Towards More Reliable Medium Access Control With Data-Driven Spectrum Allocation

by

Umuralp Kaytaz

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Electrical and Electronics Engineering

KOÇ UNIVERSITY

January 4, 2020

Towards More Reliable Medium Access Control With Data-Driven Spectrum Allocation

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Umuralp Kaytaz

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by the final

examining committee have been made.

Committee Members:

Assoc. Prof. Dr. Sinem Coleri

Asst. Prof. Dr. Baris Akgun

Assoc. Prof. Dr. Berk Canberk

Date: ____________________

To all the people that showed support and kindness over the years.

ABSTRACT

Most of the deployed wireless networking technology relies on Wi-Fi communications that depends on IEEE 802.11 standard. IEEE 802.11 consists of a set of protocols for medium access control (MAC) and physical layer (PHY) communications. Distributed Coordination Function (DCF) included in IEEE 802.11 employs a single channel in the MAC layer. However, network traffic load on a single channel leads to high packet collision and performance degradation. Single-Radio Multi-Channel (SRMC) Medium Access Control (MAC) protocols aim to transmit in parallel on distinct channels while reducing the hardware cost. Initially, we consider dynamic multi-channel access problem with independent and stochastically identical channels. We formulate the problem as a Restless Multi-Armed Bandit (RMAB) process and propose a novel load-adaptive frequency hopping method, namely Index-Based Channel Hopping (IBCH) algorithm, with the goal of optimal spectrum allocation and throughput maximization. We classify previous efforts by their spectrum decision mechanisms and rendezvous characteristics as single- and multi-rendezvous protocols. Multi-rendezvous protocols have the capability of supporting simultaneous handshaking on different channels whereas with single-rendezvous protocols only asynchronous channel negotiations are allowed. Later on, we compare performance of IBCH with various protocols from both rendezvous classes at varying transmission range, network size and spectrum availability. Simulation results show that IBCH significantly outperforms other spectrum assignment methods from both rendezvous classes. In the second part of the thesis, we focus on IEEE 802.22 based Wireless Regional Area Networks (WRANs). IEEE 802.22 WRAN standard is aimed at using cognitive radio (CR) to avoid user interference and congestion. Cognitive Radio (CR) is a promising technology for emerging intelligent wireless communication systems due to its effi-

cient utilization of the frequency bands. CR systems enhance temporal and spatial efficiency, therefore, ease spectrum scarcity by exploiting temporary idle periods in the frequency usage, known as spectrum holes. Dynamic Spectrum Access (DSA) improves spectrum utilization by allowing secondary users (SUs) to opportunistically access temporary idle periods in the primary user (PU) channels. Previous studies on utility maximizing spectrum access strategies mostly require complete network state information, therefore, may not be practical. Model-free reinforcement learning (RL) based methods, such as Q-learning, on the other hand, are promising adaptive solutions that do not require complete network information. In the second section, we tackle this research dilemma and propose deep Q-learning originated spectrum access (DQLS) based decentralized and centralized channel selection methods for network utility maximization, namely DEcentralized Spectrum Allocation (DESA) and Centralized Spectrum Allocation (CSA), respectively. Actions that are generated through centralized deep Q-network (DQN) are utilized in CSA whereas the DESA adopts a non-cooperative approach in spectrum decisions. We use extensive simulations to investigate spectrum utilization of our proposed methods for varying primary and secondary network sizes. Our findings demonstrate that proposed schemes outperform model-based RL and traditional approaches, including slotted-Aloha and Whittle index policy, while %87 of optimal channel access is achieved.

ÖZETÇE

Günümüzde kullanılan kablosuz ağ teknolojilerinin çoğu, IEEE 802.11 standardına bağlı Wi-Fi iletişimine dayanır. IEEE 802.11, orta erişim kontrolü (MAC) ve fiziksel katman (PHY) iletişimleri için bir dizi protokolden oluşur. IEEE 802.11'de bulunan Dağıtılmış Koordinasyon İşlevi (DCF), MAC katmanında tek bir kanal kullanır. Bununla birlikte, tek bir kanaldaki ağ trafiği yükü, yüksek paket çarpışmasına ve performans bozulmasına yol açar. Tek Radyo Çok Kanallı (SRMC) Orta Erişim Kontrolü (MAC) protokolleri, donanım maliyetini düşürürken farklı kanallarda paralel olarak iletim yapmayı hedefler. Başlangıçta, bağımsız ve stokastik olarak özdeş kanallarda dinamik çok kanallı erişim problemini ele alıyoruz. Sorunu Huzursuz Çoklu Kumar Makinası (RMAB) süreci olarak formüle ediyoruz ve optimum spektrum tahsisi ve verim maksimizasyonu amacıyla yeni bir yük uyarlamalı frekans atlama yöntemi, yani İndeks Tabanlı Kanal Atlama (IBCH) algoritması öneriyoruz. Önceki çabaları spektrum karar mekanizmaları ve buluşma özellikleriyle tek ve çok buluşma protokolleri olarak sınıflandırıyoruz. Çok buluşmalı protokoller farklı kanallarda eşzamanlı tokalaşmayı destekleme yeteneğine sahipken, tek buluşmalı protokollerle sadece asenkron kanal görüşmelerine izin verilir. Daha sonra, IBCH'nin performansını değişen iletim aralığında, ağ boyutunda ve spektrum kullanılabilirliğinde her iki buluşma sınıfından çeşitli protokollerle karşılaştırıyoruz. Simülasyon sonuçları IBCH'nin her iki buluş sınıfından diğer spektrum atama yöntemlerinden önemli ölçüde daha iyi performans gösterdiğini göstermektedir. Tezin ikinci bölümünde, IEEE 802.22 tabanlı Kablosuz Bölgesel Alan Ağlarına (WRAN'lar) odaklanıyoruz. IEEE 802.22 WRAN standardı, kullanıcı etkileşimini ve tıkanıklığını önlemek için bilişsel radyo (CR) kullanmayı amaçlamaktadır. Bilişsel Radyo (CR), frekans bantlarını verimli kullanması nedeniyle ortaya çıkan akıllı kablosuz iletişim sistemleri için umut verici bir teknolojidir. CR sis-

temleri, zamansal ve uzamsal verimliliği arttırır, bu nedenle spektrum delikleri olarak bilinen frekans kullanımında geçici boşta kalma sürelerinden yararlanarak spektrum kıtlığını azaltır. Dinamik Spektrum Erişimi (DSA), ikincil kullanıcıların (SU) birincil kullanıcı (PU) kanallarındaki geçici boşta kalma sürelerine fırsatçı olarak erişmesine izin vererek spektrum kullanımını geliştirir. Spektrum erişim stratejilerini en üst düzeye çıkarmaya yönelik daha önceki çalışmalar çoğunlukla ağ durumu bilgilerinin tamamını gerektirmektedir, bu nedenle pratik olmayabilir. Öte yandan, Q-öğrenme gibi modelsiz takviye öğrenme (RL) tabanlı yöntemler, tam ağ bilgisi gerektirmeyen uyarlanabilir çözümler vaat ediyor. İki bölümde, bu araştırma ikilemini ele alıyoruz ve ağ fayda maksimizasyonu için sırasıyla Q-öğrenme kaynaklı spektrum erişimi (DQLS) tabanlı merkezi olmayan ve merkezi kanal seçim yöntemlerini, yani DEcentralized Spectrum Atama (DESA) ve Merkezi Spectrum Atama (CSA) öneriyoruz. Merkezi derin Q-ağı (DQN) aracılığıyla üretilen eylemler CSA'da kullanılırken, DESA spektrum kararlarında işbirlikçi olmayan bir yaklaşım benimsemektedir. Değişen birincil ve ikincil ağ boyutlarına yönelik önerilen yöntemlerimizin spektrum kullanımını araştırmak için kapsamlı simülasyonlar kullanıyoruz. Bulgularımız önerilen planların modele dayalı RL ve oluklu Aloha ve Whittle indeks politikası dahil geleneksel yaklaşımlardan daha iyi performans gösterdiğini ve en iyi kanal erişiminin % 87'sinin elde edildiğini göstermektedir.

ACKNOWLEDGMENTS

First of all, I would like to thank to Assoc. Prof. Dr. Sinem Coleri for giving me a chance to work with her. Secondly, I would like to thank Asst. Prof. Dr. Baris Akgun for taking my enthusiasm for artificial intelligence and machine learning seriously and supporting my work. I am very thankful to my father Bulent Kaytaz for all the financial and moral support. I am very grateful to my mother Nesrin Subasi and brother Kaan Gokalp Kaytaz for their moral support and coolness. I would like to thank Seyhan Ucar, a great researcher and publication buddy, for helping and sharing wisdom. I am very grateful for sharing this journey with my friends and brothers Rıza Arda Kırmızıoğlu, Ozan Kara, Yashar (baskan), Gokhan Onder Aksu, and Omer Faruk Altundal. We have a lot of great moments and stories together. It was also tremendous experience and fun to have Cem, Şena, Ahmad, Beyza, Koray, Bugra, Serkan, Gurkan, Kemal, Tugtekin, Meltem, Fatih, Ehsan, Berker, Selin, Doğancan, Ogun, Merve, Tugce and many more as my friends, brothers, sisters and gym/football buddies throughout this journey. I would like to thank my grandmother Ayla Kaytaz and grandfather Ismail Kazım Kaytaz, whom I still miss to this day, may your souls rest in peace. Finally, special thanks to all of the past and current ENG 142 members for their kindness and help all along the way. I could not come this far without all these people supporting me, that is for sure. Rock n Roll.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Wireless communications is a fundamental part of our daily life in today's world. High use of smartphones, tablet PCs, drones and other electronic devices creates an ever increasing traffic load on wireless networks. Satisfying the need for location-independent communication and providing robust mobile services has become a implacable challenge for the global information communication technology (ICT) industry. Although, many protocols incorporating processes, requirements and constraints have been defined for network-enabled devices, reliable transport of data is one of the primary concerns to achieve a high degree of efficiency in ever-growing wireless communications. Medium access control (MAC) protocols play a central role in wireless networks by governing decisions on the optimal sensing, transmission times and coordination among network users for spectrum decisions in the data link layer. This thesis is concerned about analyzing and proposing medium access schemes for throughput maximization and reliable communication purposes in Wi-Fi and Wireless Regional Area Networks (WRANs) considering various network topologies.

1.1 Multi-Channel Communications in Wi-Fi Networks

Wi-Fi networks are based on IEEE 802.11 family of standards and used for Wireless Local Area Networking (WLAN). Greater part of ICT depends on the seamless interaction of devices using radio bands defined by the IEEE 802.11. Distributed Coordination Function (DCF) is the essential MAC scheme in IEEE 802.11 networks that employ carrier-sense multiple access with collision avoidance (CSMA/CA) with binary exponential back-off algorithm. Traditional spectrum decisions by the DCF exploit a single channel in the frequency spectrum. Proposed MAC schemes in the literature aim to enhance functionality in multi-domain for enhancing network throughput and causing less mutual interference. Using multiple transceivers for multi-channel MAC is shown be a robust approach [1]. However, providing multiple radios to each node increases the hardware cost. Considering the scalability of Wi-Fi networks and investigating reliable communication with minimum cost we only examine multi-channel MAC protocols with single transceiver operation.

1.2 Cognitive Radio Networks

Communications in WRANs are governed by IEEE 802.22 standard that use frequencies allocated to a broadcasting service referred to as white spaces. IEEE 802.22 standard define WRAN specifications for using white spaces in the television (TV) frequency spectrum. Nodes in WRANs are generally digital TV (DTV) base stations (BS) in rural areas that allocate resource blocks available by the orthogonal frequency-division multiple access (OFDMA) digital modulation scheme. Communication in WRANs relies on cognitive radio (CR) technology for efficient utilization of multiple frequency opportunities. CR allows configuration of transceivers based on best spectral opportunities available in the allocated bandwidth. CR networks ease spectrum scarcity by exploiting temporary idle periods in the frequency usage, known as spectrum holes. Cognitive networks generally consist of a primary network and a secondary network that have multiple tiers. Primary BSs are users that have exclusive rights to a certain spectrum bands. Secondary BSs allocate primary channels according to the interference that will be caused by their transmission to secondary receivers.

1.3 Main Contributions

Main contributions of this thesis are listed as follows.

- Rendezvous and spectrum access based classification of Single-Radio Multi-

Channel (SRMC) Wi-Fi MAC protocols.

- We formulate the dynamic multi-channel access problem in SRMC domain as a Restless Multi-Armed Bandit (RMAB) process and propose a novel multi-randezvous multi-channel MAC approach, namely Index-Based Channel Hopping (IBCH).

- We analyze impact of transmission range, network size and spectrum availability on aggregated throughput for the proposed and examined SRMC MAC schemes with different operational characteristics.

- Proposal of deep Q-learning originated spectrum access (DQLS) based decentralized and centralized channel selection methods for network utility maximization, namely DEcentralized Spectrum Allocation (DESA) and Centralized Spectrum Allocation (CSA).

- Performance comparison of proposed DSA methods with model-based reinforcement learning and traditional approaches, including slotted-Aloha and Whittle index policy.

1.4 Thesis Outline

This thesis is organized as follows.

Chapter 2: In this chapter, we have provide a classification of IEEE 802.11 based SRMC MAC protocols based on their spectrum decision mechanisms and rendezvous characteristics. Later on, we consider dynamic multichannel access scenario in SRMC domain with independent and stochastically identical channels and formulate the problem as a Restless Multi-Armed Bandit (RMAB) process. Based on this formulation, we propose a novel multi-randezvous frequency hopping method, namely Index-Based Channel Hopping (IBCH) algorithm, for throughput maximization with optimal spectrum allocation. Furthermore, we compare IBCH with various protocols from both rendezvous classes at varying transmission range, network size and spectrum availability. Simulation results show that IBCH results in higher throughput values compared to other protocols with its optimal policy for channel selection and frequency hopping behaviour.

Chapter 3: In this chapter, we present multi-agent deep reinforcement learning based spectrum selection with wideband sensing capability for multi-user utility maximization during DSA. Previous studies on utility maximizing spectrum access strategies mostly require complete network state information. However, model-free reinforcement learning (RL) based methods, such as Q-learning,are promising adaptive solutions that do not require complete network information. This enables practical solutions for networks with no prior information. We propose deep Q-learning originated spectrum access (DQLS) based decentralized and centralized channel selection methods for network utility maximization, namely DEcentralized Spectrum Allocation (DESA) and Centralized Spectrum Allocation (CSA), respectively. We use extensive simulations to investigate spectrum utilization of our proposed methods for varying primary and secondary network sizes. Our findings demonstrate that proposed schemes outperform model-based RL and traditional approaches, including slotted-Aloha and Whittle index policy, while %87 of optimal channel access is achieved.

Chapter 2

INDEX-BASED CHANNEL HOPPING FOR MULTI-RENDEZVOUS MULTI-CHANNEL MAC

2.1 Introduction

Most of the deployed wireless networking technology relies on IEEE 802.11 standard. Distributed Coordination Function (DCF) included in IEEE 802.11 employs a single channel in the MAC layer. However, network traffic load on a single channel leads to high packet collision and performance degradation. Exploiting the multi-channel structure with channel switching capability is a common approach for enabling simultaneous transmissions on distinct channels. Dynamic multi-channel access schemes aim to increase performance and enhance network utilization by resulting in less mutual interference, higher spatial reuse and enhanced network throughput.

Many studies have been proposed to improve functionality in the multi-channel domain [1]. These approaches can be categorized based on the number of receivers and/or handshaking methods. Although multi-transceiver enabled multi-channel MAC protocols might be more robust, providing multiple radios to each node increases the hardware cost. To investigate reliable communication with minimum cost, we only examine multi-channel MAC protocols with single transceiver operation. The classification of Single-Radio Multi-Channel (SRMC) MAC protocols is demonstrated in Table 2.1.

Single transceiver operation based on the control channel depends on maintain alternating exchange phases between the control channel (for handshaking) and data channels (for transmissions). Due to non-simultaneous handshaking resulting from a single transceiver on a single control channel both Dedicated Control Channel (DCC)

Single-Radio Multi-Channel MAC Protocols							
Single-Rendezvous				Multi-Rendezvous			
Dedicated Control Channel		Common Hopping		Frequency Hopping			
Fixed phase length	Variable phase length	Synchronous hopping	Asynchronous hopping	Pseudo-random generator based hopping	Cyclic quorum based hopping	Latin square based hopping	Whittle index based hopping
[2], [3]	[4], [5]	[6], [7]	[8]	[9], [10]	[11]	[12]	**IBCH**

Table 2.1: Classification of single-transceiver multi-channel MAC protocols

protocols [13] and split phase protocols [2, 4] perform single-rendezvous functionality in SRMC networks. Therefore, we classify these protocols as DCC-based spectrum decision schemes. Another approach that also has single-rendezvous capability is the common hopping. In common hopping, nodes are deployed with a common channel selection strategy that requires them to periodically jump to the next channel in the frequency spectrum based on the predetermined hopping sequence [6]. Compared to DCC-based schemes, common hopping protocols utilize all of the channels for data transmissions. Simultaneous frequency hopping is performed if the global clock synchronization is available at nodes. Therefore, these networks can perform synchronous common hopping. Otherwise, resulting hopping behavior becomes asynchronous [8].

Schemes employed by the frequency hopping protocols are generally designed to enable multi-rendezvous functionality. However, their approaches to determining the hopping sequence may vary. Nodes running the pseudo-random generator based protocols employ a common pseudo-random function in all nodes for determining the hopping sequence or awake time period [9, 10]. On the other hand, frequency hopping sequence of the nodes can also be created by merging non-intersecting difference sets as a cyclic quorum. Rotation closure property of difference sets enables any two nodes to communicate with a nonzero probability in a given cycle without the need of global synchronization or a dedicated control channel [11]. Another approach to designing a frequency hopping scheme is the utilization of latin squares for ensuring concurrent spectrum decisions. However, this approach guarantees overlapping transmission slots only when global synchronization is available [12].

Previous efforts have shown that multi-rendezvous protocols can outperform single-rendezvous schemes [1]. It has been also shown that latin square based frequency hopping has superior performance compare to other multi-randezvous approaches [12]. However, optimal spectrum allocation and a detailed comparison of randezvous classes is yet to be provided.

The original contributions of this section is twofold. First, we provide a detailed classification and performance comparison of SRMC MAC protocols based on their rendezvous characteristics and spectrum decision approach. Second, we propose novel performance optimal multi-randezvouz multi-channel spectrum decision scheme based on Restless Multi-Armed Bandit Formulation (RMAB), namely *Index-Based Channel Hopping* (IBCH).

2.2 System Model

Network architecture that has been chosen for dynamic multi-channel access problem is based on IEEE 802.11 DCF as the medium access scheme. Therefore, operational principles of the protocols are compatible with the IEEE 802.11 standard. System implementation requires nodes equipped with a single radio and a physical medium with multi-channel availability. Physical layer assumed to accommodate up to 13 channels in the frequency spectrum. Adopted network topology is non-hierarchical in nature and consists of uniformly deployed nodes that are operationally independent and have the same transmission range. All nodes are capable of acting as a sender or receiver. Horizontal and vertical distance between adjacent nodes is constant throughout the topology. Availability of clock synchronization is assumed among all devices for concurrent spectrum decisions.

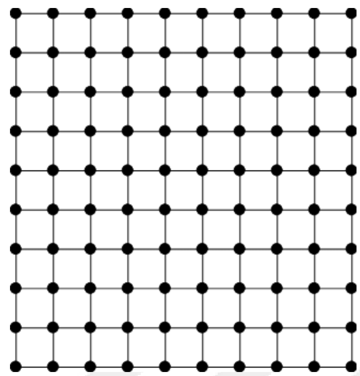

Figure 2.1: Adopted 2D Mesh Network Topology

2.3 Background

2.3.1 Dynamic Multichannel Access

Consider N independent Markov chains representing N frequency sub-bands under Gilbert-Elliot channel model [14]. Each chain i consist of two states *1 (good)* and *0 (bad)* with transition probability matrix $\mathbf{P}_i = \{p_{j,k}^{(i)}\}_{j,k\in\{0,1\}}$. Each channel ch_i is selected for a maximum available throughput Th_i $(i = 1, ..., N)$. At the beginning of each time slot t, K $(1 \leq K \leq N)$ out of N channels are selected as the *default channel* set, $\mathcal{K}(t)$, for transmission and probing purposes. Considering state-space $S_i(t) \in [0, 1]$ for each channel ch_i, $S_i(t) = 1$ represents that Th_i unit of reward/throughput obtained upon transmission. Otherwise, $S_i(t) = 0$, no throughput is obtained. Total expected reward/throughput obtained at a given time slot t from channel set $\mathcal{K}(t)$, can be calculated using

$$R_{\mathcal{K}(t)}(t) = \sum_{i\in\mathcal{K}(t)} S_i(t)Th_i \tag{2.1}$$

In order to obtain maximum amount of throughput/aggregated reward from transmissions, a policy, which determines optimal way of selecting the transmission channels at each time slot, have to be found.

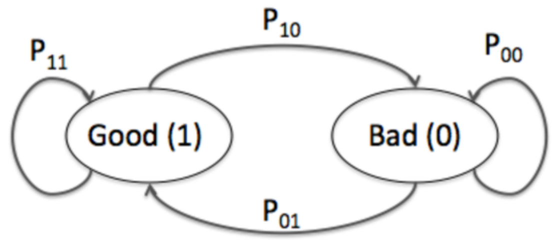

Figure 2.2: Gilbert-Elliot channel model [21]

2.3.2 Restless Multiarmed Bandit (RMAB) Formulation

Multi-Armed Bandit (MAB) problems are stochastic scheduling formulations that have been a part of reinforcement learning and probability theory literature for many years [15]. These set of problems aim to provide solutions for utility maximization while allocating pre-determined scarce resources in the presence of competing choices. MAB consists of a player with a full knowledge of the environment (states of arms) that activates *one* of N markovian arms and receives a corresponding reward. While the activated arm chances its state, other states of passive arms are frozen. MAB problems aim to find the optimal policy for activation order in order to obtain maximum expected reward over an unlimited time, *a.k.a infinite horizon.* RMAB problems are generalizations of MAB where $S_i(t) \in \{0, 1\}$ for any K such that $1 \leq K \leq N$ and $\sum_{i=1}^{N} S_i(t) = K$ [16]. Similarly, dynamic multichannel access scenario can be modeled as a RMAB problem, where K channels are assigned for transmission source and destination (S-D) pairs out of N available frequencies.

Considering channel states are not observable without spectrum sensing, each node has to infer the state of the channel based on its previous experiences. Conditional probability $w_i(t)$ of a channel ch_i is in *good* state $S_i(t) = 1$ has been shown to be sufficient statistics for optimal decision making [17]. Therefore, at each time step t a belief vector/information state $\Omega(t) = \{w_1(t), w_2(t), ..., w_N(t)\}$ can be used as the sufficient statistics in the optimal policy calculation. One-step belief update using results of spectrum decision at time t can be denoted as follows;

$$w_i(t+1) = \begin{cases} p_{11}^i & i \in \mathcal{K}(t), S_i(t) = 1 \\ p_{01}^i & i \in \mathcal{K}(t), S_i(t) = 0 \\ w_i(t)p_{11}^i + (1 - w_i(t))p_{11}^i & i \notin \mathcal{K}(t) \end{cases} \tag{2.2}$$

Using policy $\pi(\Omega(t))$ notation for transmission/default channel set $\mathcal{K}(t)$ selection under information state $\Omega(t)$, *expected total discounted reward* over infinite horizon can be calculated using

$$\mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_{\pi(\Omega(t))}(t) \middle| \Omega(1)\right] \tag{2.3}$$

where $0 \le \gamma \le 1$ is the discount factor for decreasing contribution of future rewards in the expected reward calculation. RMAB with discounted reward criterion can be denoted as a tuple $(\Omega(1), \{\mathbf{P}_i\}_{i=1}^N, \{Th\}_{i=1}^N, \gamma)$.

2.3.3 Index Policy

Index policy for dynamic spectrum access assigns index to each of the available channel states at any given time slot. Spectrum decisions are regulated based on the indices provided by the index policy. Channel indices for a given time slot are ordered based on the expected discounted reward/throughput that calculated using outcome of previously selected transmission channels. Gittin index policy has been shown to be optimal for MAB problems, where a priority index is assigned for determining the order of arm activation [18]. Whittle index policy is the generalization of Gittin index that offers an optimal solution to *indexable* RMAB problems with the relaxed constraint that average number of activated arms over the infinite horizon equals to K. Although, not all RAMBS are *indexable*, Whittle index offers an optimal solution to indexable RAMBs such as our adopted dynamic multichannel access scenario with semiuniversal structure such that $p_{11} \ge p_{10}$ for any given N stochastically identical channels and K selections [37].

Strongly decomposable index policies are used for arms and/or channels whose index solely depends on its characteristics and not on other arms and/or channels. Thus, decoupling reduces complexity of N-dimensional spectrum assignment into N independent problems of size 1-D. As an example of such policy, myopic policy ignores future rewards and focuses on the expected immediate reward. Myopic action and/or channel selection $\mathcal{A}(t)$ at time t can be calculated with following equation, where $Th(i)$ represents expected throughput of channel ch_i and $w_i(t)Th(i)$ is the myopic index

$$\hat{\mathcal{A}}(t) = \arg\max_{\mathcal{A}(t)} \sum_{i \in \mathcal{A}(t)} w_i(t)Th_i \tag{2.4}$$

2.4 Approach

Next, we propose our *Index-Based Channel Hopping* (IBCH) protocol based on RMAB formulation and corresponding Whittle index policy. Whittle index policy becomes identical to the myopic policy and has a queue-based structure for stochastically identical positively correlated channels ($p_{11}^i \geq p_{01}^i$) as shown in Fig. 2.3. Resulting indexing strategy has been shown to be the optimal policy for N and K such that $1 \leq K \leq N$ [37]. Similar to [12], at each time slot IBCH separates the available spectrum into *default channels* and *switching channels*. Nodes assigned with a *default channel*, default nodes, stay idle and wait for incoming transmissions on their *default channels*. On the other hand, switching nodes, assigned with a *switching channel*, switch to the *default channel* of their intended destinations for transmission.

Algorithm 1 is executed for implementing Whittle-index policy based frequency hopping behaviour. Queue-based transmission is triggered upon obtaining node, channel and set limit information. Channel queue ($Queue$) is updated based on the Whittle index policy provided in [37]. Initially, each node is assigned with a *default channel* based on its id (Lines 1-2). Corresponding channel number of each node is computed

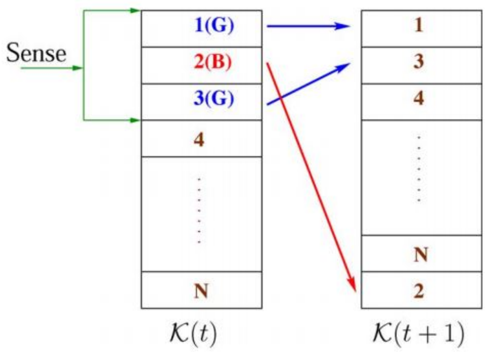

Figure 2.3: Whittle index for $p_{11} >= p_{01}$ [37]

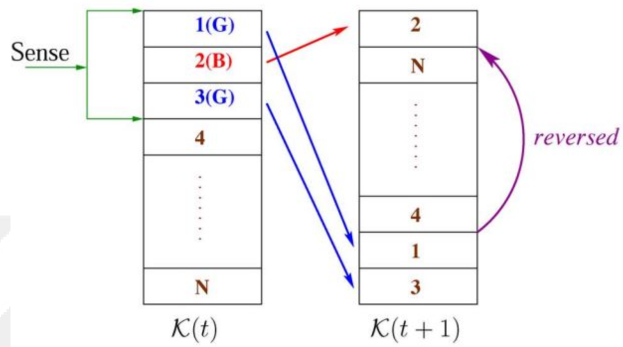

Figure 2.4: Whittle index for $p_{11} < p_{01}$ [37]

Algorithm 1: Index-Based Channel Hopping (IBCH)

Input: $\mathcal{C}$ ←Channels, $\mathcal{N}$ ← Nodes, K ← set limit
Data: *Queue*: channel queue, $\mathcal{T}$: cycle length
for *each node* $n_i \in \mathcal{N} \mid i \in \{1, ..., N\}$ **do**
  Assign $n_i.channel \leftarrow i(mod)|\mathcal{C}|$
Initialize *Queue* ←random_order($\mathcal{C}$)
Compute *default_channel_set* ← *SAT*(*Queue*)
for *each time_slot* $t_j, j \in \{2, .., \mathcal{T}\}$ **do**
  for *each channel* $ch_k \in$ *default_channel_set* **do**
    Get Tx_k^{j-1} ← transmission over ch_k at t_{j-1}
    if Tx_k^{j-1} == *Success* **then**
      Update *Queue.insert_front*(ch_k)
    else
      Update *Queue.insert_rear*(ch_k)
  Compute *default_channel_set* ← *SAT*(*Queue*)

by the formula $i(mod)|\mathcal{C}|$, where i is node id and $|\mathcal{C}|$ is the cardinality of the channel set. Assuming nodes are deployed sequentially based on their ids, this approach guarantees that neighbouring nodes do not have non-overlapping *default channels*. Channel assignment is followed by queue initialization based on random ordering for regulating transmissions at the first time slot t_1 (Line 3). Using *Switch and Transmit (SAT)* algorithm with the initialized queue structure *Queue*, nodes are separated into *default* and *switching* nodes. *Default nodes*, whose *default channels* ids are among first K elements of the queue, stay idle and wait for incoming transmissions from *switching nodes* with *switching channel* ids. Upon completion of the transmissions at the first time slot, a set for representing the *default nodes* channel ids (*default_ channel_ set*) is returned (Line 4). Rest of the procedure is iterated for remaining time slots of a given *time cycle* with length $\mathcal{T}$ (Lines 5-12). While iterating over each channel ch_k, initially, result of transmission over ch_k at the previous time slot is obtained Tx_k^{j-1} (Line 7). Later on, based on the result of the transmissions, channel ids in the queue are reordered for organizing transmission channels at the upcoming slot (Lines 7-11). After reordering, first K channel ids at the beginning of the queue become the *default channels* for the upcoming time slot [37].

Algorithm 2: Switch-And-Transmit (SAT)

Input: *Queue*: channel queue
Output: *default_ channel_ set*
Data: $\mathcal{N} \leftarrow$ Nodes, $K \leftarrow$ set limit
Initialize $default\_channel\_set \leftarrow Queue[1:K]$
Initialize $switching\_set \leftarrow Queue[K+1:N]$
for *each channel* $ch_S \in switching\_set$ **do**
    Compute $set^{switch} \leftarrow get\_ch\_nodes(\mathcal{N}, ch_S)$
    for *each source_ node* $n_i \in set^{switch}$ **do**
        Compute $n_j \leftarrow find\_destination\_node(n_i)$
        Get $ch_D \leftarrow n_j.channel$
        Set $n_i.channel \leftarrow ch_D$
        Compute $n_i.transmit\_to(n_j)$

Algorithm 2 is run for queue-based frequency switching and S-D pairing. Based on the updated queue structure ($Queue$) and set limit (K) provided by Algorithm 1 procedure, channels are separated into *default* and *switching* channels (Lines 1-2). Later on, for each of the *switching channel* ids corresponding nodes are computed set^{switch} (Line 3-4). Procedure continues by computing destination nodes and initiating transmission for each of the nodes $n_i \in set^{switch}$ corresponding to the switching node ids ch_s (Lines 5-9).

2.5 Protocols

In the interest of an elaborate examination, we have selected synchronous and asynchronous common hopping schemes, CHMA and DSMMAC, as the single-rendezvous protocols in our analysis. On the other hand, LACH has been chosen as the multi-rendezvous approach for outperforming cyclic quorum and pseudo-random generator based frequency hopping schemes [12]. A comparison of protocol strategies for channel selection is presented in Table 2.2.

Channel Selection Strategies		
CHMA	LACH	DSMMAC
$c_t = s_t \, mod \, m$	$R_i = i \, mod \, n$ $SB_i = (i + \lfloor i/n \rfloor) \, mod \, n$ $IDS_i = (SB_i + R_i) \, mod \, n$ $IDC_i = SB_i \, mod \, m$	$\Phi = (n, k, \lambda)$
3 Channels $10 \, mod \, 3 \underset{C}{\rightarrow} 1$ **5 Channels** $10 \, mod \, 5 \underset{C}{\rightarrow} 0$	**3 Channels** $10 \, mod \, 25 \underset{R_i}{\rightarrow} 10$ $(10 + \lfloor 10/25 \rfloor) \, mod \, 25 \underset{SB_i}{\longrightarrow} 10$ $(10 + 10) \, mod \, 25 \underset{IDS_i}{\longrightarrow} 20$ $10 \, mod \, 3 \underset{IDC_i}{\longrightarrow} 1$ **5 Channels** $10 \, mod \, 25 \underset{R_i}{\rightarrow} 10$ $(10 + \lfloor 10/25 \rfloor) \, mod \, 25 \underset{SB_i}{\longrightarrow} 10$ $(10 + 10) \, mod \, 25 \underset{IDS_i}{\longrightarrow} 20$ $10 \, mod \, 5 \underset{IDC_i}{\longrightarrow} 0$	**3 Channels** (25, 8,1) ={1,2,4,8,12,13,21,23} (25,8,1) ={3,10,11,15,16,19,20,22} (25,9,2) ={ 5,6,7,9,14,17,18,24,25} **5 Channels** (25, 8,1) ={1,2,4,8,12,13,21,25} (25,8,1) ={3,10,11,15,17,19,20,22} (25,7,1) ={6,7,9,14,16,18,24} {5},{23}

Table 2.2: Channel selection strategies

2.5.1 Channel-Hopping Multiple Access (CHMA)

Elimination of hidden-terminal interference requires a consensus on the timing and order of channel switching between frequency-hopping nodes in the network. CHMA exploits this fact by assuming a common frequency hopping sequence [6]. Transmission request is achieved by sending a Request-to-Send (RTS) message to intended destination on the common channel that all nodes listen to at any given time. In case of successful handshake completion, S-D pair stops their hopping sequence and continues data transmission on that channel. However, remaining idle nodes pursue hopping until an idle channel is detected. Channel of the hopping sequence at time t can be calculated with $c_t = s_t \, mod \, m$, where m represents the number of channels and s_t is the slot index at time t. After the completion of the data transmission, S-D pair calculates the channel and joins rest of the nodes at the common hopping sequence. There are several drawbacks of this approach. First of all, nodes that are engaged in a transmission are unaware of the current status in the network such that they have no prior information of which devices and channels are occupied. On the other hand, hopping behavior is frequently repeated and nodes may switch to channels that are already occupied by a transmission.

2.5.2 Difference-Set-Based Multi-Channel MAC (DSMMAC)

Difference sets have been exploited in the literature for designing power-saving sleep scheduling algorithms, in which main objective is to minimize nodes wake-up duty cycle [11]. Using difference sets in MAC design is an effective approach for maximizing number of overlapping slots between S-D pairs. DSMMAC aims to maximize rendezvous probability of the nodes by combining mutually exclusive difference sets and exploiting the rotation closure property for ensuring nodes to meet at each cycle with a certain probability regardless of global synchronization [8]. Design of the hopping sequence is accomplished by combining non-intersecting difference sets that have high rendezvous probability. Later on, each difference set is assigned to a distinct channel where elements of the difference set correspond to the time slots of a given cycle. Selected difference sets determine the spectrum decision of the nodes at any given time slot. If the combined difference sets can not span a whole cycle, remaining time slots are allocated with randomly selected channels. A difference set of length k with rendezvous probability λ calculated from a cycle of length n can be represented as $\Phi = (n, k, \lambda)$. For each $r \in \{1, ..., n-1\}$, there exists λ ordered pairs (d_i, d_j) such that $d_i, d_j \in \Phi$ and $r = (d_i - d_j) \, mod \, n$ [8]. Similar to CHMA, S-D pair that successfully completes the handshake procedure starts data transmission over the channel and stops its hopping sequence. Nodes return to the difference set based common hopping sequence only after the transmission is completed.

2.5.3 Load-Aware Channel Hopping (LACH)

Initial time and channel allocation mechanism of the LACH protocol utilizes latin squares for load sharing. Latin squares have been adopted as a MAC protocol or a network coding scheme [?]. Using latin squares for channel hopping scheduling design creates multi-rendezvous functionality. After the network initialization phase, where nodes determine their initial channel hopping sequence, nodes periodically update their hopping sequence in order to adapt to the changing network traffic. For the co-ordination of time and frequency domain allocations, nodes get assigned with default and switching slots based on the latin square that is used for the implementation. Default slots are time periods where nodes are not allowed to transmit and obliged to listen for incoming transmission requests from their neighbors. On the other hand, nodes that are willing to transmit, switch to the channel of their intended receivers during the switching slots. Initially, *row* (R_i) and *symbol* (SB_i) values are calculated for each node i. Using these two parameters, $initial\ default\ slot$ (IDS_i) and $initial\ default\ channel$ (IDC_i) values are determined for the implementation [12]. This approach ensures nodes with successive IDs to have unique default slot and default channel assignments. Continual adaptation to traffic load is maintained by changing the number of extended default slots based on the utilization function provided in the related paper [12].

2.6 Performance Evaluation

2.6.1 Simulation Setup

The goal of the simulations is to evaluate the performance of IBCH, CHMA, DSM-MAC and LACH protocols in terms of aggregate throughput using ns-3 and python simulations. Three different simulation scenarios have been conducted at varying transmission range, network size and spectrum availability. Rendezvous between a S-D pair is assumed to be realized when nodes can accomplish a successful handshake procedure on the predefined hopping sequence. In total 100 nodes are uniformly deployed in a grid area of 800 x 800 meters, forming a 2D mesh network. Ideal transmission environment have been adopted, where there is no propagation loss and default transmission range of each node is 230 m, if not otherwise specified. A bursty traffic model have been simulated with a traffic arrival probability uniformly distributed between 0 and 1. Burst length is uniformly distributed between 200 and 300 packets, where packet size is 1000 bytes. Nodes select a random neighbour from their transmission range during the handshaking process in order to initiate a transmission. Each cycle consist of 25 time slots of length 10 ms. For the implementation of LACH protocol a latin square of size 25 have been chosen in order to match the cycle length.

2.6.2 Simulation Results

Number Of Channels vs. Aggregate Throughput

Fig. 2.5 shows the aggregate throughputs resulting from IBCH, CHMA, DSMMAC and LACH protocols with respect to number of available channels. Aggregate throughput obtained from IBCH surpasses other protocols as the number of available channels increases in both of the network configurations. Although, multi-rendezvous approach LACH with 60 nodes has overall better performance compared to single-randezvous protocols, CHMA and DSMMAC, it is outperformed by CHMA as the number of available channels increases in the 140 node configuration. Nodes with LACH deployment are incapable of differing their transmissions to an upcoming slot due to

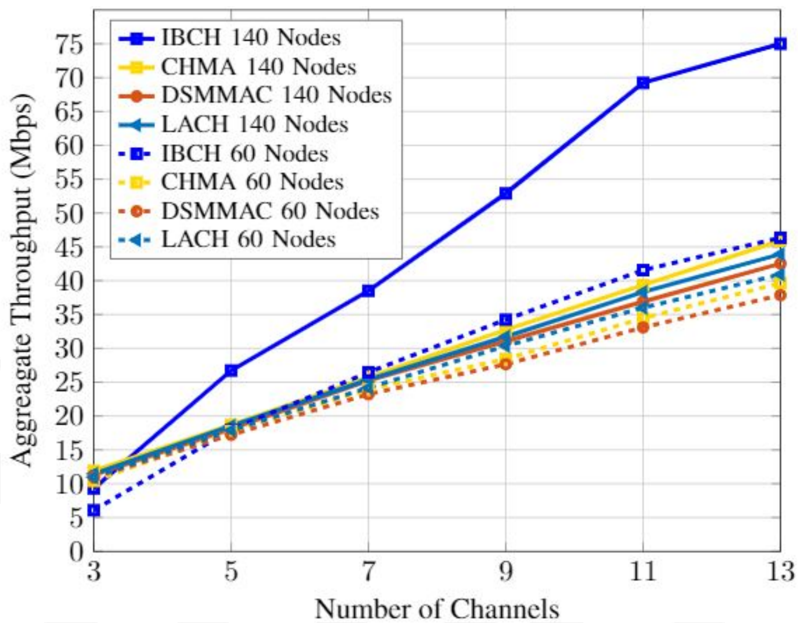

Figure 2.5: Aggregate throughput with different number of channels

its utilization based cyclic scheduling mechanism for default slots. Therefore, slot-based transmission approach of CHMA results in a better aggregate throughput than LACH in case of greater channel and node availability. DSMMAC has the worst performance among all the protocols in terms of aggregate throughput, due to nodes inability to change their difference set based assigned frequencies in case of varying channel utilization. It is observed that in a configuration of 140 nodes, IBCH obtains significantly greater throughput values.

Number Of Nodes vs. Aggregate Throughput

Fig. 2.6 shows the aggregate throughput values obtained from the experiments with 5 and 11 channels. Increasing bandwidth availability is reflected to the acceleration of the increase in the observed aggregate throughput values. Similar to the previous experiment IBCH significantly outperforms other protocols as the network size increases. Multi-rendezvous protocol, LACH, has better performance compared to single-rendezvous protocols, CHMA and DSMMAC, while 5 channels are available. On the other hand, LACH is outperformed by the CHMA protocol in case of greater

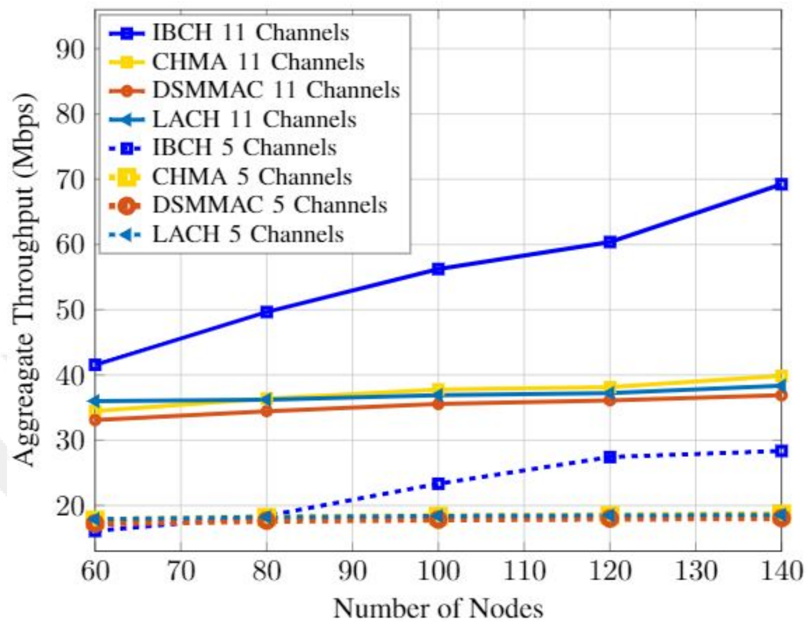

Figure 2.6: Aggregate throughput with different number of nodes

node and frequency spectrum availability due its consideration of cycle-based utilization instead of a slot-based approach, as mentioned above. Furthermore, DSMMAC is again outperformed by the CHMA and LACH protocols due its lack of adaptability to interference with increasing bandwidth availability.

Transmission Range vs. Aggregate Throughput

Fig. 2.7 shows the performance of IBCH, CHMA, DSMMAC and LACH protocols under various transmission ranges. We consider three different transmission ranges of 100, 230 and 350 meters. The main objective of the experiment was to examine the effects of changing interference and node availability. These transmission ranges allowed nodes to hear and randomly select pairs from their 1-hop, 2-hop and 4-hop neighbourhoods respectively. It's evident that multi-randezvous protocols IBCH and LACH allows better spatial reuse and generates higher aggregate throughput with decreasing transmission ranges. On the other hand, examined single-rendezvous protocols, CHMA and DSMMAC, behave completely different with the increasing trans-

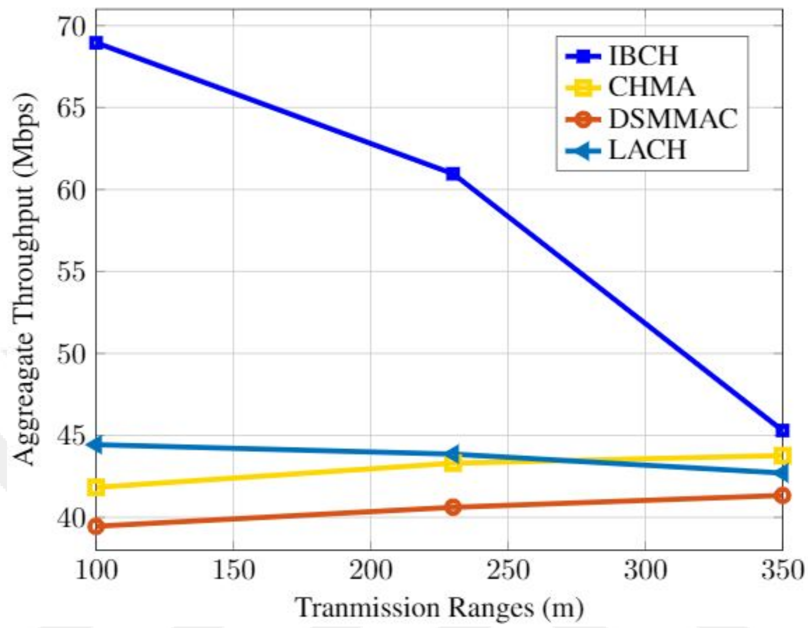

Figure 2.7: Impact of transmission range

mission range. As all of the nodes listen to the same channel at any given slot with single rendezvous protocol deployment, increasing transmission range allows them to be more aware of their surrounding transmissions. It is observed that resulting decrease in interference causes an increase in the aggregate throughput obtained from the single rendezvous protocols. However, IBCH shows greater aggreagated throughput in all of the simulated transmission scenarios.

Chapter 3

DISTRIBUTED DEEP REINFORCEMENT LEARNING WITH WIDEBAND SENSING FOR DYNAMIC SPECTRUM ACCESS

3.1 Introduction

Ever-increasing demand in wireless networks results in an excessive usage of frequency bands in the Radio Frequency (RF) spectrum. Cognitive Radio (CR) is a promising technology for emerging intelligent wireless communication systems due to its efficient utilization of the frequency bands. CR systems enhance temporal and spatial efficiency, therefore, ease spectrum scarcity by exploiting temporary idle periods in the frequency usage, known as spectrum holes. Dynamic Spectrum Access (DSA) plays a central role in CR networks by allowing secondary users (SUs) to opportunistically access these holes in the primary user (PU) channels. Throughput maximization during DSA requires efficient frequency selection and minimal interference by the users in the secondary system. Reinforcement learning is an adaptive online-learning technique that is widely-used for utility maximization in multi-agent systems [24]. Due to its complex and autonomous nature, CR environment can be modeled as a multi-agent system. Therefore, reinforcement learning is a promising solution for spectrum assignment and throughput maximization in CR networks. On the other hand, deep reinforcement learning (DRL) combines reinforcement learning with deep neural network (DNN) architectures for approximating utility values during the learning process [25]. Using DRL enables solving high-dimensional spectrum assignment problems without any prior knowledge about the environment. Up to now, several works with RL and/or deep RL (DRL) have been proposed with the aim of interference avoidance and/or utility maximization in CR networks. DRL based frequency allocation has been investigated for maximizing the number of successful transmissions in [26]. A similar approach has been proposed in [27] to maximize the network utility during spectrum access via Deep Q-network (DQN). Deep Q-learning (DQL) and narrowband sensing policy has been examined for heterogeneous networks that operate different MAC protocols [28]. However, narrowband sensing adopted in these studies limits the usage of multiple spectral opportunities. [29] examines multi-agent RL with wideband sensing for learning sensing policies in a distributed manner. Different from the proposed architecture, we focus on spectrum assignment and primary system frequency allocations. Wideband sensing with DQN based access policy has been investigated for utility maximization of a single SU [30]. However, existence of multiple cognitive agents in different primary network settings and effect of centralized policy learning is omitted in the presented work.

In this section, we propose DQL and wideband sensing based multiple SU utility maximization for the first time in the literature. The original contribution of this study is threefold. First, we develop a Markov Decision Process (MDP) formulation for utility maximization during secondary system decision making based on wideband sensing. Second, we propose novel decentralized and centralized spectrum selection methods based on wideband sensing, namely DEcentralized Spectrum Allocation (DESA) and Centralized Spectrum Allocation (CSA), respectively. Third, we investigate the spectrum utilization of the proposed spectrum allocation methods for varying primary and secondary network sizes.

3.2 System Model

Fig. 3.1 represents IEEE 802.22 standard and Wireless Regional Area Network (WRAN) adopted network architecture. PUs and SUs are assumed to be WRAN Base Station (BS) of a digital television (DTV) system. We consider N primary system users assigned to N frequency sub-bands where each PU allocates all of the resource blocks available by the orthogonal frequency-division multiple access (OFDMA) as in [30].

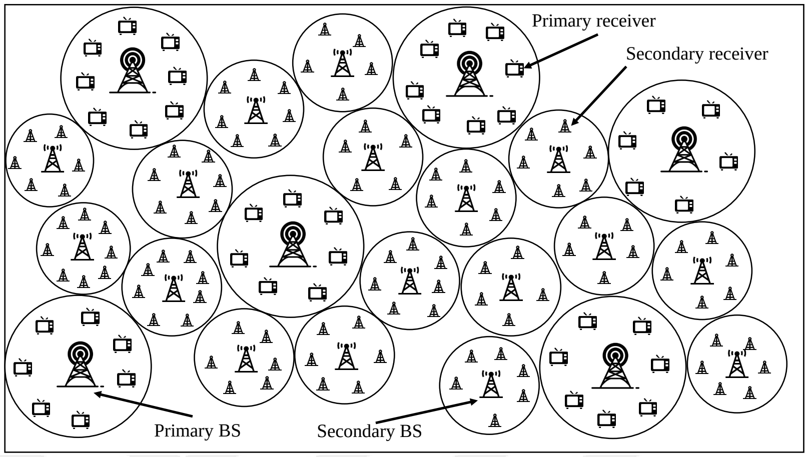

Figure 3.1: PU & SU DTV base stations

Each channel can be modeled as an independent Markov chain that is either occupied or idle at any given time slot. Transition probabilities of channel occupancy states are not known by the secondary system and uniformly distributed between 0 and 1 for each channel. Secondary BSs that are willing to transmit to their receivers are attempting to access idle PU channels without causing interference to primary system transmissions.

Secondary system consists of K cognitive agents that are capable of wideband spectrum sensing. Each of K SUs can stay idle or switch to one of N PU channels for transmission. Furthermore, each of K SU is capable of sensing N primary channels in a decentralized manner. In the centralized scenario, on the other hand, distributed sensing by the secondary system is assumed to be collected at the central DRL unit via distributed sensing [29]. Transmitters and receivers in SU WRAN cells are assumed to be synchronized and no prior knowledge of primary system spectrum occupancy patterns is provided.

3.3 Problem Formulation

MDP is a method used for modeling discrete time stochastic processes. In this section, we first present the MDP formulation of multi-agent decision making considering the secondary system spectrum access problem in DSA. Later on, we elaborate on value functions for estimating utility values of SU spectrum decisions given the primary system channel occupancy at any time slot. Policy defines way of behaving for each SU that is learning PU access patterns. We calculate spectrum decision policy approximating the utility values in the absence of switching behaviour information (transition probabilities) of PU channels. Therefore, we also discuss Q-learning for obtaining optimal spectrum decision policy by recursively approximating value functions.

3.3.1 Markov Decision Process

Our model for the RL task consists of a finite MDP that can be defined using finite set of states and set of actions as well as one-step dynamics of the environment [24]. State space $\mathcal{S} = \{\, s_1, ..., s_S \,\}$ represents possible scenarios of primary system frequency occupations where each observed $s_t \in \{0, 1\}^N$ represents the PU channel allocations at a given time slot t. Each action, $a_j \in \{0, ..., N + 1\}$, from the action space $\mathcal{A} = \{\, a_1, ..., a_K \,\}$, represents a selected PU channel for the decentralized SU j. Action $a_j = N + 1$ represents remaining idle, whereas $a_j \in \{0, ..., N\}$ are for primary channel numbers. In the centralized implementation, $N + 1$ possible actions for each of K SU results in an action space of size $(N + 1)^K$, of which a single encoded action value $a \in \{0, ..., (N + 1)^K\}$ is selected for representing channel selection of all secondary system members.

Each action taken by the SUs determines successor states that are new occupations in PU channels *a.k.a states*. An action a_j taken from state s_t results in a successor state s_k with a state transition probability of $\mathcal{P}(t, k)$ and incurs an immediate reward of $\mathcal{R}(t, j)$. Overall, we can define MDP as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$ with a reward function $\mathcal{R}$ and a discount factor γ for contribution of possible future rewards during possible action scenarios. We define our MDP model as follows:

1. $\mathcal{S}$: $\{\, s_1, ..., s_S \,\}$ state space, where each state s_t represents primary system channel occupations.

2. $\mathcal{A}$: $\{a_1, ..., a_K\}$ action space, where each action represents a decentralized spectrum decision.

3. $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ reward function that represents reward observed $r_t \in \mathcal{R}$ by the secondary BS while accessing an idle PU channel at a given time slot.

4. $\mathcal{P}$: State transition probability $P(s|s')$ that denotes probability of primary channel occupations in the upcoming slot s' given the current state s. We adopt model-free learning approach for our MDP formulation with unknown transition probabilities that are independent of actions taken by SUs.

3.3.2 Value Functions

RL is an effective way of solving stochastic optimal control problems. Therefore, we propose RL based solution for our MDP model with unknown transition probabilities. In RL, value functions are equations for estimating expected sum of discounted rewards from a given state considering the possible future actions [24]. Expected return of a state s under a policy π can be defined using value function $v^\pi(s)$. $\mathbb{E}^\pi[\cdot]$ represents the expected utility of following policy π. For any time step t, r_t, s_t and a_t represent observed reward, state and action at time t as shown in Table 3.1. Value of a state at time t, $v^\pi(s)$, can be calculated with the following equation, where discount factor γ^k is used for decreasing contribution of future state values during expected value calculation

$$v^\pi(s) = \mathbb{E}^\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s\right] \tag{3.1}$$

Similarly, *action-value function* $q^\pi(s, a)$ represents the value of taking action a at time t from state s under policy π and it can be calculated using the following equation;

$$q^{\pi}(s,a) = \mathbb{E}^{\pi}\Big[\sum_{k=0}^{\infty}\gamma^{k} r_{t+k+1}\,\Big|\, s_t = s, a_t = a\Big] \tag{3.2}$$

There exists an optimal deterministic stationary policy for each MDP model [31]. This policy maximizes the expected reward returned from a state in the MDP model with unknown transition probabilities. Optimal *state-value* function v_* and optimal *action-value* function q_* for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ under policy π can be denoted as follows;

$$v_*(s) = \max_{\pi} v^{\pi}(s) \tag{3.3}$$

$$q_*(s,a) = \max_{\pi} q^{\pi}(s,a) \tag{3.4}$$

3.3.3 Q-Learning

Q-learning is a widely-used control algorithm for policy independent approximation of optimal *action-value* function ($q*$) based on *Bellman optimality* equations. In the absence of state transition probabilities $\mathcal{P}$, optimal policy calculation requires approximation of optimal value functions, $q*$ and $v*$, for utility maximization. *Bellman optimality* equations allow policy independent representations and recursive calculation for optimal value functions [24]. These equations are;

$$v_*(s) = \max_{a} \mathbb{E}[\, r_{t+1} + \gamma v_*(s_{t+1}) \,|\, s_t = s, a_t = a\,] \tag{3.5}$$

$$q_*(s,a) = \mathbb{E}[\, r_{t+1} + \gamma \max_{a_{t+1}} q_*(s_{t+1}, a_{t+1}) \,|\, s_t = s, a_t = a \,] \tag{3.6}$$

Learning algorithm uses multiple *experiences* and one-step look-ahead approach for recursively approximating $q*$ values, in which, *experience* represents a sample taken from the MDP model. An *Experience* at time slot t can be denoted using $e_t = \langle s_t, a_t, r_t, s_{t+1} \rangle$. Q-learning uses the following update rule for solving the MDP formulation and approximating $q*$ values;

$$q(s,a) \leftarrow q(s,a) + \Delta q(s,a) \tag{3.7}$$

using learning rate α, $\Delta q(s,a)$ is defined as;

$$\Delta q(s,a) = \alpha \Big[\, r_t + \gamma \max_{a_t} q(s_{t+1}, a_t) - q(s_t, a_t) \,\Big] \tag{3.8}$$

Table 3.1: Reinforcement Learning Notation

Notation	Description
γ	Discount factor
π	Policy
r_t	Reward observed at time t
s_t	State at time t
a_t	Action taken at time t
$v^{\pi}(s)$	*State-value* function under policy π
$q^{\pi}(s,a)$	*Action-value* function under policy π
$v_*(s)$	Optimal *state-value* function
$q_*(s,a)$	Optimal *action-value* function
$e_t = \langle s_t, a_t, r_t, s_{t+1} \rangle$	*Experience* at time t

3.4 Approach

Next, we describe our DQN architecture and algorithms for DQL based spectrum allocation. First, we explain our DQN architecture detailing hyperparameters selected for the learning procedure. Later on, we present wideband sensing based DQL algorithm for single SU frequency selections. Finally, we elaborate on our proposed centralized and decentralized DRL architectures for multi-user utility maximization during secondary system spectrum decisions.

3.4.1 Deep Q-Network

Q-learning calculation requires memorization and representation of Q-values during the expected reward calculation. DQN is a deep reinforcement learning method that uses deep neural network architecture for value storage and expected utility approximation during Q-learning procedure. Using DQN allows complex mapping of high-dimensional *state-space* representations to *action*-values [26, 27]. Our DQN training procedure uses *experience replay* technique. With this approach, we ensure convergent and stable training of DQN model by sampling past experience at each batch and training on decorrelated examples [32].

Table 3.2: Hyperparameters for DQN

Hyperparameter	Value
Number of episodes	200
Number of time slots	10000
Number of layers	5
Batch size B	10
Memory size	2000
Learning rate α	0.001
Discount factor γ	0.95
Exploration rate ϵ	$1.0 \rightarrow 0.01$
Epsilon decay	0.997
Loss function	Huber
Optimizer	Adam

Feed-forward network consists of 5 fully connected layers, where input is the state s_i. Input layer is of size $1\,x\,N$ and output layer is a vector of action set A. Hidden layers are of size 30, 20, 10 and use $ReLU$ activation function that computes $f(x) = max(x, 0)$. Last layer of the network uses *softmax* activation function $\sigma(x_j) = e^{x_j} / \sum_i e^{x_i}$ for predicting Q-value of each action. We chose Adam gradient descent optimization algorithm for updating weights during DQN training [33]. Huber loss has been chosen as the loss function during back propagation due to its robustness against outliers [34]. At each iteration i, Adam algorithm is used to update weights θ_i of the DQN for minimizing loss function

$$L_i(\theta_i) = \begin{cases} \frac{1}{2}(y_i - Q_i)^2 & for\,|y_i - Q_i| \leq c \\ c|y_i - Q_i| - \frac{1}{2}c^2 & otherwise \end{cases} \tag{3.9}$$

where $y_i = \mathbb{E}[\,r_{i+1} + \gamma\max_{a'} q_*(s_{i+1}, a')]$ is calculated with the DQN using weights θ_{i-1} from previous iteration, $Q_i = q(s_i, a|\theta_i)$ and $c = 1.0$ as presented in Table 3.2.

3.4.2 DQL originated Spectrum Access Algorithm (DQLS)

Algorithm 3 is executed for learning spectrum decision policies based on detected PU channel allocations. Learning is triggered upon receiving primary system channel information from the environment using wideband sensing. First, two DQN implementations ($DQN^{train}, DQN^{target}$) are initialized for the *experience replay* procedure using state space S and action space A (Line 1). Training lasts for a predetermined amount of episodes (Lines 2-16). At each episode randomly generated PU spectrum decisions are determined using wideband sensing (Lines 3-15). Action a_t corresponding to the highest expected utility is returned by the DQN^{train} given the current PU channel occupations s_t (Line 4). Training procedure uses ϵ-greedy policy for choosing random actions over actions that return highest expected utility with probability ϵ (Line 5). *Exploration* is necessary in order to find better action selections that result in higher reward signals. We linearly decrease the exploration rate ϵ ($1.0 \xrightarrow{0.997} 0.01$)

Algorithm 3: DQL originated Spectrum Access (DQLS)

Input: $\mathcal{S} \leftarrow$ primary system channel occupation
Output: DQN^{target}
Data: Mem : Memory, b : minibatch

1 **Initialize** $DQN^{train}, DQN^{target} \leftarrow DQN(\mathcal{S}, \mathcal{A})$
2 **for** *each episode* $\mathcal{E}_t$ **do**
3   **for** *each state* $s_t \in \mathcal{E}_t$ **do**
4   **Compute** $a_t = DQN^{train}.get\_action(s_t)$
5         $a_t \leftarrow exploration$ with prob. ϵ
6   **Compute** $r_t = DQN^{train}.get\_reward(s_t, a_t)$
7   **Get** next state s_{t+1}
8   **Store** $Mem.store(e_t = \langle s_t, a_t, r_t, s_{t+1} \rangle)$
9   **Update** $s_t \leftarrow s_{t+1}$
10   **if** $|Mem| > B$ **then**
11   **Compute** $b = Mem.sample(B)$
12   **for** *each* e_i *in* b **do**
13   **Compute** $t = \mathbb{E}[\, r_{i+1} + \gamma \max_{a_{i+1}} q_*(s_{i+1}, a_{i+1})]$
14   **Compute** $DQN^{train}.train(s_i, t)$
15   **Update** $DQN^{target} \xleftarrow{weights} DQN^{train}$

by epsilon decay for a balanced *exploration-exploitation* trade-off. Immediate reward obtained by the selected action is recorded for the rest of the training procedure (Line 6).

After getting the channel occupation in the subsequent time slot current state is updated and an experience e_t is stored in the memory Mem using action a_t, states s_t, s_{t+1} and immediate reward r_t (Lines 7-9). As the experience count surpasses the predetermined batch size B, experiences are randomly sampled from the memory to form a *minibatch* and train on decorrelated experiences (Lines 10-11). Based on the information contained in each experience, Q-value $Q(s, a)$ of that *state-action pair* is calculated and DQN model DQN^{train} is trained (Lines 12-14). After each episode learned weights by the training model DQN^{train} is used for updating the target model DQN^{target} (Line 15).

Algorithm 4: Centralized Spectrum Allocation (CSA)

Input: $K \leftarrow$number of SUs, $N \leftarrow$ number of PUs
Output: Set^{action} : Secondary system actions
Data: $s_i \leftarrow$ state during i-th slot
1 **Compute** $|\mathcal{A}| \leftarrow (N+1)^K$
2 **Initialize** $Set^{action} = \{\}, assigned^{ch} = 0$
3 **Train** $DQN^{target}, DQN^{train} \leftarrow DQLS$
4 **Compute** $a = DQN^{target}.get\_action(s_i)$
5 **while** $assigned^{ch} \neq K$ **do**
6   **Update** $|\mathcal{A}| \leftarrow (N+1)^{K-assigned^{ch}-1}$
7   **Compute** $PU\_index = \lfloor (a/|\mathcal{A}|) \rfloor$
8   **Update** $Set^{action}.append(PU\_index)$
9   **Update** $a = a \mod |\mathcal{A}|$
10   **Update** $assigned^{ch}$ += 1

3.4.3 Centralized Spectrum Allocation Algorithm (CSA)

Centralized DQL architecture uses Centralized Spectrum Allocation Algorithm (CSA) for learning spectrum decision policies. After obtaining primary system channel occupation patterns using distributed sensing [29], Algorithm 4 is run for determining secondary system spectrum decisions. Initially the number of possible actions, action space size, is calculated considering each of K SUs has (N+1) actions for N detected PU channels (Line 1). An empty set of actions Set^{action} is created with a variable $assigned^{ch}$ for counting the number of assigned SU actions (Line 2). Centralized DQN is trained and an action number for encoding all individual actions of K SUs is computed using single DQN based DQLS algorithm (Lines 3-4). Action code a, generated by the central DQN for a given state s_i, is decoded into separate action numbers until all SUs have an assigned action (Lines 5-10). For each individual SU action this procedure keeps the action of remaining SUs as a subset and divides the action code to possible number of remaining actions $(N+1)^{K-assigned^{ch}-1}$ (Lines 6-7). During the rest of the procedure action code representing the cumulative actions of remaining SUs and number assigned SU actions are updated (Lines 9-10). After the calculation of the current channel assigned to the SU, immediate reward signal is obtained from the environment and total utility is updated (Lines 12-13).

3.4.4 Decentralized Spectrum Allocation Algorithm (DESA)

Decentralized policy learning architecture consists of non-cooperative secondary BSs that perform independent wideband sensing and policy learning during DSA. Initially, each secondary BS SU performs wideband sensing individually and determines primary system channel allocations. Based on spectrum occupancy information, SUs are trained using the proposed DQLS algorithm (Lines 1-2). Upon completion of the training procedure, secondary BSs choose the S action and primary channel index, which provide the highest expected utility given state s_j at the current time slot j (Lines 3-4).

Algorithm 5: Decentralized Spectrum Allocation (DESA)

Input: $\mathcal{S} \leftarrow$ primary system channel occupation
Output: a_j : PU channel at j-th slot
Data: $s_j \leftarrow$ state at j-th slot
1 **for** *each* $SU \in$ *secondary system* **do**
2   **Train** $DQN_{SU}^{target}, DQN_{SU}^{train} \leftarrow DQLS$
3 **for** *each* $SU \in$ *secondary system* **do**
4   **Compute** $a_j = DQN_{SU}^{target}.get\_action(s_j)$

3.5 Simulations

3.5.1 Simulation Setup

Simulations of CSA and DESA algorithms have been implemented for centralized and decentralized spectrum decision scenarios using open-source neural-network library Keras [35] and numerical computation library TensorFlow [36]. CSA uses a single DQN for generating secondary system actions during DSA. Central DQN in the CSA is aware of the primary system channels, which are detected by the SUs using wideband sensing approach. During the DESA, each SU has been modeled as a learning cognitive agent that is capable of performing sensing and policy learning operations independently. We have modeled PU channels as independent 2-state Markov chains that can either be in *1(occupied)* or *0 (vacant)* state. Similar to previous work in [30], namely DQN-based access policy (DQNP), we have implemented DSA for $N = 20$ channels each assigned to a different PU with randomly generated transition probability matrices $\{\mathbf{P}_i\}_{i=1}^{20}$.

Network utility performance under CSA and DESA have been compared with classical slotted-Aloha protocol, optimal policy performance and model-based RL approach. Expected channel throughput under slotted-Aloha protocol at a given time-slot i is given by $n_i p_i = (1-p_i)^{n_i-1}$, where p_i represents transmission probability and n_i is number of SUs in the CR environment [27]. Whittle index policy for stochastically identical negatively correlated channels has been implemented as the model-based RL approach [37]. For the optimal policy performance comparison we have implemented fixed-pattern channel switching based optimal policy (OP) algorithm proposed in [26]. Each of the simulations were carried out for $T = 10,000$ time slots and $E = 200$ episodes. We have used *average SU utility per episode* as the performance evaluation metric for representing average cumulative reward gained by SUs at each episode.

3.5.2 Simulation Results

We first present simulation results for spectrum decisions of a single SU coexisting with multiple PUs in the CR environment. Fig. 3.2 shows *average SU utility per episode* obtained by a single SU spectrum decisions under DQLS. Note that network scenario with 1 SU and 20 PUs corresponds to the architecture proposed for DQNP. We have two observations from Fig. 3.2. First, average SU throughput under DQLS increases with the increasing number of detected PU channels. High cardinality of the action space enables SU to be aware of more spectral opportunities and learn a better channel selection policy. Second, convergence to DQN-based access policy takes longer with increasing number of detected PU channels. Increasing complexity in the CR network results in higher dimensionality of action and state space representations,

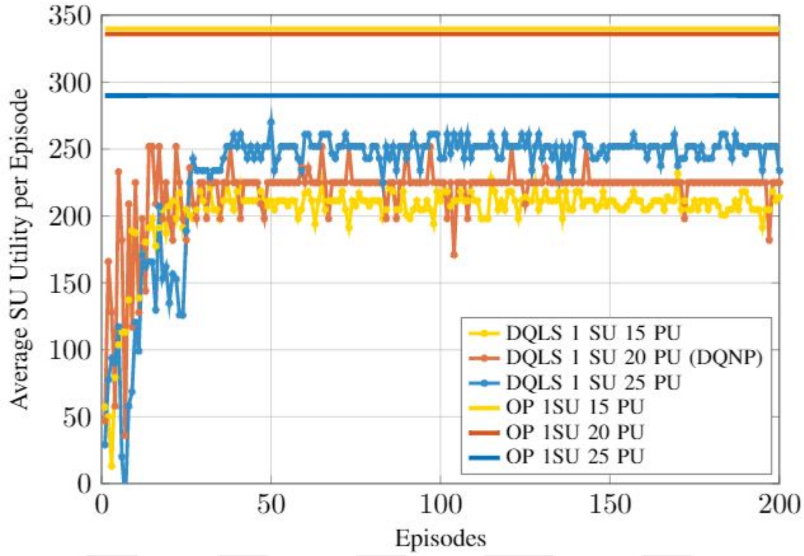

Figure 3.2: DQLS versus optimal channel access policy

hence policy learning from channel occupancy observations slows down.

Fig 3.3 shows comparison of centralized and decentralized spectrum selection performance for varying primary network sizes. Number of SUs is set to 3 while the number of primary base stations vary as 15 and 25. Similar to the results presented in Fig 3.2, decentralized SUs under DESA perform better with DQN-based channel selection strategy as the number of available PU channels per SU increases. Due to increasing complexity of primary channel allocations, centralized learning approach results in a lower performance with increasing number of primary BSs. Different from slotted-Aloha based medium access, average utility under Whittle index policy increases with increasing number of SU. Furthermore, DESA spectrum policy performance surpasses all of the other spectrum decisions approaches achieving up to 80% of optimal policy performance.

Fig 3.4 depicts the performance of proposed MAC schemes with different number of SUs. Similar to the results depicted in Fig 3.2 and Fig 3.3, DESA-based channel access results in a better performance as the number of primary channel per SU increases. It's evident that DESA-based spectrum assignment outperforms other approaches. Furthermore, performance gap with CSA increases as the number of SU reaches to

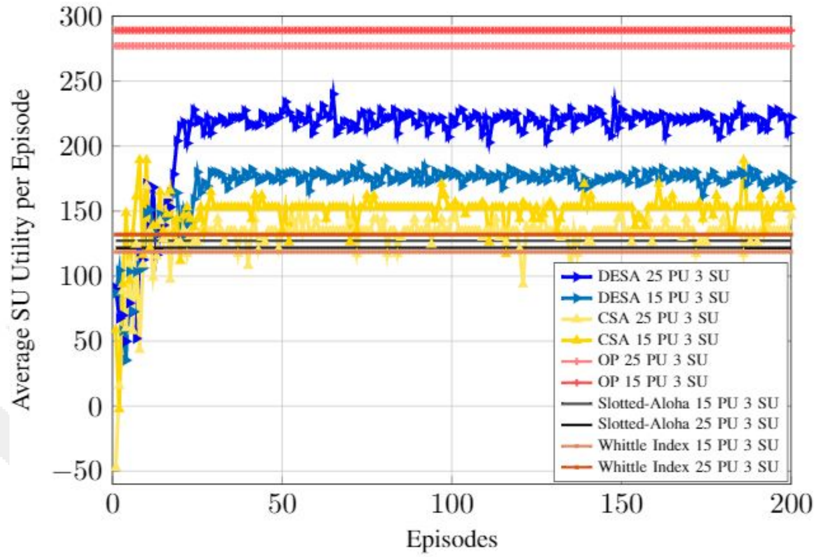

Figure 3.3: CSA versus DSA for changing primary system

4. Compared to results obtained by centralized spectrum assignment under CSA, we observe that independent policy learning in the decentralized scenario results in higher average network utility values. Average utility obtained by the proposed schemes surpasses other approaches with 2 SU availability in the CR environment.

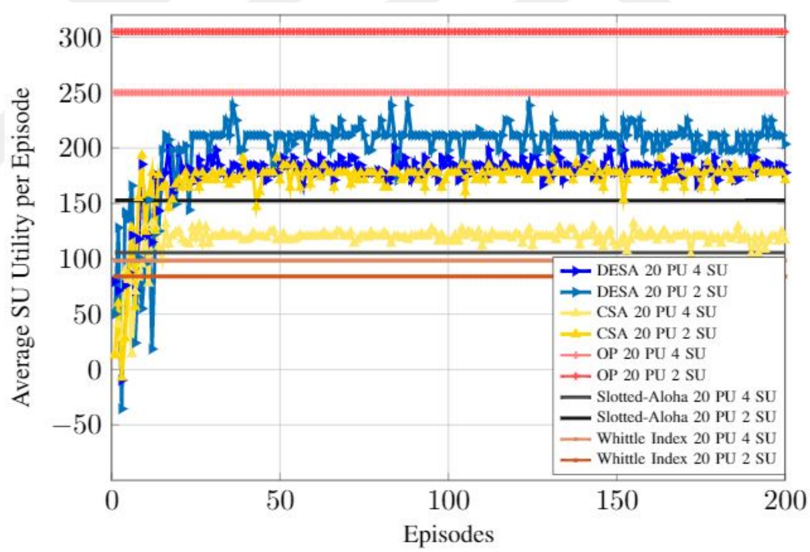

Figure 3.4: CSA versus DSA for changing secondary system

Chapter 4

CONCLUSION

This thesis addresses reliability of spectrum access problems in wireless networks focusing on Wi-Fi and wireless regional area networks. Prosing data-driven spectrum decision solutions for the reliable frequency allocations is the major contribution of the presented work.

In chapter 2, we have initially provide a classification of IEEE 802.11 based SRMC MAC protocols. Later on, we propose a novel Whittle index based multi-channel spectrum decision scheme, IBCH, using our RMAB formulation of dynamic multi-channel access problem. We also compare MAC design schemes with different operational characteristics by changing network configurations such as transmission range, network size and spectrum availability. In general, IBCH results in higher throughput values compared to other protocols with its optimal policy for channel selection and frequency hopping behaviour. On the other hand, single-rendezvous approach common hopping protocol (CHMA) outperforms LACH and DSMMAC as the computational complexity and overhearing increases due to its slot-based decision making flexibility. As future work, we aim to concentrate on mobility and traffic load adaptation of index-based channel hopping.

In chapter 3, we present multi-agent deep reinforcement learning based spectrum selection with wideband sensing capability for multi-user utility maximization during DSA. Initially, DQN-based spectrum decisions of a single SU coexisting with multiple PUs have been derived. Later on, we propose novel algorithms for DQL and wideband sensing based centralized and decentralized spectrum selection, namely DESA and CSA. Decentralized channel access results in a better performance as the number of primary channel per SU increases. Moreover, we observe that independent policy learning in non-cooperative manner under DESA results in more effective spectrum decisions than centralized spectrum assignment under CSA. Going forward, our future work will focus on extending both the system model and proposed algorithms. We will concentrate on efficient utilization of OFDMA resource blocks and analyzing aggregated interference at primary receivers during secondary system transmissions. We also plan to implement other DRL approaches such as policy gradient methods.

BIBLIOGRAPHY

[1] Jeonghoon Mo, H. So and J. Walrand, "Comparison of Multichannel MAC Protocols", IEEE Transactions on Mobile Computing, vol. 7, no. 1, pp. 50-65, 2008.

[2] J. So and N. Vaidya, "Multi-channel mac for ad hoc networks: handling multi-channel hidden terminals using a single transceiver", in In Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing (MobiHoc '04), 2019, pp. 222-233.

[3] Y. Kim, M. Lee and T. Lee, "Coordinated Multichannel MAC Protocol for Vehicular Ad Hoc Networks," in IEEE Transactions on Vehicular Technology, vol. 65, no. 8, pp. 6508-6517, Aug. 2016.

[4] Jenhui Chen, Shiann-Tsong Sheu and Chin-An Yang, "A new multichannel access protocol for IEEE 802.11 ad hoc wireless LANs," 14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003., Beijing, China, 2003, pp. 2291-2296 vol.3.

[5] B. Yang, B. Li, Z. Yan, Mao Yang and Xiaoya Zuo, "A reliable channel reservation based multi-channel MAC protocol with a single transceiver," 2015 11th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QSHINE), Taipei, 2015, pp. 265-271.

[6] A. Tzamaloukas and J. J. Garcia-Luna-Aceves, "Channel-hopping multiple access," 2000 IEEE International Conference on Communications. ICC 2000. Global Convergence Through Communications. Conference Record, New Orleans, LA, USA, 2000, pp. 415-419 vol.1.

[7] A. Tzamaloukas and J. J. Garcia-Luna-Aceves, “Channel Hopping Multiple Access with Packet Trains for Ad Hoc Networks,” In Proc.IEEE Mobile Multimedia Communications (MoMuC), Jan. 2000.

[8] F. Hou, L. X. Cai, X. Shen and J. Huang, "Asynchronous Multichannel MAC Design With Difference-Set-Based Hopping Sequences," in IEEE Transactions on Vehicular Technology, vol. 60, no. 4, pp. 1728-1739, May 2011.

[9] Paramvir Bahl, Ranveer Chandra, and John Dunagan, " SSCH: slotted seeded channel hopping for capacity improvement in IEEE 802.11 ad-hoc wireless networks ",In Proceedings of the 10th annual international conference on Mobile computing and networking (MobiCom ’04), 216-230.

[10] Hoi-Sheung, W. So, J. Walrand and Jeonghoon Mo, "McMAC: A Parallel Rendezvous Multi-Channel MAC Protocol," 2007 IEEE Wireless Communications and Networking Conference, Kowloon, 2007, pp. 334-339.

[11] C. Chao and H. Tsai, "A Channel-Hopping Multichannel MAC Protocol for Mobile Ad Hoc Networks," in IEEE Transactions on Vehicular Technology, vol. 63, no. 9, pp. 4464-4475, Nov. 2014.

[12] C. Chao, H. Tsai and C. Huang, "Load-aware channel hopping protocol design for mobile ad hoc networks," ISWPC 2012 proceedings, Dalian, 2012, pp. 1-6.

[13] Yu-Chee Tseng, Shih-Lin Wu, Chih-Yu Lin and Jang-Ping Sheu, "A multichannel MAC protocol with power control for multi-hop mobile ad hoc networks," Proceedings 21st International Conference on Distributed Computing Systems Workshops, Mesa, AZ, USA, 2001, pp. 419-424.

[14] E. N. Gilbert, "Capacity of a burst-noise channel," in The Bell System Technical Journal, vol. 39, no. 5, pp. 1253-1265, Sept. 1960.

[15] W. R. Thompson, “On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples,” Biometrika, vol. 25, no. 3/4, p. 285, 1933.

[16] Whittle, “Restless bandits: activity allocation in a changing world,” Journal of Applied Probability, vol. 25, no. A, pp. 287–298, 1988.

[17] R. D. Smallwood and E. J. Sondik, “The Optimal Control of Partially Observable Markov Processes over a Finite Horizon,” Operations Research, vol. 21, no. 5, pp. 1071–1088, 1973.

[18] J. C. Gittins, “Dynamic Allocation Indices for Bayesian Bandits,” Mathematical Learning Models — Theory and Algorithms Lecture Notes in Statistics, pp. 50–67, 1983.

[19] K. Liu and Q. Zhao, “Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access,” IEEE Transactions on Information Theory, vol. 56, no. 11, pp. 5547–5567, 2010

[20] "ns-3", ns-3, 2019. [Online]. Available: https://www.nsnam.org/. [Accessed: 15-Dec- 2019].

[21] K. Wang and L. Chen, "On Optimality of Myopic Policy for Restless Multi-Armed Bandit Problem: An Axiomatic Approach," in IEEE Transactions on Signal Processing, vol. 60, no. 1, pp. 300-309, Jan. 2012.

[22] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao and B. Krishnamachari, "Optimality of Myopic Sensing in Multichannel Opportunistic Access," in IEEE Transactions on Information Theory, vol. 55, no. 9, pp. 4040-4050, Sept. 2009.

[23] Q. Zhao, L. Tong, A. Swami and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," in

IEEE Journal on Selected Areas in Communications, vol. 25, no. 3, pp. 589-600, April 2007.

[24] Sutton, Richard S. and Barto, Andrew G." Reinforcement Learning: An Introduction", Second : The MIT Press, 2018.

[25] Mnih, V., Kavukcuoglu, K., Silver, D. et al." Human-level control through deep reinforcement learning", Nature, 529–533, 2015.

[26] S. Wang, H. Liu, P. H. Gomes and B. Krishnamachari, "Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks," in IEEE Transactions on Cognitive Communications and Networking, vol. 4, no. 2, pp. 257-265, June 2018.

[27] O. Naparstek and K. Cohen, "Deep Multi-User Reinforcement Learning for Distributed Dynamic Spectrum Access," in IEEE Transactions on Wireless Communications, vol. 18, no. 1, pp. 310-323, Jan. 2019.

[28] Y. Yu, T. Wang and S. C. Liew, "Deep-Reinforcement Learning Multiple Access for Heterogeneous Wireless Networks," 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, 2018, pp. 1-7.

[29] J. Lundén, S. R. Kulkarni, V. Koivunen and H. V. Poor, "Multiagent Reinforcement Learning Based Spectrum Sensing Policies for Cognitive Radio Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 7, no. 5, pp. 858-868, Oct. 2013.

[30] H. Q. Nguyen, B. T. Nguyen, T. Q. Dong, D. T. Ngo and T. A. Nguyen, "Deep Q-Learning with Multiband Sensing for Dynamic Spectrum Access," 2018 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), Seoul, 2018, pp. 1-5.

[31] R. Bellman, THE THEORY OF DYNAMIC PROGRAMMING. Ft. Belvoir: Defense Technical Information Center, 1954.

[32] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, and Rob Fergus (Eds.). Curran Associates Inc., USA, 5055-5065

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization,"CoRR, vol. abs/1412.6980, 2015.

[34] P. J. Huber, "Robust Estimation of a Location Parameter," The Annals of Mathematical Statistics. [Online]. Available: https://projecteuclid.org/euclid.aoms/1177703732

[35] "Keras: The Python Deep Learning library," Home - Keras Documentation. [Online]. Available: https://keras.io/

[36] "TensorFlow: A System for Large-Scale Machine Learning." [Online]. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

[37] K. Liu and Q. Zhao, "Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access," IEEE Transactions on Information Theory, vol. 56, no. 11, pp. 5547–5567, 2010.