LEARNING AIDED DYNAMIC SPECTRUM ACCESS IN DECENTRALIZED NETWORKS WITH UNKNOWN NUMBER OF SECONDARY USERS

Student Name: SUMAN PANI

IIIT-D-MTech-ECE
September 18, 2017

Indraprastha Institute of Information Technology
New Delhi

Thesis Advisor

Dr. Sumit J Darak

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Electronics & Communication,
with specialization in Communication & Signal Processing

Keywords: opportunistic spectrum access, online learning algorithms, decentralize network, unsynchronized cognitive radios.

Certificate

This is to certify that the thesis titled **"Learning Aided Opportunistic Spectrum Access in Decentralized Networks with Unknown Number of Secondary Users"** submitted by **Suman Pani** for the partial fulfillment of the requirements for the degree of *Master of Technology* in *Electronics and Communication & Engineering* is a record of the bonafide work carried out by her under my guidance and supervision in Communication and Signal Processing at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Dr. Sumit J Darak
Indraprastha Institute of Information Technology, New Delhi

Abstract

Opportunistic spectrum access (OSA) in decentralized network is a challenging problem since each unlicensed user (i.e. cognitive radio (CR)) needs to characterize frequency bands as per the occupancy statistics of multiple licensed users. In dynamic and heterogeneous networks, all CRs may not be active simultaneously and can leave or join the network at any time. This makes OSA more challenging especially in the decentralized network where active CRs do not have any knowledge of other CRs in the network. In this thesis, a new decision making policy (DMP) using online learning algorithms has been proposed for characterization of frequency bands and orthogonalization of CRs into different but optimal bands. The proposed DMP, when implemented at all CRs in decentralized network, leads to an order-optimal policy. There are two main underlying strategies 1) learning the subband availability, which helps to reduce the long waiting time due to occupancy by primary user(PU), and 2) learning the efficiency of rank, which helps in reducing collision among the secondary users(SUs). For the fixed number of SUs, the loss in throughput decreases by 63% in case of 4 SUs and 46.6% in case of 6 SUs compared to existing state-of-art DMPs, Musical Chair(MC). For varying number of SUs, the loss in throughput decreases by 81.5% in case of 4 SUs and 84.8% in case of 6 SUs compared to MC.

Acknowledgment

Towards the end of my master degree, I would like to pay my hearty gratitude to several individuals who contributed in many ways. First of all, I would like to express my deep gratitude to my thesis supervisor, Dr. Sumit J Darak for his support, guidance and motivation all along the way. I feel extremely fortunate to work with him. He has been a great source of inspiration to me and I thank him from the bottom of my heart. I would like to thank Dr. Pankaj Jalote and all my professors at IIITD for providing such a wonderful environment to work. I would like to express my gratitude towards my family for their kind co-operation and encouragement. Last but not the least, I want to thank my friends for all thoughtful discussions we had, which prompted me to think beyond the obvious.

Contents

1 Introduction 1

1.1 Motivation and objective . . . 1

1.2 Novelty . . . 2

1.2.1 Fixed number of SUs . . . 3

1.2.2 Varying number of SUs . . . 3

1.3 Terminology . . . 3

1.3.1 Multi-armed Bandit(MAB) . . . 3

1.3.2 Cognitive Radio Network(CRN) . . . 4

1.4 Outline . . . 6

2 Previous Work 7

2.1 Subband Statistics Estimation . . . 7

2.2 Subband Access Algorithm . . . 7

3 System Model 10

3.1 Regret Model . . . 11

3.2 Subband Statistic Estimation . . . 11

3.2.1 Upper Confidence Bound (UCB) . . . 11

3.2.2 Bayes-UCB . . . 12

3.3 Problem Formulation . . . 13

4 Proposed DMP for Fixed Number of SUs 15

4.1 Algorithm : Rank Learning with Unequal Rank-Reward (RLUR) . . . 15

4.1.1 Performance Comparision of UCB and BUCB . . . 17

4.2 Algorithm : SU Estimation with Rank learning (SERL) . . . 20

4.3 Simulation Result . . . 22

4.4 Inference . . . 26

5 Proposed DMP for variable number of SUs 27

5.1 Algorithm : Rank Learning with Unequal Rank-Reward (RLUR) . . . 27

5.2 Algorithm : SU Estimation with Rank learning (SERL) 28
5.3 Simulation Result 30
5.4 Inference 34

6 Mathematical Proof **35**
6.1 ρ_{rand} 35
6.2 Random Rank Selection Without the Knowledge of the no. of SU 35
6.3 Rank Selection Based on Rank Learning 36
6.3.1 Occurrence of Collision 36
6.3.2 Non-occurrence of Collision 38
6.3.3 Calculation of Probability 39
6.4 Comparision of Probabilities 40
6.4.1 Comparing with Random Rank Selection 40
6.4.2 Comparing with ρ_{rand} 41

7 Conclusion and Future Work **42**

List of Figures

1.1 Markovian model for subband status 5

3.1 Utilization of one time slot for PU 10
3.2 Utilization of one time slot for SU 10
3.3 output of *BUCB(x,z,w)* function for various values of parameter 12

4.1 Normalised regret $\frac{R(n)}{\log n}$ vs time slot for no. of SUs is 4 18
4.2 Normalised regret $\frac{R(n)}{\log n}$ vs time slot for no. of SUs is 6 19
4.3 Total no. of collision for synchronized SU 19
4.4 Total switching cost for synchronized SU 20
4.5 Normalised regret $\frac{R(n)}{\log n}$ vs time slot for synchronized SU(no. of SUs is 4) 23
4.6 Total no. of collision for synchronized SU(no. of SUs is 4) 23
4.7 Total switching cost incurred for synchronized SU(no. of SUs is 4) 24
4.8 Normalised regret $\frac{R(n)}{\log n}$ vs time slot for synchronized SU(no. of SUs is 6) 24
4.9 Total no. of collision for synchronized SU(no. of SUs is 6) 25
4.10 Total switching cost incurred for synchronized SU(no. of SUs is 6) 25

5.1 Time slot vs the subband that was selected in that time slot, for number of SU is 3 and number of subband is 8 28
5.2 Normalised regret $\frac{R(n)}{\log n}$ vs time slot for varying SU(maximum no. of SUs is 4) . . 31
5.3 Total no. of collision for varying SU(maximum no. of SUs is 4) 31
5.4 Total switching cost incurred for varying SU(maximum no. of SUs is 4) 32
5.5 Normalised regret $\frac{R(n)}{\log n}$ vs time slot for varying SU(maximum no. of SUs is 6) . . 32
5.6 Total no. of collision for varying SU(maximum no. of SUs is 6) 33
5.7 Total switching cost incurred for varying SU(maximum no. of SUs is 6) 33

List of Abbrevations

BUCB	Bayesian Upper Confidence Bound
D2D	Device-to-device
DMP	Decision Making Policy
FBS	Frequency Band Switching
IID	Independent and Identical Distribution
MAB	Multi Armed Bandit
MDP	Markovian Decision Process
MC	Musical Chair
OSA	Opportunistic Spectrum Access
PU	Primary User
RLUR	Rank Learning with Unequal Rank-Reward
SU	Secondary User
TS	Thompson Sampling

Chapter 1

Introduction

1.1 Motivation and objective

In the next-generation wireless communication, increasing the efficiency of spectrum utilization is going to be one of the major problems. Spectrum is a limited resource and with the increase in the wireless communication and the advancement of technology, it becomes progressively necessary to utilize the spectrum to its maximum capacity. Moreover, energy efficiency of such network should be as high as possible due to obvious factors such as cost, environmental concerns etc. The need to increase the utilization of the electromagnetic spectrum for economical reasons as well as to support wide range of data intensive services has generated significant interest from academia as well as industries [1]. Among the various OSA based envisioned paradigms such as device-to-device communications [2] and LTE-unlicensed network [3], dynamic spectrum access (DSA) based cognitive radio network seems to be promising solution to deal with spectrum scarcity problem mainly caused by inefficient static spectrum allocation [1, 4–6]. For instance, DARPA's spectrum collaboration challenge 2016 [7] is a significant step to bring DSA to life.

Cognitive radio network consists of licensed or primary users (PU) and unlicensed or secondary users (SU). The PU, who buys the subband does not transmit in the subband all the time. Hence, immense amount of the spectrum allocated to PU is not efficiently utilized. The SUs need to identify and transmit in the vacant frequency bands without causing any interference to active PUs. Here, SUs opportunistically use the subbands when they are unoccupied by the PUs. Hence in order to get the maximum throughput, they must choose the subband for transmission such that the chosen subband is more likely to be unoccupied by the PU. Moreover, if multiple SUs opt for the same subband at the same time, they face a collision and lose the transmitted data. Hence, it must be avoided. In the last decade, significant research efforts have been observed in the spectrum sensing to detect the status (vacant or occupied) of the frequency band [1, 4–6], wideband spectrum sensing [8] and co-operation among SUs [4–6]. However, to make DSA feasible in practical wireless communication environment, SUs need decision making policy (DMP) to characterize frequency bands based on their vacancy statistics and quality, choose better bands as well as avoid collisions among SUs to improve throughput. In addition, DMP should minimize the number of frequency band switching (FBS) as well as collisions for longer battery life of SU terminal.

DMPs for centralized approach have been designed and studied extensively in the literature. However, centralized approach has many drawbacks such as higher susceptibility to failure, reduced diversity, security issues etc. These problems are greatly reduced in case of decentralized networks due to the absence of a single central controlling unit. Hence, any dysfunction is generally confined to a section of the network instead of affecting the whole network. Moreover, it has advantages such as no communication overhead, lower complexity and ease of implementation. So, a need for designing a suitable DMP for a decentralized network arose. However, doing so is a challenging problem as SUs do not share information in case of decentralized networks leading to inaccuracy in estimation of subband statistics and frequent collisions among the SUs.

Various DMPs [9–24] have been proposed in the literature for DSA in the decentralized network. Please refer to Section 2 for more details. Existing DMPs [9, 11–16] assume complete knowledge of number of active SUs and instantaneous transmission feedback. Both these assumptions are not feasible in practical decentralized network. Furthermore, their goal is to improve the throughtput of the network without paying an attention to the number of FBS and collisions, which are critical parameters for resource constrained battery operated SU terminals. Recent state-of-the-art DMP [24] does extremely well to estimate the number of SUs. However, it does not take into account delayed feedback and follows random frequency band selection approach which has been shown to incur significant penalty in throughput, number of FBS as well as collision. The design of DMP which offers comprehensive solution to these challenges is the focus of the work presented in this thesis.

In this thesis, we alleviate the drawbacks of existing DMPs by proposing a new DMP to access the subbands. Proposed DMP takes into consideration a more practical scenario where the number of SU is variable and unsynchronized. The above algorithm is validated by comparing it with existing algorithms for different number of SUs and different vacancy probability of the subbands based on regret, switching cost and number of collisions.

1.2 Novelty

In this thesis, a rank based algorithm has been proposed where each SU chooses a rank, r and selects the r^{th} best subband according to its gathered information on subband statistic for sensing. This concept, though introduced in ρ_{rand} [9], had the drawback that the SUs needed the information about the number of SUs in the system and then chose the rank randomly. In the following section, it is highlighted how the proposed DMPs overcomes the mentioned problem both in case of 1) Fixed number of SUs, and 2) Variable number of SUs.

1.2.1 Fixed number of SUs

The novelty of the proposed algorithm stems from taking into consideration a more practical scenario where the number of SUs, U is not known to the SUs. Without that information, the SUs do not have a predefined range for rank selection and hence initialize the range of rank as $1, 2, ..., C$. The proposed DMP is designed using Bayesian Upper Confidence Bound (BUCB) based online learning algorithm which is superior to other online learning algorithms as shown in [13]. The BUCB is used to develop strategies to: 1) Characterize the frequency band availability statistics which in turn helps to reduce the long waiting time due to occupancy by PUs, 2) Orthogonalize SUs to minimize the number of SU collisions. It is shown that the proposed DMP, when implemented at all active SUs in the decentralized network, leads to an order-optimal DMP.

1.2.2 Varying number of SUs

In this thesis, a rank based DMP has been proposed, which is inspired by the DMP discussed in [9]. The novelty of the proposed algorithm stems from taking into consideration a more practical scenario where the number of SUs are variable. The time of entry and exit of the SUs in the network is not synchronized. Moreover, the proposed algorithm approaches learning of both rank and subband performances, which ensures that the SU instead of operating on a randomly chosen rank, chooses the one which leads to least number of collision. Furthermore, it uses Bayesian approach as discussed in [25] named as BUCB, unlike Upper Confidence Bound (UCB) algorithm used in [9].

1.3 Terminology

In this section, details of two key components, 1) Multi-armed Bandit, and 2)Cognitive Radio Network, used in modeling the problem are discussed. Furthermore, various parameters used to model the problem are also explained, the significance of which will be reveled in later chapters.

1.3.1 Multi-armed Bandit(MAB)

In a multi-armed bandit problem, as explained in [26]- [27], there is a slot machine with two or more arms, each of which when played by an agent j yields some reward depending on an probability p_j, unknown to the agent. The problem is to design a policy that, at each stage, a SU chooses to play one of the arms with a view to maximize the expected sum of rewards. Equivalently, an attempt can be made to minimize the regret defined as the difference between the maximum expected throughput when the probability p_j of each arm and the selection criteria is known, and the expected sum of throughput actually obtained by a particular DMP in the absence of that knowledge. The DMP necessarily needs to have a balance between exploiting the best arm as decided by learning based of previous result and exploring the quality of other arms so that the learned result is trustworthy and not biased.

1.3.2 Cognitive Radio Network(CRN)

Cognitive radio network, as explained in [28], is a heterogeneous decentralized network with licensed and unlicensed users. The licensed users/PUs are given priority over the unlicensed users/SUs. Hence, the SUs can access the spectrum only when the PUs are not transmitting over them. The U SUs opportunistically try to access and transmit data over one of the C subbands such that they do not hamper the transaction of PUs and also avoid colliding with other SUs.

This problem can be easily modeled as a MAB, details of which are explained in [29]. The C subbands are the arms played by the U SUs which can be equivalent to the agents as explained in the MAB. The SU access the subbands according to a particular DMP so that it both exploits the best subband according to its learning and explores for better subbands.

Some parameters used to model the CRN are explained below.

- Arm evolution model
- Reward model

Arm Evolution Model

When a SU chooses a subband for sensing at any instance, the status of the subband can be one of the following:

- Vacant : The PU is transmitting over the subband.
- Occupied : The PU is not transmitting over the subband.

The stochastic process governing the status of the subband can be modeled as independent and identical distribution (IID) or Markovian model.

In *IID* model, the status of the subband at any time slot is independent of the status of other subbands as well as its status in previous time slots. The subband c is vacant with some probability p_c and occupied otherwise.

In *Markovian* model, the status of the subband is independent of the status of the other subbands. However, as shown in Fig. 1.1, it does depend on the status of the given subband in previous time slot as the status of the given subband is modeled as Markov decision process. The probability $\mu := [\mu_{o-v}; \mu_{v-o}]$ signifies the subband statistics. $\mu_{o-v} := [\mu_{o-v,1}, \mu_{o-v,2}, ..., \mu_{o-v,C}]$ is the probability that the status of subband has changed from *occupied* to *vacant* and $\mu_{v-o} := [\mu_{v-o,1}, \mu_{v-o,2}, ..., \mu_{v-o,C}]$ is the probability that the status of subband has changed from *vacant* to *occupied.*

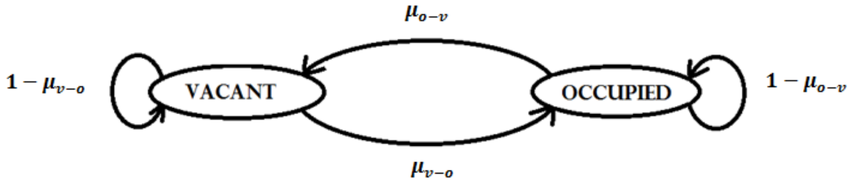

Figure 1.1: Markovian model for subband status

Reward Model

In the decentralized multi-agent network interaction between the SUs plays an important role in the performance. When two or more SUs select the same subband at the same time slot for sensing and transmission, it leads to an interaction between them. The interaction between SUs affect their ability to collect rewards. There are many models which govern the allocation of reward among the interacting SUs. Some of them are, *general symmetric interaction model, agent-specific interaction model, random sharing model* etc, as explained in [30]. The one we have used is *collision model.* In collision model, the SUs receive reward when they transmit successfully in a time slot. Collision results in a complete loss of communication for that time slot for the SUs experiencing a collision. The reward received by the SU can be awarded in two ways.

An easier approach to model the reward is to have a binary model where the reward is

- 1 if the SU successfully transmits the data
- 0 if the SU is unsuccessful in transmitting the data

A more evolved and complex model of giving reward is to take into consideration the quality of the subband and generate a number $\in (0,1)$, such that a higher rewards denotes a better quality subband. The reward is then modeled as:

- $\in (0,1)$ if the SU successfully transmits the data
- 0 if the SU is unsuccessful in transmitting the data

It is understood that the probability of vacancy as mentioned in the previous section and the reward of a subband need not be dependent on each other. More details regarding the various parameters of a CRN can be studied from [30]. The issues in implementing spectrum sensing using CRN (such as subband uncertainty, noise uncertainty, aggregate interference uncertainty and sensing interference limit) is explained in detail in [31]- [32]. A detail of how wide spectrum sensing be implemented usinf CRN is shown in [33]

1.4 Outline

The remaining thesis is organized as below:

Chapter 3 describes in detail the system model in which we are trying to solve the above discussed problems

In **Chapter 4** goes into details of the proposed DMP for efficient selection of a subband by the SUs for fixed number of SUs. The results are presented for different parameters such as the number of SUs in the system and the vacancy probability.

In **Chapter 5** the working of the proposed DMP is explained for efficiently selecting a subband by the SUs for variable number of SUs. The results are presented for different set of parameters such as the maximum number of SUs in the system and the vacancy probability.

Chapter 6 deals with the mathematical formulation of *regret* and *collision* for the proposed DMP.

Chapter 2

Previous Work

In multi-user decentralized networks, DMP must consist of two sub-policies: 1) subband statistics estimation, and 2) subband selection algorithm.

2.1 Subband Statistics Estimation

The subband statistics estimation problem is approached as a MAB algorithm which can be understood from [27]. An in-depth study of CRN and the various ways it can be modeled is done in [28]- [30]. The estimation of the statistic is based on UCB as mentioned in [9]. In recent years, Bayes algorithm (BUCB) [25] and Thompson Sampling (TS) [34] approaches had been proved to be more efficient for wireless scenario.

2.2 Subband Access Algorithm

The time division fair scheme (TDFS)-DMP [10] and ρ_{rand}-DMP [9], are the well known DMPs for the decentralized network with known number of SUs, say U. Both DMPs employ upper confidence bound (UCB) algorithm for characterization of frequency bands and selection of the optimal subset of U number of frequency bands. In both DMPs, each SU randomly and independently choses the rank between 1 to U which in turn governs the frequency band selection among the optimal U frequency bands. Such randomization approach ensures orthogonalization of SUs since the rank is re-assigned only for colliding SUs. In TDFS-DMP, SUs switch the frequency band in round robin fashion to allow fair access to optimum bands among all SUs. However, this incurs significant penalty in terms of FBS and hence, ρ_{rand}-DMP [9] is a preferred choice where SUs randomly changes their rank and hence, frequency band only when collision occurs. [9, 10] shows that the online learning algorithm offers superior performance over random frequency band selection approach. Still, they used randomization approach for rank selection. In [11, 12], ρ_{rand}-DMP is improved further by replacing randomization approach for orthogonalization with learning based approach. Furthermore, [13] shows that Bayesian online learning algorithm based DMP offers significant improvement in the number of FBS as well as collisions and hence, throughput, compared to UCB like algorithms. Also, computation complexity of former is much lower than latter.

The DMP in [14] is based on assigning unique hopping sequence to each SU thereby minimizing the collision among them. However, it is assumed that SUs have complete knowledge of frequency bands. The DMPs in [15, 16] employ auction based approach for collision free transmission. However, such DMP needs central controller and communication among SUs which may not be feasible in the practical networks. In addition, the main drawback of DMPs in [9–16] is the requirement of complete knowledge of the number of SUs which is not feasible in most of realistic scenarios.

The work in [17] uses decoupling of exploration and exploitation approach in a multi-armed bandit environment. The subband which the SU decides to transmit over for a time slot may or may not be the one whose sensing information it chooses to collect in the time slot. Hence, the SU exploits the best option of subband available to it without compromising the exploitation when it comes to gathering sensing information. However, this approach has a high complexity as the SU has to keep track of both the exploited subband and explored subband. [18] proposes an algorithm which selects the rank in two phases. It splits the rank into two parts. Each SU then selects a rank from from the first part and if it faces a collision, it again selects a rank from the second part to transmit. In [19], the authors have come up with a technique based on hopping where all the SUs select the subband according to a predefined sequence of subbands(known as the hopping sequence) as defined by [20], which eliminates the problem of collision. However, the hopping sequence consists of all the subbands irrespective of the number of SUs in the system hence, the SUs do not restrict their subband selection to the optimal subbands. So, there is no concept of exploitation of the optimal subband. In [21], the authors have come up with an algorithm where the SUs exchange the subband they are settled into only if none of them faces a loss in doing so, hence increasing the over all performance. However, the problem in this algorithm is that there needs to be some communication between the SUs in order for the exchange to occur.

The MEGA-DMP in [23] and MC-DMP in [24] are the only DMPs for the decentralized network with unknown number of SUs. It has been shown in [24] that MC-DMP outperforms DMP in [23] and it is easy to implement. Hence, the discussion hereafter is limited to [24] which is the current state-of-the-art algorithm. MC-DMP divides the time horizon in two stages: 1) Learning stage, 2) MC algorithm. In the learning stage, each SU randomly choses the frequency band and observes the total throughput as well as number of collisions on each frequency band. This information is then exploited to estimate the number of active SUs in the network. In MC algorithm stage, information gained in the learning stage is used for frequency band selection. The SU lock itself to the frequency band for which there is no collision and remains there for the rest of the time. However, it does not take into account delayed feedback and follows random frequency band selection approach which has been shown to incur significant penalty in throughput, number of FBS as well as collision. The goal of the proposed DMP presented in the next section is to offer comprehensive solution to above mentioned challenges.

Chapter 3

System Model

In this chapter, details of the system model assumed in the thesis has been discussed. Also, various parameters mentioned in Subsection 1.3.2 are defined.

Let us consider, the decentralized network is modeled as C subbands, servicing U independent SUs, where the value of U is less than C. A slotted time based approach is considered. Both the cases of SUs' clocks being synchronized and not synchronized are considered. Let H be the horizon and t the number of time slots any SU u has spent in the system, where $u \in 1, 2, ..., U$.

Each SU selects one of the C subbands and senses for the presence of PU. If the subband is unoccupied by the PU, it transmits in that time slot. Else it waits for the next time slot.

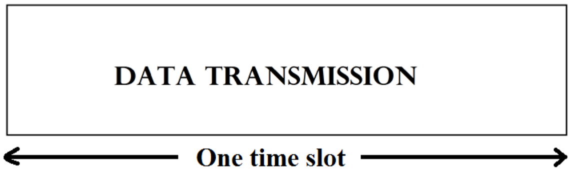

Figure 3.1: Utilization of one time slot for PU

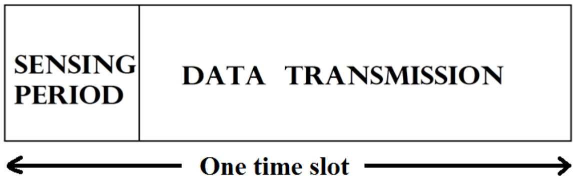

Figure 3.2: Utilization of one time slot for SU

From Fig. 3.1 and 3.2, it can be seen that during the sensing time period, a SU can detect the presence of a PU as the PU starts transmitting at the onset of the time slot. However, it can not detect the presence of other SUs as all the SUs are sensing in that period and are not transmitting any data. It is assumed that the SUs are in close proximity to each other causing an interference when two or more SUs try to transmit in the same subband at any particular time slot leading to collision. This assumption leads to another insight that the subband quality, in other words *reward* for any subband can be considered same for all the SUs.

3.1 Regret Model

$S(t)$ is the total number of successful transmission for all the U SUs in the network at t^{th} time slot for the proposed DMP.

$$S(t) = \sum_{c=1}^{C} \sum_{u=1}^{U} \mu(c)\mathbb{E}[V_{c,u}(t)], \tag{3.1}$$

where $\mathbb{E}$ is the expectation operator.
$V_{c,u}(t)$ is 1 when u is the only SU sensing subband c in t^{th} time slot.
μ_c is the probability of vacancy of the subband c.

$S^*(t)$ is the total number of successful transmission for centralized DMP where the subband statistic is known and there is no collision among the SU.

$$S^*(t) = t\sum_{u=1}^{U} \mu(u^*), \tag{3.2}$$

where $\mu(u^*)$ is the u^{th} highest value of μ.

So, the regret of the policy can be calculated as loss of reward due to the lack of knowledge.

$$R(t) = S^*(t) - S(t) \tag{3.3}$$

As it can be seen centralized policy is the ideal case scenario. So, $S^*(t) > S(t)$, implying $R(t)$ is always positive.

3.2 Subband Statistic Estimation

In subband statistic estimation, we try to estimate the probability of vacancy of the subband. However, subband statistic that is calculated takes into consideration both exploitation and exploration. Hence, its value can be high even for a subband whose vacancy of probability is not high if that subband has been selected for less number of time slots. We have used 2 different policies to calculate the subband availability statistic and the rank performance:

3.2.1 Upper Confidence Bound (UCB)

The statistic value for subband and rank performance statistic as calculated by the SU u at time slot t is:

$$\begin{aligned} g_u(m,t) &= UCB(A_{m,u}(t), B_{m,u}(t), t) \\ &= \frac{\sum_{i=1}^{t} A_{m,u}(t)}{B_{m,u}(t)} + \sqrt{\frac{2\log t}{B_{m,u}(t)}} \end{aligned} \tag{3.4}$$

In Eq. 3.4, the first term is the exploitation term which increases as the $A_{m,u}(t)$ increases. the second term is exploration term which increases as $B_{m,u}(t)$ decreases.

3.2.2 Bayes-UCB

The statistic value for subband or rank m as calculated by the SU u at time slot t, as explained in [35] , the inverse of incomplete beta function of elements $1-\frac{1}{t}, A_{m,u}(t)+1$ and $B_{m,u}(t)-A_{m,u}(t)+1$. Where, incomplete beta function is defined as $(1/BETA(z,w))\int_0^x t^{z-1}(1-t)^{w-1}dt$. So, the statistic value for subband or rank is calculated as,

$$\begin{gathered} g_u(m,t) = BUCB(x,w,z) \\ s.t\ \ x = 1/BETA(z,w)\int_0^{g_u(m,t)} t^{z-1}(1-t)^{w-1}dt, \end{gathered} \tag{3.5}$$

where, $x = 1-\frac{1}{t}, z = A_{m,u}(t)+1, w = B_{m,u}(t)-A_{m,u}(t)+1$ and $BETA(z,w) = \int_0^1 t^{z-1}(1-t)^{w-1}dt$. Here, t denotes the number of time slot the experiment has been conducted, $B_{m,u}(t)$ denotes the number of times a particular subband/rank is selected and $A_{m,u}(t)$ denotes the number of time successful transmission was achieved when a particular subband/rank is selected.

To examine how this works, a brief study of *g=BUCB(x,z,w)* is required. This is done by studying the plot of the function for various parametric values.

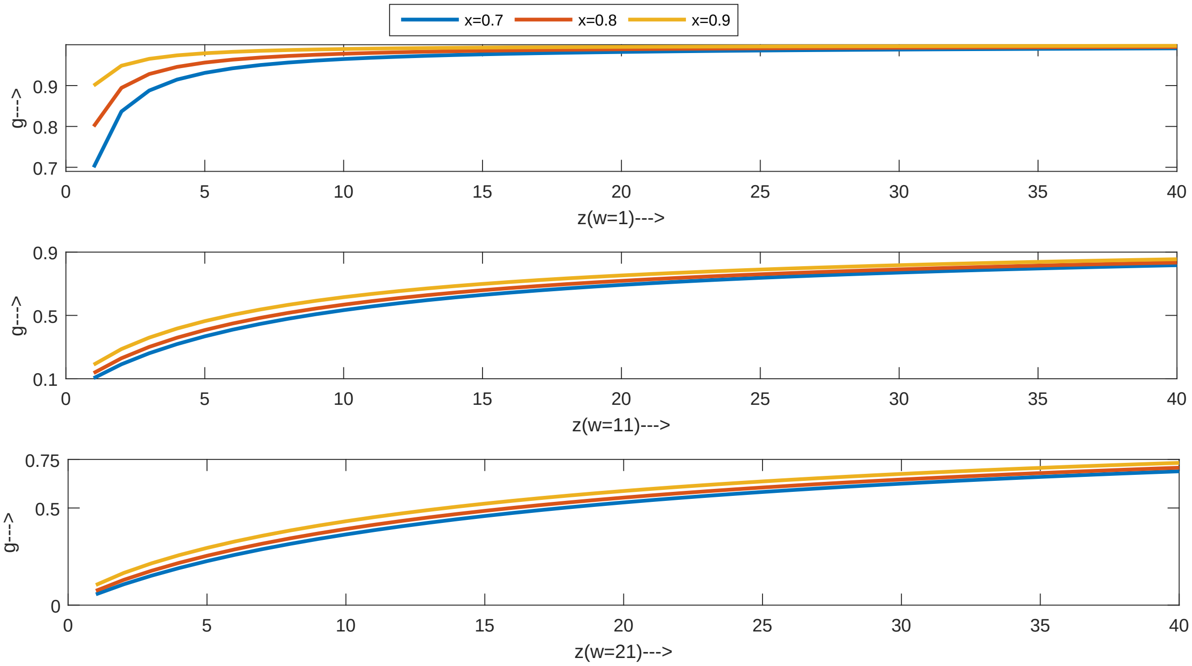

Figure 3.3: output of *BUCB(x,z,w)* function for various values of parameter

The plot in Fig. 3.3 is generated as $z\ vs\ g$ when z varies from 1 *to* 40. Each subplot is generated for a specific value of w as mentioned in the *x-label* and each subplot consists of the plot for 3 different values of x as denoted by the legend.

As it can be seen from Fig. 3.3, for specific values of x and w, g increases as z increases. Also, for fixed value of z and w, g increases as x increases. However, for a given pair of values of x and z, g decreases as w increases.

In case of our problem, there can be three scenarios which will examined individually. The first case is where a subband/rank is selected and yields a successful transmission. In that case, both $A_{m,u}(t)$ and $B_{m,u}(t)$ will increase. Which results in an increase in z. However, w remains same as in previous time slot. In this case, g increases. The second case is where a subband/rank is selected and yields an unsuccessful transmission. In that case, $B_{m,u}(t)$ will increase and $A_{m,u}(t)$ remains same. Which results in an increase in w and w remains same as in previous time slot. In this case, g decreases. The above two scenarios explains how exploitation works in BUCB. The third case is when a subband/rank is not selected. In that case, both $A_{m,u}(t)$ and $B_{m,u}(t)$ will remain same as in previous time slot. Which, in turn, implies that there will be no change in the value of z and w. However, x increase as t increases. And hence, g increases. This takes care of the exploitation of the subband/rank in BUCB.

3.3 Problem Formulation

The problem that the proposed DMP is trying to solve is to decide which subband should each SU choose in each time slot so as to maximize the spectrum utilization. This can be done by increasing the throughput by avoiding long waiting time and colliding with other SUs. This depends on an accurately estimation which among the C subbands have high vacancy probability and orthogonally settling into the U-best subbands .

Problem 1:Let there be H time slots and C subbands being access by U SUs. The regret is given by Eq. 3.3 which needs to be reduced. Hence, the problem statement is given as:

$$\begin{array}{l} \min\limits_{u} \; R(H) \\ \sum\limits_{U=1}^{C} \left(H \sum\limits_{u=1}^{U} \mu(u^*) - \sum\limits_{c=1}^{C} \sum\limits_{u=1}^{U} \mu(c)\, \mathbb{E}[V_{c,u}\,(H)] \right) \\ \mathit{Subject\ to}\ U \leq C \qquad\qquad\qquad\qquad , \end{array} \tag{3.6}$$

where, $\mathbb{E}[V_{c,u}\,(H)]$ estimated number of time slot in which SU u did not face a collision on subband c over the entire horizon, and $\mu(c)$ is the probability of the subband c not being occupied by PU.

Table 3.1: List of Notation

Symbol	Notation
C	number of channels
$c_u(t)$	The subband selected by SU u in time slot t
C_c_u	Number of successive collision SU u have faced in current rank.
$D_{r,u}$	Number of times SU u chose rank r
g_u	Vector of statistic value representing subband vacancy
H	Horizon for which experiment is conducted
H_u	Total time slot for which the user stays in the network
h_u	Vector of statistic value representing rank performance
N_u	The number of SU in the system as estimated by SU u
R	Regret of the network
$R_u(t)$	The rank SU u selected at time slot t
R_c_u	The number of times SU u had consecutively chosen the current rank
S	Number of successful transmission in proposed DMP
$S*$	Number of successful transmission in Centralized DMP
$S_u(t)$	Sensing information of SU u in time slot t
$T_{c,u}$	Number of times SU u choose subband c for sensing
t	time slot for which experiment is conducted
t_u	time slot for which a user has stayed in the network
U	number of SUs
$V_{c,u}$	Indicator of collision
$X_{c,u}$	Number of times SU u found subband c free
$Y_{r,u}$	Number of times SU u transmitted successfully on rank r
Φ	Threshold for resetting the data
μ	Probability of vacancy of the Subbands

Chapter 4

Proposed DMP for Fixed Number of SUs

In this section, we have proposed 2 different DMPs to approach the problems discussed in Chapter 1. In the proposed DMPs, U SUs opportunistically trying to access an unoccupied subband in C subbands. In the proposed DMP, H is the size of the horizon, $c_u(t)$ is the subband selected by SU u at time slot t. As we have already established that the number of SUs is not known to the SUs, the SUs have no idea of the range of the rank. Though it can be easily assured that $R_u \in 1, 2, ..., C$, as a rank $R_u > C$ does not denote to any subband. $T_{c,u}(t)$ gives the number of times subband c has been selected by SU u upto time slot t, $X_{c,u}(t)$ gives the number of times SU u found subband c free. $D_{r,u}(t)$ gives the number of times rank r has been selected by SU u upto time slot t, $Y_{r,u}(t)$ gives the number of times SU u made a collision free transmission while operating on rank r.

Consider a case where, $u \leq C$, when a subband chooses a rank $r > u$ it may reduce the collision it faces with the other SUs but it chooses a subband with lower availability, decreasing the throughput. Hence, a varying rank rewards policy is opted. As selection of a better rank gives a higher rank reward, the users are motivated to choose a better rank.

4.1 Algorithm : Rank Learning with Unequal Rank-Reward (RLUR)

For the first C time slots, each subband is selected once and when the selected subband is unoccupied by the PU, SU transmits on that subband as shown in steps 1 to 9. For the rest of the time slot *C+1* to *N*, the quality index of the rank and the subbands are computed as shown in steps 11 to 17. In step 18,SU u selects subband with the R_u^{th} highest quality index as as

$$\begin{aligned} &c_u(t) = c \\ s.t \;\; &g(c,t) \text{ is } R_u^{th} \text{ highest } \;\text{ value } \;\text{ of } g(:,t) \end{aligned} \tag{4.1}$$

Algorithm 1 Rank Learning with Unequal Rank-Reward (RLUR)

```
Input: $X_{c,u}(t-1), T_{c,u}(t-1), Y_{r,u}(t-1), D_{r,u}(t-1)$ , $R_u, S_u, Rank\_reward, N$
Output: $curr\_sel_u(t)$
for $t \leq C$ do
    if (any$T_{:,u}$ == 0 ) then
        $c_u(t) = c$ s.t $T_{i,j}(t-1) = 0$
        if $N_u(t)$ is vacant then
            $X_{c_u(t),u}(t) = 1$
        end if
        $T_{c_u(t),u}(t) = 1$
    end if
end for
for $t = C + 1$  to $H$  do
    $X_{c_u(t),u}(t) = X_{c_u(t),u}(t-1)$
    $T_{c_u(t),u}(t) = T_{c_u(t),u}(t-1)$
    $Y_{R_u(t),u}(t) = Y_{R_u(t),u}(t-1)$
    $D_{R_u(t),u}(t) = D_{R_u(t),u}(t-1)$
    for $i = 1\, to\, C$ do
        Compute
        $g(i,t) = BUCB(X_{i,u}(t), T_{i,u}(t), t)\ \ Eq.3.5\ or\ UCB(X_{i,u}(t), T_{i,u}(t), t)\ Eq.3.4$
        $h(i,t) = BUCB(Y_{i,u}(t), D_{i,u}(t), t)\ \ Eq.3.5\ or\ UCB(Y_{i,u}(t), D_{i,u}(t), t)\ Eq.3.4$
    end for
    Select band $c_u$ according to Eq 4.1
    if $c_u(t) \neq c_u(t-1)$, sense for the whole time slot and set the value of $S_u$ as in Eq 4.2
    if $c_u(t)$ is vacant, $X_{c_u(t),u}(t) = X_{c_u(t),u}(t) + 1$
    if $c_u(t) == c_u(t-1)$, $S_u == 0$ and no PU was sensed, transmit data in $c_u(t)$
    $D_{R_u(t),u}(t) = D_{R_u(t),u}(t) + 1$
    $T_{c_u(t),u}(t) = T_{c_u(t),u}(t) + 1$
    if Collision occurs in $c_u(t)$ or $S_u == 1$ then
        Select new rank $R_u$ according to Eq 4.3
        if $R_u(t) \neq R_u(t-1)$ then
            Reset $R\_c_u\, and\, C\_c_u$ to 0
        end if
        $C\_c_u = C\_c_u + 1$
    else
        $R\_c_u = R\_c_u + 1$
        $Y_{R_u(t),u}(t) = Y_{R_u(t),u}(t) + Rank\_reward(R_u(t))$
    end if
    Calculate $\Phi$ proportional to $R\_c_u$ as per 4.4
    if $C\_c_u > threshold$ then
        Reset values of $Y_{r,u}(t-1)$, $D_{r,u}(t-1)$, $R\_c_u$, $C\_c_u$ to 0
    end if
end for
```

In step 19, if the subband $c_u(t)$ is the same as that of $c_u(t-1)$, sense the subband for the whole time slot and assign value to S_u as:

$$S_u = \begin{cases} 0, & \text{if no SU is found} \\ 1, & \text{otherwise} \end{cases} \tag{4.2}$$

If the subband $c_u(t)$ is unoccupied by PU, no SU was found in the subband in the previous time slot and t is not a only sensing time slot, the SU transmits else it waits for the next time slot as shown in step 20 to 23. The reason this is done is not to disturb those SUs who have already settled into a particular subband. If SU faces a collision on the subband $c_u(t)$ or it sensed the presence of another SU, it chooses a new rank as mentioned in step 24 and 25 as,

$$\begin{aligned} &R_u(t) = r \\ &s.t. \;\; h(r,t) \geq h(i,t), \;\; \text{where}\, i \in [1, 2, ..., C] \end{aligned} \tag{4.3}$$

The number of collisions faced by the SU while operating in rank $R_u(t)$ and the number of time, the current rank R_u has been consecutively selected is updated as shown in 26 to 34. The $Rank\_reward$ mentioned in step 33 is defined as, $Rank\_reward(i) > Rank\_reward(j)\, \forall i < j$. This leads to a faster increase in $Y_{c,u}(t)$ when a better rank is selected. This ensures that the SUs are more inclined to choose a better rank. The threshold, Φ is calculated as

$$\Phi = \frac{R\_c_u}{\alpha} \tag{4.4}$$

where,R_c_u is the number of times SU, u has chosen currently used subband consecutively.
α is any positive integer. The statistics related to rank performance are reset if the number of collision exceeds the threshold. as shown in step 35 to 38. The idea behind it is that the data collected by the SU is not efficient if it leads to multiple successive collisions. However, the threshold is considered to be adaptive and increases as the amount of time a SU spends in any rank increases. The reason why it taken into consideration is that, if an SU have spent less time in a particular rank, it has not gathered reliable amount of data about it and hence, discarding all the acquired data for a few collisions is not a big lose. However, if an SU have spent long period of time in any rank, it should not be made to loose all the data for a few collisions.

4.1.1 Performance Comparision of UCB and BUCB

As mentioned in the algorithm **RLUR**, the estimation of the probability of vacancy of the subband and rank performance statistic can be done using both UCB and BUCB. Hence, in Figure. 4.1 to 4.4, a comparison of accumulated regret is done for RLUR executed with different combinations of methods for calculation of statistic determining the vacancy probability of subbands and rank performance.

The probability for the MDP, p_{mdp}, considered for this experiment is.

$p_{mdp} =$
0.14 0.01 0.23 0.50 0.45 0.22 0.15 0.10 0.45 0.33 free to busy
0.32 0.48 0.26 0.10 0.04 0.25 0.40 0.38 0.15 0.19 busy to free
0.44 0.47 0.30 0.02 0.03 0.08 0.29 0.45 0.10 0.29 free to free
0.10 0.05 0.21 0.38 0.48 0.45 0.16 0.17 0.40 0.25 busy to busy

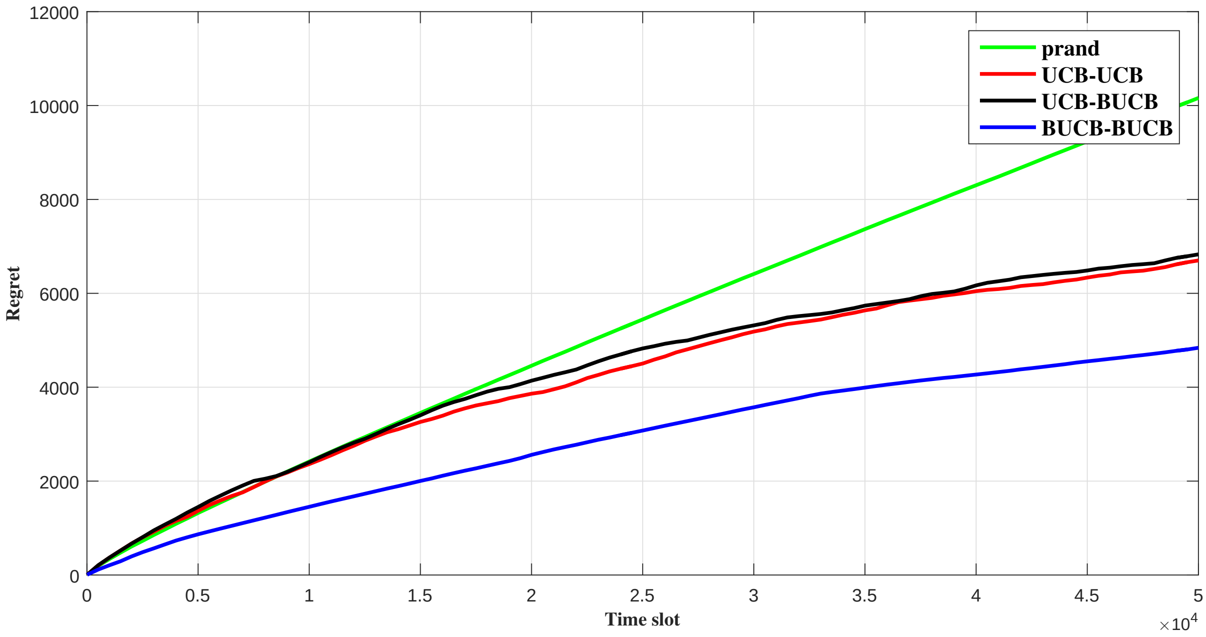

Figure 4.1: Normalised regret $\frac{R(n)}{\log n}$ vs time slot for no. of SUs is 4

As it can be seen from fig 4.1 and 4.2,there is a significant improvement from Random-UCB to BUCB-BUCB. UCB-BUCB and UCB-UCB performs almost the same. The regret in case of BUCB-BUCB improves by 51% compared to Random-UCB and 28% from UCB-BUCB and UCB-BUCB when the number of SUs is 4. For number of SUs to be 6, the regret decreases by 60% as compared to Random-UCB and 20% for both UCB and BUCB. From Fig. 4.3 collision is negligible compared to Random-UCB. However it can be seen from fig 4.4, for the collision avoidance the switching cost is very high. Keeping this in mind, the rest of the simulations, both for fixed number of SUs and varying number of SUs and that of RLUR and SERL are done for this combination, i.e. both the statistics are calculated using BUCB.

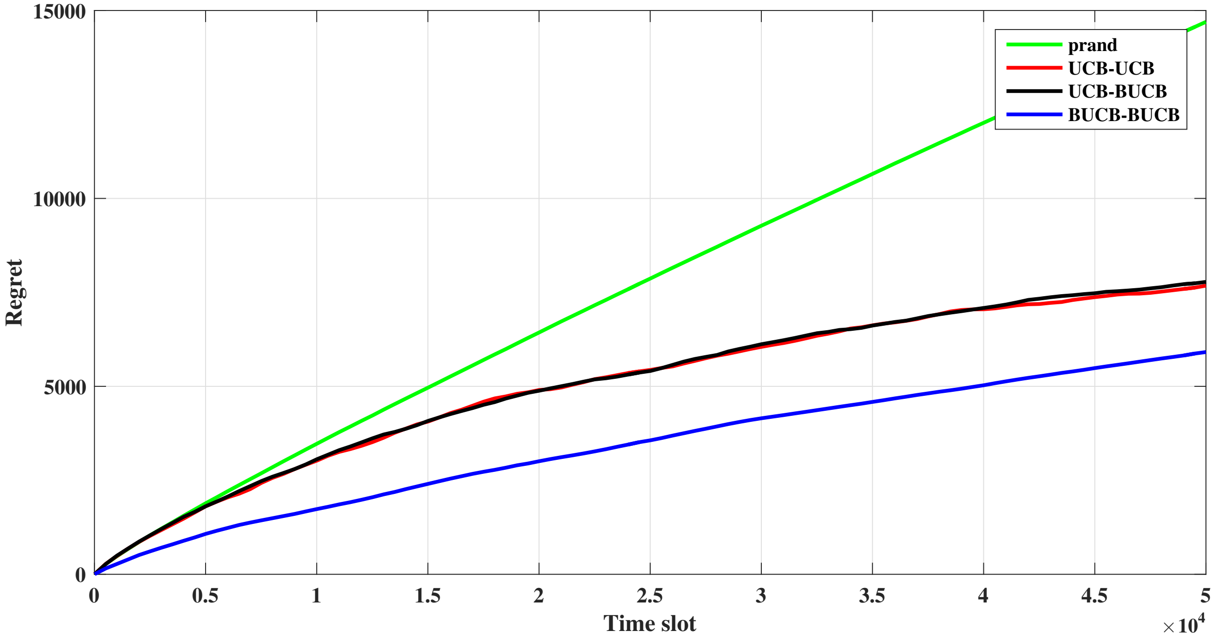

Figure 4.2: Normalised regret $\frac{R(n)}{\log n}$ vs time slot for no. of SUs is 6

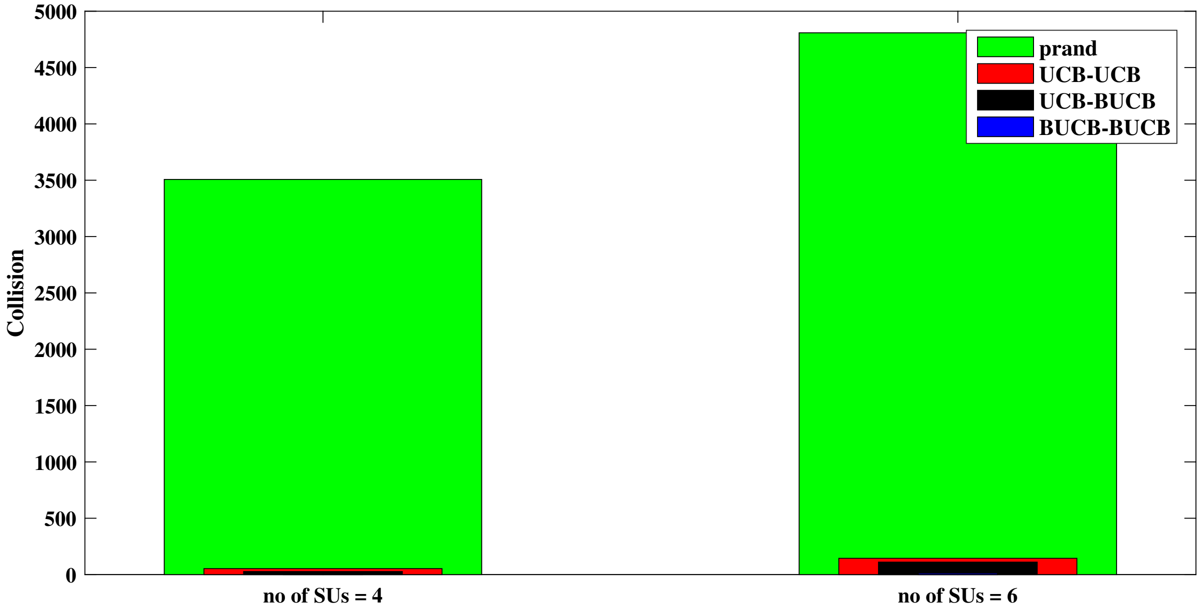

Figure 4.3: Total no. of collision for synchronized SU

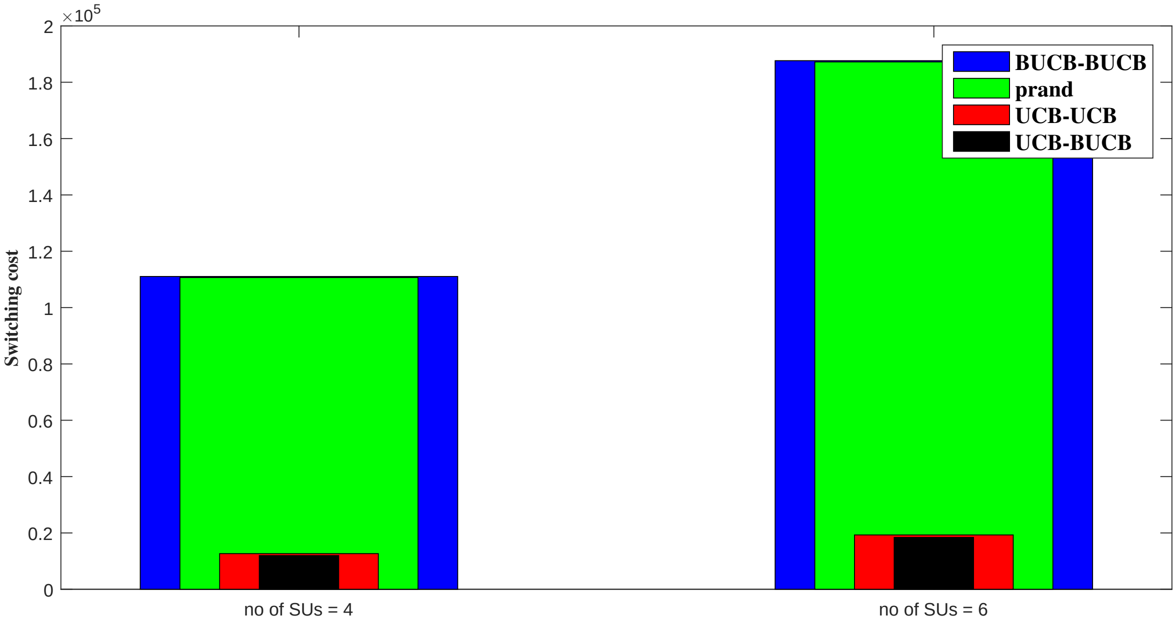

Figure 4.4: Total switching cost for synchronized SU

4.2 Algorithm : SU Estimation with Rank learning (SERL)

For the first C time slots, each subband is selected once and when the selected subband is unoccupied by the PU, SU transmits on that subband as shown in steps 1 to 9. For the rest of the time slot *C+1* to H, the quality index of the rank and the subbands are computed as shown in steps 11 to 17. User u selects subband $c_u(t)$ for transmission in step 18. If the subband $c_u(t)$ is unoccupied by PU, the SU transmits else it waits for the next time slot and updates the statistics accordingly as shown in step 19 to 23.The SUs try to estimate the number of SUs in the network N_u in step 24 to 30 as in [24].

The value of N_u is initialized to *1*. The algorithm then calculates the value of N_u after every t_{int} time slots until The computation is done as, $t \leq range$.

$$\begin{aligned} N_u \ &:= \ \min\left(round\left(\frac{\log\left(\frac{t-C_u}{t}\right)}{log\left(1-\frac{1}{C}\right)} + 1 \right), C \right) \\ and\, N_u \ &:= \ C \ \ if \ \ C_u == t \end{aligned} \tag{4.5}$$

Beyond the range i.e. $t > range$, the value of N_u is fixed at that value of N_u corresponding which the maximum reward was obtained. In case of a collision on the subband $c_u(t)$, the SU chooses a new rank as in step 31 as follows.

$$\begin{aligned} & R_u(t) = r \\ & s.t. \ \textit{h(r,t)} \ \text{ is } \ \text{max. , value } \ \text{among } [h(1,t), h(2,t) \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad ..., h(N_u,t)] \end{aligned} \tag{4.6}$$

Algorithm 2 SU Estimation with Rank learning (SERL)

```
Input: $X_{i,u}(t-1), T_{i,u}(t-1), Y_{r,u}(t-1), D_{r,u}(t-1)$ , $R_u, t_{int}, c_u, H$
Output: $c_u(t)$
for $t \leq C$ do
    if (any $T_{(:,u)}(t-1) == 0$ ) then
        $c_u(t) = i$ s.t $T_{i,j}(t-1) = 0$
        if $c_u(t)$ is vacant then
            $X_{c_u(t),u}(t) = 1$
        end if
        $T_{c_u(t),u}(t) = 1$
    end if
end for
for $t = C + 1$ to $H$ do
    $X_{c_u(t),u}(t) = X_{c_u(t),u}(t-1)$
    $T_{c_u(t),u}(t) = T_{c_u(t),u}(t-1)$
    $Y_{R_u(t),u}(t) = Y_{R_u(t),u}(t-1)$
    $D_{R_u(t),u}(t) = D_{R_u(t),u}(t-1)$
    for $i = 1$ $to$ $C$ do
        Compute
        $g(i,t) = BUCB(X_{i,u}(t), T_{i,u}(t), t)$ $Eq.3.5$
        $h(i,t) = BUCB(Y_{i,u}(t), D_{i,u}(t), t)$ $Eq.3.5$
    end for
    Select subband $c_u(t)$ as per Eq. 4.1
    if $c_u(t)$ is vacant then
        $X_{c_u(t),u}(t) = X_{c_u(t),u}(t) + 1$
        $D_{R_u(t),u}(t) = D_{R_u(t),u}(t) + 1$
    end if
    $T_{c_u(t),u}(t) = T_{c_u(t),u}(t) + 1$
    if t is a multiple of $t\_int$ and less than $range$ then
        Calculate $N_u(t)$ as in Eq. 4.5
    end if
    $S_{int,u}$ is updated
    if t == $range$ + 1 then
        $N_u = N_u$ corresponding max($S_{int,u}$)
    end if
    if Collision occurs in $c_u(t)$ then
        Select a rank $R_u(t)$ for the SU $u$ from 1 to $N_u(t)$ according to Eq. 4.6
        $C_u = C_u + 1$
    else
        $Y_{R_u(t),u}(t) = Y_{R_u(t),u}(t) + 1$
    end if
end for
```

The value of C_u is updated depending on if the SU has faced a collision or not as shown in 33. The C_u in step 33 is the total number collision faced by a SU u.

4.3 Simulation Result

The experiment is conducted 10 independent times. The horizon of the experiment is 50000 time slots. The subbands are all considered to be of the same quality, though different probability of vacancy and hence the reward is considered to be 1 for all the subbands. The number of subbands considered are 10 and simulation results for both number of SUs 4 and 6 are shown. The SUs are all synchronized and hence, enters and leaves the network at the same time.

For the musical chair experiment, the learning time duration is taken as 3000 time slots as per the assumption in [24]. The $Rank\_reward$ mentioned in RLUR is set to vary from 0.9 for the rank C to 1 for the rank *1*. Hence, $Rank\_reward = 1 : -(0.1/(C-1)) : 0.9$. The t_{int} mentioned in SERL is fixed at 400 and the range is taken as $5 * t_{int}$.

The Experiment is conducted for two sets of probability for the MDP, p_{mdp} and are denoted as case 1 and case 2.

The p_{mdp} for case 1 is
0.14 0.01 0.23 0.50 0.45 0.22 0.15 0.10 0.45 0.33 free to busy
0.32 0.48 0.26 0.10 0.04 0.25 0.40 0.38 0.15 0.19 busy to free
0.44 0.47 0.30 0.02 0.03 0.08 0.29 0.45 0.10 0.29 free to free
0.10 0.05 0.21 0.38 0.48 0.45 0.16 0.17 0.40 0.25 busy to busy

The p_{mdp} for case 1 is
0.21 0.16 0.34 0.22 0.03 0.05 0.45 0.34 0.48 0.30 free to busy
0.32 0.44 0.17 0.25 0.46 0.38 0.03 0.11 0.04 0.23 busy to free
0.30 0.32 0.06 0.27 0.48 0.47 0.02 0.20 0.08 0.26 free to free
0.17 0.08 0.43 0.26 0.03 0.10 0.50 0.35 0.40 0.21 busy to busy
The probability of vacancy is calculated from the above set of p_{mdp}

In Figure 4.5 and 4.8, the straight line is the output regret for case 1 and the marker line is the output regret for case 2.

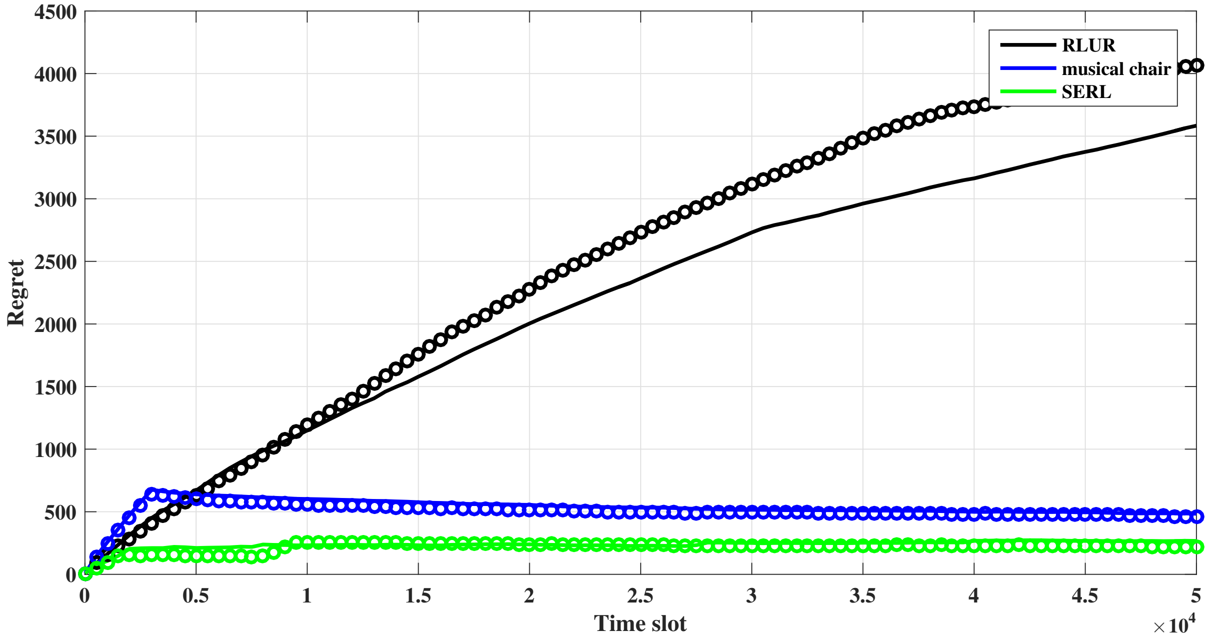

Figure 4.5: Normalised regret $\frac{R(n)}{\log n}$ vs time slot for synchronized SU(no. of SUs is 4)

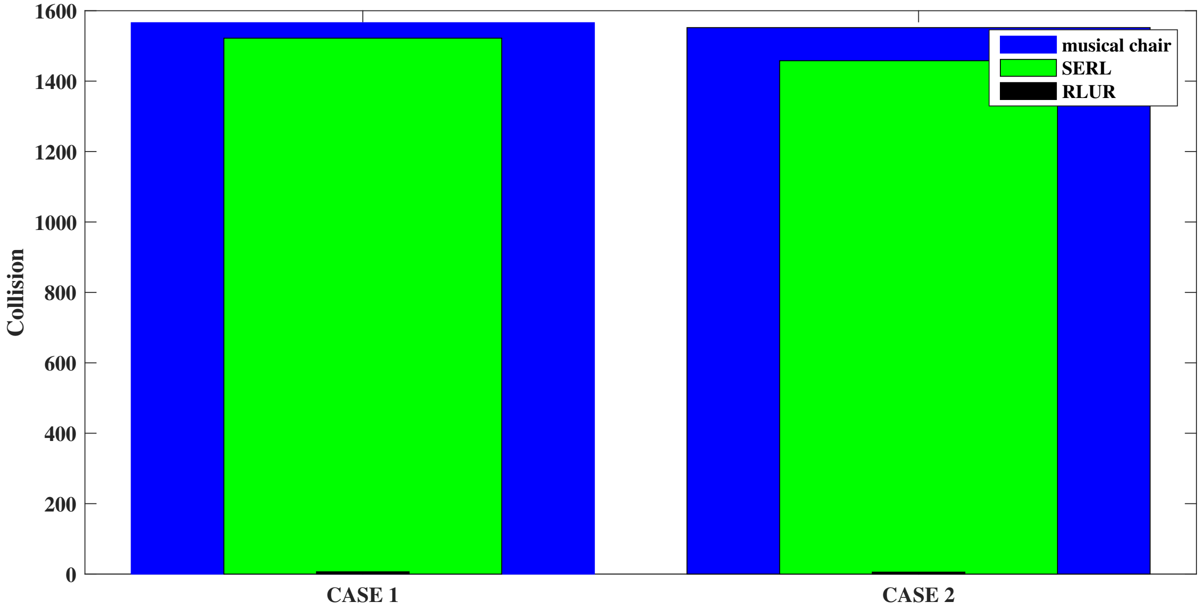

Figure 4.6: Total no. of collision for synchronized SU(no. of SUs is 4)

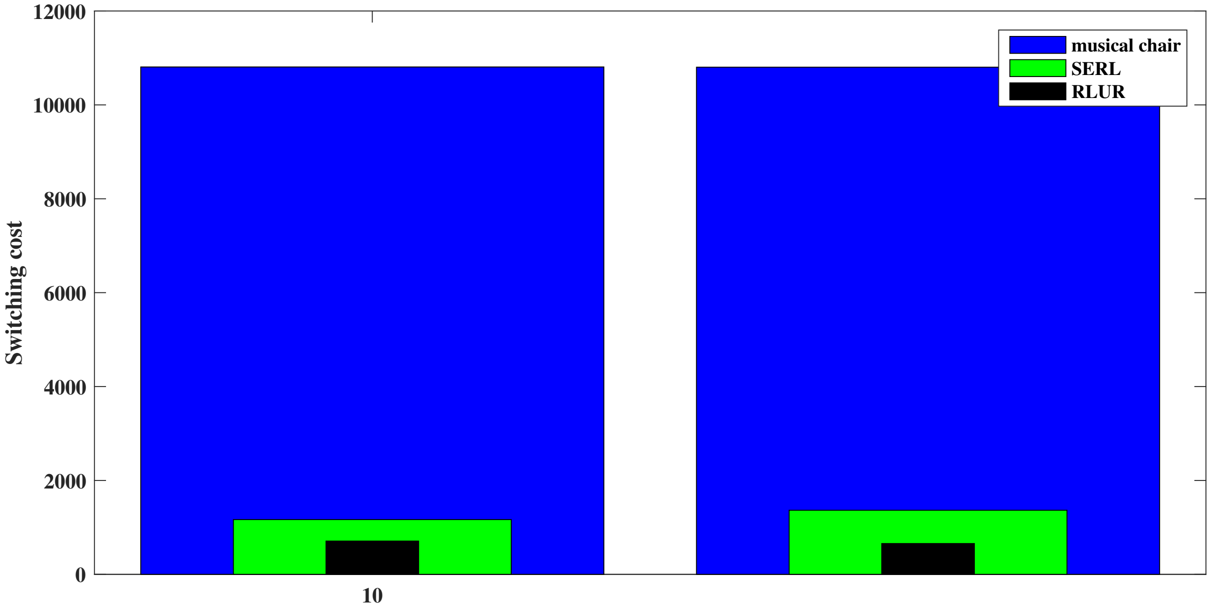

Figure 4.7: Total switching cost incurred for synchronized SU(no. of SUs is 4)

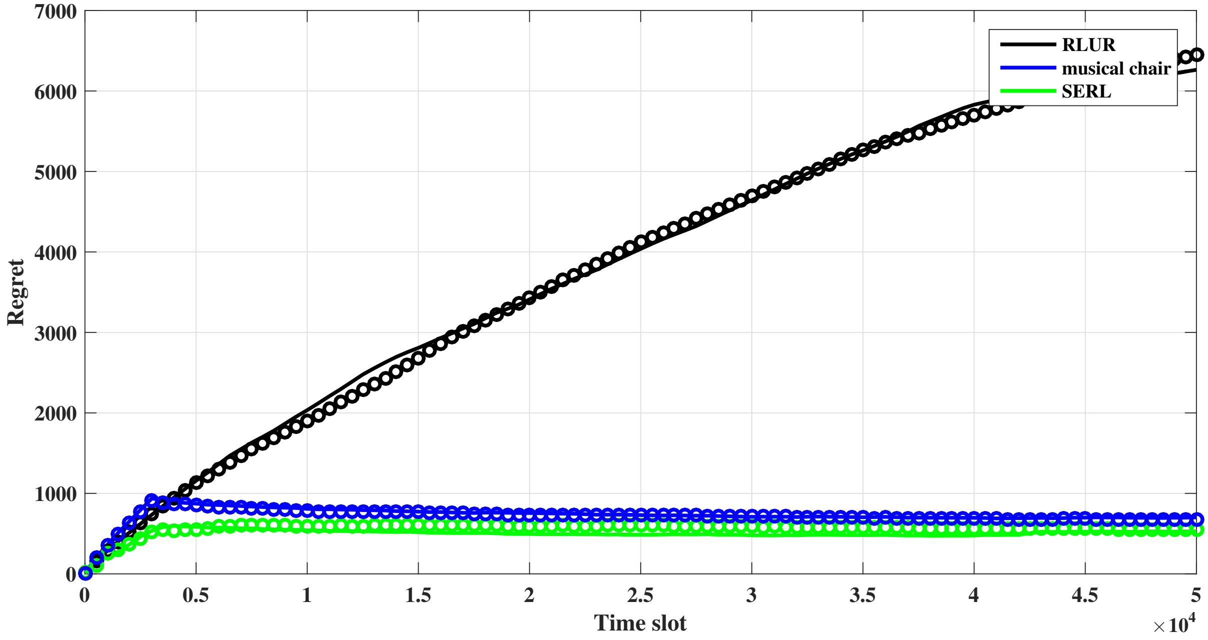

Figure 4.8: Normalised regret $\frac{R(n)}{\log n}$ vs time slot for synchronized SU(no. of SUs is 6)

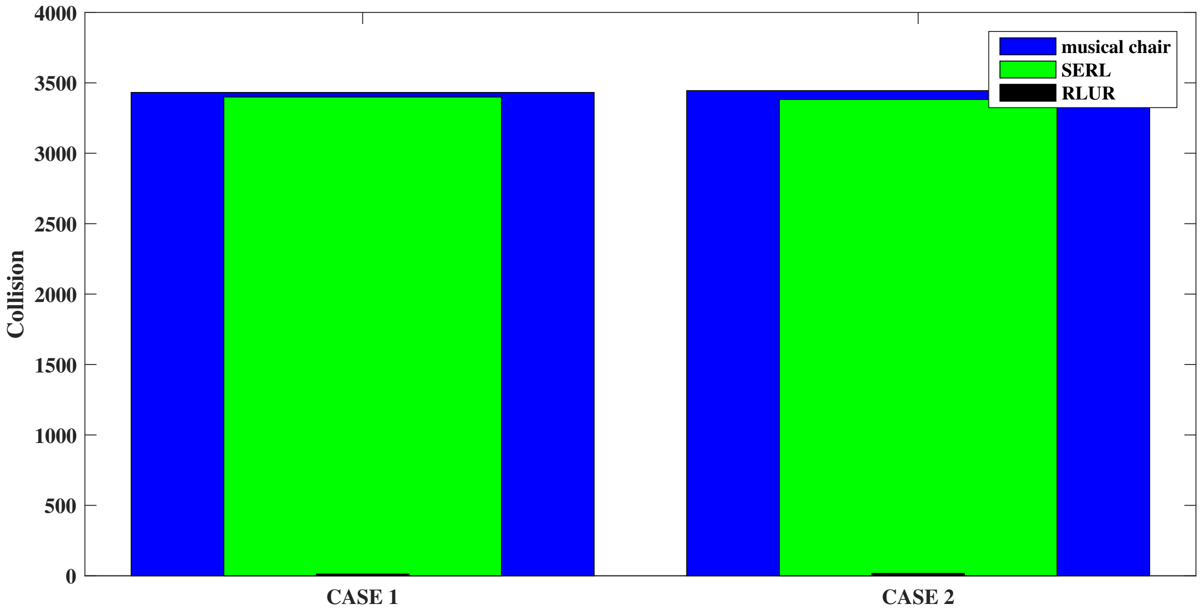

Figure 4.9: Total no. of collision for synchronized SU(no. of SUs is 6)

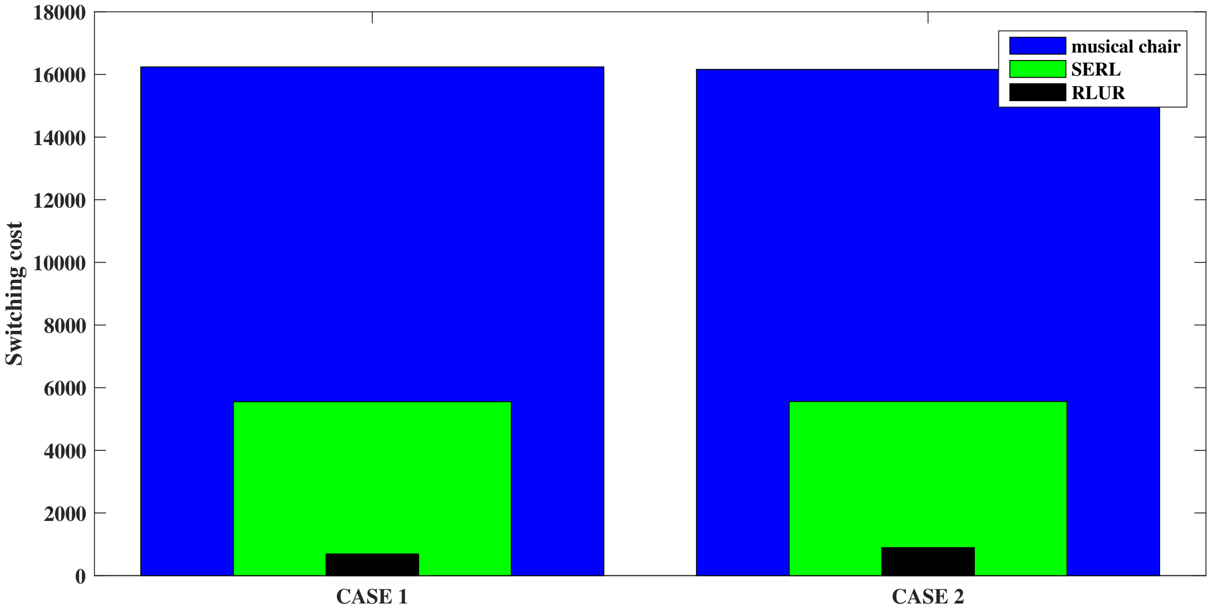

Figure 4.10: Total switching cost incurred for synchronized SU(no. of SUs is 6)

4.4 Inference

As it can be observed from Figure 4.5 and 4.8, the Musical chair algorithm increases steeply for the learning time slot 3000 and then, the number of SUs is predicted. The SUs then settles into the top subbands orthogonally as the lack of regret confirms that the allocation is the same as that of the best selection. As per the regret in SERL, the regret increases steeply for the learning time interval t_{int}, 400. That is when the number of SUs is first predicted. It then gradually settles into the optimum subbands at around 2500. However, it can be seen that although the regret in RLUR increase at a slower speed, it doesn't predict the number of SUs or settles into optimum subbands and hence, the regret keeps increases.

For SERL, the regret accumulated decreases by 45% in case of 4 SUs and 25% in case of 6 SUs compared to Musical Chair. The number of collision is almost same in both the algorithm SERL and MC. The switching cost decreases approximately by 87% in case of 4 and 65% in case of 6 SUs.

However, in real time scenario, assuming that all the SUs enter and exit the system exactly at the same instance is an unrealistic assumption. Hence, in the following chapter, the above formulated algorithms will be extended for the case where the enter and exit of SUs are not synchronized.

Chapter 5

Proposed DMP for variable number of SUs

In this chapter, the two DMPs proposed in the Chapter 4 are discussed for time varying number of SUs. In the proposed DMP, U SUs opportunistically trying to access an unoccupied subband in C subbands. Here, the number of SU in the system U represents the maximum number of SUs in the system at any time slot and can not exceed the number of subband, C. H is the size of the horizon, $c_u(t)$ is the subband selected by SU u at time slot t. t_u is the number of time spent by SU u in network upto time t and H_u is the total time slot SU u spent in the network. As we have already established that the number of SUs is not fixed, the SUs have no idea of the range of the rank. Though it can still be assured that $R_u \in 1, 2, ..., C$, as a rank $R_u > C$ does not denote to any subband.

At each time slot t a SU can enter or exit the network with probability unique to the SU. However, we make a fair assumption that once a SU enters the network, it stays there for sufficient time to learn subband statistics. Hence, $p_{entering} \geq p_{leaving}$.

5.1 Algorithm : Rank Learning with Unequal Rank-Reward (RLUR)

As already explained in Section 4.1, the unequal rank reward forces the SU to choose a better rank to operate in. So, when ever a SU which was transmitting in a good subband leaves the system, some other SU moves to the now empty rank and chooses the good subband. Hence, this algorithm takes care that the better performing subband does not stay empty when the SU using it leaves.

Here, we are trying to validate the claim that the SUs occupies the subband with high vacancy of probability for greater portion of the horizon and move to a better performing subband when the SU using it leaves. The order of subband as per decreasing order of probability of vacancy for this experiment is $1, 7, 8, 2, 3, 5, 6, 4$

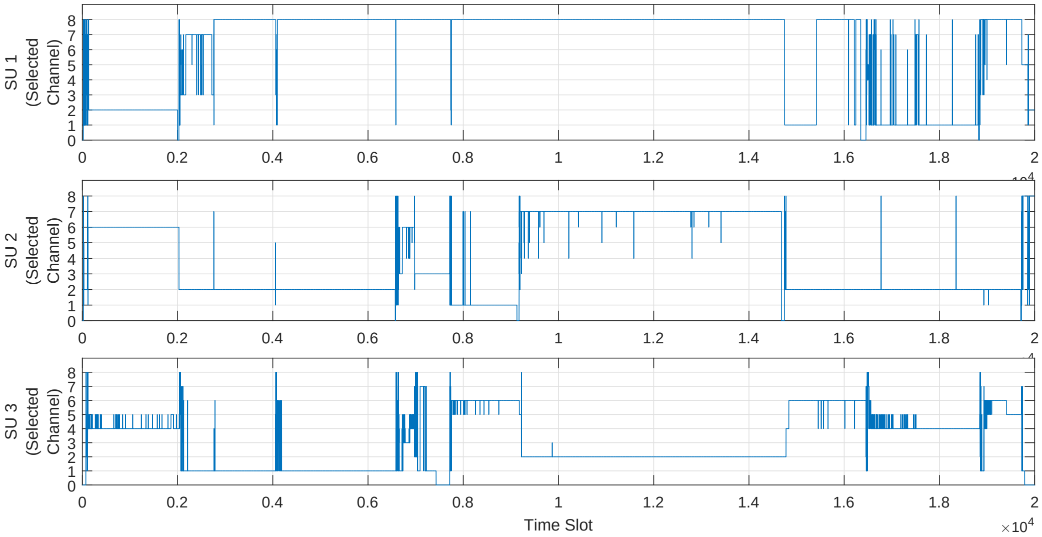

Figure 5.1: Time slot vs the subband that was selected in that time slot, for number of SU is 3 and number of subband is 8

5.2 Algorithm : SU Estimation with Rank learning (SERL)

The SERL is the same as described in Section 4.2. Except a small difference that the prediction of number of SU is not restricted to a certain range and is calculated throughout the experiment to cope with the varying number of SUs.

Algorithm 3 Dynamic SU Estimation with Rank learning (Dynamic-SERL)

```
Input: $X_{i,u}(t-1), T_{i,u}(t-1), Y_{r,u}(t-1), D_{r,u}(t-1)$ , $R_u, t_{int}, c_u, H_u$
Output: $c_u(t)$
for $t_u \leq C$ do
    if (any $T_{(:,u)}(t-1) == 0$ ) then
        $c_u(t) = i$ s.t $T_{i,j}(t-1) = 0$
        if $c_u(t)$ is vacant then
            $X_{c_u(t),u}(t) = 1$
        end if
        $T_{c_u(t),u}(t) = T_{i,j}(t-1) + 1$
    end if
end for
for $t_u = C + 1$ to $H_u$ do
    $X_{c_u(t),u}(t) = X_{c_u(t),u}(t-1)$
    $T_{c_u(t),u}(t) = T_{c_u(t),u}(t-1)$
    $Y_{R_u(t),u}(t) = Y_{R_u(t),u}(t-1)$
    $D_{R_u(t),u}(t) = D_{R_u(t),u}(t-1)$
    for $i = 1$ to $C$ do
        Compute
        $g(i,t) = BUCB(X_{i,u}(t), T_{i,u}(t), t)$  $Eq.3.5$
        $h(i,t) = BUCB(Y_{i,u}(t), D_{i,u}(t), t)$  $Eq.3.5$
    end for
    Select subband $c_u(t)$ as per Eq. ??
    if $c_u(t)$ is vacant then
        $X_{c_u(t),u}(t) = X_{c_u(t),u}(t) + 1$
        $D_{R_u(t),u}(t) = D_{R_u(t),u}(t) + 1$
    end if
    $T_{c_u(t),u}(t) = T_{c_u(t),u}(t) + 1$
    if t is a multiple of $t\_int$ then
        Calculate $N_u(t)$ as in Eq 4.5
    end if
    if Collision occurs in $c_u(t)$ then
        Select a rank $R_u(t)$ for the SU $u$ from 1 to $N_u(t)$ according to Eq. 4.6
        $C_u = C_u + 1$
    else
        $Y_{R_u(t),u}(t) = Y_{R_u(t),u}(t) + 1$
    end if
end for
```

5.3 Simulation Result

The experiment is conducted in the same environment as the Section 4. 10 independent experiments are conducted and averaged over it. The horizon of the experiment is 50000 time slots. The subbands are all considered to be of the same quality and hence the reward is considered to be 1 for all the subbands. The number of subbands considered are 10 and simulation results for both maximum number of SUs in network is 4 and 6 are shown. The SUs are not synchronized and hence, do not enter and leave the network at same time. However, the entering and exiting process happens only after an interval of a certain epoch time period. The process of enter and exit from the network is a Bernoulli process with probabilities, $p_{entering} = 0.1$ and $p_{leaving} = 0.01$

For the musical chair experiment, the learning time duration is taken as 3000 time slots as per the assumption in [24]. The epoch after which the MC restarts and the SUs can enter or exit is taken as 5000. The $Rank\_reward$ mentioned in RLUR is set to vary from 0.9 for the rank C to 1 for the rank *1*. Hence, $Rank\_reward = 1 : -(0.1/(C-1)) : 0.9$. The t_{int} mentioned in Dynamic-SERL is fixed at 400.

The Experiment is conducted for two sets of probability for the MDP, p_{mdp} and are denoted as case 1 and case 2.

The p_{mdp} for case 1 is
0.14 0.01 0.23 0.50 0.45 0.22 0.15 0.10 0.45 0.33 free to busy
0.32 0.48 0.26 0.10 0.04 0.25 0.40 0.38 0.15 0.19 busy to free
0.44 0.47 0.30 0.02 0.03 0.08 0.29 0.45 0.10 0.29 free to free
0.10 0.05 0.21 0.38 0.48 0.45 0.16 0.17 0.40 0.25 busy to busy

The p_{mdp} for case 1 is
0.21 0.16 0.34 0.22 0.03 0.05 0.45 0.34 0.48 0.30 free to busy
0.32 0.44 0.17 0.25 0.46 0.38 0.03 0.11 0.04 0.23 busy to free
0.30 0.32 0.06 0.27 0.48 0.47 0.02 0.20 0.08 0.26 free to free
0.17 0.08 0.43 0.26 0.03 0.10 0.50 0.35 0.40 0.21 busy to busy
The probability of vacancy is calculated from the above set of p_{mdp}

In Figure 5.2 and 5.5, the straight line is the output regret for case 1 and the marker line is the output regret for case 2.

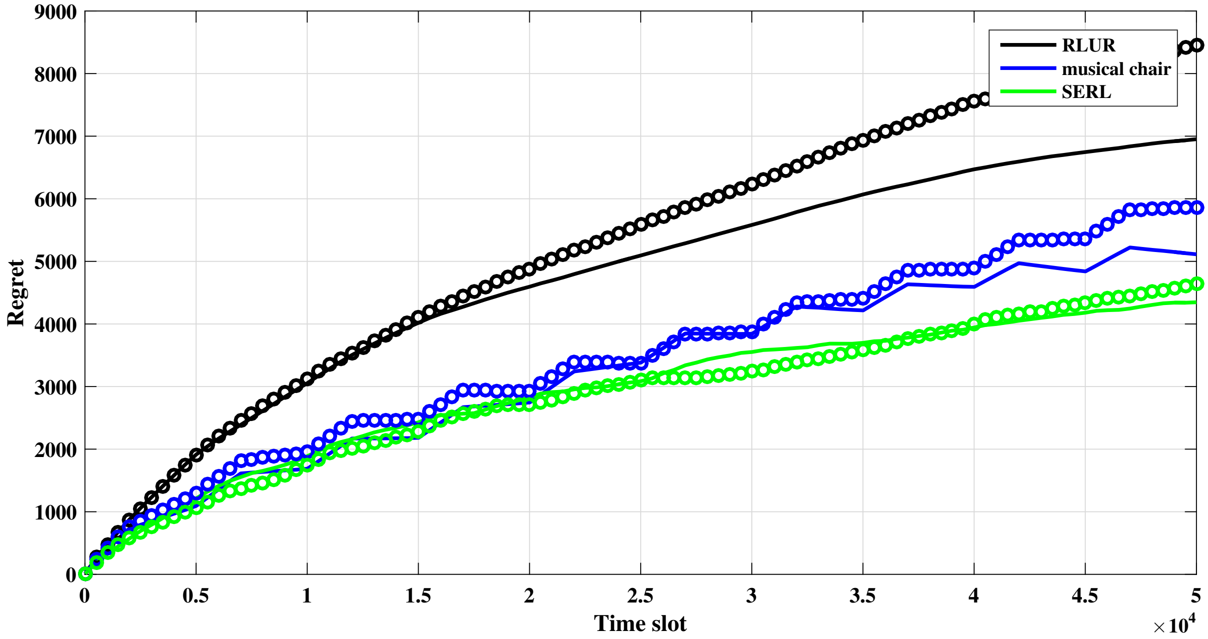

Figure 5.2: Normalised regret $\frac{R(n)}{\log n}$ vs time slot for varying SU(maximum no. of SUs is 4)

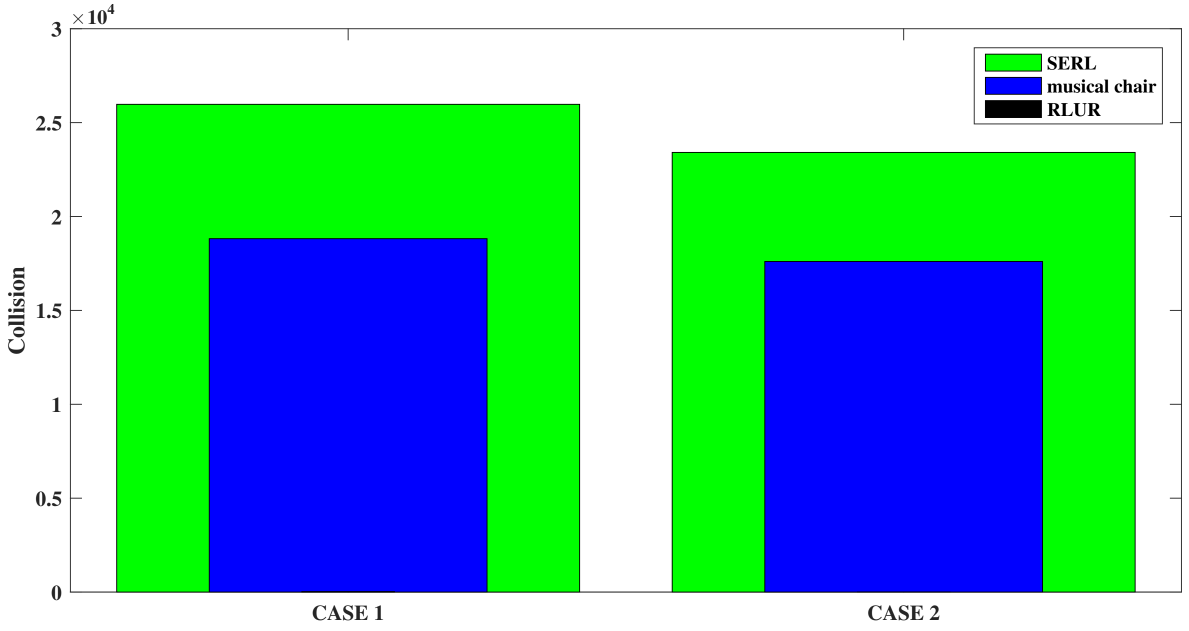

Figure 5.3: Total no. of collision for varying SU(maximum no. of SUs is 4)

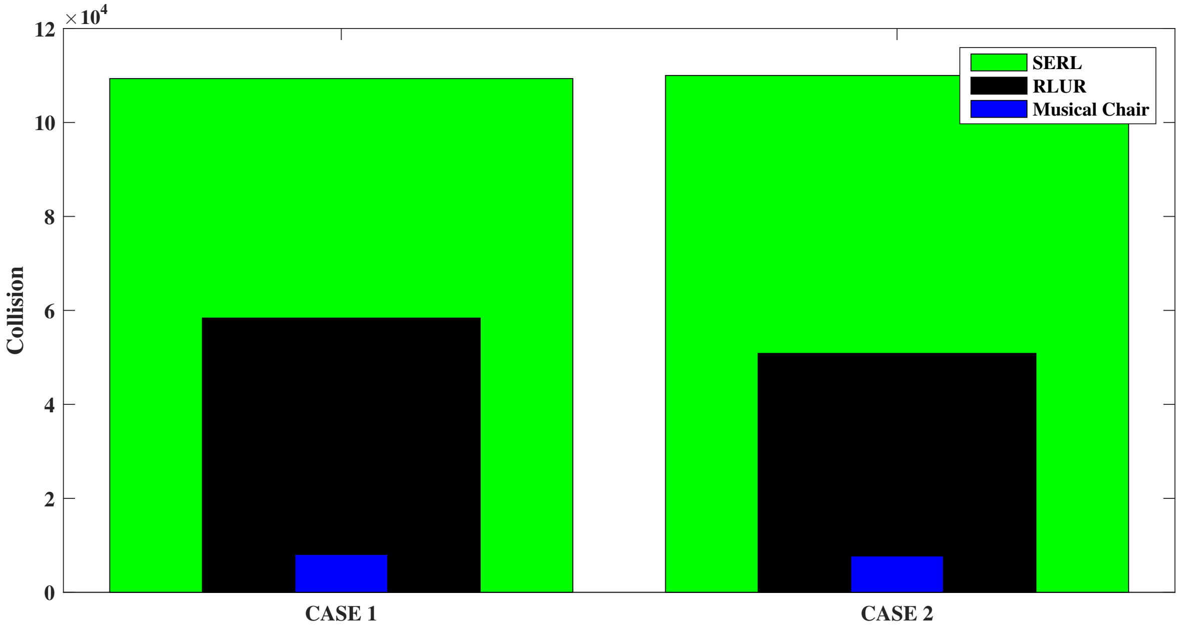

Figure 5.4: Total switching cost incurred for varying SU(maximum no. of SUs is 4)

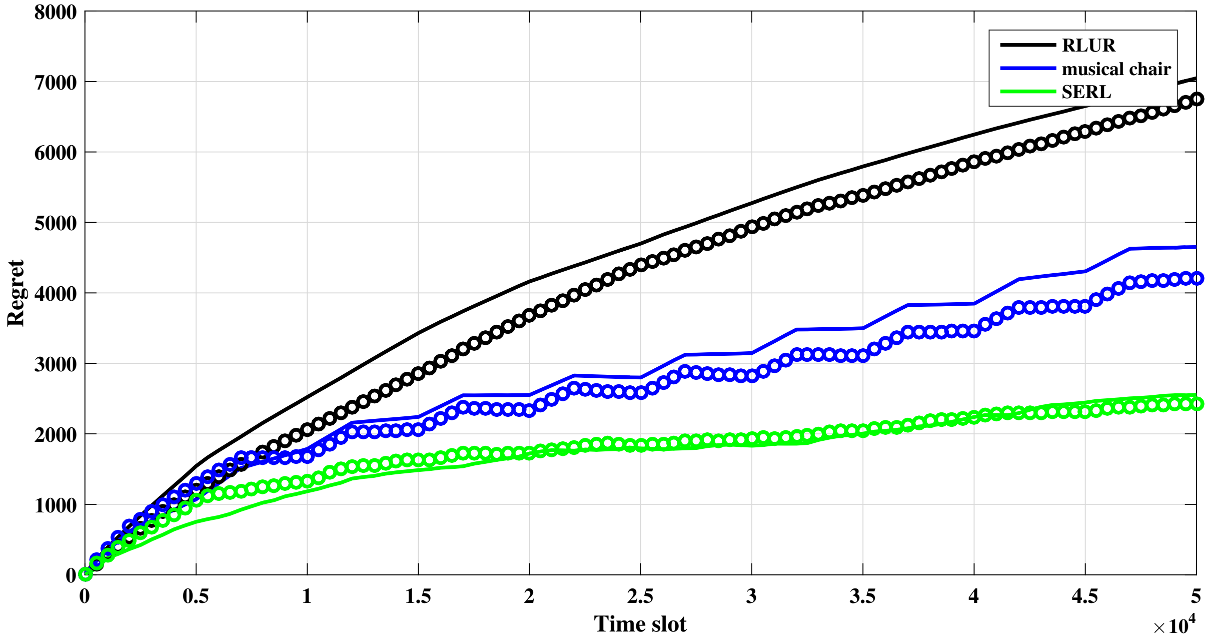

Figure 5.5: Normalised regret $\frac{R(n)}{\log n}$ vs time slot for varying SU(maximum no. of SUs is 6)

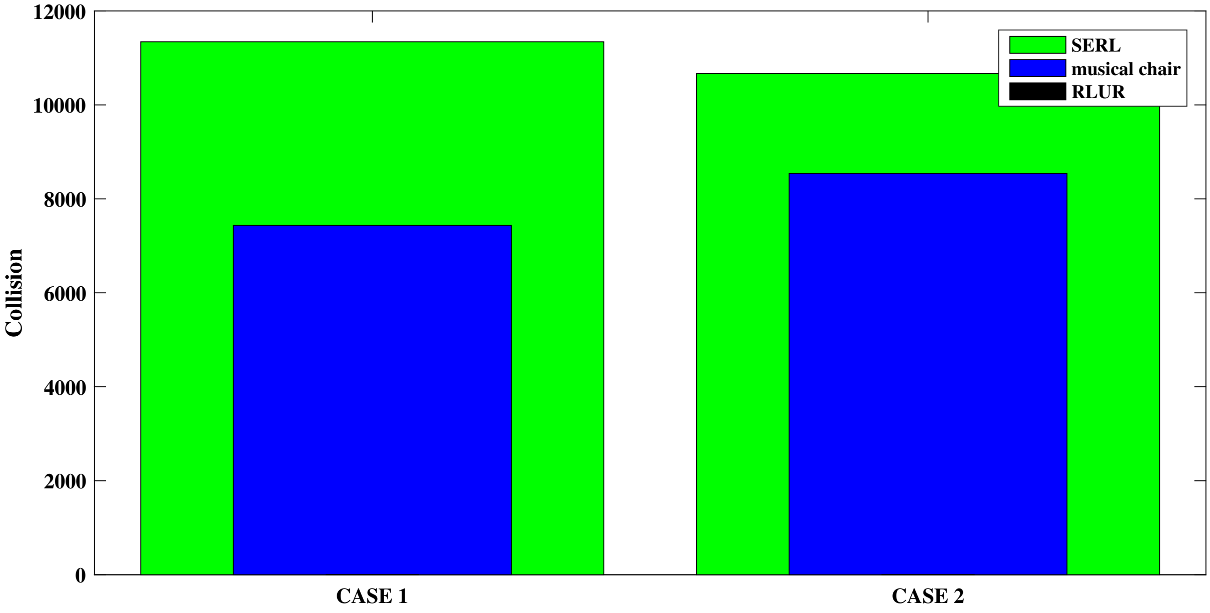

Figure 5.6: Total no. of collision for varying SU(maximum no. of SUs is 6)

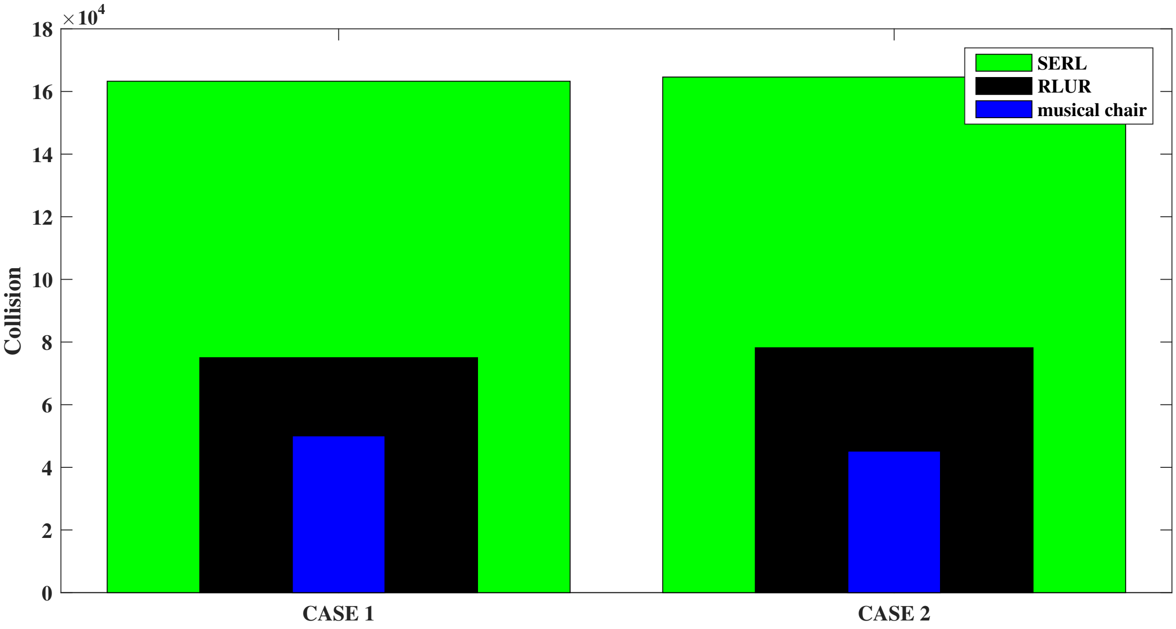

Figure 5.7: Total switching cost incurred for varying SU(maximum no. of SUs is 6)

5.4 Inference

As it can be observed from Figure 5.2 and 5.5, All the 3 algorithms increases without converging which is due to the dynamic nature of the algorithm. It can be seen RLUR still performs the worst compared to the Musical chair and the Dynamic-SERL. Also, it can be seen that Dynamic-SERL performs the best. The Musical chair algorithm increases steeply for the learning time slot 3000 in each epoch and then predicts the number of SUs. The SUs then settles into the top subbands orthogonally for the rest of the time slots. As per the regret in SERL, the regret increases steeply for the learning time interval. At this point the Dynamic-SERL performs similar to the Musical chair. However, as the experiment proceeds, Musical chair accumulates higher regret due to the learning period in each epoch. Where as the continuous learning in Dynamic-SERL accumulates lower regret. Also, it can be seen that the rate at which regret in RLUR increase slows down after a while as the SUs starts to settle into higher ranks. But it doesn't predict the number of SUs or settles into optimum subbands and hence, the regret increases at a higher rater than the other algorithms.

For Dynamic-SERL, the regret accumulated decreases by 40% in case of 4 SUs and 20% in case of 6 SUs compared to Musical Chair. The decrease in regret is 64% in case of 4 SUs and 41% for 6 SUs. The number of collision and the switching cost are both high compared to the other algorithms.

Chapter 6

Mathematical Proof

In this section, we have mathematically calculate the probability that all the SUs in the network are orthogonally assigned to the subbands. Two different scenario are considered namely, random rank selection without the knowledge of the no. of SU and the rank learning scenario again without the knowledge of the no. of SU. The calculated probabilities are then compared with that of ρ_{rand}, which is explained in [9].

6.1 ρ_{rand}

Here, a case is considered where the SUs are aware of the number of SUs in the system. Hence the rank is chosen from 1 to U. The number of Markovian states the system can be in is $\binom{2U-1}{U}$. As we know, only the orthogonal configuration is absorbing state. The number of orthogonal states is 1. When the rank is randomly chosen, the probability of reaching any state is equiprobable. Hence, the probability of reaching an absorption state is $p_{\rho_{rand}} = \binom{2U-1}{U}^{-1}$ as in Equ. 35 in [9].

$$p_{\rho_{rand}} = \frac{U!(U-1)!}{(2U-1)!} \tag{6.1}$$

6.2 Random Rank Selection Without the Knowledge of the no. of SU

Consider a case where the SUs are unaware of the number of SUs in the system. Hence, the rank can be chosen from 1 to C. The number of Markovian states the system can be in is $\binom{U+C-1}{U}$. As we know, only the orthogonal configuration is absorbing state. The number of orthogonal states are $\binom{C}{U}$. When the rank is randomly chosen, the probability of reaching any state is equiprobable. Hence, the probability of reaching an absorption state is

$$p_{random} = \binom{C}{U} * \binom{U+C-1}{U}^{-1}$$

$$p_{random} = \frac{C!(C-1)!}{(C+U-1)!(C-U)!} \tag{6.2}$$

6.3 Rank Selection Based on Rank Learning

The statistic of the rank is calculated and the rank with the highest value of statistic is selected. The statistic is calculated as:

$$h(n) = \frac{X}{T} + \sqrt{\frac{2\log n}{T}} \tag{6.3}$$

Let $h_c(n)$ is the statistic of the rank selected in n^{th} time slot and $h_r(n)$ is the statistic of any of the rest of the ranks. Let the statistic at n^{th} time slot be given by:

$$h_c(n) = \frac{X_c}{T_c} + \sqrt{\frac{2\log n}{T_c}} \tag{6.4}$$

$$h_r(n) = \frac{X_r}{T_r} + \sqrt{\frac{2\log n}{T_r}} \tag{6.5}$$

When the SU decides to transmit in a time slot it encounters one of the two scenarios, unsuccessful transmission corresponding to occurrence of collision and successful transmission corresponding to non-occurrence of collision

6.3.1 Occurrence of Collision

When collision occurs, for the rank c the rank reward accumulated is 0 and the number of times the rank is chosen will increase by 1. However, no changes will be there for the rest of the rank. Hence, at $(n+1)^{th}$ time slot the statistics update to

$$h_c(n+1) = \frac{X_c}{T_c+1} + \sqrt{\frac{2\log(n+1)}{T_c+1}} \tag{6.6}$$

$$h_r(n+1) = \frac{X_r}{T_r} + \sqrt{\frac{2\log(n+1)}{T_r}} \tag{6.7}$$

Comparing Eq. 6.5 and Eq. 6.7, it can be seen that the change occurs in the numerator of the second term. However, for a large value of n, $\log n+1$ increases very slowly. Hence $\sqrt{2\log(n+1)}$ term is almost constant, making $h_r(n)$ almost constant. Hence, the value of h_r for the i^{th} best channel can never surpass the value of h for the $(i+1)^{th}$ best rank.

Comparing Eq. 6.4 and Eq. 6.6, it can be seen that both the term decreases. Decrease in first term is,

$$\begin{aligned}\frac{X_c}{T_c} - \frac{X_c}{T_c+1} &= \frac{X_cT_c + X_c - X_cT_c}{T_c(T_c+1)} \\ &= \frac{X_c}{T_c(T_c+1)}\end{aligned} \tag{6.8}$$

Decrease in second term is,

$$\begin{aligned}\frac{\sqrt{2\log n}}{\sqrt{T_c}} - \frac{\sqrt{2\log(n+1)}}{\sqrt{T_c+1}} &= \frac{\sqrt{2\log n}}{\sqrt{T_c}}\left[1 - \frac{1}{\sqrt{1+\frac{1}{T_c}}}\right] \\ &= \frac{\sqrt{2\log n}}{\sqrt{T_c}}\left[1 - \left(1 - \frac{1}{2T_c} + \frac{3}{8T_c^2} - \frac{5}{16T_c^3} + ...\right)\right] \\ &= \frac{\sqrt{2\log n}}{\sqrt{T_c}}\left[\frac{1}{2T_c} - \frac{3}{8T_c^2} + \frac{5}{16T_c^3} - ....\right] \\ &= \frac{\sqrt{2\log n}}{2T_c\sqrt{T_c}}\end{aligned} \tag{6.9}$$

as $\frac{1}{T_c}$ is small.

Hence,

$$D1 = h_c(n) - h_c(n+1) = \frac{X_c}{T_c(T_c+1)} + \frac{\sqrt{2\log n}}{2T_c\sqrt{T_c}} \tag{6.10}$$

So, from equations 6.10 the probability of choosing c again in $(n+1)^{th}$ time slot is,

$$\begin{aligned}p &= \mathbf{P}\left[(h_c(n) - h_{r_2}(n)) \geq \left(\tfrac{X_c}{T_c(T_c+1)} + \tfrac{2\log n}{2T_c\sqrt{T_c}}\right)\right] \\ &= \mathbf{P}\left[(\tfrac{X_c}{T_c} + \sqrt{\tfrac{2\log n}{T_c}} - h_{r_2}(n)) \geq \left(\tfrac{X_c}{T_c(T_c+1)} + \tfrac{\sqrt{2\log n}}{2T_c\sqrt{T_c}}\right)\right] \\ &= \mathbf{P}\left[(\tfrac{X_c}{T_c} - \tfrac{X_c}{T_c(T_c+1)}) \geq \left(h_{r_2}(n) + \tfrac{\sqrt{2\log n}}{2T_c\sqrt{T_c}} - \sqrt{\tfrac{2\log n}{T_c}}\right)\right] \\ &= \mathbf{P}\left[(\tfrac{X_c}{T_c+1}) \geq \left(h_{r_2}(n) + \tfrac{1-2T_c}{2T_c}\sqrt{\tfrac{2\log n}{T_c}}\right)\right] \\ &\geq \mathbf{P}\left[(\tfrac{X_c}{T_c}) \geq \left(h_{r_2}(n) - \sqrt{\tfrac{2\log n}{T_c}}\right)\right] \\ &= \mathbf{P}\left[(X_c - \alpha T_c) \geq \Big(\underbrace{h_{r_2}(n) - \sqrt{\frac{2\log n}{T_c}} - \alpha}_{\text{K}}\Big)T_c\right]\end{aligned}$$

Where, α[1] is the probability of non-occurrence of collision and is very small.

Hence, using Hoeffding's inequality

[1] $\alpha = \dfrac{\dbinom{(U-1)+(C-1)-1}{U-1}}{\dbinom{U+C-1}{U}}$

$= \frac{U(C-1)}{(U+C-1)(U+C-2)}$

$$p \leq \exp(-2K^2T_c) \tag{6.11}$$

As it can be seen $2K^2T_c$ is a large quantity, the value of p is relatively very small. The probability of the second best channel being chosen is $1-p$. The other ranks can never be chosen as their value of h will never surpass that of the rank that was second best in time slot n.

6.3.2 Non-occurrence of Collision

When no collision occurs, the rank reward and the number of times the rank is chosen will both increase by 1. However, no changes will be there for the rest of the rank. Hence, at $(n+1)^{th}$ time slot the statistics update to

$$h_c(n+1) = \frac{X_c+1}{T_c+1} + \sqrt{\frac{2\log(n+1)}{T_c+1}} \tag{6.12}$$

$$h_r(n+1) = \frac{X_r}{T_r} + \sqrt{\frac{2\log(n+1)}{T_r}} \tag{6.13}$$

Comparing Eq. 6.4 and Eq. 6.12 , we can see that the first term increases and the second term decreases.

The increase in first term is,

$$\begin{aligned} \frac{X_c+1}{T_c+1} - \frac{X_c}{T_c} &= \frac{X_cT_c + T_c - X_cT_c - X_c}{T_c(T_c+1)} \\ &= \frac{T_c - X_c}{T_c(T_c+1)} \end{aligned} \tag{6.14}$$

The decrease in second term, as explained in equation. 6.9 is $\frac{\sqrt{2\log n}}{2T_c\sqrt{T_c}}$

$$D2 = h_c(n) - h_c(n+1) = -\frac{T_c - X_c}{T_c(T_c+1)} + \frac{\sqrt{2\log n}}{2T_c\sqrt{T_c}} \tag{6.15}$$

So, from equation. 6.15 the probability of choosing c again in *n+1* is,

$$\begin{aligned} q &= \mathbf{P}\left[(h_c(n) - h_{r_2}(n)) \geq \Big(-\frac{T_c - X_c}{T_c(T_c+1)} + \frac{\sqrt{2\log n}}{2T_c\sqrt{T_c}}\Big)\right] \\ &= \begin{cases} 1 & , if\left(\frac{T_c-X_c}{T_c(T_c+1)} - \frac{\sqrt{2\log n}}{2T_c\sqrt{T_c}}\right) \geq 0 \\ 1 + \left(\frac{T_c-X_c}{T_c(T_c+1)} - \frac{\sqrt{2\log n}}{2T_c\sqrt{T_c}}\right) & , otherwise \end{cases} \end{aligned} \tag{6.16}$$

As explained in earlier, $(h_c(n) - h_{r_2}(n)) \in (0,1)$. Hence, when $\left(-\frac{T_c-X_c}{T_c(T_c+1)} + \frac{\sqrt{2\log n}}{2T_c\sqrt{T_c}}\right) \leq 0$, $(h_c(n) - h_{r_2}(n)) \in (0,1)$ will always be greater and the probability $q = 1$. So, it can be inferred that q is large making is highly probable that the subband is chosen again. The second best rank is chosen with probability $1-q$

6.3.3 Calculation of Probability

Let the number of SUs not facing collision be a and those SUs are denoted as *1* to a. From Eq. 6.11 and 6.16 it can be inferred that,

- each of a SUs can choose same rank with probability q
- each of a SUs can choose 2^{nd} best rank with probability *1-q*
- each of u-a SUs can choose same rank with probability p
- each of u-a SUs can choose 2^{nd} best rank with probability *1-p*

Let the probability of the SUs to have an orthogonal configuration be $\mathbf{P}[OC]$.

$$\begin{aligned} \mathbf{P}[OC] = \mathbf{P}[OC|\text{SUs faced collsion in previous time slot}]* \\ \mathbf{P}[OC|\text{SUs didnot face collsion in previous time slot}] \end{aligned} \tag{6.17}$$

Calculation of $\mathbf{P}[OC|$SUs didnot face collsion in previous time slot$]$

Considering the worst case situation, where the second best rank of all the SUs are same. If more than one SU selects the second best rank, there can not be a orthogonal state.
$\mathbf{P}[OC|\text{SUs didnot face collsion in previous time slot}] =$
$\Big(\mathbf{P}[\text{one SU chooses } 2^{nd} \text{ best rank}] + \mathbf{P}[\text{no SU chooses } 2^{nd} \text{ best rank}]\Big) *$
$\left(\frac{\text{no of orthogonal configuration}}{\text{total no of configuration}}\right)$
$= \Big((\prod_{i=1}^{a-1} q_i) * (1 - q_a) + (\prod_{i=1}^{a} q_i)\Big) * 1$

$$= \prod_{i=1}^{a-1} q_i \tag{6.18}$$

Calculation of $\mathbf{P}[OC|$SUs faced collsion in previous time slot$]$

The number of SU who faced collision are *u-a*. Let there are M blocks of n_m SUs each. Hence, $\sum_{m=1}^{M} n_m = U - a$. It is assumed that all the SU in a block has chosen the same rank.

$\mathbf{P}[OC|\text{SUs faced collsion in previous time slot}] =$
$\Big(\mathbf{P}[\text{one SU chooses the same rank}]+\mathbf{P}[\text{no SU chooses the same rank}]\Big)*\frac{\text{no of possible orthogonal configuration}}{\text{total no of states system can move to}}$

$$= \left(\prod_{m=1}^{M} \left(\left(\prod_{i=1}^{n_m - 1} (1 - p_i) \right) * p_{n_m} \right) + \prod_{m=1}^{M} \left(\prod_{i=1}^{n_m} (1 - p_i) \right) \right) * \frac{\begin{pmatrix} C - a \\ U - a \end{pmatrix}}{\begin{pmatrix} C + U - 1 \\ U - a \end{pmatrix}}$$

$$= \Big(\prod_{m=1}^{M} \prod_{i=1}^{n_m - 1} (1 - p_i) \Big) * \frac{\begin{pmatrix} C - a \\ U - a \end{pmatrix}}{\begin{pmatrix} C + U - 1 \\ U - a \end{pmatrix}} \tag{6.19}$$

So, from equation 6.17 to 6.19

$$\mathbf{P}[OC] = \left(\prod_{i=1}^{a-1} q_i \right) * \left(\prod_{m=1}^{M} \prod_{i=1}^{n_m - 1} (1 - p_i) \right) * \frac{\begin{pmatrix} C - a \\ U - a \end{pmatrix}}{\begin{pmatrix} C + U - 1 \\ U - a \end{pmatrix}}$$

$$\mathbf{P}[OC] = \underbrace{\Big(\prod_{i=1}^{a-1} q_i \Big) * \Big(\prod_{m=1}^{M} \prod_{i=1}^{n_m - 1} (1 - p_i) \Big)}_{\text{I}} * \underbrace{\frac{(C - a)!(C + a - 1)!}{(C - U)!(C + U - 1)!}}_{\text{J}} \tag{6.20}$$

6.4 Comparision of Probabilities

From equation 6.13 and 6.16, it can be seen that term I in Eq. 6.20 tends to 1.

6.4.1 Comparing with Random Rank Selection

Now, comparing term J in Eq. 6.20 with p_{random} Eq. 6.2, we can show that J is greater than p_{random}.

'a' can vary from 0 to U and *'U'* can vary from 1 to $C - 1$.

- For $a = 0$ and $U = 1$
  $\mathbf{P}[OC] = \frac{C!(C-1)!}{(C-1)!C!} = 1$
  Comparing it with Eq. 6.2, $\mathbf{P}[OC] \geq \frac{1!(C-1)!}{C!}$
- For $a = 0$ and $U = C - 1$
  $\mathbf{P}[OC] = \frac{C!(C-1)!}{(2C-2)!}$
  Comparing it with Eq. 6.2, $\mathbf{P}[OC] \geq \frac{(C-1)!(C-1)!}{(2C-2)!}$
- For $a = U - 1$

$\mathbf{P}[OC] = \frac{C-U+1}{C+U-1}$
Hence, $\mathbf{P}[OC] \geq \frac{U!(C-1)!}{(C+U-1)!}$.

This proves that the proposed algorithm attains orthogonality with a higher probability compared to the case where the rank is randomly selected.

6.4.2 Comparing with ρ_{rand}

Now, comparing $\frac{(C-a)!(C-1)!}{(C-U)!(C+U-a-1)!}$ with Eq. 6.1, we can show that $\frac{(C-a)!(C-1)!}{(C-U)!(C+U-a-1)!}$ is converges faster than $\frac{U!(U-1)!}{(2U-1)!}$

'a' can vary from 0 to U and *'U'* can vary from 1 to $C-1$

- For $a = 0, U = 1$
  $P_{\rho_{rand}} = 1$ and $\mathbf{P}[OC] = 1$
- For $a = 0, U = C-1$
  $\mathbf{P}[OC] = \frac{C!(C-1)!}{(2C-2)!}$
  $P_{\rho_{rand}} = \frac{(C-1)!(C-2)!}{(2C-3)!}$
  $\mathbf{P}[OC] \geq \frac{(C-1)!(C-2)!}{(2C-3)!}$
- For $a = U-1$
  $\mathbf{P}[OC] = \frac{C-U+1}{C+U-1}$
  Hence, $\mathbf{P}[OC] \geq \frac{U!(U-1)!}{(2U-1)!}$.

This proves that the proposed algorithm attains orthogonality faster compared to ρ_{rand}.

Chapter 7

Conclusion and Future Work

In the thesis, we have proposed two rank based DMPs for a distributive network for both synchronized as well as varying number of SUs. The SUs are completely unaware of each other and hence there is no communication among them. The DMPs takes into consideration that the number of SUs is unknown to the SUs. The setback caused by the lack of knowledge regarding the number of SU is approached to tackle through two methods. In RLUR,the DMP motivates each SU to always opt for a better rank by giving better rank higher reward and avoiding collision, trusting that over the period of time, the SUs will settle into the best subbands orthogonally. A credibility factor is also introduced in the algorithm which ensures that whenever the acquired data seems to give unreliable results, it is discarded. In SERL, however each SU estimates the number of SUs in the system at some fixed interval for the learning period after which it fixes the number of SUs and settles in to optimum subband. Even in case of a dynamic system, SERL performs better than RLUR.

Also, unlike the state-of-art DMPs, ρ_{rand} [9] and musical chair [24], we have considered that the SUs must learn the rank for which it encounters the least collision and obtains the best throughput. Also, unlike the existing DMPs, BUCB has been used instead of UCB for calculating the statistics of the subband and the rank. The experiment for the proposed DMP was simulated for 10 subbands and, 4 and 6 number of SUs. The results were compared with musical chair [24] and it was seen that the proposed DMP out performs the existing DMPs in terms of regret, collision and switching cost. Also, we have shown that the proposed DMP converges relatively faster as compared to both the ρ_{rand} [9] and musical chair [24].

Future works involve extension of DMPs for various practical scenarios such as delayed response, in which case, the SUs get cumulative collision information in a certain interval of time. Also, the performance of the DMPs can be analyzed for the case of unequal reward of subband where the reward of the subband is dependent on the quality of the subband instead of the simple successful-unsuccessful transmission model used for the thesis. Another practical aspect that can be incorporated is that of time varying reward which takes into consideration that the quality of the subband is not fixed and depends on external factors such as environmental condition. Hence can lead to change in the reward obtained from the subband. The proposed DMPs can be extended for D2D communication and RF energy harvesting. Also, the algorithms can be validated using real radio signals.

Bibliography

[1] J. Chapin and W. Lehr, “Mobile Broadband Growth, Spectrum Scarcity and Sustainable Competition,” *TPRC*, Sept. 2011.

[2] Doppler, Klaus and Rinne, Mika and Wijting, Carl and Ribeiro, Cássio B and Hugl, Klaus, “Device-to-device communication as an underlay to LTE-advanced networks,” in *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42–49, 2019.

[3] Zhang, Haijun and Chu, Xiaoli and Guo, Weisi and Wang, Siyi, “Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum,” in *IEEE Communication Magazine*, vol. 53, no. 3, pp. 158–164,2015.

[4] X. Hong, J. Wang, C-X. Wang and J. Shi, “Cognitive Radio in 5G: A Perspective on Energy-Spectral Efficiency Trade-off,” in *IEEE Communication Magazine*, vol. 52, no. 7, July 2014.

[5] A. Asadi, Q. Wang and V. Mancuso, “A Survey on Device-to-Device Communication in Cellular Networks,” *IEEE Communications Surveys and Tutorials*, vol. 16, no. 4, pp. 1801-1819, Nov. 2014.

[6] C. Herranz, V. Osa, J. F. Monserrat, D. Calabuig, N. Cardona and X. Gelabert, “Cognitive Radio Enabling Opportunistic Spectrum Access in LTE-Advanced Femtocells,” in *Proc. IEEE International Conference on Communications (ICC’12)*, pp. 5593-5597, Ottawa, ON, June 2012.

[7] https://spectrumcollaborationchallenge.com

[8] S. K. Sharma, E. Lagunas, S. Chatzinotas and B. Ottersten, “Application of Compressive Sensing in Cognitive Radio Communications: A Survey,” *IEEE Communications Surveys and Tutorials*, vol. 18, no. 3, pp. 1838-1860, Feb. 2016.

[9] A. Anandkumar and N. Michael and A. K. Tang and A. Swami, “Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret,” in *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, April 2011.

[10] K. Liu and Q. Zhao, “Distributed Learning in Multi-Armed Bandit with Multiple Players,” *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.

[11] S. J. Darak, H. Zhang, J. Palicot and C. Moy, "An Efficient Policy for D2D Communications and Energy Harvesting in Cognitive Radios: Go Bayesian!," in 23th *European Signal Processing Conference (EUSIPCO)*, pp. 1236–1240, Nice, France, Aug. 2015.

[12] M. Zandi, M. Dong and A. Grami, "Distributed Stochastic Learning and Adaptation to Primary Traffic for Dynamic Spectrum Access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1675–1688, Mar. 2016.

[13] S. J. Darak, S. Dhabu, C. Moy, H. Zhang, J. Palicot and A. P. Vinod , "Decentralized Spectrum Learning and Access for Heterogeneous Cognitive Radio Networks," *Elsevier Digital Signal Processing*, vol. 37, pp. 13–23, Feb. 2015.

[14] G. Zhang, A. Huang, H. Shan, J. Wang, T. Q. S. Quek and Y. D. Yao, "Design and Analysis of Distributed Hopping-Based Channel Access in Multi-Channel Cognitive Radio Systems with Delay Constraints," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 11, pp. 2026–2038, Nov. 2014.

[15] M. Zandi, M. Dong and A. Grami, "Dynamic Spectrum Access via Channel-Aware Heterogeneous Multi-channel Auction with Distributed Learning," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 5913–5926, Nov. 2015.

[16] D. Kalathil, N. Nayyar and R. Jain, "Decentralized Learning for Multiplayer Multiarmed Bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, April 2014.

[17] Orly Avner and Shie Mannor and Ohad Shamir, "Decoupling exploration and exploitation in multi-armed bandits," in *arXiv preprint arXiv:1205.2874*, 2012.

[18] Kumar, Rohit and Darak, Sumit J and Sharma, Ajay K and Tripathi, Rajiv, "Twostage decision making policy for opportunistic spectrum access and validation on USRP testbed," in *Wireless Networks*, pp. 1–15, 2016.

[19] G. Zhang and A. Huang and H. Shan and J. Wang and T. Q. S. Quek and Y. D. Yao, "Design and Analysis of Distributed Hopping-Based Channel Access in Multi-Channel Cognitive Radio Systems with Delay Constraints," in *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 11,pp. 2026–2038, November 2014.

[20] G. Zhang and A. Huang and J. Wang and H. Shan and T. Q. S. Quek, "Hopping-Based Channel Access in Cognitive Radio Systems," in *Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th*, pp. 1–5,June 2013.

[21] Orly Avner and Shie Mannor, "Learning to coordinate without communication in multi-user multi-armed bandit problems," in *CoRR*, vol. abs/1504.08167, 2015.

[22] R. Kumar, S. J. Darak, A. Yadav and A. K. Sharma and R. K. Tripathi, "Channel Selection for Secondary Users in Decentralized Network of Unknown Size," in *IEEE Communications Letters*, pp. 1–1, 2017.

[23] O. Avner and S. Mannor, "Concurrent Bandit and Cognitive Radio Networks," in *Machine Learning and Knowledge Discovery in Databased*, pp. 66–81, Springer, April 2014.

[24] J. Rosenski, O. Shami and L. Szlak, "Multi-Player Bandits a Musical Chairs Approach," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1–9, New York, USA, 2016.

[25] Kaufmann, Emilie and Cappé, Olivier and Garivier, Aurélien, "On Bayesian upper confidence bounds for bandit problems," in *International Conference on Artificial Intelligence and Statistics*, pp. 592–600,2012.

[26] Auer, Peter and Cesa-Bianchi, Nicolò and Fischer, Paul, "Finite-time Analysis of the Multiarmed Bandit Problem," in *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, May 2003.

[27] Rajeev Agrawal, "Sample Mean Based Index Policies with O(log n) Regret for the Multi-Armed Bandit Problem," in *Advances in Applied Probability*, vol. 27, no. 4,pp. 1054–1078, 1995.

[28] Mushtaq, MT and Khan, MS and Naqvi, MR and Khan, RD and Khan, MA and Koudelka, Otto F, "Cognitive radios and cognitive networks: A short introduction," in *Journal of Basic & Applied Scientific Research*, 2013.

[29] K. Liu and Q. Zhao, "A Restless Bandit Formulation of Opportunistic Access: Indexablity and Index Policy," in *2008 5th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops*,pp. 1–5,June 2008.

[30] Tekin, Cem, "Online Learning in Bandit Problems," *The University of Michigan*,2013.

[31] Cabric, Danijela and Mishra, Shridhar Mubaraq and Brodersen, Robert W, "Implementation issues in spectrum sensing for cognitive radios," in *Signals, systems and computers, 2004. Conference record of the thirty-eighth Asilomar conference on*, vol. 1, pp. 772–776, 2004.

[32] Subhedar, Mansi and Birajdar, Gajanan, "Spectrum sensing techniques in cognitive radio networks: a survey," in *International Journal of Next-Generation Networks*, vol. 3, no. 2,pp. 35–51, 2011.

[33] Quan, Zhi and Cui, Shuguang and Sayed, Ali H and Poor, H Vincent, "Wideband spectrum sensing in cognitive radio networks," in *2008 IEEE International Conference on Communications*,pp. 901–906, 2008.

[34] Kaufmann, Emilie and Korda, Nathaniel and Munos, Rémi , "Thompson sampling: An asymptotically optimal finite-time analysis," in *International Conference on Algorithmic Learning Theory*, pp. 199–213, 2012.

[35] Kaufmann, Emilie and Cappé, Olivier and Garivier, Aurélien, "On the efficiency of Bayesian bandit algorithms from a frequentist point of view," *Neural Information Processing Systems (NIPS)*,2011.

[36] Y. Gai and B. Krishnamachari and R. Jain, "Combinatorial Network Optimization With Unknown Variables: Multi-Armed Bandits With Linear Rewards and Individual Observations," in *IEEE/ACM Transactions on Networking*, vol. 20, no. 5,pp. 1466–1478, October 2012.

[37] Lai, Tze Leung and Robbins, Herbert, "Asymptotically efficient adaptive allocation rules," in *Advances in applied mathematics*, vol. 6, no. 1-44, pp. 4–22, 1985.

[38] Wu, Tsung-Ying and Lin, Kuo-Wei and Huang, Po-Han and Liao, Wanjiun, "A Distributed cooperation strategy in cognitive radio networks," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 2517–2521,2013.

[39] A. Anandkumar, N. Michael, A. Tang and A. Swami, "Distributed Algorithms for Learning and Cognitive Medium Access With Logarithmic Regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, April 2011.

[40] Darak, Sumit Jagdish and Zhang, Honggang and Palicot, Jacques and Moy, Christophe, "Efficient decentralized dynamic spectrum learning and access policy for multi-standard multi-user cognitive radio networks," in *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*, pp. 271–275, 2014.

[41] Darak, Sumit Jagdish and Modi, Navikkumar and Nafkha, Amor and Moy, Christophe, "Spectrum Utilization and Reconfiguration Cost Comparison of Various Decision Making Policies for Opportunistic Spectrum Access Using Real Radio Signals," in *11th EAI International Conference on Cognitive Radio Oriented Wireless Networks, CROWNCOM 2016*, 2016.

[42] Darak, Sumit J and Moy, Christophe and Palicot, Jacques, "Proof-of-Concept System for Opportunistic Spectrum Access in Multi-user Decentralized Networks," in *EAI Endorsed Transactions on Cognitive Communications*, vol. 2, no. 7, pp. 1–10, 2016.