Game theory for tactical networks

Mahmoud Almasri

DOCTORAT BRETAGNE LOIRE MATHSTIC

ENSTA Bretagne

THESE DE DOCTORAT DE

L'ECOLE NATIONALE SUPERIEURE
DE TECHNIQUES AVANCEES BRETAGNE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Télécommunications, Informatique,*

Par

Mahmoud ALMASRI

Game Theory for Tactical Networks

Thèse présentée et soutenue à Brest, le 27 Avril 2020

Unité de recherche : Lab-STICC UMR CNRS 6285

Rapporteurs avant soutenance :

M. Karim Abed - Meraim
Prof, Université d'Orléans

Mme Nadège Thirion - Moreau
Prof, Université de Toulon

Composition du Jury :

M. Emanuel Radoi
Prof, Université de Bretagne Occidentale, Examinateur

M. Christian Jutten
Prof, Université de Grenoble, Examinateur

M. Christophe Moy
Prof, Université des Rennes 1,

M. Vincent Choqueuse
Dr, Enib Brest,

M. Ali Mansour
Prof, Ensta-Bretagne, Directeur de thèse

M. Ammar Assoum,
Prof, Université Libanaise,

To My Wonderful Parents...

Title: A PhD thesis, Game Theory for Tactical Networks

Keywords: Opportunistic Spectrum Access, Multi-Armed Bandit, Priority Dynamic Access

Since 1990's, the demand on wireless devices, mobile and wireless networks, has experienced unprecedented growth which makes the frequency bands more and more crowded. Several studies, initiated by the Federal Communications Commission (FCC) in the United States (US), have shown that the frequency bands are not well used: Some frequency bands are overlapped while others underutilized. The Opportunistic Spectrum Access (OSA) in Cognitive Radio (CR) represents one of several proposed solutions to tackle the scarcity and enhance the efficiency use of the spectrum. In OSA, two categories of users are considered: Primary Users (PUs), also known as licensed users, have the right to fully access their dedicated bandwidths; and Secondary Users (SUs), i.e. opportunistic users, would like to exploit vacant frequency bands unused by the PUs. Due to hardware limitation, a SU can access one channel at each time slot trying to reach the best channel with the highest vacancy probability. To identify the best channel, we formulate OSA as a Multi-Armed Bandit (MAB) problem, in which an agent plays one arm at each time trying to reach the optimal arm with the highest expected reward. Several MAB algorithms have been suggested to solve the MAB problem in the context of OSA, such as: Thompson Sampling (TS), Upper Confidence Bound (UCB), e-greedy, etc.

By focusing first on a single SU, we analyze the performance of the well-known MAB algorithms (i.e. TS, UCB, e-greedy) that deal with OSA. Thus, we propose our MAB algorithms based on UCB, called: e-UCB and AUCB. Both of them achieve good results compared to well-known variants of MAB algorithms, i.e. UCB and e-greedy, in which the SU can quickly learn the vacancy probability of channels without any information or prior knowledge about the available channels. Our analytical proof, as well as the simulation results, of e-UCB and AUCB show that the SU can efficiently distinguish and converge to the best channel after a finite number of time slots.

For multiple users, the big challenge of SUs remains to learn collectively (Cooperative learning) or separately (Competitive learning) the vacancy probabilities of the channels. As a matter of fact, a cooperative or competitive learning policy is required in order to manage the secondary network and decrease the number of collisions among users. Generally, the policies to manage a secondary network can be classified into two main categories: Random access or priority access. Most recent works in OSA focus on the random access while the priority access is not enough considered in the literature. In fact, the priority access can have an important role in tactical networks in which several SUs exist with some hierarchy levels.

In our work, we propose a cooperative and competitive policies for the priority access respectively called Side Channel and All-Powerful Learning (APL). In our policies, each SU has an assigned priority rank, and his target remains to access the channels according to his rank. Moreover, Side Channel and APL deal with the priority dynamic access where the users can enter into or leave the network. While, to the best of our knowledge, only the priority or dynamic access are considered in several recent works. Finally, a proof is developed to verify the performance of proposed learning policies on a real radio environment. Simulation results show that Side channel and APL can achieve better results than several recent works: the users can quickly reach their dedicated channels while decreasing the number of collisions among them.

Titre: Thèse de doctorat, Théorie des jeux pour les communications militaires tactiques

Mots-clés: Accès Opportuniste au Spectre, Multi-arm bandit , Accès prioritaire dynamique

Durant le siècle passé, les ressources spectrales ont été allouées exclusivement aux services qui sont apparus au fur et à mesure des années. Avec une augmentation soutenue, des besoins en bandes fréquentielles à allouer aux applications de communication sans fil émergentes, les opérateurs de radiocommunication se sont trouvés, face à une pénurie de ressources spectrales. Néanmoins, plusieurs études initiées par la Commission fédérale des communications (FCC : Federal Communications Commission qui est une agence indépendante du gouvernement des États-Unis) ont montré que les bandes de fréquences sont mal exploitées: certaines bandes sont peu chargées, or d'autres sont surchargées. L'Accès Opportuniste au Spectre (AOS) dans une radio cognitive (RC) représente une potentielle solution proposée pour lutter contre un manque accru du spectre et améliorer le rendement de l'utilisation. Dans un AOS, deux catégories d'utilisateurs sont définis: les utilisateurs primaires (PU), possédant les licences, ont un droit exclusif d'accéder à leurs bandes fréquentielles en permanence; et les utilisateurs secondaires (SU) ou opportunistes qui cherchent à exploiter les bandes de fréquences libérer par un PU. Dans beaucoup de situations et dû généralement à limitation matérielle ou à un coût, un SU ne peut explorer et y accéder qu'un seul canal à un instant donné. Ce SU cherche donc à trouver le meilleur canal, i.e. le canal possédant les meilleures conditions de transmission qui sera le plus disponible. Pour identifier le meilleur canal, nous avons proposé un modèle d' AOS en se basant sur un problème de multi-arm bandit (MAB), dans lequel un joueur joue une seule machine à sous à chaque tournée en espérant de découvrir la meilleure machine qui augmentera son gain. Dans la littérature, plusieurs algorithmes ont été développés pour mieux aborder le problème du MAB, notamment: Thompson Sampling (TS), Upper Confidence Bound (UCB), e-greedy, etc..

En considérant uniquement le cas d'un seul SU, nous avons analysé et comparé les performances des algorithmes bien cités du MAB (TS, UCB, e-greedy). Nous avons par ailleurs proposé deux nouveaux variétés de l'algorithme UCB: e-UCB et AUCB. Les deux derniers algorithmes ont donné une grande satisfaction en montrant des meilleures performances que les autres variantes bien connues des algorithmes UCB ou e-greedy, dans lesquelles le SU peut rapidement estimer la probabilité de disponibilité des canaux sans préalable information.

Dans un deuxième temps, nous avons penché sur un cas plus général où plusieurs utilisateurs secondaires coexistent, leur principal goal de ces SU reste à trouver la meilleur stratégie (apprentissage coopératif) ou les stratégies individuelles (apprentissage compétitif) pour mieux estimer les probabilités de disponibilité des canaux. Généralement, un réseau secondaire peut être générer selon deux types d'accès : aléatoire ou prioritaire. Plusieurs travaux récents sur l'AOS ont exclusivement considéré l'accès aléatoire. Par contre, l'accès prioritaire a été relativement négligé, alors qu'un accès prioritaire devient crucial dans des réseaux tactiques pour lesquels plusieurs SU coexistent avec un certain niveau de hiérarchie. Pour plusieurs utilisateurs, le grand défi des SU reste d'apprendre collectivement (apprentissage coopératif) ou séparément (apprentissage compétitif) les probabilités de disponibilité des canaux. En effet, une stratégie d'apprentissage coopératif ou compétitif est nécessaire pour gérer le réseau secondaire et diminuer le nombre de collisions entre les utilisateurs. Généralement, les stratégies de gestion d'un réseau secondaire peuvent être classées en deux catégories principales: accès aléatoire ou accès priorité. Les travaux les plus récents en OSA se concentrent sur l'accès aléatoire alors que l'accès priorité n'est pas suffisamment pris en compte dans la littérature. En fait, l'accès priorité peut avoir un rôle important dans les réseaux tactiques dans lesquels plusieurs SU existent avec certains niveaux de hiérarchie.

Dans nos études et pour un réseau tactique avec une certaine hiérarchie, nous avons proposé deux stratégies, l'une coopérative et l'autre compétitive Side Channel et All-Powerful Learning (APL) respectivement. Selon ces deux stratégies, chaque SU a un rang fixe, et son objectif est d'accéder aux canaux disponibles en respectant son rang. En plus, Side Channel et APL prennent en compte un accès prioritaire et dynamique, où les utilisateurs peuvent entrer ou sortir du réseau à tout moment. Dans la littérature, un accès prioritaire ou un accès dynamique ont été séparément évoqués. Finalement une étude de performance théorique a été développée pour les stratégies d'apprentissage proposées. Les simulations ont montré que Side Channel et APL ont donné les meilleurs résultats par rapport à la littérature. En appliquant l'une de ces deux stratégies, les utilisateurs secondaires peuvent rapidement identifier les canaux correspondants à leurs rangs tout en réduisant le nombre de collisions parmi eux.

Acknowledgements

First and foremost, I would like to thank my thesis director Prof. Ali MANSOUR for his continuous support, motivation and untiring guidance have made this dream come true. His vast knowledge, calm nature and positive criticism motivated me to starve for nice results. I would also express my sincere thanks to his wife "Houwaida", his son "Taha" and his daughter "Mariam" for the warm and gracious hospitality during different vacations spent in the villages, beaches and mountains. They are my second family members in France.

I would like to express my thanks to Prof. Karim ABED MERAIM and Prof. Nadège Thirion for reviewing my thesis and for all their valuable comments that helped me to improve it. Also, thank you to Prof. Emanuel Radoi and Prof. Christian Jutten as examiners. I would like to acknowledge Dr. Christophe Osswald, Prof. Christophe Moy, Dr. Denis Lejeune for their support throughout the Ph.D. years. I have improved immeasurably as a scientist and as a communicator, and the opportunity to work with them over the last three years was an honor.

Thanks go out to Prof. Ammar Assoum for his guidance (academic, scientific, and otherwise) through the course of this research, and to Dr. Vincent Choqueuse for his useful comments in the CSI and for attending my defense as an invited member.

My eternal cheerleader: I miss our interesting and long-lasting chats. My forever interested, encouraging and always enthusiastic my mother: Ghazwa Almasri, she was always keen to know what I was doing and how I was proceeding, although it is likely that she has never grasped what it was all about! I will miss your screams of joy whenever a significant momentous was reached. This moment also reminds me of the influence of my father and my siblings, who have provided me through moral and emotional support in my life. I am also grateful to my other family members and friends who have supported me along the way Jean Marrie, Houssam Barbara, Kahina Bensafia. I wished if they share with me this moment that they and I dreamt of. Thanks for all your encouragement!

Contents

List of Tables

List of Figures

Acronyms

ADC	**A**nalog to **D**igital **C**onverter
AMPS	**A**nalog **M**obile **P**hone **S**ervice
CR	**C**ognitive **R**adio
CC	**C**ognitive **C**ycle
CNET	**C**entre **N**ational d'**E**tudes des **T**élécommunications
CDMA	**C**ode **D**ivision **M**ultiple **A**ccess
CRN	**C**ognitive **R**adio **N**etwork
CFD	**C**yclostationary **F**eature **D**etection
DSA	**D**ynamic **S**pectrum **A**ccess
DAC	**D**igital to **A**nalog **C**onverter
ED	**E**nergy **D**etection
FCC	**F**ederal **C**ommunications **C**ommission
FDMA	**F**requency **D**ivision **M**ultiple **A**ccess
GPRS	**G**eneral **P**acket **R**adio **S**ervice
GSM	**G**lobal **S**ystem for **M**obile communications
ISM	**I**ndustrial, **S**cientific and **M**edical
IoT	**I**nternet **o**f **T**hings
IID	**I**ndependent and **I**dentically **D**istributed
KL-UCB	**K**ullback-**L**eibler **U**pper **C**onfidence **B**ounds
LTE	**L**ong **T**erm **E**volution
M2M	**M**achine **to** **M**achine
MFD	**M**atched **F**ilter **D**etection
MAB	**M**ulti-**A**rmed **B**andit
MEGA	**M**ulti-user **e**-**g**reedy Collision **A**voiding

OSA	**O**pportunistic **S**pectrum **A**ccess
PSD	**P**ower **S**pectral **D**ensity
PU	**P**rimary **U**ser
QoS	**Q**uality **o**f **S**ervice
RF	**R**adio **F**requency
RL	**R**einforcement **L**earning
SU	**S**econdary **U**ser
SLK	**S**elective **L**earning of the **k**-th largest expected rewards
SDR	**S**oftware-**D**efined **R**adio
TS	**T**hompson **S**ampling
TDFS	**T**ime-**D**ivision **F**air **S**hare
UMTS	**U**niversal **M**obile **T**elecommunications **S**ystem
UCB	**U**pper **C**onfidence **B**ound
UWB	**U**ltra **W**ide **B**and
WCDMA	**W**ideband **C**ode **D**ivision **M**ultiple **A**ccess
WLAN	**W**ireless **L**ocal **A**rea **N**etwork

Chapter 1

Advanced Technologies in Wireless Internet and Communications Networks

Contents

1.1 Thirty years of wireless communications technology

Demand of wireless devices, mobile and wireless networks, has experienced unprecedented advancement since 1990's. However, wireless networks started with small personal area networks (e.g. WiFi, Bluetooth, etc.) and evolved into metropolitan networks (e.g. GSM: Global System for Mobile communication). From wireless local area networks (WLANs) to mobile phones, people need to be connected to the network anytime no matter where they are. Indeed, WLANs entered in everyday life through standards. Moreover, several generations of wireless networks have been deployed, mainly dedicated to cellular telephone technology (e.g. GSM) then more oriented towards multimedia (e.g. UMTS: Universal Mobile Telecommunications System).

Nowadays, it is very easy to establish in a few minutes a wireless network allowing computers or mobile devices to communicate with each other. The difficulty of implementation is the reception area, related to the power of the transmitter, the detection of the receiver and the security of the transmitted data. The primary and essential advantage of wireless communication systems is mobility. Finally, we can say that, due to their advantages, flexibility and mobility, wireless communication networks such as WLAN or telephony networks represent the best technologies to the industrial automation. Moreover, it has been shown that the frequency bands of wireless communication are not uniformly used (some frequency bands are not well used while others are very crowded), hence the necessity to introduce novel technologies in order to achieve an efficient use of the spectrum. The Cognitive Radio represents one of the proposed technologies, widely suggested in the literature, to enhance the efficiency use of the spectrum.

1.1.1 WLAN and unlicensed standards

Nowadays, WLAN represents one of the most popular technologies used by the industry in order to provide machine-to-machine connections, help mobile devices to connect to a given network, as well as to extend the internet networks (e.g. the connection between mobile devices and routers can reach up to 150 meters indoors and 300 meters outdoors).

WLAN IEEE 802.11 (WiFi) family represents the famous standards firstly

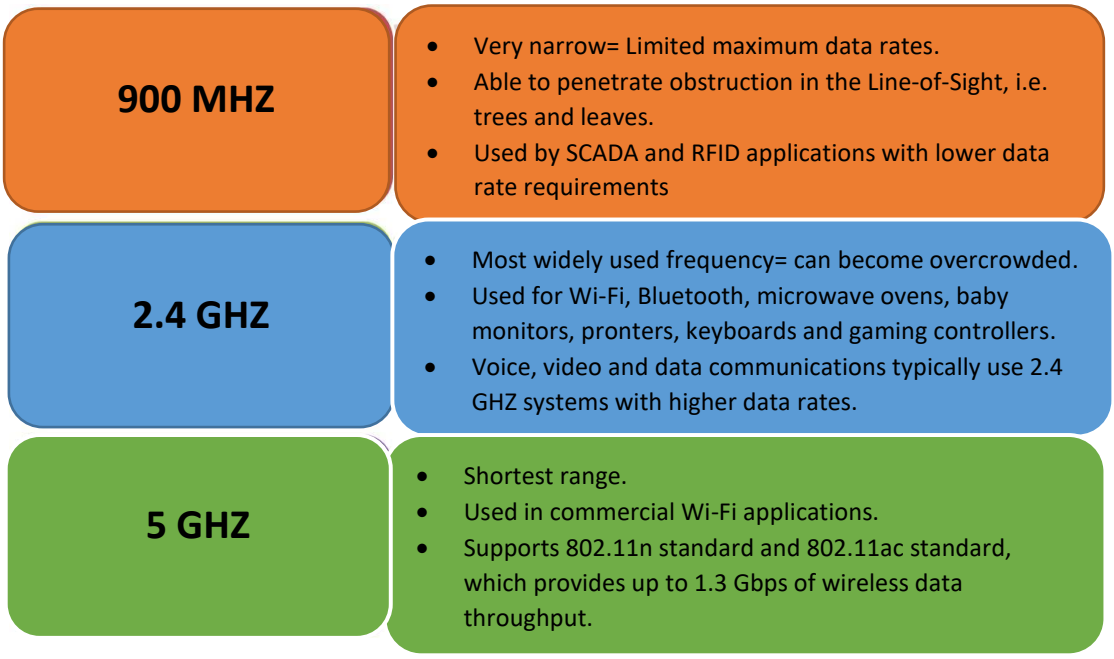

FIGURE 1.1: Use of ISM band frequencies in industrial wireless networks.

commercialized by Apple for their iBook series in 1999. Several new applications that provide new technical challenges, based on WiFi technology, have been proposed such as Voice over IP (VoIP), Video streaming, etc. Recently, these applications, becoming mature with low cost compared to cellular communication, use a broadband Internet connection in order to make unlimited phone calls and to unload overcrowded cellular networks. Even though their benefits, the arrival of new wireless applications and services makes the unlicensed frequency bands more and more crowded.

For its simplicity and efficiency, all cellphones, laptops, tablet computers and printers nowadays, having 802.11 wireless modems, use the 2.4 GHz or 5 GHz ISM (Industrial, Scientific and Medical) bands. Based on [1, 2], we make a comparison between the most commonly used ISM bands in the industrial wireless networks. Despite the intent of the original allocation, the fastest-growing uses of the ISM bands recently return to the low-power and short-range communication systems: Wireless computer networks, wireless keyboards and mice, baby monitors, cordless phones, garage door openers. The ISM bands, specifically the 2.4 GHz, have been originally designed for several applications, such as: medical equipment, microwave, types of electrodeless lamps, and process heating [3].
ISM generally represents unlicensed frequency bands that are overloaded,

noisy, and need powerful Radio Frequency (RF) system to manage the interference problems. Moreover, many technologies that use ISM bands (e.g. WiFi) do not provide any protection or warranty that may lead to unsecure communication among the devices. Over the past few years, there have been many applications and communication techniques that have adopted ISM bands. The Internet of Things (IoT) represents one of these applications and this is mainly due to the low power and cost of spectrum of ISM bands without users being directly involved. Furthermore, the future technologies that use ISM frequency bands may incorporate satellite communications, such as smallSATs (small-satellites), cubeSATs (cube-satellites), nanoSATs (nano-satellites)[4, 5].
To solve the overlap of the unlicensed bands, in which all the mentioned technologies make the frequency bands more and more crowded, the US Federal Communications Commission (FCC) has recommended several ways to manage and solve the limitation of the frequency resources. The well-known adopted solution by FCC is to open the licensed frequencies for the opportunistic use, provided that the interference with the license users are under certain limits.

1.1.2 Evolution of the mobile generations

Since a few decades, mobile wireless communications have experienced four generations of revolution and evolution as shown in Table 1.1. Each of them has a higher capacity and new features compared to the previous one. The 5G implies the whole wireless world interconnection together with very high data rates Quality of Service (QoS). From second to fifth generation, each generation has a significant improvement to the previous one (e.g. enhancement the quality of services with more satisfaction in customer experience).
The Cognitive Radio (CR) represents one of the new technologies that use the benefit of the Software-Defined Radio (SDR), based on its flexibility, in order to greatly increase the spectrum efficiency. As we will see in section 1.2, the Opportunistic Spectrum Access (OSA) in CR allows a cognitive user to access the unused licensed spectrum in opportunistic manner. As mentioned in [6, 7, 8], CR is a future technology that is considered as an enabling technology for future cellular mobile generations.

The first mobile generation (1G) , also known as AMPS (Analog Mobile Phone Service), is an analog technology of cellular network, and was introduced

in 1970's and fully implemented throughout the 1980's, purely designed for voice calls without considering data services. AMPS had several problems such as poor voice quality, lack of security, and sometimes suffered from calls dropped.
Unlike the first generation, Global System for Mobile Communications (GSM) is a digital standard developed by the CNET (Centre National d'Etudes des Télécommunications) in France, also known as the second generation cellular technology, it is used by mobile phones and tablets for full duplex voice telephony. In order to add a packet transmission capability to GSM, General Packet Radio Service (GPRS) was developed and has been available for users over GSM. GPRS is more suitable method for transmitting data over a cellular network. Indeed, in the case of GPRS, the resources are not continuously allocated for transmission but only when data is exchanged, unlike for Circuit Switched Voice Services in GSM where a virtual circuit is established and associated resources are reserved for the duration of the communication. GPRS is a data service that can be used for Multimedia Messaging Service (MMS), short message service (SMS), Wireless Access Point (WAP), Internet browsing, etc. GPRS is often called 2.5G or 2G+ where 2.5 indicates that it is a technology between GSM (second generation) and Universal Mobile Telecommunications System (UMTS: third generation). This latter is a cellular technology based on the Wideband Code Division Multiple Access (WCDMA) technique while the multiple access for GSM is done by a combination between Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA). A major improvement of UMTS over GSM is, thanks to its new coding technique, the possibility to reuse same frequencies in adjacent radio cells and consequently to assign a larger spectral width for each cell (5 MHz). Whereas in GSM, adjacent radio cells must use different frequency bands (reuse factor varying from 1/3 to 1/7) which implies (in GSM) dividing and distributing the frequencies allocated to the operator between several radio cells. Thanks to the increased speed of data transmission, UMTS opened the door to new applications and services. In particular, UMTS made it possible to transfer multimedia message such as images, sound and videos in real time. In France, SFR Company (a French telecommunication operator) was the first to launch its UMTS commercial offer on November 10th 2004, followed by Orange on December 9th 2004. Nevertheless, Bouygues Telecom has opened its UMTS license commercially at the beginning of 2007. The 4th generation or Long Term Evolution (LTE) mobile network was quickly installed in France to replace previous generations: 3G,

Technology ⇨ / Feature ⇩	1G	2G	3G	4G	5G
Start/ Deployment	1970 – 1980	1990 – 2004	2004-2010	Now	Soon (probably 2020)
Data Bandwidth	2kbps	64kbps	2Mbps	1 Gbps	Higher than 1Gbps
Technology	Analog Cellular Technology	Digital Cellular Technology	CDMA 2000 (1xRTT, EVDO) UMTS, EDGE	Wi-Max LTE Wi-Fi	WWWW(coming soon)
Service	Mobile Telephony (Voice)	Digital voice, SMS, Higher capacity packetized data	Integrated high quality audio, video and data	Dynamic Information access, Wearable devices	Dynamic Information access, Wearable devices with AI Capabilities
Multiplexing	FDMA	TDMA, CDMA	CDMA	CDMA	CDMA
Switching	Circuit	Circuit, Packet	Packet	All Packet	All Packet
Core Network	PSTN	PSTN	Packet N/W	Internet	Internet

TABLE 1.1: Efficiency of each cellular wireless generation, 1G allowed making a call, 2G allowed to send text messages and MMS, 3G added video calls and access to the internet, 4G guarantees a worthy quality of a Wi-Fi connection. The high speed of the 5G will allow to develop many communication systems in the urban landscape, such as transport, traffic, radar imagery, security control it will also ensure faster data rates, higher connection density, improved overall wireless coverage, etc.

2G and 1G. Nowadays, 4G continues its deployment in France and should cover 98% of the territory by 2024.

Following on from 4G, the 5th generation is the last proposed standard with a max speed of 35.46 Gbps, will be 35 times faster than 4G. With its high speed, the fifth-generation has the potential to enable fundamentally new applications, industries, and business models and dramatically improve the quality of life in the whole world via unprecedented use cases that require high data-rate instantaneous communications, low latency, and massive connectivity for new applications for mobile, eHealth, autonomous vehicles, smart cities, smart homes, and the IoT.

1.2 The explosive growth of wireless communications: Towards Cognitive Radio

The past decade has witnessed an explosive demand of wireless spectrum and that led to the major stress and the scarcity in the frequency bands. Moreover, the radio landscape has become progressively heterogeneous and very complex (e.g. several radio standards, diversity of services offered). Nowadays, the rise of the new applications and technologies accelerates the spectrum scarcity problem and encouraged wireless applications. The new wireless technologies (e.g. 5G) will support high speed data transfer rates including voice, video, and multimedia.
In order to combine, at a very low cost, these future technologies with the existing one, the future network infrastructure should have a sufficient flexibility. The software-defined radio (SDR), a programmable and multi-functional radio that can adapt to a wide variety of services and standards, will facilitate the dynamic air interface reconfiguration of the network nodes by software modifications. Moreover, SDR represents a low-cost power-efficient solution that is becoming more and more apparent in the commercial world and can easily switch from one standard to another (e.g. GSM, EDGE, Wi-Fi, Bluetooth, and LTE) to provide a continuous and high-speed connection.

1.2.1 Software Defined Radio: The core of Cognitive Radio

The **Software-Defined Radio (SDR)** is a radio communication system that can adapt to any frequency band and receive any modulation using the same hardware. SDR is a wireless technology in which hardware platform like

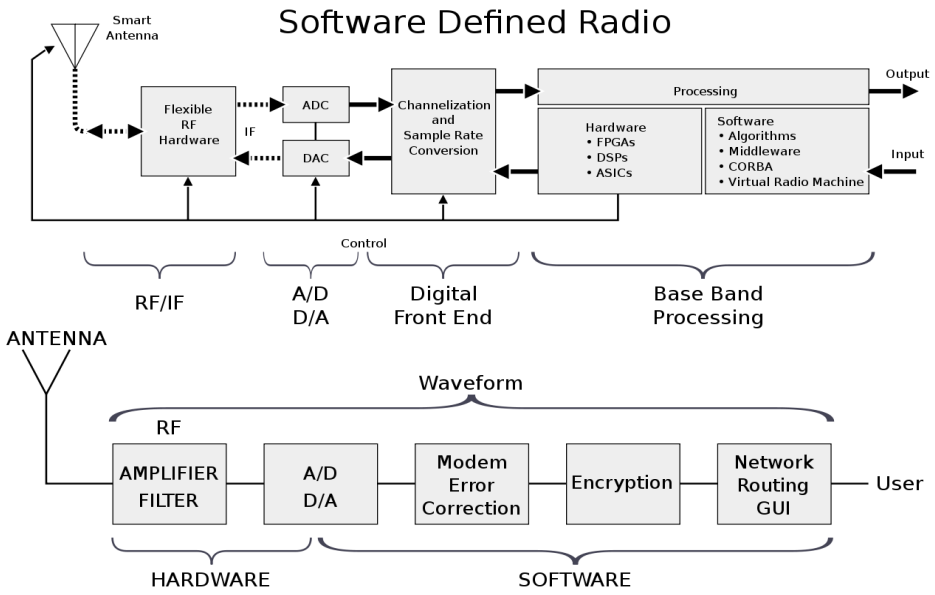

FIGURE 1.2: Software Defined Radio concept that contains: Analogue to digital converter (ADC or A/D), Digital to analogue converter (DAC or D/A) and Hardware system (i.e. Intermediate Frequency/Radio Frequency: IF/RF) [9].

modulators/demodulators, amplifiers, mixers, filters are replaced by a software or firmware operating on programmable processing devices [9].
Besides providing a very cheap radio receiver, SDR devices can easily examine the spectrum, assign of frequency distributions in an efficient manner, detect interferences, analyze the characterization of noise by bands and identify spectrum intruders.

SDR may be used for cellular networks where standards upgrade regularly. Indeed, when using a generic hardware, the upgrade of standards can be easily performed, for instance the shift from the third generation to the fourth one could be successfully achieved by updating the software of SDR and reconfiguring it without changing the hardware.
Several definitions of SDR can be found in the literature [10, 11], in which the common definition ground is that: SDR is a radio communication system in which some parameters, such as transmit power, frequency carrier and bandwidth, can be adjusted in software without changing the hardware itself. One of the important benefits of SDR, for the further technology, is its ability to offer the flexibility, reconfigurability and portability features for the Cognitive Radio (CR).

In [12], some frequency bands are saturated while others are not used depending on time and place. The cognitive radio, therefore, seems to be more and more necessary to homogenize the use of the spectrum and to allow the coexistence of different wireless communication technologies and multitude of communicating objects (e.g. M2M). More generic and powerful reasoning mechanisms are needed to enable radio:

- Adapt to the wide variety of wireless networking technologies (e.g. GSM, UMTS, WiFi, Bluetooth, etc.),
- Consider multiple conflicting goals (e.g. several competitive accesses in the network),
- Select the most relevant opportunity among a multitude of available choices (e.g. the most vacant one or tend to selecting high- quality channels),

1.2.2 The rise of Cognitive radio

The Cognitive Radio (CR) is a new paradigm firstly proposed by Mitola in 1999. One of the main features of CR is the flexibility in which the parameters of the radio (e.g. carrier frequency, power, modulation, and bandwidth) can be modified depending on: the radio environment, the situation, the state of the network and geolocation. The CR can be seen as an extension paradigm of the SDR that helps the decision-making engine to enhance its future decision where the functionality of the SDR is to help the CR to dynamically and autonomously modify its parameters and protocols accordingly [13, 14, 15, 16]. The CR represents one of the important solutions to solve the explosive growth of wireless communications and the spectrum scarcity. Since the last decade, the CR has more and more attracted the attention as a new future technology. J. Mitola introduced in his Ph.D. the definition of the CR as follows: "The cognitive radio identifies the point at which wireless personal digital assistants (PDAs) and the related networks are sufficiently computationally intelligent about radio resources and related computer-to-computer communications to:

- Detect user communication needs,
- Provide radio resources and wireless services most appropriate to those needs."

Generally, there are two types of users in a CR system: Licensed (Primary Users: PU) and unlicensed (Secondary Users: SU) users. The spectrum sharing among licensed and unlicensed users can be classified into three main models:

- Underlay Access
- Overlay Access
- Interweave Access

Underlay Access

In the underlay access, rather than detecting holes in the frequency band, the SU can access and transmit his data simultaneously with the PU. In other words, PU and SU are authorized to coexist in the same frequency band only if the interference is controlled and managed properly. Indeed, the SU, on the one hand, should use his cognitive ability by controlling his transmission power in order to keep the interference with the PU under a certain limit. On the other hand, primary transmission may be able to withstand some degrees of harmful interference. The average of the harmful interference at the primary receiver, also known as interference threshold, must not exceed the interference level defined by the FCC as interference temperature [17]. In order to keep the interference with PU under a certain threshold, underlay systems use ultra-wideband (UWB) transmission where the SU can spread his transmission signal on a very large bandwidth bands. In this case, the transmission power is below the power of the PU, and this fact limits the interference with the latter [18]. Underlay access suffers from various drawbacks such as: it can be used only for a short transmission range (i.e. limited transmission power can be reached), and requires high level of cognition in order to measure the interference power with the PU.

Overlay Access

Unlike Underlay, the Overlay access authorizes SUs and PUs to transmit simultaneously at a maximum power transmission without interference. Indeed, in Overlay access, the SU should not cause any harm to the PU's QoS because there is a high cooperation level between them. Many works focus on the overlay access in which the SU plays the role of a relay to ensure the transmission of the PU's data and to enhance the transmission rate of the latter [19, 20].

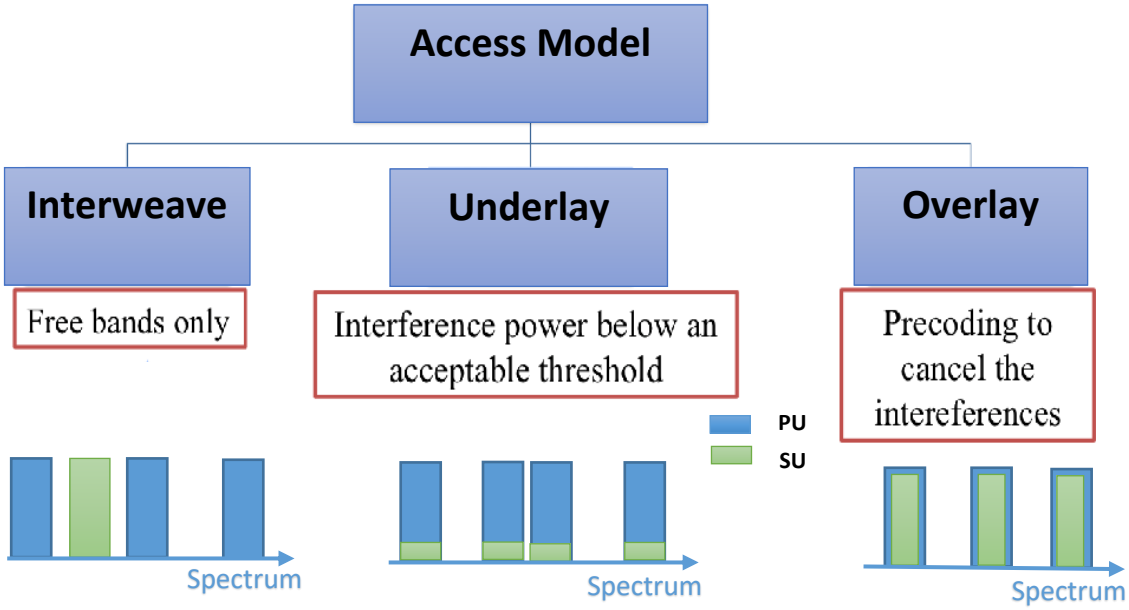

FIGURE 1.3: Spectrum access according to the three possible models: Interweave, Underlay or Overlay.

Interweave access

In this method, there is no constraint on the transmission power of a SU except that he should limit his transmission to available spectral holes. Indeed, the SU should accurately notice the unused frequency band and survey the activity of the PU in order to prevent any interference with the licensed users. In other words, a SU can exploit the spectrum holes in an opportunistic manner. Upon the localization of a spectrum hole, he can transmit his data with the maximum authorized power level.

In our work, we focus on the Interweave Access where the SU can access, in an opportunistic manner, the unused frequency band without any cooperation with the PU. This technique refers to the Opportunistic Spectrum Access (OSA) in the context of CR which is the most attractive way to access the spectrum [21, 22, 23, 24].

1.2.3 The Cognitive Cycle

A cognitive radio, based on intelligent observation of its environment, should make the best decision. However, it is impractical for a cognitive radio (also called Secondary User: SU) to observe wide frequency bands to access the

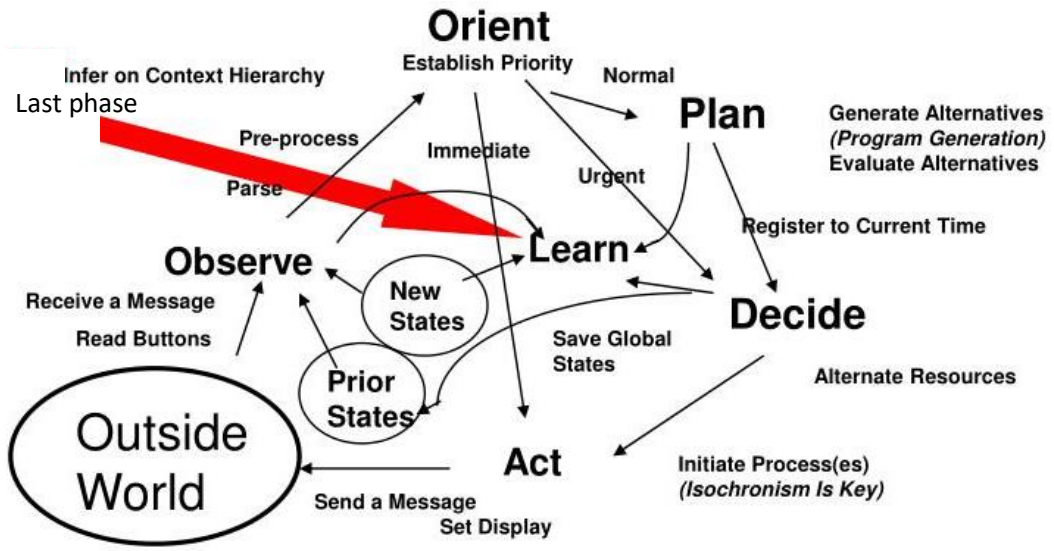

FIGURE 1.4: Cognitive cycle decision-making as introduced by Mitola [25]

best one (e.g. the band with the highest availability, quality, etc.). Subsequently, under a partial observation (e.g. one band at each time slot), SU should make the best decision in order to improve his communication performance in the most prevailing conditions.

According to Mitola [25], in a cognitive cycle, the user has the capacity to collect some information from his environment (observation), and then analyze them in order to make the best action. As shown in Fig. 1.4 proposed by Mitola, at each time slot, the cognitive radio executes six steps: Observe, Orient, Plan, Decide, Act and Learn. In the observation phase, the cognitive radio discovers its environment using different sensors (e.g. the state or the quality of the selected channel, the interference level). In the orientation phase, the cognitive radio determines its priorities, makes plans according to the characteristics of the environment found in the observation phase. In the decision phase, the cognitive radio chooses and selects the best candidate plans, and finally makes actions during the act phase (e.g. transmits the data, reconfigures its parameters, etc).

To finish the cognitive cycle, the cognitive radio should execute a last step that represents the learning phase in order to enhance the decision issues. This step can improve the future action of cognitive radio based on its past perceptions, observations, decisions and actions.

A decision-making process should consider certain parameters such as: spectrum, modulation scheme, power specification, data rate, etc. Consequently, the decision about the availability or the quality of a spectrum should be

made to access a particular spectrum bandwidth. Finally, the best decision-making of cognitive radio should depend on its environment parameters, and on the nature of interaction between users (i.e. cooperative or competitive behavior).

1.3 Our Motivations

1.3.1 Our Research Objectives

This research focuses on the decision-making in the context of cognitive radio. We mainly focus on the learning and decision-making in the cognitive cycle. The main goal of the thesis is to develop algorithms to help the cognitive radio to make an efficient decision in order to maximize a utility function (e.g. data rate, throughput capacity, etc.). Our work also considers both cooperative and competitive network trying to increase the transmission time and decrease the number of collisions among users. Moreover, we focus on some features of networks with the priority access.
Our major contributions can be summarized as follows:

- Focus on learning algorithms capable to evaluate the vacancy probabilities of channels.
- Consider multi-cognitive users.
- Determine the quality of service.
- Investigate the convergence.
- Investigate competitive or cooperative learning policies.
- Consider networks with priority access.

1.3.2 Outlines

Chapter 2 introduces the Multi-Armed Bandit game and presents its major areas of applications. It also presents briefly the Dynamic Spectrum Access (DSA) and the Opportunistic Spectrum Access (OSA), a subset of DSA. OSA allows a cognitive user to exploit opportunities in the frequency bands and enables the coexistence of both cognitive and licensed users. In addition, it focuses on the self-adaptation and learning in wireless networks and discusses the design space and networking algorithms.

In chapter 3, we evaluate the performance of the existing MAB learning algorithms suggested for a single user. We also study the performance of MAB algorithms in the multi-user case. The main contributions introduced in chapter 3 are:

- The mathematical MAB model.
- The performance of existing MAB.
- Extension of MAB algorithms to consider multi cognitive users.
- A novel cooperative policy to manage a secondary network.

In chapter 4, we propose two MAB algorithms called AUCB and e-UCB that can achieve better results compared to the well-known MAB algorithms, such as: Thompson Sampling, UCB and e-greedy. These two algorithms can help the cognitive user to increase his throughput capacity. Thus, the main contributions in chapter 4 are briefly summarized as follows:

- AUCB and e-UCB to enhance the spectrum learning of the cognitive user.
- Extend AUCB and e-UCB to consider multiple priority users.
- Investigate the performance of AUCB and e-UCB
- Investigate the analytical convergence of the proposed algorithms.

In chapter 5, we analyze the sequential decision-making applied in OSA. We introduce the system model for multiple users. We also study the quality and the availability of channels in order to optimize the performance of communication and enhance the quality of service.
Hereinafter, we investigate the competitive cognitive users' behavior in the priority cognitive networks for which two competitive policies, All-Powerful Learning (APL) and Priority Fairness Access (PFA), have been proposed. Those latter help the cognitive users to learn separately the vacant probabilities of channels and decrease the number of collisions among them. Our policies provide a steady state where each user has no interest to change its action since shifting to another action may affect his award. The main contributions in chapter 5 are briefly summarized as follows:

- Propose APL for the priority access.
- Investigate the convergence proof of APL.

- Evaluate the performance of AUCB and e-UCB with APL and compare to several recent works.
- Study the Quality of Service where the user should estimate the availability and quality of channels.
- Propose PFA to tackle two or more priority levels.

Finally, chapter 6 concludes the thesis by summarizing our contributions, providing some discussions of the main results, and suggesting future research tracks based on this thesis.

Chapter 2

Learning and Decision-Making for Opportunistic Spectrum Access

Contents

2.1 Introduction

In this chapter, we introduce the state of the art of the decision-making and learning mechanisms for the Opportunistic Spectrum Access (OSA). OSA, a particular case of Dynamic Spectrum Access (DSA), represents one of the proposed solutions to better use the frequency bands.
The main idea of OSA is to share the spectrum between Primary Users (PUs) with a privilege access to their bands and Secondary or opportunistic Users (SUs), who can access these bands when PUs are not active.
The outline for this chapter is as follows: In section 2.2, we discuss the concept of DSA in CR and the main challenge of the SU in OSA. In section 2.3, we investigate the spectrum decision and the machine learning classification. In this section, we also introduce the reinforcement learning and how it can help the SU make an optimal decision. In section 2.4, we introduce the Multi-armed bandit problem and we formulate the OSA as MAB in order to enhance the spectrum learning. In section 2.5, we introduce the different existing models in MAB that are used to formulate the obtained reward such as Independent Identical Distributed (IID) or Markovian Models. In section 2.6, we present the state of the art of MAB algorithms for single or multiple agents. Finally, section 2.7 summarizes this chapter.

2.2 Dynamic Spectrum Access and Opportunistic Spectrum Access

The radio spectrum can be classified into two different categories:

- Static spectrum used in various standards on the wireless spectrum where licensed users have a full and privilege access to a specific band.
- Dynamic access in which the spectrum can be wisely shared between users to reduce the time where any band is free.

However, Dynamic Spectrum Access can be seen as a new concept to maximize the spectral efficiency. Several studies in the United States (US) have shown that some frequency bands have become more and more crowded while others are almost not used [26]. As shown in Fig. 2.1, DSA can be categorized into three main models: Exclusive Use, Open Sharing and Hierarchical Access [27]:

FIGURE 2.1: Classification of the DSA where three models are suggested: Exclusive Use Model, Open Sharing Model, and Hierarchical Access Model.

- In Exclusive Use Model, the frequency bands are allocated to the exclusive use where a cognitive user can access them under some constraints. Indeed, in the Spectrum Property Rights, the licensed user (i.e. PU) is authorized to sell or trade a part of his frequency bands. This model encourages the opening of a secondary market. The dynamic spectrum allocation, introduced by the European DRiVE project, aims to improve the spectrum efficiency by exploiting the spatial and temporal traffic statistics of different services.

- Open Sharing Model is referred to open new portions of the frequency bands for unlicensed users with the same rights to access the spectrum.

- In Hierarchical Access Model, a secondary network is introduced in which a SU may access the spectrum as far as he does not cause any harmful interference to the PU. As mentioned before [Chapter 1, section 2], there are three Hierarchical Access Models: Interweave, Underlay and Overlay access.

In practice, Interweave is the widely used model that allows SU to access the unused frequency bands, in opportunistic manner, without any cooperation with the PU. Due to the detection cost, the SU is not able to sense simultaneously all the available frequencies. Subsequently, under a constraint detection (e.g. one channel at each time slot), the SU should make a decision and find an opportunity to send his data. In our work, we are interested to apply the MAB algorithm in OSA for which several algorithms are suggested in the literature to help SU make a decision. SU in OSA faces many challenges in order to reduce the interference with PU [28]:

- Searching for unused bands using **Spectrum Sensing** techniques.

- **Spectrum Decision**: Selecting the band with the highest vacancy probability.

- **Spectrum Mobility**: Evacuating a previously selected channel when PU reappears.

- **Spectrum Sharing**: Sharing available frequency bands with other SUs.

In our study, we choose to focus on two challenges: **Spectrum Decision** and **Spectrum Sharing**.
To help the SU make a good decision, the game theory is recently applied in CR and represents a suitable solution for: adjusting the transmission power,

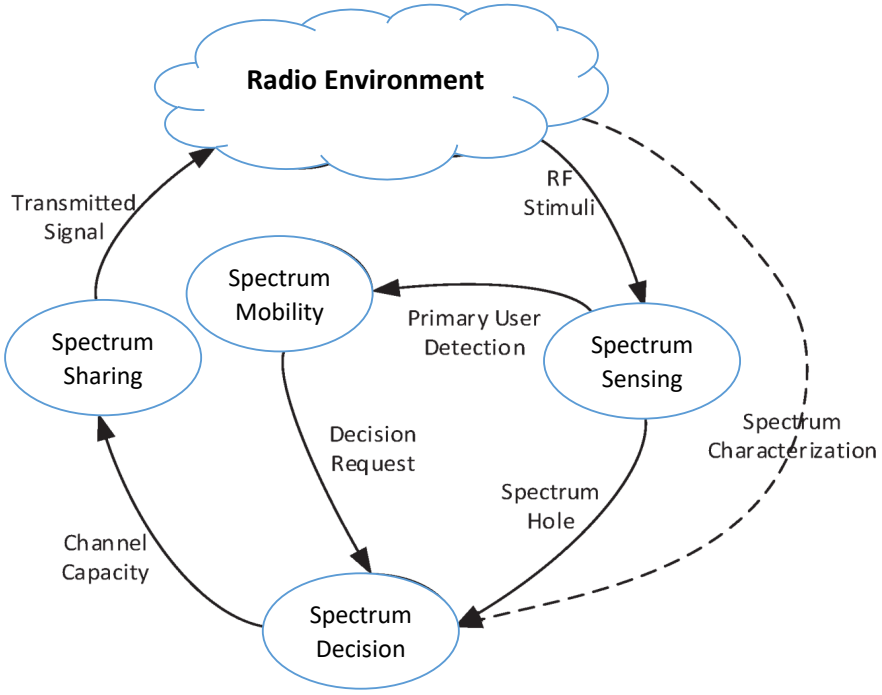

FIGURE 2.2: Cognitive radio decision making as introduced in [28].

management spectrum, allocating resources, etc. The most used games in the CR are prisoner's Dilemma, Stackelberg and Multi-armed Bandit (MAB). In our work, we mainly focus on the Multi-armed Bandit problem in which we formulate the OSA as a MAB problem which is widely used in the literature. For multiple SUs, the extended MAB should reach an equilibrium among users, use optimally the frequency bands and decrease the number of collisions among users.

2.3 Decision-Making Process and Machine Learning

Spectrum Decision is the ability to find and access the most suitable frequency band with the highest availability probability. However, in the context of CR, the main target of a SU is to increase his transmission time and rate by finding the most vacant band. At each time slot, the SU can observe the state of the frequency bands, decide which band to access, and make an action in order to be adapted to his environment (e.g. transmitting power, frequency band). To evaluate the taken action, good or bad, a SU receives a reward, the goal being to enhance its behavior at the next stage. Thereby, a SU should be a cognitive device in order to learn its environment from

past decisions and make the best action. Recently, there is a growing interest in applying machine learning in the context of cognitive radio in order to achieve an efficient resource allocation [29, 30]. However, in the cognitive cycle, the learning and the comprehensive situation represent the more important functions in order to help the SU adapt to his environment.
Generally, the machine learning problem can be defined as in [31]: "An agent learns about his environment with respect to a task A (e.g. his action to the environment at instant t) and measures the performance obtained from this task r (e.g. reward at instant t) if his performance at task A, measured by r, is improved over time".
The machine learning algorithms, as shown in Fig. 2.3, are classified into three main categories[31, 32]:

- Supervised learning: The agent can create a function model based on the input data where the desired output is known. Then, for new data input, and based on the generated function, the agent is able to predict the future output. The supervised learning algorithm can be classified into two main categories: Classification or Regression problems. In the classification problem, the data variable is considered as discrete, and the objective is to identify the classes of the data. The regression algorithm is very useful to predict a real value quantity for an input data, such as: Linear regression, logistical regression, perception.

- Unsupervised learning: Unsupervised model is widely used to draw inferences or extract features and patterns from the dataset, without any feedback. The most important unsupervised learning methods are clustering, k-Means, Self-organizing maps, etc.

- Reinforcement learning (RL): It represents another part of machine leaning and has an important role in the multi-agent system, since it allows an agent to learn from his environment by interacting with it. The agent should enhance his behavior from the feedback (e.g. reward). Indeed, RL may allow an agent to adapt to his environment by finding a suitable action to reach the best reward. Several variants of RL can be found in the literature such as Q-learning, Upper Confidence Bound (UCB), ϵ-greedy, etc.

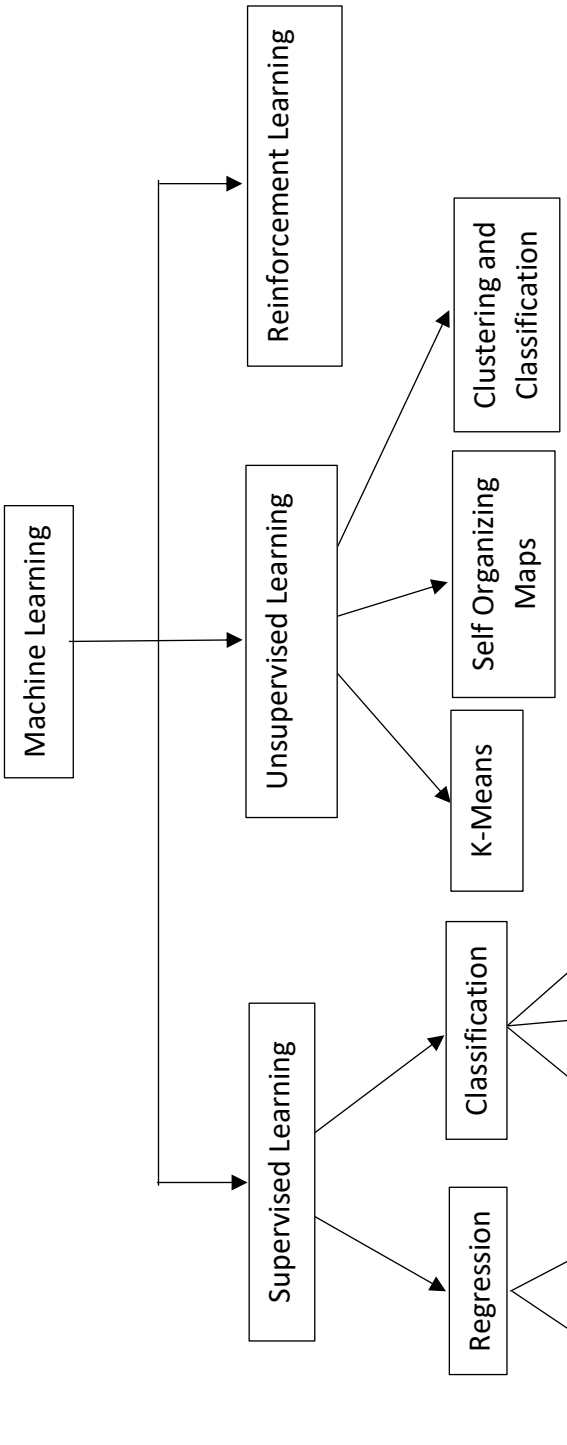

FIGURE 2.3: Machine learning algorithms.

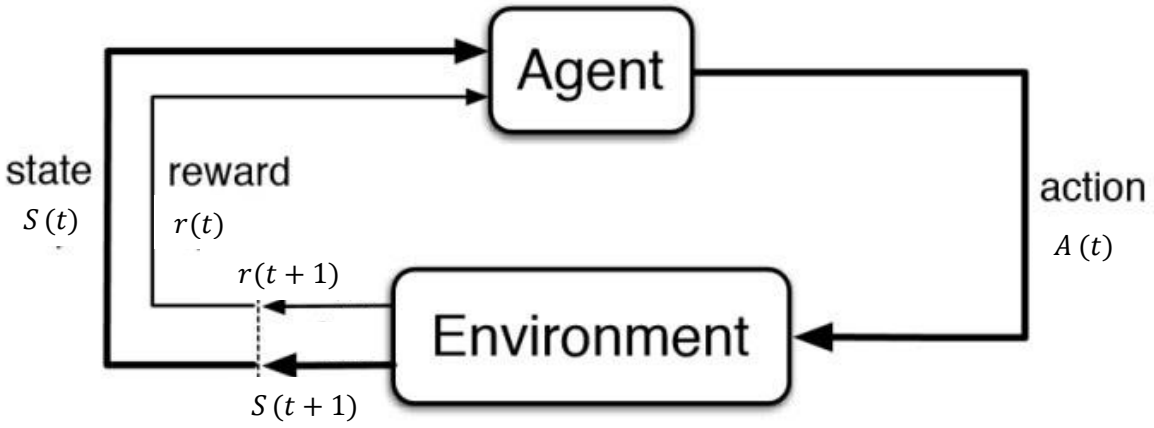

FIGURE 2.4: Reinforcement Learning Framework

2.4 Reinforcement Learning

In the Reinforcement Learning (RL), the agent can maximize his reward without any prior information about his environment. However, by memorizing the states of an environment or the actions he took, the agent can make a better decision in the future. The reward feedback, also called reinforcement signal, has an important role to help an agent to learn from its environment. As shown in Fig. 2.4, at each time slot, the agent observes the state of his environment $S(t)$, selects an action $A(t)$ and receives a reward $r(t)$. Based on the expected reward obtained up to time slot t and the state of the environment, the agent tries to enhance his action at the next slot.

RL is widely used in several domains: Robotics, Aircraft control, self-driving cars, Business strategy planning, etc. It was first developed for a single agent who should find an optimal policy that maximizes his expected reward knowing that the optimal policy depends on the environment. Unlike the case of a single agent, for multiple agents, the optimal policy depends not only on the environment but also on the policies selected by other agents. Moreover, when multiple agents apply the same policy their approaches in such systems often fail because each agent tries individually to reach a desired result. In other words, it is impossible for all agents in a certain system to maximize simultaneously their personal reward, although find an equilibrium for the system representing a point of interest. Subsequently, it is important to find a policy for each agent in order to guarantee the convergence to an equilibrium state in which no agent can gain more when modifying its own action.

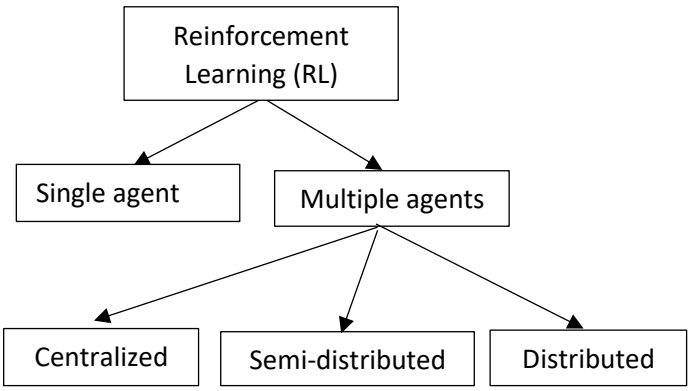

FIGURE 2.5: Taxonomy of Reinforcement Learning

Recently, the multi-agent in the reinforcement learning attracts more and more the attention in several fields where multiple agents may exist in the system, such as: robotics, distributed control, telecommunications, economics, etc. For a given system, the agent may independently find his own good policy, without coordinating with other agents neither their reward nor their actions. Fig. 2.5 shows all the possible reinforcement learning frameworks for a single or multiple agents. In a centralized algorithm, the decision is made at the network level, while in a distributed algorithm, each agent makes his own decision independently without any cooperation with others. A semi-distributed algorithm is based on the combination between centralized and distributed techniques.
In RL, Exploitation-Exploration dilemma represents an attractive problem. In order to maximize his performance (exploitation), the agent should gather some information about his environment (exploration). This is known as the Exploration-Exploitation dilemma in the reinforcement learning. If the agent spends a lot of time on the exploration phase, then he cannot maximize his reward. Similarly, when the agent focuses on the exploitation phase by exploiting his current information, then he may miss the best action that leads to the highest reward. Thus, the agent needs to balance the tradeoff between Exploration and Exploitation in order to obtain an appropriate result.

2.5 Multi-Armed Bandit problem

Due to its generic nature, Multi-armed bandit game attracts more and more attention and is widely used to solve the decision-making problem in several fields such as: Jamming communication, Clinical trials, object tracking, ads

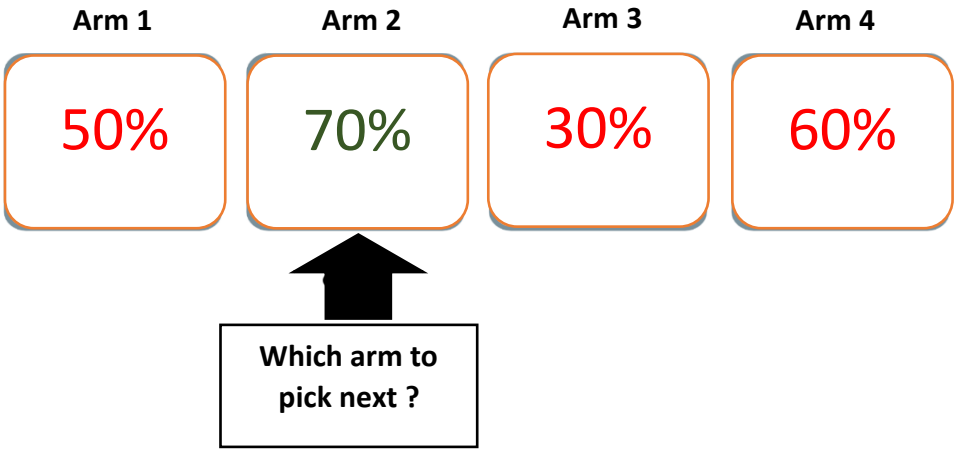

FIGURE 2.6: Several Arms with different expected reward. After a finite period of time the agent has a perception about the reward obtained from each arm.

selection on web pages.
The Multi Armed-Bandit problem represents a subset of RL in which an agent is in front of several slot machines (or arms), and at each time slot he pulls an arm and receives a fixed reward. Like most RL frameworks, the agent starts the game without any priori knowledge about the expected reward of the arms. The main goal of the agent is to find the arm with the highest expected reward. Here, we should define two classes of arms:

Optimal arm: This arm has the highest expected reward and is represented by the arm 2 in Fig. 2.6. The agent tries to reach this arm in order to maximize his expected reward.
Suboptimal arms: Include all other arms considered as non-optimal. Efficient MAB algorithms should be able to limit playing with suboptimal arms.

Fig. 2.7 presents all MAB variants in which several assumptions can be considered: Single or multiple agents, model of rewards, number of arms, with or without prior knowledge, etc.

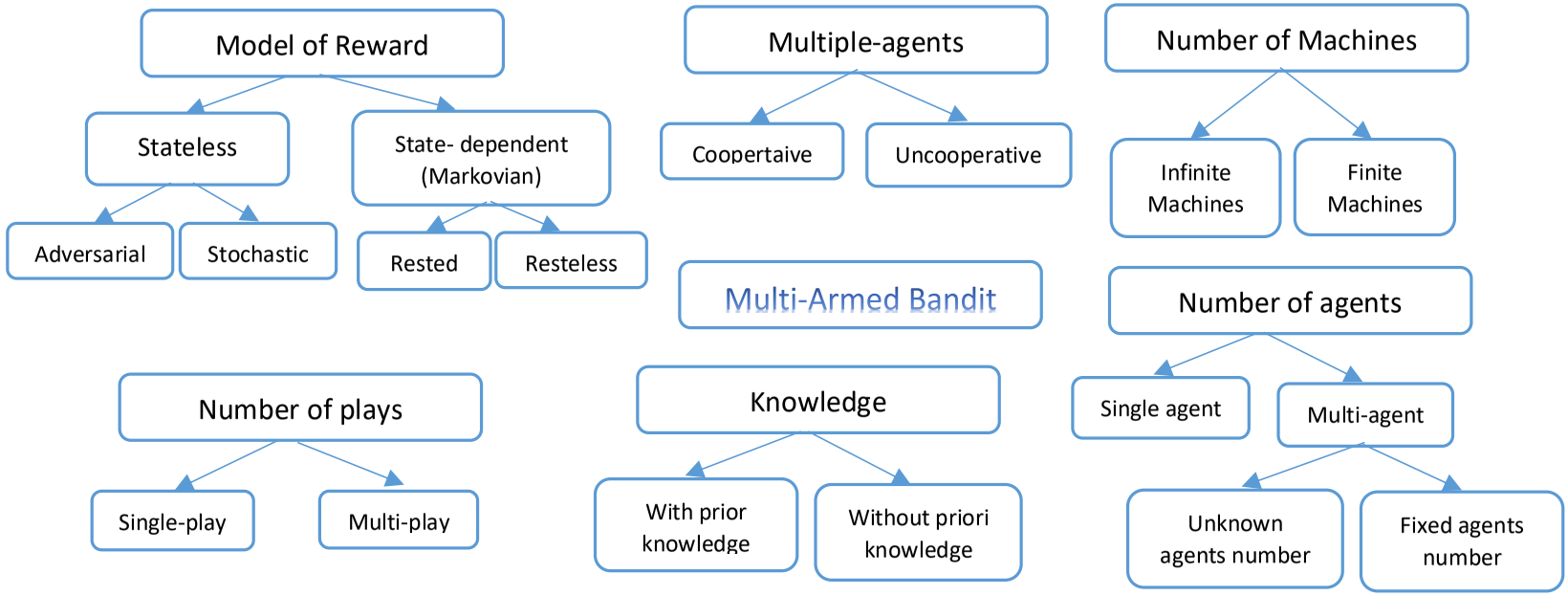

FIGURE 2.7: Classification of MAB problem.

2.6 Model of MAB reward

In order to formulate the obtained reward from the slot machines, different models exist in the literature such as: Markovian or IID. In this section, we focus on these two models, being the most widespread, in order to address the MAB problem.

2.6.1 IID Bandit Problem Formulation

Due to its simplicity, this model has received much attention to formulate the reward obtained from different arms [33, 34, 35, 36, 37]. The paper of Lai and Robbins in [38] presents one of the earliest studies to solve the MAB problem based on the Independent Identical Distributed (IID) model. In their study, the authors considered C arms and showed that the expected regret, i.e. the loss of reward by selecting non-optimal arms, achieves a logarithmic asymptotic behavior with respect to the number of plays. In [39], Anantharam et al. extended the work of Lai and Robbins while considering that the agent can play simultaneously N arms ($C > N$) and they also proved a logarithmic regret. Using IID model, the authors of [40] proposed a simple algorithm called UCB1 (Upper Confidence Bound) for a single agent inspired from the work presented in [38].

At each time slot t, the agent selects an arm $i \in \{1, \ldots, C\}$ based on the past observations and obtains an IID reward $r_i(t)$. The i-th arm can be observed in one of the two binary states $S_i(t)$: $S_i(t) = 1$ if the i-th channel is free and 0 otherwise. Without any loss of generality, we can consider that: $r_i(t) = S_i(t)$. The main target of the agent is to maximize his long-term reward. Subsequently, the optimization problem consists in maximizing the following quantity:

$$\max \sum_{t=1}^{n} r_i(t) \tag{2.1}$$

where n stands for the total number of slots. The mean reward of the i-th arm up to slot n can be defined as follows:

$$\mu_i = \frac{1}{n} \sum_{t=1}^{n} r_i(t) \tag{2.2}$$

Based on the IID model, several works recently consider the case of multiple agents [35, 36, 37] which represents a more realistic case in OSA. However,

in OSA, several SUs may exist in the network and their main objective is to learn collectively the availability of channels in order to decrease the number of collisions among them.

2.6.2 Markov Bandit Problem Formulation

In MAB problem, the received arms reward can be formulated as a Markov model where each arm has finite states space and each state provides a stationary and positive reward. Moreover, the state of the arms changes according to a stochastic process that is referred to as a Markov process. In other words, in a Markov process, the future state of the arm is related to its current as well as its previous state and to the transition probability P. In a Markov model, two modes of arms can be considered: active and passive. However, when an arm is selected by the agent during the current time step, it is considered in active mode, otherwise it is referred as passive. In Markov MAB, two sub-models can be considered (as shown in Fig. 2.7):

- **Rested MAB model**: In this model, only the state of the selected arm evolves, while the state of other arms stays frozen, i.e. does not evolve.
- **Restless MAB model**: In such model, the states of all arms continue to evolve at each time step. This model was first proposed by Whittle in 1988, in which the passive arms evolve regardless of whether there are played or not. The restless model is also sufficiently suggested in the literature and several works such as [41, 42] considered this model under the multi-user case.

Under Markovian model, each arm is modeled as aperiodic, irreducible and discrete time with finite state $S_i(t)$. In Fig. 2.8, we consider two binary states (e.g. free or busy) for each arm with a Markov chain characterized by:

- Transition matrix of probability (P_i).
- Transition diagram (Fig. 2.8).
- Stationary vector of the states (π_i).

Let P_i^{kl} be the transition probability for the i-th arm from state k to state l, then the transition matrix, P_i, of Markov model for two states becomes:

$$P_i = \begin{pmatrix} P_i^{00} & P_i^{01} \\ P_i^{10} & P_i^{11} \end{pmatrix}$$

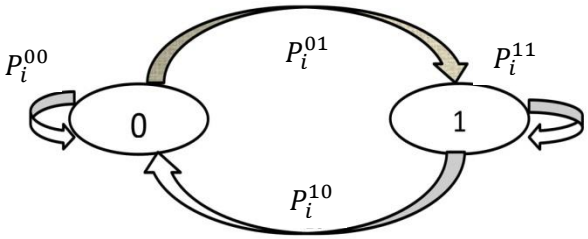

FIGURE 2.8: Channel occupancy model

For simplicity reasons, assuming $P_i^{00} = P_i^{10}$, and $P_i^{11} = P_i^{01} = \phi_i$, knowing that $P_i^{00} + P_i^{01} = P_i^{10} + P_i^{11} = 1$, then the matrix P_i can be simplified as follows:

$$P_i = \begin{pmatrix} 1-\phi_i & \phi_i \\ 1-\phi_i & \phi_i \end{pmatrix}$$

The stationary vector defines the mean rewards for a transition matrix: $\pi_i = [\pi_i^0, \pi_i^1]$, where

$$\pi_i^0 = \frac{P_i^{10}}{P_i^{10} + P_i^{01}} = P_i^{10}, \text{ and } \pi_i^1 = \frac{P_i^{01}}{P_i^{10} + P_i^{01}} = P_i^{01}$$

Consequently, the mean reward for each arm i is given below:

$$\mu_i = \frac{1}{n} \sum_{t=1}^{n} \sum_{b \in S_i} \pi_i^b r_i^b(t) \tag{2.3}$$

The main goal of the agent is to estimate μ_i in order to access the best channel with the highest vacancy probability. Based on [43], recent works in [44, 45, 46] tackle the Markov MAB for multiple agents, that achieves a logarithmic regret with respect to time slot. In our work, we focus on the IID model in which the well-known MAB algorithms proposed to address the MAB problem are based on the IID model.

In the following section, we present the MAB algorithms for a single user to tackle the MAB problem; and latter, we focus on MAB with multiple agents in order to apply it in OSA.

2.7 MAB algorithms

In this section, we discuss the state of the art of the proposed algorithms to tackle the IID MAB problem for a single or multiple agents cases. In MAB problem, an agent faces the exploration vs the exploitation dilemma. As a matter of fact, the agent may fail to pull the optimal arm, and thus should gather more information about the expected reward of arms to select the optimal one (exploration). On the other hand, the agent tries to select the current optimal arm in order to maximize his gain (exploitation).

2.7.1 MAB algorithms for a single agent

In the literature, a variety of MAB algorithms have been proposed to tackle the exploration-exploitation dilemma for a single agent. e-greedy represents the simplest MAB algorithm and was initially proposed in [47]. In e-greedy, the agent pulls a random arm with a constant probability e (exploration), and then pulls the arm that has the highest expected reward with a probability of $1 - e$ (exploitation). As a result, we conclude from e-greedy that a higher value of e may increase the exploration epoch and thus the total regret. Moreover, e-greedy does not achieve an asymptotic convergence to the optimal arm since e is considered as a constant. To solve the asymptotic convergence, the authors of [43] proposed the decreasing e-greedy in which e is not considered as a constant but a parameter decreasing with time. However, when e tends towards 0, the agent may have a good estimation about the expected reward of each arm and regularly accesses the optimal one. The authors also found the analytical convergence of the decreasing e-greedy that achieves a logarithmic asymptotic behavior. The main drawback of the decreasing e-greedy is that the exploration and exploitation epochs are independent, since the agent should stop the exploration after a finite period. Subsequently, the agent cannot follow a dynamic environment in which the mean reward of arms changes over time. Unlike e-greedy, Upper Confidence Bound (UCB) makes a tradeoff between exploration and exploitation at each time slot [48]. UCB has strong performance guarantees and represents the widely used algorithm to address the IID MAB problem. Lai and Robbins in their paper [48] have shown that the regret of UCB achieves a logarithmic convergence with respect to time. Subsequently, after a finite number of time slots, and based on UCB, the agent may find the best arm by maximizing his reward. Several versions of UCB have been proposed to achieve a better performance compared to the classical one [40, 49, 50, 51]. In [43], the authors proposed UCB1

in order to achieve a balance between exploration and exploitation epochs. In UCB1, each arm assigned an index $B_i(T_i(t))$ where $T_i(t)$ is the number of times to play with the i-th arm. The index $B_i(T_i(t))$ of the i-th arm up to slot t can be defined as follows:

$$B_i(T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t))$$

where $X_i(T_i(t))$ and $A_i(T_i(t))$ represent the exploitation and exploration epochs respectively. At each time step, the agent makes an action a_t and pulls the arm with the highest index $B_i(t, T_i(t))$ as follows:

$$a_t = argmax(B_i(t, T_i(t)))$$

After a finite time step, the agent may have a good estimation about the expected regret of each arm, and then he regularly pulls the optimal one. In the same study, the authors also proposed two other versions of UCB, called UCB2 and UCB-tuned, achieving both better results than UCB. These two versions have also the exploration-exploitation factors $X_i(T_i(t))$ and $A_i(t, T_i(t))$. In UCB2, the agent selects at each time step the arm that has the highest index $B_i(t, T_i(t))$ and then pulls this arm $[(1+\alpha)^{r_i} + 1(1+\alpha)^{r_i}]$ times where α is a constant ($0 < \alpha < 1$) and r_i denotes the number of epochs the arm i is selected. Moreover, the effect of the exploration factor, $A_i(t, T_i(t))$, is reduced compared to UCB1 in order to minimize the exploration epoch and converge towards the optimal arm.
Unlike UCB1 and UCB2, in UCB-tuned, the exploration factor depends on the variance of the reward obtained from the i-th arm while achieving better results. Similarly to UCB-tuned, and based on the variance and the expected mean of arms reward, the authors of [52] proposed a novel version of UCB called UCB-Variance (UCB-V) that achieves a better regret compared to UCB1, UCB2 and UCB-tuned.

Recent versions of UCB, such as Kullback-Leibler-UCB (KL-UCB) [53] and Bayes-UCB [50], have been proposed in order to reduce the total regret compared to classical versions of UCB. KL-UCB considers the distance between the estimated expected rewards of the arms as a factor in order to discriminate the optimal arm. Unlike KL-UCB, Bayes-UCB is a Bayesian model in which each arm is characterized as an estimate of a distribution $O = \{O_1, \ldots, O_C\}$ that is drawn from a priori distribution. Another important Bayesian algorithm is referred to Thompson Sampling (TS). Because of its optimal regret and excellent performance that can exceed the state of the

MAB algorithms	Computation complexity	Experimental performance	Asymptotic convergence	Upper bound of regret
e-first	|	||	N/A	N/A
e-greedy	|	||	N/A	N/A
Decreasing e-greedy	|	||	Yes	Log(n)
Lai and Robbins	|||	N/A	Yes	N/A
Agrawal	|	|||	Yes	N/A
UCB1	|	|||	Yes	Log(n)
UCB2	||	|||	Yes	Log(n)
UCB-Tuned	||	|||	N/A	N/A
KL-UCB	||	|||	Yes	Log(n)
Bayes-UCB	||	||	Yes	Log(n
Exp3	||	||	Yes	$\sqrt{n}$
Thompson Sampling	|	|	Yes	Log(n)

TABLE 2.1: Comparison between the well-known MAB algorithms. '|' indicates a high performance, '|||' indicates that the performance is weak and 'N/A' indicates that the information is not available.

art MAB algorithms, TS attracts more and more the attention of the machine learning community [49, 54, 55]. Recently, several studies have found a concrete bound of its optimal regret [56, 57, 58]. More details and discussion about TS and its performance are introduced in the next chapter. Moreover, we adopt TS in chapter 4 as a reference in order to evaluate the performance of our proposed AUCB and *e*-UCB.

Table 2.1 compares the performance of the three general families of the MAB algorithms: Thompson Sampling, UCB and *e*-greedy.

2.7.2 MAB algorithms to manage multiple agents

To formulate the OSA as a MAB problem, recent works extend the simple case of MAB (i.e. the case of a single agent) to consider several agents [35, 59, 60, 61, 62]. In our work, we are interested in the OSA for multiple priority access in which SUs should access the spectrum according to their ranks. Moreover, decreasing the number of collisions among SUs represents a point of interest to enhance the global performance of the secondary network. In general, when two SUs access the same channel to transmit, their data cannot be correctly received because of the interference between them.
When a collision occurs among users, several proposals can be found in the literature in order to enhance their behavior in the next slots. We present below two well-known collision models in the literature that are widely used in OSA:

- ALOHA-like model: If a collision occurs between two or more users, then none of them receives a reward, despite the selected channels is free. This model may ensure the fairness among users, and no collision avoidance mechanism is used.

- Reward sharing model: If two or more users select the same channel at the same time, the colliding users share the obtained reward from the selected channel (each of them receives the same reward).

The above models can affect the methodologies used to collect the reward from the target channel while the learning phase is not affected. In our work, we consider the most widely used, ALOHA-like.
Based on the ALOHA-like, the works of [35, 59, 60, 62, 63, 64, 65, 66] proposed semi-distributed and distributed algorithms in which users cannot exchange information with each other. Liu and Zhao in [59], proposed Time-Division Fair Share (TDFS) policy and showed that the proposed algorithm

MAB for multiple agents	Learning ALgorithm	Experiment Performance	Asymptotic Convergence	Upper Bound
Centralized	Multiple Play [39]	N/A	Yes	N/A
	Auction problem [67]	I	N/A	N/A
Semi-distributed	Cooperative [68]	I I	N/A	N/A
	Bipartie matching [69]	I I I	N/A	$\log^2(n)$
	Leader-follower[70]	I I I	Yes	N/A
Distributed	Random Rank [35]	I I	Yes	$\log(n)$
	TDFS [59]	I I	Yes	$\log(n)$
	SLK [61]	I I I	Yes	$\log(n)$
	Musical Chair [60]	I	Yes	$\log(n)$
	MEGA [63]	I	Yes	$\log(n)$

TABLE 2.2: Compares the performance of the well-known MAB algorithms for multiple agents. 'I' indicates a high performance, 'I I I' indicates that the performance is weak and 'N/A' indicates that the information is not available or no value available.

may achieve an asymptotic logarithmic behavior. In such algorithm, the users access the channels with different offsets. TDFS also ensures the fairness among users; while in our work we are interested in the priority access where users access the channels based on their prior rank. In [59], TDFS policy was been used to extend UCB1 algorithm to consider multiple users. Beside TDFS, the authors of [35] proposed Random Rank policy, based on UCB1, to manage the secondary network. Random Rank represents a distributed policy (i.e. no-information exchange among users) in which the user achieves a different throughput.

Musical Chair in [60] represents another important policy for the random access to manage a secondary network. During time T0, the user selects a random channel to have an information about the channels availability (exploration), after that, the user accesses the channels according to his priori rank (exploitation). Similarly, the authors of [63] proposed Multi-user e-greedy collision Avoiding (MEGA) algorithm for the random access. The importance of MEGA algorithm is that it can handle the random dynamic access in which the users can enter or leave the network.

Generally, the random access is widely suggested in the literature while in our work, we focus on the priority access in which there are few studies. The k-th largest expected rewards (SLK) in [61] is one of rare algorithms

proposed for the priority access based on UCB1. Table 2.2 compares the performance of the most popular MAB policies for the random access or priority access, that can be generally classified into three categories: Centralized, Semi-distributed and Distributed.

2.8 Conclusion

Opportunistic Spectrum Access (OSA) in Cognitive Radio (CR) represents a reliable solution to achieve an efficient use of the frequency bands. OSA in CR allows unlicensed users (or Secondary Users: SUs) to access a portion of the licensed frequency bands without causing any harmful interference to the licensed users (also called Primary Users: PUs).
First, we investigated the decision-making that represents one of the main challenges faced by SUs in OSA. Then, we presented the machine learning taxonomy and more precisely the Reinforcement learning to address the decision-making problem. Next, we focused on the MAB problem, as a part of reinforcement learning, that is widely suggested in OSA in order to help the SU make a good decision.
We continued to present the state of the art of MAB algorithms applied in OSA for a single user. We compared the performance of existing algorithms by comparing some features: Computation complexity, experimental performance, asymptotic convergence.
Multiple users case represents a more realistic model in a real radio environment in which a policy is required to learn collectively or separately the vacancy probabilities of channels. Thus, we first introduced the recent works of MAB policies to manage a secondary network, and then we summarized their performance.

In the next chapter, we will investigate the performance of the well-known MAB algorithms in OSA for a single user. Then, we propose a novel policy to extend the MAB algorithms to consider multiple priority users.

Chapter 3

MAB algorithms in OSA for Multiple Users

Contents

3.1 Introduction

Opportunistic Spectrum Access (OSA) represents one of the proposed solutions to enhance the spectrum efficiency by sharing the spectrum between two types of users: Primary and Secondary. A Secondary User (SU) should make the best decision in order to maximize his throughput, without any prior information about the vacancy probabilities of channels. Moreover, any interference with the Primary User (PU) should be under a certain limit. Thus, Multi-Armed Bandit (MAB) has been chosen as a suitable solution to help a SU make a decision. In MAB problem, an agent tries to find the optimal arm without any prior knowledge about the arm's reward.

To the best of our knowledge, OSA is not being used in a real radio environment but has recently received more and more attention in order to increase the spectrum efficiency. In this chapter, we investigate the performance of the well-known MAB algorithms in OSA such as: Thompson Sampling (TS), Upper Confidence Bound (UCB) and e-greedy. The performance of MAB algorithms applied in OSA depends on several criteria: The convergence speed towards the optimal channel (i.e the channel with the highest availability probability) and the number of times to access the worst channels. Thus, we evaluate the performance of MAB algorithms using: the percentage to access the best channel (P_{best}) and the loss of rewards by selecting the worse channels (Regret). We should mention that the well-known MAB algorithms are first suggested for a single user, and to use it in the OSA for multiple users, a policy is required to manage the secondary network. In this chapter, we propose a cooperative learning policy, called Side Channel, that may be used with any learning MAB algorithm. The proposed policy can help SUs to learn collectively the vacancy probabilities of channels and decrease the number of collisions among users. Moreover, this policy takes into account the priority access. It can be shown that the regret of the proposed policy, under MAB algorithms, has a logarithmic asymptotic behavior for any number of users. Subsequently, after a finite number of time slots the users are able to learn the vacancy of channels and each of them accesses his dedicated channel.

In section 3.2, we introduce and discuss in details the well-known MAB algorithms. In section 3.3, we present the major challenges faced by the SU in the licensed bands. The problem formulation in OSA, for a single or multiple

users, is introduced in section 3.4. In section 3.5, we propose our first contribution in which a novel cooperative policy is proposed to consider multiple priority users. Section 3.6 provides the simulation results of the MAB algorithms for a single or multiple users. Finally, section 3.7 concludes this chapter.

3.2 Single Agent MAB Algorithms

MAB problem is a simple case of Reinforcement Learning (RL) where an agent tries to learn the expected reward of arms. The main target of the agent is to find the best action that maximizes his long-time reward.
MAB problem was firstly suggested for a single agent and several algorithms have been proposed on the matter, such as: Thompson Sampling (TS), Upper Confidence Bound (UCB), e-greedy. In such algorithms, the reward obtained from each arm is formulated by an IID model. According to these MAB algorithms, if the mean reward of the worst arms is very close to the optimal one (i.e. arm with the highest payoff), then the agent spends a lot of time to converge towords the optimal arm, otherwise he reaches the optimal behavior faster. MAB algorithms have recently become widely suggested in the literature in several fields such as a website optimization (e.g. maximizing conversions on landing pages, the composition of a landing page may involve deciding which image to show or color background to display) [71], internet advertising [72, 73, 74]. MAB has also been adopted by several companies, including Adobe, Amazon [71], Facebook, Google [75, 76], LinkedIn [73, 74], Microsoft [72], Netflix.

3.2.1 Thompson Sampling

Thompson Sampling (TS), a Bayesian algorithm, represents the earliest learning algorithm proposed to tackle the MAB problem [77]. Despite its better performance compared to several other MAB algorithms, TS is largely ignored in the literature. Recently, TS attracts more and more attention and several works investigated the proof of its strong performance [56, 57, 58]. In TS, each arm (or channel in the context of cognitive radio) has assigned an index $B_i(t, T_i(t))$ where $T_i(t)$ stands for the number of slot times to access the i-th channel by the agent. The index of each arm follows a Beta distribution:

$$B_i(t, T_i(t)) = \beta(W_i(t, T_i(t)) + 1, Z_i(t, T_i(t)) + 1)$$

Algorithm 1: Thompson Sampling Algorithm

Input: C, n,
C: number of channels,
n: total number of slots,
Parameters: $T_i(t), W_i(t, T_i(t)), Z_i(t, T_i(t))$,
$S_i(t)$: the state of the selected channel, equals one if the channel is free and 0 otherwise,
$T_i(t)$: the number of times the i-th channel is sensed by the user up to time t,
$W_i(t, T_i(t))$: the success access up to time t depends on $T_i(t)$,
$Z_i(t, T_i(t))$: the failure access up to time t depends on $T_i(t)$,
Output: $B_i(t, T_i(t))$,
$B_i(t, T_i(t))$: the index assigned for channels,
foreach $t = 1$ *to* n **do**
  $a_t = \arg\max_i B_i(t, T_i(t))$,
  The user observes the state $S_i(t)$,
  $W_i(t, T_i(t)) = \sum_{t=0}^{n} S_i(t) 1_{a_t=i}$,
  % $1_{a_t=i}$: equals 1 if the user selects the i-th channel and 0 otherwise,
  $Z_i(t, T_i(t)) = T_i(t) - W_i(t, T_i(t))$,
  $B_i(t, T_i(t)) = \beta(W_i(t, T_i(t)) + 1, Z_i(t, T_i(t)) + 1)$,

where $W_i(t, T_i(t))$ and $Z_i(t, T_i(t))$ represent respectively the numbers of success and failure accesses. The expected value of a Beta distribution for a random variable X for $a > 1$ and $b > 1$ is given by:

$$E[X] = \frac{a}{a+b}$$

At the initialization $t = 0$, the user begins to access the channels without any prior information about their availability probability μ_i. When $t = 0$, $W_i(t, T_i(t))$= $Z_i(t, T_i(t)) = 0$, and for all channels, the assigned index $B_i(0, T_i(0))$ is set to $\beta(1, 1)$. At each slot t, the user makes an action a_t and selects the channel with the highest index:

$$a_t = \arg\max_i B_i(t, T_i(t))$$

After a finite number of time slots, the vacancy probability of the i-th channel μ_i will be very closed to $B_i(t, T_i(t))$. By choosing the channel with the highest index, the user usually selects the optimal one.

3.2.2 Upper Confidence Bound

Upper Confidence Bound (UCB) represents the most famous MAB algorithm that is widely suggested in the literature. The idea of UCB was initially introduced by Lai and Robbins in [38]. Several versions of this algorithm have been also proposed, such as: UCB1, UCB2, UCB-tuned, UCB-Normal, UCB-Bayes, KL-UCB, etc. Unlike TS, UCB is proposed with a solid mathematical background.

Algorithm 2: Upper Confidence Bound Algorithm

Input: α, C, n,
α: the exploration-exploitation factor,
C: the number of channels,
n: the total number of slots,
Parameters: $T_i(t), X_i(T_i(t)), A_i(t, T_i(t))$,
$T_i(t)$: the number of time slots the channel is sensed up to time t,
$X_i(T_i(t))$: the exploitation contribution of channels depends on $T_i(t)$,
$A_i(t, T_i(t))$: the exploration contribution of channels depends on $T_i(t)$ and t,
Output: $B_i(t, T_i(t))$,
$B_i(t, T_i(t))$: the index assigned for channels,
foreach $t \in [1, C]$ **do**
    The user senses each channel once,
    The user updates his index $B_i(t, T_i(t))$,
while $t \leq n$ **do**
    $a_t = \max_i B_i(t-1, T(t-1))$,
    $T_i(t) + +$,
    $X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{\tau=1}^{t} S_i(\tau)$,
    % $S_i(\tau)$ is the observed state from channel i at τ,
    % $S_i(\tau) = 1$ if the channel i is vacant and 0 otherwise,
    $A_i(t, T_i(t)) = \sqrt{\frac{\alpha \ln(t)}{T_i(t)}}$,
    $B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t))$,

To help the SU make a good decision, UCB1 [43] represents the most used version in the literature. The importance of UCB1 is probably due to the fact that it balances between the optimality and the simplicity. Several recent works proposed various policies to manage a secondary network with multiple users based on UCB1. Similar to TS, in UCB1, each channel has assigned an index $B_i(t, T_i(t))$, where $T(_i(t))$ has the same definition as in TS. Based on the previous states of the channels, the user should learn the vacancy probabilities of channels. As shown in algorithm 2, $B_i(t, T_i(t))$ contains two factors: $X_i(t, T_i(t))$ and $A_i(t, T_i(t))$ that represent respectively the exploitation

(also known as the expected reward) and the exploration. The role of the exploration factor $A_i(t, T_i(t))$ remains to reinforce the algorithm to examine the state of the available channels in order to gather information about their vacancy probabilities. Therefore, the exploitation factor, $X_i(t, T_i(t))$, will be very close to the vacancy of channels, μ_i.
The assigned index of each channel can be defined as follows:

$$B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t)) \tag{3.1}$$

At the initialization period (i.e. $t = 1$ to the number of channels C), the user selects each channel once to have some knowledge about the vacant probabilities of channels. After that, the user selects the channel that has the highest index $B_i(t, T_i(t))$:

$$a_t = \arg\max_i B_i(t, T_i(t))$$

In [43], the authors also suggest the upper bound of the sum of regret (i.e. the loss of reward by selecting the worst channels) and show that the regret achieves a logarithmic asymptotic behavior with respect to time. Subsequently, after a finite number of time slots the user recognizes the optimal channel and always selects it.

3.2.3 *e*-greedy

e-greedy represents one of the simplest algorithms that was first introduced in [47] to tackle the MAB problem. A recent version of this algorithm called decreasing *e*-greedy is proposed in [43] to achieve a better performance with respect to the classical version (see algorithm 3).
In order to have information about the availability of channels (exploration), the user selects a random channel if χ (i.e. a uniform random variable $\in$ [0,1]) $< \epsilon_t$ where ϵ_t = min $\{1, \frac{H}{t}\}$ and H is a constant number. Otherwise, the user selects the channel wih the highest expected reward $X_i(T_i(t))$ (exploitation). The authors have also proved that the upper bound of the regret grows linearly due to the selection of a random channel during the exploration period. Later on, the regret achieves a logarithmic asymptotic behavior (exploitation).

Algorithm 3: e-greedy Algorithm

Input: C, n,
C: number of channels,
n: total number of slots,
Parameters: $T_i(t)$,
$T_i(t)$: number of time slots the channel is sensed up to time t,
Output: $X_i(t, T_i(t))$,
$X_i(T_i(t))$: the expected reward that depends on $T_i(t)$,
foreach $t = 1$ *to* n **do**
  if $\chi < \epsilon_t$ **then**
    The user makes a random action a_t,
  else
    $a_t = \max_i X_i(T_i(t))$,
    $T_i(t) + +$,
    $X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{\tau=1}^{t} S_i(\tau)$,
    % $S_i(\tau)$ is the observed state from channel i at τ,
    % $S_i(\tau) = 1$ if the channel i is vacant and 0 otherwise,

3.3 The Challenges of the Secondary User in the Licensed Bands

For strategic and logistic reasons as well as to simplify the complexity of SU receivers in our working context, we assume that the SU is able to sense and explore one channel at each time slot to find transmission opportunities. OSA in cognitive radio networks (CRNs) presents new challenges compared to current wireless networks:

- Spectrum sensing,
- Learning and extracting information,
- Decision-making.

3.3.1 Spectrum Sensing

Through its sensors, a cognitive radio is able to gather some information from its environment. So, a SU can find and access frequency bands unused by the licensed users. Subsequently, it is possible for a SU to send his data over these empty bands. In this context, several spectrum-sensing techniques have been proposed in the literature. Matched filter detection (MFD) represents one of these techniques in which the transmission signal of the

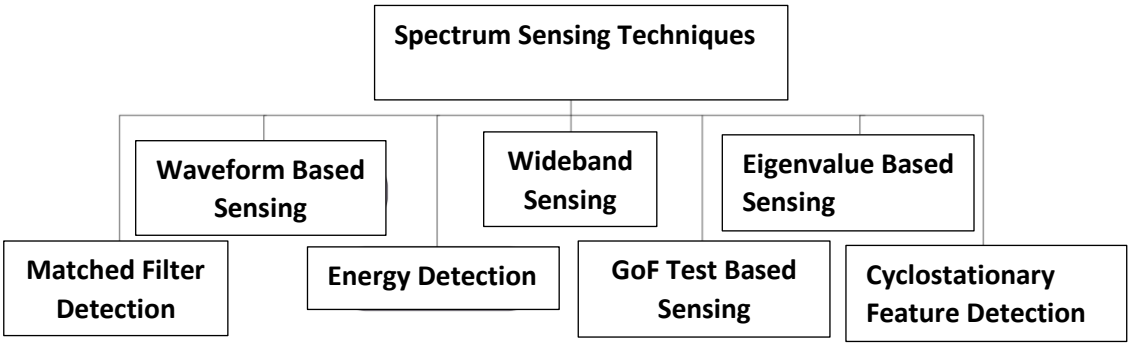

FIGURE 3.1: Spectrum Sensing Techniques

licensed users should be known to the cognitive user [78, 79, 80, 81]. Cyclostationary Feature Detection (CFD) represents another form of the spectrum sensing technique which also requires a prior information about the characteristics of the licensed user's transmission signal [79]. Moreover, it has better robustness to noise uncertainty compared to MFD method [82, 83].
In order to achieve a good detection performance, CFD technique requires significantly long observation time and complex implementation.
The most popular technique to detect the licensed transmission signal is the Energy Detection (ED). Despite its low computation complexity, it suffers from the noise uncertainty and requires a sufficient number of samples to ensure a high probability of detection [79, 84, 85, 86]. The most-known spectrum sensing techniques found in the literature are depicted in Fig. 3.1.

3.3.2 Learning and Extracting Information

The cognitive radio system would also benefit from introducing learning capabilities into equipments in order to increase their decision-making autonomy for unexpected situations. The learning abilities would allow to ensure a further optimization and an efficient resource allocation. However, after the collection of some information (e.g. the availability of channels), the SU constructs a database of the environment in order to adapt his transmission parameters then optimize the performance of the communication. Recent studies show that the cognitive user, based on machine learning and spectrum prediction, can increase his throughput compared to the randomly spectrum sensing [87]. Modern machine learning techniques [88, 89, 90] may guide the system reconfiguration in order to maximize the opportunities of the cognitive users and decrease the interference with the licensed users as much

as possible. However, when the SU interacts with his environment, it gathers a sequence of information about the history state of channels as shown in Fig. 3.2. The extracting information can be treated by the SU to find the best actions with respect to his environment. Indeed, in the literature, several Multi-Arm Bandit (MAB) learning algorithms have been suggested in order to exploit the available spectrum in the best manner. The well-known MAB learning algorithms are TS [77], UCB, and e-greedy [47]. Consequently, learning and extracted information can be considered as an important step in order to improve the cognitive user's decision and minimize the time spent trying to find the best action.

3.3.3 Decision-Making

A SU should select the best decision and decide which channel to access among the available spectrum bands. The decision-making of a SU should depend on his past successes and failures decision. When a SU appears in the network, his best decision becomes to increase his transmission time without causing any harm to the licensed user. The decision-making of a SU should consider the following conditions:

- A SU can only access unused licensed spectrum bands.
- When a SU accesses a bad channel (e.g. with respect to the vacancy probability of that channel), he should modify his future decision and switch to the best band in order to improve the throughput, the quality of service and the energy consumption.

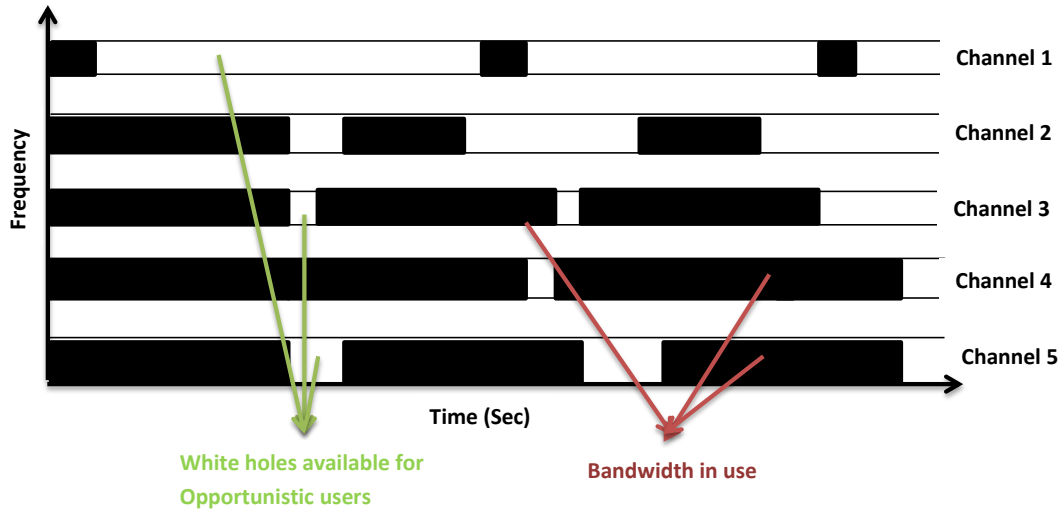

FIGURE 3.2: Vacancy of licensed channels

- Due to the presence of multiple users, a collision may occur among them, therefore, colliding users should change their decision in order to avoid interference and collision in their scheduled future transmissions.

Consequently, the optimal decision depends on the activities of licensed users, as well as on the competitive or cooperative behavior of the cognitive users. In competitive access, the cognitive user acts individually, without the need for any interaction with other users, in order to maximize his personal benefits (e.g. throughput, transmission time). Although the competitive access may increase the number of collisions among users, it reduces the complexity of the system. On the other hand, in a cooperative access, the cognitive users should collaborate and exchange some information about their environment in order to make a collective decision to increase the performance of their cognitive network. The collaborative access may decrease the number of collisions among users and avoid conflicts with each other.

3.4 Problem Formulation

3.4.1 Single User

In OSA, there are several existing problems not yet considered, to the best of our knowledge, in the context of a MAB approach, such as: multiple users, collisions among users, power transmission levels, sensing techniques, etc.
Let us start to formulate the classical OSA as a MAB problem in which we consider a Secondary User (SU) that needs to access opportunistically the frequency band. Later on, we will consider multiple users with more realistic scenarios. Suppose that, the licensed band, reserved to licensed users, contains C Independent Identical Distributed (IID) channels. We denote $i \in \{1, ..., C\}$ the i-th best channel. Each channel, when observed by the SU, appears in two possible states $\{\text{free, occupied}\}$. In the remainder of this work, we consider the numerical values $\{1, 0\}$ that denote the states of the channels: 1 if the observed channel is free and 0 otherwise. Thus, we consider that the occupancy of all channels $\Theta = \{\theta_1, ..., \theta_C\}$ follows a Bernoulli distribution. Moreover, we assume that the vector Θ is supposed stationary.
In our work , we tackle the case in which the activity of the PUs is supposed to be synchronized and the time is slotted ($t = 1, 2, ...n$). Let $S_i(t)$ be the state of the i-th channel at slot t, considered as a random variable drawn from a given distribution θ_i.
At each time slot, the user chooses a channel to observe its state $S_i(t)$, and

receives a reward $r_i(t)$. Without any loss of generality, in our work we consider that the obtained reward from the i-th channel equals to its binary state, i.e. $S_i(t) = r_i(t)$. If the channel selected by the SU is free, then he may transmit his own data, otherwise he should wait for the next slot to sense another channel. Let μ_i be the expected availability of the i-th channel defined as follows:

$$\mu_i = E[\theta_i] = \text{Probability}(\text{channel } i \text{ is free}) = P(S_i(t) = 1) \tag{3.2}$$

We consider that the channels are ordered by their vacancy probabilities $\mu_1 > \mu_2 > ... > \mu_C$. As the vacancy probabilities of channels represented by the vector $\Gamma = \{\mu_1, \mu_2, ..., \mu_C\}$ are unknown, the user has to learn Γ based on the information obtained over time. After evaluating the vector Γ, the target of the user remains to access the best available channel, i.e. μ_1, in order to increase his transmission time and rate. As μ_i is unknown to the user, let us define the regret as the loss of reward between an ideal scenario where the user has a prior knowledge about the vacancy probability of channels and he can always accesses the best one, and the reward obtained from using a particular MAB algorithm.
The regret factor is widely used in the literature in order to evaluate the performance of a given algorithm. A MAB algorithm is said to be optimal if the obtained regret is minimal. For a single user, the regret $R(n, \beta)$ using a given algorithm β up to the total number of slots n can be defined as follows:

$$R(n, \beta) = n\mu_1 - \sum_{t=1}^{n} \mu_i^{\beta}(t) \tag{3.3}$$

where $\mu_i^{\beta(t)}(t)$ represents the vacancy probability of the selected channel at slot t using the algorithm β.
The regrets of most suggested MAB algorithm, such as Thompson Sampling, UCB and e-greedy achieve a logarithmic asymptotic behavior. Consequently, after a finite number of time slots, the user may have a good estimation of the channels' vacancy probabilities and always selects the optimal one.

3.4.2 Multi-Users

Let us consider U SUs trying to learn the vacant probabilities of channels in order to access only the U best channels. At each time slot t, each user can sense and access a channel when available to transmit. Multiple SUs can work in cooperative or uncooperative modes. The regret for the multi-user

case, no matter if it is under cooperative or competitive modes, can be written as follows:

$$R(n, U, \beta) = n \sum_{k=1}^{U} \mu_k - \sum_{t=1}^{n} E\left(S^{\beta(t)}(t)\right) \tag{3.4}$$

where μ_k is the mean availability of the k^{th} best channel; $S^{\beta(t)}(t)$ stands for the global reward obtained by all users at the time slot t; $E(.)$ represents the mathematical expectation, and $\beta(t)$ represents all the selected channels[1] by users at t. We can define $S^{\beta(t)}(t)$ by:

$$S^{\beta(t)}(t) = \sum_{j=1}^{U} \sum_{i=1}^{C} S_i(t) I_{i,j}(t) \tag{3.5}$$

where the state variable[2] $S_i(t) = 0$ indicates that the channel i is occupied by the PU at slot t, otherwise $S_i(t) = 1$; $I_{i,j}(t) = 1$ if the j^{th} user is the sole occupant of channel i at the slot t and 0 otherwise. In the multi-user case, the regret may be affected by the collision among the SUs and the vacancy of channels which allows us to define the regret for the U SUs as follows:

$$R(n, U, \beta) = n \sum_{k=1}^{U} \mu_k - \sum_{j=1}^{U} \sum_{i=1}^{C} P_{i,j}(n) \mu_i \tag{3.6}$$

where $P_{i,j}(n) = \sum_{t=1}^{n} E\left[I_{i,j}(t)\right]$ stands for the expectation of times when the user j is the sole occupant in the channel i up to time n. In the following section, we propose a novel cooperative policy for the priority access to manage a secondary network.

3.5 Cooperative learning with a Side Channel policy

The coordination among the SUs can enhance the efficiency of their network, instead of dealing with their partial information about the environment. To manage a cooperative network, we propose a policy based on the use of a Side Channel in order to exchange simple information among SUs with a very low information rate. The Side Channels are widely used in wireless telecommunication networks to share data among the base-stations [91],

[1] $\beta(t)$ indicates the channel selected by the user at instant t in the single user case while in the multi-access it indicates the channels selected by all users at slot t.

[2] The variable $S_i(t)$ may represent the reward of the i^{th} channel at slot t.

and specifically in the context of cognitive network. However, in [19] and [20], the authors considered the cooperative spectrum sharing among PUs and SUs to enhance the transmission rate of the PUs using a side channel. The signaling channel in our policy is not wide enough to allow high data rate transmission unlike the ones proposed in [19] and [20] which should have a high rate to ensure the data transmission among PUs and SUs. In our policy, the transmission is done over periods. During the first period, i.e. Sub-Slot1, SU_1 (the highest priority user) searches the best channel by maximizing his index. At the same time, and via the secure channel, SU_1 must inform the other users to evacuate his selected channel in order to avoid any collision with him (see algorithm 4). While avoiding the first selected channel, the second user SU_2 should repeat the same procedure and so on. If SU_2 does not receive the choice of SU_1 in the first Sub-Slot1 (suppose that SU_1 does not need to transmit during this Sub-Slot), it can choose directly the first suggested channel by maximizing his index $B_{i,2}(t, T_{i,2}(t))$. To the best of our knowledge, all proposed policies, such as SLK, k^{th} MAB consider a fixed priority, i.e. the k^{th} best channel is reserved for the k^{th} user all the time. Then, if SU_1 does not transmit for a certain time, the other users cannot select better channels. Subsequently, the main advantages of the cooperation in this policy are:

- An efficient use of the spectrum where best channels are constantly accessed by users.
- An increase in the users' transmission time by avoiding the collision among them.
- Reaching a lower regret compared to several existing policies.

Algorithm 4: Side Channel policy based on Upper Confidence Bound,

Input: α, C, n,

Parameters: k, I_j, $a(t,j)$, $T_{i,j}(t)$

k: cyclic sub-slot,

I_j : indicates the activity of the j^{th} secondary user,

% $I_j = 1$ means SU active, otherwise $I_j = 0$,

$a(t,j)$: indicates the channel selected by user j at time t,

$T_{i,j}(t)$: the number of time the j^{th} user senses the i^{th} channel,

Output: $B_{i,j}(t, T_{i,j}(t))$

$B_{i,j}(t, T_{i,j}(t))$: the assigned index of the i^{th} channel for the j^{th} user,

foreach $t \in [1, C]$ **do**

  Each SU should visit each channel one time, Each SU should update his $B_{i,j}(t, T_{i,j}(t))$,

while $t \leq n$ **do**

  $t = t + 1$,

  $k = k\ mod(j) + 1$,

  if $I_j = 0$ **then**

    There is no transmission; And for all $j^{'} > j$, each $SU_{j^{'}}$ has the chance to exploit the ($j^{'}$-1)$-th$ channel,

  else

    if $k = 1$ **then** SU_1 %test 1

      Applies a MAB algorithm and searches the channel $a(t,1)$ that maximizes his index $B_{i,1}(t, T_{i,1}(t))$,

      Broadcasts $a(t,1)$ to other users,

      Transmits his data.

    ⋮

    if $k = U$ **then** SU_U %test U

      Eliminates all channels selected by other users, i.e. $a(t,1)$,..., $a(t, U-1)$

      Applies a MAB algorithm and searches the channel $a(t,U)$ that maximizes his index $B_{i,U}(t, T_{i,U}(t))$,

      Transmits his data.

3.6 Simulation Results

In this section, we focus first on the well-known MAB algorithms, such as: Thompson Sampling, UCB and e-greedy that are suggested to tackle the OSA

for a single user. Later on, we evaluate the performance of the proposed learning policy to help users to learn collectively the vacancy probability of channels.

In OSA, the user is supposed to select one channel at each time slot and transmits his data if the targeted channel is free. Otherwise, he should wait the next slot to select another channel. We evaluate the performance of the MAB algorithms in OSA using three important criteria: Regret, P_{best} and the throughput. The Regret measures the speed of the convergence and represents the difference between the reward obtained from the ideal scenario and that obtained from a given policy. P_{best} represents the percentage of times to access the best channel; this is the guarantee that the learning MAB algorithm converges and always selects the best channel. Finally, the throughput represents the percentage of successful transmission by the SU. Without a prior knowledge about the vacancy of the channels, the best successful transmission achieved by the SU may not exceed the availability probability of the best channel. Indeed, a given policy is considered as optimal whereas the best channel is distinguished and always been selected by the user. Moreover, in the ideal scenario where the vacancy probabilities of channels are considered as known, the best strategy for users is to regularly select the best channel with the highest vacancy probability. Subsequently, the best throughput achieved by user is related to the vacancy probability of the best channel.

In our simulations, we consider 9 IID channels ordered by their availability as follows

$$\Gamma = [0.9 \;\; 0.8 \;\; 0.7 \;\; 0.6 \;\; 0.5 \;\; 0.4 \;\; 0.3 \;\; 0.2 \;\; 0.1]$$

and we suppose that only one SU is trying to learn the availabilities of channels in order to access the most vacant one, i.e. $\mu_1 = 0.9$. For UCB, the exploration-exploitation factor α is set to 1.5. According to several studies [40, 92, 93, 94], α should be higher than 1 in order to ensure the convergence of the UCB. When α increases, UCB gives more weight to the exploration factor $A(t, T_i(t))$ in order to gather more information about the vacancy probability of channels while the exploitation factor $X_i(t, T_i(t))$ becomes less important. Unlike UCB, in e-greedy the exploration-exploitation phases are separated and the period of each of them depends on the exploration factor H. According to [43], the value of H should be higher than 90 in order to have a good estimation about the vacancy probability of channels and then access the optimal one.

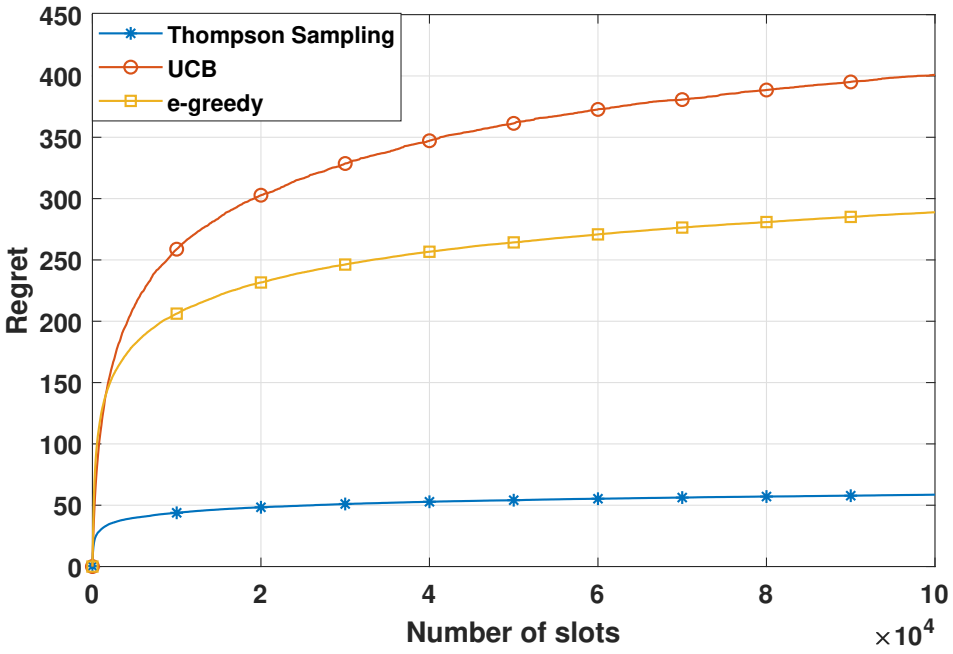

FIGURE 3.3: Regret comparison of Thompson Sampling, UCB1 and e-greedy

Fig. 3.3 compares the regret of the well-known MAB algorithms Thompson Sampling, UCB1 and e-greedy in the context of OSA. As we can see, the regrets of the three MAB algorithms achieve logarithmic asymptotic behaviors with respect to time. So, after a finite number of time slots the user will be able to learn the vacancy probability of the channels and then access the most vacant one. Fig. 3.3 also shows that Thompson Sampling outperforms UCB and e-greedy may reach the lowest regret.

In Fig. 3.4, we compare the percentage of times to access the best channel, P_{best}, by the SU that is given as follows:

$$P_{best} = 100 \times \sum_{t=1}^{n} \frac{\mathbb{1}_{(\beta(t)=\mu_1)}}{t}$$

where $\mathbf{1}_{(\mathbf{a}=\mathbf{b})} = \begin{cases} 1 \text{ if a =b} \\ 0 \text{ otherwise} \end{cases}$

In Fig. 3.4, P_{best} shows three main parts:

- The initialization part from 1 to C, where the user selects each channel once in order to have a prior information about the vacancy probability of channels.

- The adaptation part that goes from C+1 to 2000 slots approximately.

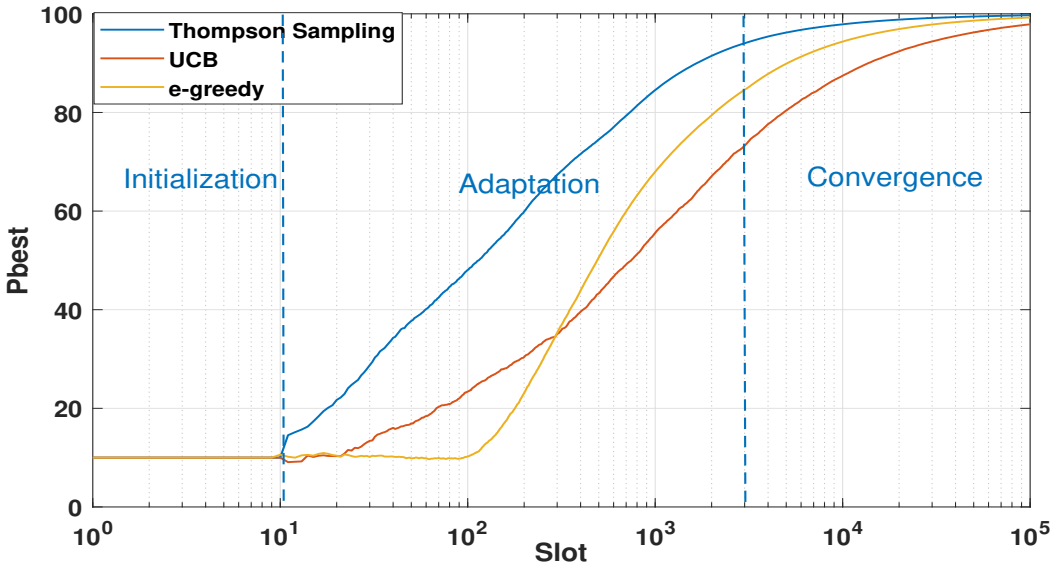

FIGURE 3.4: Pbest comparaison of Thompson Sampling, UCB and *e*-greedy

- The last part in which the user converges to the best channel μ_1.

After the initialization part, P_{best} approximately increases in a similar way for the three algorithms. After hundreds of slots, Thompson Sampling outperforms UCB and *e*-greedy. Up to 1000 slots, Thompson Sampling achieves 85 % of the best channel, while *e*-greedy achieves around 70 % and finally UCB produce the worst result with only 55 %.

Another important metric, widely used in the literature in order to evaluate the performance of the MAB algorithms in OSA, is the throughput capacity (also known as successful transmission). According to Fig. 3.5, we can conclude that the user can find more opportunities using Thompson Sampling with a successful transmission that can reach around 90 %. While under UCB or *e*-greedy around 85 % of the transmission is achieved.

In the following simulations, we consider 3 SUs with 10 channels and their vacancy probabilities are given by:

$$\Gamma = [0.9\ 0.8\ 0.7\ 0.6\ 0.5\ 0.45\ 0.4\ 0.3\ 0.2\ 0.1]$$

The percentage of times where each SU_k selects his optimal channel using a given approach can be defined as follows:

$$P^k_{best}(t) = \sum_{t=1}^{n} \frac{1_{(if\ a(t)=k)}}{t}$$

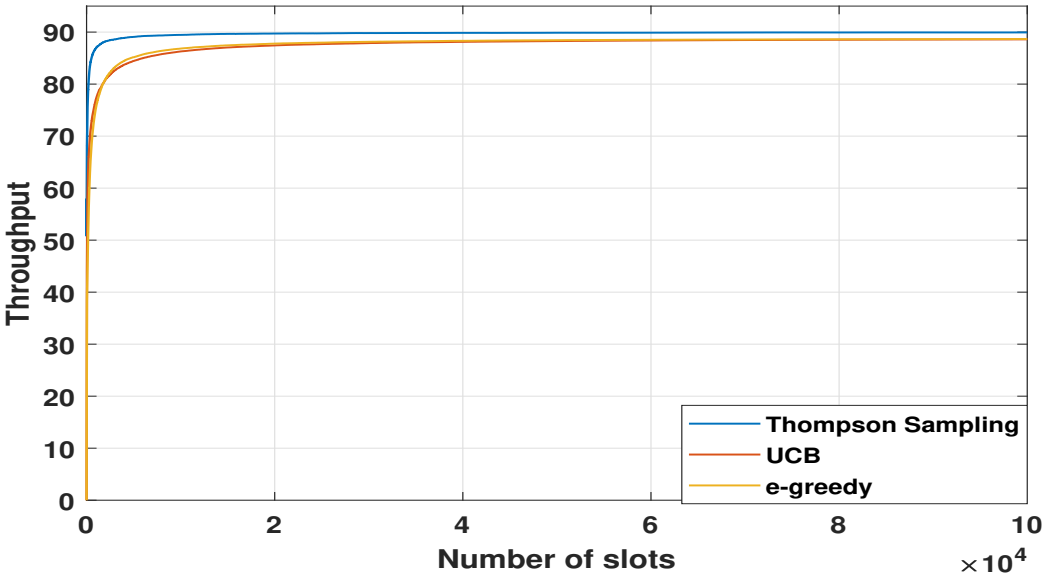

FIGURE 3.5: Throughput capacity of Thompson Sampling, UCB and e-greedy

Fig. 3.6 presents $P^k_{best}(t)$ of our policy under the three learning algorithms. It shows, as obtained for the single user case, that the $P^k_{best}(t)$ is divided into three parts: Initialization, adaptation and convergence. Subfigures (a), (b) and (c) show $P^k_{best}(t)$ for our policy under Thompson Sampling, UCB and e-greedy respectively. In each case, three users with different priority levels are considered. Subfigure (d) compares $P^k_{best}(t)$ of the first priority user under each algorithm. From our simulations, and based on our policy, one can conclude that users can reach their dedicated channels according to their prior rank. As we can observe, in subfigures (a), (b) and (c), the first priority user SU_1 converges towards the best channel μ_1, followed by SU_2 and SU_3 towards the second and third best channels μ_2 and μ_3 respectively. In addition, we notice in subfigure (d) a high convergence speed of our policy with TS compared to UCB and e-greedy.

Fig. 3.7 depicts the regret defined in equation (3.6), where the proposed approach achieves a logarithmic regret for the three learning algorithms (i.e. Thompson Sampling, UCB and e-greedy). Under our policy, the same figure shows that Thompson Sampling achieves a lower regret compared to UCB or e-greedy. This is due to the fact that Thompson Sampling can learn the vacancy probabilities of channels in a more efficient way and then achieves better results.

Random Rank [35] and Selective Learning of the k-th largest expected rewards (SLK) [61] are two policies proposed in the literature to manage a

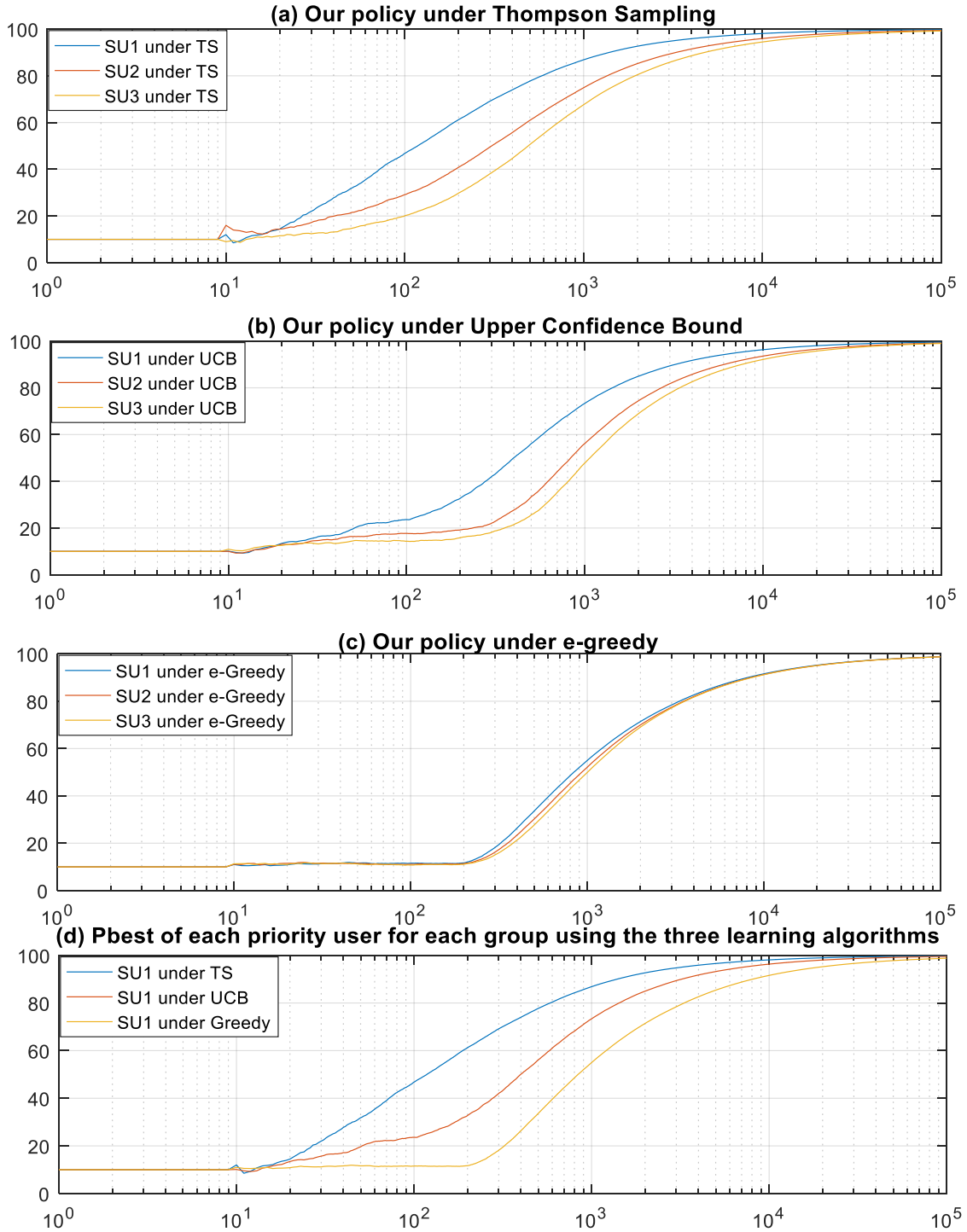

FIGURE 3.6: P_{best} of the proposed policy under TS, UCB and e-greedy for three priority users.

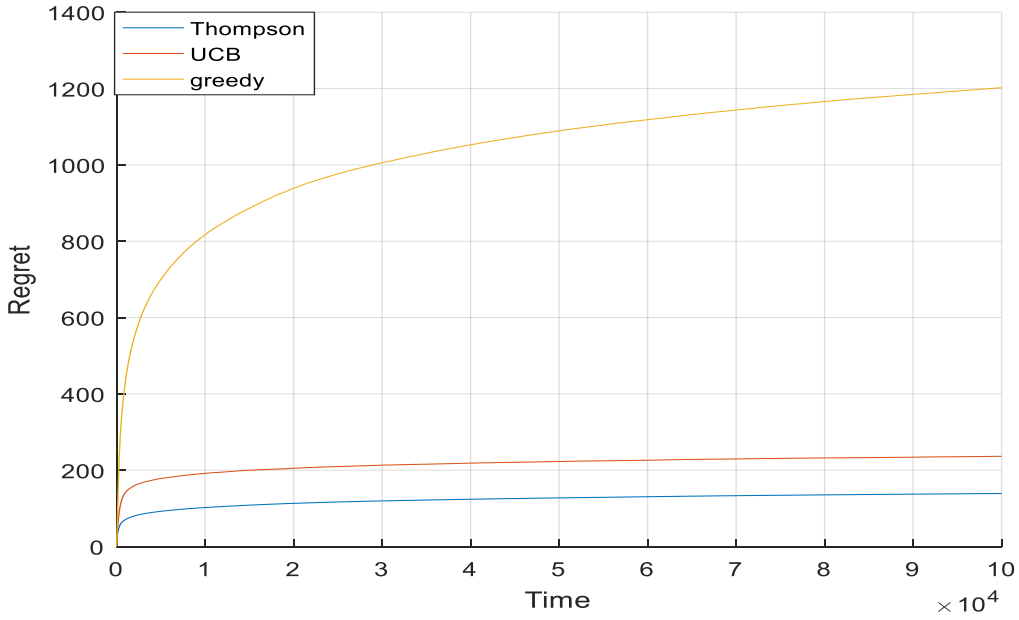

FIGURE 3.7: Logarithmic Regret of our policy under the three learning algorithms, for three secondary users.

secondary network in which several users share the frequency band.

These policies have been suggested based on UCB algorithm. Our proposed policy and SLK take into account the priority access in which each user has a prior rank. However, in Random Rank, no fixed or prior rank is considered, and after each collision, the users regenerate uniformly their ranks from $\{1, ..., U\}$. As we can see in Fig. 3.8, our policy outperforms Random Rank and SLK policies by achieving the lowest regret.

3.7 Conclusion

In this chapter, we introduced the well-known algorithms: Thompson Sampling (TS), Upper Confidence Bound (UCB) and e-greedy in order to tackle the MAB problem. Then, we investigated the performance of the mentioned algorithms in the context of OSA. We started by modeling the basic OSA as a MAB problem in which one SU is considered to learn the vacancy probability of channels and then access the best one. Then, we proposed a cooperative learning policy called Side Channel in order to consider the multi-access case in OSA. Indeed, in OSA several SUs may exist in the network and the main issue is to learn collectively the vacancy probability of channels while decreasing the collision among users. In our policy, the users can exchange their choices using a Side Channel in order to avoid any collision among them.

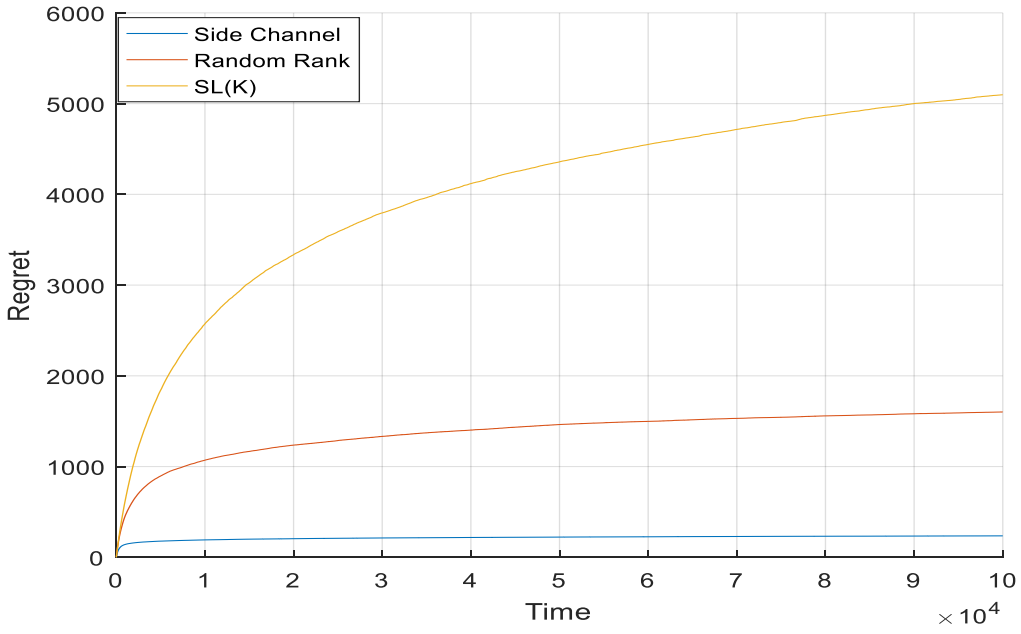

FIGURE 3.8: Global regret for the three plocies (Our proposed one, Random Rank, SLK) applied on UCB with 3 SUs.

Moreover, in our work we are interested in the priority access in which, and based on our proposed policy, the users have different priority levels.

In order to evaluate the performance of MAB algorithms applied in OSA, for a single or multiple users, we used two main factors: the regret (i.e. loss of reward) and the P_{best}. The former represents the widely used parameter in the literature to study the convergence speed towards the optimal choice while the latter confirms that a given MAB algorithm is converged to the optimal channel. In our simulation, we investigated the performance of our proposed policy under TS, UCB and e-greedy. The obtained results showed that a better performance can be reached using TS. Based on our policy for the priority access, the first priority user converges towards the first optimal channel followed by the second and third users to their dedicated channels. Finally, we compared the regret of our policy with those of Random Rank and SLK that have a logarithmic asymptotic behavior and proved that Side Channel achieved the optimal regret.

Chapter 4

Distributed Learning for the Priority Cognitive Access

Contents

4.1 Introduction

In order to enhance the spectrum learning and obtain better results compared to several well-known Multi-Armed Bandit (MAB) algorithms, we propose two novel algorithms based on Upper Confidence Bound (UCB): *e*-UCB and AUCB.
Well-known MAB algorithms, as well as *e*-UCB and AUCB, contain two phases: exploration and exploitation. These algorithms generally try to decrease the total regret defined as the gap between the reward obtained in the ideal scenario and that obtained using a given MAB algorithm. Based on *e*-UCB and AUCB, the Secondary User (SU) can quickly reach the optimal channel (the most vacant one). Subsequently, the SU is not only able to find an opportunity in the licensed band but he will also increase his transmission time and rate in the long-term. Analytical convergences of *e*-UCB and AUCB are investigated and we will show that the regret can achieve a logarithmic asymptotic behavior with respect to time. That means, after a finite time slots the user is able to learn the vacancy probabilities of channels and will always select the optimal one. Later on, we extend the proposed algorithms to consider multiple users by considering two possible models: Cooperative or competitive. The cooperative access enables users to exchange information with each other in order to maximize a common function (e.g. the throughput of the network). While, the competitive access is more challenging: The collision among users can cause a loss of reward and all users need to learn separately the vacancy probabilities of channels.

The organization of this chapter is as follows: Section 4.2 introduces the proposed MAB algorithm *e*-UCB and investigates the upper bound of its regret. In section 4.3, we present AUCB and the upper bound of its regret. In section 4.4, the proposed MAB algorithms are extended to consider multiple users under two models: Cooperative and competitive. In section 4.5, we evaluate the performance of our algorithm for a single or multiple users. Finally, section 4.6 concludes this chapter.

4.2 Distributed Learning and Access Algorithms

In this section, we propose a novel learning algorithm called *e*-UCB, to tackle the OSA problem and help a SU to make a decision. In the literature, several algorithms have been proposed to solve the MAB problem: Thompson

Sampling (TS), UCB, e-greedy, EXP3, etc. TS represents a simple but effective algorithm that can exceed the state of the art of other MAB algorithms. However, TS is almost neglected in the literature because it was first proposed without any analytical proof. Recently, TS attracts more and more attention and is being successfully applied in a wide variety of domains. Nevertheless, we have the formal proof of its upper bound and optimal regret. In this section, we propose e-UCB that can achieve better results compared to UCB or e-greedy. We also show the analytical proof of e-UCB for a single user.

4.2.1 e-UCB for Opportunistic Access

UCB represents a popular MAB algorithm that is proposed with a solid mathematical background. Several versions of UCB have been proposed to tackle the MAB problem and improve the performance compared to the classical UCB: UCB1, UCB2, UCB-tuned, UCB-Bayes. UCB1, proposed in [40], is widely used in Opportunistic Spectrum Access (OSA) to manage a secondary network for a single or multiple users. The importance of UCB1 is due to the fact that this algorithm achieves a trade-off between the optimality and the complexity.

Like all MAB algorithms, UCB1 provides two phases in order to learn the vacancy probabilities of channels: exploration and exploitation. In UCB1, each channel has assigned an index, $B_i(t, T_i(t))$, where $T_i(t)$ denotes the number of times that the i-th channel was selected by the user up to the slot t. $B_i(t, T_i(t))$ is mainly based on the exploitation, $X_i(T_i(t))$[1], and the exploration, $A_i(t, T_i(t))$:

$$B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t)) \tag{4.1}$$

The two factors $X_i(T_i(t))$ and $A_i(t, T_i(t))$ can be defined as follows:

$$X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{j=1}^{t} r_i(j) \tag{4.2}$$

$$A_i(t, T_i(t)) = \sqrt{\frac{\alpha \ln(t)}{T_i(t)}} \tag{4.3}$$

where $r_i(j)$ stands for the reward obtained from the i-th channel at the instant j, $1 \leq j \leq t$; $r_i(j) = 1$ if the channel i is free, and 0 otherwise; α represents the exploration-exploitation factor.

[1] $X_i(T_i(t))$ can be also considered as the expected reward

Besides UCB1, e-greedy represents another important algorithm first proposed in [47]. According to a recent version of e-greedy proposed in [40], the user selects a random channel if χ (a random variable $\in$ [0,1]) $< \epsilon_t$ = min $\{1, \frac{H}{t}\}$ and H is a constant number, else SU selects the most vacant channel.

During the learning epoch, the exploration phase represents a fundamental step to gather information about the vacancy probabilities of channels but it should loss its importance over time[2]. Indeed, the exploration factor $A_i(t, T_i(t))$ in UCB1 plays an important role in order to explore the vacancy of channels. Moreover, this factor has the same weight at any given time up to the total number of slots. On the other hand, in the case of e-greedy, the exploitation and exploration are separated. Indeed, the user explores the vacancy of channels if a uniform random variable $\chi < \epsilon_t$, where $\epsilon_t = \min\{1, \frac{H}{t}\}$ otherwise, the user exploits the gathered information by selecting the channel with the highest expected reward $X_i(T_i(t))$. Subsequently, the user may select many bad channels in the learning phase, i.e. $\chi < \epsilon_t$. Moreover, due to the random selection process during the exploration phase, a large number of collisions may occur and many transmissions can be lost. To solve the mentioned limitations of UCB1 and e-greedy, we propose e-UCB that may learn the vacancy of channels faster than UCB1 and e-greedy. Similarly to UCB1 and e-greedy, in e-UCB, each channel has assigned an index $B_i(t, T_i(t))$ and at each time slot, the user tries to select the channel with the highest index $B_i(t, T_i(t))$ if $\chi < \epsilon_t$, otherwise the user selects the channel with the highest expected reward $X_i(T_i(t))$, see algorithm 5.

4.2.2 Regret Analysis For e-UCB

In this section, we investigate the upper bound of regret for e-UCB for a single user. Let C be the number of independent identically distributed (I.I.D.) channels and $\Gamma = \{\mu_i\}$ stands for the availability vector. For a single user, the regret $R(n, \beta)$ (i.e. the loss of reward by selecting a non-optimal channel) up to the total number of slots n under a given policy, β, can be expressed as follows:

$$R(n, \beta) = n\mu_1 - \sum_{t=1}^{n} \mu_i^{\beta}(t) \tag{4.4}$$

where $n\mu_1$ means that the channel μ_1 has always been selected up to time n in an ideal scenario; $\mu_i^{\beta}(t)$ is the mean of reward (considered as an estimator

[2]The impact of the exploration phase during and after the learning period has not been considered in the well-known algorithms such as: UCB1 or e-greedy.

Algorithm 5: e-UCB for a single user

Input: H, C, n,
H: the exploration factor,
C: the number of channels,
n: the total number of slots,
Parameters: $T_i(t), A_i(t, T_i(t))$,
$T_i(t)$: the number of time slots the channel is sensed up to time t,
$A_i(t, T_i(t))$: the exploration contribution of channels that depends on $T_i(t)$ and t,
Output: $B_i(t, T_i(t))$, $X_i(T_i(t))$:
$B_i(t, T_i(t))$: the index assigned for channels,
$X_i(T_i(t))$: the exploitation contribution of channels that depends on $T_i(t)$,
Initialization
for t = *1 to* C **do**
  SU senses each channel once,
  SU updates $B_i(t, T_i(t))$, $X_i(T_i(t))$, $A_i(t, T_i(t))$ according to eq. (4.1), (4.2), (4.3),
for t = *C+1 to* n **do**
  if $\chi < \epsilon_t$, **then**
    $a_t = \arg\max_i \; B_i\big(t-1, T_i(t-1)\big)$,
  else
    $a_t = \arg\max_i \; X_i\big(T_i(t-1)\big)$,
  $T_i(t) + +$,
  SU updates $B_i(t, T_i(t))$, $X_i(T_i(t))$, $A_i(t, T_i(t))$ according to eq. (4.1), (4.2), (4.3)

to vacancy probability) obtained from the i^{th} channel selected at time slot t using a given MAB algorithm β. Let $T_i(n)$ denote the total number of times that the i-th channel was selected by the user up to the total time of slots n. Due to hardware constraints, we suppose that the user can only sense one channel at each time slot, then:

$$\sum_{i=1}^{C} T_i(n) = n$$

Hereinafter, we show that the upper bound of regret for e-UCB achieves a logarithmic asymptotic behavior. So, after a finite number of time slots, the user can identify the best channel, μ_1. The regret in eq. (4.4) can be expressed

as follows:

$$\begin{aligned} R(n,\beta) &= n\mu_1 - \sum_{i=1}^{C} T_i[n]\mu_i \\ &= \sum_{i=1}^{C} E\big[T_i(n)\big]\mu_1 - \sum_{i=1}^{C} E\big[T_i(n)\big]\mu_i \\ &= \sum_{i=1}^{C} E\big[T_i(n)\big]\Delta_i \end{aligned} \tag{4.5}$$

where $E[.]$ is the expectation and $\Delta_i = \mu_1 - \mu_i$. According to e-UCB, the user selects each channel once during the initialization phase and every time $a_t = i$; then, $T_i(n)$ can be written as follows:

$$T_i(n) = 1 + \sum_{t=C+1}^{n} \mathbb{1}_{\{a_t=i\}} \tag{4.6}$$

where $\mathbb{1}_{\{a_t=i\}}$ equals 1 if $a_t = i$ and 0 otherwise. Up to time n, the user may select each channel at least l times; then, according to (4.6), $T_i(n)$ can be bounded as follows:

$$T_i(n) \leq l + \sum_{t=C+1}^{n} \mathbb{1}_{\{a_t=i;T_i(t-1)\geq l\}} \tag{4.7}$$

In e-UCB, the user may select the i^{th} non-optimal channel either in the exploration or exploitation phases. Let M_i and N_i be the events that the user selects the i^{th} channel during exploration and exploitation respectively, and let $\mathbb{D}$ be the event that $T_i(t-1) \geq l$. Then, $T_i(n)$ can be expressed as follows:

$$T_i(n) \leq l + \sum_{t=K+1}^{n} \mathbb{1}_{\{M_i(t);\mathbb{D}\}} + \sum_{t=K+1}^{n} \mathbb{1}_{\{N_i(t);\mathbb{D}\}} \tag{4.8}$$

In the above equation, the second and third terms follow the Bernoulli distribution $\Big($i.e. $E\{X\} = p\{X=1\}$ where X is a random variable in $\{0,1\}\Big)$. In this case, the expectation of $T_i(n)$ can be written as:

$$E[T_i(n)] \leq l + \sum_{t=K+1}^{n} \underbrace{p\{M_i(t);\mathbb{D}\}}_{\mathbb{A}} + \sum_{t=K+1}^{n} \underbrace{p\{N_i(t);\mathbb{D}\}}_{\mathbb{B}} \tag{4.9}$$

According to e-greedy, the user selects the i^{th} channel during the exploration phase if at $(t-1)$, $B_i(t-1, T_i(t-1)) > B_1(t-1, T_1(t-1))$. Subsequently, $\mathbb{A}$ can be expressed as follows:

$$\mathbb{A} = p\{\chi < \epsilon_t; B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1)); \mathbb{D}\}$$

The event $\chi < \epsilon_t$ in the above equation is independent of the selection procedure. Then, we obtain:

$$\mathbb{A} = \epsilon_t \times p\{B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1)); \mathbb{D}\}$$

So, we get:

$$\mathbb{A} \leq 2H \times t^{-2\alpha+1}$$

Proof. appendix A □

According to Cauchy theorem [95], a series of the form $\sum_{t=1}^{n} t^{-2\alpha+1}$ can converge if $\alpha > 1$. Let $\alpha = 2$ (in order to achieve a balance between the exploration-exploitation phases), then we obtain:

$$\sum_{t=C+1}^{n} \mathbb{A} \leq 2H \times \sum_{t=1}^{n} t^{-3} \leq \frac{\pi^2 H}{3}$$

Let us find an upper bound of $\mathbb{B}$ in eq. (4.9) that referred to the the probability to access the i-th channel in the exploitation phase. Indeed, during the exploitation phase, the user may select the i-th channel at time slot t whereas $X_i(T_i(t-1)) > X_1(T_1(t-1))$ at $t-1$. Then, we get the following inequality:

$$\mathbb{B} = p\{\chi \geq \epsilon_t; X_i(T_i(t-1)) \geq X_1(T_1(t-1)); \mathbb{D}\} \tag{4.10}$$

Since the two events $\chi \geq \epsilon_t$ and $X_i(T_i(t-1)) \geq X_1(T_1(t-1))$ are completely independent, we obtain:

$$\mathbb{B} = (1-\epsilon_t) \times p\{X_i(T_i(t-1)) \geq X_1(T_1(t-1)); \mathbb{D}\} \tag{4.11}$$

The probability p in the above equation can be bounded as follows:

$$p\{X_i(T_i(t-1)) \geq X_1(T_1(t-1)); \mathbb{D}\} \leq Y + Z \tag{4.12}$$

where $Y = p\{X_i(T_i(t-1)) \geq a; \mathbb{D}\}$, $Z = p\{X_1(T_1(t-1)) \leq a; \mathbb{D}\}$, and a is a constant number that can be chosen as: $a = \frac{\mu_1+\mu_i}{2} = \mu_1 - \frac{\Delta_i}{2} = \mu_i + \frac{\Delta_i}{2}$.
Let us first consider, the first term of eq. (4.12):

$$\begin{aligned} Y &= \sum_{y=l}^{n} p\{X_i(T_i(t-1)) \geq \mu_i + \frac{\Delta_i}{2}; T_i(t-1) = y\} \\ &= \sum_{y=l}^{n} p\{X_i(y) \geq \mu_i + \frac{\Delta_i}{2}; T_i(t-1) = y\} \\ &\leq \sum_{y=l}^{n} p\{X_i(y) \geq \mu_i + \frac{\Delta_i}{2}\} \end{aligned} \tag{4.13}$$

Using the Chernoff-Hoeffding theorem in [96][3], we can upper bound the above equation as follows:

$$Y \leq \sum_{y=l}^{n} \exp^{-\frac{2\Delta_i^2 y^2}{4y}} \leq n \exp^{\frac{-l\Delta_i^2}{2}}$$

According to the proof provided in appendix A, we consider $l = \frac{8\ln(n)}{\Delta_i^2}$. So, we get:

$$Y \leq n \exp^{-4\ln n} = \frac{1}{n^3} \tag{4.14}$$

The upper bound of Z can be expressed as: $Z \leq \frac{1}{n^3}$

Proof. appendix B □

Finally, $E[T_i(n)]$ can be upper bounded by:

$$E[T_i(n)] \leq \frac{8\ln n}{\Delta_i^2} + \frac{\pi^2 H}{3} + \frac{2}{n^3} \tag{4.15}$$

From the above inequation, we conclude that the user plays each arm no more than $\frac{8\ln n}{\Delta_i^2}$ plus a constant number. Finally, based on eq (4.5) and (4.15), the regret of e-UCB, $R(n, e\text{-UCB})$, can be upper bounded by the following equation:

$$R(n, e\text{-UCB}) = 8\ln n \sum_{i=2}^{C} \frac{1}{\Delta_i} + \left(\frac{\pi^2 H}{3} + \frac{2}{n^2}\right) \sum_{i=1}^{C} \Delta_i \tag{4.16}$$

[3]According to [96], Chernoff-Hoeffding theorem is stated as follows: Let $X_1, ..., X_n$ be random variables in $\{0, 1\}$, and $E[X_t] = \mu$, and let $S_n = \sum_{i=1}^{n} X_i$. Then $\forall\ a \geq 0$, we have $P\{S_n \geq n\mu + a\} \leq \exp^{\frac{-2a^2}{n}}$ and $P\{S_n \leq n\mu - a\} \leq \exp^{\frac{-2a^2}{n}}$.

4.3 Exploration-Exploitation Dilemma of UCB

4.3.1 Lower Exploration Impact with AUCB

In this section, we propose an improved version of UCB called Arctan-UCB (AUCB). AUCB achieves better performance with respect to previous versions of UCB. In the previous section, the exploration factor of *e*-UCB becomes approximately equal to zero after the learning period in which the user gets sufficient information about the vacancy probabilities of channels. While in AUCB, and after the learning period, the impact of the exploration factor decreases without reaching zero as in *e*-UCB. Indeed, in a dynamic environment in which the availabilities of channels change over time, the exploration factor should keep some importance in order to follow dynamic channels, and thus be adapted to this environment.
Hereinafter, we prove analytically that the regret for a single user (the analytically proof for multiple users is included in the next chapter) can achieve a logarithmic asymptotic behavior with respect to time. Subsequently, a SU may quickly find and access the optimal channel (the most vacant one), and then maximizes its transmission time and rate. In the exploration factor of UCB1, $A_i(t, T_i(t))$, a non-linear function is used in order to ensure the convergence towards the best channel:

$$A_i(t, T_i(t)) = \sqrt{\frac{\alpha \ln(t)}{T_i(t)}} \tag{4.17}$$

where α denotes to the exploration-exploitation factor. This latter has large impact on the behavior of many versions of UCB (e.g. UCB1, UCB2, etc.). For instance, decreasing the factor α may reduce the effect of the exploration factor $A_i(t, T_i(t))$ of UCB1. Moreover, by selecting the channel with the highest index $B_i(t, T_i(t))$, an additional weight can be added to the exploitation $X_i(T_i(t))$. On the other hand, by increasing the value of α, the algorithm spends more time to gather information about the vacancy probabilities of channels. For a large value of α, the expected reward $X_i(T_i(t))$ becomes more and more close to the vacancy probabilities of channels. According to several studies [40, 92, 93, 94], the value of α should be in the range]1, 2] in order to get a balance between exploration-exploitation phases. The impact of α on UCB1 is widely studied in the literature [92, 97, 98]. In this section, we focus on another way to influence the exploration factor by using another non-linear function. Indeed, the **square-root** function introduced in eq. (4.17) is

widely considered [33, 40, 93, 94] for the following reasons:

- Being a positive function with respect to time t.
- It increases non-linearity which allows to restrict the effect of the exploration after the learning period.

In the case of UCB1, the exploration factor has the same weight at any given time. Thus, the big challenge remains to reduce the effect of the exploration factor after the learning period using another increasing non-linear function with the following features:

- It should have a high derivative with respect to time at the beginning to boost the exploration factor during the learning phase in order to accelerate the estimation of channels availability.
- It should also have a strong asymptotic behavior in order to restrict the exploration factor $A_i(t, T_i(t))$ under a certain limit, after the user has collected some information about the vacancy probabilities of channels.

Our study proved that the exploration factor can be adjusted by using the **arctan** function which is endowed with the above features. Indeed, an improved convergence rate to the best choice can be reached with the **arctan** compared to the one obtained with the **square-root**. On the one hand, the effect of the exploration factor $A_i(t, T_i(t))$ can be reduced after the learning phase. In the same way, maximizing the index $B_i(t, T_i(t))$ can produce some additional weight to the exploitation factor $X_i(T_i(t))$. Like the case of e-UCB, we also proved that the regret of AUCB achieves a logarithmic asymptotic behavior.

4.3.2 Regret Analysis

In this section, we prove that the upper bound of regret of AUCB can achieve a logarithmic asymptotic behavior while providing better results compared to e-UCB. According to eq. (4.5) the regret for a single user can be expressed as follows:

$$R(n, \beta) = \sum_{i=1}^{C} E\big[T_i(n)\big]\Delta_i \tag{4.18}$$

As the vacancy probabilities of channels are supposed to be constant, then the upper bound of $E\big[T_i(n)\big]$ can imply the upper bound of the regret. $T_i(n)$ can be upper bounded by the following expression (also shown is eq (4.7)):

$$T_i(n) \leq l + \sum_{t=C+1}^{n} \mathbb{1}_{\{a_t=i; T_i(t-1)\geq l\}} \tag{4.19}$$

As AUCB selects at each time slot the channel with the highest index obtained in the previous slot, the user may access, at the slot t, a non-optimal channel if the index of this channel at $(t-1)$, $B_i(t-1, T_i(t-1))$, is higher than the index of the best channel $B_1(t-1, T_1(t-1))$. In this case, we can develop further eq (4.19) as follows:

$$T_i(n) \leq l + \sum_{t=C+1}^{n} \mathbb{1}_{\{B_1(t-1,T_1(t-1))<B_i(t-1,T_i(t-1)) \text{ and } T_i(t-1)\geq l\}} \tag{4.20}$$

The index of channels $B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t))$ is based on:

- The exploitation factor $X_i(T_i(t))$.
- The exploration factor $A_i(t, T_i(t))$. This factor under AUCB is defined as follows: $A_a(t, T_i(t)) = \arctan\left(\frac{\alpha \ln(t)}{T_i(t)}\right)$,

Using eq (4.20), we can prove that:

$$T_i(n) \leq l + \sum_{t=C+1}^{n} 1_{\{X_1(T_1(t-1))+A_a(t-1,T_1(t-1))< X_i(T_i(t-1))+A_a(t-1,T_i(t-1)) \text{ and } T_i(t-1)\geq l\}} \tag{4.21}$$

The summation argument in the above equation follows Bernoulli's distribution. In this case, the expectation of $T_i(n)$ should satisfy the following constraint:

$$E[T_i(n)] \leq l + \sum_{t=C+1}^{n} P\Big\{ X_1(T_1(t-1)) + A_a(t-1, T_1(t-1)) < X_i(T_i(t-1)) + A_a(t-1, T_i(t-1)) \text{ and } T_i(t-1) \geq l \Big\} \tag{4.22}$$

The probability in eq (4.22) becomes:

$$Prob = P\Big\{ X_1(T_1(t-1)) - X_i(T_i(t-1)) \leq \arctan\left(\frac{\alpha \ln(t)}{T_i(t-1)}\right) - \arctan\left(\frac{\alpha \ln(t)}{T_1(t-1)}\right) \text{ and } T_i(t-1) \geq l \Big\} \tag{4.23}$$

After the learning period where $T_i(t-1) \geq l$, the user will have a good estimation of channels availability and thus may access regularly the best channel. Therefore, $T_i(t-1) \ll T_1(t-1)$; and $\arctan\left(\frac{\alpha \ln(t)}{T_i(t-1)}\right) \geq \arctan\left(\frac{\alpha \ln(t)}{T_1(t-1)}\right)$. Using the asymptotic behaviors of the non-linear functions **sqrt** and **arctan**, the probability in eq (4.23) becomes bounded by:

$$Prob \leq P\Bigg\{ X_i(T_1(t-1)) - X_i(T_i(t-1)) \leq \sqrt{\frac{\alpha \ln(t)}{T_i(t-1)}} - \sqrt{\frac{\alpha \ln(t)}{T_1(t-1)}} \text{ and } T_i(t-1) \geq l \Bigg\} \quad (4.24)$$

By taking the minimum value of $X_i(T_1(t-1)) + \sqrt{\frac{\alpha \ln(t)}{T_1(t-1)}}$ and the maximum value of $X_i(T_i(t-1)) + \sqrt{\frac{\alpha \ln(t)}{T_i(t-1)}}$ at each time slot, we can upper bound eq (4.22) by the following equation:

$$E[T_i(n)] \leq l + \sum_{t=C+1}^{n} P\Bigg\{ \min_{0<S_1<t} \left[X_1(S_1) + \sqrt{\frac{\alpha \ln(t)}{S_1}} \right] \leq \max_{l \leq S_i < t} \left[X_i(S_i) + \sqrt{\frac{\alpha \ln(t)}{S_i}} \right] \Bigg\} \quad (4.25)$$

where $S_i \geq l$ to fulfill the condition $T_i(t-1) \geq l$. Then we obtain:

$$E[T_i(n)] \leq l + \sum_{t=1}^{n} \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} P\Big\{ X_1(S_1) + A_1(t,S_1) < X_i(S_i) + A_i(t,S_i) \Big\} \quad (4.26)$$

The above probability can be upper bounded by:

$$P\Big\{ X_1(S_1) + A_1(t,S_1) < X_i(S_i) + A_i(t,S_i) \Big\} \leq P\Big\{ X_1(S_1) + A_1(t,S_1) \leq \mu_1 \Big\} + P\Big\{ \mu_1 < \mu_i + 2A_i(t,S_i) \Big\} + P\Big\{ X_i(S_i) + A_i(t,S_i) \geq \mu_i + 2A_i(t,S_i) \Big\} \quad (4.27)$$

Using the ceiling operator $\lceil \rceil$, let $l = \lceil \frac{4\alpha \ln(n)}{\Delta_i^2} \rceil$, where $\Delta_i = \mu_1 - \mu_i$ and $S_i \geq l$, then the event $\mu_1 < \mu_i + 2A_i(t,S_i)$ in eq (4.27) becomes false, in fact:

$$\begin{aligned}\mu_1 - \mu_i - 2A_i(t, S_i) &= \mu_1 - \mu_i - 2\sqrt{\frac{\alpha \ln(t)}{S_i}} \\ &\geq \mu_1 - \mu_i - 2\sqrt{\frac{\alpha \ln(n)}{l}} \\ &\geq \mu_1 - \mu_i - \Delta_i = 0\end{aligned}$$

Based on eq (4.26) and (4.27), we obtain:

$$E[T_i(n)] \leq \left\lceil \frac{4\alpha \ln(n)}{\Delta_i^2} \right\rceil + \sum_{t=1}^{n} \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} \left\{ P\left\{ X_1(S_1) \leq \mu_1 - A_1(t, S_1) \right\} + P\left\{ X_i(S_i) \geq \mu_i + A_i(t, S_i) \right\} \right\} \quad (4.28)$$

Using Chernoff-Hoeffding bound [96], we can prove that:

$$P\left\{ X_1(S_1) \leq \mu_1 - A_1(t, S_1) \right\} \leq \exp^{\frac{-2}{S_1}\left[S_1\sqrt{\frac{\alpha \ln(t)}{S_1}}\right]^2} = t^{-2\alpha} \quad (4.29)$$

$$P\left\{ X_i(S) \geq \mu_i + A_i(t, S_i) \right\} \leq \exp^{\frac{-2}{S_i}\left[S_i\sqrt{\frac{\alpha \ln(t)}{S_i}}\right]^2} = t^{-2\alpha} \quad (4.30)$$

The two equations above and eq (4.28) lead us to:

$$\begin{aligned}E[T_i(n)] &\leq \left\lceil \frac{4\alpha \ln(n)}{\Delta_i^2} \right\rceil + \sum_{t=1}^{n} \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} 2t^{-2\alpha} \\ &\leq \frac{4\alpha \ln(n)}{\Delta_i^2} + 1 + 2\sum_{t=1}^{n} t^{-2\alpha+2}\end{aligned} \quad (4.31)$$

According to the Cauchy theorem [95], a series of the form $\sum_{t=1}^{n} t^{-2\alpha+2}$ converges if $\alpha > \frac{3}{2}$. Let $\alpha = 2$ (in order to achieve a balance between exploration and exploitation phases), then we obtain:

$$\begin{aligned}E[T_i(n)] &\leq \frac{8 \ln(n)}{\Delta_i^2} + 1 + 2\sum_{t=1}^{n} t^{-2} \\ &\leq \frac{8 \ln(n)}{\Delta_i^2} + 1 + \frac{\pi^2}{3}\end{aligned} \quad (4.32)$$

Proof. appendix C □

Finally, we obtain an upper bound of the regret of AUCB, $R(n, AUCB)$:

$$R(n, \text{AUCB}) \leq 8 \sum_{i=2}^{C} \left[\frac{\ln(n)}{\Delta_i} \right] + \left(1 + \frac{\pi^2}{3} \right) \sum_{i=1}^{C} \Delta_i \tag{4.33}$$

According to the above equation, AUCB can yield better results compared to e-UCB in which the upper bound of regret of e-UCB is higher than the one of AUCB (see eq (4.16) and (4.33) for the upper bound of e-UCB and AUCB respectively).

4.4 MAB Algorithms with Multiple Users

In this section, we extend the proposed MAB algorithms (AUCB and e-UCB) to consider the case of multiple users trying to learn collectively the channels availability. We consider two possible operation modes: Cooperative and competitive access.

4.4.1 Cooperative Side Channel Policy

We extend our proposed AUCB using the cooperative side channel policy presented in Chapter 3 (Section 3.5) in which users can exchange simple information with a very low information rate. Moreover, the side channel policy takes into consideration the priority access in which, after a finite number of time slots, the k-th user usually accesses the k-th highest entry in the index $B_i(t, T_i(t))$ (i.e. the k-th best channel). Based on the side channel policy, the priority user should broadcast the choice of his channel to other users in order to avoid any collision. It should be noticed that the broadcast packet of the priority user has the risk to be lost, then a collision may occur among users and the amount of regret increases. In order to simplify finding an upper bound of AUCB under side channel policy, we consider an error-free channel in which broadcast packet losses do not occur. However, considering no error-free can produce collisions among users and results in adding some constant values to the regret.
Hereinafter, we show that the upper bound of the regret has a logarithmic asymptotic behavior. In the case of multiple users, the regret for U users

under a policy β, without considering the collision among users can be expressed as follows[4]:

$$R(n, U, \beta) = n \sum_{k=1}^{U} \mu_k - \sum_{i=1}^{C} \sum_{j=1}^{U} \mu_i E\left[T_{i,j}(n)\right] \tag{4.34}$$

where μ_k represents the k-th best channel; n is the total number of slots and $T_{i,j}(n)$ stands for the number of times that the user j accesses the channel i up to time n.

Let $E[T_i(n)] = \sum_{j=1}^{U} E[T_{i,j}(n)]$ be the number of times that the i-th channel is sensed by all users up to time n. By considering that each user can sense one channel at each time slot, we obtain:

$$\sum_{i=1}^{C} E[T_i(n)] = nU \Rightarrow n = \frac{1}{U} \sum_{i=1}^{C} E[T_i(n)]$$

In this case, the regret can be expressed as follows:

$$\begin{aligned} R(n, U) &= \frac{1}{U} \sum_{i=1}^{C} E[T_i(n)] \sum_{k=1}^{U} \left(\mu_k - \mu_i\right) \\ &= \frac{1}{U} \sum_{i=1}^{C} E[T_i(n)] \sum_{k=1}^{U} \Delta_{(k,i)} \end{aligned} \tag{4.35}$$

where $\Delta_{(k,i)} = \mu_k - \mu_i$. To simplify the above equation, we consider the summation over worst and best channels as follows:

$$R(n, U) = \frac{1}{U} \sum_{i=1}^{U} E[T_i(n)] \sum_{k=1}^{U} \Delta_{(k,i)} + \frac{1}{U} \sum_{i=U+1}^{C} E[T_i(n)] \sum_{k=1}^{U} \Delta_{(k,i)} \tag{4.36}$$

The first term of the regret in eq (4.36) equals 0 whenever $i \in$ best channels, then we obtain:

$$R(n, U) = \frac{1}{U} \sum_{i=U+1}^{C} E[T_i(n)] \sum_{k=1}^{U} \Delta_{(k,i)} \tag{4.37}$$

[4]Under no-collision assumption, the regret represents the difference between the reward collected in the ideal scenario where each user accesses all-time his dedicated channel and the one obtained using a given policy.

For a single user, using AUCB or *e*-UCB, we previously found an upper bound of $E[T_i(n)]$ in eq (4.15) and (4.33).
First, let us find an upper bound of AUCB under Side Channel policy. Based on eq. (4.33), $E[T_i(n)]$ for U users becomes:

$$E[T_i(n)] = \sum_{j=1}^{U} E[T_{i,j}] \leq \sum_{j=1}^{U} \left[\frac{8\ln(n)}{\Delta^2_{(j,i)}} + 1 + \frac{\pi^2}{3} \right] \tag{4.38}$$

and the regret of our AUCB under the Side Channel policy can be written as:

$$R(n, U, \text{AUCB}) \leq \sum_{i=U+1}^{C} \sum_{k=1}^{U} \left[\frac{8\ln(n)}{\Delta_{(k,i)}} + \Delta_{(k,i)}(1 + \frac{\pi^2}{3}) \right] \tag{4.39}$$

According to the above equation, the global regret of AUCB under the Side Channel policy has a logarithmic upper bound, which means that after a period of time, each user will have a good estimation of the channels availability and each one accesses a channel based on his rank.
Similarly to *e*-UCB, the upper bound of regret for multiple users can be expressed as follows:

$$R(n, U, e\text{-UCB}) \leq \sum_{i=U+1}^{C} \sum_{k=1}^{U} \left[\frac{8\ln n}{\Delta_{(k,i)}} + \left(\frac{\pi^2 H}{3} + \frac{2}{n^2} \right) \Delta_{(k,i)} \right] \tag{4.40}$$

According to the above upper bound expression of AUCB or *e*-UCB, we conclude that AUCB can achieve a better result while the users quickly reach their dedicated channels.

4.4.2 Competitive Random Rank Policy

The random access in OSA, cooperative or competitive, is widely suggested in the literature and several policies have been proposed to manage a secondary network. As a matter of fact, if each user tries to reach selfishly the best channel using one of the mentioned MAB algorithms (*e*-UCB or AUCB), then, a large number of collisions may occur and many transmission data will be lost. Hence, we introduce a simple (but optimal) policy from the literature for the competitive random access. This policy, called Random Rank, was initially proposed under UCB1 (see algorithm 7). During the initialization, and based on the Random Rank policy, each user draws a rank from the set $\{1, ..., U\}$ where U is the number of users in the network. All users have the same chance to sense and access the i-th channel. As in UCB1, each channel

has an assigned index $B_i(t, T_i(t))$, and the user should make a decision based on his generated rank. For instance, a user, with a rank equals to 4, should access the fourth highest entry in the index $B_i(t, T_i(t))$.

Algorithm 6: Random Rank policy

Input: C, U, n, $\xi_k(t)$,
C: number of channels,
U: number of users,
n: total number of slots,
$\xi_k(t)$: indicates a collision for the k^{th} user at time t,
Parameters: $T_i(t)$, $A_i(t, T_i(t))$,
$T_i(t)$: number of time slots the channel is sensed up to time t,
$A_i(t, T_i(t))$: the exploration contribution of channels that depends on $T_i(t)$ and t,
Output: $B_i(t, T_i(t))$, $X_i(T_i(t))$:
$B_i(t, T_i(t))$: the index assigned for channels,
$X_i(T_i(t))$: the exploitation contribution of channels that depends on $T_i(t)$,
Initialization:
$k = 0$,
for $t = 1$ *to* C **do**
    SU_k senses each channel once,
    SU_k updates $B_i(t, T_i(t))$, $X_i(T_i(t))$, $A_i(t, T_i(t))$,
for $t = C + 1$ *to* n **do**
    SU_k senses the k-th highest entry in the index $B_i(t, T_i(t))$,
    if $\xi_k(t) = 1$ **then**
        SU_k generates randomly a new rank from the set $\{1, ..., U\}$,
    else
        SU_k keeps his rank.
    SU_k updates $B_i(t, T_i(t))$, $X_i(T_i(t))$, $A_i(t, T_i(t))$

4.5 Simulation and results

In this section, we evaluate the performance of e-UCB and A-UCB for single and multiple users over 9 channels with the following availability probabilities:

$$\Gamma = [0.9\ 0.8\ 0.7\ 0.6\ 0.5\ 0.4\ 0.3\ 0.2\ 0.1]$$

Under the two scenarios, single and multiple users, we evaluate the performance of e-UCB or AUCB using two main parameters: the regret and the percentage of access to the best channel.

4.5.1 Test for a Single User

Let us consider a SU in the secondary network trying to estimate the vacancy probabilities of channels in order to access the best one. In Fig. 4.1, we compare the regret that achieves a logarithmic asymptotic behavior for the 4 MAB algorithms: TS, AUCB, *e*-UCB, UCB. This result deals with our analytical proof which showed that the upper bound of regret has a logarithmic asymptotic behavior for *e*-UCB and AUCB. The same figure shows that TS achieves the lower regret. In fact, the good performance of TS is widely suggested for a single user and several studies found an upper bound for its optimal regret that also achieves a logarithmic asymptotic behavior. Despite its high performance for a single user, TS may not achieve a good result for multiple users as shown in the next section.

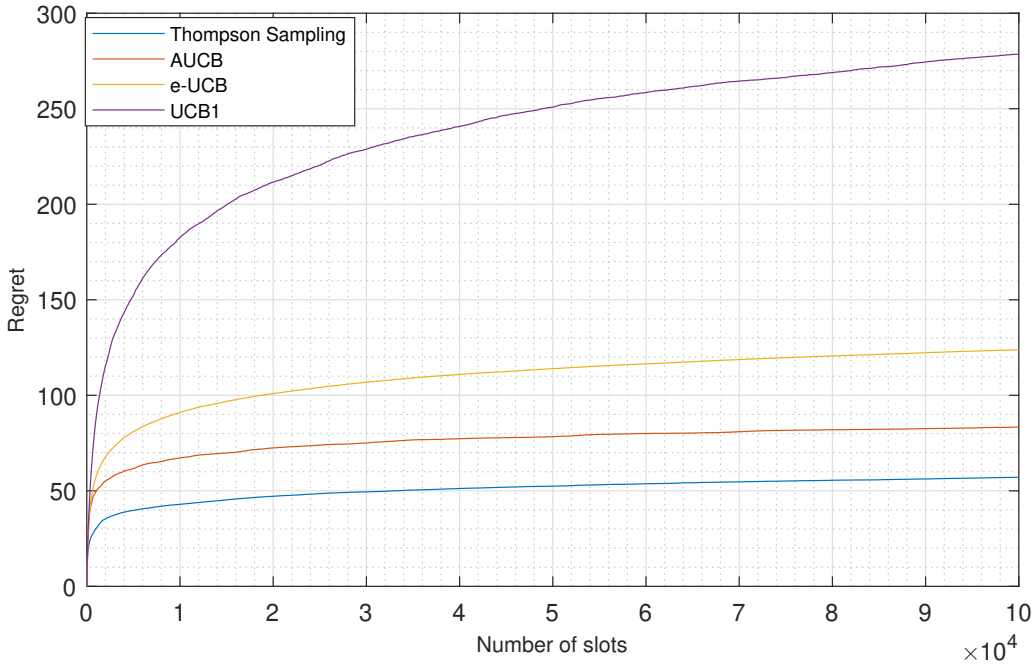

FIGURE 4.1: The regret of the 4 MAB algorithms

Fig. 4.2 compares the percentage to access the best channel using the 4 MAB algorithms. As mentioned before, TS seems to be a good algorithm that can largely exceed the performance of the state of the art MAB algorithms. For this reason, we adopt TS as a reference in order to evaluate the performance of the two proposed MAB algorithms AUCB and *e*-UCB. As it can be seen in Fig. 4.2, the 4 MAB algorithms can reach the best channel after a finite number of time slots, while TS represents the best one.

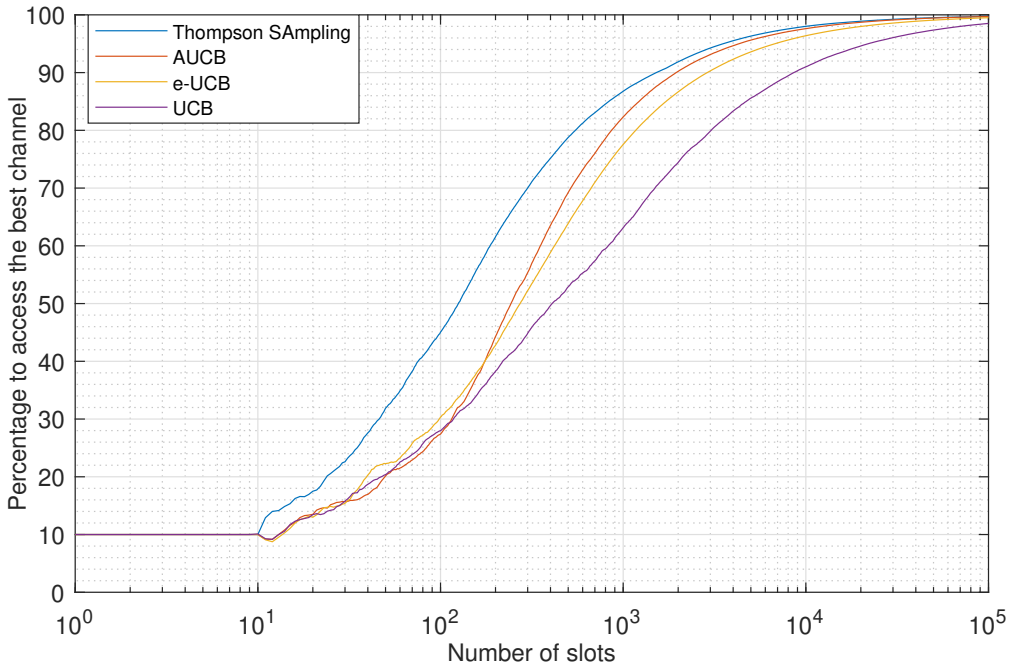

FIGURE 4.2: The selection percentage of the best channel using the 4 MAB algorithms TS, AUCB, *e*-UCB and UCB

4.5.2 Test for Multiple Users

In this section, we consider the case of multiple users trying to learn the vacancy probabilities of channels. To fix our idea, let us consider 3 SUs that want to learn and access the 3 best channels (i.e. $\mu_1 = 0.9$, $\mu_2 = 0.8$, $\mu_3 = 0.7$) under two possible scenarios: cooperative and competitive access. The cooperation among users has an important role to enhance the spectrum efficiency and to guarantee an optimal sharing of available spectrum. Indeed, the collaboration among users, on the one hand, may lead to a full and quick estimation of the vacancy probabilities of channels. On the other hand, it is necessary to mitigate the harmful interference with PUs. While in the competitive access, each user selfishly makes its own action without any cooperation with others. Despite the increasing number of collisions compared to the cooperative access, the competitive access may decrease the complexity of the network and can be useful in several scenarios.

Let us first consider the cooperative priority access based on the Side Channel policy. Fig. 4.3 represents the regret of the 4 MAB algorithms AUCB, TS, *e*-UCB and UCB1. As it can be seen, the regrets for all algorithms have logarithmic asymptotic behaviors in which the users are able to estimate the vector Γ and often access their dedicated channels. Moreover, based on the Side Channel policy in which no-collision can occur among users, the TS

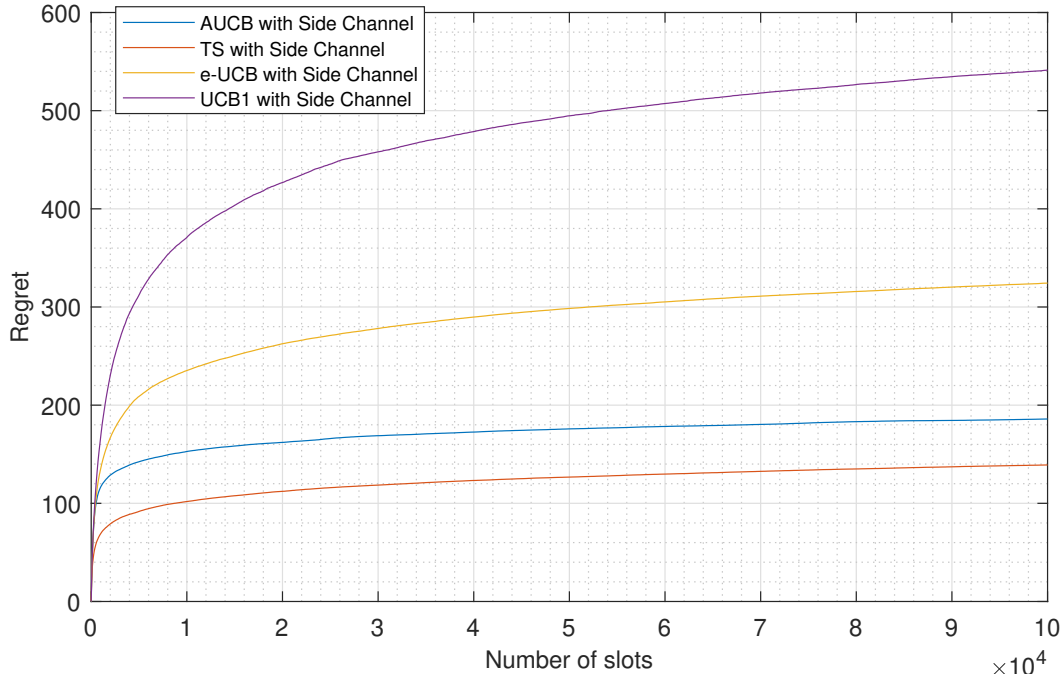

FIGURE 4.3: The regret of the 4 MAB algorithms TS, AUCB, e-UCB and UCB

achieves best results followed by AUCB, e-UCB and UCB1 respectively.

Fig. 4.4 represents the percentage of times to access the best channels by each user. As we can see, using our Side Policy policy, the users are able to converge towards their dedicated channels: SU_1 converges to the best channel "μ_1" followed by SU_2 and SU_3 to the second "μ_2" and third "μ_3" best channels respectively. Moreover, the users converge quickly to their dedicated channels under TS followed by AUCB, e-UCB and UCB1.

It remains to compare the performance of the 4 MAB algorithms under competitive access. The regrets of the mentioned algorithms are depicted in Fig. 4.5 with the Random Rank policy for multiple users. In the latter policy, the target of the k-th user is to access one of the U best channels and not a specific one as is the case of the Side Channel policy. Despite its optimal convergence for a single user or multiple cooperative users, TS may not achieve the best result for multiple competitive users as shown in Fig. 4.5. In fact, for competitive multiple users, the performance for a given MAB algorithm depends on the access of bad channels and also on the number of collisions among users. Those two factors are related to the exploration impact of MAB algorithms.

Indeed, the impact of the exploration factor should decrease after the learning period, as in AUCB, after the user has got sufficient information

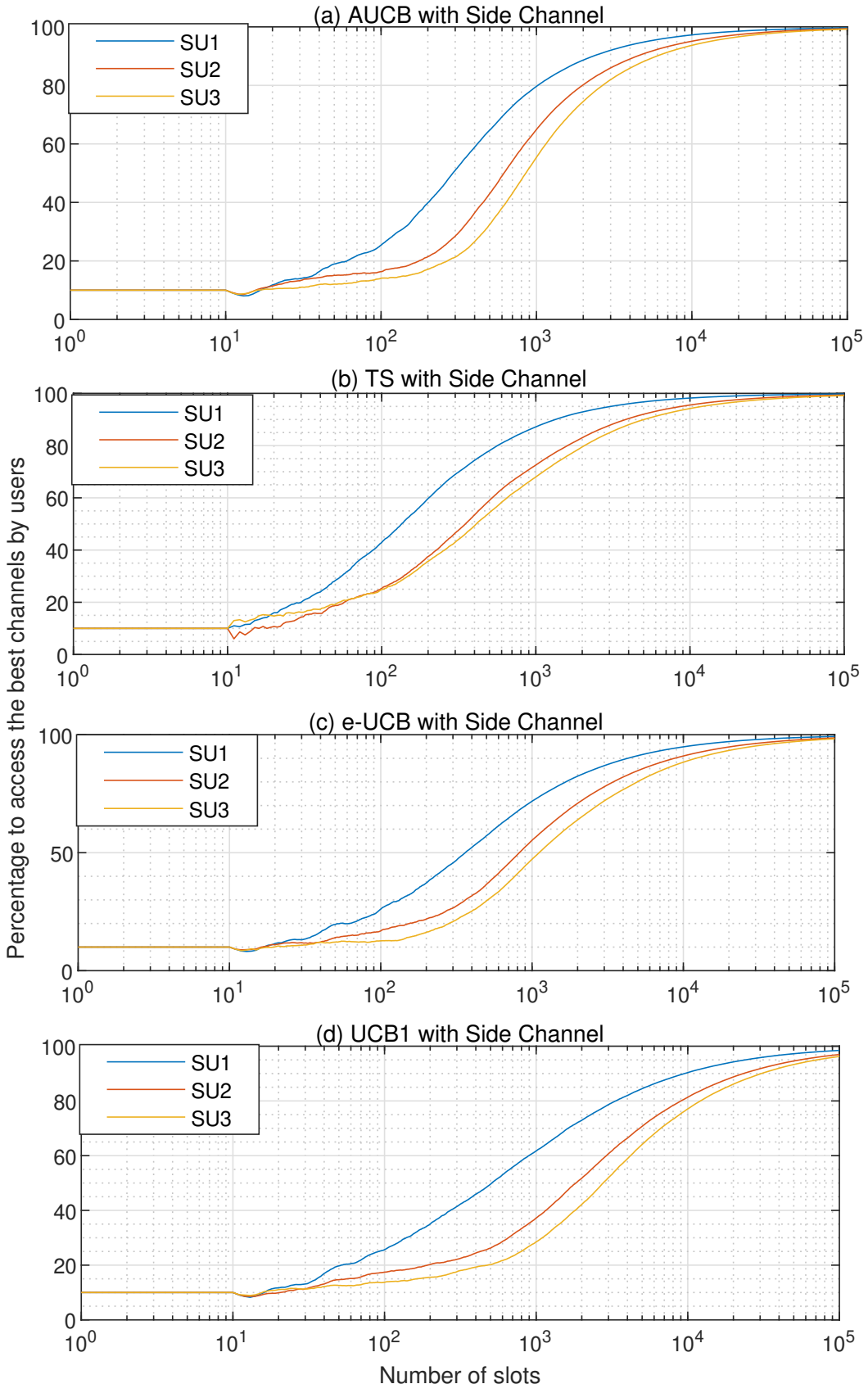

FIGURE 4.4: Access the best channels by 3 SUs using the 4 MAB algorithms AUCB, TS, *e*-UCB and UCB

about the vacancy probabilities of channels. While in the case of TS, the exploration factor is still having the same weight at any given time, which basically produces a large number of collisions compared to AUCB. This explains why, for multiple users using Random Rank policy, AUCB attains a lower regret and reaches better performance compared to TS. Fig. 4.5 also shows that TS gives better results than *e*-UCB. Indeed, due to the high exploration impact during the learning period, users access many bad channels which increases the regret. Moreover, a high exploration level may significantly increase the number of collisions among users and by the same way the global regret.

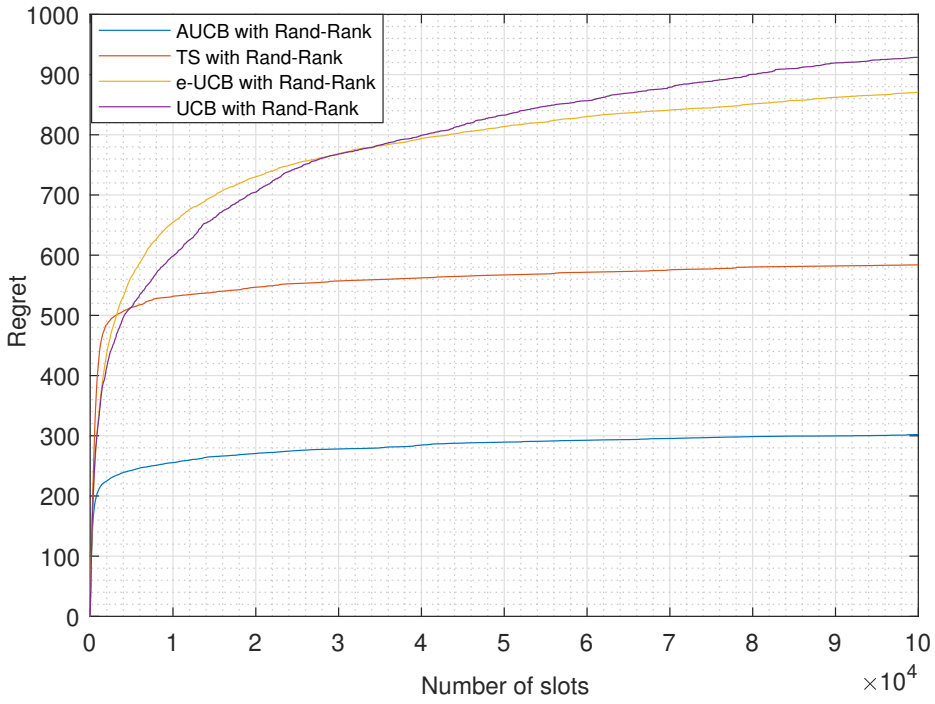

FIGURE 4.5: The regret of the 4 MAB algorithms TS, AUCB, *e*-UCB and UCB

According to our simulations, TS represents an optimal solution for a single user but not necessary for multiple users. More precisely, TS is very sensitive to the number of collisions. Indeed, a high level of collisions number among users, under a given policy, can decrease rapidly the convergence speed towards the optimal channels.

4.6 Conclusion

In this chapter, we proposed two Multi-Armed Bandit (MAB) algorithms, called Arctan-Upper Confidence Bound (AUCB) and *e*-UCB, in order to apply them in the Opportunistic Spectrum Access (OSA). In OSA, a SU tries to estimate the vacant probabilities of channels in order to reach the optimal one with the highest vacancy probability. It has been shown that the proposed MAB algorithms achieves a good result in OSA compared to several well-known MAB algorithms such as UCB1 or *e*-greedy. However, AUCB and *e*-UCB cannot exceed the performance of the Thompson Sampling (TS) that seems to exceed the state of the art of MAB algorithms. We also investigate the upper bound of regret (i.e. the loss of reward due to the selection of worse channels) that achieves a logarithmic asymptotic behavior for AUCB and *e*-UCB. According to the obtained upper bound, AUCB achieves a lower regret compared to *e*-UCB and thus, can reach the best channel faster than this latter. We also studied the performance of the proposed MAB algorithms for multiple users under two operation modes: Cooperative or competitive access. For the cooperative access, we use our proposed Side Channel policy that takes into account the priority access. We also proved the analytical upper bound of the two proposed MAB algorithms under our policy. For the competitive access, we used Random Rank, an existing policy in the literature to manage a secondary network for the random access. Through our simulations, we can say that TS represents a suitable solution for single or multiple cooperative users compared to AUCB, *e*-UCB, UCB1 and *e*-greedy. While, for the competitive access, TS may not achieve the best performance, as a result of the collisions among users, and thus our proposed AUCB achieves better results.

Chapter 5

Competitive Priority Cognitive Access

Contents

5.1 Introduction

As mentioned in previous chapters, Multi-Armed Bandit (MAB) algorithms in Opportunistic Spectrum Access (OSA) can represent a suitable solution to enhance the spectrum usage. The well-known MAB algorithms, such as: Thompson Sampling (TS), Upper Confidence Bound (UCB) and e-greedy have been firstly suggested for a single user case. Recently, several policies have been proposed to extend the MAB algorithms to consider multiple users under cooperative or Competitive accesses. Moreover, these different existing policies can be classified into two main categories: Random access or priority access. In our work, we are interested in the priority access with cooperative or competitive users access that may represent a suitable solution in a tactical network. In the previous chapters, we proposed a novel policy for the cooperative priority access; while in this chapter, we focus on the Competitive priority access. The kind of priority access was not well studied in the literature since the most recent works focus on the random access. Moreover, we particularly focus on the priority dynamic access in which the priority users can leave and enter to the network at any time, while, to the best of our knowledge, only the priority or the dynamic access are considered in several existing works. Later on, we consider the quality of service (QoS) in the secondary network. In this new model, a SU should be able to estimate the availability of channels and consider as well as their qualities. After gathering sufficient information about available channels, the users should only access optimal ones characterized by their highest availability and quality.
The organization of this chapter is as follows. Section 5.2 introduces a novel competitive priority policy, called All-Powerful Learning (APL). In section 5.3, we prove that the upper bound of the regret of APL achieves a logarithmic asymptotic behavior. In section 5.4, we study the QoS of AUCB. In section 5.5, we suggest a novel policy to consider different priority levels. Numerical results are presented in section 5.6 in which the performance of APL compared to several existing policies is shown. Finally, section 5.7 concludes the chapter.

5.2 Multiple Users for Competitive Access

In the literature, few studies have considered the cooperative access in OSA while the Competitive access is widely treated and many competitive policies have been proposed. The main advantage of the competitive access is

the reduction of the network complexity compared to a cooperative access. Indeed, in the latter case, the users should collaborate and exchange some information about their environment in order to make a collective decision which can increase the performance of the cognitive network.

The main challenge in the competitive access remains to learn collectively the vacant probabilities of channels while decreasing the number of collisions among users. In this section, we focus on the competitive access and more precisely on the priority competitive access. First, we discuss the existing policies to manage a secondary network and their drawbacks; then, we propose a novel policy for the priority dynamic access.

5.2.1 Random Access

In a random access, the main target of users is to access one of the optimal channels and not necessarily the most vacant one. Indeed, if all users try to reach the same highly vacant channel, this would cause a large number of collisions among them. A Random Rank policy represents one of the important policies for the random access [35]. Based on the Random Rank policy, each user should make an action depending on a rank generated randomly from a set $\{1, ..., U\}$ after each collision. Only collided users should seek a novel rank, while the others keep their current ranks.
Musical Chair represents another important policy for the random access [60]. In this policy, the users select the channels randomly during the learning period in order to gather information about the vacant probabilities of channels (Exploration phase). Moreover, during the learning period and based on the collision number, users should estimate their number U in the network using the following equation [60]:

$$U \approx min\left(round\left(\frac{log\left(\frac{T_0 - C_{T_0}}{T_0}\right)}{log\left(1 - \frac{1}{C}\right)} + 1\right), C\right) \tag{5.1}$$

where C is the number of channels; T_0 represents the duration of the learning period; C_{T0} is the number of collisions during T_0. After the learning period, the user accesses a random channel from the U best channels. When a collision occurs, the user chooses randomly another best channel; otherwise, he keeps selecting the same channel.
Besides Musical Chair, the Multi-user e-greedy collision Avoiding (MEGA) policy proposed in [63] can achieve a random access. MEGA policy is an extension of the e-greedy firstly proposed in [47] for a single user. According

to [60], Musical Chair achieves better results compared to MEGA. For this reason, we will compare the performance of our proposed policy to Musical Chair.

5.2.2 Priority Access

The priority access in OSA is not well considered in the literature. Moreover, and to the best of our knowledge, existing MAB algorithms cannot be used for the priority dynamic access in which other users can't enjoy a dedicated channel of a leaving user, as shown in Fig. 5.1.

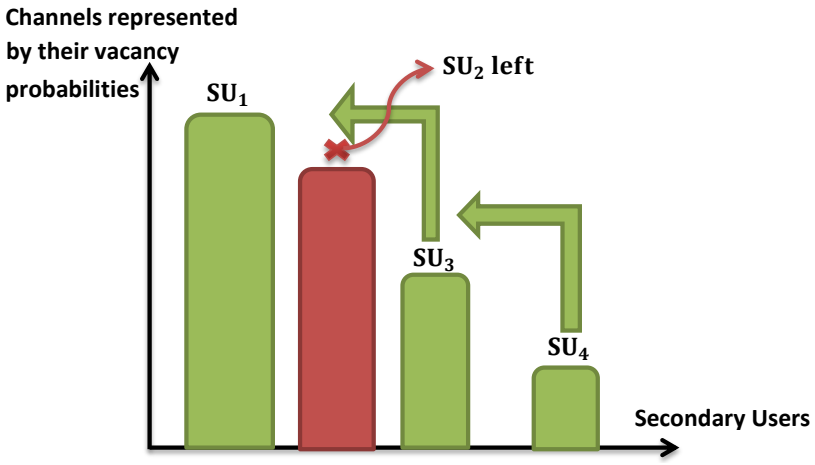

FIGURE 5.1: Priority access after a user left his dedicated channel

k-th MAB [62] and SLK (Selective learning of the K-th largest expected rewards) [99] represent two MAB policies proposed in the literature to address the priority access in the secondary network. Based on the *k*-th MAB, each user has a known rank and he should access the channels respecting his rank. The main idea consists on slotting in the time and each slot is divided into multi sub-slots depending on the user priority ranks. Therefore, the transmission time under a large number of users tends towards zero. For the higher ranking users, this fact becomes the major limitation of this policy. In addition, this policy does not consider the dynamic access and the number of users should be fixed and known in advance. Based on UCB1, the authors of [99] proposed the SLK algorithm, an efficient algorithm for the priority access, that does not have the limitation of the *k*-th MAB. However, as in the latter policy the number of users must be fixed and previously known by each user.

5.3 All-Powerful Learning for the Priority Access

In this section, we propose a new policy for the priority access. This policy enables a secondary user to learn the vacant probabilities of channels and ensures the convergence to his dedicated channel. Moreover, it can be used with all learning MAB algorithms such as: Thompson Sampling (TS), Upper Confidence Bound (UCB), AUCB, e-UCB, e-greedy, etc. We should highlight that our proposed policy does not require prior knowledge about the channels as in the case for other policies, such as: Musical Chair [60], SLK [99], k-th MAB [62], MEGA [63], Side Channel[100], etc. Indeed, existing policies to manage a secondary network suffer from one or more of the following disadvantages:

1. The number of users should be fixed and known to all users.
2. SUs should have a prior information about the number of channels.
3. Expected transmission time should be known.
4. The dynamic access is not suggested. To recall, in a dynamic access, the users can at any given time enter or leave the network.
5. Some algorithms consider a restricted dynamic access, where a SU can't leave the network during the learning or the exploration phases.
6. The vacant probabilities of channels should be static; otherwise, users cannot adapt to their environment.
7. The priority access is seldomly suggested in the literature, while the random access represents the most used model.

Unlike SLK and k-th MAB, our proposed policy for the priority access, called All-Powerful Learning algorithm (APL), doesn't suffer from the above mentioned drawbacks. As a matter of fact, SLK and k-th MAB policies suffer from the 1^{st}, 2^{nd} and 4^{th} mentioned drawbacks.

In a classical priority access, each channel has assigned an index $B_i(t)$ and the highest priority user SU_1 should sense and access the channel with the highest index $B_i(t)$ at each time slot. Indeed, the best channel, after a finite number of time slots, will have the highest index $B_i(t)$.

As the second priority user SU_2 should avoid the first best channel and try to access the second best one. To reach his goal, SU_2 should sense the first and second best channels at each time slot in order to estimate their vacant

Algorithm 7: All-Powerful Learning algorithm

Input: $k, \xi_k(t), r_i(t)$,
k: indicates the $k-th$ user or $k-th$ best channel,
$\xi_k(t)$: indicates a presence of collision for the $k-th$ user at instant t,
$r_i(t)$: indicates the state of the $i-th$ channel at instant t, $r_i(t) = 1$ if the channel is free and 0 otherwise,
Initialization
$k = 1$,
for $t = 1$ *to* C **do**
  SU_k senses each channel once,
  SU_k updates his index $B_i(t)$,
  SU_k generates a rank of the set $\{1, ..., k\}$,
  $k + 1$,
for $t = K{+}1$ *to* n **do**
  SU_k senses a channel in his index $B_i(t)$ according to his rank,
  if $r_i(t){=}1$ **then**
    SU_k transmits his data,
    if $\xi_k(t){=}1$ **then**
      SU_k regenerates his rank of the set $\{1, ..., k\}$,
    else
      SU_k keeps his previous rank,
  else
    SU_k refrains from transmitting at instant t,
  SU_k updates his index $B_i(t)$

probabilities and then access the second best channel if available. In this case, the complexity of the hardware is increased, and we conclude that a classical priority access represents a costly and impractical method to settle down each user to his dedicated channel. In the case of APL, at each time slot, the user senses a channel and transmits his data if the channel is available (see algorithm 1). In our policy, each SU_k has a prior rank, $k \in \{1, ..., U\}$, and his target is to access the k-th best channel. The major problem of the competitive priority access is that each user should selfishly estimate the vacant probabilities of the available channels. Our policy can intelligently solve this issue by making each user generate a rank around his prior rank to get information about the channels availability. For instance, if the rank generated by the k-th user equals 3 (considering that $k > 3$), then he should access the channel that has the third index, i.e. $B_3(t)$. In this case, SU_k can examinate the states of the k best channels and his target is the k-th best one.

However, if the rank created by SU_k is different than k, then he selects a channel with one the following probabilities: $\{\mu_1, \mu_2, ..., \mu_{k-1}\}$ and he may collide with a priority user, i.e. $SU_1, SU_2, ..., SU_{k-1}$. Therefore, SU_k should

avoid regenerating his rank at each time slot; otherwise, a large number of collisions may occur among users and transmitted data can be lost. So, after each collision, SU_k should regenerate his rank from the set $\{1, ..., k\}$. Thus, after a finite number of slots, each user settles down to his dedicated channel. It remains to investigate the analytical convergence of APL to verify its performance in a real radio environment.

As a second-major contribution of this chapter, we propose and develop MAUCB which is an extension of our previous algorithm AUCB presented in chapter 4. To present MAUCB, we first need to define the regret. Let us introduce the definition of the regret for multiple users that takes into consideration not only the access to the bad channels but also the collision among users. Under a policy β, the regret represents the difference between the obtained reward in the ideal scenario, $S^*(n, U)$, in which best channels are known and accessed by the SUs, and the reward obtained using a given policy, $S(n, U, \beta)$:

$$R(n, U, \beta) = S^*(n, U) - S(n, U, \beta) \tag{5.2}$$

$S^*(n, U)$ and $S(n, U, \beta)$ are defined as follows:

$$S^*(n, U) = n \sum_{k=1}^{U} \mu_k \tag{5.3}$$

$$S(n, U, \beta) = \sum_{i=1}^{C} \sum_{j=1}^{U} \mu_i \, E[P_{i,j}(n)] \tag{5.4}$$

where n stands for the total number of slots; μ_k represents the vacant probability of the k^{th} best channel; $E[.]$ is the expectation and $P_{i,j}(n)$ represents the number of times where user j is the only occupant of the channel i up to n.

Let $T_{i,j}(n)$ be the total number of times where the j^{th} user senses the i^{th} channel up to n. For the sake of notational simplicity, we consider $T_i(n) = \sum_{j=1}^{U} T_{i,j}(n)$ and $P_i(n) = \sum_{j=1}^{U} P_{i,j}(n)$ as the total number of times where the i-th channel is sensed by all users, and the total number of times where the users access the i-th channel without producing any collision up to n. Let $O_k(n)$ be the number of collisions in the k-th best channel as well $T_k(n)$ and $P_k(n)$ being respectively the total number of times where the k-th best channel is sensed by all users and the total number of times where the users access the k-th best channel without making any collision up to n. $O_k(n)$ can be expressed as follows:

$$O_k(n) = T_k(n) - P_k(n) \tag{5.5}$$

It is worth mentioning that the number of channels C should be higher than the number of active users U, otherwise:

- Using a learning algorithm to find the best chan- nels does not make any sense, since all channels need to be accessed.
- Considering that the user should sense one channel at each time slot, at least one collision may occur among users, and users cannot converge to free-collision state under any learning policy.

Subsequently, by considering that $C \geq U$ and $\mu_1 \geq \mu_i, \forall\, i$, we can upper bound the regret in eq (5.2) of MAUCB using APL as follows:

$$\begin{aligned} R_{APL}(n, U, \text{MAUCB}) &\leq n \sum_{k=1}^{U} \mu_k - \sum_{k=1}^{U} \mu_k E\left[P_k(n)\right] \\ &\leq \mu_1 \left(Un - \sum_{k=1}^{U} E\left[P_k(n)\right] \right) \end{aligned} \tag{5.6}$$

Based on the assumption that the user selects only one channel at each time slot, we obtain the following equality:

$$\sum_{i=1}^{C} \sum_{j=1}^{U} T_{i,j}(n) = \sum_{i=1}^{C} T_i(n) = Un \tag{5.7}$$

From eq (5.6) and (5.7), the regret can be bounded as follows:

$$R_{APL}(n, U, \text{MAUCB}) \leq \mu_1 \left(\sum_{i=1}^{C} E\left[T_i(n)\right] - \sum_{k=1}^{U} E\left[P_k(n)\right] \right) \tag{5.8}$$

We can break $\sum_{i=1}^{C} E[T_i(n)]$ into two terms:

$$\sum_{i=1}^{C} E[T_i(n)] = \sum_{k=1}^{U} E[T_k(n)] + \sum_{i=U+1}^{C} E[T_i(n)] \tag{5.9}$$

Based on eq (5.8) and (5.9), we obtain the following equation:

$$R_{APL}(n, U, \text{MAUCB}) \leq \mu_1 \left[\sum_{i=U+1}^{C} E[T_i(n)] + \sum_{k=1}^{U} E[O_k(n)] \right] \tag{5.10}$$

As mentioned before, the regret in the multi-user case mainly depends on the access to bad channels and the collisions produced among users. Similarly, the upper bound of $R_{APL}(n, U, \text{MAUCB})$ in eq (5.10) deals with the global

definition of the regret in which $T_i(n)$ represents the access of bad channels, and $O_k(n)$ stands for the number of collisions in the k-th best channel. In order to bound the regret, we need to bound the two terms $E[T_i(n)]$ and $E[O_k(n)]$.
In eq (4.32) in the previous chapter, we proved that the upper bound of $E[T_i(n)]$ in AUCB for a single user is expressed as follows:

$$E[T_i(n)] \leq \frac{8\ln(n)}{\Delta_i^2} + 1 + \frac{\pi^2}{3} \tag{5.11}$$

Then, for each user in the multi-user case, we obtain:

$$E[T_{i,1}(n)] \leq \frac{8\ln(n)}{\Delta_{(1,i)}^2} + 1 + \frac{\pi^2}{3}$$

$$\vdots$$

$$E[T_{i,U}(n)] \leq \frac{8\ln(n)}{\Delta_{(U,i)}^2} + 1 + \frac{\pi^2}{3}$$

Therefore, the upper bound of $E[T_i(n)]$ for all users becomes:

$$E[T_i(n)] = \sum_{j=1}^{U} E[T_{i,j}] \leq \sum_{k=1}^{U} \left[\frac{8\ln(n)}{\Delta_{(k,i)}^2} + 1 + \frac{\pi^2}{3} \right] \tag{5.12}$$

In order to obtain an upper bound of the total regret, it remains to find an upper bound of $E[O_U(n)] = \sum_{k=1}^{U} E[O_k(n)]$ which stands for the expectation of the total number of collisions that may occur in the U best channels. Let $E[O_C(n)] = \sum_{i=1}^{C} O_i(n)$ be the expectation of the total number of collisions that may occur in all channels, and $E[D_k(n)]$ be the expectation of the total number of collisions encountered by the k^{th} priority user in all channels. To explain our idea and to identify the difference between these latter parameters, Table (5.1) presents a case study with corresponding $D_k(n)$ and $O_k(n)$. $E[O_U(n)]$ can be defined as follows:

$$E[O_U(n)] = \sum_{k=1}^{U} E[O_k(n)] \leq \sum_{i=1}^{C} E[O_i(n)] = \sum_{k=1}^{U} E[D_k(n)] \tag{5.13}$$

We should mention that, when the users get a good estimation about the vacant probabilities of channels and each of them accesses his dedicated channel, then non-collision state can occur. On the other hand, the k^{th} user may collide with other users in two cases:

	SU1	SU2	$D_{SU1}(t)$	$D_{SU2}(t)$	$O_{C1}(t)$	$O_{C2}(t)$	$O_{C3}(t)$
t=1	C1	C1	1	1	2	0	0
t=2	C2	C3	0	0	0	0	0
t=3	C2	C2	1	1	0	2	0
t=4	C3	C2	0	0	0	0	0
t=5	C3	C3	1	1	0	0	2
t=6	C1	C1	1	1	2	0	0

TABLE 5.1: Two SUs access three available channels. In this case, the total number of collisions for the two users is $D(n) = \sum_{k=1}^{U} D_k(n) = D_{SU1}(n) + D_{SU2}(n)$. The number of collisions in all channels produced by the users is $O_C(n) = \sum_{i=1}^{C} O_i(n) = O_{C1}(n) + O_{C2}(n) + O_{C3}(n)$, while the number of collisions in the best channels, i.e. $C1$ and $C2$, is $O_U(n) = \sum_{k=1}^{U} O_k(n) = O_{C1}(n) + O_{C2}(n)$.

- If he does not identify well his dedicated channel.
- If he does not respect his prior rank[1].

Let $T_k^{'}(n)$ and S_s be respectively the total number of times, where the k^{th} user badly identifies his dedicated channel and the time he needs to return to his prior rank. After each bad estimation, the user will change his dedicated channel. In this case, he may collide with other users until convergence to his prior rank. Subsequently, for all values of n, the total number of collisions for the k^{th} user $D_k(n)$ can be upper bounded by:

$$D_k(n) \leq T_k^{'}(n) S_s \tag{5.14}$$

As $T_k^{'}(n)$ and S_s are independent, we have:

$$E[D_k(n)] \leq E[T_k^{'}(n)] E[S_s] \tag{5.15}$$

Let us find an upper bound of $E[T_k^{'}(n)]$, and let $\mathbb{A}_k(t)$ be the event that the k^{th} user identifies his dedicated channel, the k^{th} best one, at instant t. Then, $\forall\, k+1 \leq i \leq C$ and $1 \leq m \leq k-1$, the event $\mathbb{A}_k(t)$ takes place when the following condition is satisfied:

$$\mathbb{A}_k(t) : B_i(t) \leq B_k(t) \leq B_m(t)$$

For a bad estimation event at instant t, $\exists\, i \in \{k+1, ..., C\}$ and $\exists\, m \in \{1, ..., k-1\}$, $\bar{\mathbb{A}}_k(t)$ is true when we have the following condition:

[1] After each collision and according to our policy APL, the user should regenerate a rank.

$$\bar{\mathbb{A}}_k(t) : \left[B_i(t) > B_k(t)\right] \; or \; \left[B_k(t) > B_m(t)\right]$$

Then, the total number of times of a bad estimation where the k^{th} priority user does not access his channel up to n, $E[T_k^{'}(n)]$, can be upper bounded as follows:

$$E[T_k^{'}(n)] \leq E[T_{B_i>B_k}(n)] + E[T_{B_m<B_k}(n)] \tag{5.16}$$

where $T_{B_i>B_k}(n)$ stands for the total number of times in which the index of the i^{th} channel exceeds that of the k^{th} best one for all $i \in \{k+1, ..., C\}$ up to n; $T_{B_m<B_k}(n)$ stands for the total number of times in which the index of the k^{th} best channel exceeds the m^{th} best one for all $m \in \{1, ..., k-1\}$. We should mention that, for the first priority user, $E[T_{B_m<B_k}(n)]$ should equal 0, since his dedicated channel has the highest vacant probability. Based on eq (4.32) in chapter 4, $T_{B_i>B_k}(n)$ for the k^{th} user can be upper bound by:

$$E[T_{B_i>B_k}(n)] \leq \frac{8\ln(n)}{\Delta^2_{(k,i)}} + 1 + \frac{\pi^2}{3} \tag{5.17}$$

where $\Delta_{(k,i)} = \mu_k - \mu_i$. Since $\mu_i \leq \mu_{k+1}$ for all $i \in \{k+1, ..., C\}$ and $\mu_k \geq \mu_{k+1} \geq ... \geq \mu_C$, then $\Delta_{(k,i)} \geq \Delta_{(k,k+1)}$. Subsequently, the upper bound of $E[T_{B_i>B_k}(n)]$ can be expressed by:

$$E[T_{B_i>B_k}(n)] \leq \frac{8\ln(n)}{\Delta^2_{(k,k+1)}} + 1 + \frac{\pi^2}{3} \tag{5.18}$$

Similarly, the second term $E[T_{B_m<B_k}(n)]$ in eq (5.16) should satisfy:

$$E[T_{B_m<B_k}(n)] \leq \frac{8\ln(n)}{\Delta^2_{(m,k)}} + 1 + \frac{\pi^2}{3} \tag{5.19}$$

where $\Delta_{(m,k)} \geq \Delta_{(k-1,k)}$ for all $m \in \{1, ..., k-1\}$. Then, we obtain:

$$E[T_{B_m<B_k}(n)] \leq \frac{8\ln(n)}{\Delta^2_{(k-1,k)}} + 1 + \frac{\pi^2}{3} \tag{5.20}$$

Based on eq (5.16), (5.18) and (5.20), $E[T^{'}(n)]$ can be expressed as follows:

$$E[T^{'}(n)] \leq \sum_{k=1}^{U} \left(\frac{8\ln(n)}{\Delta^2_{(k,k+1)}} + 1 + \frac{\pi^2}{3} \right) + \sum_{k=2}^{U} \left(\frac{8\ln(n)}{\Delta^2_{(k-1,k)}} + 1 + \frac{\pi^2}{3} \right) \tag{5.21}$$

The expectation of time S_s for U SUs can be upper bounded by:

$$E[S_s] \leq \binom{U}{2U-1} - 1 \tag{5.22}$$

Proof. appendix D □

Based on eq (5.13), (5.15), (5.21) and (5.22), the total number of collisions in the best channels for U SUs can be upper bounded by:

$$E[O_U(n)] \leq \left[\binom{U}{2U-1} - 1 \right] \cdot \left[\sum_{k=1}^{U} \left(\frac{8\ln(n)}{\Delta^2_{(k,k+1)}} + 1 + \frac{\pi^2}{3} \right) + \sum_{k=2}^{U} \left(\frac{8\ln(n)}{\Delta^2_{(k-1,k)}} + 1 + \frac{\pi^2}{3} \right) \right] \tag{5.23}$$

Finally, the global regret of U users for MAUCB using APL can be expressed as follows:

$$R_{APL}(n, U, \text{MAUCB}) \leq \mu_1 \left[\sum_{k=1}^{U} \sum_{i=U+1}^{C} \left(\frac{8\ln(n)}{\Delta^2_{(k,i)}} + 1 + \frac{\pi^2}{3} \right) + \frac{1-p}{p} \left[\sum_{k=2}^{U} \left(\frac{8\ln(n)}{\Delta^2_{(k-1,k)}} + 1 + \frac{\pi^2}{3} \right) + \sum_{k=1}^{U} \left(\frac{8\ln(n)}{\Delta^2_{(k,k+1)}} + 1 + \frac{\pi^2}{3} \right) \right] \right] \tag{5.24}$$

The above upper bound of regret contains three main components: The first one depends on the loss of reward due to the selection of bad channels. The second and third components stand for the loss of reward resulting from collisions among users in the U best channel. Moreover, the three components have a logarithmic asymptotic behavior, which means that, after a finite number of slots each user can converge to his dedicated channel.

5.4 Study the Quality of Service Using AUCB

In wireless communication field and more precisely in the context of OSA, taking into consideration two criteria, i.e. vacancy and quality of channels,

is necessary when a certain level of QoS is required. Recent works, such as the one proposed in [42], have studied the QoS for multiple users using the Random Rank policy [35]. Based on the QoS-UCB, the user can learn the vacancy probabilities of channels and recognize their qualities. For its optimality with respect to other existing works, we adopt AUCB to study the QoS of secondary network. Moreover, we use in our simulations APL policy for the priority access in order to help the users to share the available spectrum under a competitive access.
Let each channel have the instantaneous quality $q_i(t)$ at the slot t, then the expectation of the quality collected from the channel i up to n is given as follows:

$$G_i(T_i(n)) = \frac{1}{T_i(n)} \sum_{\tau=1}^{T_i(n)} q_i(\tau) \tag{5.25}$$

The global mean reward, that takes into account the quality as well as the availability of all channels, can be defined as follows [42]:

$$\mu_i^Q = G_i(T_i(n)).\mu_i \tag{5.26}$$

The index assigned to the i^{th} channel, $B_i^Q(t, T_i(t))$, that takes into account the availability, $X_i(T_i(t))$, and quality, $Q_i(t, T_i(t))$ of the i^{th} channel can be defined by:

$$B_i^Q(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t)) - Q_i(t, T_i(t)) \tag{5.27}$$

where $X_i(T_i(t)), Q_i(t, T_i(t))$ and $A_i(t, T_i(t)))$ are given by:

$$X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{\tau=1}^{t} S_i(\tau) \tag{5.28}$$

$$A_i(t, T_i(t))) = \arctan\left(\frac{\alpha \ln(t)}{T_i(t)}\right) \tag{5.29}$$

$$Q_i(t, T_i(t)) = \frac{\gamma M_i(t, T_i(t)) \ln(t)}{T_i(t)} \tag{5.30}$$

where S_i is the state of the i-th channel; α being the exploration-exploitation factor; γ represents the impact of the quality factor; $M_i(t, T_i(t)) = G_{\max}(t) - G_i(T_i(t))$ is the difference between the maximum expected quality over channels at time t, i.e. $G_{\max}(t)$, and the quality collected from channel i up to time slot t, i.e. $G_i(T_i(t))$. However, when the i^{th} channel has a good quality $G_i(T_i(t))$ as well as a good availability $X_i(T_i(t))$ at a time t then the quality

factor $Q_i(t, T_i(t))$ decreases while $X_i(T_i(t))$ increases. Subsequently, by selecting the maximum of his index $B_i^Q(t, T_i(t))$, the user has a large choice to access the i^{th} channel with a high quality and availability.

5.5 Priority and Fairness Access

Existing works, as well as our proposed policy APL for the dynamic access consider the case of U users with U priority level. In this section, we propose another policy called Priority and Fairness Access (PFA) to tackle different priority levels (not necessarily U levels) for U users, and each level may contain at least one user as shown in Fig. 5.2. Users in the same level should have the same chance to access similar channels related to their level.

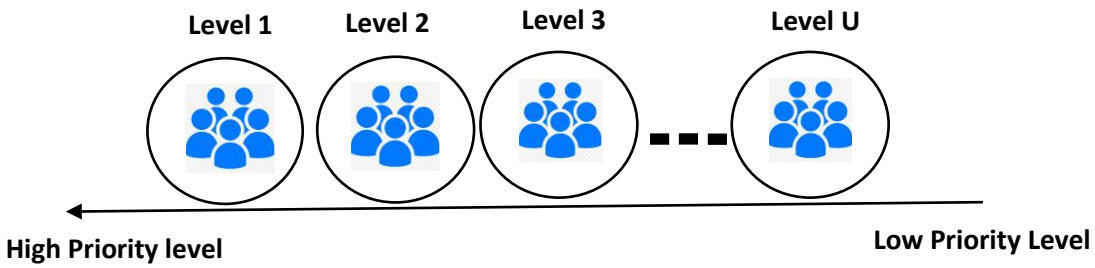

FIGURE 5.2: Different priority levels in the secondary network

5.5.1 PFA for Two Sets of Priority Levels

To simplify our discussion, we consider at first two sets of secondary users: priority (SU_p) and ordinary (SU_o) users. Our policy PFA ensures the fairness among users belonging to the same set. For instance, SU_p have an equal right to access the best channels while SU_o will try to fairly share the almost best ones. All users try to learn selfishly the vacancy probabilities of channels.
Hereinafter, we propose a cooperative transmission technique in order to extend the transmission range of SU_o and enhance the throughput by requesting the SU_p to act as relays.
PFA takes into consideration the competitive learning (to decrease the complexity of the system) with a cooperative transmission, while, our previous policy Side Channel proposed in Chapter 3 is considered as a cooperative learning that requires a certain cooperation level among users with a competitive transmission.
Based on PFA, let P and $O = (U - P)$ be respectively the number of priority and ordinary users in the secondary network and suppose that the vacant

probabilities of channels are ordered as: $\mu_1 > \mu_2 > ... > \mu_C$ where μ_1 represents the channel with the highest vacant probability and $\Gamma = \{\mu_i\}$. First, the vector Γ is unknown and the users should estimate it after a finite time observation. Let $\Gamma_P = \{\mu_1, \mu_2, ..., \mu_P\}$ represents the P best channels reserved for the priority users and $\Gamma_O = \{\mu_{P+1}, \mu_{P+2}, ..., \mu_U\}$ depicts the almost best channels for the ordinary users, then the priority users should examinate only the states of the P best channels at each time slot in order to gather some information about their vacant probabilities. Although, for specific constraints related to hardware, a priority user is not able to sense wide frequency band at each time slot. To solve the latter issue and make each priority user scan the P channels, let each user generate a rank in the set $\{1, ..., P\}$ in order to scan the P best channels and obtain information about their vacancy, while, the rank generated by the ordinary users should be in the set $\{1, ..., U\}$. After the learning period, if the rank generated by the priority users belongs to the set $\{1, ..., P\}$, then he accesses one of the best channels and may collide with a priority user (see algorithm 8).

Algorithm 8: Priority and Fairness access (PFA)

Input: $\xi_p(t), \xi_o(t)$
$\xi_p(t)$: indicates a presence of collision under the *pth* priority user at time t,
$\xi_o(t)$: indicates a presence of collision under the *oth* ordinary user at time t,
Initialization
for t = *1 to* C **do**
  Each user senses each channel once,
  Each SU updates his index $B_i(t)$,
SU_p generates a rank of the set $\{1, ..., P\}$,
SU_o generates a rank of the set $\{1, ..., U\}$,
for t = *C+1 to* n **do**
  Each SU senses a channel in his index $B_i(t)$ according to his rank,
  if $\xi_p(t)$=*1* **then**
    SU_p regenerates his rank of the set $\{1, ..., P\}$,
  else
  $\overline{SU}_p$ retains his previous rank,
  if $\xi_o(t)$=*1* **then**
    SU_o regenerates his rank of the set $\{1, ..., U\}$,
  else
  $\overline{SU}_o$ retains his previous rank,
  Each SU updates his index $B_i(t)$

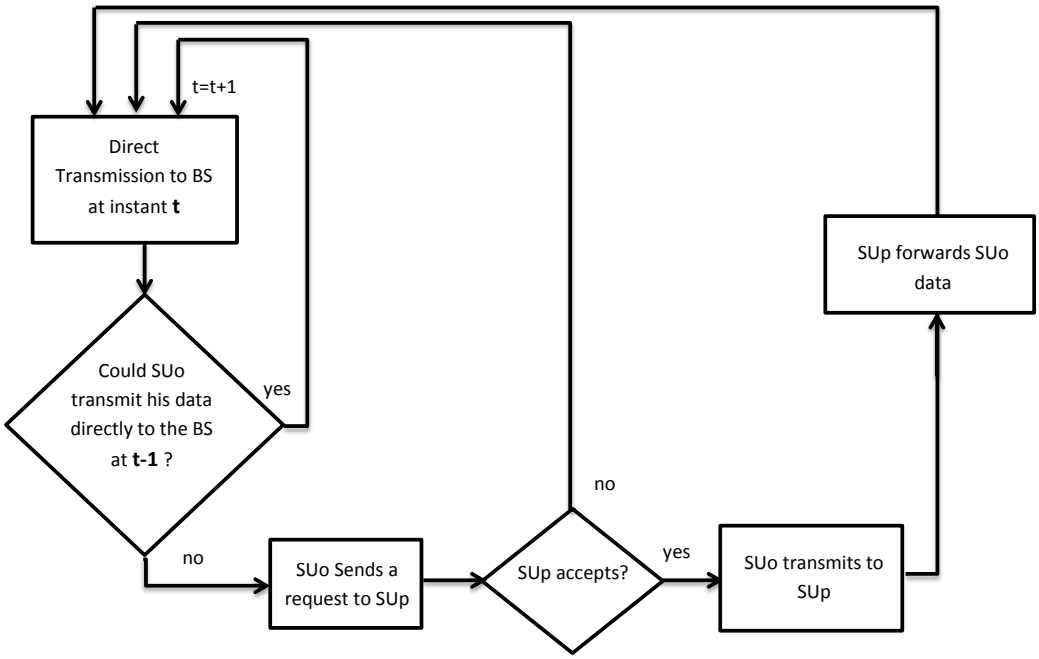

FIGURE 5.3: Novel transmission technique to enhance the transmission rate of the secondary network

5.5.2 Transmission Technique Based on PFA

In [19, 101, 102], the authors proposed a cooperative spectrum sharing among primary and secondary users in order to enhance the transmission rate of the former, and thus reduce his spectrum usage.

In this section, we propose a novel transmission technique based on our policy PFA in order to enhance the data rate and extend the transmission range of the secondary network.

When a secondary user is located in a fading location and having access only to poor channels, then he is not able to perform a direct transmission to the primary base station. Thus, this user has the ability to communicate with priority users SU_p in order to ensure the transmission of his data. For a simplicity sake, let's consider one priority (SU_p) and two ordinary secondary users (SU_{o1} and SU_{o2}) in the secondary network as shown in Fig. 5.4. In that figure, SU_{o1} and SU_{o2} might be interested in communicating with SU_p to increase their transmission rate and range[2].

If, SU_{o1} and SU_{o2} send a request to SU_p at the same time, this latter is

[2]The nearest users to the primary base station are considered as priority ones, and a given user can change his set by changing the interval of his generated rank

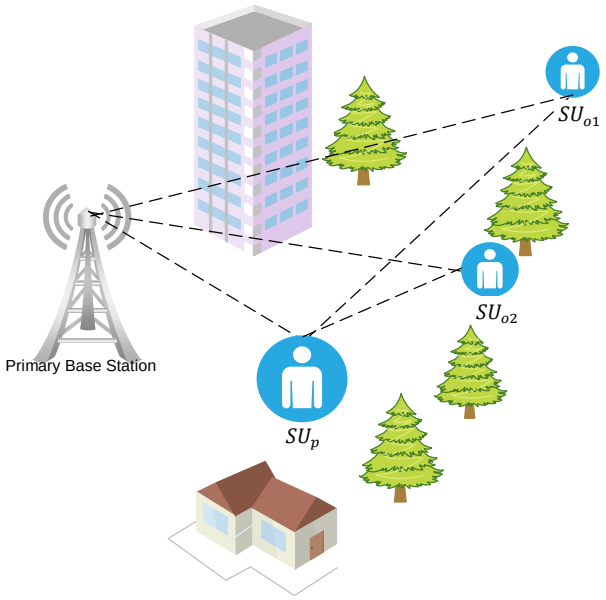

FIGURE 5.4: Cooperative transmission with one priority user

authorized to accept only one request based on various parameters[3], such as: the transmission power, the signal to noise ratio (SNR), the distance to primary BS, etc. The role of the priority user is to maximize the transmission rate of the network while the aim of the ordinary users is to selfishly maximize their transmission rate. SU_o with a refused request selects a channel based on his observation and transmits directly his own data to the primary BS. If the request of an ordinary user is accepted by his corresponding SU_p, this latter selects the best available channel based on his observation to send the ordinary user's data with his own data. SU_o should also make a direct transmission to BS to have an information about his SNR (Signal-to-Noise Ratio) in order to make a decision in the next slot[4]. Indeed, if his direct transmission is accomplished successfully at previous slot, then he continues transmitting to BS in the next slot; Otherwise, he sends a request to the priority users. Therefore, the achievable data rate at time t can be expressed as follows:

$$R_o(t) = \frac{W}{2} log_2\big(1+\rho_{SU_p,BS}(t)\big) + \frac{W}{2} log_2\big(1+\rho_{SU_o,BS}(t)\big) \tag{5.31}$$

$$= \frac{W}{2} log_2\Big(\big(1+\rho_{SU_p,BS}(t)\big)\big(1+\rho_{SU_o,BS}(t)\big)\Big) \tag{5.32}$$

[3]In our case, we consider the approval of a request is based on the SNR value, then the priority user can accept the request of an ordinary user who has the lowest SNR because it is more difficult for this user to make a direct transmission to the primary BS in the next slot.

[4]Assuming that each user can have a notification on his SNR after each transmission time.

where W is the bandwidth of the selected channel at time t. $\rho_{SU_p,BS}(t)$ and $\rho_{SU_o,BS}(t)$ stand for the SNR of the link between SU_p-BS and SU_o-BS respectively. If $\rho_{SU_p,BS}(t)$ is much higher than $\rho_{SU_o,BS}(t)$, the achievable rates can be simplified as follows:

$$R_o(t) \approx \frac{W}{2} log_2(1 + \rho_{SU_p,BS}(t)) \tag{5.33}$$

The transmission time of SU_p will be divided into several segments: Tp_1, Tp_2, Tp_3 and Tp_4; where, during Tp_1, the user senses a channel and exchanges information with the ordinary users; Tp_2 is the time during which data is received from the ordinary user; Tp_3 stands for the time spent to relay the data received by SU_p to the PU's channels and depends on hardware constraints; finally, Tp_4 is the time required to send the data of the ordinary user to the primary BS.

5.6 Simulation Results

In this section, we evaluate separately the performance of the two proposed policies APL and PFA. As mentioned before, the existing policies, as well as APL, tackle U priority levels for U users where each level contains only one user while the case of two or more priority levels in which each level contains at least one user, to the best of our knowledge, is not yet considered. One of the benefits of PFA is that it can enhance the transmission rate of users and extend the transmission range of the network as discussed in the previous section.

In our simulations, we evaluate separately the performance of APL and PFA since it is clear that APL achieves better results compared to PFA. Indeed, in the case of APL, when the k-th user regenerates his rank in the set $\{1, ..., k\}$, he may collide with all the top priority users, i.e. $SU_1, SU_2, ..., SU_{k-1}$, before settling down to his dedicated channel, i.e. the k-th best channel. However, in the case of PFA, when the k-th user in the i-th set regenerates his rank, he may collide with all users in the different sets before settling down to his specific set. First, we compare the performance of APL to the recent works such as SLK and Musical Chair. Hereinafter, we introduce PFA and we evaluate its performance that depends on its ability to make the priority users uniformly access the best channels and to enable the ordinary users to access uniformly the almost best channels.

5.6.1 Evaluating the Performance of APL

In this section, we evaluate the performance of the APL and its ability to make each user selects his dedicated channel after a finite number of time slots. We evaluate the performance of APL compared to the existing learning policies such as Musical Chair and SLK. To make this comparison, we use two main performance indexes: the regret related to the access of worst channels and the percentage of times to access best channels by each user. A collision may occur when two or more users try to access the same channel. We adopt in our simulations the ALOHA model, widely used one in OSA, in which none of the collided users receives a reward. After each collision, and based on our policy APL, the collided users should regenerate their rank. First, we consider a static setting of users, then we investigate the dynamic access in which the priority users can enter or leave the network.
Let us consider 4 priority users trying to access 9 communication channels with the following vacant probabilities:

$$\Gamma = [\,0.9\ 0.8\ 0.7\ 0.6\ 0.5\ 0.45\ 0.3\ 0.25\ 0.1]$$

Each user can access one of the available channels according to his prior rank. For instance, the first priority user, SU_1, should regularly access the best channel, $\mu_1 = 0.9$, while the targets of the users SU_2, SU_3 and SU_4 are to access respectively the channels 0.8, 0.7, 0.6. Fig. 5.5 shows the regret of our proposed APL with the well-known MAB algorithms: AUCB, TS, UCB1 and e-greedy. As we can see, the regrets of the four MAB algorithms with APL have a logarithmic asymptotic behavior that means the 4 users are able to estimate the vector Γ, and then access their dedicated channels. According to several recent works, TS seems to outperform the current state-of-the-art MAB algorithms by achieving a lower regret for a single user but not necessarily for multiple users. Indeed, despite its good performance compared to UCB1 and e-greedy shown in Fig. 5.5, TS cannot give better results compared to AUCB. Similarly, in the previous chapter, AUCB achieved better results under Random Rank compared to TS. The same figure shows that the regret under e-greedy, varies as an increasing function over time at the beginning with a high derivative, provides worst results. Indeed, the random access by each user, based on e-greedy, during the learning period may lead to a large number of collisions among users. Then, the regret that depends on the access of worse channels and the number of collisions among users may increase significantly.

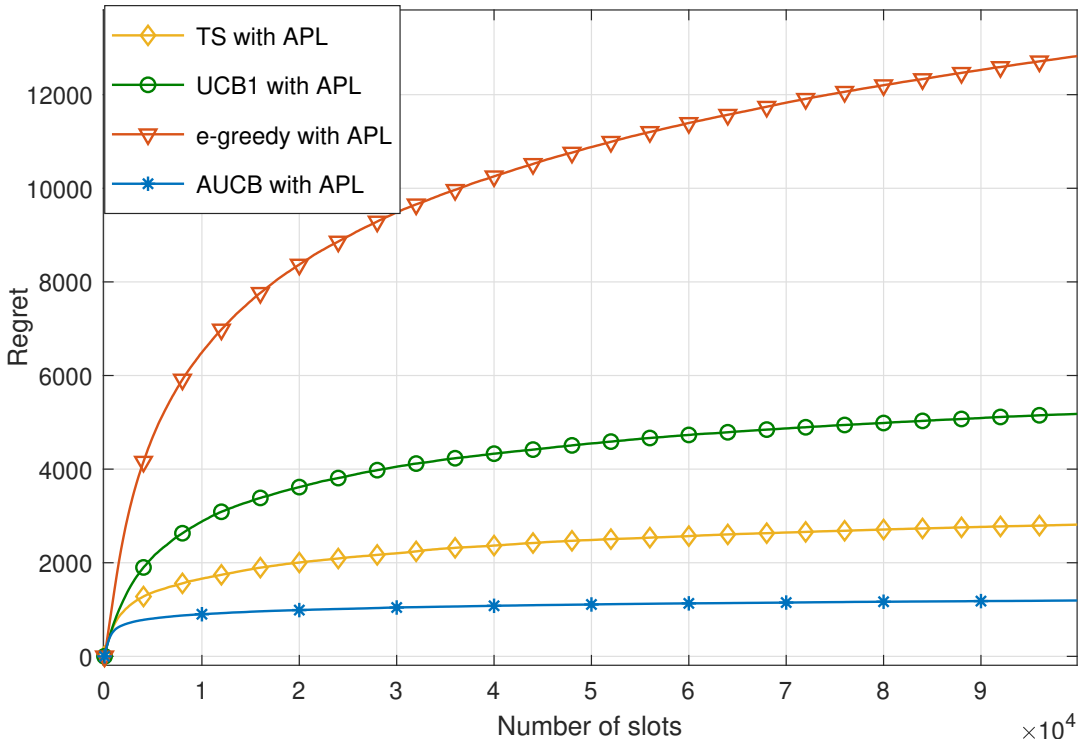

FIGURE 5.5: Regret of APL compared to SLK and Musical Chair policies

In Fig. 5.6, we compare the regret of APL to SLK and Musical Chair. APL and SLK take into consideration the priority access while Musical Chair is proposed for the random access. Despite the regret of APL and SLK has a logarithmic asymptotic behavior, the regret of Musical Chair has two parts:

- A linear part at the beginning, during the learning period, due to the large number of collisions resulting from the random selection.
- A constant part in which the users exploit the U best channels .

As we can see from Fig. 5.6, APL using AUCB and TS outperforms Musical Chair and SLK by achieving the lower regret.

Fig. 5.7 shows the percentage of times that the k-th user accesses his dedicated channel based on our policy APL up to n, $P_k(n)$. This latter is given by:

$$P_k(n) = 100 \times \sum_{t=1}^{n} \frac{\mathbb{1}_{(a_{t,k}=\mu_k)}}{t}$$

where $a_{t,k}$ represents the channel selected by the k-th user at instant t;

and $\mathbf{1}_{(\mathbf{a}=\mathbf{b})} = \begin{cases} 1 \text{ if a =b} \\ 0 \text{ otherwise} \end{cases}$

In Fig. 5.7, P_{Best} shows three main parts:

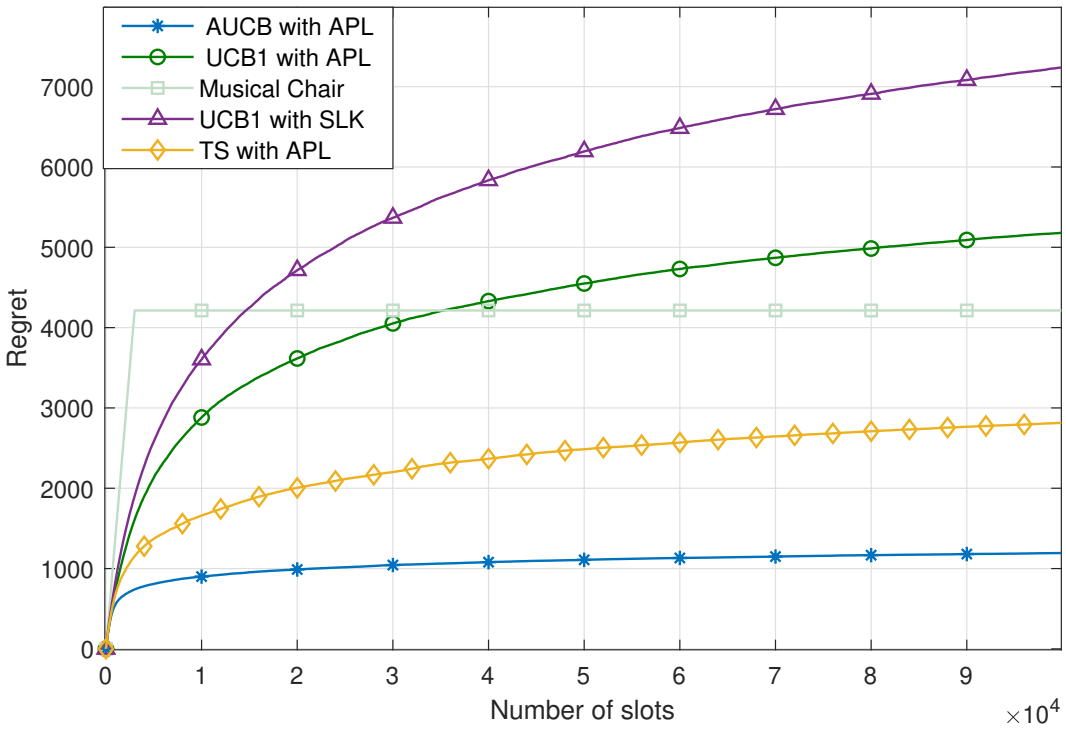

FIGURE 5.6: Regret of APL compared to SLK and Musical Chair policies

- The first part from 1 to the number of channels C represents the initialization phase, where each user selects each channel once in order to collect some information about the vacancy of channels.
- The second part from slot C+1 to 2000 represents the adaptation phase.
- In the last part, the users converge asymptotically to their dedicated channels.

As we can see, based on our policy APL, the users are able to converge to their targeted channels: The first priority user SU_1 converges to the best channel μ_1, followed by SU_2, SU_3 and SU_4 to the channels μ_2, μ_3 and μ_4 respectively. In addition, we can observe a fast converges of APL using AUCB compared to TS.

In [42], the authors proposed a novel version UCB1 that considers not only the availability but also the quality of channels. Based on QoS-UCB1, the user is able to estimate the vacant probability as well as the quality of channels. In this section, we study the QoS-AUCB that gives better results

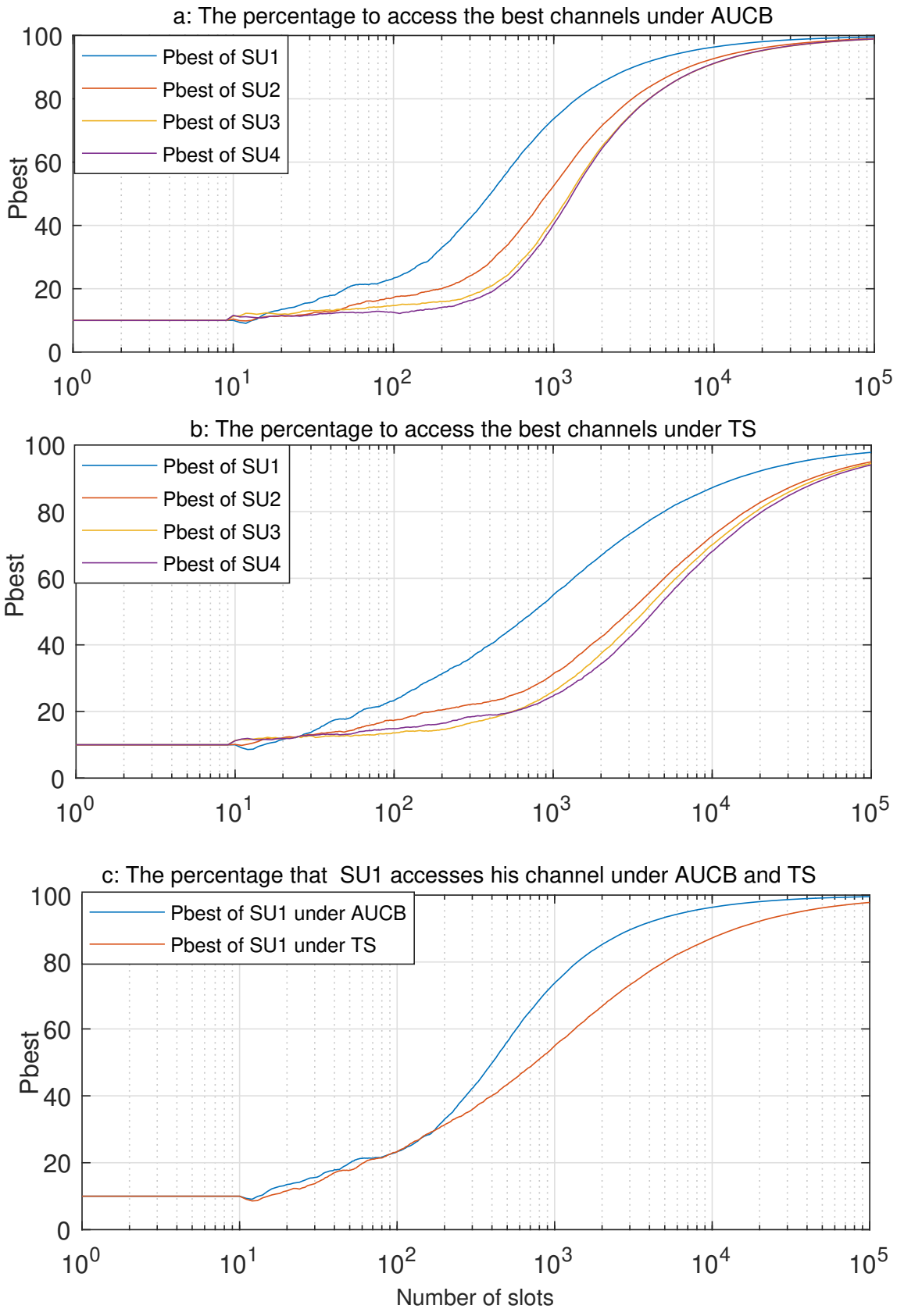

FIGURE 5.7: Pbest of APL using TS and AUCB

compared to QoS -UCB. Beside the vector Γ, let us define a vector G to represent the empirical mean of the quality observed from the channels:

$$G = [\,0.75\ 0.99\ 0.2\ 0.8\ 0.9\ 0.7\ 0.75\ 0.85\ 0.8]$$

Using the two vectors Γ and G, we obtain the global mean, Γ_Q, that considers the vacancy as well as the quality of the channels:

$$\Gamma_Q = [0.67\ 0.79\ 0.14\ 0.48\ 0.37\ 0.28\ 0.22\ 0.17\ 0.08]$$

After estimating Γ_Q, and based on our policy APL, the first priority user SU_1 selects the channel with the highest global mean, i.e. the second entry in vector Γ_Q, while the target of SU_2, SU_3 and SU_4 are the channels 1, 4 and 5 respectively. This result can be observed in Fig. 5.8, where each user selects his assigned channel based on QoS-UCB or QoS-AUCB. Moreover, QoS-AUCB enables the users to select their dedicated channels more frequently than in the case of QoS-UCB.

Fig. 5.9 depicts the regret of QoS-AUCB and QoS-UCB for the 4 priority

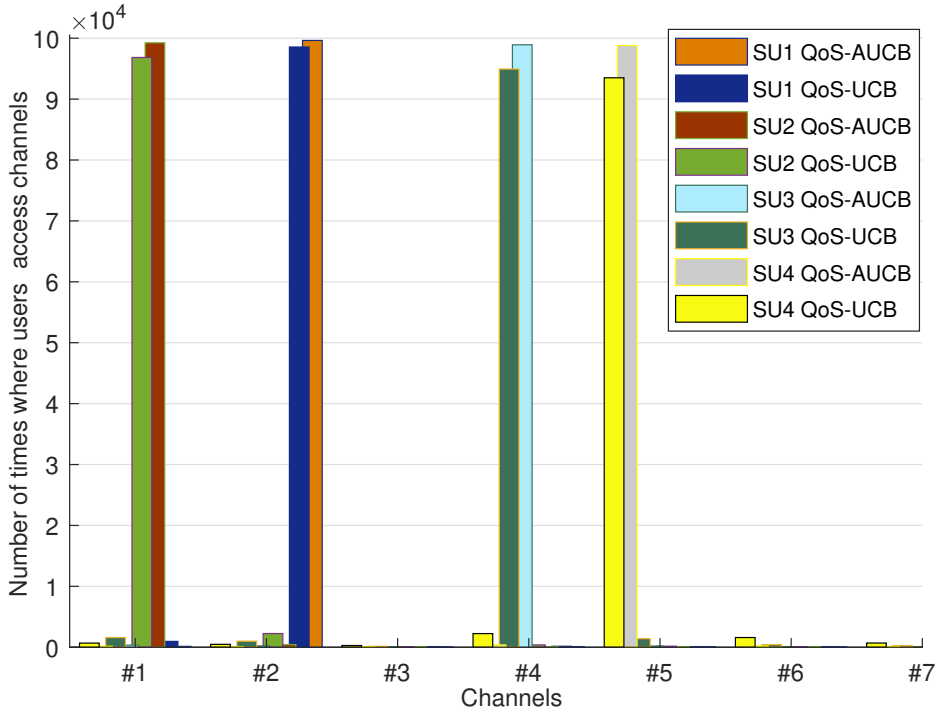

FIGURE 5.8: Access channels by the priority users using QoS-AUCB and QoS-UCB

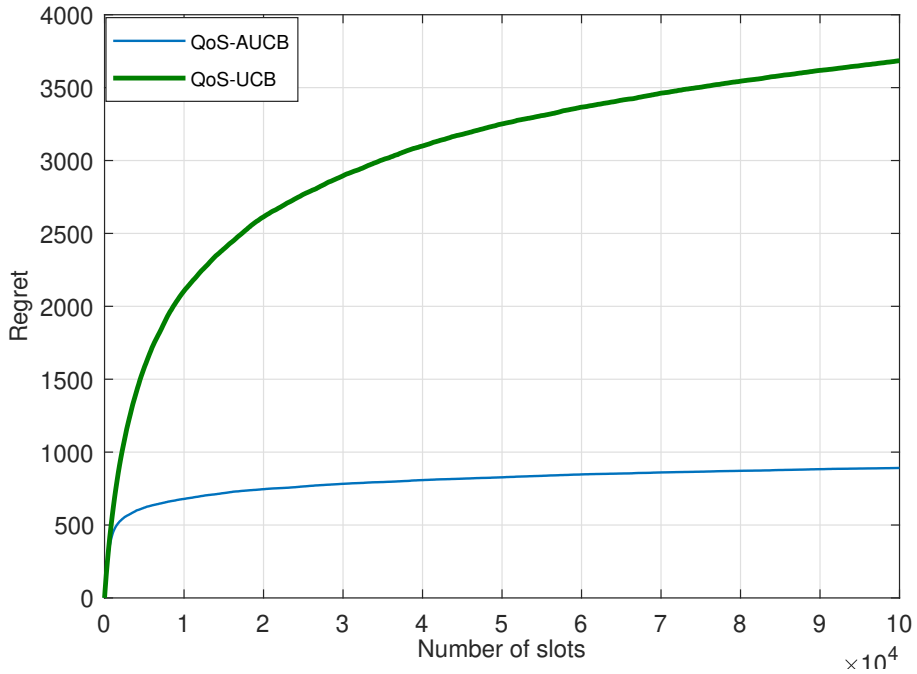

FIGURE 5.9: Regret of QoS-AUCB and QoS-UCB

users. In [42], the authors show that QoS-UCB achieves good results compared to several existing algorithms, such as Regenerative Cycle Algorithm (RCA) [45], restless UCB (RUCB) [103] and Q-learning [104]. From Fig. 5.9, one can observe that QoS-AUCB achieves better results compared to QoS-UCB.

5.6.2 Evaluate the Performance of PFA

According to PFA, two or more sets of users can be considered such that users from the same set have the same choice to select and access the channels. Assuming that two sets (or levels) of users are considered: Priority and ordinary, where the priority set contains one user ($P = 1$), while the ordinary contains two users ($O = 2$). The three users try to learn selfishly the vacant probabilities of 9 channels:

$$\Gamma = [0.9\ 0.8\ 0.7\ 0.6\ 0.5\ 0.4\ 0.3\ 0.2\ 0.1]$$

If the users have a prior knowledge about the vacant probabilities of channels Γ, then the target of the priority user SU_p remains to access the most vacant channel, i.e. $\mu_1 = 0.9$, while the two ordinary users SU_{o1} and SU_{o2} access uniformly the two almost-best channels, i.e. $\mu_2 = 0.8$ and $\mu_3 = 0.7$.

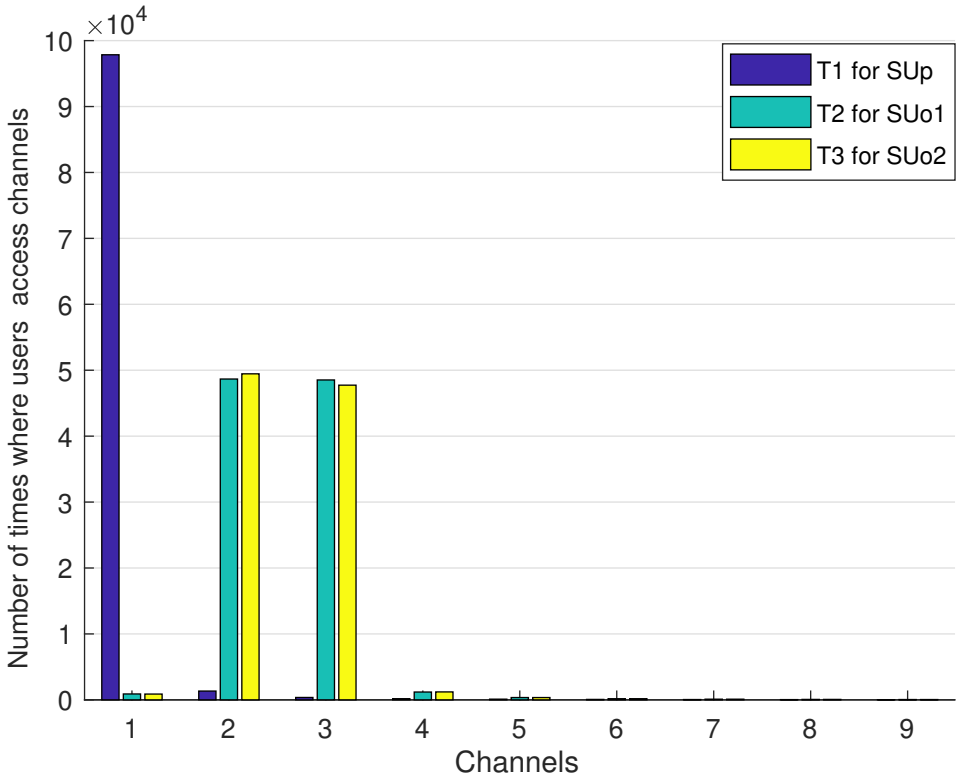

FIGURE 5.10: Number of times that the users access the channels

The performance of PFA depends on its ability to ensure the priority and the fairness: Could PFA provide the priority between two different sets? Or could PFA ensure the fairness among users in the same set? In Fig. 5.10, we display $T_k(n)$ that represents the accessing of all channels by each user up to the total number of slots n. We can clearly notice that the priority user often accesses the best channel while the ordinary users uniformly access the two almost-best channels.

Fig. 5.11 compares the regret of PFA using TS, UCB1 and e-greedy and SLK for the three users. SLK considers U priority levels for U users where the k-th user tries to select and access the k-th best channel. Although the two policies (i.e. SLK and PFA) have a logarithmic asymptotic behavior, PFA achieves a lower regret. Moreover, obtaining a lower regret can provide more opportunity for users in the licensed network which increases the transmission rate.

As mentioned before, PFA can be used to extend the transmission range of the network, more precisely for the ordinary users, based on the transmission technique proposed in Section (5.5.2). Let us define the throughput of the o-th ordinary users $H_o(t)$ that also represents the percentage of successful transmission. Generally, the throughput depends on several factors: the

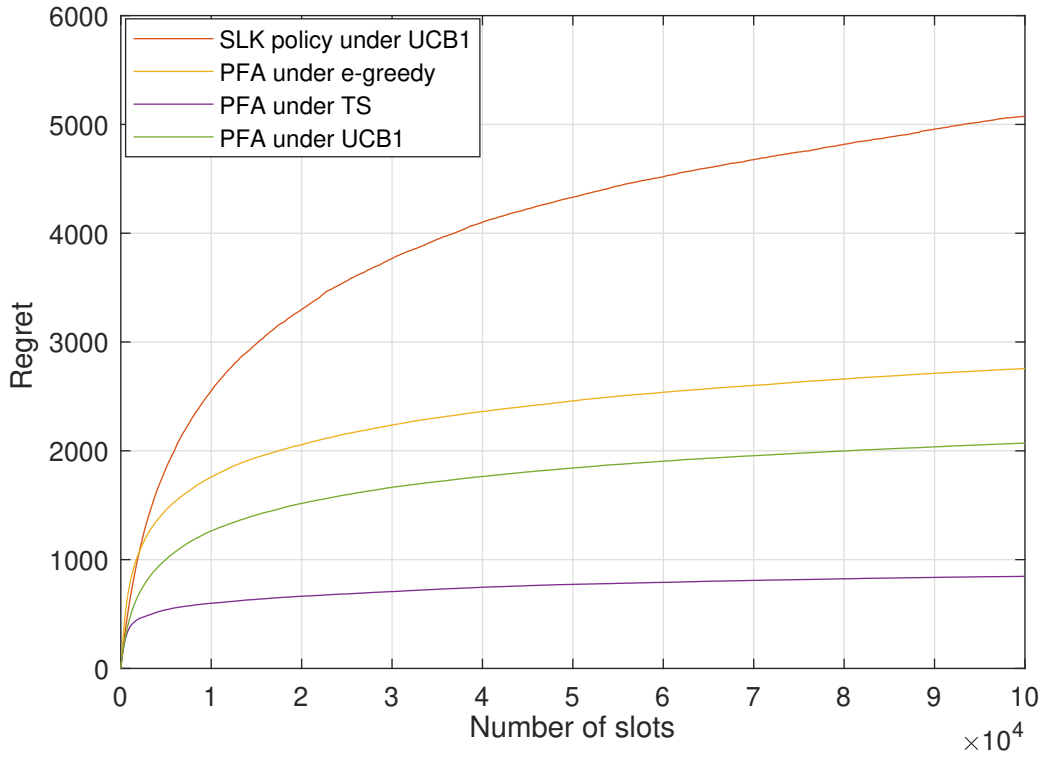

FIGURE 5.11: The regret of our policy PFA under TS, UCB1 and *e*-greedy compared to SLK

vacant probabilities of channels, collisions among users and the SNR values. The vacant of channels and collisions among users have been discussed and enhanced using PFA. While, to enhance the SNR, we use the proposed transmission technique. Let SNR_ζ stand for the SNR threshold and $SNR_o(t)$ be the SNR of the o-th ordinary user SU_o at time t. If $SNR_o(t) < SNR_\zeta$, then this user may lose his transmitted data at time t; then, in the next time slot, SU_o should send a request to the priority users. A priority user accepts the request of the SU_o with the lowest SNR. Let $H_o^L(t)$ indicate whether the transmission was successful at instant t for the o-th user. $H_o^L(t)$ can be expressed as follow:

$$H_o^L(t) = r_{i,o}(t).C_{i,o}(t).\mathbf{1}_{(SNR_o(t)>SNR_\zeta)} \; || \; r_{i,p}(t).C_{i,p}(t).\mathbf{1}_{(SNR_o(t)<SNR_{o'(t)})} \quad (5.34)$$

where $r_{i,o}(t)$ and $r_{i,p}(t)$ stand for the state of the i-th channel selected by ordinary or priority users respectively, and they equal 1 if the channel is free and 0 otherwise. $C_{i,o}(t)$ and $C_{i,p}(t)$ indicate whether a collision occurs in the i-th channel, and they equal 1 if the user is the sole in occupied in the $i-$th

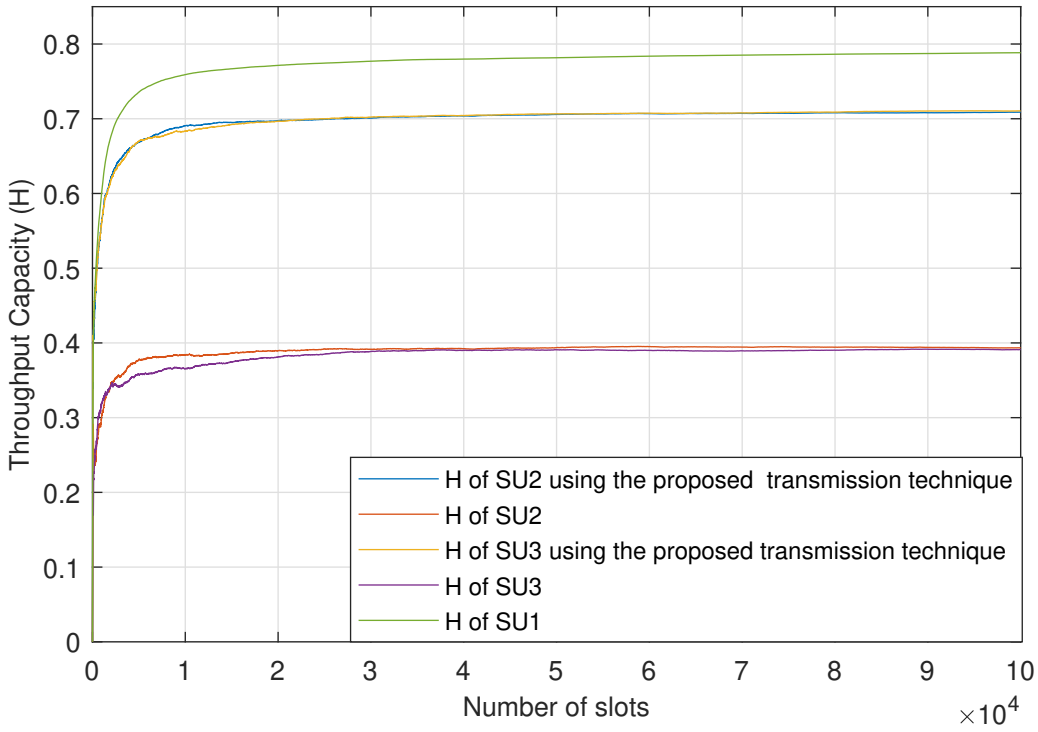

FIGURE 5.12: Throughput capacity of the primary and ordinary users

channel and 0 otherwise. The logical expression $\mathbf{1}_{(a)}$ equals 1 if a is true and 0 otherwise, and $||$ indicates the OR operation. Then, the maximum data throughput achievable by an ordinary user SU_o up to n can be expressed as follows:

$$H_o(n) = \sum_{t=1}^{n} \frac{1}{t} \left(H_o^L(t) \right) \tag{5.35}$$

In Fig. 5.12, the throughput capacity of the priority users SU_p exceeds that of the two ordinary users SU_{o1} and SU_{o2}. After a finite number of time slots, and based on our policy PFA, the priority user converges towards the most vacant channel, i.e. $\mu_1 = 0.9$. and his percentage of successful transmission will be around 90%. Although the throughput achievable by the priority user should be around 90%, SU_p achieves only %80 as shown in Fig. 5.12. This throughput decrease is referred to the collisions produced among users. Fig. 5.12 also shows the throughput capacity of the ordinary users with and without the proposed transmission technique. Beside the two ordinary achieve the same throughput, there is an enhancement of 30% when the proposed transmission technique is considered. Indeed, when the ordinary

users transmit directly to the primary Base Station, their reached throughput is around 40% of the throughput while 70% of the throughput can be achieved by the ordinary users using the proposed transmission technique. Finally, the improvement of the SNR using the proposed transmission technique can enhance the throughput capacity of the ordinary users.

5.7 Conclusion

This chapter focuses on the multiple Secondary Users (SUs) scenario in the Opportunistic Spectrum Access (OSA). In order to help the SUs make a decision, Multi-Armed Bandit (MAB) has recently attracted the attention and seems to be a suitable solution to solve the OSA problem. The well-known MAB algorithms such Thompson Sampling (TS), Upper Confidence Bound (UCB) or e-greedy applied in OSA, tackle the case of a single user, while several SUs may exist in the network trying to learn separately the vacant probabilities of channels. For multiple users, accessing the spectrum can be cooperative or competitive. In this chapter, we focused on the competitive access in which a novel learning policy, called All-Powerful Learning (APL), for the priority access is proposed. Our proposed APL does not require any cooperation or prior knowledge about the existing channels as do several existing policies, such as: Selective learning of the K-th largest expected rewards (SLK), Musical Chair, Multi-user e-greedy collision Avoiding (MEGA) and k-th MAB. It should be noticed that, the priority access has not been studied sufficiently in the literature although SLK and k-th MAB are two of the rare algorithms for the priority access based on UCB1 and e-greedy. While APL is not based on any specific MAB algorithm and can be used with any MAB algorithm, such as: AUCB, TS, UCB1, e-greedy. APL also considers the priority dynamic access while, to the best of our knowledge, only the priority or the dynamic access are considered in several recent works. It has been shown that when a priority user leaves the network, a little increase of regret can be observed.

Unlike APL that takes into consideration U priority levels for U users, we proposed a novel policy, called Priority Fairness Access (PFA), to consider two or more hierarchical levels for U users. PFA represents a novel issue, not yet considered in the literature, to tackle the priority and the fairness access. In our simulations, we considered two sets of secondary users: Priority users set with one member and ordinary users set with two members. It has been

shown that the priority user often accesses the most vacant channel while the ordinary users access uniformly the almost-best channels.

Chapter 6

Overview, General Conclusions and Future Work

Contents

6.1 Conclusion

Cognitive Radio (CR) can reduce the impending spectrum scarcity problem and increase its efficiency by allowing Opportunistic Spectrum Access (OSA) to the licensed spectrum bands. Indeed, in OSA, Secondary users (SUs) search for holes (white space) by sensing the spectrum. The presence of such holes in the primary users licensed channels are exploited by the SUs in order to transmit their data. In this thesis, we investigate several fundamental decision-making techniques used in CR, and more precisely in the Opportunistic Spectrum Access in order to help the SUs make good decisions.
In Chapter 1, we discussed, on the one hand, the technological advancement in wireless networks and mobile telephony that makes the frequency bands more and more crowded. On the other hand, the recent studies about the usage of the spectrum showed that up to 60% of the frequency bands is not used in several regions in the United States (US). We also investigate the rise of CR, the Cognitive Cycle (CC) and the Software-Defined Radio (SDR) that represents the core of CR.
In Chapter 2, we formulate the access to channels as Multi-Armed Bandit problem in order to help a SU make a good decision. This chapter also investigates the state-of-the-art of MAB algorithms that are widely used in OSA namely: Thompson Sampling (TS), Upper confidence Bound (UCB), ϵ-greedy, etc.
Our contributions detailed in Chapters 2, 3, 4 and 5 are summarized as follows:

- New MAB algorithm, called ϵ-UCB, based on the classical UCB in order to learn the vacant probabilities of the available communication channels. ϵ-UCB can achieve better results compared to several MAB algorithms by quickly finding the best channel with the highest vacant probability. In ϵ-UCB, the exploration-exploitation phases are separated. Indeed, during the learning period, ϵ-UCB gives a particular importance to the exploration in order to gather enough information about the vacant probabilities of channels. Then, ϵ-UCB focuses on the exploitation by regularly accessing the best channel. We have also proven that the upper bound of the sum regret achieves a logarithmic asymptotic behavior. It has shown that ϵ-UCB represents a suitable solution for a static environment where the vacant probabilities of the channels do not evolve over time while for dynamic environment,

the performance of *e*-UCB decreases significantly as most of MAB algorithms.

- To solve the main drawback of *e*-UCB in a dynamic environment, we propose AUCB that can achieve better results compared to *e*-UCB as well to several existing versions of UCB. Unlike *e*-UCB, the main advantage of AUCB that balances between exploration and exploitation during the learning period. After this period, the user gives the most important effect to the exploitation with a little impact on the exploration. Thanks to this latter property, AUCB can detect the dynamism of the vacant probabilities of the channels and thus easily adapt to a given environment.

- A novel cooperative policy, called Side Channel policy, is proposed in order to help users to learn collectively the vacant probabilities of the available channels. This policy requires some cooperation level among users to learn in an unknown environment. It has been shown that, after a finite number of slots, each user converges towards his dedicated channel that corresponds to his prior rank. Moreover, a significant improvement is achieved with Side Channel policy compared to several existing policies. Indeed, based on our policy, the users can reach their targeted channels faster than other policies that manage a secondary network such as Random Rank and SLK. Our proposed policy can also be used for dynamic priority access where the priority users can enter or leave the network.

- For the competitive access, we proposed a novel policy called All-Powerful Learning (APL) for the priority access. This former does not require any cooperation or prior knowledge about the vacant probabilities of channels. APL can achieve better results compared to several existing policies for the competitive access such as Musical Chair and SLK. We also investigated the upper bound of regret for APL and showed that it achieves a logarithmic asymptotic behavior. That means that, after a finite number of time slots the users are able to learn the vacant probabilities of channels and often access their targeted channels. Finally, it has been shown that APL represents a good solution for competitive priority dynamic access.

- We proposed another competitive policy, called Priority Fairness Access (PFA), that considers different priority levels where each level contains one or more users. PFA achieves the fairness for the users who have the same priority level by letting them uniformly access the available channels. The main advantage of the PFA is that it can be used to extend their transmission range as well to enhance the users' transmission rate.

- We investigated the quality of service in the secondary network where the users are able to learn not only the vacant probabilities of channels but also their quality. Thereby, the main target of a user is to find the best channel with the best quality and highest availability.

- Finally, a proof of concept has been developed for e-UCB, AUCB, Side Channel and APL in order to ensure the convergence of our proposed methods in a real radio environment. Via the analytical convergence analysis as well as the experimental results, it has been shown that the regret achieves a logarithmic asymptotic behavior.

6.2 Perspective and Future Work

Opportunistic Spectrum Access in Cognitive Radio is still an attractive research to reach a high efficiency in the current networks. Our proposed MAB, for a single user or multiple-user cases, can be considered to be an effective solution for enhance system reliability. These methods that we developed in this thesis have the ability to be improved further:

- Generally, like most important works based on the Multi-Armed Bandit technique, this work focused on the Independent Identical Distributed (IID) model in which the state of each channel is supposed to be drawn from IID process. While Markov process represents another important technique that may represent a more realistic model to describe the state of the available channels, although it is a complex process compared to IID.

- In the Side Channel policy, each user broadcasts his choice of channels to other users in order to avoid any collision in the next slot. However, the broadcast packet of the user risks to be lost, then a collision may occur among users. Therefore, for a more realistic model, considering

the case of non-error free in the Side Channel can be included in the future work.

- Consider a dynamic environment where the vacancy of channels evolves over time. Indeed, unlike *e*-UCB and several existing MAB algorithms, we believe that AUCB offers an important solution to adapt to a dynamic environment where the exploration is still having some impact after the learning period.

- In PFA, different priority levels may be considered with one or more users for each level. In our study, and for the simplicity sake, we considered a simplistic case of PFA with only two levels and the scenario with more than two levels should be investigated. Moreover, a concept proof of PFA should be done.

- The analytical convergence of APL has been investigated for a fixed number of users, while, for dynamic access, its performance has been tested only via simulation. Then, developing an analytical proof of APL for dynamic access is required in order to show its performance in a real radio environment.

- For a more realistic model, the future work may also investigate the effect of using the state-of-the-art spectrum sensing techniques to detect the activity of the Primary Users on the performance of the learning and decision-making. Moreover, considering the imperfect sensing, i.e. the probability of false alarm and miss detection, represents a new challenge to developing a more realistic network.

Appendix A

Upper Bound of $\mathbb{A}$ in e-UCB

In this Appendix, we investigate the upper bound of $\mathbb{A} = \epsilon_t \times Prob$ in e-UCB where $Prob$ can be expressed as follows:

$$Prob \leq P\big\{B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1)); T_i(t-1) \geq l\big\}$$

The index of the i-th channel $B_i(t, T_i(t))$ is based on the exploration, $X_i(T_i(t))$, and the exploitation, $A_i(t, T_i(t))$:

$$B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t)) \tag{A.1}$$

Then, we obtain:

$$Prob \leq P\Bigg\{X_1(T_1(t-1)) + A_1(t-1, T_1(t-1)) \leq X_i(T_i(t-1)) + A_i(t-1, T_i(t-1)) \text{ and } T_i(t-1) \geq l\Bigg\} \tag{A.2}$$

By taking the minimum value of $X_1(T_1(t-1)) + A_1(t-1, T_1(t-1))$ and the maximum value of $X_i(T_i(t-1)) + A_i(t-1, T_i(t-1))$ at each time slot, we can upper bound $Prob$ by the following equation:

$$Prob \leq P\Bigg\{\min_{0<S_1<t}\Big[X_1(S_1) + A_1(t, S_1)\Big] \leq \max_{l \leq S_i < t}\Big[X_i(S_i) + A_i(t, S_i)\Big]\Bigg\} \tag{A.3}$$

where $S_i \geq l$ to fulfill the condition $T_i(t-1) \geq l$. Then, we obtain:

$$Prob \leq \sum_{S_1=1}^{t-1}\sum_{S_i=l}^{t-1} P\Big\{X_1(S_1) + A_1(t, S_1) < X_i(S_i) + A_i(t, S_i)\Big\} \tag{A.4}$$

The above probability can be upper bounded by:

$$Prob \leq \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} P\Big\{X_1(S_1) + A_1(t,S_1) \leq \mu_1\Big\} + P\Big\{\mu_1 < \mu_i + 2A_i(t,S_i)\Big\} + P\Big\{X_i(S_i) + A_i(t,S_i) \geq \mu_i + 2A_i(t,S_i)\Big\} \quad (A.5)$$

Using the ceiling operator $\lceil\rceil$, let $l = \lceil \frac{4\alpha \ln(n)}{\Delta_i^2} \rceil$, where $\Delta_i = \mu_1 - \mu_i$ and $S_i \geq l$, then the event $\mu_1 < \mu_i + 2A_i(t,S_i)$ in eq (A.5) becomes false, in fact:

$$\begin{aligned} \mu_1 - \mu_i - 2A_i(t,S_i) &= \mu_1 - \mu_i - 2\sqrt{\frac{\alpha \ln(t)}{S_i}} \\ &\geq \mu_1 - \mu_i - 2\sqrt{\frac{\alpha \ln(n)}{l}} \\ &\geq \mu_1 - \mu_i - \Delta_i = 0 \end{aligned}$$

Based on eq (A.5), we obtain:

$$Prob \leq \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} P\Big\{X_1(S_1) \leq \mu_1 - A_1(t,S_1)\Big\} + P\Big\{X_i(S_i) \geq \mu_i + A_i(t,S_i)\Big\} \quad (A.6)$$

Using Chernoff-Hoeffding bound[1] [96], we can prove that:

$$\begin{aligned} P\Big\{X_1(S_1) \leq \mu_1 - A_1(t,S_1)\Big\} &\leq \exp^{\frac{-2}{S_1}\left[S_1\sqrt{\frac{\alpha \ln(t)}{S_1}}\right]^2} \\ &= t^{-2\alpha} \end{aligned} \quad (A.7)$$

$$\begin{aligned} P\Big\{X_i(S_i) \geq \mu_i + A_i(t,S_i)\Big\} &\leq \exp^{\frac{-2}{S_i}\left[S_i\sqrt{\frac{\alpha \ln(t)}{S_i}}\right]^2} \\ &= t^{-2\alpha} \end{aligned} \quad (A.8)$$

[1] According to [96], Chernoff-Hoeffding theorem is defined as follows: Let $X_1, ..., X_n$ be random variables in [0,1], and $E[X_t] = \mu$, and let $S_n = \sum_{i=1}^{n} X_i$. Then $\forall\ a \geq 0$, we have $P\{S_n \geq n\mu + a\} \leq \exp^{\frac{-2a^2}{n}}$ and $P\{S_n \leq n\mu - a\} \leq \exp^{\frac{-2a^2}{n}}$.

The two inequations above and inequation (A.6) lead us to:

$$Prob \leq \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} 2t^{-2\alpha} \leq 2t^{-2\alpha+2} \tag{A.9}$$

Finally, we obtain:

$$\mathbb{A} \leq \frac{H}{t} \times 2t^{-2\alpha+2} = 2H \times t^{-2\alpha+1} \tag{A.10}$$

Appendix B

Upper Bound of Z in e-UCB

This appendix stands for finding an upper bound of Z that contributes to finding an upper bound of e-UCB:

$$Z = p\{X_1(T_1(t-1)) \leq a; T_1(t-1) \geq l\} \tag{B.1}$$

where a is a constant number that can be chosen as follows: $a = \frac{\mu_1+\mu_i}{2} = \mu_1 - \frac{\Delta_i}{2} = \mu_i + \frac{\Delta_i}{2}$, and $\Delta_i = \mu_1 - \mu_i$. After the learning period where $T_i(t-1) \geq l$, we have: $T_1(t-1) >> T_i(t-1)$. Then Z can be upper bounded by:

$$\begin{aligned} Z \leq & p\{X_1(T_1(t-1)) \leq a; T_1(t-1) \geq l\} \qquad \text{(B.2)} \\ \leq & \sum_{z=l}^{n} p\{X_1(T_1(t-1)) \leq \mu_1 - \frac{\Delta_i}{2}; T_1(t-1) = z\} \\ \leq & \sum_{z=l}^{n} p\{X_1(z) \leq \mu_1 - \frac{\Delta_i}{2}\} \end{aligned}$$

(B.3)

Using the Chernoff-Hoeffding in [125] [1], we can upper bound the above equation as follows:

$$Z \leq \sum_{z=l}^{n} \exp^{-\frac{2\Delta_i^2 z^2}{4z}} \leq n \exp^{\frac{-l\Delta_i^2}{2}} \tag{B.4}$$

According to the proof provided in appendix A, we have $l = \frac{8\ln(n)}{\Delta_i^2}$. So, we obtain:

$$Z \leq n \exp^{-4\ln n} = \frac{1}{n^3} \tag{B.5}$$

[1] According to [96], Chernoff-Hoeffding theorem is defined as follows: Let $X_1, ..., X_n$ be random variables in [0,1], and $E[X_t] = \mu$, and let $S_n = \sum_{i=1}^{n} X_i$. Then $\forall\ a \geq 0$, we have $P\{S_n \geq n\mu + a\} \leq \exp^{\frac{-2a^2}{n}}$ and $P\{S_n \leq n\mu - a\} \leq \exp^{\frac{-2a^2}{n}}$.

Appendix C

Upper Bound of $T_i(n)$ in AUCB

In this Appendix, we investigate the upper bound of $E[T_i(t)]$ in order to obtain an upper bound of Regret of AUCB. $E\left[T_i(n)\right]$ under AUCB can be upper bounded by the following expression:

$$E[T_i(n)] \leq \frac{8\ln(n)}{\Delta_i^2} + 1 + 2\sum_{t=1}^{n} t^{-2} \tag{C.1}$$

To resolve $\sum_{t=1}^{n} t^{-2}$, we consider the Taylor's series expansion of $\sin(t)$:

$$\sin(t) = t - \frac{t^3}{3!} + ... + (-1)^{2k+1}\frac{t^{2k+1}}{(2k+1)!} + ... \tag{C.2}$$

As $\sin(t) = 0$ when $t = \pm k\pi$, then we obtain:

$$\begin{aligned}\sin(t) &= t \times \left(1 - \frac{t^2}{\pi^2}\right) \times ... \times \left(1 - \frac{t^2}{k^2\pi^2}\right)... \\ &= t - \Big(\sum_{i=1}^{n} \frac{1}{i^2\pi^2}\Big)t^3 + ...\end{aligned}$$

where q_k is a general coefficient. By comparing the above equation with eq (C.2), we obtain $\sum_{i=1}^{n} \frac{1}{i^2} = \frac{\pi^2}{3!}$. Finally, we obtain the upper bound of $E[T_i(n)]$ as follows:

$$E[T_i(n)] \leq \frac{8\ln(n)}{\Delta_i^2} + 1 + \frac{\pi^2}{3} \tag{C.3}$$

Appendix D

Upper Bound of S_S in MAUCB

Let us estimate the time S_s and let us consider U users with different priority levels based on our policy APL. At a certain moment, supposing that each user has a random rank, then at least two of them may have the same rank, and a collision may occur. In this case, the collide users should regenerate a random rank around their prior rank[1]. After a finite number of collisions, the system will converge to the steady state where each user has a unique rank, i.e. its prior rank. Let S_s be a random variable with a countable set of finite outcomes $1, 2, ...$ occurring with the probability p_1, p_2... respectively, where p_t represents a non-collision at instant t. The expectation of S_s can be expressed as follows:

$$E[S_s] = \sum_{t=1}^{\infty} tp[S_s = t] \tag{D.1}$$

where the random variable S_s follows the probability $p[S_s = t]$:

$$p[S_s = t] = (1-p)^t p$$

and p indicates the probability of non-collision at an instant t, while $(1-p)^t$ indicates the probability of having collisions from the instant 0 till $t-1$. Then we obtain:

$$E[S_s] = \sum_{t=1}^{\infty} t(1-p)^t p \tag{D.2}$$

Let $I_a(x)$ be defined as follows:

$$I_a(x) = (1-a)\sum_{t=1}^{\infty}(ax)^t \tag{D.3}$$

[1]For SU_k, it should regenerate a rank in the set $\{1, ...k\}$.

TABLE D.1: Three SUs trying to converge toward a steady state where each one finds its prior rank. The roman number indicates the number of users selecting the same rank

Cases	Rank 1	Rank 2	Rank 3
1	|	||	0
2	|||	0	0
3	|	|	|
4	||	|	0
5	||	0	|

where a is a constant number such that $ax < 1$. $I(x)$ can converge to:

$$I_a(x) = \frac{(1-a)ax}{1-ax}$$

Based on the previous equation, we have:

$$\frac{dI_a(x)}{dx} = \frac{(1-a)a}{(1-ax)^2}$$

Using the previous equation, we obtain:

$$\frac{dI_a(x)}{dx}_{x=1} = \frac{a}{(1-a)} \tag{D.4}$$

Considering that $a = 1-p$, we conclude that $E[S_s] = \frac{1-p}{p}$. To clarify the idea and estimate the probability p, we consider that three SUs are trying to find their prior rank where the Table (D.1) displays all the possible cases. Subsequently, the probability to converge to a steady state, i.e. the case 3, is $p = \frac{1}{5}$, and $E[S_s] = 4$.

To estimate the value of p as well $E[S_s]$, let us introduce the problem suggested in [105, Chapter 5], to count the number of ways of putting U identical balls into U different boxes.

According to [105, Chapter 5], the probability p to converge to a steady state where each box has just one ball is $p = \frac{1}{\binom{U}{2U-1}}$ and $E[S_s] = \binom{U}{2U-1} - 1$. However, our problem of convergence to a steady state represents a restricted case of the problem introduced in [105]. Then, the expected time to converge to a steady state of our policy APL for U SUs can be upper bounded by:

$$E[S_s] \leq \binom{U}{2U-1} - 1 \tag{D.5}$$

List of publications

Journal Papers:

[1] **M. Almasri**, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune "Distributed algorithm under Cooperative or Competitive Users with Priority Access in Cognitive Networks," in EURASIP Journal on Wireless Communications and Networking, vol. 2020, no. 1, p. 1-31 (2020).

[2] **M. Almasri**, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune, "Priority Dynamic Access based on Multi-Armed Bandits, in Advances in Science, Technology and Engineering Systems, vol. 5, no. 4, p. 223-233 (2020).

[3] **M. Almasri**, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune, "Dynamic DecisionMaking Process in the Opportunistic Spectrum Access" in Wireless Personal Communications. https://doi.org/10.1007/s11277-020-08064-w.

International Conference Papers:

[1] **M. Almasri**, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune, "Opportunistic spectrum access in cognitive radio for tactical network," in European Conference on Electrical Engineering and Computer Science (EECS), Bern, Switzerland, December 2018.

[2] **M. Almasri**, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune, "All-powerful learning algorithm for the priority access in cognitive network," in European Signal Processing Conference (EUSIPCO), A Coruna, Spain, September 2019.

[3] **M. Almasri**, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune, "Distributed algorithm to learn osa channels availability and enhance the transmission rate of secondary users," in International Symposium on Communications and Information Technologies (ISCIT), Ho Chi Minh, Vietnam, September 2019.

[4] **M. Almasri**, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune, "Managing Single or Multi-User Channel Allocation for the Priority Cognitive Access", in European Signal Processing Conference (EUSIPCO), Amsterdam, Pays-Bas, Janvier 2021.

[5] **M. Almasri**, A.Mansour, L.White, "Stackelberg and MAB Models for Decision-Making Process" in Conference on Cognitive Radio Oriented Wireless Networks (EUSIPCO), (Submitted in April 2020)

Bibliography

[1] A. Kumbhar. Overview of ism bands and software-defined radio experimentation. *Wireless Personal Communications*, 97(3):3743–3756, 2017.

[2] H. Mazar. International, regional and national regulation of SRDs. In *Proceedings of ITU Workshop on Short Range Devices and Ultra Wide Band (UWB)*, 2014.

[3] Federal Communications Commission. FCC online table of frequency allocations. *Online: http://www. fcc. gov/oet/spectrum/table/fcctable*, 2008.

[4] R. Simons and K. Goverdhanam. Applications of nano-satellites and cube-satellites in microwave and rf domain. In *MTT-S International Microwave Symposium*, pages 1–4, 2015.

[5] C. Frost, E. Agasid, et al. Small spacecraft technology state of the art. *NASA Technical Report TP-2014-216648/REV1, NASA Ames Research Center*, 2014.

[6] X. Hong, J. Wang, C. Wang, and J. Shi. Cognitive radio in 5g: a perspective on energy-spectral efficiency trade-off. *IEEE Communications Magazine*, 52(7):46–53, 2014.

[7] Y. Liang, K. Chen, G. Li, and P. Mahonen. Cognitive radio networking and communications: An overview. *IEEE transactions on vehicular technology*, 60(7):3386–3407, 2011.

[8] E. Hossain, D. Niyato, and D. Kim. Evolution and future trends of research in cognitive radio: a contemporary survey. *Wireless Communications and Mobile Computing*, 15(11):1530–1564, 2015.

[9] A. Volkwin. *Suitability of a commercial software defined radio system for passive coherent location*. PhD thesis, University of Cape Town, 2008.

[10] J. Palicot. *Radio engineering: From software radio to cognitive radio*. John Wiley & Sons, 2013.

[11] SDR Forum. SDRF cognitive radio definitions. http://data.memberclicks.com/site/sdf.

[12] M. Marcus, C.J. Burtle, B. Franca, A. Lahjouji, and N. McNeil. Federal Communications Commission: Spectrum policy task force. In *ET Docket no. 02-135*, November 2002.

[13] V. Garg. *Wireless communications & networking*. Elsevier, 2010.

[14] S. Benedetto, L. Correia, and M. Luise. *The Newcom++ Vision Book*. Springer, 2011.

[15] B. Benmammar and A. Amraoui. *Radio resource allocation and dynamic spectrum access*. Wiley Online Library, 2013.

[16] H. Harada. Software defined radio prototype toward cognitive radio communication systems. In *International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005.*, pages 539–547, 2005.

[17] Federal Communications Commission. Establishment of interference temperature metric to quantify and manage interference and to expand available unlicensed operation in certain fixed mobile and satellite frequency bands. *Et Docket*, (03-237), 2003.

[18] B. Wang and KJ. Liu. Advances in cognitive radio networks: A survey. *IEEE Journal of selected topics in signal processing*, 5(1):5–23, 2010.

[19] M. López-Martínez, J. Alcaraz, L. Badia, and M. Zorzi. A superprocess with upper confidence bounds for cooperative spectrum sharing. *IEEE Trans. on Mobile Computing*, 15(12):2939–2953, 2016.

[20] Xinxin Feng, Gaofei Sun, Xiaoying Gan, Feng Yang, Xiaohua Tian, Xinbing Wang, and Mohsen Guizani. Cooperative spectrum sharing in cognitive radio networks: A distributed matching approach. *IEEE Trans. on Com.*, 62(8):2651–2664, 2014.

[21] I. Akyildiz, W. Lee, M. Vuran, and S. Mohanty. Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *Computer networks*, 50(13):2127–2159, 2006.

[22] A. Goldsmith and I. Maric. Capacity of cognitive radio networks. In *Principles of Cognitive Radio Networks*. Cambridge University Press, 2012.

[23] T. Yucek and H. Arslan. A survey of spectrum sensing algorithms for cognitive radio applications. *IEEE communications surveys & tutorials*, 11(1):116–130, 2009.

[24] I. Akyildiz, B. Lo, and R. Balakrishnan. Cooperative spectrum sensing in cognitive radio networks: A survey. *Physical communication*, 4(1):40–62, 2011.

[25] J. Mitola. *An integrated agent architecture for software defined radio*. PhD thesis, Royal Inst. of Technology, 2000.

[26] P. Kolodzy. Spectrum policy task force report. *Federal Communications Commission ET Docket 02, vol. 135*, 2002.

[27] Q. Zaho and BM. Sadler. A survey of dynamic spectrum access: Signal processing, networking and regulatory policy. *IEEE Signal Processing Magazine*, 55(5):2294–2309, 2007.

[28] S. Haykin. Brain-empowered wireless communications. *IEEE Journal on Selected Areas in Commun*, 23(2):201–220, 2005.

[29] C. Clancy, J. Hecker, S. Erich, and T. Shea. Applications of machine learning to cognitive radio networks. *IEEE Wireless Communications*, 14(4):47–52, 2007.

[30] A. Mody, S. Blatt, N. Thammakhoune, T. McElwain, J. Niedzwiecki, D. Mills, M. Sherman, and C. Myers. Machine learning based cognitive communications in white as well as the gray space. In *MILCOM 2007-IEEE Military Communications Conference*, pages 1–7, 2007.

[31] S. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Malaysia Pearson Education Limited, 2016.

[32] E. Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[33] R. Agrawal. Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

[34] W. Jouini, D. Ernst, C. Moy, and J. Palicot. Upper confidence bound based decision making strategies and dynamic spectrum access. In *ICC*, Cape Town, South Africa, May 2010.

[35] A. Anandkumar, N. Michael, A. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Sel. Areas in Com.*, 29(4):731–745, 2011.

[36] K. Liu and Q. Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Trans. Signal Processing*, 58(11):5667–5681, 2010.

[37] Y. Gai, B. Krishnamachari, and R. Jain. Learning multiuser channel allocations in cognitive radio networks: a combinatorial multi-armed bandit formulation. In *IEEE Symp. on Dynamic Spectrum Access Networks*, Singapore, April 2010.

[38] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[39] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Trans. on Autom. Cont.*, 32(11):968–976, 1987.

[40] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[41] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays, part ii: Markovian rewards. *IEEE Trans. on Autom. Cont.*, 32(11):977–982, 1987.

[42] N. Modi, P. Mary, and C. Moy. Qos driven channel selection algorithm for cognitive radio network: Multi-user multi-armed bandit approach. *IEEE Trans. on Cog. Com. & Networking*, 3(1):1–6, 2017.

[43] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[44] N. Modi, P. Mary, and C. Moy. Qos driven channel selection algorithm for opportunistic spectrum access. In *Globecom Workshops SDRAN-CAN*, December 2015.

[45] C. Tekin and M. Liu. Online learning in opportunistic spectrum access: A restless bandit approach. In *INFOCOM*, April 2011.

[46] H. Liu, K. Liu, and Q. Zhao. Logarithmic weak regret of non-bayesian restless multi-armed bandit. In *IEEE (ICASSP)*, pages 1968–1971, 2011.

[47] C. John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge, 1989.

[48] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[49] B. Giuseppe, J. Loeppky, and R. Lawrence. A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*, 2015.

[50] E. Kaufmann, O. Cappé, and A. Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, La Palma, Canary Islands, April 2012.

[51] O. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Annual conf. On Learning Theory*, Budapest, Hungary, July 2011.

[52] J. Audibert, S. Rémi Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. In *International conference on algorithmic learning theory*, pages 150–165. Springer, 2007.

[53] O. Cappé, A. Garivier, O. Maillard, R. Munos, and G. Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

[54] S. Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

[55] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, Granada, Spain, December 2011.

[56] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *conf. on Learning Theory*, Edinburgh, Scotland, June 2012.

[57] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conf. on Algorithmic Learning Theory*, Lyon, France, October 2012.

[58] S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, Scottsdale, USA, April 2013.

[59] K. Liu, Q. Zhao, and B. Krishnamachari. Decentralized multi-armed bandit with imperfect observations. In *Annual Allerton Conference on Communication, Control, and Computing*, Monticello, USA, October 2010.

[60] J. Rosenski, O. Shamir, and L. Szlak. Multi-player bandits-a musical chairs approach. In *ICML*, New York, USA, June 2016.

[61] Y. Gai and B. Krishnamachari. Decentralized online learning algorithms for opportunistic spectrum access. In *GLOBECOM*, Texas, USA, December 2011.

[62] N. Torabi, K. Rostamzadeh, and V. C. Leung. Rank-optimal channel selection strategy in cognitive networks. In *GLOBECOM*, California, USA, December 2012.

[63] O. Avner and S. Mannor. Concurrent bandit and cognitive radio networks. In *European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Nancy, France, September 2014.

[64] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune. All-powerful learning algorithm for the priority access in cognitive network. In *EUSIPCO*, A Coruña, Spain, September 2019.

[65] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune. Distributed algorithm under cooperative or competitive users with priority access in cognitive networks. *EURASIP journal on wireless communications and networking*, (Accepted).

[66] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune. Distributed algorithm to learn osa channels availability and enhance the transmission rate of secondary users. In *International Symposium on Communications and Information Technologies (ISCIT)*, Ho Chi Minh, Vietnam, September 2019.

[67] F. Fu and M. Schaar. Learning to compete for resources in wireless stochastic games. *IEEE Transactions on Vehicular Technology*, 58(4):1904–1919, 2008.

[68] M. Di, K. Chowdhury, and L. Bononi. Learning with the bandit: A cooperative spectrum selection scheme for cognitive radio networks. In *2011 IEEE Global Telecommunications Conference-GLOBECOM 2011*, pages 1–6. IEEE, 2011.

[69] D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

[70] S. Kar, V. Poor, and S. Cui. Bandit problems in networks: Asymptotically efficient distributed allocation rules. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 1771–1778. IEEE, 2011.

[71] D. Hill, H. Nassif, Y. Liu, A. Iyer, and S. Vishwanathan. An efficient bandit algorithm for realtime multivariate optimization. In *International Conference on Knowledge Discovery and Data Mining*, pages 1813–1821, 2017.

[72] T. Graepel, J. Candela, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. Omnipress, 2010.

[73] D. Agarwal. Computational advertising: the linkedin way. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1585–1586, 2013.

[74] D. Agarwal, B. Long, J. Traupman, D. Xin, and L. Zhang. Laser: A scalable response prediction platform for online advertising. In *ACM international conference on Web search and data mining*, pages 173–182, 2014.

[75] S. Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

[76] S. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31(1):37–45, 2015.

[77] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3):285–294, 1933.

[78] D. Cabric, S. Mishra, and R. Brodersen. Implementation issues in spectrum sensing for cognitive radios. In *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.*, volume 1, pages 772–776. Ieee, 2004.

[79] D. Bhargavi and C. Murthy. Performance comparison of energy, matched-filter and cyclostationarity-based spectrum sensing. In *2010 IEEE 11th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, 2010.

[80] S. Kapoor, S. Rao, and G. Singh. Opportunistic spectrum sensing by employing matched filter in cognitive radio network. In *2011 International Conference on Communication Systems and Network Technologies*, pages 580–583, 2011.

[81] I. Glover and P. Grant. *Digital communications*. Pearson Education, 2010.

[82] M. Oner and E. Jondral. Cyclostationarity based air interface recognition for software radio systems. In *Proceedings. 2004 IEEE Radio and Wireless Conference*, pages 263–266. IEEE, 2004.

[83] P. Urriza, E. Rebeiz, and D. Cabric. Multiple antenna cyclostationary spectrum sensing based on the cyclic correlation significance test. *IEEE Journal on Selected Areas in Communications*, 31(11):2185–2195, 2013.

[84] Harry Urkowitz. Energy detection of unknown deterministic signals. *Proceedings of the IEEE*, 55(4):523–531, 1967.

[85] J. Wu, T. Luo, and G. Yue. An energy detection algorithm based on double-threshold in cognitive radio systems. In *2009 First International Conference on Information Science and Engineering*, pages 493–496. IEEE, 2009.

[86] C. Liu, M. Li, and M. Jin. Blind energy-based detection for spatial spectrum sensing. *IEEE Wireless Communications Letters*, 4(1):98–101, 2014.

[87] F. Qi, Y. Zhihui, and S. Keqin. Spectrum environment machine learning in cognitive radio. *Procedia Engineering*, 29:4181–4185, 2012.

[88] M. Alsheikh, S. Lin, D. Niyato, and H. Tan. Machine learning in wireless sensor networks: Algorithms, strategies, and applications. *IEEE Communications Surveys & Tutorials*, 16(4):1996–2018, 2014.

[89] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. Chen, and L. Hanzo. Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications*, 24(2):98–105, 2016.

[90] M. Sun, G. Lebanon, and P. Kidwell. Estimating probabilities in recommendation systems. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 734–742, 2011.

[91] J.G. Van Bosse and F.U. Devetak. *Signaling in telecommunication networks*. John Wiley & Sons, Canada, 2 edition, 11 2006.

[92] W. Jouini, D. Ernst, C. Moy, and J. Palicot. Multi-armed bandit based policies for cognitive radio's decision making issues. In *International conf. on Signals, Circuits and Systems*, Djerba, Tunisia, November 2009.

[93] C. Tekin and M. Liu. Online algorithms for the multi-armed bandit problem with markovian rewards. In *Annual Allerton Conf. on Com., Control, and Computing*, Monticello, USA, September 2010.

[94] B. Giuseppe, J. Loeppky, and R. Lawrence. A survey of online experiment design with the stochastic multi-armed bandit. In *arXiv preprint arXiv:1510.00757*, 2015.

[95] AL Cauchy. *Sur la convergence des séries*, volume 2. 1889.

[96] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[97] L. Melian-Gutierrez, N. Modi, C. Moy, F. Bader, I. Perez-lvarez, and S. Zazo. Hybrid ucb-hmm: A machine learning strategy for cognitive radio in hf band. *IEEE Trans. on Cog. Com. & Networking*, 1(3):347–358, 2015.

[98] L. Melián-Gutiérrez, N. Modi, C. Moy, I. Pérez-Álvarez, F. Bader, and S. Zazo. Upper confidence bound learning approach for real hf measurements. In *ICC Workshop*, London, UK, June 2015.

[99] Y. Gai and B. Krishnamachari. Decentralized online learning algorithms for opportunistic spectrum access. In *GLOBECOM*, Texas, USA, December 2011.

[100] Mahmoud Almasri, Ali Mansour, Christophe Moy, Ammar Asoum, Christophe Osswald, and Denis Lejeune. Opportunistic spectrum access in cognitive radio for tactical network. In *European Conference on Electrical Engineering and Computer Science*, Bern, Switzerland, December 2018.

[101] O. Simeone, I. Stanojev, S. Savazzi, Y. Bar-Ness, U. Spagnolini, and R. Pickholtz. Spectrum leasing to cooperating secondary ad hoc networks. *IEEE Journal on Selected Areas in Com.*, 26(1):203–213, 2008.

[102] J. Zhang and Q. Zhang. Stackelberg game for utility-based cooperative cognitiveradio networks. In *ACM international Symp. on Mobile ad hoc networking computing*, New Orleans, USA, May 2009.

[103] H. Liu, K. Liu, and Q. Zhao. Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Trans. on Inf. Theory*, 59(3):1902–1916, 2013.

[104] X. Chen, Z. Zhao, and H. Zhang. Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks. *IEEE Trans. on mobile computing*, 12(11):2155–2166, 2013.

[105] M. Bóna. *A walk through combinatorics: An Introduction to Enumeration and Graph Theory, 2nd edition*. World Scientific Publishing Company, London, 2006.

Title: A PhD thesis, Game Theory for Tactical Networks

Keywords: Opportunistic Spectrum Access, Multi-Armed Bandit, Priority Dynamic Access

Since 1990's, the demand on wireless devices, mobile and wireless networks, has experienced unprecedented growth which makes the frequency bands more and more crowded. Several studies, initiated by the Federal Communications Commission (FCC) in the United States (US), have shown that the frequency bands are not well used: Some frequency bands are overlapped while others underutilized. The Opportunistic Spectrum Access (OSA) in Cognitive Radio (CR) represents one of several proposed solutions to tackle the scarcity and enhance the efficiency use of the spectrum. In OSA, two categories of users are considered: Primary Users (PUs), also known as licensed users, have the right to fully access their dedicated bandwidths; and Secondary Users (SUs), i.e. opportunistic users, would like to exploit vacant frequency bands unused by the PUs. Due to hardware limitation, a SU can access one channel at each time slot trying to reach the best channel with the highest vacancy probability. To identify the best channel, we formulate OSA as a Multi-Armed Bandit (MAB) problem, in which an agent plays one arm at each time trying to reach the optimal arm with the highest expected reward. Several MAB algorithms have been suggested to solve the MAB problem in the context of OSA, such as: Thompson Sampling (TS), Upper Confidence Bound (UCB), e-greedy, etc.

By focusing first on a single SU, we analyze the performance of the well-known MAB algorithms (i.e. TS, UCB, e-greedy) that deal with OSA. Thus, we propose our MAB algorithms based on UCB, called: e-UCB and AUCB. Both of them achieve good results compared to well-known variants of MAB algorithms, i.e. UCB and e-greedy, in which the SU can quickly learn the vacancy probability of channels without any information or prior knowledge about the available channels. Our analytical proof, as well as the simulation results, of e-UCB and AUCB show that the SU can efficiently distinguish and converge to the best channel after a finite number of time slots.

For multiple users, the big challenge of SUs remains to learn collectively (Cooperative learning) or separately (Competitive learning) the vacancy probabilities of the channels. As a matter of fact, a cooperative or competitive learning policy is required in order to manage the secondary network and decrease the number of collisions among users. Generally, the policies to manage a secondary network can be classified into two main categories: Random access or priority access. Most recent works in OSA focus on the random access while the priority access is not enough considered in the literature. In fact, the priority access can have an important role in tactical networks in which several SUs exist with some hierarchy levels.

In our work, we propose a cooperative and competitive policies for the priority access respectively called Side Channel and All-Powerful Learning (APL). In our policies, each SU has an assigned priority rank, and his target remains to access the channels according to his rank. Moreover, Side Channel and APL deal with the priority dynamic access where the users can enter into or leave the network. While, to the best of our knowledge, only the priority or dynamic access are considered in several recent works. Finally, a proof is developed to verify the performance of proposed learning policies on a real radio environment. Simulation results show that Side channel and APL can achieve better results than several recent works: the users can quickly reach their dedicated channels while decreasing the number of collisions among them.