8, 9, 10] to performance optimization [11, 12], and from market-oriented access strategies [13, 14] to new management and architecture paradigms [15, 16, 17, 18]. More recently, some effort has also been given to the development of adaptive learning-based approaches [19] - [30]. Zhao et al. [19] have developed a model for predicting the dynamics of the OSA environment when periodic channel sensing is used. A simple two-state Markovian model is assumed for the activities of PUs on each channel. Using this model, Zhao et al. derive an optimal access policy that can be used to maximize channel utilization while limiting interference to PUs. In [20], Unnikrishnan and Veeravalli propose a cooperative channel selection and access policy for OSA systems under interference constraints. In this paper, the PUs activities are assumed to be stationary Markovian, and the Markovian statistics are assumed to be known

to all SUs. A centralized approach is considered, where all cooperating SUs report their observations to a decision center, which makes the decision regarding when and which channels to sense and access at each time slot. In [21], the authors develop channel-decision policies for two SUs in a two-channel OSA system. The PUs activities are modeled as discrete-time Markov chains. Liu and Zhao [22] consider the case of multiple non-cooperative SUs in OSA systems where SUs are assumed not to exchange information among themselves. The occupancy of primary channels is modeled as an independent and identically distributed Bernoulli process, and OSA is formulated as a multiarmed bandit problem where agents are not cooperative with each other. Chen et al. [30] develop a cross-layer optimal access strategy for OSA that integrates the physical layers sensing with the medium access control (MAC) layers sensing and access policy. They establish a separation principle, meaning that the physical layers sensing and the MAC layers access policy can be decoupled from the MAC layers sensing without losing optimality. The developed framework assumes that the spectrum occupancy of PUs also follows a discrete-time ON/OFF Markov process.

In most of these works, the models developed to derive optimal channel-selection policies assume that the PUs activities follow the Markovian process model. Although analytically tractable, the Markovian process may not accurately model the dynamics of the PUs activities. In fact, the OSA environment has very unique characteristics that make it too difficult to construct models that predict its dynamics, and it is therefore important to develop techniques that can achieve approximately optimal behaviors without requiring models of the environments dynamics.

Indeed, reinforcement learning (RL) [31], a sub-field of artificial intelligence (AI), is a foundational idea built on the basis of learning from interaction without

requiring models of the environment's dynamics, yet can still achieve approximately optimal behaviors. RL techniques require *experience* only, which can be acquired from an *online* or *a simulated* interaction. While learning from an online interaction requires no models of the environment's dynamics, learning from a simulated interaction requires a model that just generates samples of the behavior, not the complete probability distribution. In OSA, for instance, it is easy to generate samples of the environment's behavior according to the desired distribution, but it may be too difficult, or even impossible, to obtain the explicit form of the distribution. For example, a user can easily generate samples of the occupancy of a particular spectrum band through periodic sensing, but it may be infeasible to derive the explicit distribution of traffic behavior.

Based on the aforementioned facts, in this work, we formulate the OSA, using an RL framework. In order to test the effectiveness of the Q-learning scheme in terms of exploiting the spectrum opportunities, we evaluate the learning algorithm for a single secondary-user group and compare the algorithm's performance under different environmental conditions with the random access method [32, 33]. Further, we evaluate two multi-agent RL schemes, namely the non-cooperative and cooperative Q-learning schemes, and compare their performances with the random scheme. Simulation results show that the partial and fully cooperative schemes perform better than the non-cooperative and the random schemes in terms of achieved throughput and balanced traffic loads. Depending on the communication overhead due to the extra traffic in exchanging information between the cooperating users, different levels of partial cooperation can be used. Overall, the proposed learning technique achieves high throughput performance by learning through experience from interaction with the environment and intelligently locating and exploiting spectrum oppor-

tunities. Therefore, it obviates the need for prior knowledge of the environment's characteristics and dynamics.

This thesis is organized as follows. In Chapter 2., we state the OSA problem and discuss its requirements. In Chapter 3., we present our single-agent RL framework for efficient OSA. In Chapter 4., we evaluate and compare the proposed single-agent RL approach with random access approach. In Chapter 5., we present the Multi-Agent RL framework for efficient OSA. We study the effect of having multiple secondary-user groups and evaluate the three different access schemes in Chapter 6. Finally, we conclude the thesis in Chapter 7.

2. OPPORTUNISTIC SPECTRUM ACCESS (OSA)

The spectrum has traditionally been divided by FCC into frequency bands. These spectrum bands are assigned to licensees (or *primary users* (PUs)) who have exclusive and flexible rights to use these bands. PUs are also protected against interference when using their assigned bands. Due to recent findings, showing that large portions of the licensed bands are lightly used or unused at all, and in order to address the spectrum scarcity problem, FCC opens up for the so-called *opportunistic spectrum access* (OSA).

The basic idea behind OSA is to allow unlicensed users, also referred to as *secondary users* (SUs), to exploit unused licensed spectrum on an instant-by-instant basis, but in a manner that limits interference to PUs so as to maintain compatibility with legacy systems. In OSA, an agent is a group of two or more secondary users also known as *secondary-user group* (SUG) who want to communicate together. We assume that all SUs are associated with a *home* band to which they have usage rights at all time. In order to communicate with each other, all SUs in the group must be tuned to the same band, being either their home band or another unused licensed band. While communicating in the home band, the SUG may decide to seek for spectrum opportunities in another band. This typically happens when, for example, any of the SUs judge that the quality of their current band is no longer acceptable. This can be done by continuously assessing and monitoring the quality of the band via some quality metrics, such as signal-to-noise ratio (SNR), packet success rate, achievable data rate, etc. That is, when the monitored quality metric drops below a threshold that can be defined *a priori*, the SUG is triggered to start seeking for

spectrum opportunities. When a new opportunity is discovered on another band, the group switches to that band and starts communicating on it. Now suppose the group is currently using a licensed band, not the home band. Then, upon the return of PUs to their band and/or when the quality drops below the threshold, SUs must vacate the licensed band by either switching back to their home band or by searching for new opportunities. Hereafter, we say that an *exploration event* is triggered when either (i) PUs return back to their licensed band, and/or (ii) the band's quality is degraded below the threshold. In the RL terminology, we therefore consider that the agent and the environment interact at each of a sequence of discrete time steps, each of which takes place at the occurrence of an exploration event.

Prior to using a licensed band, SUs must first sense the band to assess whether it is vacant, and if it is, then they can switch to and use it for so long as no PUs are present. Upon the detection of the return of PUs to their band, SUs must immediately vacate the band. OSA has great potentials for improving spectrum efficiency, but in order to enable it, SUs must be capable of *sensing*, the ability to observe and locate spectrum opportunities; *identifying*, the ability to analyze and characterize these opportunities; and *switching*, the ability to configure and tune to the best available opportunities.

In this work, we propose an OSA scheme that self-learns from interaction with the environment, and uses its acquired knowledge to locate the best spectrum opportunities (i.e., spectrum bands that are most likely to be available), thus achieving efficient utilization of spectral resources.

3. PROPOSED SINGLE-AGENT REINFORCEMENT LEARNING (RL) FOR OSA

Reinforcement Learning (RL) is the concept of learning from past and present experience to decide what to do best in the future. That is, the learner, also referred to as *agent*, learns from experience by interacting with the environment, and uses its acquired knowledge to select the *action* that maximizes a cumulative *reward* signal (the total reward that the environment gives rise to in the long run). RL is well suited for systems whose behaviors are, by nature, too complex to predict, but the reward, or reinforcement, resulting from taking an action can easily be assessed or observed. For example, in OSA, albeit it may be difficult to predict which spectrum band will be available in the near future, the reward resulting from the use of a spectrum band can easily be determined. The reward can, for example, be assessed through the amount of obtained throughput, the experienced interference, the packet success rate, etc. Thus, RL techniques are a natural choice for OSA where it is difficult to precisely specify an explicit model of the environment, but it is easy to provide a reward function.

RL is typically formalized in the context of Markov Decision Processes (MDPs). An MDP represents a dynamic system, and is specified by giving a finite set of states ($\mathcal{S}$), representing the possible states of the system, a set of control actions ($\mathcal{A}$), a transition function (δ), and a reward function (r). The transition function specifies the dynamics of the system, and gives the probability $\mathcal{P}_{ij}^{k}$ of transitioning to state s_j after taking action a_k while in state s_i. The dynamics are Markovian in the sense that the probability of the next state s_j depends only on the current state

s_i and action a_k, and not on any previous history. The reward function assigns real-numbers $r(s_i, a_k)$ to state-action pairs (s_i, a_k) so as to represent the immediate reward of being in state s_i and taking action a_k.

In this chapter, we provide an RL formulation of the OSA problem, and propose an RL scheme as a possible solution.

3.1. Markov Decision Process (MDP)

We formulate OSA as a finite MDP, defined by its state set $\mathcal{S}$, action set $\mathcal{A}$, transition function δ, and reward function r as follows:

State set. $\mathcal{S}$ consists of $m + 1$ states, $\{s_0, s_1, \ldots, s_m\}$. The SUG is said to be in state s_i when it is using band b_i at the current time step; i.e., no PUs are currently using band b_i. Note that state s_0 corresponds to when the group is communicating on its home band b_0. Throughout this work, the terms agent and SUG will be used interchangeably to mean the same thing. The same also applies to the terms state and band.

Action set. At every time step (i.e., an exploration event), while in state s_i, the agent can either choose to *exploit* by switching back to its home band b_0, or choose to *explore* by searching for new spectrum opportunities. If a decision is made in favor of exploration, then the agent senses an ordered sequence of bands $\{b_{k_1}, b_{k_2}, \ldots, b_{k_n}\}$, where $n = 1, 2, \ldots, m$, on a one-by-one basis until it finds, if any, the first available band. If there is one available, the agent switches to and starts using it until the the next time step. If none are available, then the agent switches back to b_0 at the end of the search. At the next time step, the same exploration

versus exploitation process repeats again. We will refer to n as the exploration index as it balances between exploration and exploitation; i.e., the larger the n, the more the exploration. Now by letting a_0 denote the action of returning to the home band b_0, and $a_k = \{b_{k_1}, b_{k_2}, \ldots, b_{k_n}, b_0\}$ the action of exploring new opportunities, the set $\mathcal{A}$ of all actions is $\mathcal{A} = \{a_0, a_1, \ldots, a_p\}$, where $p = \frac{m!}{(m-n)!}$. The index n can be viewed as a design parameter to be set *a priori*.

Transition function. $\delta : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the transition function, specifying the next state the system enters provided its current state and the action to be performed. Given any state, s_j, for action a_0, the transition function $\delta(s_j, a_0)$ equals s_0, and for any action $a_k = \{b_{k_1}, b_{k_2}, \ldots, b_{k_n}, b_0\}$, $k = 1, 2, \ldots, p$, the transition function $\delta(s_j, a_k)$ equals

$$\delta(s_j, a_k) = \begin{cases} s_0 & \text{w/ prob. } \prod_{i=1}^{n} \eta_{k_i} \\ s_{k_1} & \text{w/ prob. } 1 - \eta_{k_1} \\ s_{k_l} & \text{w/ prob. } \prod_{i=1}^{l-1} \eta_{k_i}(1 - \eta_{k_l}) \\ & \text{for } l = 2, 3, \ldots, n \end{cases}$$

For example, when $n = 2$, and the SU is in state s_j. If action $a_k = \{b_2, b_3, b_0\}$ is taken, then the user ends up in state s_2 (i.e., band b_2) with probability $1 - \eta_2$ (i.e., b_2 is available), ends up in state s_3 (i.e., band b_3) with probability $\eta_2(1 - \eta_3)$ (i.e., b_2 is occupied and b_3 is not), or ends up in state s_0 (i.e., band b_0) with probability $\eta_2\eta_3$ (i.e., both bands are not available).

It is important to reiterate that this function is only provided to generate samples of the OSA environment so as to evaluate our RL algorithm. That is, although in practice our RL technique will not need models to perform, we use models here

to generate samples of the environment's behavior to mimic an OSA environment. For example, in the evaluation section, it is assumed that the PU traffic follows a Poisson distribution, and hence, an ON/OFF renewal process model is used to mimic such an environment.

Reward function. $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ defines the reward function $r(s_i, a_k)$, specifying the reward the agent earns when taking action $a_k \in \mathcal{A}$ while in state $s_i \in \mathcal{S}$. The reward $r(s_i, a_k)$ also depends on the next state $s_j = \delta(s_i, a_k)$ the agent enters as a result of taking a_k while in state s_i. More specifically, the reward perceived by the agent when entering state s_j is a function of the quality level the SUG receives when using the band it ends up selecting. We therefore assume that each band b_j is associated with a quality level q_j, which can be determined via metrics like SNR, packet success rate, data rates, etc, and let $\phi(q_j)$ denote the reward (without including the cost of exploration yet) resulting from receiving q_j.

It is important to note that exploration also comes with a price. Recall that SUs are allowed to use any licensed band only if the band is vacant (no PUs are using it), and that discovery of opportunities is done through spectrum sensing. That is, SUs periodically (or proactively) switch to and sense certain bands to find out whether any of them is vacant or not. Unfortunately, during the sensing process, the system incurs some "sensing overhead", which can be of multiple types: energy consumed to perform sensing, delays resulting from switching across bands, throughput reduced as a result of ceasing communication, etc. By letting c_{ij} denote the cost incurred as a result of exploring band b_j while in band b_i, and s_j denote

the next state, $\delta(s_i, a_k)$, the reward function $r(s_i, a_k)$ can now be written as

$$r(s_i, a_k) = \begin{cases} \phi(q_{k_1}) - c_{ik_1} \\ \qquad\qquad \text{w/ prob. } 1 - \eta_{k_1} \\ \\ \phi(q_{k_l}) - c_{ik_1} - \sum_{t=1}^{l-1} c_{k_t k_{t+1}}, l = 2, 3, \ldots, n \\ \qquad\qquad \text{w/ prob. } \prod_{t=1}^{l-1} \eta_{k_t}(1 - \eta_{k_l}) \\ \\ -c_{ik_1} - \sum_{t=1}^{n-1} c_{k_t k_{t+1}} - c_{k_n 0} \\ \qquad\qquad \text{w/ prob. } \prod_{t=1}^{n} \eta_{k_t} \end{cases}$$

where $a_k = \{b_{k_1}, b_{k_2}, \ldots, b_{k_n}, b_0\}$, $k = 1, 2, \ldots, p$.

Consider a special scenario where the PU traffic load is the same and equal to η for all bands b_j. Suppose that $\phi(q_j) = q$ for all bands b_j, and that the cost c_{ij} incurred when switching from band b_i to band b_j is equal to c for all i, j. Let $\bar{E}$ denote the expected value of the reward function $r(s_i, a_k)$ normalized with respect to c (i.e., $\bar{E} = E[r(s_i, a_k)]/c$). One can now express $\bar{E}$ as

$$\bar{E} = (\frac{q}{c} - 1)(1 - \eta) + \frac{q}{c}(\eta - \eta^n) + \frac{\eta^{n+1} - 2\eta + \eta^2}{1 - \eta} \tag{3.1}$$

Using Eq. (3.1), one can easily see that the reward that the agent receives increases monotonically with the exploration index n when $\frac{q}{c} > \frac{\eta}{1-\eta}$ (or equivalently $\eta < \frac{q}{q+c}$), decreases monotonically with the index n when $\frac{q}{c} < \frac{\eta}{1-\eta}$ (or equivalently $\eta > \frac{q}{q+c}$), and is independent of the index n when $\frac{q}{c} = \frac{\eta}{1-\eta}$ (or equivalently $\eta = \frac{q}{q+c}$). Therefore, for a given PU traffic load, the optimal exploration index n that the agent should use so as to maximize its reward depends on the ratio q/c (or equivalently $\frac{q}{q+c}$).

Intuitively, when the network is lightly loaded (η is small), the chances of finding available bands are high, and hence, it is rewarding to explore for more bands.

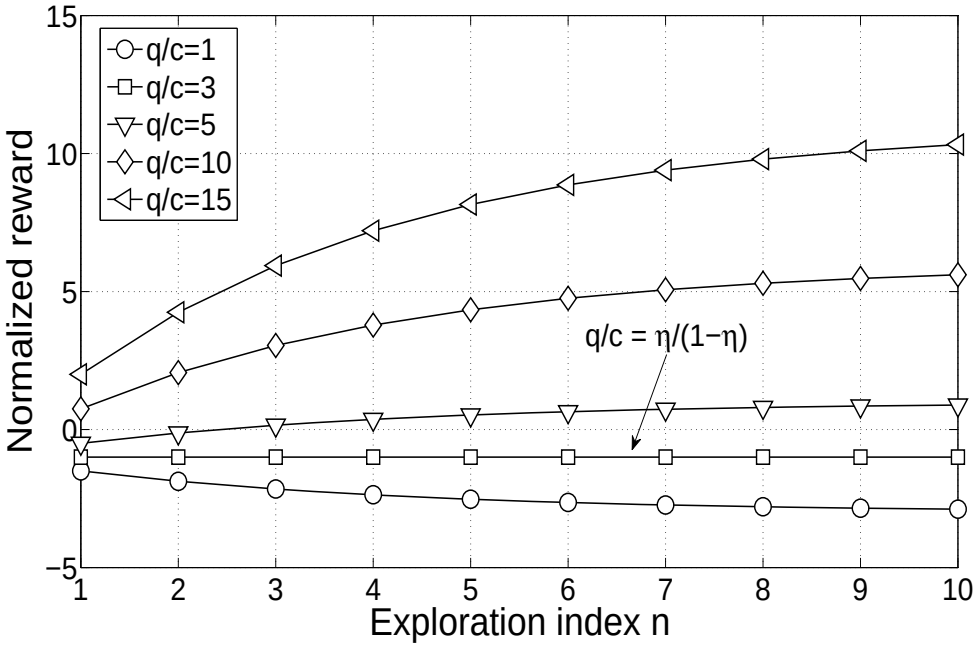


FIGURE 3.1: Reward as a function of exploration index n: $\eta = 0.75$

This explains why for small η values (i.e., $\eta < \frac{q}{q+c}$), the higher the exploration index, the higher the reward. Now when the network is heavily loaded (η is large), the chances of finding empty bands are low, and hence, it is not rewarding to explore for more bands. This explains why for high values of η (i.e., $\eta > \frac{q}{q+c}$), the lower the exploration index, the higher the reward. That is, the expected reward is not worth the exploration cost for high values of η. Note that as the cost c goes to zero, $\frac{q}{q+c}$ goes to 1. Therefore, when the cost is negligible, $\eta < \frac{q}{q+c}$ holds for all η since $\frac{q}{q+c} \approx 1$, and thus, the reward increases monotonically with the exploration index n regardless of the PU load η.

As an example, we plot in Fig. 3.1 the reward as a function of the index n for different values of the q/c ratio. The PU traffic load η is set to 0.75 (i.e.,

$\frac{\eta}{1-\eta} = 3$). As expected, when $\frac{q}{c} = \frac{\eta}{1-\eta} = 3$, the index value has no effect on the reward value. On the other hand, when $\frac{q}{c} > 3$, the higher the index, the higher the reward; whereas, when $\frac{q}{c} < 3$, the higher the index, the lower the reward.

3.2. Learning-Based OSA Scheme

The goal of the agent is to learn a policy, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, for choosing the next action a_i based on its current state s_i that produces the greatest possible expected cumulative reward. A cumulative reward R is typically defined through a discount factor γ, $0 \leq \gamma < 1$, as $\sum_{t=0}^{\infty} \gamma^t r(s_{i+t}, a_{i+t})$. Because it is naturally desirable to receive rewards sooner than later, the reward is expressed in a way that future rewards are discounted with respect to immediate rewards.

A function, $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, is defined for each state-action (s_i, a_k) pair as the maximum discounted cumulative reward that can be achieved when starting from state s_i and taking action a_k according to the optimal policy. Hence, given the Q-function, it is possible to optimally act by selecting actions that maximize $Q(s_i, a_k)$at each state. Q can be recursively constructed as follows. The Q-learning algorithm learns an estimate $\hat{Q}$ of the optimal Q-function by selecting actions and observing their effects. In particular, each step in the environment involves taking an action a_k in state s_i and then observing the following state and the resulting reward. Given this information, Q is updated via the following equation:

$$\hat{Q}_l(s_i, a_k) \leftarrow \quad (1 - \alpha_l)\hat{Q}_{l-1}(s_i, a_k) + \alpha_l\{r(s_i, a_k) + \gamma \max_{k'} \hat{Q}_{l-1}(\delta(s_i, a_k), a_{k'})\}$$

where $\alpha_l = 1/(1 + visits_l(s_i, a_k))$ and $visits_l(s_i, a_k)$ is the total number of times this state-action pair has been visited up to and including the lth iteration. This

approximation algorithm is guaranteed to converge to the optimal Q-function in any MDP, given the appropriate exploration during learning [31].

4. EVALUATION OF SINGLE-AGENT RL

In this chapter, we study the proposed single-agent Q-learning scheme by evaluating and comparing its performance to a random access scheme. The random access scheme will be used here as a baseline for comparison, and is defined as follows. Whenever an exploration event is triggered, the SUG, using the random access approach, selects a spectrum band among all bands randomly. If the selected band is idle, then the group uses it until the return of a PU. Otherwise, i.e., if the selected band happens to be busy, then the group goes back to its home band. This process repeats until an idle band is found.

4.1. Simulation Settings

We consider that the spectrum is divided into m non-overlapping bands, and that each band is associated with a set of PUs. We model PUs' activities on each band as a renewal process alternating between ON and OFF periods, which represent the time during which PUs are respectively present (ON) and absent (OFF). For each spectrum band b_j, we assume that ON and OFF periods are exponentially distributed with rates λ_j and μ_j, respectively. Note that the primary traffic load η_j on band b_j can be expressed as $\mu_j/(\mu_j + \lambda_j)$. Recall that the power of RL lies in its capability to converge to approximately an optimal behavior without needing prior knowledge of PUs' traffic behavior. The exponential distributions will, however, be used to generate samples in-order to evaluate our learning techniques using simulated interaction. Throughout this section, we characterize the PU traffic system load by

$\bar{\eta} = \frac{1}{m}\sum_{i=1}^{m}\eta_i$ (denoted as `pbar` in figures) and $CoV = \sigma/\bar{\eta}$, which respectively denote the average and the coefficient of variation of PU traffic loads across all bands, where σ denotes the standard deviation of traffic loads.

At every exploration event, while in state s_i, the agent can either choose to *exploit* by switching back to its home band b_0, or choose to *explore* by searching for new spectrum opportunities. If a decision is made in favor of exploration, then the agent senses an ordered sequence of bands $\{b_{k_1}, b_{k_2}, \ldots, b_{k_n}\}$, where $n = 1, 2, \ldots, m$, on a one-by-one basis until it finds, if any, the first available band. If there is one available, the agent switches to and uses it until the the next time step. If none are available, then the agent switches back to b_0 at the end of the search. At the next time step, the same exploration vs. exploitation process repeats again. We will refer to n as the exploration index as it balances between exploration and exploitation; i.e., the larger the n, the more the exploration.

4.2. Effect of Primary-User Traffic Load

We begin by studying the effect of PU traffic load $\bar{\eta}$ on the achievable throughput. Figure 4.1 plots the total throughput, normalized with respect to the maximal achievable throughput[1], that the SUG achieves as a result of using our Q-learning and the random access schemes for two different PU traffic loads: $\bar{\eta} = 0.5$ and $\bar{\eta} = 0.8$. The measured throughput is based on what the SUG receives from the

[1]The maximal/ideal achievable throughput corresponds to when the agent knows exactly where spectrum opportunities are; i.e., the agent always knows which bands are available, and thus, it exploits them without any cost.

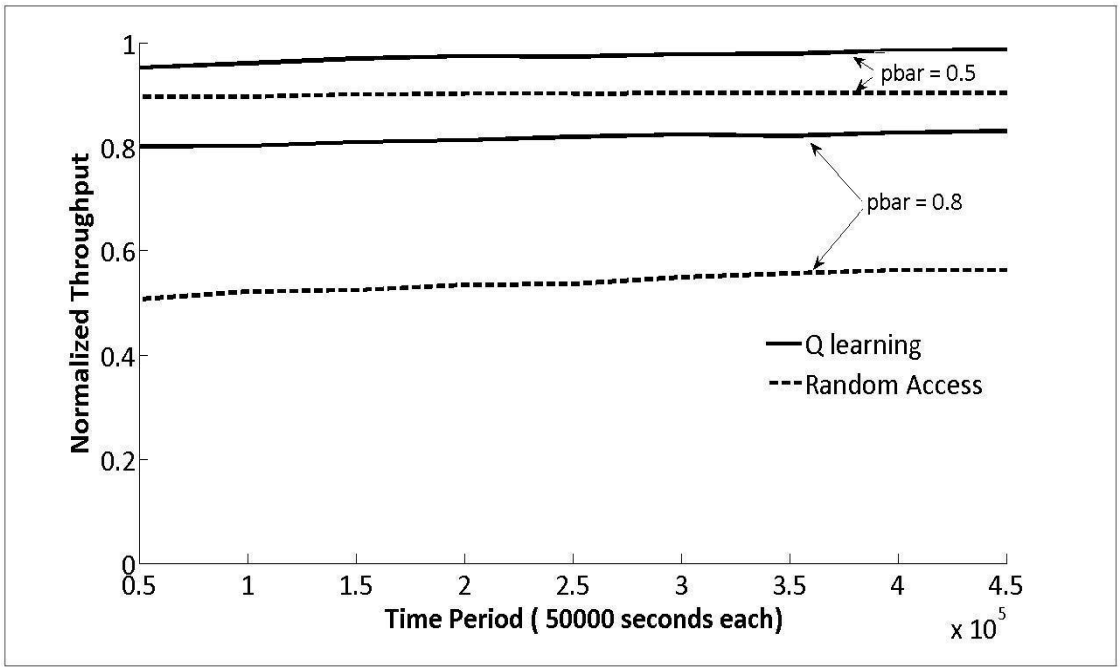


FIGURE 4.1: Throughput behavior under two different primary-user traffic loads pbar $\equiv \bar{\eta} = 0.5$ and 0.8: $m = 7$, $CoV = 0.5$

m licensed bands only; i.e., not accounting for the home band. In this simulation scenario, CoV is set to 0.5, exploration index n is set to 3, and the total number of bands m is set to 7. First, as expected, note that the higher the $\bar{\eta}$, the lesser the achievable throughput under both schemes. However, regardless of the PU load, the Q-learning scheme always outperforms the random scheme. Also, note that the more loaded the system is, the higher the difference between the throughput achievable under Q-learning and that achievable under random access (e.g., the throughput gain is higher when $\bar{\eta} = 0.8$).

To further illustrate the effect of $\bar{\eta}$ on the performance of the proposed Q-learning scheme, we plot in Fig. 4.2 the throughput gain as a function of $\bar{\eta}$. Note that the throughput gain increases as the PU traffic load increases. In other words,

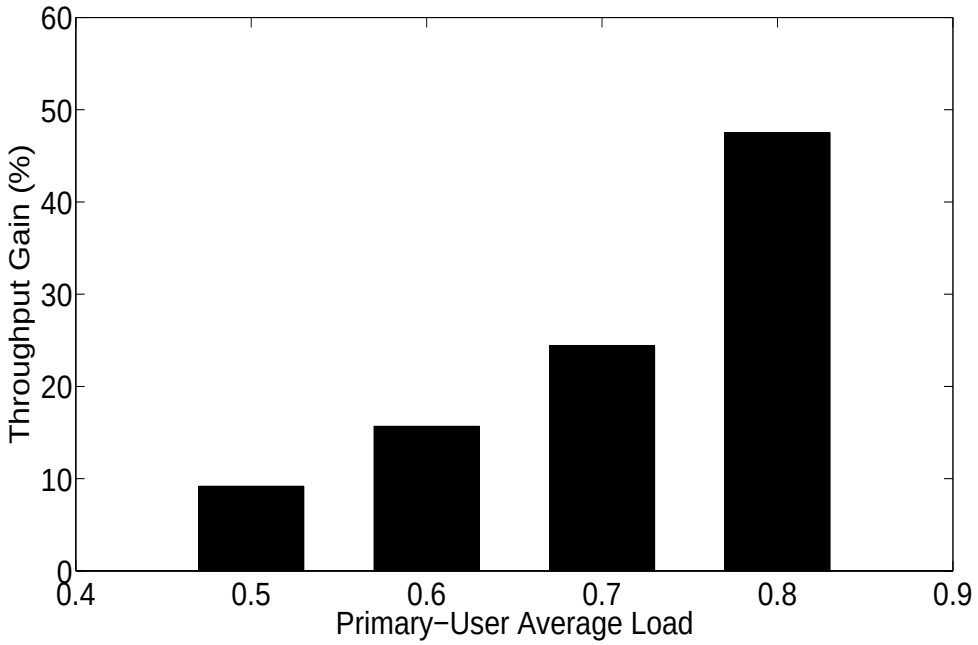


FIGURE 4.2: Throughput gain as a function of the primary-user average loads $\bar{\eta}$: $m = 7$, $CoV = 0.5$

the Q-learning scheme performs even better under heavily loaded systems. This can be explained as follows. When $\bar{\eta}$ is high; i.e., when spectrum opportunities are scarce, the learning capability of the Q-learning scheme allows the OSA agent to efficiently locate where the opportunities are, whereas the random access scheme leads to a lesser throughput since it accesses the bands randomly. When $\bar{\eta}$ is small, on the other hand, the random access scheme is able to achieve high throughput since spectrum opportunities are too many to miss even when bands are selected unintelligently.

To summarize, these obtained results show that the proposed Q-learning scheme is capable of achieving anywhere between 80% to 95% of the maximal achievable throughput by learning from experience, and without requiring prior knowledge of

the environment. The results also show that the scheme achieves high throughput performance even under heavy traffic loads.

4.3. Effect of Primary-User Load Variability

Figure 4.3 plots the total throughput that the SUG achieves under our proposed Q-learning and the random access schemes for two different PU load variations: $CoV = 0$ and $CoV = 0.6$. (Recall that CoV reflects the variation of loads across different bands; i.e., the higher the CoV, the higher the variation.) Note that when the $CoV = 0.6$, the Q-learning scheme achieves about 90% of the maximal/ideal throughput by simply locating and exploiting unused opportunities through learning from experience, whereas the random access scheme achieves only about 60%. When $CoV = 0$ (i.e., all bands experience identical loads), the Q-learning and the random access achieve approximately about 64% and 55%, respectively. As expected, the throughput gain increases with the coefficient of variation. As shown in Fig. 4.3, the gain is higher when $CoV = 0.6$ than when $CoV = 0$.

To further illustrate the effect of PU load variability on the achievable throughput, we show in Fig. 4.4 the throughput gain for different values of CoVs. The CoV is varied from 0 to 0.6. The average PU traffic load, $\bar{\eta}$, is set to 0.8 (which implies that only 20% of the spectrum is available for the SUG). The total number of bands is set to $m = 7$ and the exploration index is taken to be $n = 3$. Observe that the higher the variation of PU loads across different bands, the higher the throughput gain; i.e., the higher the throughput the agent/group can achieve when compared with that achievable under the random access scheme. This can be explained as follows. When the average of PU traffic loads is maintained the same, a high vari-

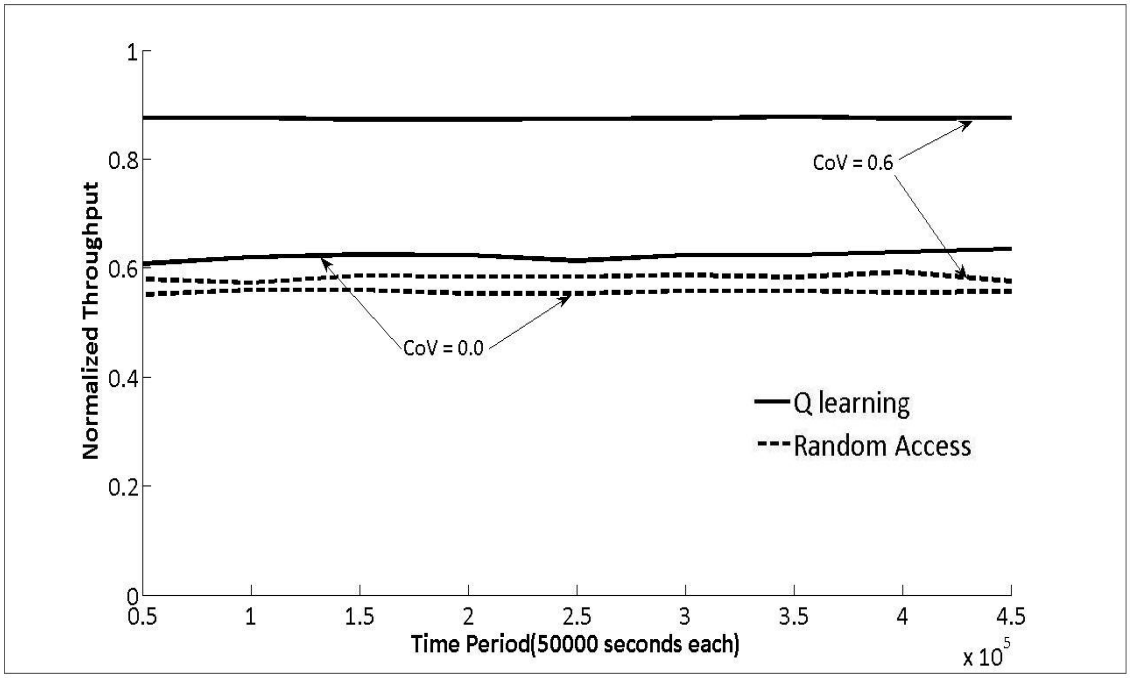


FIGURE 4.3: Achievable throughput under Q-learning and random access schemes: $\bar{\eta} = 0.8$, $m = 7$, $n = 3$.

ation in the loads across different bands increases the likelihood of finding highly available spectrum bands. This, on the other hand, also increases the likelihood of finding spectrum bands with lesser opportunities. With experience, the Q-learning scheme learns about, and starts exploiting, these more available bands, yielding then more throughput. When the load variation is low, on the other hand, the learning algorithm achieves less throughput because all bands are equally-loaded, and hence, there is no special (i.e., more available) bands that the algorithm can learn about. This explains why both the Q-learning and the random access achieve similar performances when all bands have identical loads. The gain can, however, reach up to 50% when bands have different loads (e.g., $CoV = 0.6$), as shown in Fig. 4.4.

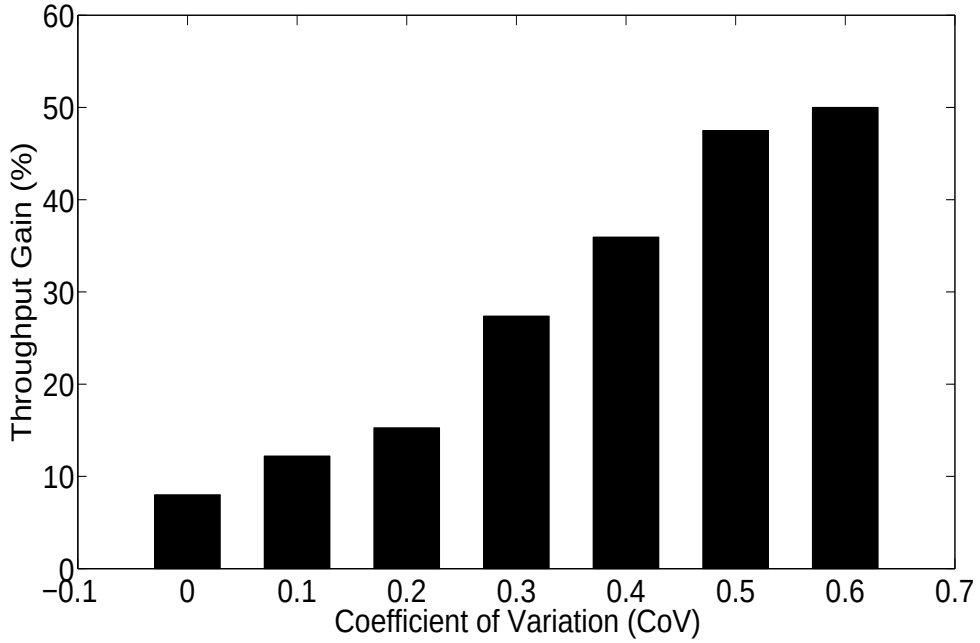


FIGURE 4.4: Throughput gain as a function of primary-user load variability: $m = 7$, $\bar{\eta} = 0.8$.

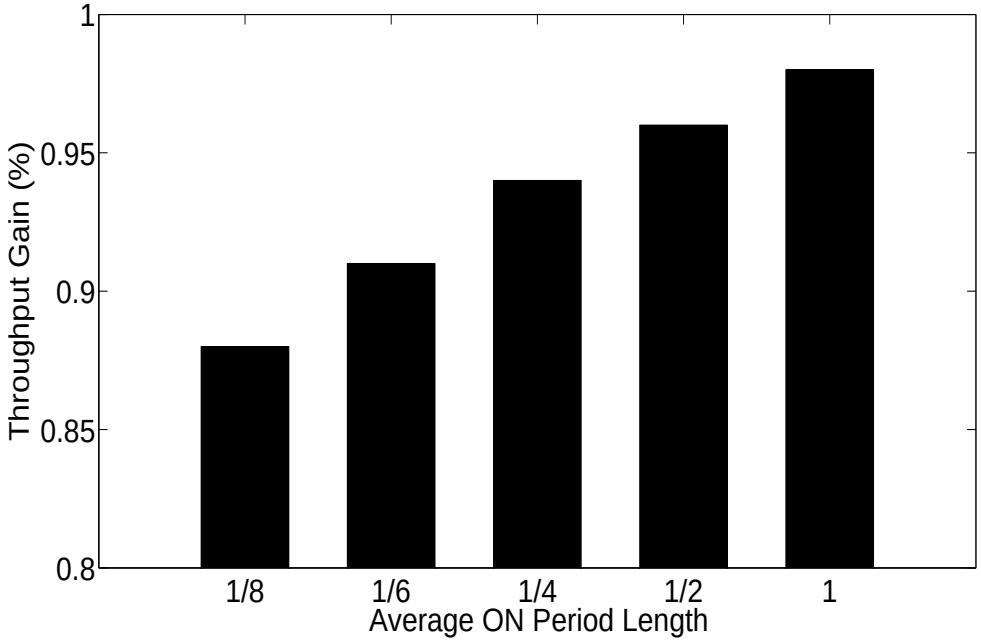


FIGURE 4.5: Throughput gain as a function of ON/OFF period lengths: $\bar{\eta} = 0.5$, $CoV = 0.2$, $m = 7$, $n = 3$.

4.4. Effect of Primary-User Load ON/OFF Period

In this section, we study the effect of ON/OFF period lengths on the performance of the Q-learning scheme. We vary the lengths of ON and OFF periods while keeping the PU traffic loads η_i the same for all i. Since the PU load is kept the same, an increase in OFF periods leads to an increase in ON periods as well, and vice versa. The normalized throughput that the Q-learning scheme achieves is shown in Fig. 4.5 for different values of ON period lengths. Here, CoV is set to 0.2, $\bar{\eta}$ is set to 0.5, n is set to 3, and m is set to 7.

Note that the higher the length of ON/OFF periods, the higher the throughput gain. Note also that having short ON/OFF periods forces the agent to make frequent transitions so as to find available spectrum bands. Whereas, when ON/OFF periods are long, the transitions are not that often, thus leading to less switching overhead, which yields more achievable throughput. In other words, when the length of ON/OFF periods increases, the SUG can possess the available spectrum bands for longer periods of time. When the lengths of ON/OFF periods are low, the SUG has the spectrum band available to it only for a short period of time, leading to frequent transitions across different bands.

4.5. Q-learning Optimality: Exploration Index n

In this section, we study the effect of the exploration index n on the behavior of the Q-learning scheme. Recall that the index n is a design parameter to be chosen and set *a priori*, which can take on any number less than or equal to the number

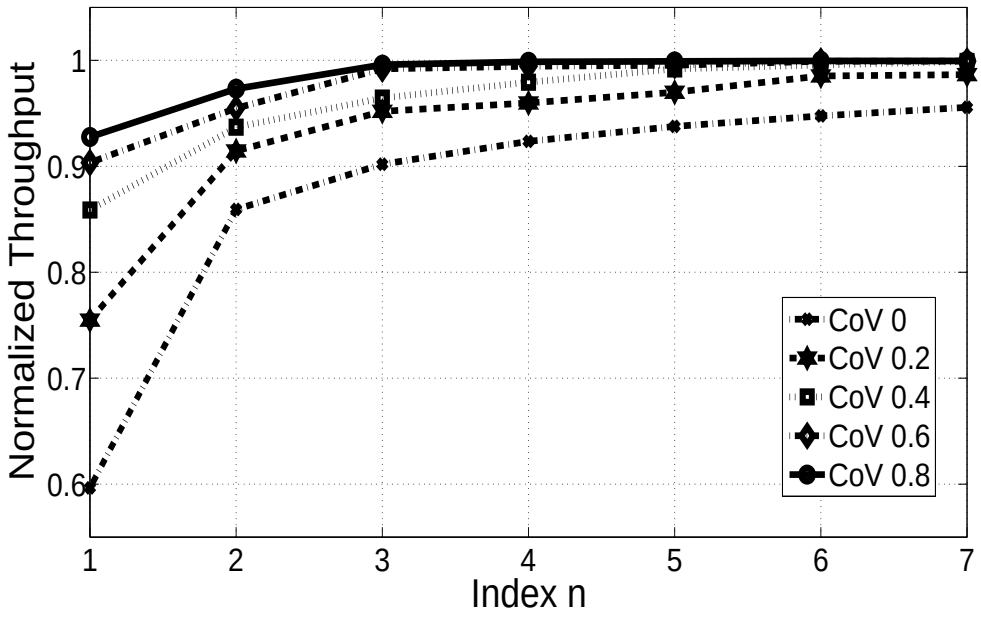


FIGURE 4.6: Effect of index n on throughput: $\bar{\eta} = 0.8$, $m = 7$.

of available bands m. This parameter balances between two conflicting objectives: the desire of increasing the chances of finding available bands (i.e., by increasing n), and the desire to reduce the incurred overhead/cost due to scanning (i.e., by decreasing n).

Figure 4.6 plots the normalized throughput as a function of n for different values of CoV. Note that as the index n increases, the achievable throughput first increases with n, then flattens out. This means that increasing the number of scanned/searched bands beyond a certain threshold does not necessarily yield more achievable throughput. For example, when CoV is above 0.6, the figure shows that the SUG can no longer benefit from increasing its exploration index n when the index reaches approximately 3. As explained in Section 4.3., note that the higher

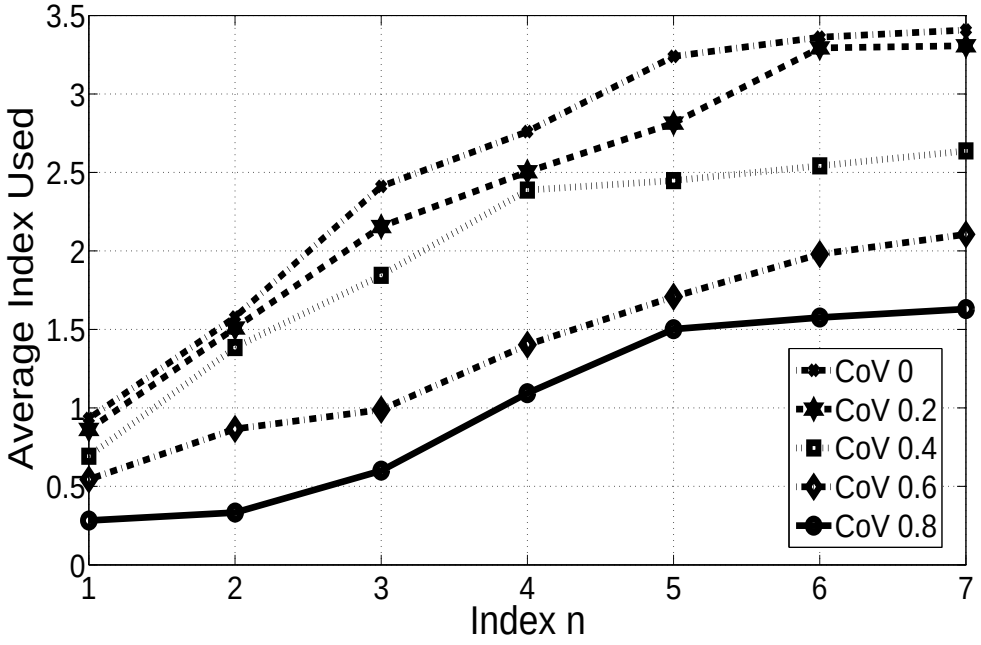


FIGURE 4.7: Index used as a function of index n: $\bar{\eta} = 0.8$, $m = 7$.

the CoV, the higher the throughput.

To further study this behavior, for each exploration index n scenario, we measured the average number of bands that are actually scanned before finding one available band. We refer to this number as *average index used*. Figure 4.7 shows the average index used for finding available bands as a function of the exploration index n for different values of CoV. Note that as the exploration index n (i.e., the number of allowable bands that can be scanned) increases, the average index used to find an available band (i.e., the actual, measured number of scanned bands) first increases then flattens out. This means that even when the SUG is allowed to scan all bands, it ends up visiting only a few before finding an available one as a result of using its learning capabilities. The figure also shows that the higher the CoV,

the smaller the actual index used to find an available band. Therefore, the learning capabilities allow to find spectrum opportunities quickly, thus limiting the incurred exploration overhead.

To summarize, we conclude that there exits an optimal index beyond which throughput can no longer be increased even when the agent is allowed to scan more bands. This optimal index is decided by the Q-learning by striking a balance between the need to increase the chances of finding opportunities and the desire to keep the searching overhead minimum. It is important to mention that setting the exploration index n higher than the optimal index still allows the agent to achieve the maximum throughput; i.e., the throughput that would also be achieved when the exploration index is set to an optimal one. However, the lower the n, the lesser the complexity of the Q-learning in terms of action set space and convergence time. Therefore, it is very crucial that one determines the optimal (or near-optimal) index so as to set the Q-learning scheme accordingly.

Let us now study how this optimal index varies under different PU traffic loads. Figures. 4.8 and 4.9 plot the optimal index n as a function of CoV (variation of primary-traffic loads) and $\bar{\eta}$ (average of primary-traffic loads), respectively. Figure 4.8 shows that the optimal index decreases as the coefficient of variation CoV increases. When the average of PU traffic loads is kept the same, high values of CoV (i.e., high variations in the loads across different bands) increase the chances of finding highly available spectrum bands. With experience, the Q-learning scheme can quickly learn and locate where these more available bands are, requiring then lesser number of scanned bands; i.e., a lower optimal index. Now when the load variation is low, the Q-learning scheme needs to scan more bands to be able to find one available since all of them are equally-loaded, and hence, there is no special

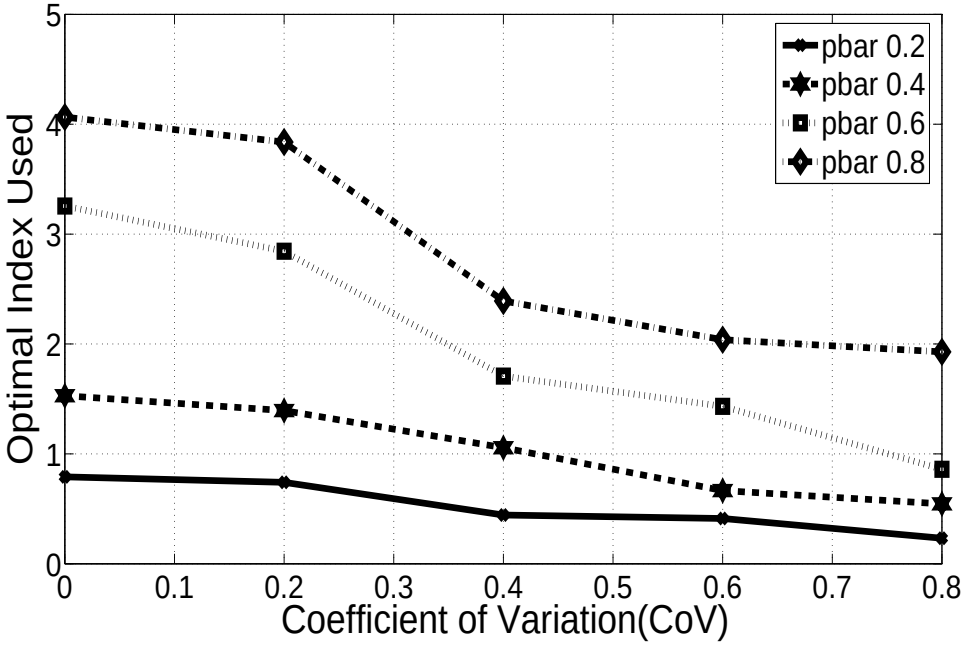


FIGURE 4.8: Index used as a function of CoV.

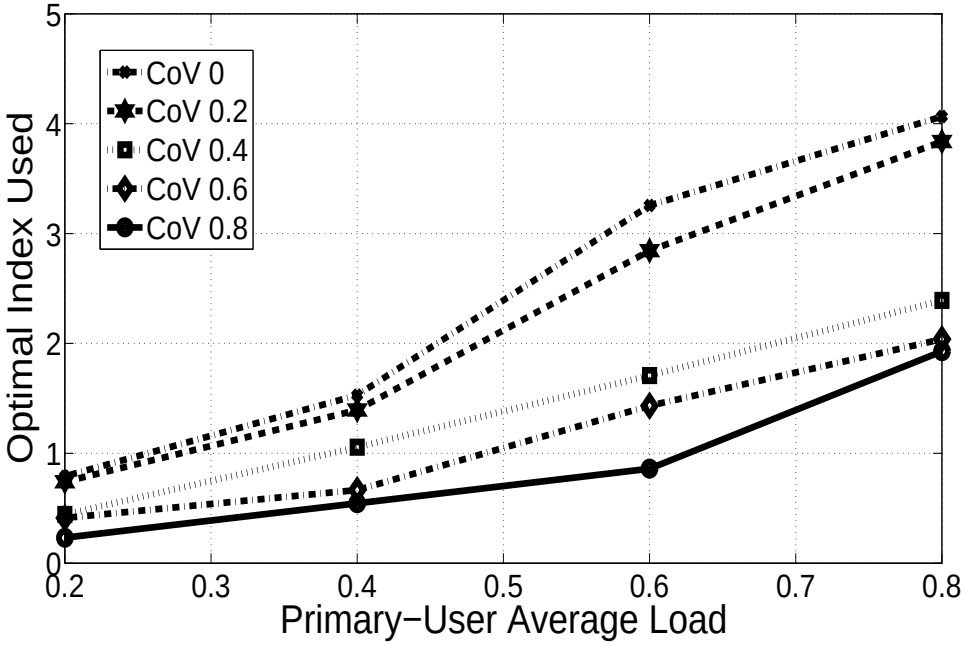


FIGURE 4.9: Index used as a function of `pbar` ($\bar{\eta}$).

(i.e., more available) bands that the algorithm can learn about. This explains why the optimal used index is relatively high when bands have similar loads.

Figure 4.9 shows that the optimal index increases with the average $\bar{\eta}$ of the PU traffic loads, which can be explained as follows. When the system is highly-loaded (i.e., $\bar{\eta}$ is high), spectrum opportunities are scarce. Therefore, regardless of how good the learning capabilities are, the SUG still needs to scan quite a few bands before finding an available band. It is when the system is lightly-loaded that learning can be effective as it can now quickly locate where these available bands are, thus needing lesser bands to scan to find one available. This explains why the optimal index is small under lightly loaded systems.

5. PROPOSED MULTI-AGENT RL FOR OSA

In the previous chapters, we have tested the effectiveness of the Q-learning scheme in terms of exploiting the spectrum opportunities. This is done by evaluating the Q-learning scheme for a single secondary-user group and comparing the Q-learning scheme's performance under different environmental conditions with the random access scheme. In this chapter we study the effect of multiple secondary-user groups in the OSA environment and compare its performance under the three different access schemes : Non-cooperative, Cooperative, and Random.

For this work, we formulate OSA as a finite MDP, defined by its state set $\mathcal{S}$ consisting of one state s only ($\mathcal{S} = \{s\}$), the action set $\mathcal{A}$ and the reward function r described as follows.

Action set. At each time step, the agent chooses an action from the action set $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$, where m is the number of bands. The number of actions is equal to the number of spectrum bands in the system. Taking action a_i while using spectrum band b_j makes an SUG enter and use spectrum band b_i.

Reward function. The reward perceived by the agent when taking action a_i and entering spectrum band b_i is a function of the quality level the SUG receives when using the band. We assume that each band b_i has its own bandwidth capacity V_i, and when more than one SUG uses a spectrum band, the bandwidth is equally divided among all the SUGs using the band. For example, if there are a total of 3 SUGs, A, B, and C, each taking action i, j, and k respectively, then the reward of SUG A, denoted by ra_{ijk}, can be calculated as

$$ra_{ijk} = \begin{cases} V_i/3 & \text{when } i = j = k \\ V_i/2 & \text{when } i = j \neq k \text{ or } i = k \neq j \\ V_i & \text{when } i \neq j \neq k \end{cases}$$

Non-cooperative Q-learning. The goal of the agent is to learn a policy, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, for choosing the next action a_i that produces the greatest possible expected cumulative reward. A function, $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, is defined so that its value for each state-action (s, a_i) pair corresponds to the maximum discounted cumulative reward that can be achieved when starting from state s and taking action a_i. Q can be constructed recursively [31] as follows.

$$Q(s, a_i)(t+1) = \quad Q(s, a_i)(t) + \alpha \times (r(s, a_i) - Q(s, a_i)(t))$$

where $0 < \alpha < 1$ is the learning rate. When using the non-cooperative Q-learning scheme, each SUG calculates its Q table independently from other SUGs.

Action selection. The action selection mechanism plays a very important role in Q-learning. During the learning process, this selection mechanism is what enables the agent to choose its actions. We consider the ϵ-greedy exploration as the action selection mechanism, where the action corresponding to the highest Q value in that time step is chosen with a probability of $(1-\epsilon)+\epsilon/m$, and any other action from the action set $\mathcal{A}$ is chosen with a probability of ϵ/m. The ϵ-greedy mechanism balances between exploration and exploitation.

Probability vector. Based on the ϵ-greedy exploration, we define the probability vector over the action set as follows. $X = (x_1, x_2, \ldots, x_m)$, where x_i is the probability

of taking action i

$$x_i = \begin{cases} (1-\epsilon) + \epsilon/m & \text{if } Q_i \text{ is the highest value} \\ \\ \epsilon/m & \text{otherwise} \end{cases}$$

where again m is the number of actions.

Cooperative Q-learning. Our multi-agent cooperative scheme is based on the multi-agent Q-learning approach derived in [34]. To illustrate, suppose that SUG A with probability vector X is going to cooperate with two other SUGs, B and C, with probability vectors Y and Z respectively. The Q table entry for SUG A choosing action i can be calculated as [34]:

$$Q(s,a_i)(t+1) = Q(s,a_i)(t) + x_i(t)\alpha[(\sum_{j=1}^{j=m} y_j(t) \sum_{k=1}^{k=m} (ra_{ijk})(z_k(t))) - Q(s,a_i)(t)]$$

Similarly, each SUG can compute its Q table values based on the probability vectors of the other SUGs.

6. EVALUATION OF MULTI-AGENT RL

In this chapter, we evaluate the performance of the proposed schemes. We show the importance of cooperation in multi-agent OSA systems by comparing the per SUG average received throughput of the cooperative scheme with that of a non-cooperative scheme. Specifically, we study the effect that cooperation has on network load balancing by allowing SUGs to make better action decision, leading to more effective exploitation of bandwidth opportunities. This also ensures fairness among SUGs by making sure that all SUGs receive (approximately) equal throughput shares.

6.1. Simulated Access Schemes

We consider that the spectrum is divided into m non-overlapping spectrum bands with ϕ SUGs. We mimic the presence of PUs by considering different spectrum bands with different bandwidth capacities. Let V_j denote the bandwidth capacity of band j. A spectrum band with a higher bandwidth capacity is meant to have a lower PU activity, and vice versa. We consider a time-slotted system, and assume that SUGs interact with the environment in accordance with these time slots. That is, SUGs can only enter or leave a band at the beginning and at the end of these time steps. We now summarize the three access schemes that are evaluated in this section.

Random Access Scheme. At the end of each time slot/step, an SUG using the random access scheme selects a spectrum band among the m available bands

randomly, and uses it during the next time slot. If more than one SUG happen to select the same spectrum band, they share the bandwidth of the band equally.

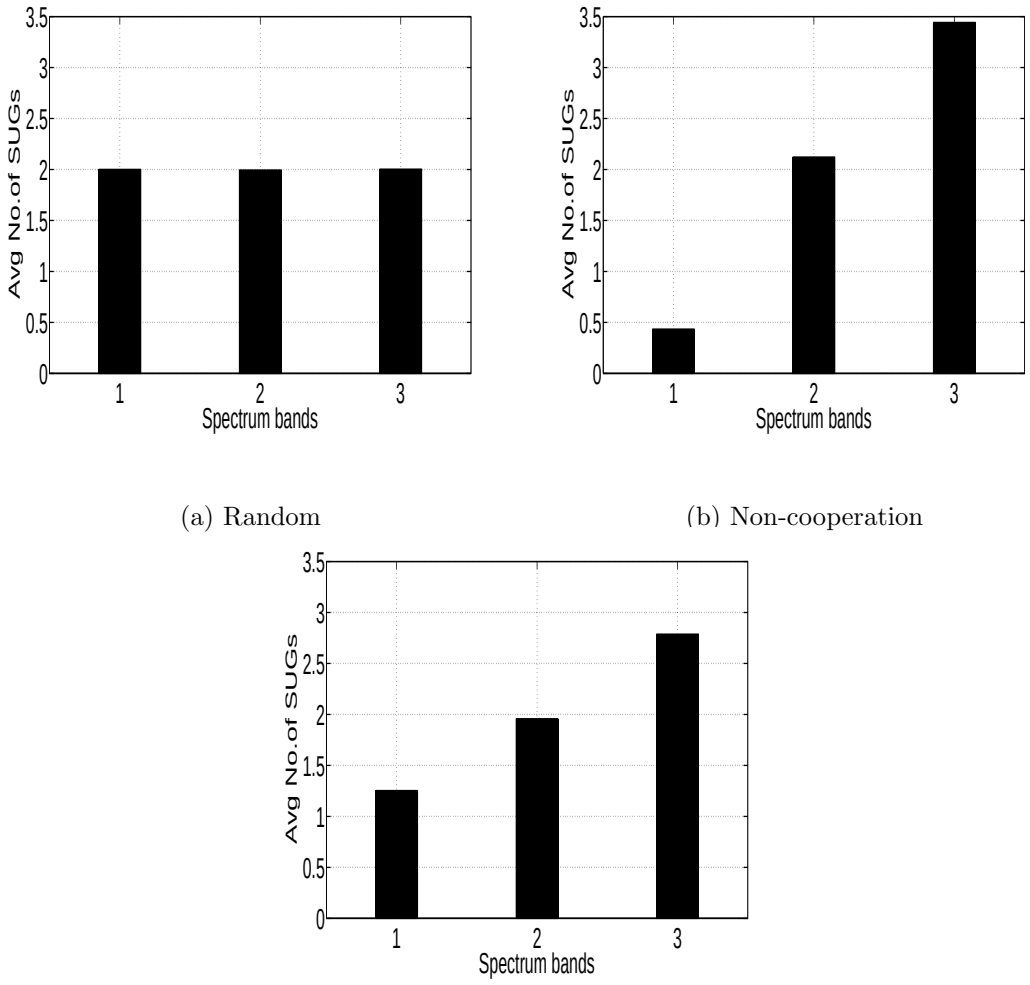
Non-cooperative Access Scheme. In the non-cooperative access scheme, each SUG uses the non-cooperative Q-learning policy discussed in Chapter 5. to create and update its own Q table. Each SUG enters the environment and takes actions based on its own Q table without cooperating with any of the other SUGs. When two or more SUGs choose the same band during the same time step, they share its bandwidth equally. Although the SUGs are typically unaware of the other agent's actions and act independently, the effect of the other SUG's actions are reflected in the reward that the SUGs receive from the spectrum band.

Cooperative Access Scheme. In the cooperative access scheme, each SUG maintains its own Q table using the cooperative multi-agent Q-learning, discussed in Chapter 5.. Here, an agent's Q table is formulated by taking into account the probabilities associated with the actions of the other SUGs with which it cooperates. In this case, at each time step, the SUG is provided with the probability vector of every other SUG with which it cooperates. The tradeoff here is between the communication overhead caused by extra traffic needed for exchanging the probability vectors among the cooperating SUGs and the performance gains due to improved action selections because of cooperation.

6.2. Cooperation Vs. Non-cooperation

First, we consider a OSA system with $m = 3$ spectrum bands and $\phi = 6$ SUGs. Bandwidth capacities are set to $V_j = [5\ \ 10\ \ 15]$. In this scenario, an ideal balanced

spectrum load is reached when each of the SUGs get a reward of 5 units, which implies that the 1^{st} band has 1 SUG, the 2^{nd} has 2 SUGs, and the 3^{rd} band has 3 SUGs. We simulate the three different access schemes for this scenario, and plot the average number of SUGs (averaged over 10000 episodes) in each of the three spectrum bands (i.e., the distribution of SUGs) in Fig. 6.1.



(a) Random

(b) Non-cooperation

(c) Full cooperation

FIGURE 6.1: SUG distribution: $m = 3$, $\phi = 6$, $V_j = [5\ \ 10\ \ 15]$.

The figure shows the average number of SUGs that end up choosing each of the three spectrum bands for each of the three studied schemes. It can be observed that the fully cooperative access scheme leads to the ideal balanced system load. As explained earlier, this is because in the fully cooperative method, each SUG accounts for all the possible actions that could be taken by its counterparts when making a decision. On the other hand, when SUGs do not cooperate, they may not select the best available band, as they have no clue what other SUGs will select, leading to a lesser balanced load distribution when compared with that of the cooperative scheme. Clearly and as expected, the Random access scheme results in an equally distributed SUGs among all bands, leading to the worst load balance when compared with the other two schemes[2].

Fairness is another important metric that we also evaluate in this work. To do this, we plot in Fig. 6.2 the coefficient of variation (CoV) of the received rewards of all the SUGs as a function of time period (each time period corresponds to 500 epochs). Observe that the fully cooperative access scheme has the lowest CoV among the three schemes. The lower the CoV, the closer the SUGs' received rewards are to one another, indicating a fairer access scheme. It can also be seen that the CoV of the non-cooperative access scheme is approximately twice that of the fully cooperative access scheme, and the CoV of the random access scheme is substantially higher than the other two. Therefore, cooperation improves performances not only in terms of network load balancing, but also in terms of ensuring fairness among all SUGs.

[2]We want to mention that these above results do not account for the communication overhead caused by message exchange needed to share the probability vectors among cooperative SUGs.

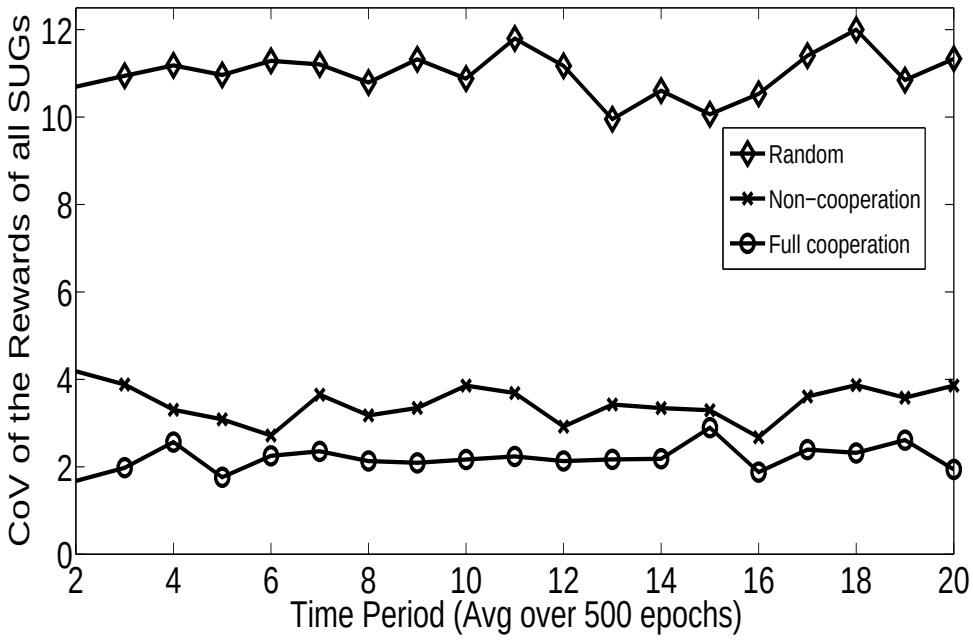


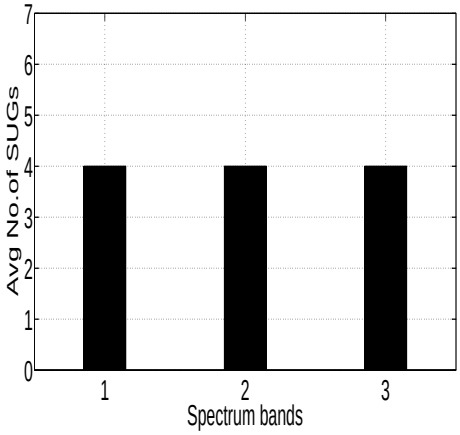
FIGURE 6.2: Coefficient of variation of the rewards of all the SUGs at each time period: $m = 3$, $\phi = 6$, $V_j = [5 \;\; 10 \;\; 15]$.

6.3. Impact of Degree of Cooperation

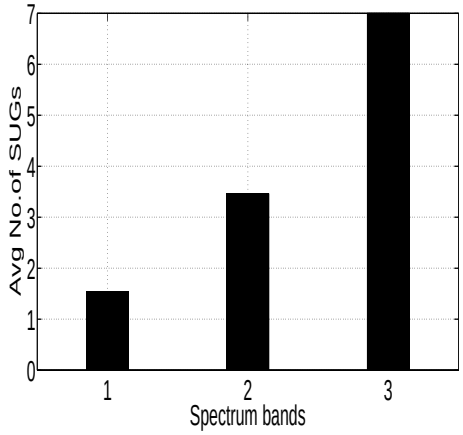
Recall that cooperation increases the performance because it allows the SUGs to make a better decision when selecting their next actions. This is because the SUGs take into account what other SUGs will select when making their action decisions. However, acquiring such information would necessitate the exchange of messages among cooperative SUGs, which clearly incurs extra overhead. Therefore, the challenge is to strike a good balance between the desire for a higher level of cooperation that enables a better action selection and the need for a lower level of cooperation so as to keep the cooperation overhead to a minimum. Cooperation overhead comes from the extra traffic needed to exchange the probability vectors and also from the computing delay/time resulting from solving the complex equations involved in updating the Q table entries of the cooperative SUGs.

We now study the impact of degree of cooperation on the achievable performances of a OSA system with $m = 3$ spectrum bands and $\phi = 12$ SUGs. The bandwidth capacities of the spectrum bands are set to $V_j = [10 \quad 20 \quad 30]$. In this scenario, an ideal balanced load is reached when each of the SUGs earn a reward of 5 units, corresponding to when the 1^{st} band houses 2 SUGs, the 2^{nd} band 4 SUGs, and the 3^{rd} band 6 SUGs. For this simulation scenario, we evaluate and compare the performances of the cooperative access scheme by considering three degrees of cooperation: 2 (i.e, each SUG cooperates with 2 other SUGs), 4 (i.e, each SUG cooperates with 4 other SUGs), and 6 (i.e, each SUG cooperates with 6 other SUGs).

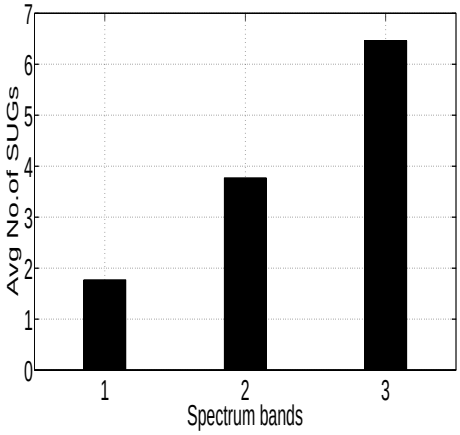
Fig. 6.3 shows the average number of SUGs that end up choosing each of the three spectrum bands for the random scheme, non-cooperative scheme, and



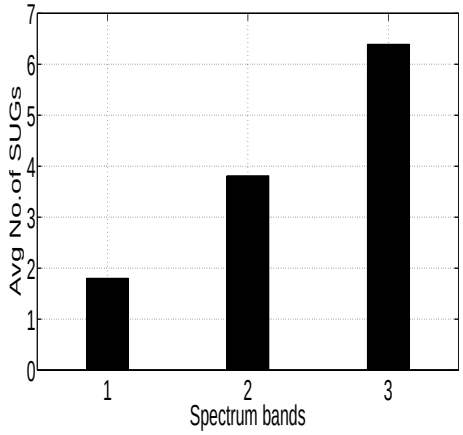
(a) Random



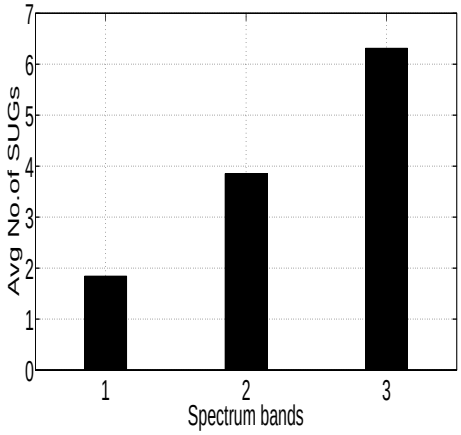
(b) Non-cooperation



(c) Cooperation with 2 SUGs



(d) Cooperation with 4 SUGs



(e) Cooperation with 6 SUGs

FIGURE 6.3: SUG distribution: $m = 3$, $\phi = 12$, $V_j = [10 \;\; 20 \;\; 30]$.

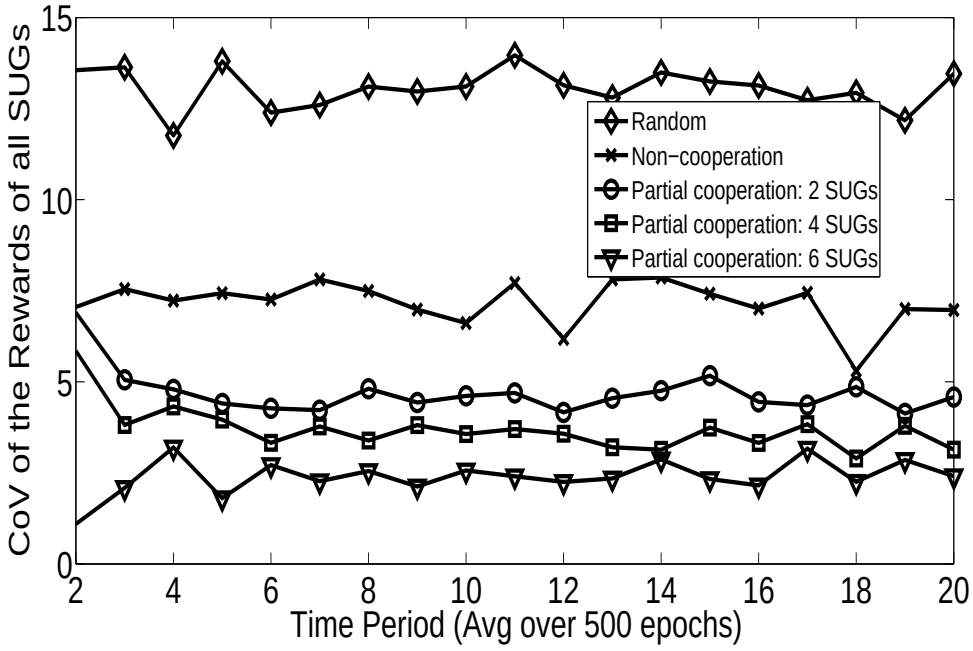


FIGURE 6.4: Coefficient of variation of the rewards of all the SUGs at each time period: $m = 3$, $\phi = 12$, $V_j = [10\ \ 20\ \ 30]$.

cooperative access scheme with 2, 4 and 6 degree of cooperation. Note that as the degree of cooperation increases, the system load becomes more balanced. That is, the cooperative access scheme with degree of cooperation equal to 6 leads to a better balanced system load when compared with the other two degrees of cooperation.

We also study fairness achieved under each of the three cooperation degrees, and plot the CoV of the received rewards of the SUGs in Fig. 6.4. Observe that cooperation with a degree of 6 has the lowest CoV, followed by a degree of 4, and then followed by a degree of 2. This indicates that a higher degree of cooperation leads to a lower CoV, meaning that SUGs receive closer amounts of rewards, thus ensuring fairness among SUGs. Therefore, cooperation improves performances not only in terms of network load balancing, but also in terms of ensuring fairness among

all SUGs. Note that each of the three degrees of cooperation has a lower CoV when compared with the non-cooperative and random access schemes.

It is important to mention again that although higher degree of cooperation results in improved action selection decisions, it also incurs more communication overhead and execution times. Therefore, one must choose the degree of cooperation that balances between good selection decision and minimum overhead so as to lead to an increased overall system performance.

7. CONCLUSION

Technological advances enabled cognitive radios, which have recently been recognized as the key technology for realizing OSA. Cognitive radios are viewed as intelligent systems that can self-learn from their surrounding environments, and auto-adapt their operating parameters in real-time to improve spectrum efficiency.In this thesis, we have developed a reinforcement learning-based framework for OSA and have evaluated and compared the performance of a single-agent RL algorithm with the random scheme and in addition we have formulated and compared the throughput performance of two multi-agent RL algorithms, namely the non-cooperative and cooperative Q-learning scheme along with the random scheme. It is shown from simulations that partial and fully cooperative access schemes perform better than the non-cooperative and the random schemes in terms of achieving higher throughput and a better balanced traffic loads. We also showed that cooperation improves performances not only in terms of network load balancing, but also in terms of ensuring fairness among all users. The proposed learning techniques do not require prior knowledge of the environment's characteristics and dynamics, yet can still achieve high performance by learning from interaction with the environment.

BIBLIOGRAPHY

1. M. Vilimpoc and M. McHenry, "Dupont Circle Spectrum Ulitzation During Peak Hours," in www.newamerica.net/files/archive/Doc_File_183_1.pdf, 2006.

2. M. McHenry, "Reports on Spectrum Occupancy Measurements, Shared Spectrum Company," in www.sharedspectrum.com/?section nsf_summary.

3. M. McHenry and D. McCloskey, "New York City Spectrum Occupancy Measurements," Shared Spectrum Conference, Sept. 2004.

4. "Second report and order and memorandum opinion and order," in http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-08-260A1.pdf.

5. R. J. Lackey and D. W. Upmal, "Speakeasy: the military software radio," IEEE Communications Magazine, May 1995.

6. J. Mitolla III and G. Maguire, "Cognitive radio: making software radios more personal," IEEE Personal Communications, Aug. 1999.

7. S. Haykin, "Cognitive radio: brain-empowered wireless communications," IEEE on Selected Areas in Communications, Feb. 2005.

8. A. Ghasemi and E. S. Sousa, "Interference aggregation in spectrum-sensing cognitive wireless networks," IEEE of Selected Topics in Signal Processing, vol. 2, no. 1, pp. 41 - 56, Feb. 2008.

9. Z. Quan and S. Cui and A. H. Sayed, "Optimal linear cooperation for spectrum sensing in cognitive radio networks," IEEE of Selected Topics in Signal Processing, vol. 2, no. 1, pp. 28 40, Feb. 2008.

10. H. Su and X. Zhang, "Cross-layer based oppotrunitic MAC protocols for QoS provisionings over cognitive radio wireless networks," IEEE on Selected Areas in Communications, vol. 26, no. 1, pp. 118 129, Jan. 2008.

11. C.-T. Chou and S. Shankar and H. Kim and K. G. Shin, "What and how much to gain by spectrum agility?" IEEE on Selected Areas in Communications, vol. 25, no. 3, pp. 576 588 April 2007.

12. S. Srinivasa and S. A. Jafar, "Cognitive radio networks: how much spectrum sharing is optimal?" Proceedings of IEEE GLOBECOM, pp. 3149 3153, Nov. 2007.

13. Z. Ji and K. J. R. Liu, "Belief-assisted pricing for dynamic spectrum allocation in wireless networks with selfish users," Proceedings of IEEE SECON, vol. 1, pp. 119 127, Sept. 2006.

14. Z. Ji and K. J. R. Liu, "Multi-stage pricing game for collusion-resistant dynamic spectrum allocation," IEEE on Selected Areas in Communications, vol. 26, no. 1, pp. 182 191, Jan 2008.

15. S. Delaere and P. Ballon, "Flexible spectrum management and the need for controlling entities for reconfigurable wireless systems," Proceedings of IEEE DySPAN, pp. 347 362, April 2007.

16. Y. T. Hou and Y. Shi and H. D. Sherali, "Spectrum sharing for multi-hop networking with cognitive radios," IEEE on Selected Areas in Communications, vol. 26, no. 1, pp. 146 155, Jan. 2008.

17. A. Ginsberg and J. D. Poston and W. D. Horne, "Toward a cognitive radio architecture: integrating knowledge representation with software defined radio technologies," Proceedings of IEEE MILCOM, pp. 1 7, Oct. 2006.

18. S. Yarkan and H. Arslan, "Exploiting location awareness toward improved wireless system design in cognitive radio," IEEE Communications Magazine, vol. 46, no. 1, pp. 128 136, Jan. 2008.

19. Q. Zhao and S. Geirhofer and L. Tong and B. M. Sadler, "Opportunistic spectrum access via periodic channel sensing," IEEE Trans. Signal Process., vol. 56, no. 2, pp. 785 - 796, Feb. 2008.

20. J. Unnikrishnan and V. V. Veeravalli, "Dynamic spectrum access with learning for cognitive radio," in Proc. Asilomar Conf. Signals Syst. Comput., Oct. 2008, pp. 103 - 107.

21. H. Liu and B. Krishnamachari and Q. Zhao, "Cooperation and learning in multiuser opportunistic spectrum access," Proceedings of IEEE ICC, 2008, pp. 487 - 492.

22. K. Liu and Q. Zhao, "Distributed learning in cognitive radio networks: Multi-armed bandit with distributed multiple players," Proceeding of IEEE Int.Conf. Acoust., Speech, Signal Process., 2010.

23. Z. Han and C. Pandana and K. J. R. Liu, "Distributive opportunistic spectrum access for cognitive radio using correlated equilibrium and no-regret learning," Proceedings of IEEE WCNC, pp. 11 - 15, March 2007.

24. K. E. Nolan and P. Sutton and L. E. Doyle, "An encapsulation for reasoning, learning, knowledge representation, and reconfiguration cognitive radio elements," Proceedings of Intl Conference on Cognitive Radio Oriented Wireless Networks and Communications, June 2006, pp. 1 - 5.

25. H. Kim and K. G. Shin, "Fast discovery of spectrum opportunities in cognitive radio networks," in Proc. IEEE DySPAN, Oct. 2008, pp. 1 - 12.

26. H. Kim and K. G. Shin, "Efficient discovery of spectrum opportunities with MAC-layer sensing in cognitive radio networks," IEEE Trans. Mobile Comput., vol. 7, no. 5, pp. 533 - 545, May 2008.

27. U. Berthold and M. Van Der Schaar and F. K. Jondral, "Detection of Spectral Resources in Cognitive Radios Using Reinforcement Learning," Proceedings of IEEE DySPAN, pp. 1 - 5, Oct. 2008.

28. M. Maskery, V. Krishnamurthy, and Q. Zhao, "Decentralized dynamic spectrum access for cognitive radios: Cooperative design of a noncooperative game," IEEE Trans. Commun., vol. 57, no. 2, pp. 459 - 469, Feb. 2009.

29. J. Unnikrishnan and V. V. Veeravalli, "Cooperative sensing for primary detection in cognitive radio," IEEE J. Sel. Topics Signal Process., vol. 2, no. 1, pp. 18 - 27, Feb. 2008.

30. Y. Chen and Q. Zhao and S. Ananthram, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors, IEEE Trans. Inf. Theory, vol. 54, no. 5, pp. 2053 - 2071, May 2008.

31. R. S. Sutton and A. G. Barto, Reinforcement Learning, The MIT Press, 1998.

32. B. Hamdaoui and P. Venkatraman and M. Guizani, "Opportunistic Exploitation of Bandwidth Resources through Reinforcement Learning," IEEE Global Telecommunications Conference, pp. 1 - 6, Nov. 2009.

33. P. Venkatraman and B. Hamdaoui and M. Guizani, "Opportunistic Bandwidth Sharing Through Reinforcement Learning," IEEE Transactions on Vehicular Technology, vol. 59, no. 6, pp. 3148 - 3153, July 2010.

34. E. R. Gomes and R. Kowalczyk, "Dynamic analysis of multiagent Q-learning with ϵ-greedy exploration," Proceedings of the 26th International Conference on Machine Learning, pp. 369 - 376, 2009.