Machine Learning and Statistical Decision Making for Green Radio

Navikkumar Modi

▶ **To cite this version:**

Navikkumar Modi. Machine Learning and Statistical Decision Making for Green Radio. Autre. CentraleSupélec, 2017. Français. NNT : 2017CSUP0002 . tel-01668536

HAL Id: tel-01668536

https://theses.hal.science/tel-01668536

Submitted on 20 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre : 2017-04-TH ANNÉE 2017

CentraleSupélec

UNIVERSITE BRETAGNE LOIRE

THÈSE / CENTRALESUPÉLEC
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE CENTRALESUPÉLEC

Mention : Télécommunications

Ecole doctorale 359 « Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Electronique (MATISSE) »

présentée par

Navikkumar MODI

Préparée à l'UMR 6164 - IETR (Equipe SCEE)
Institut d'Electronique et de Télécommunications de Rennes

Machine Learning and Statistical Decision Making for Green Radio

Thèse soutenue à Rennes le 17 May 2017

devant le jury composé de :

Visa KOIVUNEN
Professeur, Université d'Aalto, Finlande/ *Rapporteur*
Samson LASAULCE
Directeur de recherche CNRS, L2S / *Rapporteur*
Damien ERNST
Professeur, Université de Liège, Belgique/ *Président du jury, examinateur*
Emilie KAUFMANN
Chargée de Recherche CNRS, Université de Lille/ *Examinateur*
Philippe MARY
Maître de conférence, INSA, Rennes/ *Co-encadrant*
Christophe MOY
Professeur, CentraleSupélec, Rennes/ *Directeur de thèse*

I would like to dedicate this thesis to my loving parents, wife and daughter ***ARYA*** *...*

Institutional Acknowledgements

This work has received a French government support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference No. ANR-10-LABX-07-01. The authors would also like to thank the Region Bretagne, France, for its support of this work.

Acknowledgements

I would like to thank Christophe Moy and Philippe Mary, my thesis advisors, and Jacques Palicot, head of SCEE department, for giving me the chance to pursue research in a warm and productive environment. Their helpful comments, corrections and critical minds guided me all along this work providing me with a certain amount of independence. The fruitful conversations we had during this period have helped me develop the thematic outline for this thesis. I am very grateful for their motivation, guidance and patience. Their guidance and friendship made my 3 years in Rennes a great life experience.

I would also like to extend my gratitude to the SCEE members: Carlos Bader, Amor Nakfha, Yves Louet, Pascal Cotret and all professors of CentraleSupélec, Rennes, France. I would like to thank all the members of TEPN project from CominLabs for their fruitful discussions and suggestions. I extend my sincere thanks to Karine Bernard, Jeannine Hardy and Emilie Gesnys for their administrative support.

Moreover, along these years, I had the pleasure to work with very stimulating and friendly teams. I would like to thank Laura Melian-Gutierrez and Ivan Pérez-Álvarez, from IDeTIC, Universidad de Las Palmas de Gran Canaria, Spain, and as well as Sumit Darak, from Department of Electronics and Communication Engineering, IIIT-Delhi, India, for all the discussions and guidance.

I wish to thank especially to my current and previous office mates Quentin Bodinier, Rémi Bonnefoi, Malek Naoues, Marwa Chafii, Vincent Gouldieff, Lilian Besson, Muhammad Abdul Wahab, Rami Othman, Yifan Zhou, Xiguang Wu, Oussama Lazrak, Abir Amri, Vincent Savaux, Eren Unlu, Jerome Manceau, Samba Traore, Lamarano Mamadou Diallo, Hussein Kobeissi and all fellows in CentraleSupélec-Rennes for their help and kindness.

I am thankful to the jury members of my thesis committee. I would like to thank Prof. Visa Koivunen and Prof. Samson Lasaulce, for reviewing this Ph.D. work and providing me with useful comments and suggestions. I would also like to thank Prof. Damien Ernst, Dr. Emilie Kaufmann and my Ph.D. supervisors, for participating to the committee of my PhD defense. I thank them all for their valuable time and suggestions.

Finally, I would like to thank my family for their unconditional support and sacrifices. My mother and father have always been my source of inspiration and they always encouraged me for better education. I thank my wife (Dipal Modi) for her constant support, encouragement and understanding. I thank all my relatives and friends (Hrishikesh Deshpande, Deepak Subramanian, Yogesh Karpate and all my friends) who indirectly contributed in my well-being and this work.

Abstract

Future cellular network technologies are targeted at delivering self-organizable and ultra-high capacity networks, while reducing their energy consumption. This thesis studies intelligent spectrum and topology management through cognitive radio techniques to improve the capacity density and Quality of Service (QoS) as well as to reduce the cooperation overhead and energy consumption. This thesis investigates how reinforcement learning can be used to improve the performance of a cognitive radio system.

In this dissertation, we deal with the problem of opportunistic spectrum access in infrastructure-less cognitive networks. We assume that there is no information exchange between users, and they have no knowledge of channel statistics and other user's actions. This particular problem is designed as multi-user restless Markov multi-armed bandit framework, in which multiple users collect a priori unknown reward by selecting a channel. The main contribution of the dissertation is to propose a learning policy for distributed users, that takes into account not only the availability criterion of a band but also a quality metric linked to the interference power from the neighboring cells experienced on the sensed band. We also prove that the policy, named distributed restless QoS-UCB (RQoS-UCB), achieves at most logarithmic order regret. Moreover, numerical studies show that the performance of the cognitive radio system can be significantly enhanced by utilizing proposed learning policies since the cognitive devices are able to identify the appropriate resources more efficiently.

This dissertation also introduces a reinforcement learning and transfer learning frameworks to improve the energy efficiency (EE) of the heterogeneous cellular network. Specifically, we formulate and solve an energy efficiency maximization problem pertaining to dynamic base stations (BS) switching operation, which is identified as a combinatorial learning problem, with restless Markov multi-armed bandit framework. Furthermore, a dynamic topology management using the previously defined algorithm, RQoS-UCB, is introduced to intelligently control the working modes of BSs, based on traffic load and capacity in multiple cells. Moreover, to cope with initial reward loss and to speed up the learning process, a transfer RQoS-UCB policy, which benefits from the transferred knowledge observed in historical periods, is proposed and provably converges. Then, proposed dynamic BS switching operation is demonstrated to reduce the

number of activated BSs while maintaining an adequate QoS. Extensive numerical simulations demonstrate that the transfer learning significantly reduces the QoS fluctuation during traffic variation, and it also contributes to a performance jump-start and presents significant EE improvement under various practical traffic load profiles.

Finally, a proof-of-concept is developed to verify the performance of proposed learning policies on a real radio environment and real measurement database of HF band. Results show that proposed multi-armed bandit learning policies using dual criterion (e.g. availability and quality) optimization for opportunistic spectrum access is not only superior in terms of spectrum utilization but also energy efficient.

Contents

Acknowledgements v

Abbreviations xiii

Mathematical Notations xv

Résumé en Français xxi

1 Introduction 1
- 1.1 Green Radio Resource Management . . . 1
  - 1.1.1 Spectrum Management . . . 2
  - 1.1.2 Green Wireless Networks . . . 3
- 1.2 Major Contributions and Outline . . . 4

2 On-line Learning for Green Cognitive Radio 9
- 2.1 Introduction . . . 9
- 2.2 Cognitive Radio . . . 10
  - 2.2.1 Definitions . . . 10
  - 2.2.2 Basic Cognitive Cycle . . . 11
- 2.3 Decision Making Process . . . 13
  - 2.3.1 Machine Learning . . . 13
  - 2.3.2 Reinforcement Learning . . . 15
  - 2.3.3 Transfer Learning . . . 17
- 2.4 Multi-armed Bandit – Theory, Variants and Policies . . . 19
  - 2.4.1 Stateless (Classical iid) Bandit Problem Formulation . . . 21
  - 2.4.2 State-dependent (Markovian) Bandits Problem Formulation . . . 22
  - 2.4.3 Stateless Bandits Policies . . . 23
  - 2.4.4 Markovian Bandits Policies . . . 27
- 2.5 On-line Learning for Green Cognitive Radio . . . 30
  - 2.5.1 On-line Learning for Dynamic Spectrum Access . . . 30
  - 2.5.2 On-line Learning for Energy Efficient Wireless Networks . . . 32
- 2.6 Conclusion . . . 34

3 MAB Performance Evaluation Methodologies 35
- 3.1 Introduction . . . 35
- 3.2 Performance Evaluation Methodologies . . . 36
  - 3.2.1 Lempel-Ziv (LZ) Complexity . . . 37

3.2.2 Optimal Arm Identification (OI) factor . . . 38
3.3 Verification of LZ and OI Complexity Impact on OSA Scenario . . . 39
3.3.1 Classical MAB System Model for OSA . . . 39
3.3.2 UCB1 and Thomson-Sampling Policies . . . 40
3.3.3 Numerical Results . . . 41
3.4 Conclusion . . . 46

4 A New Single-player and Multi-player CR Learning Algorithms for OSA 47
4.1 Introduction . . . 48
4.2 Problem Formulation . . . 48
4.2.1 Markovian MAB System Model . . . 48
4.2.2 Wireless Network Model . . . 50
4.3 Novel Single-player and Multi-player Policy Design . . . 54
4.3.1 Single-player Rested Bandit Policy: QoS-UCB . . . 54
4.3.2 Single-player Restless Bandit Policy: RQoS-UCB . . . 56
4.3.3 Multi-player Restless Bandit Policy: Distributed RQoS-UCB . . . 57
4.4 Regret Analysis . . . 59
4.4.1 Preliminaries for Regret Analysis . . . 59
4.4.2 Regret of Single-player Policy . . . 62
4.4.3 Regret of Multi-player Policy . . . 64
4.5 Numerical Results and Analysis . . . 67
4.5.1 Simulation Settings . . . 67
4.5.2 Simulation Results and Discussions . . . 70
4.6 Conclusion . . . 77

5 TRQoS-UCB Policy: Energy Efficient Cellular Network 81
5.1 Introduction . . . 81
5.2 Network Model and Problem Formulation . . . 82
5.2.1 Network Model . . . 82
5.2.2 Problem Formulation . . . 85
5.3 RL Framework for Energy Efficient Network . . . 86
5.3.1 System Model . . . 86
5.3.2 Transfer Learning: Transfer RQoS-UCB Policy . . . 88
5.4 Convergence Analysis: Transfer RQoS-UCB Policy . . . 91
5.5 Numerical Results and Analysis . . . 92
5.5.1 Convergence Analysis . . . 94
5.5.2 Performance under Periodic Traffic Load . . . 96
5.6 Conclusion . . . 100

6 Proof of Concept: Learning for OSA in Real Radio Environment 101
6.1 Introduction . . . 101
6.2 Validation of Learning on Real HF Measurements . . . 102
6.2.1 Experimental Results and Analysis . . . 102
6.3 Demonstrators: OSA on Real Radio Spectrum . . . 106
6.3.1 USRP Testbed . . . 107
6.3.2 Experimental Results and Analysis . . . 110
6.4 Conclusion . . . 114

7 General Conclusions and Perspectives **117**
7.1 Conclusions and Overview . . . 117
7.2 Perspectives and Future Work . . . 119

A List of Publications **121**
A.1 Patent . . . 121
A.2 Journal Papers . . . 121
A.3 International Conference Papers . . . 122
A.4 National Conference Papers . . . 123
A.5 Demonstration . . . 123

B Efficient Learning in Non-stationary Scenario **125**
B.1 Introduction . . . 125
B.2 Non-stationary Markov MAB Problem Formulation . . . 126
B.2.1 Non-stationary Regret . . . 127
B.2.2 Discounted QoS-UCB (DQoS-UCB) Policy . . . 128
B.3 Simulation Results . . . 129
B.3.1 Non-stationary Environment . . . 130
B.4 Conclusion . . . 132

C Proof of Theorems and Lemmas **135**
C.1 Proof of Theorem 4.1 . . . 135
C.2 Proof of Theorem 4.2 . . . 141
C.3 Proof of Theorem 4.3 . . . 142
C.4 Proof of Lemma 4.6 . . . 143
C.5 Proof of Theorem 4.4 . . . 144

List of Figures **148**

List of Tables **153**

Bibliography **155**

Abbreviations

LTE	Long Term Evolution
WLAN	Wireless Local Area Network
ITU	International Telecommunication Union
ICT	Information and Communication Technology
IoT	Internet of Things
SDR	Software Defined Radio
CR	Cognitive Radio
CC	Cognitive Cycle
ETSI	European Telecommunications Standards Institute
FCC	Federal Communications Commission
AWGN	Additive White Gaussian Noise
IETF	Internet Engineering Task Force
QoS	Quality of Service
BER	Bit Error Rate
SINR	Signal to Interference Noise Ratio
EE	Energy Efficiency
SE	Spectral Efficiency
Tx	Transmitter
Rx	Receiver
REM	Radio Environment Mapping
WRAN	Wireless Regional Area Network
HCN	Heterogeneous Cellular Network
TVWS	TV WhiteSpace
DSA	Dynamic Spectrum Access
OSA	Opportunistic Spectrum Access
PU	Primary User
SU	Secondary User
BS	Base Station
MS	Mobile Station
ML	Machine Learning

RL	Reinforcement Learning
MPRL	Multi-Player Reinforcement Learning
MAB	Multi-armed Bandit
IID	Independent Identical Distributed
UCB	Upper Confidence Bound
TS	Thomson-Sampling
QoS-UCB	Quality of Service Upper Confidence Bound
RQoS-UCB	Restless Quality of Service Upper Confidence Bound
TL	Transfer Learning
TRQoS-UCB	Transfer Restless Quality of Service Upper Confidence Bound
D-UCB	Discounted Upper Confidence Bound
TDFS	Time Division Fare Share
RUCB	Restless Upper Confidence Bound
RCA	Regenerative Cycle Algorithm
DSEE	Deterministic Sequencing of Exploration and Exploitation
ACT	Actor CriTic
TACT	Transfer Actor CriTic
LZ	Lempel-Ziv complexity
OI	Optimal Arm Identification factor
RCS	Random Channel Selection
R2PO	Round Robin Perfect Order
ED	Energy Detector
SB	Sub-Block
BLOS	Beyond Line-Of-Sight
CSC	Channel Switching Cost
QPSK	Quadrature Phase Shift Keying
RF	Radio Frequency
USRP	Universal Software Radio Peripheral
GRC	GNU Radio Companion

Mathematical Notations

$\mathcal{A}$	Learning policy
$\mathbb{1}_{(\cdot)}$	Indicator function which returns 1 if statement inside $(\cdot)$ is true or 0
i	Arm index
j	User index
n	Observation time step
$\mathcal{A}(n,j)$	Arm which has to be played in the next time slot by user j
$\mathcal{K}$	Set of arms, $\mathcal{K} = \{1, \cdots, K\}$
$\mathcal{U}$	Set of users, $\mathcal{U} = \{1, \cdots, U\}$
K	Number of arms (or available actions)
U	Number of secondary users (SU)
$\{1, \cdots, U\}$	Set of optimal arms out of K arms
$\{U+1, \cdots, K\}$	Set of suboptimal arms
P^i	Transition matrix of the Markov chain modeling arm i
S^i	State space of the Markov chain modeling arm i
$q, q \in S^i$	Observed states
$T^{i,j}(n)$	total number of times arm i has been played by user j up to time n
$T^i(n)$	$\sum_{j=1}^{U} T^{i,j}(n)$, total number of times arm i has been played up to time n
π_q^i	Stationary distribution for state q of the Markov chain of arm i
$S^{i,j}(n)$	Observed state of arm i at time n for user j at time n
$r_q^{i,j}(n)$	Reward achieved in state $q \in S^i$ from a arm i by user j at time n
$R_q^{i,j}(n)$	Instantaneous observed quality of arm i by user j at time n
μ_i^R	$\sum_{q \in S^i} G_q^i r_q^i \pi_q^i$, global mean reward
$\Delta\mu_i^R$	$\mu_1^R - \mu_i^R$
$\Delta\mu_{i,j}^R$	$\mu_j^R - \mu_i^R$
$C_o^j(i,n)$	Indicator of collision at the $n-$th slot at channel i for SU j
$\Phi^{\mathcal{A}}(n)$	Regret which is defined as the reward loss with selected policy compared to ideal policy
α	Exploration coefficient with respect to availability criteria

β	Exploration coefficient with respect to quality criteria
$B^{i,j}(n, T^{i,j}(n))$	Learning policy index for arm i at player j at time n
$\bar{S}^i(T^i(n))$	Empirical mean of the states observation of the $i-$th channel at time n
$A^i(n, T^i(n))$	Bias term, defined same as classical UCB policy
$G_q^{i,j}(T^{i,j}(n))$	$\frac{1}{T^{i,j}(n)} \sum_{k=1}^{T^{i,j}(n)} R_q^{i,j}(k)$, the empirical mean of quality observations $R_q^{i,j}$
$G_{\max}^{q,j}(n)$	$\max_{i \in \mathcal{K}} G_q^{i,j}(T^{i,j}(n))$, maximum expected quality in state q from K channels
$G_{\max}^{j}$	$\max_{q \in S^i} G_{\max}^{q,j}$
$M^{i,j}(n, T^{i,j}(n))$	$G_{\max}^{q_1,j}(n) - G_{q_1}^{i,j}(T^{i,j}(n))$
$\zeta^{i,j}$	Regenerative state that determines regenerative cycles for arm i and user j
n_2	Total number of time slots spend in SB2 block.
$T_2^{i,j}(n_2)$	Total number of time arm i is played by user j during SB2 block up to n_2 time.
W	Frame size for multi-user distributed RQoS-UCB policy
$Rank(j)$	$Rank-$th highest entry in $B^{i,j}(n, T^i(n)), \forall i \in \mathcal{K}$ for user j
$S_{\max}$	$\max_{i \in \mathcal{K}} \lvert S^i \rvert$, where $\left\lvert S^i \right\rvert$ stands for the cardinality of the state space of arm i
$\left(P^i\right)'$	$\left\{\pi_l^i p_{lk}^i / \pi_k^i\right\}, \forall k, l \in \mathcal{S}$, being the adjoint of P^i
$\tilde{P}^i$	$\left(P^i\right)' P^i$ as the multiplicative symmetrization of P^i
$\pi_{\min}^i$	$\min_{q \in S^i} \pi_q^i$
$\pi_{\min}$	$\min_{i \in \mathcal{K}} \pi_{\min}^i$
$\hat{\pi}_q^i$	$\max\left\{\pi_q^i, 1 - \pi_q^i\right\}$
$\hat{\pi}_{\max}$	$\max_{q \in S^i, i \in \mathcal{K}} \hat{\pi}_q^i$
$r_{\max}$	$\max_{q \in S^i, i \in \mathcal{K}} r_q^i$
$M_{\min}^j$	$\min_{i \in \mathbb{K}} M^{i,j}\left(n, T^{i,j}(n)\right)$
$M_{\max}^j$	$\max_{i \in \mathbb{K}} M^{i,j}\left(n, T^{i,j}(n)\right)$
λ_2	Second largest eigenvalue of the matrix $\tilde{P}$
γ^i	Eigenvalue gap of the $i-$th arm
$\gamma_{\min}$	$\min_{i \in \mathcal{K}} \gamma^i$
$\gamma_{\max}$	$\max_{i \in \mathcal{K}} \gamma^i$
$\Omega_{k,l}^i$	Mean hitting time of state l starting from an initial state k for the ith arm.
$\Omega_{\max}^i$	$\max_{k,l \in S^i, k \neq l} \Omega_{k,l}^i$
$\Omega_{\max}$	$\max_{i \in \{1, \cdots, K\}} \Omega_{\max}^i$
$\mathbf{h}$	Non zero initial distribution
$N_{\mathbf{h}}$	$l_2(\pi)$-norm of the initial distribution $\mathbf{h}$
τ	Stopping time of a Markov chain S

$C_{\mathbf{S},\mathbf{P},\mathbf{G},\mathbf{r}}$	Constant that depends on S^i, P^i, $\mathbf{r}_q^i$ and $\mathbf{G}_q^i$
$b(n)$	Total number of completed blocks (consisting SB1, SB2 and SB3) up to current time n
$f(n)$	Total number of frames with fixed frame size W up to time n
$F^{i,j}(b(n))$	Total number of block in which arm i is played by SU j up to last block b at time n
$C_o(n)$	Total number of collisions up to time n in the U optimal channels
Υ	Maximum time required to reach to the absorbing state starting from any initial distribution
$f_w(n)$	Number of frame where any one of the U optimal arm's estimated ranks is wrong
$T_w(n)$	Number of time steps where any one of the U optimal channel's estimated ranks is wrong
n^b	Time at the end of the last completed block $b(n)$
$V^{i,j}(n)$	Total number of times where an SU j is the only one to sense and access the channel i up to time n
$S_1^{i,j}(k)$	vector of observed states from SB1 of $k-$th block in which band i is sensed by SU j
$S_2^{i,j}(k)$	vector of observed states from SB2 of $k-$th block in which band i is sensed by SU j
$S^{i,j}(k)$	vector of observed states from $k-$th block $S^i(j) = \left[S_1^i(j), S_2^i(j), \zeta^i\right]$
$\bar{b}^{i,j}$	Total number of joined blocks up to current block b for optimal band i for SU j
$\bar{X}(j)$	j-th combined block in which the optimal band is sensed, i.e. $\bar{X}(j) = \left[\bar{X}_1(j), \bar{X}_2(j), \zeta^i\right]$
$\bar{X}_1(j)$	SB1 block in j-th combined block ($\bar{X}(j)$) in which the optimal band is sensed
$\bar{X}_2(j)$	SB2 block in j-th combined block ($\bar{X}(j)$) in which the optimal band is sensed
λ	Discount factor for calculation of empirical mean
N_{change}	Number of change points in non-stationary reward distribution
T_{total}	Total transmission frame size for slotted CR network
T_{sens}	SU Sensing time out of T_{total}
T_{lear}	SU learning time out of T_{total}
T_{trans}	SU transmits out of T_{total}
$\mathcal{H}_0$	Hypothesis when channel is used by PU
$\mathcal{H}_1$	Hypothesis when channel is vacant
$y_q^i[m]$	Received signal at SU Rx when PU Tx transmits
$p^i[m]$	Zero mean i.i.d. PU signal component

$u^i[m]$	Zero mean and complex Gaussian distributed noise components
$z_q^i[m]$	Received signal at SU Rx when SU Tx transmits
$s^i[m]$	SU signal, zero mean and iid with variance $\mathbb{E}[\|s^i[m]\|^2] = \sigma_{s,i}^2$
P_{md}	Probability of miss detection
P_f	Probability of false alarm
Γ_0^i	$\frac{\sigma_{s,i}^2}{\sigma_{p,i}^2+\sigma_{u,i}^2}$, SNR under hypothesis $\mathcal{H}_0$
Γ_1^i	$\frac{\sigma_{s,i}^2}{\sigma_{u,i}^2}$, SNR under hypothesis $\mathcal{H}_1$
υ	Energy detector threshold above which the channel is decided to be occupied
F_s	Sampling frequency at the SU receiver
N_s	Number of samples during the sensing phase, where $N_s = T_{\text{sens}}F_s$
$\mathcal{P}_q^i$	Power measure of the $i-$th channel
R_q^i	Quality information metric (or reward associated with quality)
c_e	Constant added to the received power from energy detector
Ξ^i	SU's average total achievable throughput in the $i-$th channel
$h(S)$	Normalized LZ complexity, reflects the arising rate of new patterns in observations
H_{OI}	Optimal arm identification factor
P_{free}	Probability that there exists at least one free channel from a set of K arm.
$W^i(n, T^i(n))$	Total number of free state observed up to time n from arm i
$Z^i(n, T^i(n))$	Total number of occupied state observed up to time n from arm i
P_{Succ}	Number of times each vacant channel is explored over the number of iterations
$\mathbb{R}^2$	Two dimensional area covering all BSs and MSs
k	Index of BS
$\mathcal{Y}$	Set of BSs $\mathcal{Y} = \{1, \cdots, Y\}$
$\mathcal{Y}_n^{on}$	Set of active (ON) BSs at $n-$th iteration
$\mathcal{I}_k^n$	Cell coverage of BS k at time n
x^k	Two-dimensional Cartesian coordinates, denoting the locations of the MS in coverage $\mathcal{I}_k^n$ of $k-$th BS at time n
x^{on}	Two-dimensional Cartesian coordinates, denoting the locations of active MS in two dimensional area in $\mathbb{R}^2$
$\Lambda(x^k, n)$	Traffic arrival rate at a location x^k in BS k following a Poisson point process at $n-$th iteration
$L(x^k)$	Instantaneous traffic load at location x^k
$L^k(n)$	Instantaneous traffic load served by the BS $k \in \mathcal{Y}_n^{on}$
$f_{\text{Traffic}}(k)$	Traffic load profile function

P_e	Period of a traffic load in daily traffic profile
V_a	Variance in daily traffic profile
M_e	Mean arrival rate in daily traffic profile
$\Theta_j(x^{on}, n)$	Service rate at location x^{on} from BS k at $n-$th iteration
$g^n(k, x_k^{on})$	Average channel gain from BS k to active MS at location x_k^{on} at $n-$th time slot
Ba	Denotes the system bandwidth
ϕ	Orthogonality (or self interference) factor
$\text{SINR}_k(x^{on}, n)$	Received signal to interference plus noise ratio (SINR) at active MS location x^{on} from BS k at $n-$th iteration
$\rho_k(n)$	System load of BS k at $n-$th iteration
ρ^{th}	System load threshold
P_k^{tx}	Transmission power of BS k
P_f^k	Fixed operational power of BS k
P_T^k	Total operational power of BS k
a_k	BS power scaling factor
$EE(n)$	Network energy efficiency (EE) in bits per joule
$\Theta^{\min}$	Prescribed minimum data rate to continue data transmission
$1/h(x^j, n)$	Average call duration (or file size) at $n-$th iteration
h_2	Total number of iterations in historic period
H_2^i	Total number of iterations spend in arm i in SB2 block in historic period
$S^{i,h}(n)$	State observed by action i at $n-$th iteration in the historical period
$R_S^{i,h}(n)$	Immediate reward with action i at $n-$th iteration in the historical period
$B^{i,h}(n_2, T_2^i(n_2))$	Learning policy index for arm i taking into an account historic observations at time n

Résumé en Français

La densité du trafic des communications sans fil n'a cessé d'augmenter depuis 2 décennies. Les futures technologies de réseaux cellulaires visent à supporter un trafic toujours plus élevé, grâce à des réseaux auto-organisés offrant une meilleure capacité tout en réduisant la consommation d'énergie. Cependant, la capacité des réseaux va particulièrement être contrainte à courte échéance par la disponibilité du spectre, en raison de la demande conjointe en débit et qualité de service (QoS - Quality of Service) pour un nombre d'utilisateurs toujours plus grand [30]. En outre, il est désormais établi que la question de l'efficacité énergétique va devenir primordiale et qu'il va en résulter un maillage de plus en plus dense de stations de bases (hétérogènes souvent). Par ailleurs, de nouveaux réseaux sans fil vont connaître un développement extraordinaire, tels que ceux qui concerneront l'Internet des Objets (IoT – Internet of Things). On prédit pour l'IoT des milliards d'objets aux capacités radio très différentes, avec des schémas d'accès divers (accès plus ou moins libre, QoS plus ou moins importante) qui vont encore plus densifier l'utilisation du spectre radio fréquence. Ainsi on pressent que doivent émerger de nouvelles solutions d'accès au spectre, plus intelligentes que celles utilisées depuis 100 ans que la radio existe commercialement [39]. La radio intelligente (CR – Cognitive Radio) [127], un nouveau concept apparu à la toute fin du 20ème siècle, étudie et propose des solutions visant à insérer de l'intelligence dans les réseaux et équipements de communications sans fil, afin notamment de répondre à ces défis de rareté spectrale et d'efficacité énergétique qui sont au cœur de cette thèse. En combinant les facultés d'intelligence et de flexibilité, la radio intelligente ouvre la voie à l'auto-adaptation des systèmes de communications sans fil aux changements de leur environnement. Quand il s'agit d'améliorer l'efficacité énergétique et de mieux utiliser le spectre, on parle de radio verte, ou d'éco-radio (Green Radio).

Les principaux axes d'action de la radio verte peuvent donc être : la gestion du spectre et les réseaux sans fil économes en énergie.

Gestion du spectre

L'utilisation efficace du spectre radio-fréquence est un problème fondamental des communications sans fil. Les relevés statistiques de l'usage fréquentiel et temporel des fréquences présentés

dans [39, 127] établissent que le spectre n'est tout d'abord pas complètement utilisé à tout instant quand il est assigné, et que cela peut fortement dépendre du lieu. D'après la Federal Communications Commission (FCC), cela concerne entre 15% et 85% du spectre alloué [38]. Qu'une telle proportion du spectre soit sous-utilisée alors que la demande spectrale est toujours plus forte, appuie les arguments des tenants d'une utilisation plus flexible des ressources spectrales, basée sur la radio intelligente.

Les technologies d'accès dynamique au spectre (DSA - Dynamic Spectrum Access) visent à répondre à cette question [162]. L'accès opportuniste au spectre (OSA – Opportunistic Spectrum Access), en particulier, est un cas de DSA pour lequel un utilisateur secondaire (SU) exécute des techniques de détection de présence d'un signal (sensing) et de prise de décision pour accéder aux ressources spectrales quand elles sont laissées vacantes par les utilisateurs primaires (PUs) [53, 67]. Les SU sont donc dotés de capacités de radio intelligente. Le DSA permet ainsi à ces SUs de s'adapter au trafic fluctuant des PUs, afin de profiter des opportunités laissées par ceux-ci, et au final de mieux utiliser les ressources spectrales. Autrement dit, les SUs peuvent combler les « trous » laissés par les PUs. Plus les SUs sont capables d'anticiper les opportunités grâce à l'apprentissage, meilleure est l'efficacité spectrale au niveau global, et meilleure est l'efficacité énergétique des SUs qui limitent leurs tentatives de transmissions et leurs risques de collision entre eux.

En OSA, il est primordial que les SUs n'interfèrent pas avec les PUs qui doivent garder la même QoS que s'il n'y avait pas de SUs (conditions d'acceptation par les possesseurs de bandes licenciées de ce genre d'approche), mais il est aussi important que les SUs aient également une certaine QoS, qui prise au sens large peut consister à ne pas seulement rechercher à utiliser des canaux vacants, mais aussi dans des conditions offrant une certaine qualité. La qualité peut aussi bien concerner le taux d'interférence dans un canal, que la consommation d'énergie que requiert une transmission dans ce canal pour un taux d'erreur donné. Ce sera l'un des objectifs de cette thèse de prendre en considération la qualité dans l'apprentissage de systèmes radio intelligents.

Les réseaux sans fil économes en énergie

De la croissance permanente de la demande en communications découle une augmentation régulière des émissions de CO2, issue de la consommation électrique des points d'accès radio, des routeurs du réseau, ainsi que des centres de calcul et de stockage. Ils constituent les principaux consommateurs d'énergie de l'industrie des technologies de l'information et des communications (TIC) qui représentait il y a encore peu de temps 2%, et bientôt 10%, de la consommation énergétique mondiale [101]. Dans les réseaux mobiles, les stations de base représentent 60 à 80% de la consommation totale [42] et les opérateurs doivent régler une facture annuelle de

plus de 10 milliards de dollars en consommation électrique [130, 141]. Il y a par conséquent de forts intérêts économiques et écologiques à prendre en considération l'efficacité énergétique dans les réseaux de communication sans fil. Il est important de constater que ces réseaux sont actuellement dimensionnés pour fonctionner en permanence au pire cas correspondant aux pics de trafic auquel ils ne doivent faire face que de temps en temps. Ainsi en raison des fluctuations du trafic dans le temps et du déplacement des utilisateurs, le réseau est par moments surdimensionné par rapport à la demande instantanée, et par conséquent sa consommation ne diminue pas pour autant en proportion du nombre d'utilisateurs connectés. C'est tout l'enjeu du projet TEPN (Towards Energy Proprtionnal Networks) du Labex Cominlabs dans lequel se situe cette étude. En outre, on remarque qu'il serait possible par moment de mettre en veille certaines stations de base, dont le ratio entre la puissance radiative effectivement émise par son antenne et la puissance totale consommée de la station de base (incluant tous les traitements numériques mais aussi l'air conditionné du local) n'atteint péniblement que 3% [74]. Ainsi certaines études ont montré que les stations de base sont très souvent en sous-charge, et que pendant 30% du temps en semaine (45% les week-ends), la charge est inférieure à 10% de la charge maximale que la station de base peut supporter, en termes de trafic et de nombre d'utilisateurs [122]. Durant ces périodes de faible charge il a été montré [141] qu'il est particulièrement économe d'éteindre ou mettre en veille certaines des stations de base et de déporter le faible trafic qu'elles devraient supporter sur leurs voisines, elles-mêmes en sous charge.

C'est le but de la planification dynamique des réseaux, qui vise à contrôler le nombre de stations actives en fonction du trafic. Le projet FP7 EARTH illustre des cas d'utilisation pour le LTE [44]. Un enjeu important dans ce cas, et c'est ce que propose d'étudier les présents travaux, est de définir des stratégies de prise de décision, basées sur de l'apprentissage, pour contrôler le nombre de stations de base qui doivent être laissées en fonctionnement afin de maintenir un service adéquat (avec une certaine QoS).

Plan de la thèse et contributions

Nous montrons dans ce manuscrit que l'apprentissage et la prise de décision sont matures pour un déploiement réel, en termes de rapidité de convergence, de complexité de mise en œuvre et de performance, à la fois dans le cas d'un terminal seul et dans le cas de nombreux terminaux en réseau. Nous considérons les deux cas de l'efficacité de l'utilisation des ressources spectrales et de l'efficacité énergétique.

Après un chapitre introductif, le **chapitre 2** du manuscrit effectue un état de l'art des travaux relatifs au sujet de notre étude. Dans un premier temps une analyse de la littérature de la prise de décision pour la radio intelligente dresse la liste des contraintes associées au problème posé, et identifie les solutions potentielles d'apprentissage machine, notamment l'apprentissage

par renforcement. L'analyse a permis de déduire que le modèle des bandits manchots (MAB – Multi-Armed Bandit) et les algorithmes qui y répondent sont adaptés à la prise de décision pour la gestion du spectre et l'efficacité énergétique des réseaux sans fil. Les principales contributions de cette thèse se trouvent dans les chapitres 3 à 6.

Le **chapitre 3** introduit un nouveau critère de caractérisation de l'efficacité potentielle de l'apprentissage, le facteur OI (Optimal Identification). Ce facteur va être utilisé ensuite pour évaluer les performances de l'apprentissage de manière plus objective. Il va notamment permettre de pouvoir mieux comparer les performances apportées par l'apprentissage par rapport à un comportement aléatoire, mais aussi de caractériser un scénario en termes d'honnêteté : est-ce un scenario facile ou difficile pour l'apprentissage?

Le **chapitre 4** introduit un nouvel algorithme d'apprentissage dans le cadre de l'accès opportuniste au spectre (mais qui peut s'appliquer au-delà), qui combine deux critères : la disponibilité des canaux fréquentiels et un critère de qualité (par exemple le rapport signal à bruit sur interférence des canaux). Cet algorithme est basé sur l'extension des précédents travaux ne prenant en compte que la disponibilité des canaux avec des algorithmes de type UCB (Upper Confidence Bound) [67] avec une approche de modélisation du problème MAB par chaines de Markov, dans un cas « restless » : d'où son nom « restless Quality of Service UCB » ou RQoS-UCB. Dans ce modèle, deux récompenses sont prises en compte (disponibilité et qualité) et combinées pour hiérarchiser les solutions entre elles et permettre à un SU d'optimiser la sélection des canaux dans un cas OSA (et tout autre problème répondant au même modèle). La preuve analytique de convergence, c'est-à-dire une borne supérieure de la récompense d'ordre logarithmique dans les cas de chaines de Markov « rested » et « restless », est donnée tout d'abord dans le cas d'un seul joueur, pour valider l'approche, puis dans le cas de multi-joueurs non coordonnés. Il est montré par des simulations que l'algorithme RQoS-UCB proposé permet l'optimisation de l'exploitation d'opportunités de transmission minimisant les collisions entre des SUs alors que ceux-ci ne sont pas coordonnés, c'est-à-dire qu'ils n'échangent pas d'information entre eux (ce qui consommerait de manière pénalisante une partie significative de la bande passante qu'ils essaient justement de trouver de manière opportuniste). **Les principales contributions du chapitre 4 sont** :

- La modélisation du scenario OSA sous la forme de problèmes MAB markoviens « rested » et « restless » prenant en compte deux critères (qualité et disponibilité) pour faire l'apprentissage des opportunités de transmission parmi plusieurs canaux, et permettant de sélectionner à chaque instant le meilleur canal.
- Un nouvel algorithme RQoS-UCB mono-joueur et la démonstration mathématique que la récompense a une borne supérieure d'ordre logarithmique dans le cas markovien « restless ».

- L'extension de l'utilisation de l'algorithme RQoS-UCB au cas multi-joueurs et sa capacité à permettre aux joueurs d'éviter les collisions de manière non coordonnée.
- Une validation par des simulations (et dans un prochain chapitre par des démonstrations sur signaux radio réels) de nombreux scenari qui valident l'approche proposée et montrent qu'elle surpasse les solutions de l'état de l'art.

Le **chapitre 5** vise à utiliser l'algorithme RQoS-UCB dans un autre contexte que l'OSA, celui de l'efficacité énergétique des réseaux, avec des stations de base (homogènes ou hétérogènes) pouvant être mises en veille. Ce problème est en phase avec le projet TEPN (Towards Energy Proprtionnal Networks). Le problème de commutation dynamique de l'état (allumé, veille) des stations de base est modélisé par une approche de type MAB markovien. Les performances de l'algorithme RQoS-UCB sont évaluées puis une évolution basée sur le transfert de connaissance (TL - Transfer Learning) est introduite pour le cas des problèmes MAB : TRQoS-UCB. L'analyse de la preuve de convergence de la solution proposée est très similaire à celle de l'algorithme RQoS-UCB et des résultats de simulation mettent en valeur les gains obtenus en termes d'efficacité énergétique. **Les principales contributions du chapitre 5 sont** :

- La modélisation du problème de commutation des stations de base en un problème MAB markovien et l'utilisation de l'algorithme RQoS-UCB pour le résoudre.
- L'utilisation du principe de transfert d'apprentissage pour initialiser les algorithmes de prise de décision du problème MAB.
- L'évaluation de ces principes par des simulations pour l'efficacité énergétique des réseaux sans fil.

Dans le **chapitre 6** sont présentées les preuves de concept réalisées lors de ces travaux de recherche afin de démontrer la faisabilité et la pertinence des approches proposées. Tout d'abord, l'approche OSA a été appliquée au canal de transmission HF (Hautes Fréquences) qui est un canal de communication transhorizon utilisé par les radio amateurs et les militaires pour communiquer à l'échelle du globe, en profitant de phénomènes de propagation particuliers intervenant dans la gamme des fréquences HF (3 MHZ-30 MHz). Comme ces communications sont naturellement trans-frontalières, de nombreuses collisions peuvent intervenir entre les utilisateurs et il n'existe pas à l'heure actuelle de solution de coordination pour les mitiger. Des algorithmes d'apprentissage UCB ont été évalué sur des bases de données de mesures réelles issues de canaux HF, effectuées par l'Université de Las Palmas de Gran Canaria, et les résultats valident la solution proposée. Ensuite des preuves de concept du cas d'accès opportuniste au spectre ont été menées sur des signaux radio réels émis en laboratoire. Plusieurs algorithmes de bandits (UCB, Thomson Sampling, KL-UCB, QoS-UCB) sont comparés en temps-réel, dans

différentes conditions de comportement du réseau primaire (i.i.d., markovien, pour différents schémas d'occupation). Le cas mono-joueur permet de bien comprendre le gain apporté par l'apprentissage, d'évaluer sa pertinence en termes d'efficacité (pourcentage de succès de transmission opportuniste) et de vitesse de convergence et de comparer les différents algorithmes entre eux, en termes de coût de commutation. Ensuite, les démonstrations multi-joueurs permettent là encore de valider la pertinence des propositions faites dans cette thèse, notamment en mesurant le taux de collisions entre utilisateurs secondaires qui sans se coordonner (aucun échange de message) arrivent à se répartir sur les canaux de manière avantageuse pour l'ensemble des utilisateurs.

Enfin le **chapitre 7** conclut cette thèse et propose des perspectives aux travaux de recherche qui y ont été menés. Ces travaux ont abouti à la production de 3 revues internationales dont deux IEEE Transactions on Cognitive Communications and Networking, un brevet, 6 conférences internationales et deux conférences nationales.

Chapter 1

Introduction

Contents

1.1 Green Radio Resource Management 1
1.1.1 Spectrum Management 2
1.1.2 Green Wireless Networks 3
1.2 Major Contributions and Outline 4

1.1 Green Radio Resource Management

Traffic density in wireless communication systems keeps growing significantly. Future cellular network technologies are still targeted at delivering self-organizable and ultra high capacity networks, which will support an increasing number of mobile subscribers, while reducing their energy consumption. Network capacity will be heavily constrained by spectrum availability in the near future, because of the high throughput and Quality of Service (QoS) requirement from a growing number of users [30]. Moreover, it can be anticipated that the energy issue will become even more serious in near decade, because of large number of BSs are densely deployed. Furthermore, future Internet of Things (IoT) networks are also expected to be used by thousands of devices with various wireless abilities, and hence collision-free spectrum access and QoS requirement from a growing number of IoT devices also requisite more intelligent spectrum allocation solution. Cognitive Radio (CR), a new paradigm of wireless communication, has been considered as a potential way to meet the dense network demand while achieving energy efficiency. By combining the abilities of intelligence and radio flexibility, CR will be able to adapt itself to the changes in the local environment.

The major issues with respect to Green Radio Resource Management can be categorized as follows:

1.1.1 Spectrum Management

Efficient utilization of the physical radio spectrum is a fundamental issue of wireless communications. Spatial and temporal spectrum usage statistics presented in [39, 127] states that the spectrum is not fully utilized and the usage depends on location. According to Federal Communications Commission (FCC), 15% to 85% of the assigned spectrum is utilized with large temporal and geographical variations [38]. Meanwhile, the wireless communication requirements have increased both in number of users and quality of service. The conflict between the inefficient usage of spectrum and the rapid growth of wireless devices advocates for a more flexible management of the spectrum resources. It is foreseen that a large amount of underutilized spectrum will be efficiently used by applying cognitive radio techniques.

Radio Environment Mapping (REM)

Most of the current channel assignment schemes are largely based on fixed channel assignment, however it requires radio environment mapping (REM) in order to limit the interference. REM maintains a dynamic database which contains the spectrum activity, information of BS locations and/or operational parameters [163].

Recently, the FP7 FARAMIR project ([45]), FP7 ABSOLUTE project ([43]), IEEE 802.22 Wireless Regional Area Network (WRAN) standard ([62]), IEEE 802.19 standard ([61]) and IETF PAWS protocols ([63]) have done a comprehensive work on implementation of a radio environment mapping (REM) based spectrum management.

This scheme effectively controls interference to primary users (PU), however, signaling or information exchange between distributed BSs/MSs and central server (containing database) could be excessively high, and storage of such huge databases remains an issue. Thus, as an alternative method, researchers focus on dynamic spectrum access (DSA) technologies which utilize spectrum sensing and decision making techniques which sense and learn the spectrum and use it for communication when available.

Dynamic Spectrum Access

Dynamic Spectrum Access (DSA) was introduced as a better solution to serve the fluctuating traffic demand. In a DSA scheme, instead of having a fixed radio environment mapping, all channels are placed into a channel pool and potentially available to all local users [162]. Hence, DSA concept is proposed to assign channels for opportunistic and occasional access. In this manner, the spectrum is expected to be well utilized, means spectral and energy efficiencies improvement.

Opportunistic Spectrum Access (OSA), a subset of DSA, is conventionally designed to allow opportunistic unlicensed users, referred as "secondary users (SUs)" to access the spectrum occupied by licensed users, referred as "PU" [53, 67]. The only constraint for OSA implementation is to have a reliable QoS and prioritized access for PUs (no interference from SUs). SUs have to identify the holes in the PUs spectrum, with the help of sensing and decision making approach, and then continue transmissions in a specific channel. However, SUs should stop their transmissions and release the channel when requested by PU. On the other hand, best effort QoS should also be guaranteed for SUs in order to reduce their transmit power consumption, and also it could attract unlicensed users to bear extra computational power and time associated with spectrum sensing and decision making process.

OSA paradigm has been widely explored by applying distributed reinforcement learning algorithms. However, a serious drawback of these algorithms is that a number of immature decisions should be carried out prior to achieve an acceptable solution, which cannot guarantee QoS. Most of these algorithms learn to find optimal channel following a joint optimization of both availability and QoS criterion. The issue of delivering a QoS as well as reducing the cooperation overhead in a dynamic radio network are the main research topics of this thesis.

1.1.2 Green Wireless Networks

The increasing popularity of portable smart devices has flared up increasing the traffic demand for radio access network and implying massive energy consumption, which leads in the exhaustion of energy resources and causes potential increase in the CO2 emissions. Data centers, back-haul routers and cellular access networks are the main source of energy consumption in information and communication technology (ICT) industry, which is equivalent of 2% to 10% of the global power consumption [101]. In cellular networks, the energy consumption of base stations (BSs) is about 60% to 80% of the overall power consumption of the cellular network [42]. Besides, cellular network operators require to spend more than 10 billion dollars to meet current energy consumption of the cellular network [130, 141]. Thus there exists a high economical pressure for cellular network operators to take into account an energy efficiency aspect of the communication networks. The main reason behind a high energy consumption of wireless networks is because they have been designed for peak traffic load.

In fact, due to the traffic load variation in time domain and the motion of users among cells, there are opportunities for some BSs to be put from time to time in sleep mode in order to achieve higher energy efficiency. The BS's static components, i.e. baseband signal processor, controller, air-conditioner, etc., are the main sources of energy consumption compared to effectively radiated transmit power which consumes about 3% of the BS power consumption [74]. Recent studies on the temporal traffic, as shown in Fig. 1.1, have stated that BSs are largely

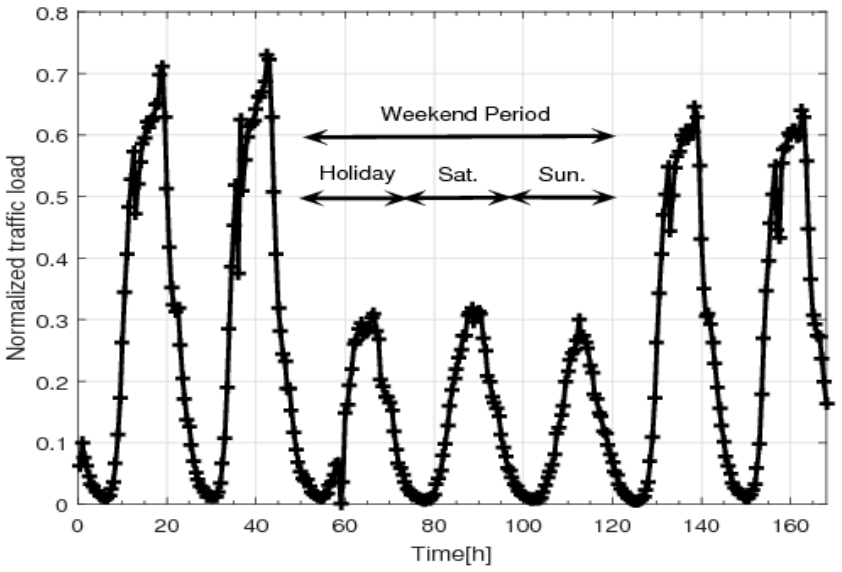


FIGURE 1.1: Normalized real traffic load during one week that are recorded by cellular operator. The data captures voice call information over one week with a resolution of one second in a metropolitan urban area, and are averaged over 30 minute time-scale [123].

underutilized. The time duration when the traffic load is below 10% of the peak load during the day is about 30% and 45% on weekdays and weekends, respectively [122]. Thus, instead of just turning off radio transceivers, the BSs operators may prefer to turn off the underutilized BSs and transfer the imposed traffic loads to the neighbor active BSs during low traffic periods in order to reduces the energy consumption [141].

Dynamic network planning aims at effectively controlling the number of active base stations according to traffic variations. An important amount of energy can be saved by only activating a minimum number of base stations that provides sufficient system capacity. Energy models of various types of LTE BS have been studied by the FP7 EARTH project in [44], where the dynamic strategy for switching a BS to sleep mode has been left for further research. However, there are very few works to achieve dynamic network planning with the help of reinforcement learning approaches, and almost negligible with the help of multi-armed bandit approaches. An important task of this thesis is to design a learning and decision making strategy following multi-armed bandit framework which controls the number of activated BSs based on the users traffic, while maintaining an adequate QoS.

1.2 Major Contributions and Outline

In this Section, we present the structure of the thesis and the contribution of each chapter. We will prove in the manuscript that the presented learning and decision making approach is mature for real deployment in terms of efficiency, convergence speed and implementation complexity

for the benefit of a single device or a network. We will particularly apply it here in a green perspective, e.g. energy efficiency improvement of cellular network or a better use of spectrum resources.

Chapter 2 provides a literature review on the background and established work related to this thesis. First, we focus on literature that aims at defining Cognitive Decision Making cycle. Next, classification of machine learning and reinforcement learning models is presented which tackle the CR decision making problem. Then, state-of-the-art contributions addressing the multi-armed bandit (MAB) framework is introduced. We continue our literature review with the discussion of existing MAB variants and available solution for it. Finally, the state-of-the-art decision making algorithms is discussed with their application to the energy-efficient spectrum management, and also to the energy efficiency optimization of wireless networks.

The main contributions of this thesis range from Chapter 3 to Chapter 6. **Chapter 3** introduces a new criteria, optimal arm identification (OI) factor, that attempts to characterize how a scenario is suitable for a learning policy. It will help to characterize the learning improvement in the rest of the thesis, and can be used in general for a fair evaluation of learning capability (compared to a random behavior for instance). Finally, numerical verification of OI factor and Lempel-Ziv (LZ) factor on the state-of-the-art learning policies applied to the OSA scenario is presented in this chapter. As a result, this chapter is published in EAI CrownCom international conference [113].

In, **Chapter 4**, the notion of learning on two different criteria is introduced. Then, an index based learning policy, referred as restless quality of service UCB (RQoS-UCB), is proposed to tackle the problem of OSA when channels are characterized not only by their availability but also by their quality. The problem is hence modeled by a modified Markov MAB framework that takes into an account rewards associated with two separate quantities, i.e. channel quality and availability. The major contribution of this thesis is to design RQoS-UCB policy for multi-armed bandit framework which separately optimize the availability and quality criterion. Following RQoS-UCB policy, its multi-player extension, named distributed RQoS-UCB policy, is proposed to provide collision-free channel access to several non-cooperative selfish users. Then, theoretical results of our proposed single-player and multi-player policies, i.e. logarithmic order upper bound on regret, are presented for the rested and restless Markov MAB framework. Finally, Extensive numerical analysis of state-of-the-art and proposed policies is performed to validate their efficiency in OSA scenario. Thus, **the main contributions of Chapter 4 are:**

- RQoS-UCB algorithm is proposed, which finds the best channel following a separate optimization of both channel quality and availability with "tunable" coefficients.
- Distributed RQoS-UCB policy, multi-player extension of the single-player, is proposed to provide channel access to non-cooperative secondary users (SUs).

- Our proposed policies are proved to converge to an optimal configuration in the restless Markov MAB framework.
- Extensive simulations of our proposed policies are performed in OSA scenario which demonstrate that RQoS-UCB policy outperforms other state-of-the-art policies.
- Parts of this chapter are published in IEEE Globecom 2015 [110] and IEEE Transaction on Cognitive Communications and Networking [112]

Chapter 5 introduces the concept of a flexible network architecture, which enables the base stations to switch between active and sleep modes. Dynamic BSs switching problem is modeled with the Markov MAB framework, and then the performance of our proposed RQoS-UCB class of algorithms is illustrated to answer the dynamic BSs switching problem. Next, the transfer learning (TL) concept is introduced for classical MAB problems. These transfer learning algorithm is designed to prioritize the action space and map it with corresponding transfered knowledge base. Then, the analytical convergence analysis of the proposed transfer RQoS-UCB (TRQoS-UCB) policy is presented. Finally, extensive numerical analysis is performed to validate the efficiency of the proposed schemes. Thus, **the main contributions of Chapter 5 are:**

- Dynamic BS switching problem is modeled as Markov MAB framework, and solved with the RQoS-UCB policy.
- Transfer Learning (TL) concept has been discussed to initialize the classical MAB decision making policy and to achieve a performance jump-start.
- Evaluation of our proposed policies is performed by simulation for dynamic BS switching framework.
- Work in this chapter is submitted to IET communications journal [116].

Next, **Chapter 6** presents a proof-of-concept of our findings in a real radio environment. First, we perform OSA with MAB approaches on the real measurement of HF band rather than simulated channel occupation patterns. This work is communicated in IEEE ICC 2016 [106] and IEEE Transactions on Cognitive Communications and Networking [105]. Then, a proof-of-concept for OSA in the decentralized network consisting of multiple secondary users (SUs) is proposed. We implement MAB learning and decision making policies on the proposed testbed consisting of a USRP boards from Ettus Research, a GNU radio companion for PUs and Matlab/Simulink for SUs. Then, experimental results are presented which depict successful transmission percentage, number of SU collisions and the number of channel switchings making SU terminals energy efficient. The presented proof-of-concept has received Best Demo (Booth)

award in EAI CrownCom 2016 conference [34] and will also support next IEEE ICC 2017 conference tutorial "On-Line Learning for Real-Time Dynamic Spectrum Access: From Theory to Practice".

Finally, **Chapter 7** concludes the thesis and summarizes original contributions. It presents possible future work based on this thesis.

Appendices detail some works which have not been included in above discussed chapters, such as Appendix B which deals with non-stationary multi-armed bandit framework and algorithm to solve it. This work has been published in EAI Endorsed Transactions on Wireless Spectrum [111]. Moreover, Appendix C gives details of mathematical development of Chapter 4 which have been also included in IEEE Transactions on Cognitive Communications and Networking [112].

Chapter 2

On-line Learning for Green Cognitive Radio

Contents

2.1 Introduction **9**
2.2 Cognitive Radio **10**
2.2.1 Definitions 10
2.2.2 Basic Cognitive Cycle 11
2.3 Decision Making Process **13**
2.3.1 Machine Learning 13
2.3.2 Reinforcement Learning 15
2.3.3 Transfer Learning 17
2.4 Multi-armed Bandit – Theory, Variants and Policies **19**
2.4.1 Stateless (Classical iid) Bandit Problem Formulation 21
2.4.2 State-dependent (Markovian) Bandits Problem Formulation 22
2.4.3 Stateless Bandits Policies 23
2.4.4 Markovian Bandits Policies 27
2.5 On-line Learning for Green Cognitive Radio **30**
2.5.1 On-line Learning for Dynamic Spectrum Access 30
2.5.2 On-line Learning for Energy Efficient Wireless Networks 32
2.6 Conclusion **34**

2.1 Introduction

This chapter introduces state-of-the-art works on decision making process for CR technologies and their advances since early days up to present. Due to the large number of publications

about CR and its functionality, we do not intend to discuss it in great details. However to be consistent, we present a brief overview of main CR definitions considered by the community. In this chapter, we discuss a CR decision making problem by following two applications context: i) first problem aims at find out effective decision making tools for Dynamic Spectrum Access (DSA) problem. ii) second problem tackles the urgent need of decision making tools development for CR to achieve energy efficiency in cellular networks by means of dynamic network planning.

The outline of the rest of this chapter is the following: Section 2.2 presents the main definitions and concepts associated with CR and also introduces the basic cognitive cycle. In Section 2.3, we present a brief overview and general classification of machine learning based decision making approaches for CR. Moreover, following the same notions from machine learning in general, we present the multi-armed bandit (MAB) problem formulation and preliminaries in Section 2.4. This section also introduces the state-of-the-art learning policies for multi-armed bandit problem. Finally, Section 2.5 briefly reviews published works related to CR applications, e.g. dynamic spectrum access, energy efficient wireless network. Finally, Section 2.6 summarizes the chapter.

2.2 Cognitive Radio

The transition from hardware-based radio to software defined radio (SDR) architectures and, eventually, their enhancement with machine learning capabilities and self-adaptability, is a gradual process that have been cultivated in the early 1990s and then accelerated in the early 2000s. Even though, SDR paradigm was initially of interest primarily for the defense research, over the past years with the eve of SDR in the industry, CR have got significant attention from the academia and industry as well. CR has been the first answer to solve spectrum scarcity issue which collected a huge attention in early 21st century.

2.2.1 Definitions

Just like SDR, CR is not a standardized technology. For this reason, there is no precise definition of what CR technology exactly constitutes. Thus, it is useful to present CR by summarizing some definitions found in the literature [127] of the CR research, development and regulation:

- Joseph Mitola [109]: *A really smart radio that would be self-, RF- and user-aware, and that would include software technology and machine learning capabilities along with a lot of high-fidelity knowledge of the radio environment.*
- Simon Haykin [57]: *A CR is an intelligent wireless communication system that is capable of being aware of its surroundings, learning, and adapting its operating parameters (e.g.*

transmit power and carrier frequency) on the fly with an objective of providing reliable anytime, anywhere, and spectrally efficient communication.

- The Institute of Electrical & Electronic Engineers (IEEE): *A radio frequency transmitter/receiver that is designed to intelligently detect whether a particular segment of the radio spectrum is currently in use, and to jump into (and out of, as necessary) the temporarily-unused spectrum very rapidly, without interfering with the transmissions of other authorized users.*

- ITU's Radiocommunication Study Group [64]: *A CR is a system that allows to observe, and to be aware of, its operational environment and can dynamically and autonomously adjust its operating parameters and protocols accordingly.*

Clearly, the definitions presented above somewhat vary which is only a matter of understanding, because sometimes the definition is pertinent to a specific application of the CR technology. However, from above definitions, we can clearly state two main functionalities of CR device: *reasoning* and *self-adaptivity*. The work presented in this thesis is focused mainly on exploring how the combination of self-adaptivity and reasoning can be applied for CR in general, and in particular to achieve energy efficient communications in cellular network.

2.2.2 Basic Cognitive Cycle

Cognitive capability is the most distinguishing feature of CR when compared with SDR. The cognition ability tries to capture the variations of the environment in which CR devices is operating over a period of time or space [57].

Two levels of cognitive cycle are suggested in [36]: node-level and network-level. At node-level, each CR device processes a cognitive cycle and performs its own decision in a cooperative or non-cooperative manner. Conversely, at network-level, each CR device (or centralized player on behalf of all CRs) performs a cognitive cycle and makes its decision in a cooperative manner for the entire network. The network-level cognitive cycle can be used in centralized decision making problem such as self-organization of cellular networks.

We can see the main functions of cognitive cycle in Fig. 2.1 [4, 57, 142]: Observe, Analysis, Decision Making/Learning and Adaptation. The details of the functions are as follows [4, 57, 142]:

- Observe: Through its sensors a CR device collects information about its environment. At a particular time instant, each CR device observes the current state, which is the representation of the operating environment (e.g. interference level, regulator rules, etc.).

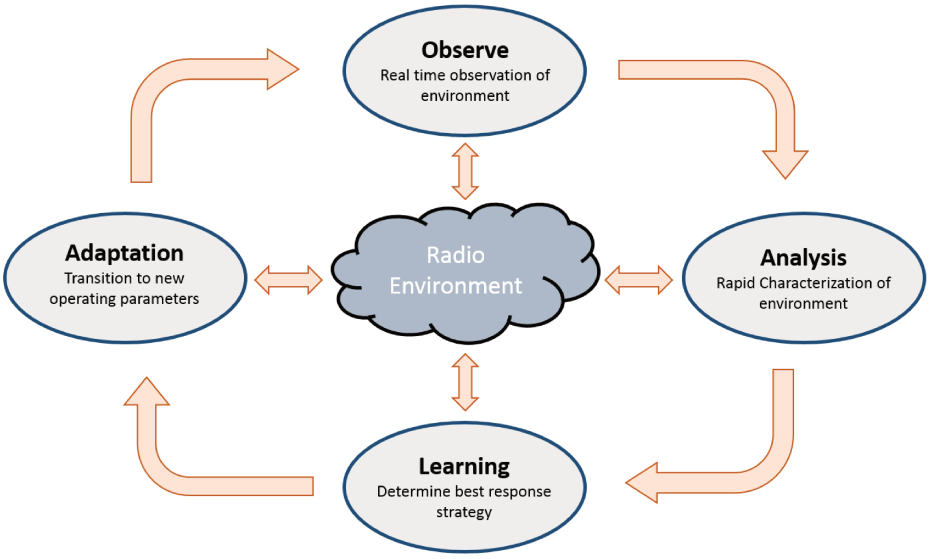


FIGURE 2.1: Cognitive Cycle of [127]

- Analysis/Representation: Based on the information provided by sensors, CR device analyzes the observed state and finds suitable representation (e.g. channel capacity, energy efficiency, spectral efficiency, etc.) to be utilized by the decision making engine to decide the next action.
- Decision Making/Learning: Decision making is the main research area in this thesis. According to the information provided by the observation part, a CR device determines which action to perform [57].
- Adaptation: The operating parameters of CR device are selected as suggested by the decision making process in order to fulfill some goals of interest.

According to [109, 127], the operating environment can be an internal entity such as the instantaneous queue size, or an external entity, such as the wireless medium usage. The current action of CR could affect the operating environment (or future state) for better or for worse, or maintains the status quo; and this in turn affects the next course of actions of CR. Hence, at any time instant, the CR device aims at improving its behavior at next time instant through carrying out a proper action. In terms of learning, the reward parameter is the criteria used to evaluate the improvement of behavior.

Another important feature of CR is the capability of adaption [108, 109]. The CR device or system will adapt its internal states, thanks to SDR capabilities, to the variations of the wireless environment by adjusting certain operating parameters, i.e. carrier frequency, transmission power, modulation, etc. The on-line adaptation of the operating parameters provides the basis for CR to dynamically interact with the environment.

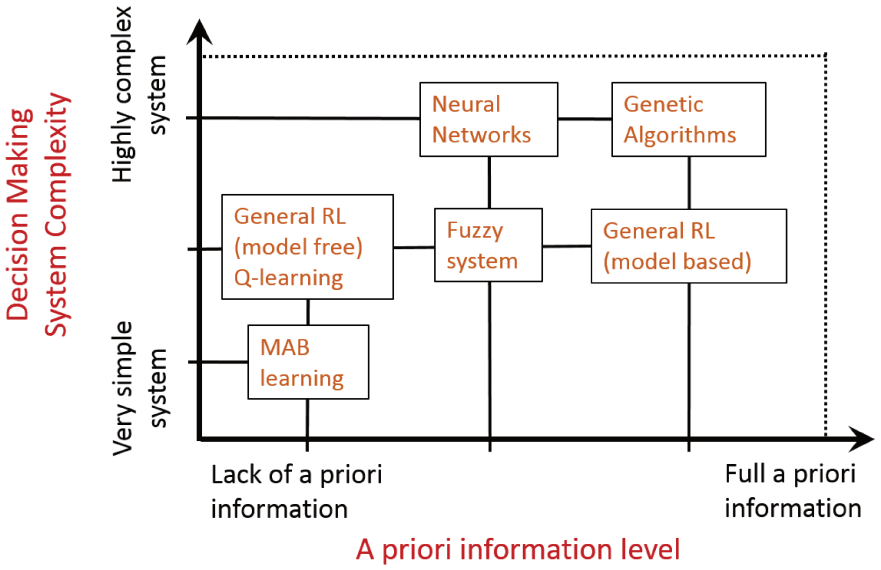


FIGURE 2.2: Decision making techniques depending on the two-dimensional design space {*System Complexity, a priori information level*}.

2.3 Decision Making Process

The decision making process is the core entity of CR that acts as a "brain" of radio system [57]. There are a large number of machine learning strategies developed in the computer science field for robotics, which can be applied to the CR decision making process.

In this section, we propose and discuss two-dimensional design space to classify the decision making approaches for CR. Figure 2.2 provides the classification of machine learning techniques in the x-y plot where the a priori information level is on the x-axis and decision making system complexity is on the y-axis.

On the one hand, a priori information defines a knowledge about the underlying model and representation of the available information. Whereas on the other hand, the system complexity is estimated by the dimension of state space and reward process modeling the environment in which CR is operating and also the number of users operating in the environment.

2.3.1 Machine Learning

Machine learning is a field that is concerned with design and development of algorithms and techniques that improve automatically with experience [8, 134]. It is a multidisciplinary field that relies on results and concepts from artificial intelligence, probability and statistics, computational complexity theory, control theory, information theory, signal processing, etc. Machine

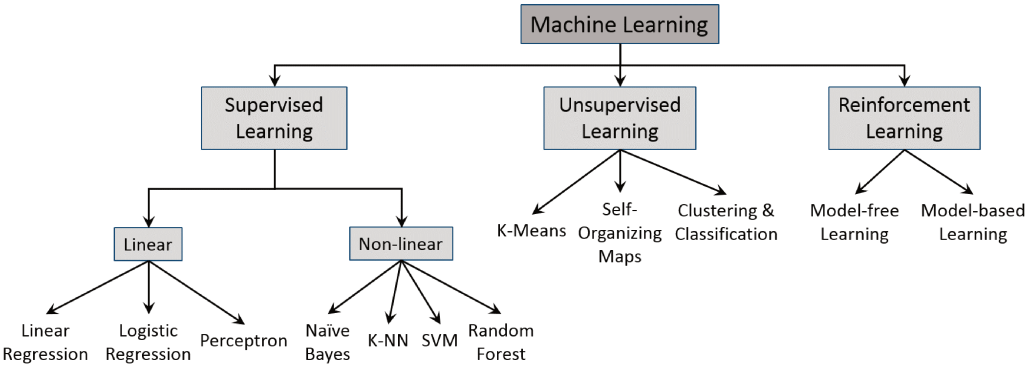


FIGURE 2.3: Taxonomy of Machine Learning Algorithms

learning is widely applied to several critical problems such as natural language processing, pattern recognition, search engine optimization, medical diagnosis, bioinformatics, brain-machine interfaces, speech and handwriting recognition, object recognition in computer vision, game playing and robot locomotion.

Machine learning problem can be generally defined as of [134]:

Definition 2.1 (Machine Learning)**.** A player is said to learn from environment with respect to some class of tasks $\mathcal{A}$ (e.g. actions) and performance measure r (e.g. state and reward), if its performance at tasks $\mathcal{A}$, as measured by r, improves with experience.

Taxonomy of machine learning algorithms are organized as in Figure 2.3, based on the desired outcome, the available input and learning style of the algorithm. Common algorithm types include [8, 134]:

- Supervised learning: besides input, the desired output is also given to player. The objective of a player is to generate a function that maps inputs to the desired outputs. Several algorithms, e.g. regression, perception, decision trees, neural network, bio-inspired learning, nearest neighbor, support vector machines, etc., are available in literature to solve the supervised learning problem.

- Unsupervised learning: the player receives the inputs, but does not possess any supervised target output. A model is prepared by deducing structures present in the input data, without any feedback. Unsupervised class of problems include clustering, dimensionality reduction and association rule learning, and respective example algorithms are the self-organizing maps and K-Means.

- Reinforcement learning: the player learns a policy of how to act given an observation (feedback) of the unpredictable environment. Every action has some impact in the environment, and the environment provides feedback (reward) that guides the learning algorithm. A

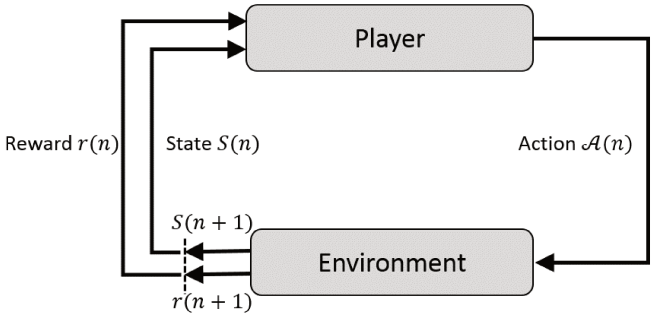


FIGURE 2.4: Reinforcement Learning Framework

list of examples include Q-learning, Temporal Difference (TD) learning, Upper Confidence Bound (UCB), etc.

2.3.2 Reinforcement Learning

Reinforcement learning (RL) is a subclass of machine learning approach which learns on-line to maximize a long-term reward without any a priori information. The methodology is to discover which actions yield the most valuable reward by trying them. The operation method of RL is trial-and-error and delayed reward [142]. The player should observe the state and reward available from the environment and takes actions that maximize them [142]. The RL model [142] where a player is interacting with the environment, as illustrated in Fig. 2.4, is defined as:

Definition 2.2 (Reinforcement Learning)**.** At each time n, the player receives some representation of environment *state*, $S(n) \in \mathcal{S}$, where $\mathcal{S}$ is the state space, a numerical *reward* $r(n)$, and on that basis a player selects an *action* $\mathcal{A}(n)$ with policy $\mathcal{A}$ to collect a *reward* $r(n+1)$ in state $S(n+1)$ at time $n+1$. The objective is to develop the RL algorithm $\mathcal{A} : \mathcal{S} \rightarrow \mathcal{A}(\mathcal{S})$ that maximizes the reward r in state S.

In the standard RL algorithm, the value $V^{\mathcal{A}}(S)$ of the state S under policy $\mathcal{A}$ is the basis to choose the action $\mathcal{A}(S)$. An optimal policy is the one which maximizes $V^{\mathcal{A}}(S)$ at each trial [142]:

$$V^{\mathcal{A}}(S) = \mathbb{E}\left[\sum_{n=0}^{\infty} \lambda^n r\Big(S(n), \mathcal{A}(S(n))\Big) | S(n) = S\right] \tag{2.1}$$

where $\mathbb{E}$ is expectation operator, $\lambda, 0 < \lambda < 1$ is a discount factor and $r\left(S(n), \mathcal{A}(S(n))\right)$ is the immediate reward with action $\mathcal{A}(S(n))$ given the state $S(n)$. (2.1) can also be expressed as:

$$V^{\mathcal{A}}(S) = \left[R\left(S, \mathcal{A}(S)\right) + \lambda \sum_{S'} P\left(S' | S, \mathcal{A}(S)\right) V^{\mathcal{A}}(S')\right]$$

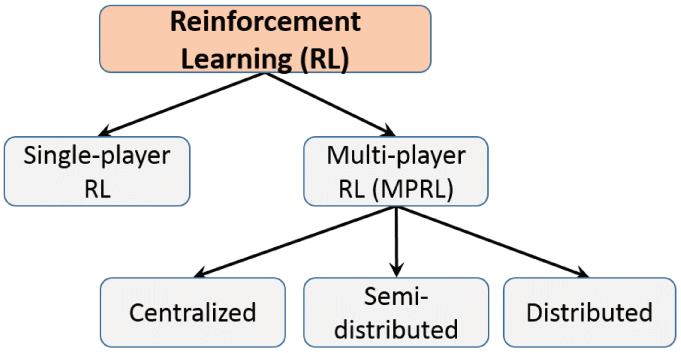


FIGURE 2.5: Classification of Reinforcement Learning Framework

where $R\left(S,\mathcal{A}(S)\right)=\mathbb{E}\left[r\left(S,\mathcal{A}(S)\right)\right]$ is the expected value of $r\left(S,\mathcal{A}(S)\right)$, S' stands for the goal state in which current state S will transit to by taking action $\mathcal{A}(S)$ with policy $\mathcal{A}$. Given the fact that there may be several successor goal states S', the probability $P\left(S'|S,\mathcal{A}(S)\right)$ defines the probability of making transition from state S to different goal states S'.

Furthermore, RL algorithms can be, as in Figure 2.5, classified along several dimensions, such as single and multi-player RL, stem from properties of underlying systems in general. However, single-player RL is not sufficient enough to justify and solve problems arising in current cellular networks, thus, multi-player RL are getting more and more attention from academia and industry as well.

Definition 2.3 (Multi-player Reinforcement Learning)**.** At each time n, the player $j\in U$, where U is a group of autonomous, interacting or non-interacting players, receives some representation of environment *states* $S^j(n)\in\mathcal{S}$ and numerical *rewards* $r^j(n)$, and on that basis player j selects *actions* $\mathcal{A}^j(n)$ with the goal of maximizing its *individual reward* $r^j(n+1)$ in state $S^j(n+1)$ or *global reward* $\sum_{j=1}^{U}r^j(n+1)$.

As in Fig. 2.5, multi-player RL algorithms are classified into three sub-classes: i) centralized algorithms where decisions are made at a centralized controller (i.e. network-level CR), ii) distributed algorithms where decisions are made by individual players without any cooperation among them (i.e. node-level CR), and iii) semi-distributed algorithms which is a mixture of centralized and distributed techniques (i.e. network-level CR).

Exploration vs Exploitation Dilemma In order to maximize the player's performance (i.e. exploitation), it has to gather information about the environment by trying out different alternative actions (i.e. exploration). Exploitation and exploration decisions, however, have to be carefully made. If the player focuses solely on exploration, it will possess accurate information about the environment, but might not be able to maximize its cumulated rewards. On the contrary, by putting more effort on exploitation, the player might miss a chance to find a best

action. Thus, one of the most crucial challenges in decision making with RL is the problem of finding a trade–off between exploration and exploitation. Some RL algorithms perform only exploration first, and then only exploitation. Others mix both exploration and exploitation all along the process. However in both cases there is a trade-off to be smartly considered between exploration and exploitation.

Literature prior to the introduction of CR largely focused on centralized approaches due to its simpler implementation. Recently, distributed algorithms gained more attention after CR has been introduced since it presents more flexibility and reduced computational power for a CR in a fully-distributed manner [148, 152]. However, this advantages comes with certain critical challenges, such as fairness and competition, since each CR device is fully-distributed and decisions are made only according to the local measurements. Thus, several works try to model CR decision making process with semi-distributed approaches where decisions are also made at node-level with certain functionalities of network-level centralized approaches.

2.3.3 Transfer Learning

Transfer Learning (TL) is a general machine learning problem and it is difficult to provide formal definition able to take into account all the possible perspective and approaches to the problem. Our definition of TL focuses on applying knowledge learnt from one problem (historical or source task) to a different (target task) but related problem [128, 149]. Let us first define formal definition of source task and target task as:

Definition 2.4 (Source Task)**.** It is defined as a task or period from which the knowledge is transferred to a current (or target) task.

Definition 2.5 (Target Task)**.** It is a current task in which learning is improved through knowledge transfer from a source task.

Each time a CR system starts operating in a network, RL algorithm has to build its knowledge base starting from scratch and adapt to specific environment, and thus the CR system makes random decisions in the beginning that may be avoided by successfully utilizing transfered knowledge. TL is not designed to replace traditional learning algorithms, but it acts as a supplement to the learning systems on different tasks. Figure 2.6 presents different combinations of traditional machine learning and TL [128]. The idea of TL is perfectly applicable for CR applications in cellular networks.

In this section, we propose a taxonomy of the major transfer learning approaches. We define three main dimensions: the transfer setting, the transferred knowledge, and the objective, as in Figure 2.6. First, we will distinguish among two different categories of transfer settings:

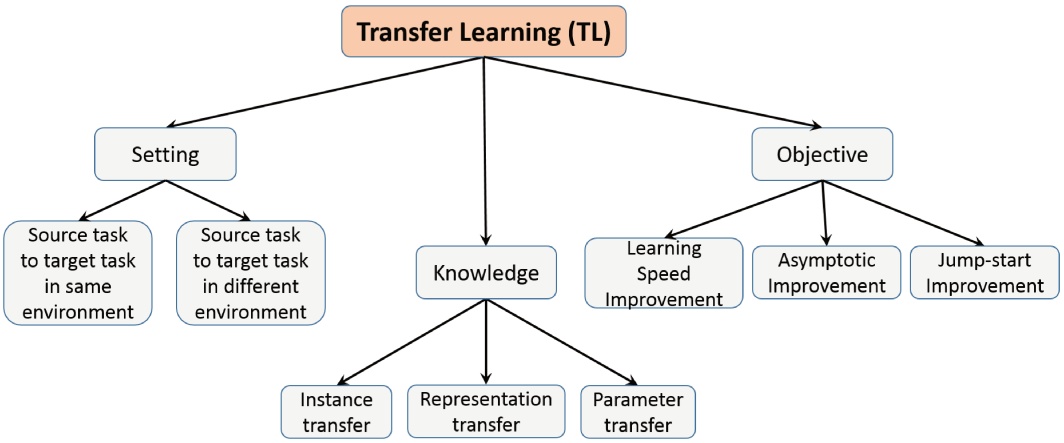


FIGURE 2.6: Taxonomy of transfer learning problem according to three main dimensions: the transfer setting, the transferred knowledge, and the objective [128]

- Transfer from (single or multiple) source task to target task with fixed environment: Most of the literature in TL focuses on the setting in which the environment is fixed and only two tasks are involved: a source task and a target task [128]. We expect that, as the number of source tasks increases, the TL algorithm is able to improve the average performance on the target task.

- Transfer across tasks with different environments: In this setting, the source and target tasks have a different environment, that is they might have different state-action variables.

Next, we classify the TL approaches into three different categories according to transferred knowledge: instance transfer, representation transfer and parameter transfer:

- Instance transfer: In this case, the RL algorithms rely on a set of samples collected from a direct interaction with the environment to build a solution for the target task at hand [84, 143].

- Representation transfer: Each RL algorithm uses a specific representation of the task and of the solution, such as state-aggregation, neural networks, or a set of basis functions for the approximation of the optimal value function.

- Parameter transfer: Most of the RL algorithms are characterized by a number of parameters which define the initialization and the behavior of the algorithm itself. For instance, in Q-learning the Q-table is initialized with arbitrary values. In this case, parameters of target RL approach is changed and adapted according to the source tasks.

Finally, we introduce a number of metrics to measure the improvement with TL over RL approaches [129, 144]. Here we discuss three main TL improvements, as defined in Figure 2.7:

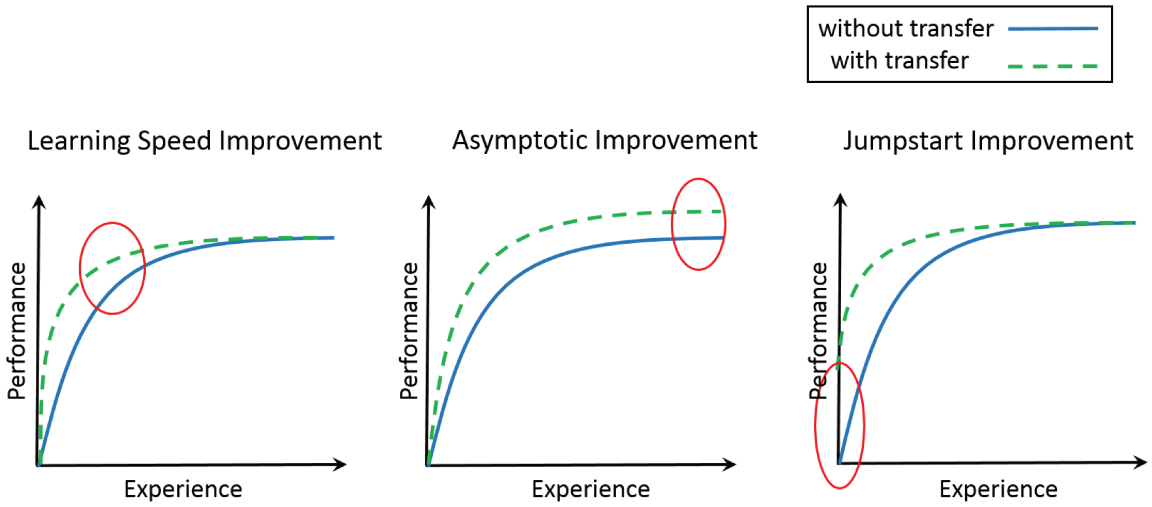


FIGURE 2.7: The three main objectives of transfer learning [129]. The red circles highlight the performance improvement in the learning process.

- Learning speed improvement: this objective is about the reduction of the amount of the time needed to learn the optimal solution of the target task at hand.
- Asymptotic improvement: in most of the problems of practical interest, a perfect approximation of the optimal value function or policy is not possible (e.g., problems with continuous state-action spaces) and the use of approximation techniques is mandatory.
- Jump-start improvement: the learning process usually starts from either a random or an arbitrary hypothesis. In this case, the improvement of the initial performance can be obtained by initializing the learning algorithm to the optimal policy of the source task, which may lead to improve or worsen the learning speed of the target task.

2.4 Multi-armed Bandit – Theory, Variants and Policies

Multi-armed bandit (MAB) is a special class of sequential decision making problems (subclass of reinforcement learning), where, in the most classical form, given a set of gambling machines (arms), a gambler (player) has to play an arm at each iteration in order to collect some reward. In a case, where player possesses a priori information about the expected rewards of the different arms, it would always play the one which maximizes his cumulative reward. Same as the most common RL framework, the expected rewards are not known to the player a priori; however, upon playing any arm, an instantaneous reward of the selected arm is revealed. In such an unknown environment, at each trial, the player may suffer from reward loss (or incur some cost)

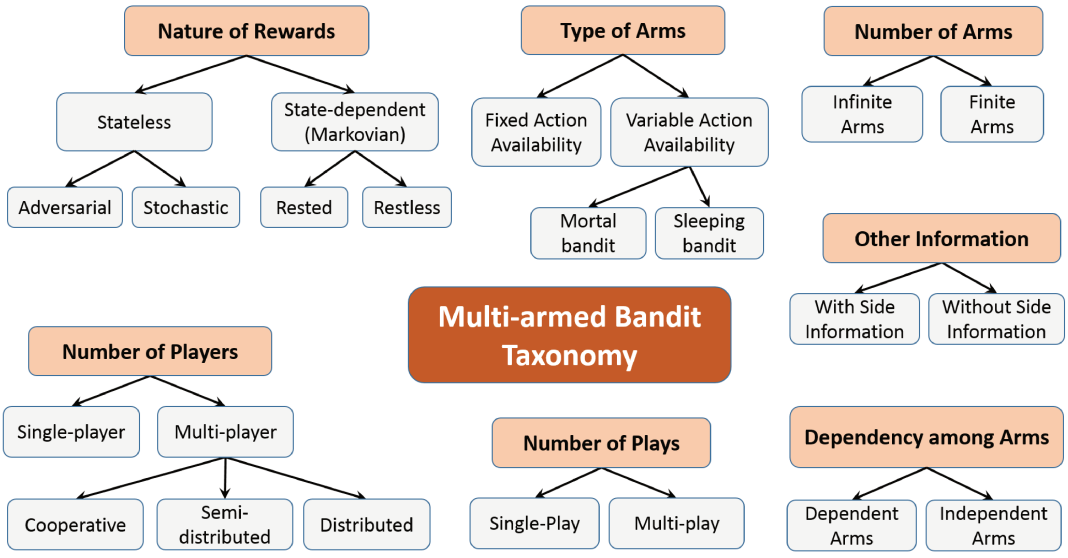


FIGURE 2.8: Classification of MAB problem formulation

due to not playing the optimal arm instead of the played arm. Let's first define optimal and suboptimal arms as:

Definition 2.6 (Optimal Arm)**.** An optimal arm is an arm with maximum expected rewards from the set of arms $\mathcal{K}$. In the multi-player case with U users, the optimal arm set contains U arms with maximum expected rewards.

Definition 2.7 (Suboptimal Arms)**.** All other arms, which are not in the set of optimal arms, are said suboptimal arms.

In this thesis, we provide a brief overview of both MAB variants and policies designed for it. In Fig. 2.8, we propose a general classification of MAB problem formulations. Although we do not consider all of these variants in this thesis, many of them may form the basis for future work. The MAB model can be varied from a number of aspects, such as changing the set of arms, the nature of the rewards, or side information that the player can take into account. In this thesis, we discuss bandit variants depending on reward characteristics which seem more relevant for the CR decision making problems. When arms possess different states, the reward depends on the arms' current states, and the MAB model is referred to state-dependent (Markovian) model. Otherwise, the MAB model is stateless, which itself can be further divided into few subsets. As state-dependent and stateless MAB models are inherently different, we discuss them separately in the following.

2.4.1 Stateless (Classical iid) Bandit Problem Formulation

In a stateless MAB model, arms do not have any specific states, and rewards generated from each arm are independently drawn from an unknown distribution, when played. A player must choose which arms to play. Let, there are K independent arms in the set $\mathcal{K} = \{1, 2, \cdots, K\}$, and U players (or users) indexed in $\mathcal{U} = \{1, 2, \cdots, U\}$. At each time step n, player $j \in \mathcal{U}$ selects an arm $i \in \mathcal{K}$ with action $A^j(n)$, and receives reward $r^{i,j}(n)$, drawn from an unknown distribution of arm i. Without loss of generality, we assume that the reward collected from arm i has a bounded support $r^{i,j} \in [0, 1]$. Now, let μ_i denote the mean reward of arm i. Without loss of generality, let's consider that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. Finally, let $n > 0$ denotes the time horizon in which the player operates and makes decision.

In the multi-player setting, interaction between the players plays a significant role in the global performance. In general, players need to interact when they select the same arm at the same time. These types of interaction between players may change their mechanism to observe and collect rewards from arms. For example, in an OSA scenario, a CR transmitter-receiver pair can be regarded as a player. If two CR transmitters pick the same channel for opportunistic transmission, due to interference between both transmissions, they may fail to correctly receive the data. As a result, we expect significant delay in the communication, and also significant energy used in signal transmission is wasted. Thus, below we present some of the state-of-the-art reward collection methodologies for multi-player MAB model.

Definition 2.8 (Strict collision model)**.** Let an ALOHA-like protocol be considered under which if two or more players use the same arm then they get zero reward, and no collision avoidance mechanisms are considered.

The reward collection methods presented above only affect the ability of a player to collect a reward, however, they do not affect the ability of a player to observe the reward. Thus, the mean reward μ^i of arm i is the same as the mean reward in the single-player model.

In the general RL framework, the performance of learning policies is analyzed using two important performance criteria: convergence and *regret* bounds [25]. The former is the guarantee that the learning algorithm, at the end, finishes to play *almost* always the optimal arm, and the latter measures the speed at which the convergence is achieved and is defined as the difference between the expected reward obtained using a full knowledge strategy and the selected learning algorithm. Given this, in order to analyze the performance of a policy we compare its performance with the *best arm selection* policy, which knows optimal arm a priori. The regret $\Phi^{\mathcal{A}}(n)$ of any policy $\mathcal{A}$ after n plays can be defined as:

$$\Phi^{\mathcal{A}}(n) = \sum_{j=1}^{U} n\mu_j - \sum_{j=1}^{U} \sum_{t=1}^{n} r^{i,j}(t) \tag{2.2}$$

Another class of stateless bandits is Adversarial or non-stochastic MAB model. It is similar to the stochastic ones in being stateless, with the difference that the observed rewards cannot be associated to any specific density function; in simple words, rewards are non-stochastic. Generally, within this setting, the player selects a finite repeated game against nature, or an adversary, in which at each time step n, the adversary draws a reward, which is unknown to the player.

Although, state-independent (IID) model is a rather simple yet elegant mathematical model for which sharp results can be derived, realistic modeling of several real-world practical applications requires incorporation of temporal information. A more complicated, yet analytically tractable state-dependent model is the Markovian model.

2.4.2 State-dependent (Markovian) Bandits Problem Formulation

In a state-dependent MAB model, every arm is associated with some finite state space. Upon being played, each arm pays some positive reward that is generated from some stationary distribution associated to the current state of that arm. The states of arms change over time according to some stochastic process, which is considered to be a Markov process. Roughly speaking, a process satisfies the Markov property if its future state solely depends on its current state rather than its full history. This type of MAB model is also referred to as Markovian MAB model. Note that in this formulation, at each iteration, the reward as well as the state of only the played arm is revealed.

In the Markovian setting, we consider that an arm has two modes, i.e. *active* and *passive*. An arm is referred as in active mode if it is selected by a player at current time step, otherwise it is in passive mode. Markovian MAB model is further classified in rested and restless problems as in Fig. 2.8:

- **Rested Markovian MAB model:** The state of only active arm evolves according to a Markov process. However the state of arms which are in passive mode remains frozen, i.e. it does not change [12, 145].
- **Restless Markovian MAB model:** Reward state of each arm evolves dynamically following unknown stochastic processes no matter it is played or not, and state evolution of the arms in passive mode is independent of the actions of the player [94, 146].

Let each arm in Markovian MAB problem be modeled as an aperiodic, irreducible and discrete time Markov chain with finite state space S^i. When arm i is activated, transitions to next state occurs according to the transition probability matrix $P^i = \left\{p^i_{kl}, k, l \in S^i\right\}$, where p^i_{kl} is the $i-$th arm's transition probability from state k to state l. Moreover, the Markov state space is S^i,

where $q \in S^i$ is the observed states. Markov chains are independent from each other and $\boldsymbol{\pi}^i$ is the stationary distribution of the $i-$th Markov chain with $\pi_q^i(n) = \pi_q^i \ \forall n$. Let a learning policy be defined as a one-to-one mapping $\mathcal{A}$ such as at each time n, a channel i is selected:

$$\begin{aligned} \mathcal{A} : \mathbb{N} &\longrightarrow \mathcal{K} \\ n &\longmapsto i \end{aligned}$$

At time slot n, player $j \in \mathcal{U}$ selects an arm $i \in \mathcal{K}$ according to some policy $\mathcal{A}$. Upon selecting an arm i at time n, player j observes the current state $q \in S^i$ of arm i, and it also collects the reward $r_q^{i,j}(n)$ associated with state q of the $i-$th arm. The mean reward μ_i of the $i-$th arm under stationary distribution $\boldsymbol{\pi}^i$ is given by:

$$\mu_i = \sum_{q \in S^i} r_q^i \pi_q^i$$

.

Same as stateless bandits, state-dependent bandit policy $\mathcal{A}$ is evaluated with the regret notion:

$$\Phi^{\mathcal{A}}(n) = \sum_{j=1}^{U} n\mu_j - \sum_{j=1}^{U} \mathbb{E}\left[\sum_{t=1}^{n} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}(t)\right] \tag{2.3}$$

where the expectation $\mathbb{E}$ is taken over the states. Let $q_{\mathcal{A}(t)}$ being the state observed by using the policy $\mathcal{A}$ at time n. Arms whose mean reward is strictly less than μ_j are referred as suboptimal arms.

2.4.3 Stateless Bandits Policies

Single-player Stateless Bandit Policies

Within this section, we discuss the policies that solve stateless MAB problem in more details. Recall that a fundamental trade-off in the MAB model is the exploration vs exploitation dilemma. On the one hand, if the player exclusively selects the arm that it thinks is the optimal (i.e. exploitation), it may fail to find that one of the other arm actually has a higher mean reward. On the other hand, if it spends a lot time exploring out all the arms and gathering statistics (i.e. exploration), it may fail to choose the optimal action often enough to maximize its gain.

Researchers have proposed a variety of approaches that tries to address this exploration-exploitation dilemma from different aspects. Initially, [37, 142, 151] proposed the most simple policies, greedy algorithm and its variants, in which the player follows the greedy policy, i.e. it plays the arm with the highest estimate, with probability $(1-\epsilon)$, and it plays a random arm with probability ϵ.

ϵ is fixed a priori and can be understood as an exploration parameter; that is, higher ϵ indicates more exploration and vice versa. However, this implies that ϵ–greedy policy never achieves asymptotic convergence to the optimal arm, since it is necessary to stop exploration phase once the optimal arm is found. [15] tackled the asymptotic convergence issue and proposed an algorithm, named decreasing ϵ–greedy, where the exploration parameter ϵ is decreasing with time. Beside the asymptotic convergence, [15] also presented that decreasing ϵ–greedy has a strong finite–time performance.

Then, Lai and Robbins [82] proposed a policy, which they named uniformly good policy, and it achieves $\mathcal{O}(\log n)$ regret bounds for some specific families of distributions, as the time horizon tends to infinity. This bound guarantees that the algorithm achieves convergence asymptotically. This study was later took forward by Agrawal in [2], and he introduced a class of algorithms that are distribution independent, i.e. they do not possess any restriction on reward distribution. In particular, [2] considered a concept called *optimism in the face of uncertainty*, first proposed in [70], that allows the player to select the arms by using a combination of the estimate of the mean reward values and the uncertainty of those estimates, such that arms having high uncertainty are selected more often. With the help of this, Agrawal in [2] showed that the proposed policies are much easier to compute than Lai and Robbins' in [82], and he presented the same asymptotic regret bound, but with a significantly larger constant factor.

To provide finite–time regret bounds, Auer et al. in [15] improved Agrawal's technique by designing a computationally simple policy called upper confidence bound (UCB). This approach achieves not only finite-time $\mathcal{O}(\log n)$ regret bound, but also asymptotic convergence. However, the constant term of the presented bound is significantly larger than that of Lai and Robbins' policy in [82]. UCB1 policy was first proposed in [15] which can be described as follows in Algorithm 1. In Algorithm 1, $\bar{S}^i(T^i(n))$ is the estimate of arm i's mean reward value, $T^i(n)$ is the number of times arm i has been played up to n, $r^i(n)$ is the instantaneous reward collected from arm i at time n, and $\mathbb{1}$ is a indicator function. To improve the performance of UCB1 policy in real–world applications and to decrease the large constant factor within the regret bound, [15, 17] also introduced modified versions of UCB1 policy, named as UCB–tuned and UCB2.

The regret constant factor improvement of UCB class of policies were later achieved, in UCB-V policy, by using modified arm indices based on the empirical mean and variance of the respective arms in [14]. For problems with finite support distributions on rewards, the authors in [23] proposed a policy, named KL-UCB, that uses the Kullback-Leibler confidence bounds to achieve asymptotically optimal sequential allocations. Alongside KL-UCB, Bayes-UCB [76], an explicitly Bayesian variant of UCB, represents the current state-of-the-art of UCB algorithms. In the Bayesian approach, each arm is represented as an estimate of a distribution that is updated in the Bayesian manner and the decision is selected which has the highest score.

Algorithm 1 UCB1 Single-player policy

Input: K, α
Output: $\mathcal{A}(n)$

for $n = 1$ **to** n **do**
  if $n \leq K$ **then**
    $\mathcal{A}(n) = n + 1$
  else
    $T^i(n) = \sum_{m=0}^{n-1} \mathbb{1}_{\mathcal{A}(m)=i}, \forall i$
    $\bar{S}^i(T^i(n)) = \frac{\sum_{m=0}^{n-1} r^i(m)\mathbb{1}_{\mathcal{A}(m)=i}}{T^i(n)}, \quad A^i(n, T^i(n)) = \sqrt{\frac{\alpha \log n}{T^i(n)}}, \forall i$
    $B^i(n, T^i(n)) = \bar{S}^i(T^i(n)) + A^i(n, T^i(n)), \forall i$
    $\mathcal{A}(n) = \arg\max_i (B^i(n, T^i(n)))$
  end if
end for

Then, POKER [151], a non UCB class of algorithm, is a generalizable economic analysis motivated approach to solve the MAB problem. While above presented approaches presents a strong generalization of many of the problems faced in real world, recently, a simpler technique, Thompson-sampling, has been shown to perform competitively to the state-of-the-art approaches in a MAB framework [3, 135]. Table 2.1 presents general overview on the presented approaches. The same principle as of UCB is employed in various sequential decision making problems, ranging from optimization to planning, as stated in recent surveys [22, 115]. Besides the direct improvements concerning the convergence behavior, the classical MAB model has been driven by multiple extensions.

Non-stationary Stateless Bandit Policies

Though the stationary formulation of MAB problem permits to address exploration vs exploitation dilemma more appropriately, it may fail to justify a changing environment where the observed reward distribution undergoes changes in time. As an example, the reward distribution of each arm is likely to experience changes in time, which exhibits the limitation of stationary MAB models. In many application domains, abrupt changes in the reward distribution are an intrinsic characteristic of the problem. Standard UCB class of policies are not well adapted for abruptly changing environment as stated in [56]. Authors, in [16, 18], introduced soft-max action selection policies, i.e. EXP3, EXP3.S, for non-stationary stateless MAB problem, where distribution of reward undergoes abrupt changes in time. Moreover, several policies such as discounted-UCB (D-UCB), sliding-window UCB (SW-UCB), belonging to the wider family of UCB, are designed to address abruptly changing non-stationary stateless MAB problems [51, 87, 137]. These policies are also consistent with more extreme settings, such as the one presented in [138] where reward distribution follows a Brownian motion.

MAB Single-player Learning Algorithms	Computation Complexity	Experimental Performance	Asymptotic Convergence	Finite-time Bound
$\epsilon-$first [37]	$\uparrow$	$\uparrow$	–	–
$\epsilon-$greedy [154]	$\uparrow$	$\uparrow$	–	–
Decreasing $\epsilon-$greedy [15]	$\uparrow\uparrow$	$\uparrow$	Yes	$\mathcal{O}(\log n)$
Lai and Robbins' [82]	$\uparrow\uparrow\uparrow$	–	Yes	–
Agrawal's[2]	$\uparrow$	–	Yes	–
UCB1 [15]	$\uparrow$	$\uparrow\uparrow$	Yes	$\mathcal{O}(\log n)$
UCB-tuned [15]	$\uparrow\uparrow$	$\uparrow\uparrow$	–	–
UCB2 [15]	$\uparrow\uparrow$	$\uparrow$	Yes	$\mathcal{O}(\log n)$
Improved UCB [17]	$\uparrow\uparrow$	–	Yes	$\mathcal{O}(\log n)$
Maillard *et al's*	$\uparrow\uparrow$	–	Yes	$\mathcal{O}(\log n)$
POKER [151]	$\uparrow\uparrow$	$\uparrow$	Yes	–
SoftMax	$\uparrow$	$\uparrow\uparrow$	–	–
Exp3 [16]	$\uparrow\uparrow$	$\uparrow$	Yes	$\mathcal{O}(\sqrt{n})$
KL-UCB [23]	$\uparrow\uparrow$	$\uparrow\uparrow$	Yes	$\mathcal{O}(\log n)$
Bayes-UCB [76]	$\uparrow\uparrow$	$\uparrow\uparrow$	Yes	$\mathcal{O}(\log n)$
Thomson-Sampling [3]	$\uparrow$	$\uparrow\uparrow$	Yes	$\mathcal{O}(\log n)$

TABLE 2.1: An overview of the single-player policies in the stateless multi-armed bandit framework. The symbols have the following meaning: '$\uparrow$' ('$\uparrow\uparrow$') means that the respective property is (strongly) satisfied. In addition, '$-$' means the property is not known.

Multi-player Stateless Bandit Policies

Initially, classical single-player MAB problem was extended by Anantharam *et al* in [11] to the setting of centralized multi-player case. In a centralized access schemes, in [11, 46], multiple arms are selected by a single centralized player, at each iteration, and it receives the reward which is a linear combination of the collected rewards from selected arms. However, it requires extensive information flow among players, and this type of learning cannot generally be used in problem where multiple players act selfishly and their collected rewards are affected by the actions of other players. Moreover, a large body of literature assumes that players are able to observe the actions of others, and decide about their actions with these observations. However, the increasing importance of networked systems warrants the development of fully-distributed algorithms for multiple communicating players faced with MAB problems.

Recently, many researchers [10, 41, 47, 71, 73] have studied the semi-distributed or distributed multi-player MAB problem, as in Table 2.2. Primarily motivated by wireless communication networks, these researchers assume no extensive information exchange among independent players and design efficient distributed policies. Liu and Zhao, in [97], proposed a distributed access policy, time-division fair share (TDFS), and proved that it has a $\mathcal{O}(\log n)$ regret for iid distributed rewards. In [97], players orthogonalize their access with different offsets in their time-sharing schedule, while we consider in Chapter 4 that players orthogonalize into different arms. TDFS

policies consider that each player collects almost the same time-average reward while our proposed policy, referred as *distributed RQoS-UCB* in Chapter 4, achieves probabilistic fairness in reward collection. Moreover, TDFS policies assign a fixed offset to each participating player in the beginning of learning in order not to collide with other players, while in our work we consider that players arrive and leave the system independently. Then, Anandkumar *et al.* in [10] formulated the distributed learning and access problem for multiple players, however they considered MAB framework with iid reward distribution only.

Multi-player Setting	Learning Algorithms	Experiment Performance	Asymptotic Convergence	Finite-time Bound
Centralized	[11]: Multiple plays	–	Yes	–
	[46]: Auction problem	↑	–	–
Semi-distributed	[41]: Cooperative	↑	–	–
	[73]: Leader-Follower	↑↑	Yes	–
	[71]: bipartite matching	↑↑	–	$\mathcal{O}(\log^2 n)$
Distributed	[10]: Random Access	↑↑	Yes	$\mathcal{O}(\log n)$
	[47]: Prioritize ranking	↑↑	–	$\mathcal{O}(\log n)$

TABLE 2.2: An overview of the multi-player policies in the stateless multi-armed bandit framework. The symbols have the following meaning: '↑' ('↑↑') means that the respective property is (strongly) satisfied. In addition, '–' means the property is not known.

Finally, there are the Bandits with the state-dependent (Markovian) reward, which is no longer iid. Motivated by practical applications and leading to an important number of interesting problems and solutions, state-dependent bandits to the MAB model are, at the moment, very popular.

2.4.4 Markovian Bandits Policies

Single-player Markovian Bandit Policies

There has been relatively less work on Markovian MAB model. Table 2.3 surveys the most important advances in Markovian MAB policies. Anantharam *et al.* in [12] wrote one of the earliest paper with the rested Markovian MAB setting. It also proposed a learning policy to play U out of the K arms at each time slot and also presented the lower bound and asymptotic upper bound on regret. However, the rewards in this work are assumed to be drawn from rested Markov chains with transition probability matrices defined by a single parameter θ with identical state spaces.

Recent work by Tekin in [145] has extended the previous results of [12] to the case with multiple parameters and no identical state spaces across the arms. They utilized UCB1 from [15]

for rested Markovian MAB problem and proved a finite time $\mathcal{O}(\log n)$ upper bound on the regret under some conditions on the Markov chain. Recently, restless bandit problem got huge attention from researchers due to its wide range of applications. State-of-the-art work on restless Markov MAB problems with independent arms can be found in the following literature [33, 93, 94, 117, 125, 126, 146]. Initially, [146] has proposed a regenerative cycle algorithm (RCA) that achieves $\mathcal{O}(\log n)$ regret under certain assumptions on the underlying Markov process. Regenerative process can be defined as:

Definition 2.9 (Regenerative Process)**.** A stochastic process $\{S(n); n \geq 0\}$ is intuitively a regenerative process if it can be split into iid cycles. That is, we assume that a collection of time points exists, so that between any two consecutive time points in this sequence, (i.e. during a cycle), the process $\{S(n); n \geq 0\}$ has the same probabilistic behavior.

Classic examples of regenerative processes are ergodic Markov chains and processes. For example, take a Markov chain or a Markov process with a countable state space S and fix a state q. Assume that the process started at state q. Then every time that q is encountered is a time of regeneration for the Markov process. Our restless formulation in Chapter 4, uses some elements of the policy and proof from [146], which is however quite different in its operation in a manner that it separates decision engine for state and reward quantity compared to joint decision process in UCB1.

Next, [93] modeled the same problem as in [146], and proposed a restless UCB (RUCB) policy, achieving finite time $\mathcal{O}(\log n)$ regret when certain system settings/parameters are known. [93], constructed a policy using deterministic sequencing of exploration and exploitation (DSEE) with an epoch structure to tackle the restless nature of arm evolution. Specifically, the policy in [93] partitions the time horizon into interleaving exploration and exploitation epochs with carefully controlled epoch lengths in order to achieve $\mathcal{O}(\log n)$ regret bound when some of the system parameters are known. Then, [94] extended the analytical studies by proving near logarithmic order regret ($\mathcal{O}(f(n)\log n)$ for any increasing divergent function $f(n)$), when system parameters are unknown. Whereas, in this thesis, we use a regenerative cycles to effectively solve the restless bandit problem achieving $\mathcal{O}(\log n)$ regret bound when no system knowledge is available. On the same line, [125] also introduced a simple index based policy for restless MAB formulation which has been proven to achieve an asymptotic $\mathcal{O}(\log n)$ upper bound on the regret, however, finite-time analysis of their policy was also missing. Then, [32] also considered the same formulation, and proposed a continuous exploration and exploitation (CEE) policy. When no information is available about the dynamics of the arms, the CEE algorithm guarantees $\mathcal{O}(f(n)\log n)$ regret uniformly over time, where $f(n)$ is increasing sequence ($f(n) \to \infty$ as $n \to \infty$), however the constant factor in regret is comparatively larger than RCA (in [146]) and RUCB (in [93]) policies.

Markovian MAB Setting	Learning Algorithms	Exp. Perf.	Asymp. Conv.	Finite-time Bound
Single-player	[12]: Rested problem	–	Yes	–
	[145]: Rested problem	↑↑	–	$\mathcal{O}(\log n)$
	[146, 147]: Restless problem	↑↑	–	$\mathcal{O}(\log n)$
	[32, 33]: Restless problem	↑	–	$\mathcal{O}(f(n)\log n)$
	[125]: Restless problem	↑↑	Yes	–
Multi-player Centralized	[12]: Rested problem Known statistics	–	Yes	–
	[95]: Restless problem Known statistics	↑	–	–
	[147]: Restless problem	↑↑	–	$\mathcal{O}(\log n)$
Multi-player Semi-distributed	[96]: Cooperative Game	↑↑	–	–
	[157]: Cooperative Game	↑	–	–
	[71]: bipartite matching Rested problem	↑↑	–	$\mathcal{O}(\log^2 n)$
Multi-player Distributed	[97, 98]: Prioritize ranking	↑	–	–
	[94]: Epoch structure Time-sharing fairness	↑↑	–	$\mathcal{O}(\log n)$
	[47]: Prioritize ranking	↑↑	–	$\mathcal{O}(\log n)$

TABLE 2.3: An overview of the learning policies in the Markovian multi-armed bandit framework. The symbols have the following meaning: '↑' ('↑↑') means that the respective property is (strongly) satisfied. In addition, '–' means the property is not known.

Multi-player State-dependent Policies

Table 2.3 presents a brief overview of the available approaches to tackle the multi-player Markovian MAB problem in a centralized, semi-distributed and distributed way. Initially, the works in [95, 147] considered restless centralized access schemes in contrast to distributed access in Chapter 4.

Moreover, Markovian MAB framework has also been designed to address the semi-distributed access problem, as discussed in [71, 96, 157]. However, in these works, the authors have assumed that each player declares its actions to others e.g. the selected arm, which can be a strong constraint and would produce extra signaling overhead in the wireless communication context. Recent work on semi-distributed rested Markov MAB problem by Kalathil *et al.* in [71] proposed a semi-distributed policy, UCB4, for both iid and rested Markovian rewards, which utilizes distributed bipartite matching algorithm as an access mechanism.

Next, [94] also introduced distributed DSEE policy which follows a similar epoch structure as of single-player policy. In the exploration epochs, the players play all arms in a round-robin fashion with different timing offsets determined by the pre-agreement. In the exploitation epochs, each player calculates the sample mean of every arm based on its own local observations and plays some arms with the largest sample mean in a round-robin fashion with a certain offset. Note that

even though the players have different time-sharing offsets, collisions occur during exploitation epochs since the players may arrive at different sets of the arms due to the randomness in their local observations. Moreover, it achieves near logarithmic ($\mathcal{O}(f(n)\log n)$) order regret. Next, [94] also extended distributed DSEE policy to the case when no pre-agreement exist among players, and it proved that proposed policy achieves the same near logarithmic order regret. However, in this thesis, we extend the single player policy (which itself quite different in operation compared [94]) to multi-player, when no system knowledge is available. We prove that it still achieves $\mathcal{O}(\log n)$ order regret. Recently, Gai *et al.* [47] proposed a policy under which players select their preallocated rank-optimal channels to achieve orthogonalization.

2.5 On-line Learning for Green Cognitive Radio

As discussed before, cognitive radio (CR) technologies can be seen as potential approach for wireless networks to maximize spectral efficiency and energy efficiency. Within this context, dynamic spectrum access (DSA) technologies have been proposed to solve spectrum scarcity problem, and thereafter, dynamic network planning, with the help of CR technologies, has also been discussed to maximize the energy efficiency in wireless networks.

2.5.1 On-line Learning for Dynamic Spectrum Access

Several machine learning algorithms have been utilized for CR decision making problem from past decade, yet some of them could be more suitable than others. Different supervised and/or unsupervised class of decision making approaches can be applied in CR such as Fuzzy Logic, Neural Networks, Hidden Markov Model, Genetic Algorithms, or Classification Algorithms [9, 9, 31, 58, 65]. However, these approaches require significant training in order to make it applicable for dynamic CR scenario, and also complexity of them increases with the increase in the uncertainty of the environment. Moreover, these approaches posses strong a prior information to build the initial knowledge base.

From the beginning of dynamic spectrum access era, the multi-player reinforcement learning approaches have been proposed in many papers as an effective solution for CR, mainly in two aspects: improving the selection policy and speeding up convergence. Initially [156, 158, 159] extended the study of single-player RL to multi-player RL scenario where more SUs implement a semi-distributed RL algorithm. However, they did not present the analytical and numerical studies on system overheads and computational complexity due to regular information exchange, and the quality of service (QoS) requirement for SUs was neglected in the design of reward model. Moreover, [28, 49] utilized Q-learning to improve SINR through improving power allocation on semi-distributed CR, however, channel allocation was ignored in this work. Then, [156]

presented a Q-learning approach for joint channel and power allocation. The methodology in [156] defined the states of the system as combinations of the transmit power level and the channel availability information, hence it takes a very long time to converge to an optimal solution. Moreover, the presented model requires system level information exchange which leads to a large amount of system overhead, and analytical study on the computational cost associated with this overhead was missing.

Next, [114, 120, 150] presented a fully distributed version of Q-learning approaches, however, these algorithms did not guarantee the convergence to optimal policy. Although, numerical studies illustrated that when using the distributed Q-learning schemes, convergence can still be achieved given a sufficiently large number of iterations. In order to strike a balance between the simplicity of the model (namely, the distributiveness of learning) and the optimality of RL algorithm, careful modeling is needed with respect to different network scenarios. However, traditional RL approaches have several states which models the system, but it may be very difficult to find and manage states in some scenarios of wireless systems. This issue occurs particularly in the distributed CR scenario. Recently, Jouini *et al.* in [67] proposed the use of MAB framework for CR learning and decision making problem.

In [67], authors utilized UCB1 policy (initially proposed in [15]) as a MAB learning algorithm to learn about the most available channel from the given set of channels in a single-player CR scenario. Detailed discussion and reading on single-player stateless MAB model as a CR learning framework can be found in [67–69, 133]. Recently, Markovian (state-dependent) MAB problems are gaining a lot of interest to model opportunistic spectrum access (OSA) scenario. Initially, the works in [41, 46, 96, 147] considered centralized spectrum access schemes in contrast to distributed spectrum access in this thesis. Moreover, MAB framework has also been designed to address the distributed channel selection problem, as discussed in [71, 94, 157]. In [71], the authors assumed that each SU declares its actions to others e.g. the selected channel, which can be a strong constraint in terms of signaling overhead. In [94], the SUs orthogonalize their transmission with different offsets in their time-sharing schedule, while we consider that SUs orthogonalize into different channels. Next, in [10], the authors formulated the OSA problem of distributed learning and spectrum access for multiple SUs, however they have considered MAB framework with iid reward distribution only in contrast to restless Markov MAB framework in Chapter 4 of this thesis.

Initially, [66, 80, 81] works considered concept of considering data rate as a reward for decision making policies. Later in [96, 98, 99, 124, 125, 147] authors considered Markov multi-armed bandit approaches in which reward process models throughput (or data rate) achieved from respective channels. These state-of-the-art approaches considered reward as a combination of

both channel quality (data rate) and availability statistics, defined as:

$$r^{i,j} = (1 - S^{i,j}(n))R^{i,j}(n).$$

Moreover in the above discussed state-of-the-art approaches, state (availability) of channel was assumed to evolve following Markov process whereas quality information is assumed to be fixed for specific state and channel which is a restrictive hypothesis since it ignores temporal quality variations. On the contrary, we considered channel quality and availability as separate rewards with "tunable" coefficients for learning process. In wireless communication field, the separation of both functionalities, i.e. availability and quality, is necessary for applications where certain level of transmission quality is requisite. Indeed, by separating the engine on the decision and on the quality, we can choose to emphasize on one and/or the other criteria. Recently, the authors in [153] modeled the OSA problem with POMDP framework which considers channel quality along with availability statistics to decide about the channel to sense. However, the presented system model has comparatively higher computational complexity and also there is no theoretical guarantee on the convergence of proposed policies. On the contrary, we model OSA problem with the MAB framework which turns to be very easy and less complex to implement. Channel qualities are moreover taken into account as rewards to different users.

2.5.2 On-line Learning for Energy Efficient Wireless Networks

Recently, there has been a substantial body of work towards traffic load-aware BS adaptation, and the authors, in [29, 118], have validated the possibility of improving energy efficiency from different perspectives and also showed the possibility of energy saving by simulations. According to the traffic loads variation, dynamically putting a BSs into sleep mode is one of the effective ways to reduce the total energy consumption of the network. In [52, 91, 119], authors proposed how to dynamically adjust the sleeping status of BSs, depending on the learned and predicted traffic loads of the network. Then, [102, 103] introduced some BSs switching strategies for dynamic BSs operations depending on daily traffic variation. However, reliable prediction of BSs traffic load is still an important challenge for network operators, which limits its' usefulness in practical applications. An alternative energy-efficient procedure is employed in [6, 7], where some BSs being turned to sleep mode and switched on in a low-powered relay station mode during low-traffic intervals. Next, authors in [121] introduced single-agent RL algorithms as an application of dynamic BS switching operation, however, these algorithms are highly dependent on a priori knowledge of the traffic loads.

As discussed in Chapter 1, and in [50, 75], the problem of energy efficiency maximization with dynamic BS switching operation is a combinatorial problem, and it has been proven to be NP-hard. One of the several possible alternative to solve combinatorial problems generally

require a central controller. Due to the complexity of solving combinatorial problem, several works, e.g. [55, 139], adopted fixed BSs switching patterns and then evaluated QoS metrics, i.e. call blocking probability and the outage probability. On the same track, in [52, 78, 141], some greedy algorithms have been introduced to tackle the BSs switching operation without presenting sufficient theoretical guarantees of convergence to the optimal configuration. The greedy algorithm presented in [141] maintains a trade-off between delay (QoS requirement) and energy consumption. Then, [78] took forward a greedy algorithm to handle trade-off between energy consumption and the revenue in heterogeneous cellular networks.

Most of the prior works dealing with machine learning for the energy saving problem, were generally lacking of solid theoretic analysis on the convergence, which creates room for improvement. However, [91] presented a solution to the Markov decision process (MDP) process, which modeled traffic load pattern, with the help of actor-critic algorithm, a centralized RL approach. They furthermore showed that a prior information about the traffic load of the cellular networks is not necessary [89]. Recently, distributed schemes for dynamic BS switching operation [54, 123, 155] were proven to be more beneficial as they do not require a central controller, but combinatorial problem optimization in a distributed manner demands extensive information exchange and computational complexity. However, all the existing distributed schemes did not present solid theoretic analysis on the convergence, which makes them less appealing. In Chapter 5 of this thesis, we consider a MAB learning policy to solve the dynamic BS switching problem in a centralized manner and also we present theoretical guarantee on the convergence to the optimal solution.

In general, RL algorithms require a long time to converge when the available action set is large which makes them less appealing for BS switching operation. As discussed in [164], traffic load variations in heterogeneous cellular network follow the same temporal and spatial pattern over several days, thus the use of previous learning information about traffic load may be beneficial to make the RL algorithms quickly convergent to the optimal solution compared to naive algorithm which has no information at the beginning. transfer learning (TL) concept, [24, 89, 144], deals with the previous learning experience applied to same or similar problem. Recently, the authors of [89] extended the actor-critic algorithm with a new algorithm named as Transfer Actor CriTic (TACT). The convergence of TACT, in [89], was established via ordinary differential equation and stochastic approximation theory, thus, the optimality of the obtained solution cannot be guaranteed. However, TL concept for MAB problem has not been very well studied in literature, and it is intuitively appealing and cognitive inspired.

2.6 Conclusion

Cognitive radio (CR) has been proposed as an effective approach to achieve energy efficiency in wireless networks. This chapter has reviewed state-of-the-art decision making approaches as applications to dynamic spectrum access, and also energy efficient wireless network.

Within this chapter, we proposed an original classification of the main decision making tools to tackle the CR field. We presented that machine learning and reinforcement learning models tackle the same CR decision making problem; however, the system complexity and available prior knowledge differ. As a result, the selected decision making tools also differ according to the application scenario constraints and requirements. Next, research on transfer learning has been reviewed, which has been shown as a supplement of conventional reinforcement learning approach, to improve the decision making performance. Then, we reviewed the state-of-the-art literature of relevance in the field of simple and efficient sequential decision making framework, referred as multi-armed bandit (MAB) framework. We continued the literature review with the discussion of existing MAB variants and available solution for it.

Next, traditional machine learning and reinforcement learning models and algorithms have been reviewed with their application to energy-efficient channel allocation for single-player and multi-player cases. Furthermore, energy efficiency optimization of wireless networks has been discussed. State-of-the-art works on application of several decision making tools to dynamic network planning based on traffic load patterns has been reviewed as a solution. Before tackling directly the main topics of this work, let's investigate, in the next chapter, an important issue, that will be used in the rest of this work, to evaluate the learning efficiency.

Chapter 3

MAB Performance Evaluation Methodologies

Contents

3.1 Introduction 35
3.2 Performance Evaluation Methodologies 36
3.2.1 Lempel-Ziv (LZ) Complexity 37
3.2.2 Optimal Arm Identification (OI) factor 38
3.3 Verification of LZ and OI Complexity Impact on OSA Scenario . . 39
3.3.1 Classical MAB System Model for OSA 39
3.3.2 UCB1 and Thomson-Sampling Policies 40
3.3.3 Numerical Results 41
3.4 Conclusion 46

3.1 Introduction

Given the description of the research objectives and state-of-the-art literature in the previous chapters, we now present the key performance evaluation methodologies, which will be utilized for next chapter. As seen in Fig. 2.2, decision making process shifts towards reinforcement learning in a case of very limited a priori information and also very simplistic modeling, hence leaving few degrees of freedom to radio equipment to search suitable decision making strategy. Thus, multi-armed bandit (MAB) paradigm can be seen as a perfect answer to CR decision making problem which possesses very limited or almost no a priori information on spectrum band occupancy by primary users (PU).

The performance of MAB policies is influenced by the amount of structure present in the observed rewards, i.e. the transition probability of the Markov process, from which MAB policy is learning [131, 160]. For example, a secondary user (SU) is trying to learn about the probability for a channel to be vacant in OSA scenario, the success of SU implementing MAB policies is strongly affected by the amount of structure obtained in the PUs activity pattern in these channels [100]. In this chapter, we address the following fundamental questions: i) When is it advantageous to apply MAB learning framework? ii) Is considered scenario, in which an SU is learning and predicting, to be declared as an easy or hard one in terms of prediction?, and iii) Is the use of MAB policies for CR decision making problem always justified over a simple non-intelligent approach, i.e. round-robin scheme?

In almost all cases, the performance of MAB learning policies has been studied with respect to stationary distribution of underlying system, e.g. probability that PUs occupy radio channels in OSA context [67, 94]. In certain cases, the underlying system is modeled as an iid (stateless) reward process [67], hence it does not take into account the likely sequential activity patterns. To address the first question raised above, the Lempel-Ziv (LZ) complexity has been introduced in [100] to characterize the sequential pattern for general reinforcement learning (RL) problem. However, MAB framework is a special kind of RL game where a player maximizes its long term reward by making action to learn about the optimal arm, contrarily to general RL framework where a player is interacting with a system by making actions and learns about the underlying structure of the system. As a consequence, in Section 3.2 of this Chapter, we propose the *Optimal Arm Identification (OI) factor* to identify the difficulty associated with prediction of an optimal arm. Finally in Section 3.3, the last question raised above is answered by comparing the performance of MAB policies against the random channel selection (RCS) approach (a non-intelligent approach) in OSA context. Finally, the conclusion is given in Section 3.4.

3.2 Performance Evaluation Methodologies

In general, multi-armed bandit (MAB) algorithms are evaluated with the stationary distribution of the underlying Markov process, i.e. the occupancy of the channels. However, stationary distribution of Markov process alone is not sufficient for evaluating the efficiency of MAB policies. In fact, the performance of MAB policies leverages on the structure of the sequential pattern of state observation and also on the difficulties associated with identification of the optimal arm, i.e. the arm with optimal mean reward distribution. In the case of OSA scenario, the ON/OFF PUs activity models the spectrum usage pattern as depicted in Fig. 3.1. If the separation between the mean reward distribution of the optimal and sub-optimal arms is large, a player should be able to converge to the optimal arm faster, i.e. it achieves a higher number of opportunistic accesses. Therefore, estimating the amount of structure present in the sequential

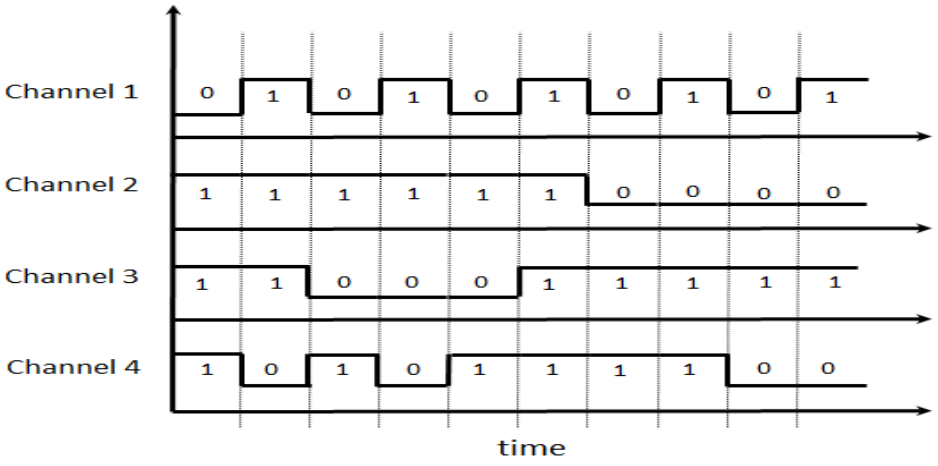


FIGURE 3.1: PUs activity pattern

pattern of the observed states of Markov process is of essential interest for applying machine learning strategies to CR applications.

3.2.1 Lempel-Ziv (LZ) Complexity

Lempel-Ziv (LZ) complexity was proposed by Lempel and Ziv, in [85], which measures the rate of production of new patterns in a sequence. It has been widely adopted in several research areas such as biomedical signal analysis, data compression and pattern recognition. The complexity coefficient a is computed by scanning the sequence and incrementing a every time a new substring of consecutive symbols is found. At time n, the number of different patterns, the measure of complexity, is $a(n)$. In order to obtain a complexity measure which is independent of the sequence length, $a(n)$ must be normalized. If the length of sequence is n and the number of different states is 2, it has been proved that the upper bound of $a(n)$ is given by [1, 85]

$$c(n) < \frac{n}{(1-\epsilon_n)\log_2(n)}$$

where, ϵ_n is small quantity and $\epsilon_n \to 0 (n \to 0)$. In general, $n/\log_2(n)$ is the upper bound of $a(n)$, where the base of logarithm is 2 represents total number of states.

$$\lim_{n\to\infty} a(n) = b(n) = \frac{n}{\log_2(n)}$$

and $a(n)$ can be normalized via $b(n)$:

$$h(n) = \frac{a(n)}{b(n)}$$

where $h(n)$, the normalized LZ complexity, reflects the arising rate of new patterns in the sequence.

LZ complexity is a property of individual sequences and it can be estimated regardless of any assumption about the underlying process that generated the data. In [100], the authors have applied the LZ definition to the production rate of new patterns in Markovian processes. This is of particular interest when the environment, in which CR device is operating, is modeled as a Markov process to evaluate the efficiency of MAB policies.

For an ergodic source, LZ complexity equals to the entropy rate of the source, which for a Markov chain S is given by [100]:

$$h(S) = -\sum_{k,l} \pi_k p_{k,l} \log p_{k,l}, \tag{3.1}$$

where $p_{k,l}$ is the transition probability between state k and l, π_k being the stationary distribution of the Markov chain in state k, i.e. $\pi_k^i(n) = \pi_k^i \ \forall n$. A system with LZ complexity equal to 1 implies very high rate of new patterns production and thus it could make difficult for the learning policy to predict the next sequence. For example in Fig. 3.1, channels 1 to 4 have different PUs activity pattern characterized by the normalized LZ complexity of 0.05, 0.30, 0.60 and 0.66, respectively. It is clear that prediction of next vacancy is an easy task in case of channel 1 which has the lowest LZ complexity, whereas it becomes more and more difficult to predict the next state in channel 4 which has the highest LZ complexity. However, we will see, later in this chapter, that LZ complexity does not correctly predicts the benefits of learning policy in OSA scenario, thus we propose another performance evaluation criteria below.

3.2.2 Optimal Arm Identification (OI) factor

The problem of finding an optimal arm in MAB framework has been studied since the 1950s under the name 'ranking and identification problems' [13, 72]. As stated before, the performance of MAB policy also leverages on the separation between the mean reward distribution of optimal and suboptimal arms. As MAB policy learns to find an optimal arm in terms of reward distribution, thus the gap between optimal and sub-optimal arm significantly affects the convergence behavior of MAB policy. Here in this chapter, we propose another criteria to characterize the difficulty for a MAB to learn the reward distribution of arms.

In recent advances in MAB context, an important focus was set on different perspective, in which each observation is considered as a reward: the user tries to maximize his cumulative reward. Equivalently, its goal is to minimize the expected regret $\Phi^{\mathcal{A}}(n)$, as defined before in (2.3). As stated in [15, 145], regret $\Phi^{\mathcal{A}}(n)$, understood as the reward loss due to the selection of sub-optimal arm with policy $\mathcal{A}$, up to time n is upper bounded uniformly by a logarithmic

function:

$$\Phi^{\mathcal{A}}(n) \leq a \sum_{i:\mu_i<\mu_1} \frac{\log n}{(\mu_1 - \mu_i)} + b \sum_{i:\mu_i<\mu_1} (\mu_1 - \mu_i), \tag{3.2}$$

where μ_i and μ_1 are the mean reward of sub-optimal and optimal arm, respectively. a and b are constants independent from parameters related to the arm's state process, the reward distribution and also the current time n. As stated in (3.2), the upper bound on regret of MAB policies is scaled by the change in mean reward optimality gap $\Delta_i = (\mu_1 - \mu_i)$. Intuitively, decreasing Δ_i makes the upper bound looser and thus increases the uncertainty on MAB policies performance. In this thesis, we propose the OI factor H_{OI} as a measure of difficulty associated with finding an optimal arm among several other arms:

$$H_{OI} = 1 - \sum_{i=1}^{K} \frac{(\mu_1 - \mu_i)}{K}, \tag{3.3}$$

where, K is the number of arms. H_{OI} measures how close the mean reward of all sub-optimal arms are from the mean reward of the optimal arm. If H_{OI} is close to 1 then all arms have very closely distributed mean reward, thus it becomes almost impossible for a learning policy to identify the optimal arm from the set of arms. Proposed OI factor is a simplified version of the complexity measure of regret minimization for generic one-parameter model proposed by Lai and Robbins in [83], and it is more easier to generalize OI factor for other reward laws.

3.3 Verification of LZ and OI Complexity Impact on OSA Scenario

In this section, we study the impact of LZ and OI complexity on the CR learning and decision making with MAB framework applied to OSA problems, which is a special case of dynamic spectrum access (DSA) problem.

3.3.1 Classical MAB System Model for OSA

We consider a network with a secondary transceiver pair (Tx-Rx) and a set of channel $\mathcal{K} = \{1, \cdots, K\}$. The SU can access one of the K channels if it is not occupied by PUs. The i-th channel is modeled by an irreducible and aperiodic discrete time Markov chain with finite state space S^i. $P^i = \left\{p^i_{kl}, (k, l \in \{0, 1\})\right\}$ denotes the state transition probability matrix of the i-th channel, where 0 and 1 are the Markov states, i.e. occupied and free respectively. Let, $\boldsymbol{\pi}^i$ be

the stationary distribution of the Markov chain, and it is defined as 2-state Markov chain:

$$\boldsymbol{\pi}^i = [\pi_0^i, \pi_1^i] = \left[\frac{p_{10}^i}{p_{10}^i + p_{01}^i}, \frac{p_{01}^i}{p_{10}^i + p_{01}^i}\right]. \tag{3.4}$$

$S^i(n)$ being the state of the channel i at time n and $r_q^i(n) \in \mathbb{R}$ is the positive reward associated to the channel i in state q at time n. Without loss of generality we can assume, that $r_q^i(n) = S^i(n)$, i.e. $S^i(n) = 1$ if sensed free and $S^i(n) = 0$ if sensed occupied. The stationary mean reward μ_i of the i-th channel under stationary distribution π^i is given by: $\mu_i = \pi^i$. A channel with μ_1, such that $\mu_1 > \mu_i$, is an optimal channel. The mean reward optimality gap is defined as $\Delta_i = \mu_1 - \mu_i$.

3.3.2 UCB1 and Thomson-Sampling Policies

We consider two different reinforcement learning (RL) strategies, i.e. UCB1 and Thomson-Sampling (TS), in order to evaluate the learning efficiency of MAB policies on channel set containing different PUs activity patterns. These policies are based on RL algorithms introduced in [3, 15, 136] as approaches to solve MAB problems and they attempt to identify the most vacant channel in order to maximize their long term reward. In figure 3.1, all channels do not have the same occupancy ratio and it seems intuitively clear that the more different the channel occupations are, the more beneficial learning approach can provide compared to non-intelligent approaches.

Upper Confidence Bound (UCB) Policy It has been shown previously in [67, 94] and in Algorithm 1 that UCB1 allows spectrum learning and decision making in OSA context in order to increase the transmission opportunities. UCB1 learns about the optimal channel from previously observed rewards starting from scratch, i.e. without any a priori knowledge of the activity within the set of channels. For each time n, UCB1 policy updates indices named as $B^i(n, T^i(n))$, where $T^i(n)$ is the number of times the i-th channel has been sensed up to time n. UCB1 returns the channel index $\mathcal{A}(n) = i$ of the maximum index. UCB1 has been detailed in Algorithm 1 in the previous chapter where α is the exploration-exploitation coefficient. If α increases, the bias $A^i(n, T^i(n))$ increases and UCB1 policy explores new channels. Otherwise, if α decreases, the index computation is more governed by $\bar{S}^i(T^i(n))$ and the policy tends to exploit the previously observed optimal channel.

Thomson-Sampling (TS) Policy introduced in [3, 136] and detailed in Algorithm 2, selects a channel having the highest $B^i(n, T^i(n))$ index, sampled from a β probability density function w.r.t. two arguments, $W^i(n, T^i(n)) = \sum_{m=0}^{n-1} S^i(m)\mathbb{1}_{\mathcal{A}(m)=i}$ and $Z^i(n, T^i(n)) =$

$T^i(n) - W^i(n, T^i(n))$, where $T^i(n)$ has the same meaning than previously. The former argument is the total number of free state observed up to time n from channel i and the latter is the total number of occupied state. For start, no prior knowledge on the mean reward of each channel is assumed (uniform distribution) and hence the index for all channels is set to $\beta(1,1)$. TS policy updates the distribution on mean reward μ_i as $\beta\left(W^i(n, T^i(n)) + 1, Z^i(n, T^i(n)) + 1\right)$.

Algorithm 2 Thomson-Sampling (TS) policy

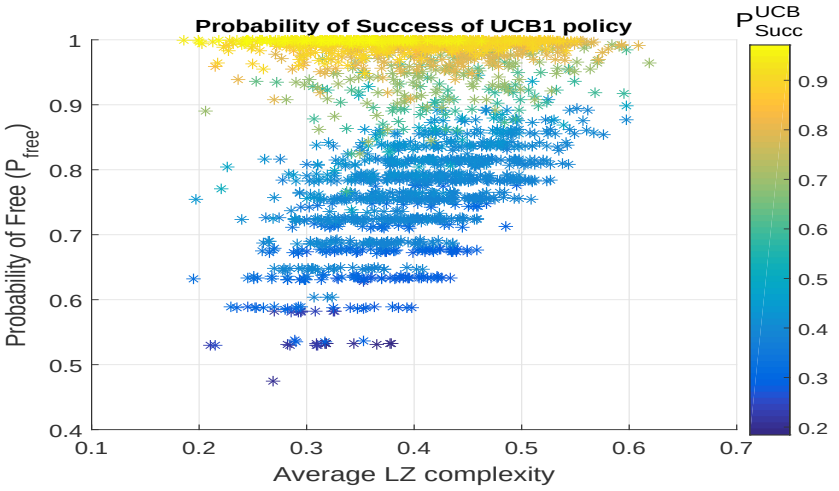
Input: K, $W^i(1) = 0$, $Z^i(1) = 0$
Output: $\mathcal{A}(n)$
1: **for** $n = 1$ **to** N **do**
2: $B^i(n, T_i(n)) =$ draw a sample from $\beta(W^i(n, T^i(n)) + 1, Z^i(n, T^i(n)) + 1)$
3: Sense channel $\mathcal{A}(n) = \arg\max_i \left(B^i(n, T_i(n))\right)$
4: Observe state $S^i(n)$
5: $T^i(n) = \sum_{m=0}^{n-1} \mathbb{1}_{\mathcal{A}(m)=i}, \forall i, \qquad W^i(n, T^i(n)) = \sum_{m=0}^{n-1} S^i(m)\mathbb{1}_{\mathcal{A}(m)=i}, \forall i$
6: $Z^i(n, T^i(n)) = T^i(n) - W^i(n, T^i(n)), \forall i$
7: **end for**

3.3.3 Numerical Results

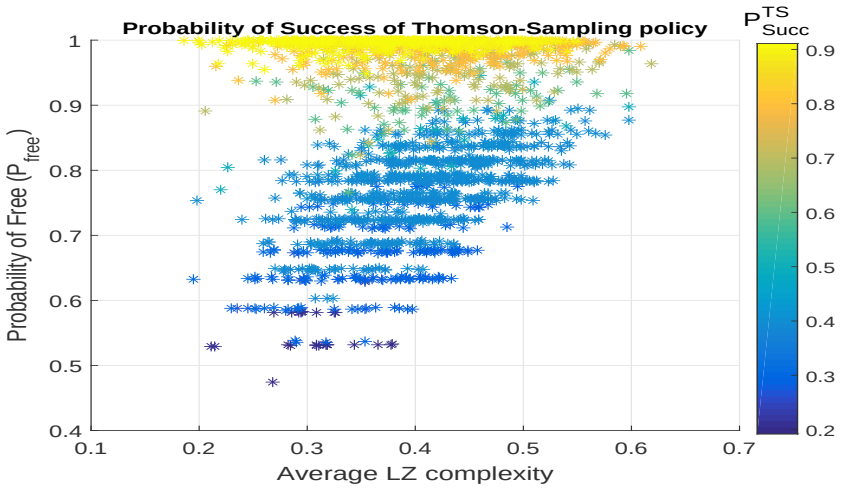
The two MAB policies, UCB1 and Thomson-Sampling (TS), are investigated and the performance they achieve are put in correlation with the information given by the LZ complexity and the OI factor H_{OI}. Markov chains with several levels of stationary distribution $\boldsymbol{\pi} = [\pi_0, \pi_1]$, LZ complexity and H_{OI} factor are generated for further numerical analysis. For simulation convenience, some parameters need to be set. Indeed, (3.4) is an undetermined system with two unknowns p_{01} and p_{10}. Therefore, as a side step, we considered 9 different levels of π_1, i.e. the probability of being vacant, as $0.1, 0.2, \cdots, 0.9$. For these values of π_1, we obtained 45 different transition probability matrices P, each corresponding to different LZ complexity. A total of $\binom{45}{5}$ combinations are obtained by considering $K = 5$ channels, and those correspond to various H_{OI} factor. Finally, MAB policies are applied to randomly selected 2000 combinations from a total of $\binom{45}{5}$ combinations. Every point in each figure corresponds to one instance of the MAB policy. For each instance, policy is executed over 10^2 iterations of 10^4 time slots each. Moreover, the exploitation-exploration coefficient in UCB1 is set to $\alpha = 0.5$ which has been proved to be efficient for maintaining a good tradeoff between exploration and exploitation [106].

Probability of Success

Probability of success P_{Succ} is computed by considering the number of times each vacant channel is explored over the number of iterations. The success probability depends on the probability



(a) UCB1



(b) Thomson-Sampling

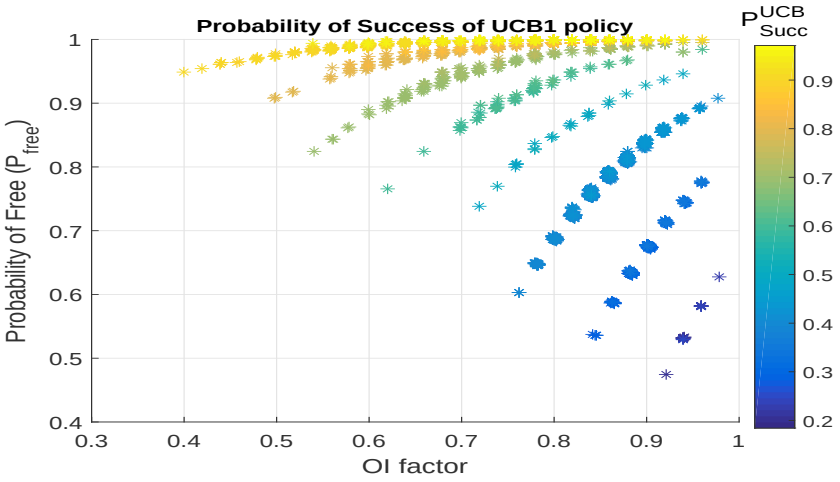
FIGURE 3.2: (a), (b) Probability of success P_{Succ} of MAB policies, i.e. UCB1 and TS, with respect to the average LZ complexity and the probability of free P_{free}. Each point denotes a particular instance of MAB policies applied to $K = 5$ channels. The number of random combinations which we analyzed is 2000.

P_{free} that there exists at least one free channel from a set of K channels. Considering that the channel occupation is independent from one channel to another, we have [100]:

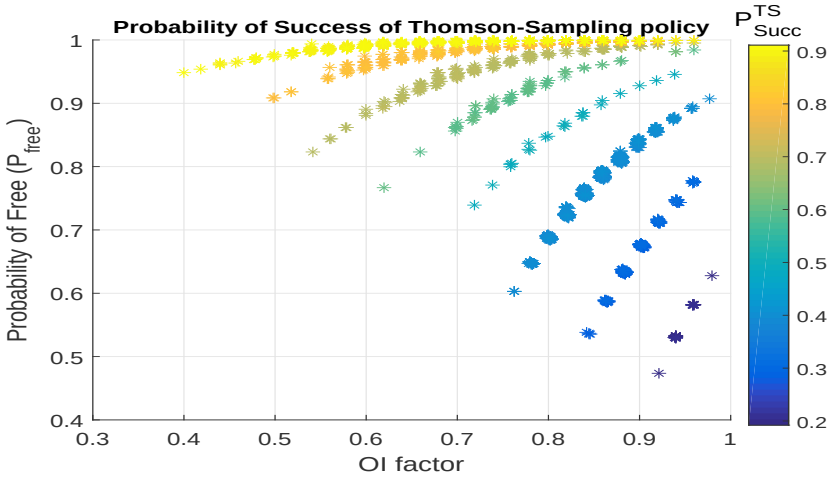
$$P_{\text{free}} = 1 - \prod_{i=1}^{K} \pi_0^i, \tag{3.5}$$

where π_0^i is the probability that the i-th channel is occupied.

Figs. 3.2(a) and 3.2(b) depict the probability of success P_{Succ} of UCB1 and TS policies, that means the probability that these policies access to a free channel, according to the probability that at least one channel is free, i.e. P_{free} and LZ complexity. In both figures, the success probability increases with P_{free} for a given level of LZ complexity. However, in Figs. 3.2(a) and



(a) UCB1

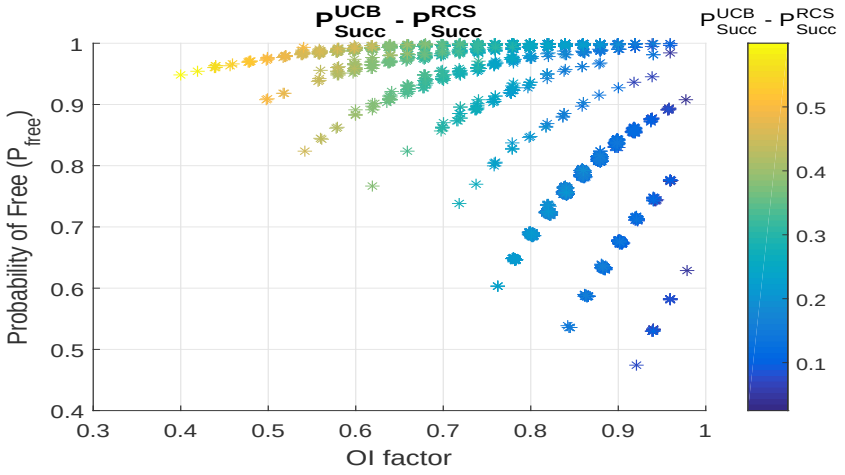


(b) Thomson-Sampling

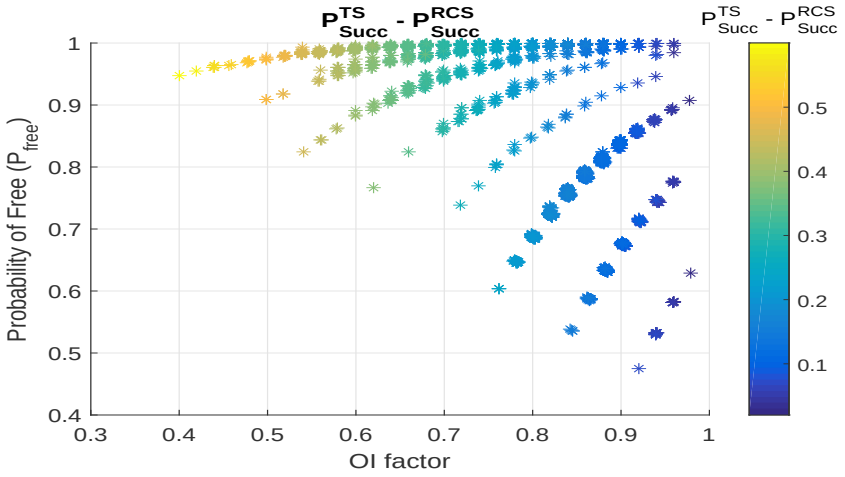
FIGURE 3.3: (a), (b) Probability of success P_{Succ} of MAB policies, i.e. UCB1 and TS, with respect to the OI factor H_{OI} and the probability of free P_{free} applied to same ensemble of channels.

3.2(b), for a given P_{free}, several values of LZ complexity lead to the same level of performance for UCB1 and TS algorithms. This reveals that the LZ complexity is not really related to the ability of UCB1 and TS policies to learn the scenario. For instance in Figs. 3.2(a) and 3.2(b), an SU is able to achieve more than 90% of probability of success on a channel set with LZ complexity of 0.2 and $P_{\text{free}} = 0.98$, whereas it only achieves 75% of probability of success on a channel set with LZ complexity of 0.6 and $P_{\text{free}} = 0.98$. In that case, the variation in the probability of success is up to 15% however the variation of P_{Succ} along the x-axis can be even less important for lower values of P_{free}.

On the other hand, Figs. 3.3(a) and 3.3(b) show the probability of success of UCB1 and TS policies according to OI factor H_{OI} and the probability of free P_{free}. As we can see H_{OI} is highly correlated to the performance of UCB1 and TS policies on a given scenario. In order to achieve



(a) UCB1



(b) Thomson-Sampling

FIGURE 3.4: (a), (b) Each point denotes the difference between the probability of success of MAB policies, i.e. UCB1 and TS, and the probability of success of the random channel selection (RCS) approach applied to $K = 5$ channels, respectively. The number of random combinations which we analyzed is 2000.

very high level of P_{Succ}, H_{OI} is required to be low. For instance in Figs. 3.3(a) and 3.3(b), an SU is able to achieve more than 90% of P_{Succ} on a channel set when H_{OI} is 0.4 and $P_{\text{free}} = 0.95$, whereas it only achieves 50% of P_{Succ} on a channel set when $H_{OI} = 0.95$ and $P_{\text{free}} = 0.95$. Thus, we can state that $P_{\text{Succ}}^{\text{UCB}}$ varies up to 40% according to the changes in H_{OI}, along x-axis, for certain values of P_{free}.

Comparison with Random channel selection policy

Figs. 3.4(a) and 3.4(b) compare the probability of success of MAB policies, i.e. UCB1 and TS, and random channel selection (RCS) approach. As expected, MAB policies outperform RCS approach in general, but difference becomes negligible for very high H_{OI} regime, i.e. the mean

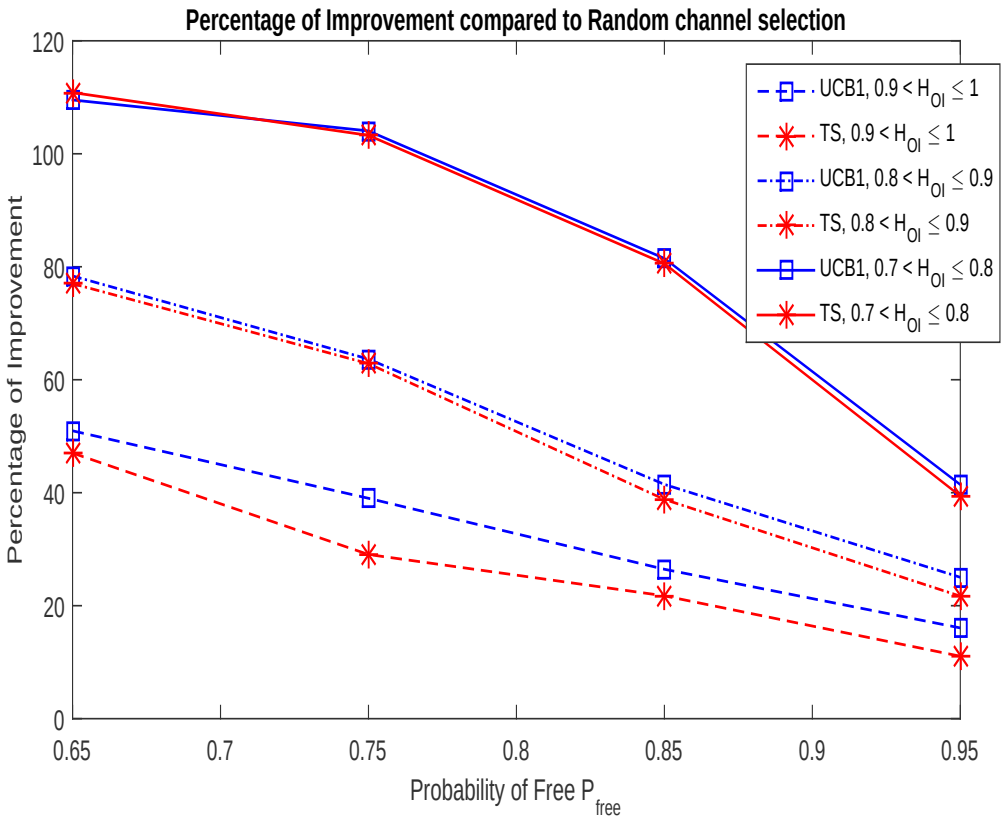


FIGURE 3.5: Average percentage of improvement in the probability of success of MAB policies, i.e. UCB1 and TS, with respect to RCS policy as a function of the OI factor H_{OI} and the probability of free P_{free}. Each point denotes an average percentage of improvement achieved by MAB policies for different H_{OI} and P_{free}.

reward of sub-optimal and optimal channels becomes equivalent. For a given P_{free}, performance of MAB policies decreases when H_{OI} increases. For instance in Figs. 3.4(a) and 3.4(b), we can notice that $P_{\text{Succ}}^{\text{UCB}} - P_{\text{Succ}}^{RCS}$ and $P_{\text{Succ}}^{\text{TS}} - P_{\text{Succ}}^{RCS}$ vary up to 50% along the x-axis with respect to the OI factor for fixed P_{free}. However, for lower values of P_{free}, e.g. more channels are occupied, performance gain achievable with MAB learning policies with respect to RCS approach reduces, thus makes learning ineffective.

Fig. 3.5 shows the average percentage of improvement in the probability of success achieved by MAB policies, i.e. UCB1 and TS, with respect to RCS approach under various PUs activity pattern. As we stated before, the percentage of improvement of MAB policies compared to RCS approach decreases when P_{free} increases, because the RCS approach is able to find more opportunities in high P_{free} regime. On the contrary, the average percentage of improvement of MAB policies also decreases when P_{free} decreases after a certain limit. It is due to the fact that there are not so many opportunities available to exploit for MAB policies in low P_{free} regime. As stated in Fig. 3.5, combinations with medium H_{OI}, i.e. $0.7 < H_{OI} \leq 0.8$, increases the percentage of improvement of MAB policies compared to the RCS approach. Even for high H_{OI}, i.e. $0.9 < H_{OI} \leq 1$, the relative improvement of learning policies is still noticeable, i.e.

more than 15%. It also reveals that all MAB policies, UCB1 and TS, achieve nearly the same level of percentage of improvement for medium H_{OI}, i.e. $0.7 < H_{OI} \leq 0.8$, whereas in case of high H_{OI}, i.e. $0.9 < H_{OI} \leq 1$, UCB1 policy significantly outperforms TS policy. Figs. 3.4 and 3.5 prove that OI factor H_{OI} is rather well suitable compared to LZ complexity to analyze the learning capability of MAB policies in OSA context.

Finally, we are able to answer the questions raised in the beginning of this chapter. We found out that, for several spectrum utilization patterns, MAB policies can be beneficial compared to non-intelligent approach, but the percentage of improvement is highly correlated with the level of OI factor and very little affected by the level of LZ complexity. Furthermore, OI factor proposed in this chapter, can also be seen as a measure to characterize the scenario, e.g. easy or hard scenario for learning and prediction, which has not been discussed in literature before.

3.4 Conclusion

In this chapter, we proposed a new criterium to evaluate the *a priori* performance of two MAB policies on a given scenario. We found out that, in diverse structure of the state observation from Markov process, the application of MAB policies can be beneficial compared to non-intelligent approach, but the percentage of improvement is highly correlated with the level of OI factor and very little affected by the level of LZ complexity. This results do not just emphasize aphorism that performance of MAB policies extremely depends on the OI factor. It also shows that LZ complexity alone is not a very effective measure for analyzing performance of the MAB policies. On the contrary, it also shows the importance of evaluating the efficiency of MAB policies with respect to the OI factor associated with optimal arm identification, a measure that has been overlooked in the previous research. While MAB policies, e.g. UCB1 and Thomson-Sampling for instance, are often assumed to be beneficial in OSA context, the problem of characterizing the scenarios where they are effective is barely studied. We evaluated the performance of UCB1 and Thomson-Sampling on various scenarios, and correlated this to the output of OI factor and LZ complexity.

Chapter 4

A New Single-player and Multi-player CR Learning Algorithms for OSA

Contents

4.1 Introduction 48
4.2 Problem Formulation 48
4.2.1 Markovian MAB System Model 48
4.2.2 Wireless Network Model 50
4.3 Novel Single-player and Multi-player Policy Design 54
4.3.1 Single-player Rested Bandit Policy: QoS-UCB 54
4.3.2 Single-player Restless Bandit Policy: RQoS-UCB 56
4.3.3 Multi-player Restless Bandit Policy: Distributed RQoS-UCB 57
4.4 Regret Analysis 59
4.4.1 Preliminaries for Regret Analysis 59
4.4.2 Regret of Single-player Policy 62
4.4.3 Regret of Multi-player Policy 64
4.5 Numerical Results and Analysis 67
4.5.1 Simulation Settings 67
4.5.2 Simulation Results and Discussions 70
4.6 Conclusion 77

4.1 Introduction

As mentioned in the previous chapters, MAB approach is particularly well adapted to tackle the OSA problem due to its low implementation complexity [67]. Indeed in [67] which preceded this work, each unlicensed or secondary user (SU) senses, at each time step, one of the K channels available in the primary users (PUs) spectrum, according to a given policy, and searches for learning the optimal channel given a certain criterium, e.g. availability. In this chapter, we consider that PUs activity is modeled with Markov process compared to iid in [67]. In this chapter, we propose a novel QoS-UCB policy for the Markovian MAB setting, which takes into account not only the availability but also the channel quality for rating a channel. This new metric allows not only to opportunistically use the spectrum holes but also maximizing the data rate achieved by the unlicensed users.

Later on in this chapter, we consider a distributed framework with several SUs where there is no information exchange or prior agreement among the different SUs, thus it introduces additional challenges such as: loss in the collected reward due to collisions among the different SUs trying to access the same channel, and also competition among different SUs since they all sense and access a channel with the highest reward in the long run. The analytical studies on the regret bound for both cases, single-player and multi-player, are provided and discussed.

The organization of this chapter is as follows. Section 4.2 introduces the problem definitions and notations. In Section 4.3, the new RQoS-UCB algorithm taking into account channel availability and quality is presented for the single-player Markov MAB problem and then this section also extends it to the distributed multi-player case. In Section 4.4, regret law has been derived for single-player and multi-player cases. Numerical results are presented in Section 4.5, which validate the efficiency of the proposed distributed RQoS-UCB policy compared to the state-of-the-art algorithms under several scenarios identified with different performance criteria, LZ complexity, OI factor, introduced in previous Chapter 3. Finally, Section 4.6 concludes the chapter.

4.2 Problem Formulation

4.2.1 Markovian MAB System Model

We consider Markovian MAB framework with $U \geq 1$ SUs which opportunistically access the spectrum, and K PUs' channels. We consider that each SU can sense only one channel in each time slot. Moreover, SUs operate in a completely uncorrelated manner w.r.t. primary users, hence the actions of SU do not affect the PUs policy. At time n, the CR device $j \in \mathcal{U}$ senses

channel $i \in \mathcal{K}$, where $\mathcal{U} = \{1, \cdots, U\}$ and $\mathcal{K} = \{1, \cdots, K\}$ define the set of SUs and channels respectively.

In classical Markov MAB problem formulation, playing a slot-machine presents single fixed positive reward dependent on observed state, whereas special case of Markovian MAB model considered in this thesis assumes that another positive random reward with a fixed iid distribution is drawn when a slot-machine is activated.

Furthermore, the band quality is rated according to the interference temperature recorded on it. We assume that the band quality in a given state is stationary in the wide sense, meaning that its statistical properties, i.e. first and second moment, are not evolving over time, but the instantaneous value $R_q^{i,j}(n)$ may vary. $G_q^{i,j}(T^{i,j}(n)) = \frac{1}{T^{i,j}(n)} \sum_{k=1}^{T^{i,j}(n)} R_q^{i,j}(k)$ denotes the empirical mean of quality observations $R_q^{i,j}$ collected from band i by SU j in state q and $T^{i,j}(n)$ denotes the total number of times band i has been sensed up to time n by SU j. The global mean reward, i.e. taking into account the quality as well as the state of each band i, is defined as:

$$\mu_{i,j}^R = \sum_{q \in S^{i,j}} G_q^{i,j} r_q^{i,j} \pi_q^i. \tag{4.1}$$

Without loss of generality, let us consider that $\mu_{1,j}^R > \mu_{i,j}^R > \mu_{K,j}^R$, $\forall\ i \in \{2, \cdots, K-1\}$. It is important to note that the optimal bands are the ones having the highest global mean reward, i.e. $\{\mu_{i,j}^R\}_{\forall i,j \in \mathcal{U}}$. The global mean reward can be seen as the expectation of the reward function of the user j in band i, i.e. $G_q^{i,j} r_q^{i,j}$. This function can be seen as a weighting of the channel availability reward, i.e. $r_q^{i,j}$, by a random variable reflecting its quality, i.e. $G_q^{i,j}$. Other combinations of quality and availability might be envisaged but fall out of the scope of this work and are left for further works.

Let ALOHA-like protocol be considered among SUs operating in network under which if two or more SUs transmit in the same channel then none of the transmissions are successful, and no collision avoidance mechanisms are considered[1]. $C_o^j(i,n)$ is the indicator of collision at the $n-$th slot at channel i for SU j. At the end of each slot n, each SU j receives reward $r_q^{i,j}(n)$ and $R_q^{i,j}(n)$. Under this model, we are interested in designing a policy $\mathcal{A}$, maximizing the expected number of successful transmissions with better quality in the long run. Let $\Phi^{\mathcal{A}}(n)$ be the regret and defined as the reward loss after n slots for U SUs and policy $\mathcal{A}$. In the ideal scenario, we assume that the channel mean reward statistics μ_i^R are known a priori by a central agent and it selects U optimal channels for U SUs.

[1]The effect of employing CSMA-CA is not taken into an account here although the use of CSMA-CA increases spectrum usage and consecutively should decrease the regret.

We are interested in minimizing the *regret* $\Phi^{\mathcal{A}}(n)$ associated with the learning and access scheme, defined as:

$$\Phi^{\mathcal{A}}(n) = \sum_{j=1}^{U} n\mu_j^R - \sum_{j=1}^{U} \mathbb{E}\left[\sum_{t=1}^{n} G_{q_{\mathcal{A}(t,j)}}^{\mathcal{A}(t,j)}(t) r_{q_{\mathcal{A}(t,j)}}^{\mathcal{A}(t,j)}(t)\right] \tag{4.2}$$

where the expectation $\mathbb{E}$ is taken over the states and qualities. Let $q_{\mathcal{A}(t,j)}$ being the state observed by SU j under the policy $\mathcal{A}$ at time t. Frequency bands whose mean reward is strictly less than $\{\mu_j^R\}_{\forall j \in \mathcal{U}}$ are referred as suboptimal frequency bands[2].

4.2.2 Wireless Network Model

A centralized primary network is considered where a radio access point serves K channels as in Fig. 4.1. In the same cell, a Tx-Rx pair of cognitive SU is considered, i.e. an SU Tx transmitting opportunistically to an SU Rx. The CR system seeks to interweave the SU's signals with the PUs transmissions in the set of channels. Due to the frequency reuse factor and partial utilization of the channel i by PUs in the neighboring cells, the interference level is not the same for all channels which leads to a varying quality according to the channel considered. This noise process is considered as stationary in wide sense. SU transmissions on adjacent channels cause the mutual interference to PUs and similarly PUs transmissions cause mutual interference to SUs. We assume that the mutual interference from the PU's transmission can be estimated perfectly at the SU receiver.

A slotted frame structure for CR is considered as in Fig. 4.2. During the frame duration, i.e. T_{total}, SU senses the channel during T_{sens}, learns during T_{lear}, and transmits (or not depending on the result of the channel sensing) during T_{trans}. The aim of a decision making policy is to decide which channel should be explored in the next time slot and can be implemented in parallel with transmission. Moreover, it requires a very small time compared to sensing and transmission duration [106].

At the current time slot, SU performs spectrum sensing on the $i-$th channel and utilizes it for communication only when there is no PU. The spectrum sensing part is error-prone but the imperfect sensing has negligible effect on the learning process as discussed in [69]. The discrete received signal at SU Rx can be written according to both hypothesis: i.e. $\mathcal{H}_0$ channel i is used by a PU and $\mathcal{H}_1$ channel i is vacant:

$$\mathcal{H}_0 : y_{q_0}^i[m] = p^i[m] + u^i[m], \tag{4.3}$$

$$\mathcal{H}_1 : y_{q_1}^i[m] = u^i[m] \tag{4.4}$$

[2]For notation simplicity, we refer μ_i^R instead of $\mu_{i,j}^R$, because the set of channels is fixed *a priori* for all SUs operating in CR network.

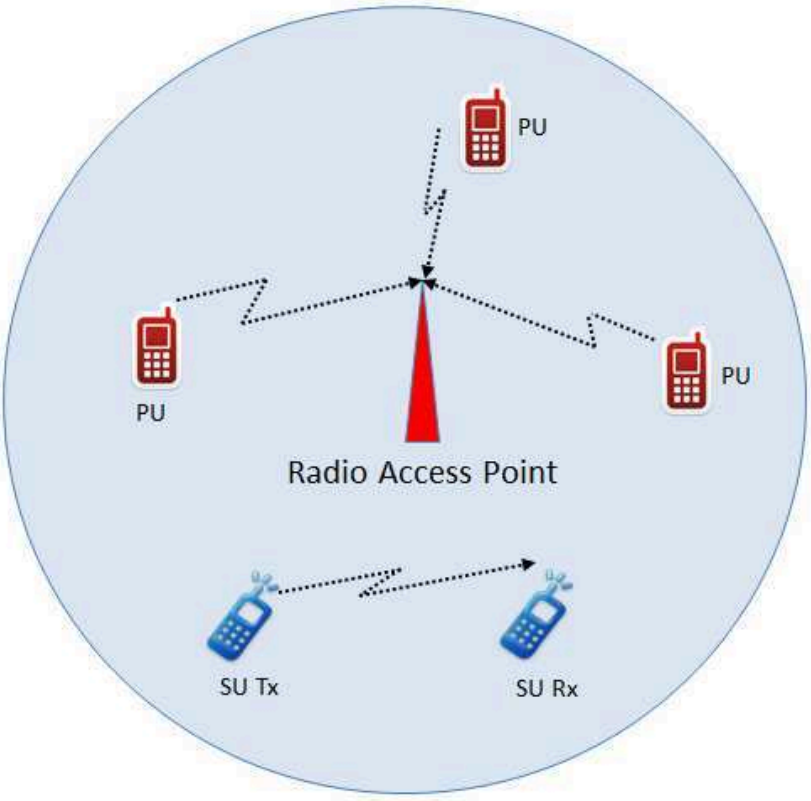


FIGURE 4.1: Interweave cognitive cell scenario

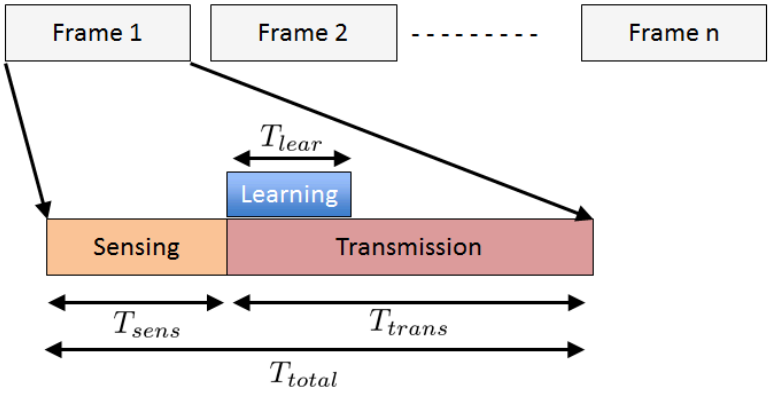


FIGURE 4.2: Cognitive radio frame structure with repetitive sensing, learning and transmission slots

where, $p^i[m]$ and $u^i[m]$ are the signal and noise component respectively for the $i-$th channel. The PU signal $p^i[m]$ is zero mean iid random process with variance $\mathbb{E}[\|p^i[m]\|^2] = \sigma^2_{p,i}$. The noise components $u^i[m]$ are assumed to be zero mean and complex Gaussian distributed with variance $\mathbb{E}[\|u^i[m]\|^2] = \sigma^2_{u,i}$ and independent from the primary users' signal p^i. We remind that $u^i[m]$ counts for the other cells interference and background noise. Theoretically, SU should transmit only when no PU occupies channel i. However, it may miss the detection of a primary user and transmits anyway. Under both hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$, the received signal at SU Rx when SU Tx transmits is

$$\mathcal{H}_0 : z^i_{q_0}[m] = s^i[m] + p^i[m] + u^i[m], \text{ under } P_{\text{md}} \tag{4.5}$$

$$\mathcal{H}_1 : z^i_{q_1}[m] = s^i[m] + u^i[m], \text{ under } 1 - P_f \tag{4.6}$$

where $s^i[m]$ being the SU signal, zero mean and iid with variance $\mathbb{E}[\|s^i[m]\|^2] = \sigma^2_{s,i}$. P_{md} in (4.5) means the probability of miss detection and P_f in (4.6) is the false alarm probability and will be given in the upcoming section. If a miss detection occurs in the $i-$th channel, SU transmits under SINR $\Gamma^i_0 = \frac{\sigma^2_{s,i}}{\sigma^2_{p,i}+\sigma^2_{u,i}}$. On the other hand, if channel i is available, SU transmits under SINR $\Gamma^i_1 = \frac{\sigma^2_{s,i}}{\sigma^2_{u,i}}$.

Energy Detector (ED) a state and quality information metric

Let's consider in the following chapter that sensing is performed by an ED with no loss of generality, but any other kind of spectrum sensing technique could also be used. ED senses channel i and measures a power level. If the measured power level is above a certain threshold υ, ED decides the channel is occupied and if the power level is below υ the channel is decided to be available. Moreover, in our work, the measured spectrum level is recorded and used as a quality information metric for RQoS-UCB policy. This is a special case here where both vacancy and quality are given by the same sensor (ED), whereas this could be different with separate sensor for both criteria in other scenarios. If the sensed channel is detected as free, the interference plus noise power level, induced by the dynamic use of the channel by PUs of neighboring cells and/or dynamic use of adjacent channel by PUs, is used to rate the quality of the channel. Let F_s being the sampling frequency at the receiver, then N_s is the number of samples during the sensing phase, where $N_s = T_{\text{sens}} F_s$. A power measure of the $i-$th channel is

$$\mathcal{P}^i_q = \frac{1}{N_s} \sum_{m=1}^{N_s} \|y^i_q[m]\|^2.$$

Using the central limit theorem (CLT), the distribution of the ED test statistic can be accurately approximated with a normal distribution for a sufficiently large number of samples, i.e. when $N_s \to \infty$. Hence, the false alarm probability $P_f(\upsilon, N_s)$ under the threshold υ and number of

samples N_s is given as [19, 92]:

$$P_f(\upsilon, N_s) = Q\left(\left(\frac{\upsilon}{\sigma_{u,i}^2} - 1\right)\sqrt{N_s}\right)$$

where $Q(\cdot)$ being the Gaussian Q-function. The channel with the highest quality has the lowest interference plus noise power level. Hence, the quality information metric (or reward associated with quality) should be inversely proportional to the ED output:

$$R_q^i = \frac{1}{\mathcal{P}_q^i + c_e} \tag{4.7}$$

where c_e is a constant added to the received power in order to avoid taking the inverse of very small numbers.

Achievable Throughput Analysis

The SU's average achievable throughput Ξ^i in the $i-$th channel is given by the sum of the achievable throughput under $\mathcal{H}_0$ and the achievable throughput under $\mathcal{H}_1$, which can be written as:

$$\Xi^i(\upsilon, N_s) = \Xi_{q_1}^i(\upsilon, N_s) + \Xi_{q_0}^i(\upsilon, N_s), \tag{4.8}$$

where $\Xi_{q_1}^i$ and $\Xi_{q_0}^i$ are defined as:

$$\Xi_{q_1}^i(\upsilon, N_s) = \frac{T_{\text{total}} - T_{\text{sens}}}{T_{\text{total}}}\pi_{q_1}^i C\left(\Gamma_1^i\right)\left(1 - P_f(\upsilon, N_s)\right) \tag{4.9}$$

$$\Xi_{q_0}^i(\upsilon, N_s) = \frac{T_{\text{total}} - T_{\text{sens}}}{T_{\text{total}}}\pi_{q_0}^i C\left(\Gamma_0^i\right)\left(1 - P_d(\upsilon, N_s)\right) \tag{4.10}$$

where $C(\Gamma)$ denotes the achievable capacity with SINR Γ, i.e. $C(\Gamma) = \log_2(1+\Gamma)$. Hence, $C\left(\Gamma_1^i\right)$ is the achievable rate in the $i-$th channel without any PU and $C\left(\Gamma_0^i\right)$ is the achievable rate in the $i-$th channel when a PU has not been detected. The achievable throughput under hypothesis $\mathcal{H}_1$ is hence $C\left(\Gamma_1^i\right)$ multiplied by the probability channel i is available, i.e. $\pi_{q_1}^i$, and the probability not to generate a false alarm, i.e. $1 - P_f(\upsilon, N_s)$. On the other hand, the achievable throughput under $\mathcal{H}_0$ is $C\left(\Gamma_0^i\right)$ weighted by the probability channel i is occupied, i.e. $\pi_{q_0}^i$ and the miss detection probability i.e. $P_{\text{md}}(\upsilon, N_s) = 1 - P_d(\upsilon, N_s)$ where $P_d(\upsilon, N_s)$ is the probability of correct detection. Both rates are weighted by the effective transmission time ratio, i.e. $\frac{T_{\text{total}} - T_{\text{sens}}}{T_{\text{total}}}$

4.3 Novel Single-player and Multi-player Policy Design

The motivation of this section is to design efficient allocation rules for the modified rested and restless Markovian MAB model in single-player and multi-player cases, as discussed in the problem formulation. A straightforward way to tackle the modified Markovian model, introduced above, is to consider the reward associated with the state and quality jointly, and apply UCB1, [145], regenerative cycle algorithm (RCA), [48, 146], or restless UCB (RUCB), [94] policies. In this thesis, we assume that channel set consists in different transmission quality, and also a CR may have certain preferences, i.e., quality or availability or both, for the channel selection. The list of notations used for the single and multi-player algorithms is available at the beginning of the manuscript, where the dependence on index j vanishes for single-player policies.

4.3.1 Single-player Rested Bandit Policy: QoS-UCB

We propose a *quality of service upper confidence bound (QoS-UCB)* policy that aims at finding a channel which is optimal in terms of both vacancy and quality in OSA scenario modeled with rested Markovian MAB. Rested Markovian MAB assumes that the arm which is in passive mode (or not activated by a player) remains frozen and only the arm in an active mode (or activated by a player) evolves. Although, rested framework is quite dissimilar from the way current wireless network evolves, it presents however a good basis for obtaining analytical results for more complicated models, such as restless and multi-player models. One of the contribution of this chapter is stated in Algorithm 3.

QoS-UCB policy computes for each channel i, the index $B^i(n, T^i(n))$ corresponding to the score of the $i-$th channel at time n. At time n, Algorithm 3 returns the channel index $\mathcal{A}(n)$ which has to be sensed in the next time slot. $T^i(n)$ defines the number of times channel i has been sensed up to time n. Initially, all the channels are sensed at least once, in order to acquire some statistics about them, Steps 1-3. After $n > K$ iterations, the indexes $B^i(n, T^i(n))\ \forall\ i \in \mathcal{K}$ are updated, step 5, as:

$$B^i(n, T^i(n)) = \bar{S}^i(T^i(n)) - Q^i(n, T^i(n)) + A^i(n, T^i(n)),\ \ \forall i \tag{4.11}$$

where $\bar{S}^i(T^i(n))$ being the empirical mean of the states of the $i-$th channel (occupied or free) at time n, defined as:

$$\bar{S}^i(T^i(n)) = \frac{S^i(1) + S^i(2) + \cdots + S^i(T^i(n))}{T^i(n)}, \forall i. \tag{4.12}$$

Algorithm 3 QoS-UCB policy

Input: α, β, $\mathcal{A}(0)$, $T^i(0) = 0$, $R^i_{q_1}(0)\ \forall\ i \in \mathcal{K}$.
Output: $\mathcal{A}(n+1)$

for $n = 1$ **to** K **do**
  Initialize policy by sensing each channel for at least one time.
end for
while $n > K$ **do**
  $B^i(n, T^i(n)) = \bar{S}^i(T^i(n)) - Q^i(n, T^i(n)) + A^i(n, T^i(n)),\ \forall i$
  $\mathcal{A}(n) = \arg\max_i B^i(n, T^i(n))$,
  Sense channel $i = \mathcal{A}(n)$ and Observe current state $S^i(n)$
  if $\mathcal{A}(n)$ free **then**
    Transmit and Observe quality $R^i_{q_1}(n)$
  else
    Do Not Transmit and $R^i_{q_1}(n) = R^i_{q_1}(n-1)$
  end if
  Update $T^i(n) = \sum_{t=1}^{n} \mathbf{1}_{\mathcal{A}(t)=i}$
  Update $\bar{S}^i(T^i(n))$, $Q^i(n, T^i(n))$ and $A^i(n, T^i(n))$ according to (4.12), (4.13) and (4.14)
end while
Update $T^i(n) \leftarrow T^i(n) + 1$ and $n \leftarrow n + 1$

The second term, i.e. $Q^i(n, T^i(n))$, represents the quality term and is computed with the observed instantaneous quality $R^i_{q_1}(n)$ in state q_1. Finally, if the scheme leads to a channel which is occupied, then the third bias term $A^i(n, T^i(n))$ forces to explore the other channels. $\bar{S}^i(T^i(n))$ is the exploitation contribution, $Q^i(n, T^i(n))$ maintains exploration and exploitation, and finally $A^i(n, T^i(n))$ represents only exploration contribution.

An important contribution of this thesis compared to traditional UCB1 policies is the term $Q^i(n, T^i(n))$ which defines the quality information of the channel i and which can be written as:

$$Q^i(n, T^i(n)) = \frac{\beta M^i(n, T^i(n)) \log(n)}{T^i(n)},\ \ \forall i \tag{4.13}$$

where,

$$M^i(n, T^i(n)) = G^{q_1}_{\max}(n) - G^i_{q_1}(T^i(n)),\ \ \forall i$$

and $G^i_{q_1}(T^i(n)) = \frac{1}{T^i(n)} \sum_{k=1}^{T^i(n)} R^i_{q_1}(k)$ denotes the empirical mean of quality observations $R^i_{q_1}(n)$ collected from channel i, $G^{q_1}_{\max}(n) = \max_{i \in \mathcal{K}} G^i_{q_1}(T^i(n))$ is the maximum expected quality within the set of channels. Thanks to this formulation, QoS-UCB tends to select a channel with the highest quality and probability to be vacant.

The bias term $A^i(n, T^i(n))$, defined in classical UCB policy [15, 67, 145] is defined as

$$A^i(n, T^i(n)) = \sqrt{\frac{\alpha \log(n)}{T^i(n)}}, \quad \forall i. \tag{4.14}$$

Two coefficients come into play in (4.13) and (4.14), i.e. β and α respectively, and are introduced to balance the trade-off between exploration and exploitation. Parameter α in (4.14) forces the exploration of other channels to check their availability while the new parameter β forces the algorithm to give some weight to the quality in the index computation. If α and β increase, exploration is preferred in order to search for channels with better quality and higher availability. However, if α and β decrease, the empirical mean of observed states, i.e. $\bar{S}^i(T^i(n))$ of channel i, dominates the learning process and bias term forces to exploit the best available channel found in previous iterations. There are three different kinds of behavior during the selection process:

- Case 1: The selected channel is optimal in terms of availability, i.e. $\bar{S}^i(T^i(n))$, and quality, i.e. $Q^i(n, T^i(n))$, then both these terms ensure to exploit this channel.
- Case 2: The selected channel is only optimal in term of availability $\bar{S}^i(T^i(n))$ or quality $Q^i(n, T^i(n))$. In this case, the respective optimal term leads to the exploitation of this channel, but the other term forces to explore the others.
- Case 3: The selected channel is not optimal both in terms of availability and quality, then the bias $A^i(n, T^i(n))$ forces to explore the other channels.

However as already discussed, rested assumptions limits the modeling capability of OSA scenario thus in the next section we focus on the restless problem.

4.3.2 Single-player Restless Bandit Policy: RQoS-UCB

In practical situations, the bands should evolve even if they are not sensed which is captured by the restless formulation. We construct an algorithm called *restless QoS-UCB (RQoS-UCB)*, as shown in Algorithm 4. RQoS-UCB policy operates on a regenerative cycles, thus instead of the actual sample path of Markov chain S^i from an arm i, we limit it to a sample path constructed (or rather stitched together) with the observations from only regenerative cycles as defined previously in Definition 2.9 and introduced in [146].

RQoS-UCB operates in a block structure as shown in Fig. 4.3. For each arm, a state ζ^i is chosen and defined as a *regenerative* state, e.g. free state. Each block is further divided into three sub-blocks (SBs), i.e. SB1, SB2 and SB3. SB1 consists in all time slots from the start of the block to right before the first visit to ζ^i, SB2 contains all time slots from the first visit to ζ^i up to but excluding the second visit to ζ^i where state and quality of the band are recorded, and

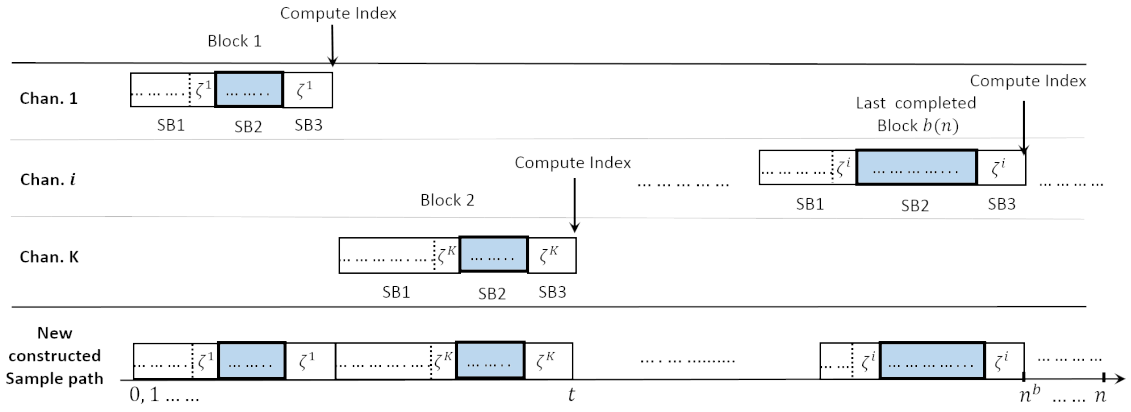


FIGURE 4.3: Example of block (i.e. SB1, SB2 and SB3 sub-blocks) operation of RQoS-UCB policy. At the end of block 1, RQoS-UCB policy computes the index based on the observations collected in SB2 block, finds a channel having the highest index among the set of channels $\mathcal{K}$, and moves to the channel (for example K) with the highest index for block 2.

finally SB3 consists in a single time slot with the second visit to state ζ^i. At the end of SB3, the policy index is computed and is compared with the index of other arms and the highest one gives the next arm to play, e.g. arm K for the second block in Fig. 4.3. Note that the sub-block division is relevant for regret analysis purpose. Indeed, all SB2 blocks are virtually assembled to construct a regenerative cycle of the Markov chains. The newly constructed sample path has exactly the same statistics as the original transition probability matrix P^i which translates restless problem into a tractable problem [146]. However, it is important to emphasize that the CR does not run only during SB2 block but also in SB1 and SB3 blocks in which channels are sensed and transmissions are performed, if bands are found free.

Initially, all the channels are observed at least once and ζ^i is fixed as a first state observed for each arm, i.e. steps 1 to 3 in Algorithm 4. After the initialization, at the beginning of a new block b, RQoS-UCB selects the channel which maximizes the policy index $B^i(n_2, T_2^i(n_2))$ $\forall$ $i \in \mathcal{K}$, step 5, as:

$$B^i(n_2, T_2^i(n_2)) = \bar{S}^i(T_2^i(n_2)) - Q^i(n_2, T_2^i(n_2)) + A^i(n_2, T_2^i(n_2)), \quad \forall i \tag{4.15}$$

where n_2 is defined as the total time spent in SB2 sub-block and $\bar{S}^i(T_2^i(n_2))$, $Q^i(n_2, T_2^i(n_2))$ and $A^i(n_2, T_2^i(n_2))$ have the same meaning than for the rested policy but computed only on the SB2 sub-block.

4.3.3 Multi-player Restless Bandit Policy: Distributed RQoS-UCB

In this part, we extend the previous approach to the distributed multi-player version of RQoS-UCB, sometimes referred as *distributed* RQoS-UCB policy in the following. If each SU applies naively the single player RQoS-UCB presented in Algorithm 4, then the number of collisions will

Algorithm 4 Single SU RQoS-UCB policy

Input: $b = 1$, $n = 0$, $n_2 = 0$, $T_2^i = 0$, α, β, $\mathcal{A}(0)$, $R_{q_1}^i(0)\ \forall\ i \in \mathcal{K}$.
Output: $\mathcal{A}(n+1)$

for n_2 = b **to** K **do**
  Initialize policy by sensing each channel for at least one block (i.e. SB1, SB2 and SB3)
end for
while (1) **do**
  $B^i(n_2, T_2^i(n_2)) = \bar{S}^i(T_2^i(n_2)) - Q^i(n_2, T_2^i(n_2)) + A^i(n_2, T_2^i(n_2))$, $\forall i$
  $\mathcal{A}(n) = \arg\max_i B^i(n_2, T^i(n_2))$
  Sense $i = \mathcal{A}(n)$ and Observe $S^i(n_2)$
  while $S^i(n_2) \neq \zeta^i$ **do**
    $n = n+1$, $\mathcal{A}(n) = i$ // Start SB1 sub-block
    Sense channel i and Observe $S^i(n_2)$
  end while
  $n = n+1$, $n_2 = n_2+1$, $T_2^i(n_2) = T_2^i(n_2)+1$, $\mathcal{A}(n) = i$; // End of SB1, start SB2
  Observe current state $S^i(n_2)$ and update $R_{q_1}^i(n_2)$
  Update $\bar{S}^i(T_2^i(n_2))$, $Q^i(n_2, T_2^i(n_2))$ and $A^i(n_2, T_2^i(n_2))$
  while $S^i(n_2) \neq \zeta^i$ **do**
    $n = n+1$, $n_2 = n_2+1$, $T_2^i(n_2) = T_2^i(n_2)+1$, $\mathcal{A}(n) = i$; // Start SB2 sub-block
    Observe current state $S^i(n_2)$ and update $R_{q_1}^i(n_2)$
    Update $\bar{S}^i(T_2^i(n_2))$, $Q^i(n_2, T_2^i(n_2))$ and $A^i(n_2, T_2^i(n_2))$
  end while
  $b = b+1$, $n = n+1$ // Start of SB3 sub-block
end while

likely increase since all SUs go for the best channel. Hence, we introduce the so-called channel access *rank* [10], defined as:

Definition 4.1 (Channel Access Rank)**.** It is defined as the number of entry which is selected from ordering set of policy index, i.e. a player having a *rank* equal to 2, goes to the channel having the second entry in the decreasing ordering set of policy index.

Let assume that each SU j keeps a set $B^{i,j}, \forall i \in \mathcal{K}$ of indexes in a decreasing order. Moreover, suboptimal channels are expected to be played as less as possible, in order not to increase the regret of the system.

Fig. 4.4 illustrates the functioning of the distributed RQoS-UCB policy for 2 SUs. As before in RQoS-UCB, a player continues to play the same arm till it completes a regenerative block, upon which it updates the indices for the arms using observations from SB2's. Here, a block in Fig. 4.4 should be understood containing the three sub-block SB1, SB2 and SB3. Within this regenerative block, player may collide in any of the time slots, and it does not receive any reward. Let us consider a slotted system with a frame size W, where each SU can be synchronized for their index calculation. Each SU computes its own $B^{i,j}$ index at the end of W if possible, i.e. the sub block SB3 has been encountered for player j in channel i when the frame ends. If not, the computation is delayed to the next frame and player j continues to play

the same channel. If two or more players go to the same channel, they collide and then they draw a random number from the channels' set $\mathcal{K}$ as their new rank. Letting a player randomize among its U optimal ranked arms can help alleviate this problem, and focuses to eventually orthogonalize the U players in their choice of arms.

This random rank is the same idea used in Anandkumar *et al.* [10]. The difference in this thesis is that, in Anandkumar *et al.* [10] the randomization is performed under an iid reward model, whereas in our case the randomization is performed at the end of a completed frame of W length and is therefore less frequent as block lengths are random. The reason for this is because with the Markovian reward model, index updates can only be performed after finishing a regenerative cycle, and switching a channel before a completion of regenerative block will waste the state observations made within that incomplete block. Algorithm 5 summarizes the steps followed by each player. Note that policy operation in each block follows the same steps as the single player policy detailed in Algorithm 4, but not reported here for the sake of simplicity.

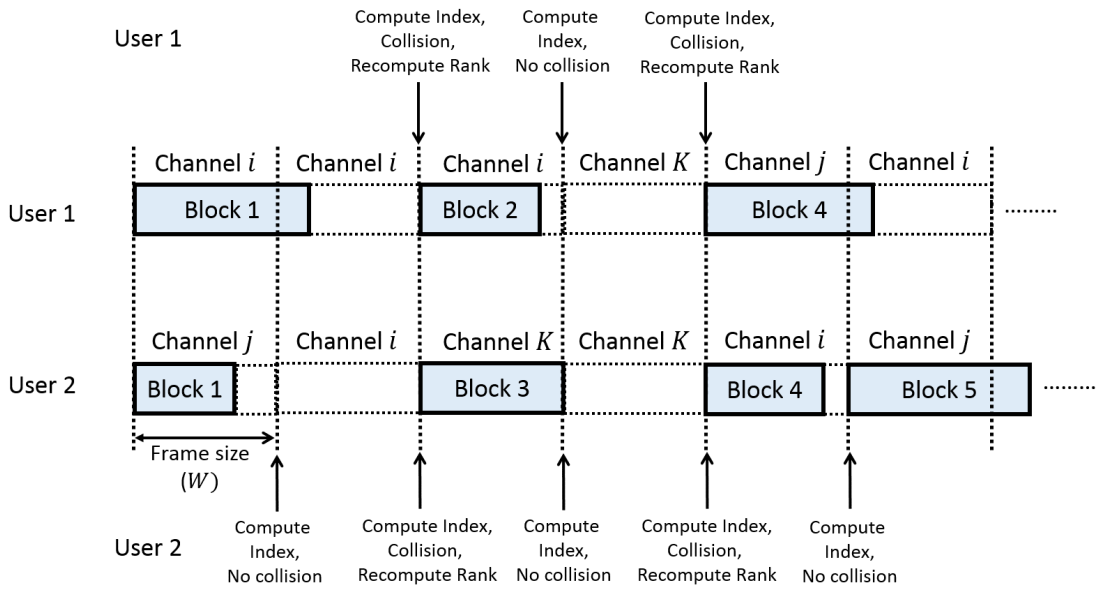


FIGURE 4.4: Running cycle for 2 different SUs using distributed RQoS-UCB policy. Actions of player 1 and 2 are listed on top and bottom, respectively

4.4 Regret Analysis

4.4.1 Preliminaries for Regret Analysis

Let us define some preliminaries which will be used in the upcoming Lemmas and Theorems. We invite the reader to refer to the list of notations to follow with this section. The upper

Algorithm 5 Distributed Multi-player Random Channel Access policy

Input: U: Number of players, K: Number of channels,
$B^{i,j}(n, T^i(n))$: single-player policy index for each SU $j \in U$ and channel $i \in K$,
$C_o^j(i, n)$: indicator of collision at $n-$th slot at channel i for SU j,
$Rank(j)$: $Rank(j)-$th highest entry in $B^{i,j}(n, T^i(n)), \forall i \in K$ for SU j
Output: $\mathcal{A}(n, j)$

if SB3 sub-block observed by SU j in last frame **then**
  Calculate RQoS-UCB policy index $B^{i,j}(n, T^i(n))$ same as Algorithm 4
  if $C_o^j(\mathcal{A}(n-1, j), n-1) = 1$ **then**
    Draw a new $Rank(j)$ randomly from the set $\{1, \cdots, U\}$ for SU j
  else
    Maintain same $Rank(j)$ for SU j
  end if
  $\mathcal{A}(n, j)$: channel having $Rank(j)-$th highest entry in $B^{i,j}(n, T^i(n))$
  Sense $\mathcal{A}(n, j)$ channel
  if collision **then**
    $C_o^j(\mathcal{A}(n, j), n) \leftarrow 1$
  else
    $C_o^j(\mathcal{A}(n, j), n) \leftarrow 0$
  end if
else
  $\mathcal{A}(n, j) = \mathcal{A}(n-1, j)$
  Follow on SB1 and SB2 sub-block same as in Algorithm 4
end if

bound on regret of rested and restless single-player Markovian MAB framework is established under the following condition on the arms.

Condition 4.1. All arms are finite-state, irreducible, aperiodic Markov chains whose transition probability matrices have irreducible multiplicative symmetrization, and the state of non-played arms may evolve. Let, $G_q^i \geq \frac{1}{\hat{\pi}_{\max} + \pi_q^i}$ and $\beta \geq 84 S_{\max}^2 r_{\max}^2 G_{\max}^2 \hat{\pi}_{\max}^2 / \left(\gamma_{\min} \Delta \mu_i^R M_{\min}\right)$.

The following Theorem and lemmas come from literature and are used for the analysis of regret bound.

Lemma 4.1. *[88]. Let $\{S(n)\}_{n \geq 0}$ be an irreducible, aperiodic Markov chain on a finite state space S with transition probability matrix P, non zero initial distribution $\mathbf{h}$ and a stationary distribution $\boldsymbol{\pi}$. Let $N_{\mathbf{h}} = \|(\frac{h_q}{\pi_q}, q \in S)\|_2$, where $\|\cdot\|_2$ denotes $l_2(\pi)$-norm. Let $\gamma = 1 - \lambda_2$ be the eigenvalue gap, where λ_2 is the second largest eigenvalue of the matrix $\tilde{P}$. Let the function $f : S \to \mathbb{R}$ be such that $\sum_{q \in S} \pi_q f(q) = 0, \|f\|_\infty \leq 1, \|f\|_2^2 \leq 1$. If $\tilde{P}$ is irreducible then for any integer ϵ,*

$$\mathbb{P}\left(n^{-1} \sum_{t=1}^{n} f(S(t)) \geq \epsilon\right) \leq N_{\mathbf{h}} e^{-n\epsilon^2 \gamma / 28}$$

The constant $N_{\mathbf{h}}$ is the $l_2(\pi)$-norm of the density of initial distribution $\mathbf{h}$ related to the stationary distribution $\boldsymbol{\pi}$.

Definition 4.2 (Stopping Times [21, 88])**.** Let $\{S(n)\}_{n\geq 0}$ be a stochastic process. A stopping time with respect to S is a random time such that for each $n \geq 0$, the event $\{\tau = n\}$ is completely determined by (at most) the total information known up to time $n, \{S(0), \cdots, S(n)\}$.

Lemma 4.2. *[12]. Let $\{S(n)\}_{n\geq 0}$ be an irreducible, aperiodic Markov chain on a finite state space S with transition probability matrix P, non zero initial distribution $\mathbf{h}$ and a stationary distribution $\boldsymbol{\pi}$. Let $\mathcal{F}_n$ be the σ-algebra generated by $S(1), S(2), \cdots, S(n)$. Let $\mathcal{G}$ be a σ-algebra independent of $\mathcal{F} = \bigvee_{n\geq 1} \mathcal{F}_n$, the smallest σ-algebra containing $F_1, F_2, \cdots$[3]. Let τ be a stopping time of $\mathcal{F}_n \vee \mathcal{G}$. Let $\mathbf{N}(q, \tau) := \sum_{t=1}^{\tau} \mathbf{1}_{(S(n)=q)}$. Then,*

$$|\mathbb{E}[\mathbf{N}(q, \tau)] - \pi_q \mathbb{E}[\tau]| \leq C_P$$

where C_P is a constant.

The following Lemma, which can be found in [21], will be used to bound regret of restless bandit policy.

Lemma 4.3. *[21]. If $\{S(n)\}_{n\geq 0}$ is a positive recurrent homogeneous Markov chain with state space S, stationary distribution π and τ is a stopping time that is finite almost surely for which $S(\tau) = q$ then for all $q \in S$*

$$\mathbb{E}\left[\sum_{t=0}^{\tau-1} \mathbb{1}_{(S(t)=q|S(0)=q)}\right] = \mathbb{E}[\tau|S(0) = q]\pi_q$$

Based on the previous result, we can first state and prove the following lemma.

Lemma 4.4. *The reward of each arm is given by a Markov chain satisfying the condition 4.1. Under any policy, the expected time difference between two successive samples from the selected arm is finite. We have*

$$\Phi^{\mathcal{A}}(n) \leq \sum_{i=2}^{K} (\mu_1^R - \mu_i^R)\mathbb{E}\left[T^i(n)\right] + C_{\mathbf{S},\mathbf{P},\mathbf{G},\mathbf{r}}, \tag{4.16}$$

where, $C_{\mathbf{S},\mathbf{P},\mathbf{G},\mathbf{r}}$ is a constant that depends on all the state spaces S^i, transition probability matrices P^i, the set of rewards $\mathbf{r}_q^i$ associated with states, and the set of rewards associated with quality $\mathbf{G}_q^i$, $i = 1, \cdots, K$

[3] Union of a increasing σ-algebra $\bigcup_{n\geq 1} \mathcal{F}_n$ is not generally a σ-algebra. Generation of a σ-algebra is done by using a join of σ-algebra typically defined $\bigvee_{n\geq 1} \mathcal{F}_n$.

Proof. Let $S^i(1), S^i(2), \cdots$ denoting the successive states observed from arm i. Let $\mathcal{F}_t^i$ be the σ-algebra generated by $S^i(1), S^i(2), \cdots, S^i(t)$ where $S^i(t)$ is the t-th observation from arm i. Let, $\mathcal{F}^i = \bigvee_{t\geq 1} \mathcal{F}_t^i$ and $\mathcal{G}^i = \bigvee_{j\neq i} \mathcal{F}^j$. Since arms are independent, $\mathcal{G}^i$ is independent from $\mathcal{F}^i$. $T^i(n)$ is a stopping time w.r.t. $\{\mathcal{G}^i \vee \mathcal{F}_n^i, n \geq 1\}$. Let, $S^i(1), \cdots, S^i(T^i(n))$ denoting the successive states observed from arm i up to time n. The reward associated with states and quality observed under policy $\mathcal{A}$ up to time n is given by:

$$\sum_{t=1}^{n} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} = \sum_{i=1}^{K} \sum_{j=1}^{T^i(n)} \sum_{q\in S^i} r_q^i G_q^i \mathbf{1}_{(S^i(j)=q)}. \tag{4.17}$$

By definition of the regret,

$$\Phi^{\mathcal{A}}(n) = n\mu_1^R - \mathbb{E}\left[\sum_{t=1}^{n} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right] \tag{4.18}$$

Therefore from (4.18),

$$\begin{aligned}
&\left|\Phi^{\mathcal{A}}(n) - n\mu_1^R + \sum_{i=1}^{K} \mu_i^R \mathbb{E}\left[T^i(n)\right]\right| = \left|-\mathbb{E}\left[\sum_{t=1}^{n} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right] + \sum_{i=1}^{K} \mu_i^R \mathbb{E}\left[T^i(n)\right]\right| \\
= &\left|\mathbb{E}\left[\sum_{i=1}^{K} \sum_{j=1}^{T^i(n)} \sum_{q\in S^i} r_q^i G_q^i \mathbf{1}_{(S^i(j)=q)}\right] - \sum_{i=1}^{K} \sum_{q\in S^i} r_q^i G_q^i \pi_q^i \mathbb{E}\left[T^i(n)\right]\right|, \text{Using (4.17), (4.1)} \\
\leq &\sum_{i=1}^{K} \sum_{q\in S^i} \left|\mathbb{E}\left[\sum_{j=1}^{T^i(n)} r_q^i G_q^i \mathbf{1}_{(S^i(j)=q)}\right] - r_q^i G_q^i \pi_q^i \mathbb{E}\left[T^i(n)\right]\right|, \text{from triangular inequality} \\
= &\sum_{i=1}^{K} \sum_{q\in S^i} r_q^i G_q^i \left|\mathbb{E}\left[\mathbf{N}(q, T^i(n))\right] - \pi_q^i \mathbb{E}\left[T^i(n)\right]\right|, \quad \text{using } r_q^i G_q^i > 0 \\
\leq &\sum_{i=1}^{K} \sum_{q\in S^i} r_q^i G_q^i C_{P^i}, \qquad \text{from Lemma 4.2} \\
= &\; C_{S,P,G,r}
\end{aligned}$$

□

4.4.2 Regret of Single-player Policy

We are now able to state the regret of the RQoS-UCB policy for restless single-player Markov MAB problem.

First, we upper bound the total expected number of plays of suboptimal arms in Theorem 4.1 which is then used to upper bound the regret of the RQoS-UCB policy.

Theorem 4.1. *Assume all arms follow condition 4.1. Let* $\pi_{\min}$, $\hat{\pi}_{\max}$, $S_{\max}$, $r_{\max}$, $\gamma_{\min}$, $M_{\min}$, $\Delta\mu_i^R$ *and* $\Omega_{\max}^i$ *defined in the List of notations. We can upper bound the total expected number of block spent in suboptimal arms as:*

$$\mathbb{E}[F^i(b(n))|b(n)=b] \leq \frac{4\alpha \log n}{(\Delta\mu_i^R)^2} + \frac{|S^1|+|S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2},$$

and the total timespent in sub-optimal arm as:

$$\sum_{i\in K} (\mu_1^R - \mu_i^R)\mathbb{E}[T^i(n)] \leq Z_1 \log n + Z_2$$

where,

$$\begin{aligned} Z_1 &= \sum_{i=2}^{K} \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + 1 \right) \frac{4\alpha}{\Delta\mu_i^R} \\ Z_2 &= \sum_{i=2}^{K} \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + 1 \right) \Delta\mu_i^R \left[\frac{|S^1|+|S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right] \end{aligned}$$

Proof. The proof is given in Appendix C.1. □

Now we present our main results on the regret of RQoS-UCB policy in Theorem 4.2.

Theorem 4.2. *Assume all arms follow condition 4.1, the regret of RQoS-UCB policy for restless Markov MAB can be bounded by*

$$\Phi^R(n) \leq Z_3 \log n + Z_4, \tag{4.19}$$

where,

$$\begin{aligned} Z_3 &= Z_1 + Z_5 \\ Z_4 &= Z_2 + Z_6 + Z_7 \\ Z_5 &= \sum_{i=2}^{K} \frac{4\alpha}{\left(\Delta\mu_i^R\right)^2} \left[\mu_i^R \left(1 + \Omega_{\max}^i\right) + \mu_1^R \Omega_{\max}^1 \right] \\ Z_6 &= \sum_{i=2}^{K} \left[\frac{|S^1|+|S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right] \left[\mu_i^R \left(1 + \Omega_{\max}^i\right) + \mu_1^R \Omega_{\max}^1 \right] \\ Z_7 &= \mu_1^R \left(\frac{1}{\Pi_{\min}} + \max_{i\in\{1,\ldots,K\}} \Omega_{\max}^i + 1 \right) \end{aligned}$$

Proof. Proof of Theorem 4.2 is given in Appendix C.2. □

We can note that the presented upper bound does not depend on the regenerative state ζ. Thus, with minimal information about the channels, an SU can still guarantee logarithmic regret by selecting appropriate exploration coefficient.

4.4.3 Regret of Multi-player Policy

We will show that using Algorithm 5, the sum regret of all players with respect to the optimal centralized (coordinated) solution is logarithmic in time. Our analysis follows a similar approach as in Anandkumar *et al.*, in [10], adapted to blocks rather than time slots and with a several of technical differences. In particular, the proof of Lemma 4.6 and Theorem 4.4 is significantly different since a single block of some player may collide with multiple blocks of other players; thus we need to consider the actions of the players jointly in order to bound the regret.

We now present a logarithmic upper bound on the number of time one of the suboptimal channel $i \in \{U+1, \cdots, K\}$, as defined in Definition 2.7, is sensed by one of the U SUs employing the same distributed RQoS-UCB policy[4].

Theorem 4.3 (Time spent in suboptimal channels)**.** *Assume all arms follow condition 4.1. Under the distributed RQoS-UCB scheme, total time spent by any SU $j \in \{1, \cdots, U\}$ in any suboptimal channel $i \in \{U+1, \cdots, K\}$ is given by:*

$$\begin{aligned} \mathbb{E}[T^{i,j}(n)] &\leq \sum_{j=1}^{U} \left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + W \right) \mathbb{E}\left[F^{i,j}(f(n)) \right], \forall i \in \{U+1, \cdots, K\} \\ &\leq \sum_{j=1}^{U} \left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + W \right) \left[\frac{4\alpha \log n}{(\Delta\mu^R_{i,j})^2} + \left[\frac{|S^j| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right] \right] \end{aligned} \tag{4.20}$$

Proof. Proof is given in Appendix C.3. □

Let now focus on the analysis of the number of collisions $C_o(n)$ up to time n in the U optimal channels. First, we state a bound on the expected number of collisions in the ideal scenario where each SU has perfect knowledge of the mean reward μ^R_i. In this case, SUs try to reach an orthogonal configuration by uniformly randomizing over the U optimal channels.

The considered stochastic process here is a finite-state Markov chain in which a state corresponds to a configuration of U SUs in U number of channels as shown in Fig. 4.5. The total number of states in the above Markov chain is the total number of combinations of U SUs, given by $\binom{2U-1}{U}$ as in Theorem 5.1 in [20]. We consider that the orthogonal access of all SUs corresponds to the absorbing state. For any other state, if there are more than one SUs or no SU in any of the channels, then the transition probability to any other (including self) state of the Markov chain is uniform. For a state, if certain channels have exactly one SU, then there are only transitions (with uniform transition probabilities) to states which consist of at least one SU in

[4]Note that the upper bound on $\mathbb{E}[T^{i,j}(n)]$ is still achieved in a similar way for an SU j even if other SUs are using a channel selection policy different from distributed RQoS-UCB policy. However on the contrary, we need to ensure that every SU must implement the same random access mechanism in order to analyze the expected number of collisions

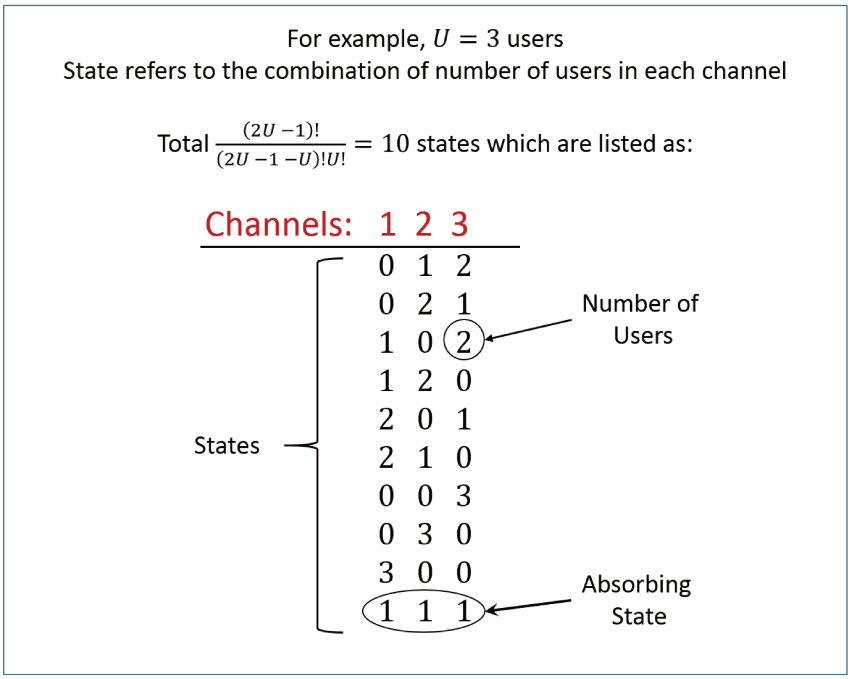


FIGURE 4.5: The list of state corresponds to a configuration of U SUs in U number of channels.

that channel. Let Υ refers the maximum time required to reach to the absorbing state starting from any initial distribution.

We use the following Lemma from [10] and [20] to bound the number of collisions arising due to the distributed scenario:

Lemma 4.5 (Number of collisions under perfect knowledge of μ_i^R, [10, 20])**.** *Given all agents have the correct ordering of the arms and do not change this ordering anymore, the expected number of collisions under random allocation access scheme in Algorithm 5 is bounded by*

$$\mathbb{E}[C_o(n)|\mu_i^R] \quad \leq \quad U\mathbb{E}[\Upsilon] \leq U\left[\binom{2U-1}{U}-1\right].$$

The above Lemma 4.5 states that there is a finite number of collisions, bounded by $U\mathbb{E}[\Upsilon]$ under the perfect knowledge of μ_i^R. However as stated before, there are no collisions in a case where all SUs have the perfect knowledge of μ_i^R in the presence of pre-allocated ranks. Thus, $U\mathbb{E}[\Upsilon]$ gives a bound on the additional number of collisions due to the absence of pre-allocated ranks or the lack of direct communication among SUs to negotiate their ranks. To analyze the number of collisions under multi-player distributed RQoS-UCB learning, we show that all SUs are able to learn the correct order of the different channels with only logarithmic regret, and then we show that only an additional finite number of collisions occurs before reaching collision-free configuration.

Let define $T_w(n)$ and $f_w(n)$ as the number of time and frame where any one of the U optimal channel's estimated ranks is wrong under distributed RQoS-UCB policy.

Lemma 4.6 (Wrong order of distributed RQoS-UCB index statistics without perfect knowledge of μ_i^R)**.** *Under the distributed RQoS-UCB scheme in Algorithm 5, the total expected number of frames and time slots for which the estimated $B^{i,j}(n, T^{i,j}(n))$ indices of distributed RQoS-UCB policy is not in a same order as the mean reward μ_i^R, is:*

$$\mathbb{E}[T_w(n)] \leq U \sum_{a=1}^{U} \sum_{b=1}^{K} \left(\frac{1}{\pi_{\min}^b} + \Omega_{\max}^b + W \right) \left[\frac{4\alpha \log n}{(\Delta\mu_{a,b}^R)^2} + \frac{|S^a| + |S^b|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right] \tag{4.21}$$

Proof. Proof of Lemma 4.6 is given in Appendix C.4. □

We are now able to discuss an upper bound on the total number of collisions $C_o(n)$ in the U optimal channels by considering the above result on $\mathbb{E}[T_w(n)]$, $\mathbb{E}[T^{i,j}(n)]$ and $U\mathbb{E}[\Upsilon]$.

Lemma 4.7 (Logarithmic number of collisions [10])**.** *The total expected number of collisions in the U optimal channels under distributed RQoS-UCB scheme satisfies,*

$$\mathbb{E}[C_o(n)] \leq U(\mathbb{E}[\Upsilon] + 1)\, \mathbb{E}[T_w(n)] \tag{4.22}$$

Next, we proceed to prove one of the main results of this chapter that the sum regret under distributed RQoS-UCB policy is logarithmic for restless bandit.

Theorem 4.4 (Regret analysis of multi-player distributed RQoS-UCB policy)**.** *Assume all arms follow condition 4.1. Then, the regret of the distributed RQoS-UCB policy can be bounded by $O(\log n)$*

$$\Phi^M(n) \quad \leq \quad X_3 \log n + X_4 \tag{4.23}$$

where,

$$X_3 = X_1 + X_5 \quad \text{and} \quad X_4 = X_2 + X_6 + X_8$$

$$X_1 = \left(\frac{\left[K^2 - KU + U\right]}{\pi_{\min}} + 1 \right) \mu_1^R U^2 (\mathbb{E}[\Upsilon] + 1) \sum_{a=1}^{U} \sum_{b=1}^{K} \frac{4\alpha}{\Delta\mu_{a,b}^R} X_9$$

$$X_2 = \left(\frac{\left[K^2 - KU + U\right]}{\pi_{\min}} + 1 \right) \mu_1^R U^2 (\mathbb{E}[\Upsilon] + 1) \sum_{a=1}^{U} \sum_{b=1}^{K} \left[\frac{|S^a| + |S^b|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right] X_9$$

$$X_5 = \sum_{i=U+1}^{K} \sum_{k=1}^{U} \frac{4\alpha}{\left(\Delta\mu_{i,k}^R\right)^2} X_7$$

$$X_6 = \sum_{i=U+1}^{K} \sum_{k=1}^{U} \left[\frac{|S^k| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right] X_7$$

$$X_7 = \left[\mu_i^R \left(W + \Omega_{\max}^i \right) + \sum_{j=1}^{U} \mu_j^R \Omega_{\max}^j + \mu_1^R X_9 \right]$$

$$X_8 = U \mu_1^R \left(\frac{1}{\pi_\zeta} + \Omega_{\max} + W \right)$$

$$X_9 = \left(\frac{1}{\pi_\zeta} + \Omega_{\max} + W \right)$$

Proof. Proof of Theorem 4.4 is given in Appendix C.5. □

We can note that the presented upper bound does not depend on the regenerative state ζ. Thus, with minimal information about the bands, an SU can still have a logarithmic regret by selecting appropriate exploration coefficients.

4.5 Numerical Results and Analysis

4.5.1 Simulation Settings

Analyzed Approaches for Performance Evaluation

Due to the application to OSA context, RQoS-UCB is compared to several other learning policies found in literature:

- Q-learning policy [27, 150] takes into account both channel quality and availability for opportunistic access. Q-learning does not follow the MAB framework in general, but the purpose of this algorithm is the same than ours and and it is a more general class of reinforcement learning policy.
- Regenerative Cycle Algorithm (RCA) [146]: It is the restless version of baseline UCB1 policy (in [15, 67, 145]) with taking into an account regenerative cycle behavior, and it plays the optimal channel in terms of only availability, i.e. $\pi_{q_1}^i$.
- *Restless Upper Confidence Bound (RUCB)* [94]: It is also the restless version of baseline UCB1 policy (in [15, 67, 145]) with taking into an account increasing exploration and exploitation epoch, and it also plays the optimal channel in terms of only availability, i.e. $\pi_{q_1}^i$.

- Best channel selection: This policy always selects the optimal channel in terms of availability and quality, i.e. channel with the highest mean reward, and if selected optimal channel is occupied it does not transmit.

- Best opportunistic selection: The Best opportunistic selection policy is a "god driven" policy which knows *a priori* all holes in the spectrum as well as the ordering of channels mean reward, considering availability and quality. This is the most efficient use of the spectrum holes, however it needs a prior information about channels statistics.

- Round-Robin approach: Deterministic sequence of trials in channel sensing. This comparison allows to know if 'Is it worth learning channel statistics?' If so, 'How much can we improve with respect to non-intelligent approach?'

The classical RCA and RUCB policies are modified such as the unique reward takes into account the combination of availability and quality, i.e. $s_q^{i,j} = R_q^{i,j} r_q^{i,j}$. Here, $r_q^{i,j}$ is the fixed reward selected for channel i and player j in state q, whereas $R_q^{i,j}$ is the sample drawn from the fixed iid reward distribution modeling the quality of channel i in state q for player j which is not considered in previous works.

Simulation Scenarios

The performance of distributed RQoS-UCB policy is investigated on a set of 10 bands and with different number of SUs $U \in \{1, \cdots, 10\}$. Simulation is performed over 10^3 runs, each with a duration about 10000 frames. QPSK signaling are assumed to be used for PU signals like in [92]. The threshold υ is set to have a high detection probability for each channel i. Constant c_e, in (4.7), is set to 0.2 which is used to upper bound the reward with value of 5. Note that the learning phase represents very low computational complexity that it may be neglected compared to sensing [106]. Learning can be done in parallel with the transmission, and so uses no time that could prevent from transmitting frames, and thus be considered as having no overhead impact on bandwidth allocation. The exploration coefficients of RQoS-UCB policy are $\alpha = 0.25$, $\beta = 0.32$, according to Condition 4.1, and will be used throughout the numerical analysis unless otherwise mentioned. As discussed in Chapter 3, the performance of MAB learning policies depends on the probability P_f that there exist at least one free channel from set of channels, and it also depends on the optimal identification (OI) factor associated with the considered scenario. Below, we present several scenarios under which the performance of MAB learning policies are analyzed.

Scenario 1: High Probability of Free P_{free} First, we consider a scenario with very high probability of free $P_{free} = 0.999$ and OI factor $H_{OI} = 0.81$. Table 4.1 summarizes the Markov

chain parameters modeling the primary network, such as the transition probabilities P^i, selected arbitrarily, for each channel on the first and second rows, the vacancy probability $\pi^i_{q_1}$ calculated from P^i, the empirical average of the band quality G^i calculated as explained in Section 4.2 and estimated at Rx and feedback to Tx, on the fourth row and the global mean reward, μ^R_i calculated with (4.1) taking into account availability and quality on the fifth row. As we have already seen in numerical analysis of Chapter 3 that, in high P_{free} regime, impact of different levels of OI factor and LZ complexity is negligible on the performance of learning.

channel	1	2	3	4	5	6	7	8	9	10
$p^i_{q_0q_1}$	0.3	0.65	0.75	0.6	0.8	0.4	0.2	0.65	0.45	0.35
$p^i_{q_1q_0}$	0.7	0.35	0.25	0.4	0.2	0.6	0.8	0.35	0.55	0.65
$\pi^i_{q_1}$	0.3	0.65	0.75	0.6	0.8	0.4	0.2	0.65	0.45	0.35
G^i	0.67	0.64	0.79	0.77	0.70	0.80	0.67	0.63	0.64	0.79
μ^R_i	0.31	0.48	0.60	0.47	0.58	0.33	0.26	0.46	0.39	0.29

TABLE 4.1: State transition probabilities, mean availability, empirical mean quality and global mean reward for scenario 1 with $P_{free} = 0.999$ and OI factor $H_{OI} = 0.81$.

Other Scenarios: Different level of OI factor Next, we consider several other scenarios with probability of free $P_{free} = 0.91$ and different OI factor H_{OI} as mentioned in Table 4.2. Markov chain parameters modeling the primary network, such as the vacancy probability $\pi^i_{q_1}$, empirical average of the band quality G^i and the global mean reward μ^R_i are estimated like in Scenario 1 and listed in Table 4.2.

	channel	1	2	3	4	5	6	7	8	9	10
Scenario 2	$\pi^i_{q_1}$	0.85	0.95	0.90	0.75	0.95	0.90	0.80	0.40	0.85	0.8
	G^i	0.63	0.68	0.73	0.67	0.48	0.84	0.54	0.55	0.99	0.85
$H_{OI} = 0.76$	μ^R_i	0.23	0.22	0.10	0.22	0.22	0.10	0.26	0.46	0.26	0.18
Scenario 3	$\pi^i_{q_1}$	0.85	0.95	0.79	0.75	0.75	0.90	0.85	0.55	0.85	0.8
	G^i	0.63	0.68	0.73	0.67	0.48	0.84	0.54	0.55	0.99	0.85
$H_{OI} = 0.85$	μ^R_i	0.23	0.22	0.19	0.22	0.30	0.10	0.23	0.39	0.26	0.18
Scenario 4	$\pi^i_{q_1}$	0.75	0.95	0.79	0.85	0.82	0.73	0.55	0.67	0.55	0.8
	G^i	0.63	0.68	0.73	0.67	0.48	0.84	0.54	0.55	0.99	0.85
$H_{OI} = 0.91$	μ^R_i	0.28	0.22	0.19	0.15	0.27	0.23	0.38	0.34	0.39	0.18

TABLE 4.2: Mean availability, empirical mean quality and global mean reward for 3 scenarios with $P_{free} = 0.91$ and different level of OI factor H_{OI}.

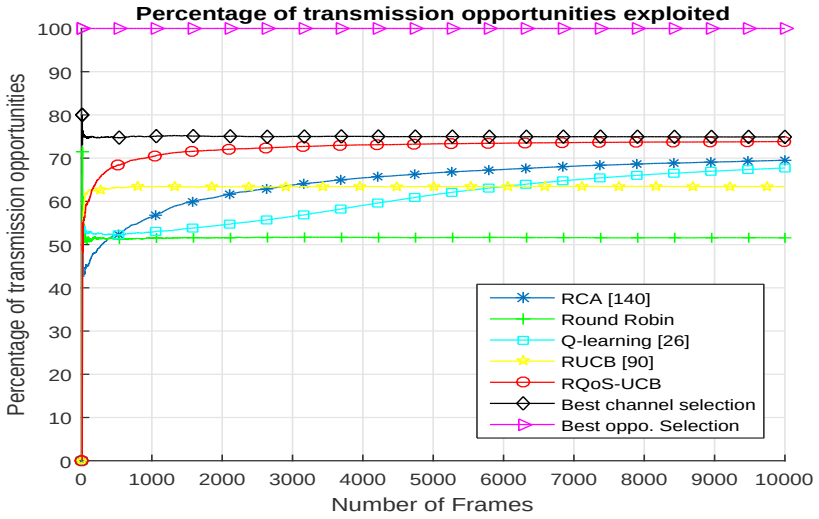
4.5.2 Simulation Results and Discussions

Single user case: Scenario 1

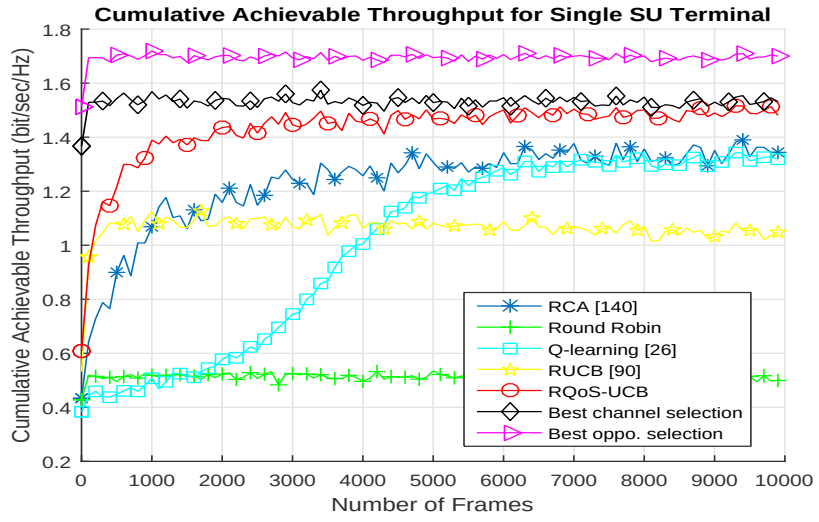
Initially, only one cognitive transceiver pair is considered trying to exploit the frequency bands of a primary network. Theorems 4.1 and 4.2 state that the new metric introduced to rate the quality in the learning phase does not prevent from achieving a logarithmic order regret in the restless case. Due to the application to OSA context, regret analysis is not sufficient to characterize the performance of a learning policy and its ability to provide high reliable data rate is of great interest for telecommunication purposes. Hence, RQoS-UCB policy is compared to several other learning policies found in literature, such as RCA [146], RUCB [94] and Q-learning [26] with the optimal exploration parameters suggested by the authors.

Fig. 4.6(a) presents the percentage of transmission opportunities defined as the ratio between the number of times a policy selects an available channel and the total number of trials. Best opportunistic transmission policy is an upper bound since it possesses a prior information about the spectrum occupancy. The proposed RQoS-UCB but also RCA [146] and Q-learning policies are able to find an optimal channel in the long run, go close to 75 % of transmission opportunity, and match the performance of Best channel selection policy which always selects a channel having the highest mean reward μ_i^R. However, the proposed RQoS-UCB policy greatly outperforms other approaches in terms of convergence speed and achieves the higher number of opportunities. This result highlights the benefits of using separate optimization of both availability and quality as we propose with QoS-UCB compared to use a single reward in the UCB policy.

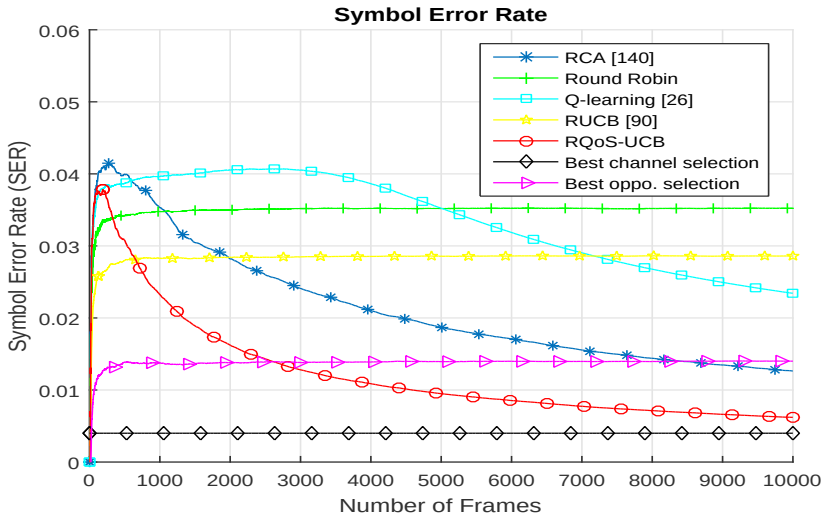
The achievable throughput is investigated in Fig. 4.6(b) and is computed with (4.9) and (4.10). The best opportunistic selection policy logically upper bounds the performance due to the higher number of transmission opportunities it exploits as in Fig. 4.6(a). RQoS-UCB, RCA, RUCB and Q-learning converge toward the best channel selection because of their ability to learn the band with the best weighted combination of quality and availability but with different convergence speed. RQoS-UCB achieves 87% of the best channel selection policy rate in 1000 frames while more than 4000 frames is needed to achieve less than 87% of this value for RCA and Q-learning policies. Moreover, Fig. 4.6(b) also depicts that RQoS-UCB outperforms RUCB, in [94], in convergence speed which is relatively far from the best channel and the best opportunistic transmission policies. This behavior is probably due to the exploitation-exploration epoch structure which has an exponentially growing length. Hence, if the channel selected is not the optimal one, the policy in [94] has difficulties to change in a real communication scenario. On the other hand, the exploration and exploitation are done in the *regenerative cycle* of a near constant length in our policy, which makes it more suitable to try other channels and hence converge faster to the optimal one. These results also demonstrate the efficiency of controlling the learning phase with two rewards instead of one when channels are characterized by not only



(a) Opportunities w.r.t. the number of frames



(b) Achievable throughput w.r.t. the number of frames



(c) Symbol error rate (SER) w.r.t. the number of frames

FIGURE 4.6: Percentage of transmission opportunities exploited, achievable throughput and SER for single SUs w.r.t. the number of frames.

their availability but also by their quality. As expected, a simple round-robin technique cannot compete in this scenario.

In Fig. 4.6(c), the average symbol error rate (SER) is investigated. The SER of RQoS-UCB, RCA and Q-learning converge toward the SER obtained with the best channel selection policy, i.e. $5 \cdot 10^{-3}$ which is the SER of QPSK signaling under 9 dB of SNR, i.e. SNR of band 3. Again, the slower convergence characteristic of Q-learning and RCA, w.r.t. RQoS-UCB, can be emphasized in this figure. This figure also reveals the limitation of the RUCB policy, which likely does not select the band with the best SNR as often as the competing policies, i.e. RQoS-UCB, RCA and Q-learning. We can even notice that the SER of the best opportunistic transmission policy is higher than RQoS-UCB policy, because it selects suboptimal bands to continue transmission when optimal band is occupied. Indeed, the best opportunistic policy is an ideal scheme which exploits at each time the best (in term of quality) available channel, but not necessarily the best one globally. Hence, rather than stopping transmission because the best channel in terms of quality is occupied, it goes to a suboptimal (in quality) channel but available. Hence, during the transmission the secondary link experiences a degraded SINR which increases the SER compared to the case where it would have used the optimal channel, however it allows to transmit anyway. In (4.8), (4.9) and (4.10) the achievable throughput depends not only on SINR but also on the transmission opportunities exploited. In other words, transmitting on a link with a better SINR leads to an increase of throughput but attenuated by the log function. On the other hand, if the link is of bad quality, this results in a decrease of throughput but marginally due to the log function which can be compensated by using more often this link (if it is more available than another one with a better SINR). This what explains that the best opp. policy may have a larger SER than the best channel selection policy but a larger throughput than the latter in Fig. 4.6(b).

Single user case: Other Scenarios

In this part, we compare the performance of RQoS-UCB and other learning approaches under several other scenarios with different OI factor. As we already discussed with the help of numerical analysis in Chapter 3, the level of OI factor significantly affects the performance of MAB learning approaches.

Fig. 4.7 present the percentage of transmission opportunities achieved by the RQoS-UCB, RCA, Q-learning, RUCB, Round Robin and best channel selection policies w.r.t. 3 scenarios with different level of OI factor. The best opportunistic selection policy logically upper bounds the performance. From Fig. 4.7, one can remark that RQoS-UCB policy achieves better performance compared to other approaches achieving similar performance as of the best channel selection policy due to separation of two optimization criteria, i.e. availability and quality.

Fig. 4.7 also depicts that the performance of learning (intelligent) approaches, i.e. RQoS-UCB, RCA, RUCB and Q-learning, reduces with the increase in OI factor, associated with scenario considered. This behavior is due to the fact that high OI factor represents more closer value of mean reward of suboptimal and optimal channels, and thus it becomes significantly difficult for learning approaches to identify and transmit in an optimal channel quickly. On the contrary, performance of non-intelligent approach, like Round Robin, increases with the increase in OI factor because it has equal chance of transmission from all channels in case of high OI factor.

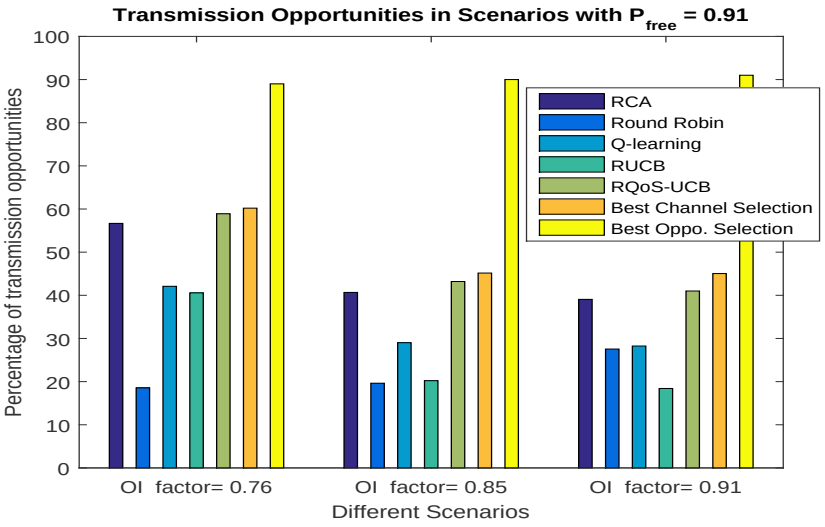


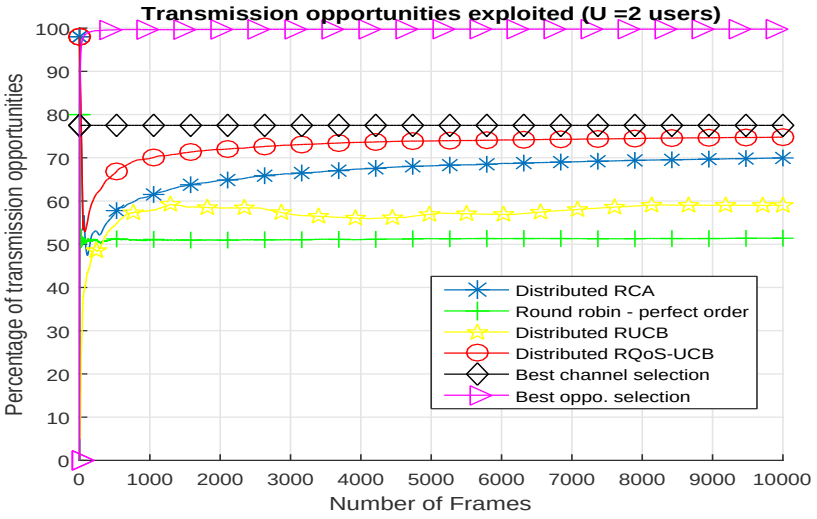
FIGURE 4.7: Percentage of transmission opportunities exploited after 10000 time steps over 100 iterations for single SU w.r.t. scenarios with different OI factor.

Multiple players case: Scenario 1

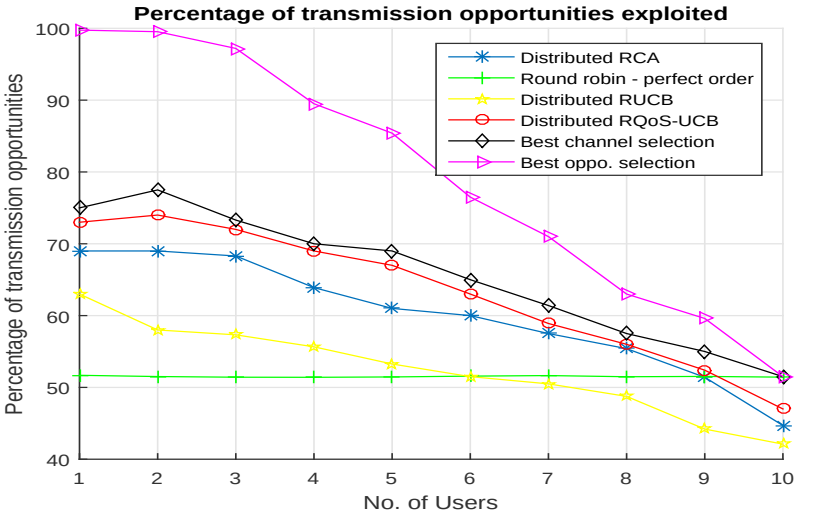
In this part, RQoS-UCB is implemented as presented in Algorithm 5 and compared to *distributed* RCA, distributed RUCB, and round robin-perfect order (R2PO) policy. Only the single player version of RCA can be found in literature, however multi-player version can be implemented easily following the same structure as distributed RQoS-UCB leading to the *distributed RCA* in the following. Moreover, the *round robin-perfect order* policy plays each channel consecutively without suffering from collisions, and *best channel selection* and *best opportunistic selection* are defined in a same manner as for the single-player policy.

Fig. 4.8 presents the average percentage of transmission opportunities exploited by the RQoS-UCB, RCA, RUCB, R2PO and best channel selection policies w.r.t. the number of frames for 2 SUs, Fig. 4.8(a), and w.r.t. the number of players in Fig. 4.8(b). The best opportunistic selection policy logically upper bounds the performance. From Fig. 4.8(a), one can remark that distributed RQoS-UCB policy achieves better performance compared to RCA converging towards the best channel selection policy due to the separation of two optimization criteria, i.e.

availability and quality. All learning approaches significantly outperform R2PO which finds less transmission opportunities even if all SUs have a perfect knowledge about each others' actions. This is also confirmed by Fig. 4.8(b) where the percentage of transmission opportunities is investigated w.r.t. the number of SUs. Distributed RQoS-UCB outperforms RCA in general, and also percentage of opportunities decreases for both when the number of SUs in the network increases and achieves similar performance as R2PO when 9 secondary transceivers are considered. Note that R2PO is not an interesting solution for CR systems, because requiring a predefined agreement or information exchange among SUs. Hence, the distributed implementation of RQoS-UCB is able to find sufficiently high number of transmission opportunities without additional signaling overhead.



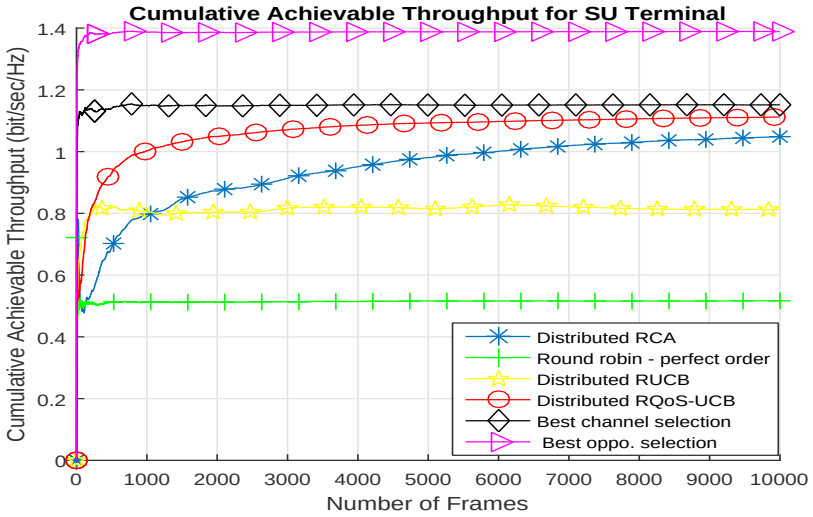
(a) Opportunities w.r.t. the number of frames



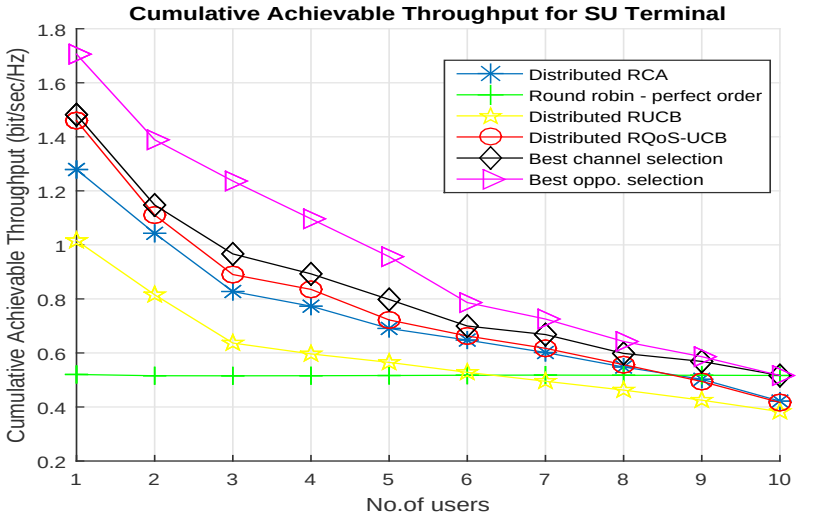
(b) Opportunities w.r.t. the number of SUs

FIGURE 4.8: Percentage of transmission opportunities for 2 SUs w.r.t. the number of frames a) and w.r.t. the number of SUs operating in network b).

The average achievable throughput is investigated in Fig. 4.9 w.r.t. the number of frames for 2 players in the network, in Fig. 4.9(a), and w.r.t. the number of SUs in Fig. 4.9(b). Distributed RQoS-UCB converges rapidly towards the best channel selection policy while distributed RCA achieves similar performance after a larger learning time. Our policy, RQoS-UCB, outperforms RUCB for all values of the number of players and converges to the same performance than RCA when the number of players is larger than or equal to 6, Fig. 4.9(b). Note that these particular values are not absolute but depend on the scenario considered and the total number of primary users' channels. Hence, our proposal is not only beneficial for up to 6 SUs but can be greater if the number of channels is higher. The achievable throughput decreases as the number of players increases but distributed RQoS-UCB matches with the best channel selection policy whatever the considered number of players and channels.



(a) Throughput w.r.t. the number of frames



(b) Throughput w.r.t. the number of SUs

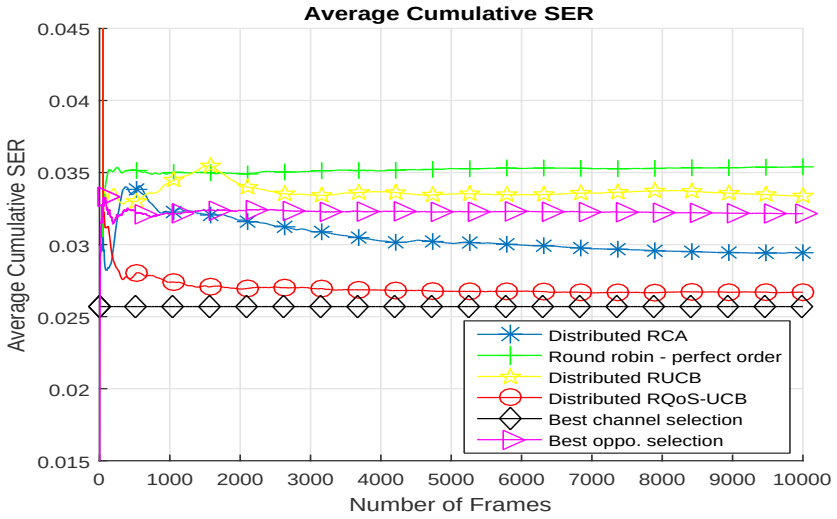
FIGURE 4.9: Average achievable throughput for 2 SUs w.r.t. the number of frames a) and w.r.t. the number of SUs operating in the network b).

The average SER obtained with several policies is investigated in Fig. 4.10, w.r.t. the number of frames and 2 SUs in Fig. 4.10(a), and w.r.t. the number of SUs in 4.10(b). The SER of distributed RQoS-UCB converges towards the SER obtained with the best channel selection policy, i.e. 26×10^{-3} which is the average SER of QPSK signaling under SNR of the first optimal (band 3) and the second optimal (band 5) bands respectively. Like in Figs. 4.6, 4.8 and 4.9, the distributed RUCB [94] and RCA [146] converge to an optimal band but with a slower convergence speed due to the single reward mixing the availability and quality, and hence the average SER achieved with RUCB and RCA is poorer than the one with RQoS-UCB. We remark in Fig. 4.10(b) that RQoS-UCB nearly matches the best channel selection policy which lower bounds the SER of all other strategies when the number of players is varying. The SER of RQoS-UCB increases as the number of players increases, since the number of channels with worse quality increases but is still lower than the SER of RCA and RUCB. At a point, i.e. more than 7 SUs in network, the difference between all learning approaches and round robin-perfect order becomes negligible as they finish to select an important proportion of the same channels. Furthermore, if K would be much larger with larger dissimilarities among channel qualities, the number of players at which our policy would offer better performance than round-robin would also be larger.

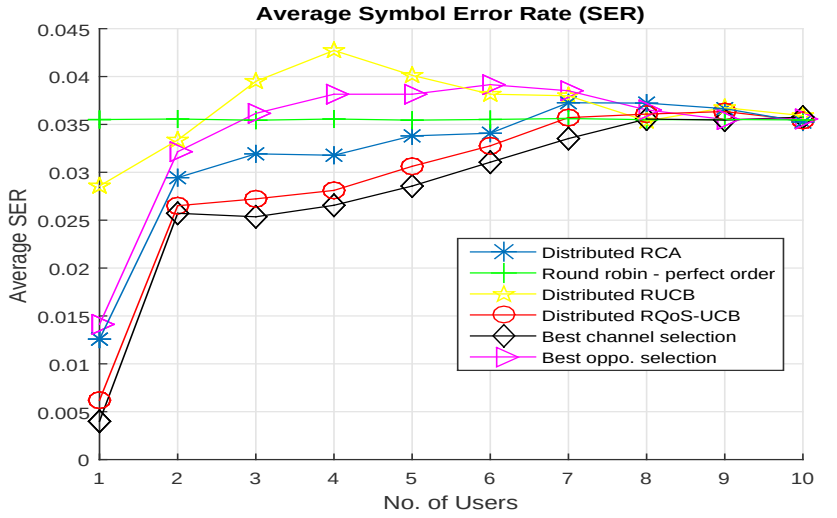
Fairness and Complexity Analysis of the Proposed Policies

One of the important features of the proposed restless RQoS-UCB policy is that it does not favor one specific player over another in order to access optimal arms. In the proposed distributed RQoS-UCB approach, each player has an equal chance to sense and transmit in any one of the U optimal channels. Fig. 4.11(a) illustrates the percentage of opportunities exploited and Fig. 4.11(b) the optimal arm selection percentage for 4 SUs in $K = 10$ channels w.r.t. the number of frames. As it can be observed, each player exploits approximately the same amount of transmission opportunities and selects the optimal arm more or less the same proportion of time. This demonstrates that the proposed distributed RQoS-UCB scheme is indeed fair in player allocation.

To conclude this part, a comparative study, in terms of complexity, learning criteria and convergence speed, between the three learning algorithms, i.e. RQoS-UCB, Q-learning and RCA, has been summarized in Table 4.3. The running time complexity is the number of operation performed and space complexity is related to the storage space (memory) needed to run [79]. For the running time complexity, RCA, RQoS-UCB and Q-learning policies behave in $\mathcal{O}(NK)$ for large N and K, where N and K are the number of time slots and channels. Time complexity of these algorithms are comparable however, RQoS-UCB performs better than the others in numerical analysis. Time complexity of reinforcement learning is negligible and it is approximately 1% of the sensing time complexity of energy detector that is also required for OSA [106]. On the



(a) SER w.r.t. the number of frames



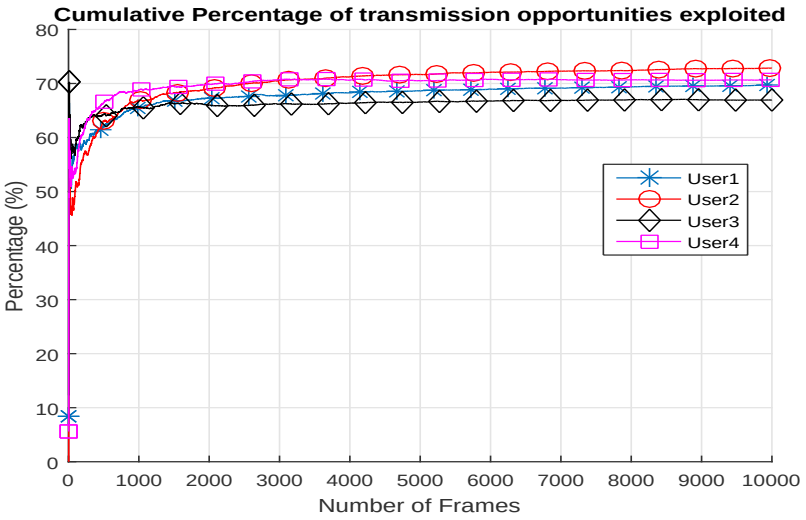
(b) SER w.r.t. the number of SUs

FIGURE 4.10: Average SER for 2 SUs w.r.t. the number of frames a) and w.r.t. the number of SUs operating in the network b).

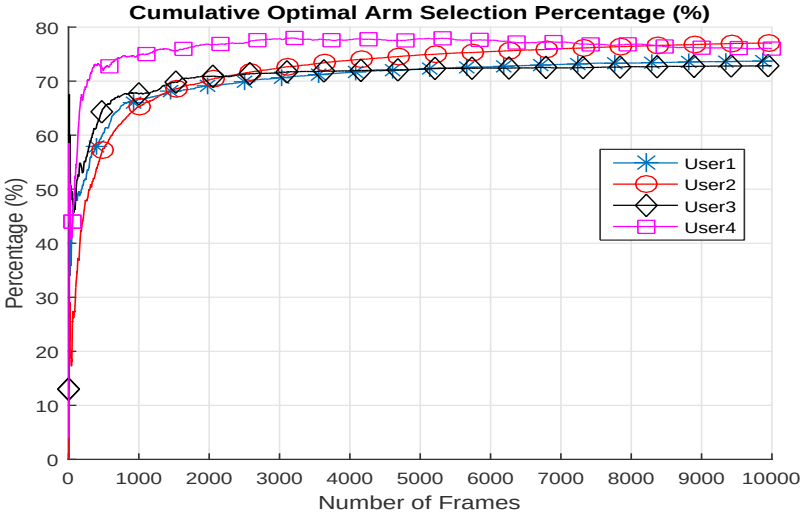
other hand, it is clear that space complexity (expected memory requirement) is the drawback of Q-learning that needs to store all past observations contrary to RCA and RQoS-UCB policies whose complexity is about $O(K)$.

4.6 Conclusion

This chapter has described the Markov MAB problem taking into account another reward distribution associated with each state modeling quality metric. The proposed Markov MAB framework dealt with CR decision making problem applied to opportunistic spectrum access (OSA) problem in single-player and multi-player infrastructure-less cognitive wireless networks.



(a) Opportunities w.r.t. the number of frames



(b) Optimal arm selection percentage w.r.t. the number of frames

FIGURE 4.11: Fairness analysis for 4 SUs implementing RQoS-UCB policy w.r.t. the number of frames.

Learning Algorithms	Running time complexity	Space complexity	Learning Criteria	Theoretical guarantee	Conv. speed
UCB1 [145]	$\mathcal{O}(N(5K+4))$	$\mathcal{O}((4K+2))$	Availability	Rested only	Very Fast
QoS-UCB	$\mathcal{O}(N(10K+5))$	$\mathcal{O}((4K+4))$	Availability and Quality	Rested and Restless	Very Fast
Q-learning	$\mathcal{O}(N(6K+3))$	$\mathcal{O}(N(3K+3))$	Availability and Quality	Rested and Restless	Medium
RCA [146]	$\mathcal{O}(N(3K+3))$	$\mathcal{O}((4K+5))$	Availability	Rested and Restless	Fast
RQoS-UCB	$\mathcal{O}(N(8K+6))$	$\mathcal{O}((4K+7))$	Availability and Quality	Rested and Restless	Fast

TABLE 4.3: Algorithms complexity for K channels and N time slots

A new MAB policy, named RQoS-UCB, has been proposed to address the single-player OSA problem modeled as a restless Markov MAB formulation. The proposed policy takes into account a quality information metric instead of only availability for most of algorithms.

Following RQoS-UCB policy, we also propose its multi-player extension, named distributed RQoS-UCB, to provide collision-free channel access to several non-cooperative selfish SU. Moreover, we have proved that our proposed policies achieve a logarithmic order regret uniformly over time for the restless Markov MAB. In cognitive radio applications, the ability to learn on a different criteria than the traditional free or occupied status of a channel is of particular interest in order to improve the QoS of SUs transmissions. Using the soft-output of ED as a quality metric for the sensed band, we have shown that the proposed policies are able to achieve a larger throughput than the state-of-the-art algorithms which suffer from a larger convergence time compared to the proposed policies.

The idea proposed in this Chapter can be used to learn on many other criteria such as energy efficiency or actual SINR on the secondary link and can be investigated in future works. Moreover, our model ignores dynamic traffic at the secondary nodes and extension to a queueing-theoretic formulation is left for future work.

Chapter 5

TRQoS-UCB Policy: Energy Efficient Cellular Network

Contents

5.1 Introduction 81
5.2 Network Model and Problem Formulation 82
5.2.1 Network Model 82
5.2.2 Problem Formulation 85
5.3 RL Framework for Energy Efficient Network 86
5.3.1 System Model 86
5.3.2 Transfer Learning: Transfer RQoS-UCB Policy 88
5.4 Convergence Analysis: Transfer RQoS-UCB Policy 91
5.5 Numerical Results and Analysis 92
5.5.1 Convergence Analysis 94
5.5.2 Performance under Periodic Traffic Load 96
5.6 Conclusion 100

5.1 Introduction

Chapter 4 has dealt with dynamic spectrum access scenario: a fully distributed learning for channel selection with better quality of service (QoS) for transmission has been proposed. Although, MAB approaches are particularly well adapted to tackle opportunistic spectrum access problem in wireless communications, we show in this Chapter that MAB model can also be applied to tackle other more complex optimization problems arising in wireless communication area, such as dynamically switching ON-OFF radio resource in cellular network.

Recently, there has been rising interest on the work related to traffic load aware BSs adaptation or self-organized cellular network. However, it is essential for network operators to guaranty radio coverage and QoS to the cellular users. Dynamically switching the BSs' operation mode to 'ON' and 'OFF' with respect to traffic load fluctuation is considered to be one of the effective methods to reduce the total energy consumption of cellular network. In this chapter, the learning process of the best network topology, i.e. sets of BSs being ON and OFF, is investigated. The BS switching operation controller employs a learning process based on the RQoS-UCB policy and attempts to find the best configuration of BSs being ON in order to estimate the traffic load variations during the day and maximize the energy efficiency (EE) of the network.

Furthermore in this Chapter, we propose to use the transfer learning (TL) concept for upper confidence bound class of policies, where it could exploit the temporal and spatial relationship in the traffic load and increase the convergence speed of the learning. As an outcome, the MAB learning framework of BS switching operation is further extended by introducing the idea of TL into the RQoS-UCB policy, and referred as Transfer RQoS-UCB (TRQoS-UCB) policy.

The organization of this chapter is as follows. Section 5.2 introduces network model and problem formulation. In Section 5.3, the reinforcement learning (RL) framework for energy efficient operation of the network is introduced. It also introduces RQoS-UCB, TRQoS-UCB algorithms for energy efficiency. Section 5.4 presents upper bound on the sub-optimal plays by the proposed TRQoS-UCB algorithm. Section 5.5 numerically evaluates and compares the proposed schemes with the state-of-the-art methods and presents their validity and effectiveness. Finally in Section 5.6, we conclude this chapter and present the future works to be investigated.

5.2 Network Model and Problem Formulation

5.2.1 Network Model

In this work, we consider a heterogeneous wireless cellular network comprising of a mixture of macro and small cells where set of BSs $\mathcal{Y} = \{1, 2, \cdots, Y\}$ lies in a two dimensional area in $\mathbb{R}^2$. In addition, we consider that there exists a BS switching operation controller, which knows the traffic loads in these BSs at current iteration and computes the energy efficiency of BSs at next stage in a centralized way. Let us assume that all BSs operate in an open access mode, meaning that any MS is allowed to connect to any BSs from any tier [139]. We focus on the downlink communication as it is the primary usage for the mobile Internet application. The network area is divided according to the Voronoi tessellation with BS acting as seeds for each cell. Each cell coverage in the wireless cellular network is denoted as $\mathcal{I}_k^n (k = 0, 1, 2, \cdots)$ at iteration n.

Traffic Profile

The same packet based traffic model as in [89, 123] is used for our analysis and simulations. Let, $x^k \in \mathcal{I}_k^n$ be two-dimensional Cartesian coordinates, denoting the locations of MS in coverage $\mathcal{I}_k^n$ of the $k-$th BS at iteration n. An MS is active when it is receiving a file transfer. When the call ends, MS becomes inactive and is considered departed from the network. Then, x^{on} is two-dimensional Cartesian coordinates, denoting the locations of active MS in two dimensional area in $\mathbb{R}^2$. Traffic load of a BS is measured in terms of the number of active MSs and their respective call duration.

At each iteration n, file transmission request at location $x^k \in \mathcal{I}_k^n$ of $k-$th BS arrives following a Poisson point process (PPP) with arrival rate $\Lambda(x^k, n)$, and the associated file size is exponentially distributed with mean $1/h(x^k, n)$. Then the instantaneous traffic load $L(x^k)$ at location x^k can be estimated as

$$L(x^k, n) = \frac{\Lambda(x^k, n)}{h(x^k, n)}$$

at iteration n. By setting different arrival rates for MSs located in different BSs, this model can capture the temporal and spatial traffic variability. Thus, when the set of BSs $\mathcal{Y}_n^{on}$ is turned ON at iteration n, the instantaneous traffic load served by BS $k \in \mathcal{Y}_n^{on}$ can be expressed as:

$$L^k(n) = \sum_{x^k \in \mathcal{I}_k^n} \frac{\Lambda(x^k, n)}{h(x^k, n)}.$$

On the contrary when BS k is turned OFF, the instantaneous traffic load served by BS k is zero, i.e. $L^k(n) = 0$. The total arrival rate (or total call duration) of a BS k is the composition of all Poisson arrivals at different locations in $\mathcal{I}_k$, which again forms a Poisson process [161]. Then at each iteration n, the total arrival rate and total call duration of a BS k can be estimated as $\Lambda_n^k = \sum_{x^k \in \mathcal{I}_k^n} \Lambda(x^k, n)$ and $h_n^k = \sum_{x^k \in \mathcal{I}_k^n} h(x^k, n)$, respectively.

The daily traffic profile of the whole cellular network is the same and repeats periodically as shown in Fig. 5.1, which can be approximated by a sinusoidal-like periodic behavior as follows [7, 122]:

$$f_{\text{Traffic}} = V_a \cdot [\cos(2\pi(l + 10)/P_e)] + M_e \tag{5.1}$$

where f_{Traffic} is the traffic load function at period $l, l \in [1, \cdots, P_e]$, P_e is the total duration of a traffic load profile, V_a is the standard deviation of traffic profile and M_e is the mean arrival rate. From the BS perspective, it is worthwhile to mention that the traffic load can be considered as one of the QoS (quality of service) requirement because it is the amount of traffic BS should serve.

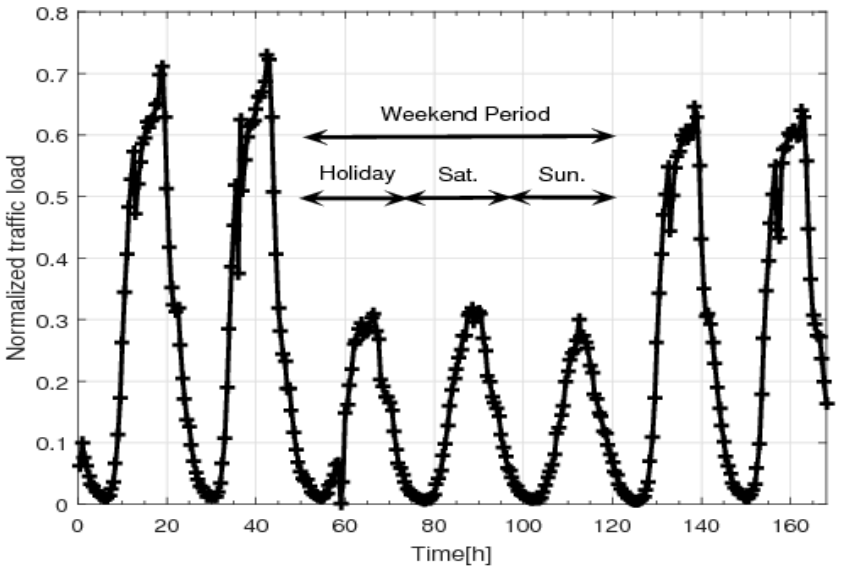


FIGURE 5.1: Normalized real traffic load during one week that are recorded by cellular operator. The data captures voice call information over one week with a resolution of one second in a metropolitan urban area, and are averaged over 30 minute time-scale [123].

BS Selection Rule

We assume that a MS connects with the nearest BS, which would suffer the least path loss during wireless transmission. An active MS located at x_k^{on} is connected with and served by the BS $k \in \mathcal{Y}_n^{on}$ which presents the best received signal strength at each iteration n[1],

$$\arg \max_{k \in \mathcal{Y}_n^{on}} g^n(k, x_k^{on}) P_k^{tx} \tag{5.2}$$

where $g^n(k, x_k^{on})$ is the average channel gain from BS k to active MS at location x_k^{on} at the $n-$th iteration and P_k^{tx} is the transmission power of BS k.

Channel Model

Assuming the physical capacity is defined as in Shannon sense, the service rate of active MS at location x^{on} from BS k at the $n-$th iteration is calculated as

$$\Theta_k(x^{on}, n) = B_a \cdot \log_2 \left(1 + \text{SINR}_k(x^{on}, n)\right) \tag{5.3}$$

where Ba denotes the system bandwidth, $\text{SINR}_k(x^{on}, n)$ is the received signal to interference plus noise ratio (SINR) at active MS location x^{on} from BS k at the $n-$th iteration, and is

[1]Denote that an other user association metric could also be utilized. The optimal user association problems has been well addressed in [59, 77, 140], however, we focus on the BS sleeping scheme rather than user association in this thesis.

defined as

$$\text{SINR}_k(x^{on}, n) = \frac{g^n(k, x^{on})P_k^{tx}}{\phi g^n(k, x^{on})P_k^{tx} + \sum\limits_{m \in \mathcal{Y}_n^{on} - k} g^n(m, x^{on})P_m^{tx} + \sigma^2} \tag{5.4}$$

where σ^2 is the noise power and the term $\sum\limits_{m \in \mathcal{Y}_n^{on} - k} g^n(m, x^{on})P_m^{tx}$ is the interference power experienced by MS x^{on} from its neighboring BSs at the $n-$th iteration. ϕ is the orthogonality (or self interference) factor and $\phi g^n(k, x^{on})P_k^{tx}$ models intra-cell interference [140].

System Load

In order to satisfy the MS's QoS requirement, a BS should provide a certain amount of resources (e.g., time or frequency) according to its traffic load and service rate. From the system's perspective, the load of BS k at the $n-$th iteration is the fraction of resource required to serve the total traffic load in its coverage,

$$\rho_k(n) = \sum_{x^{on} \in \mathcal{I}_k^n} \frac{L(x^{on}, n)}{\Theta_k(x^{on}, n)} \tag{5.5}$$

Power Consumption Model

The total power consumed $P_T(n)$ by each BS k at the $n-$th iteration can be expressed as [40]:

$$P_T^k(n) = a_k P_k^{tx}(n) + P_f^k \tag{5.6}$$

where, P_f^k denotes the static power consumption independent of $P_k^{tx}(n)$ and includes all electronic circuit power dissipations due to site cooling, signal processing hardware as well as battery backup systems. a_k is a BS power scaling factor which reflects both amplifier and feeder losses.

5.2.2 Problem Formulation

In this chapter, we aim at proposing a BS switching algorithm which maximizes network energy efficiency (EE) in bits per joule, and it can be defined as in [5]:

$$EE(n) = \sum_{k \in \mathcal{Y}_n^{on}} \frac{\sum\limits_{x^{on} \in \mathcal{I}_k^n} \Theta_k(x^{on}, n)}{P_T^k(n)} \tag{5.7}$$

The cellular network energy saving problem with BSs switching operation ON-OFF can be formulated as:

$$\max_{\mathcal{Y}_n^{on}} EE(n) = \max_{\mathcal{Y}_n^{on}} \left[\sum_{k \in \mathcal{Y}_n^{on}} \frac{\sum\limits_{x^{on} \in \mathcal{I}_k^n} \Theta_k(x^{on}, n)}{P_T^k(n)} \right] \tag{5.8}$$

$$s.t. \quad 0 \leq \rho_k(n) \leq \rho^{th}, \forall k \in \mathcal{Y}_n^{on} \tag{5.8a}$$

$$\mathcal{Y}_n^{on} \neq \phi \tag{5.8b}$$

$$\Theta_k(x^{on}, n) \geq \Theta^{\min}, \forall x^{on} \in \mathcal{I}_k^n, \forall k \in \mathcal{Y}_n^{on} \tag{5.8c}$$

Similar to [89, 123], we introduce a system load threshold $\rho^{th} \leq 1$ on the system load, in (5.8a), to maintain the dilemma between the system stability/reliability and the energy efficiency as shown in the constraint of the above problem formulation [155]. For instance, low threshold value ρ^{th} indicates that BSs would operate in a more conservative manner with a very low system load on average (i.e., large spare capacity). Thus, there would be less delay and also less call drops for MSs since BSs become more robust to bursty traffic arrivals. On the contrary, with a threshold ρ^{th} close to 1, we expect more energy saving at the cost of slight performance reduction. The constraint in (5.8b) states that there is at least one BS whose activity status is ON mode. We must guarantee a prescribed minimum data rate $\Theta^{\min}$ for MS as shown in (5.8c).

Remark 5.1. At any iteration n, the motivation of the energy efficiency maximization problem defined above in (5.8) is to search for the set of active BSs subject to the system load and data rate constraints. This problem can be proved to be NP-complete by reducing from a vertex cover problem [75, 123]. Finding an optimal set of active BS in this problem arises the following two difficulties: i) it has high computational complexity in order to find the optimal active BS set among 2^Y ON/OFF combinations, specifically when the number of BSs under observation is large; ii) finding an optimal solution of this problem requires a centralized controller which collects the channel state information and traffic load information from all BSs in practice.

5.3 RL Framework for Energy Efficient Network

5.3.1 System Model

We model the dynamic BS switching as a Markov MAB problem as shown in Fig. 5.2, where an arm represents different configuration of BSs status. An Markov decision process (MDP) for BS switching operation is defined as a tuple $\mathcal{M} =< \mathcal{S}, \mathcal{K}, \mathcal{P}, R >$, where $\mathcal{S}$ denotes the state space, $\mathcal{K}$ denotes the action space, $\mathcal{P}$ denotes a state transition probability matrix, and finally R is a reward function associated with $\mathcal{S}$,$\mathcal{K}$ and $\mathcal{P}$. Specifically, at the $n-$th iteration, the traffic load is in state $S(n)$. We consider that centralized controller selects an action (or an am) which

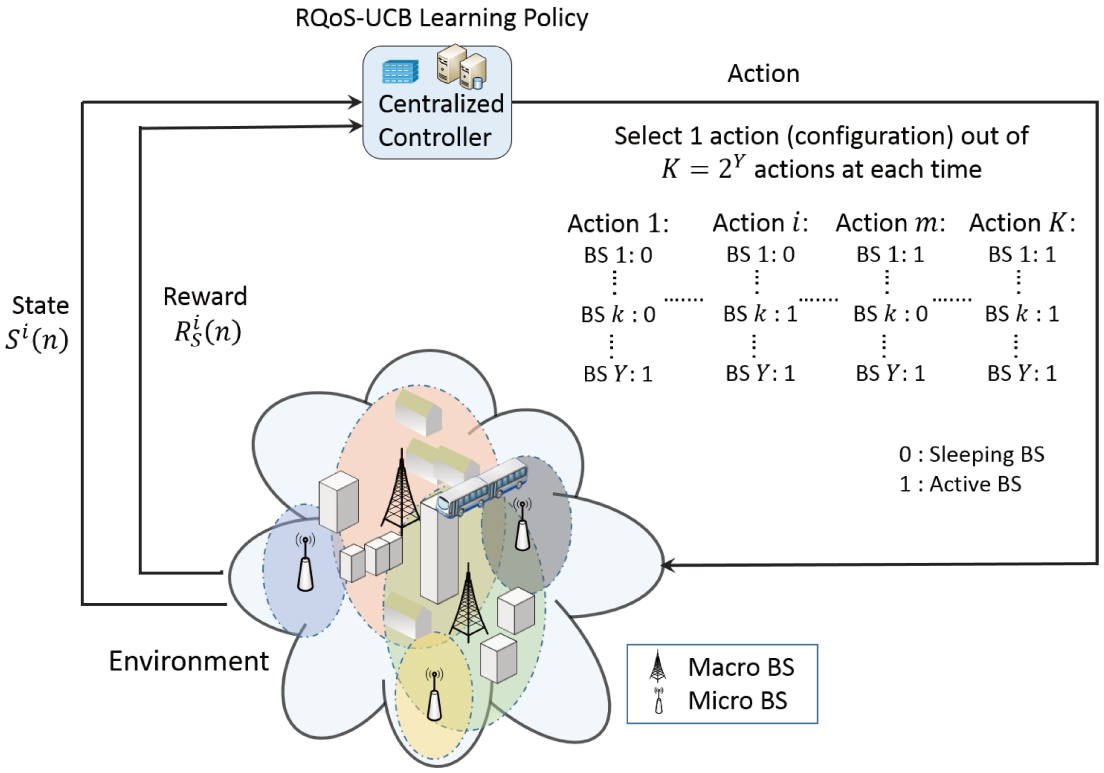


FIGURE 5.2: Reinforcement learning (RL) framework for BS switching operation.

consists of an activity status (ON or OFF) of all BSs for the next iteration. As per the selected action $\mathcal{A}(n) = \{\mathcal{A}^1(n), \cdots, \mathcal{A}^Y(n)\}$ by policy $\mathcal{A}$, the controller decides to turn a BS $k \in \mathcal{Y}$ into sleeping mode if $\mathcal{A}^k(n) = 0$. Otherwise, if $\mathcal{A}^k(n) = 1$, it remains in active mode. Let, $i \in \mathcal{K}$ denotes an index of an arm.

In general, the state $S^i(n)$ of MDP is associated with global traffic load of network, i.e. $S^i(n) = \{S^{i,1}(n), \cdots, S^{i,Y}(n)\} = \{L^1(n), \cdots, L^Y(n)\}$, at time n, which is computationally complex to solve due to large number of states. Hence to ease the analytical and numerical analysis, we consider that the global traffic load state takes value 0 or 1, where 0 represents the case where at least one in (5.8) is not satisfied and while 1 indicates the other cases.

Let assume that the traffic load state $S^i(n)$ emerges with time n, and the traffic load state transforms into $S^i(n+1)$. This behavior is determined by considering a finite volume of varying traffic loads from iteration n to $n+1$ and the related serving BSs, with the transition probability $\mathcal{P}^i = \{P^i_{k,l}, k, l \in \mathcal{S}, i \in \mathcal{K}\}$. Moreover, the steady state distribution of this MDP is defined as $\pi^i_S(n) = \pi^i_S, \forall n$. Meanwhile, the immediate reward with BS switching operation i after n iteration is $R^i_S(n) = EE(n)$, from (5.7), which is generated by the environment in state $S^i(n)$ and is fed back to the BS switching controller. Moreover, $T^i(n)$ refers to the number of times BS switching operation i has been performed by centralized controller. The mean reward μ^i

associated with the BS switching operation i under stationary distribution $\boldsymbol{\pi}_S^i$ is given by: $\mu^i = \sum_{S \in \mathcal{S}} r_S^i R_S^i \pi_S^i$.

The motivation for RL is to find an optimal BS switching operation for the problem as defined in (5.8). The regret of the policy $\mathcal{A}$ is:

$$\Phi^{\mathcal{A}}(n) = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} r_{S^{\mathcal{A}(t)}}^{\mathcal{A}(t)}(t) R_{S^{\mathcal{A}(t)}}^{\mathcal{A}(t)}(t)\right] \tag{5.9}$$

where the expectation $\mathbb{E}$ is taken over the states and observed rewards. Let $S^{\mathcal{A}(t)}$ being the state observed by using the policy $\mathcal{A}$ at iteration t. Like in the previous chapter, BS switching operation whose mean reward is strictly less than μ^* are referred as suboptimal BS switching operation. With the use of the proposed learning framework, the BS switching controller can know the BS switching strategy at last without the prior knowledge of traffic loads.

Restless Quality of Service UCB (RQoS-UCB) Policy

In this chapter, we employ MAB approaches to solve the MDP problem without requiring a priori information on traffic load and specifically we adopt the RQoS-UCB algorithm, proposed in Chapter 4.

At a given iteration n, the policy $\mathcal{A}$ selects a BS switching operation i with the highest index $B^i(n, T^i(n))$. The execution of a BS switching operation i may transform the current state $S^i(n)$ (e.g. traffic load) of network to another state $S^i(n+1)$ with certain probability $\mathcal{P}$, and also feeds back the reward $R_S^i(n)$ (e.g. energy efficiency) to the controller. Then, finally, the policy $\mathcal{A}$ updates the index $B^i(n, T^i(n))$. The algorithm repeats the above procedure until convergence to the best BS configuration during each hour of operation.

It's worthwhile to state that though there may be very high possibility in the beginning that the set of active BSs in current configuration is not sufficient to serve the traffic loads at current iteration n. In this case, the centralized controller can start an emergent response paradigm to quickly turn some BSs in active mode to handle the current traffic load.

5.3.2 Transfer Learning: Transfer RQoS-UCB Policy

In the previous section, we presented the methodology to exploit the classical RQoS-UCB policy as a solution of the BS switching operation problem. The number of arms increases exponentially with the increase in the number of BSs under centralized controller, which leads to increase the convergence time. In this section, we present the solution to the EE maximization problem in such a way the controller utilizes the knowledge of previous learned strategies during historical

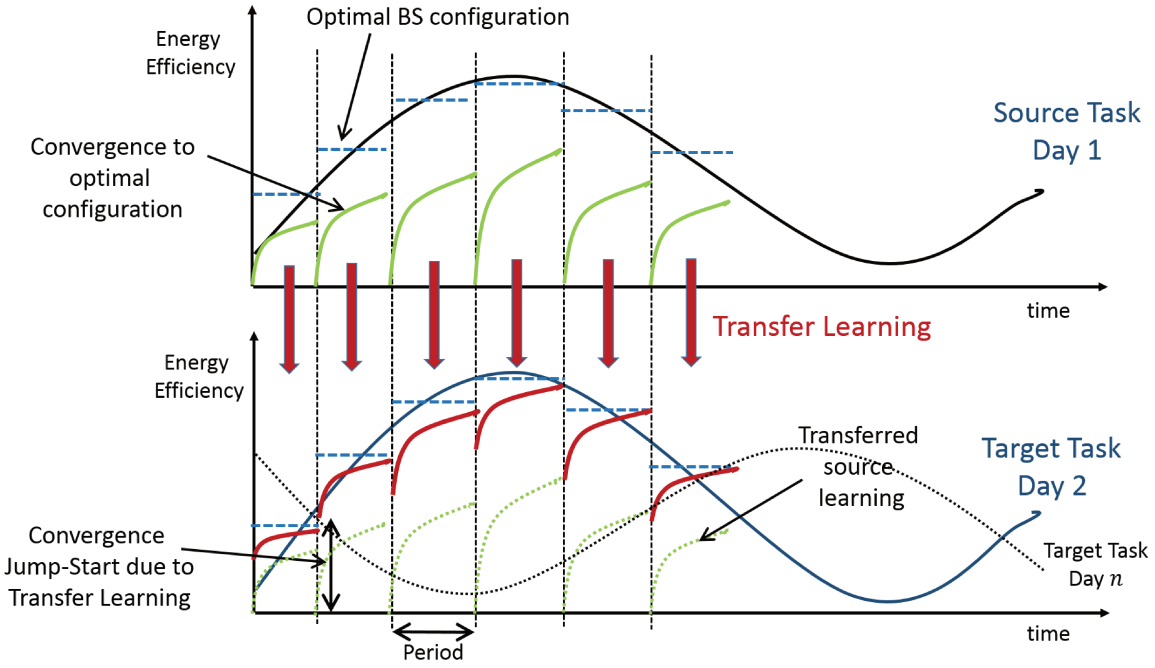


FIGURE 5.3: Transfer learning for Transfer RQoS-UCB (TRQoS-UCB) policy.

periods to find the current optimal BS switching operation. The availability of such historic data increases the curiosity of algorithm and reduces the regret in future learning. In a case of periodic traffic load profile, we consider that the day is divided in 24 periods of 1 hour, and the traffic arrival intensity remains the same during one hour as shown in Fig. 5.3. At the beginning of each period, learning policy reinitializes and learns to find an optimal BS configuration for that period.

The motivation for transfer learning is to utilize previous days (source task, Definition 2.4) the learned knowledge to achieve performance jump-start in the current learning process (target task, Definition 2.5) as shown in Fig. 5.3. In other words, it means that if the BS switching operation is performed based on transferred knowledge, the energy efficiency of the whole system tends to be optimized in the long run. However, the traffic loads in the target task period may be different than that in the source task period, i.e. target task configuration at Day n in Fig. 5.3, hence, instead of directly selecting the historical optimal BS switching operation, the controller in target task utilizes transfered knowledge, learns in current environment and selects an optimal BS switching operation.

Taking the above considerations into account, we propose a new policy update method, named *Transfer Restless QoS-UCB (TRQoS-UCB)* policy detailed in Algorithm 6. In the TRQoS-UCB algorithm, learning experience from previous historical periods is assumed to have been acquired (source task) and will be used to achieve the optimal configuration faster in target task or current learning. In the source policy, the reward achieved in state $S^{i,h}$ from a BS switching operation $i \in \mathcal{K}$ during the source task H_2^i is $r_S^{i,h}(H_2^i)$, where H_2^i is the number of iterations

Algorithm 6 Transfer RQoS-UCB policy

Input: Transferred Observations : h : total historic time, h_2 : total historic time in SB2 block,
H_2^i : total historic time action i has been selected, b^h : total blocks in historical observations
$R_S^{i,h}(t)$, $S^{i,h}(t)$ $\forall i \in \mathcal{K}, 1 \leq t \leq H_2^i$: Reward and state observed in historic data,
$\mathcal{A}^h(t), 1 \leq t \leq h$: Action performed in historic observations
Current policy initialization: $b = 0$, $n = 0$, $n_2 = 0$, $T_2^i = 0$, α, β, ζ^i $R_S^i(0)$ and $S^i(0)$.
Output: $\mathcal{A}(n+1)$

1: **while** (1) **do**
2: $B^{i,h}(n_2, T_2^i(n_2)) = \bar{S}^{i,h}(T_2^i(n_2)) - Q^{i,h}(n_2, T_2^i(n_2)) + A^{i,h}(n_2, T_2^i(n_2))$, $\forall i$
3: $\mathcal{A}(n) = \arg\max_i B^{i,h}(n_2, T^i(n_2))$
4: **while** $S^i(n_2) \neq \zeta^i$ **do**
5: $n = n + 1$ and $\mathcal{A}(n) = i$ // Start SB1 sub-block
6: Activate configuration i and Observe $S^i(n_2)$
7: **end while**
8: $n = n + 1$, $n_2 = n_2 + 1$, $T_2^i(n_2) = T_2^i(n_2) + 1$ and $\mathcal{A}(n) = i$; // End of SB1, start SB2
9: Observe current state $S^i(n_2)$ and update $R_S^i(n_2)$
10: Update $\bar{S}^{i,h}(T_2^i(n_2))$, $Q^{i,h}(n_2, T_2^i(n_2))$ and $A^{i,h}(n_2, T_2^i(n_2))$ as of (5.11), (5.12) and (5.13), respectively
11: **while** $S^i(n_2) \neq \zeta^i$ **do**
12: $n = n + 1$, $n_2 = n_2 + 1$, $T_2^i(n_2) = T_2^i(n_2) + 1$ and $\mathcal{A}(n) = i$; // Start SB2 sub-block
13: Observe current state $S^i(n_2)$ and update $R_S^i(n_2)$
14: Update $\bar{S}^{i,h}(T_2^i(n_2))$, $Q^{i,h}(n_2, T_2^i(n_2))$ and $A^{i,h}(n_2, T_2^i(n_2))$ as of (5.11), (5.12) and (5.13), respectively
15: **end while**
16: $b = b + 1$, $n = n + 1$ and $\mathcal{A}(n) = i$ // Start of SB3 sub-block
17: **end while**

in which operation i has been selected in SB2 block from total observations time h_2 in source task. Without loss of generality, one can consider that $r_S^{i,h}(H_2^i) = S^{i,h}(H_2^i)$ for a given source period H_2^i in SB2 block. Meanwhile, the observed reward associated with energy efficiency by selecting the BS switching operation i is $R_S^{i,h}(H_2^i)$ in H_2^i source task observations.

At the end of each block b in target policy, Algorithm 6 returns a BS switching operation index i which has to be applied in the next block of operation. At the beginning of new block b, the TRQoS-UCB policy selects BS switching operation i which maximizes the policy indexes $B^{i,h}(n_2, T^i(n_2))$ $\forall$ $i \in \mathcal{K}$, step 2, according to three terms:

$$B^{i,h}(n_2, T_2^i(n_2)) = \bar{S}^{i,h}(T_2^i(n_2)) - Q^{i,h}(n_2, T_2^i(n_2)) + A^{i,h}(n_2, T_2^i(n_2)), \quad \forall i \tag{5.10}$$

and $\bar{S}^{i,h}(T_2^i(n_2))$ being the joint empirical mean of the observed states of the BS switching operation i from the source task iterations H_2^i and current iteration n_2, defined as:

$$\bar{S}^{i,h}(T_2^i(n_2)) = \frac{\sum_{t=1}^{T_2^i(n_2)} S^i(t) + \sum_{t=1}^{H_2^i} S^{i,h}(t)}{T_2^i(n_2) + H_2^i}, \forall i. \tag{5.11}$$

The second term, i.e. $Q^{i,h}(n_2, T_2^i(n_2))$, is computed same as previous RQoS-UCB policy but with the source task observations:

$$Q^{i,h}(n_2, T_2^i(n_2)) = \frac{\beta M^{i,h}(n_2, T_2^i(n_2)) \log(n_2 + H_2^i)}{T_2^i(n_2) + H_2^i}, \forall i, \tag{5.12}$$

where,

$$M^{i,h}(n_2, T_2^i(n_2)) = G_{\max}^S - G_S^{i,h}(T_2^i(n_2)), \ \ \forall i,$$

and $G_S^{i,h}(T_2^i(n_2)) = \frac{1}{T_2^i(n_2)} \sum_{k=1}^{T_2^i(n_2)} R_S^i(k) + \frac{1}{H_2^i} \sum_{k=1}^{H_2^i} R_S^{i,h}(k)$ denotes the empirical mean of reward R_S^i collected by applying the BS switching operation i in state S up to the current iteration $T_2^i(n_2)$ plus total source task iteration H_2^i. Moreover, $G_{\max}^S = \max_{i \in \mathcal{K}} G_S^{i,h}(T_2^i(n_2))$ is the maximum reward within the set of BS switching operations from current and historical observations in state S. Finally, the bias term $A^{i,h}(n_2, T_2^i(n_2))$, is defined as

$$A^{i,h}(n_2, T_2^i(n_2)) = \sqrt{\frac{\alpha \log(n_2 + H_2^i)}{T_2^i(n_2) + H_2^i}}, \ \ \forall i. \tag{5.13}$$

5.4 Convergence Analysis: Transfer RQoS-UCB Policy

Analytical results in this section is discussed by considering learning in 1 period only. Also as discussed in previous section, RQoS-UCB policy reinitializes itself at the beginning of new period. We upper bound total number of suboptimal plays in specific period in Theorem 5.1, and this is established under the condition 4.1 on the arms.

Theorem 5.1. *Assume all arms follow condition 4.1. Let* $\pi_{\min}$, $\hat{\pi}_{\max}$, $S_{\max}$, $r_{\max}$, $\varepsilon_{\min}$, $M_{\min}$, $\Delta\mu^i$ *and* $\Omega_{\max}^i$ *defined as in List of Notations Chapter of thesis. Upper bound on total number of suboptimal plays is defined as:*

$$\mathbb{E}[T^{i,h}(n)] \leq \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + 1\right) \left(l^+ + \frac{|S^*|}{\pi_{\min}} \sum_{t=1}^{\infty} (t + H_2^*)^{-2} + \frac{|S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} \left(t + H_2^i\right)^{-2}\right)$$

where,

$$l^+ = \max\left(0, \frac{4\alpha \ln\left(n_2 + H_2^i\right)}{(\Delta\mu^i)^2} - H_2^i\right) \tag{5.14}$$

Proof. Proof of Theorem 5.1 is achieved by following the same steps as in Theorem 4.1 with consideration of transferred observations. □

Note that the above bound reduces to the bound of RQoS-UCB policy when the transferred knowledge is not available (i.e. $H_2^i = 0, \forall i$). At first glance of the index calculation, one might

think that $\sqrt{\frac{\alpha \ln (n_2+h_2)}{T_2^i(n_2)+H_2^i}}$ is the most natural choice for the index calculation with transferred knowledge.

Corollary 5.1. *It can be easily noticed from the proof of Theorem 5.1, that taking into an account* $A^{i,h}(n_2, T_2^i(n_2)) = \sqrt{\frac{\alpha \ln (n_2+h_2)}{(T_2^i(n_2)+H_2^i}}$ *for index calculation leads to the following bound:*

$$\mathbb{E}[T^{i,h}(n)] \leq \left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + 1\right)\left(\max\left(0, \frac{4\alpha \ln (n_2+h_2)}{(\Delta\mu^i)^2} - H_2^i\right) + \frac{|S^*|+|S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} (t+h_2)^{-2}\right)$$

The setting in Corollary 5.1 has two major drawbacks: i) it does not take into account that there could be different numbers of selections for different actions in the historical observations. ii) when H_2^i (number of selection of action i) is very small and h_2 (total number of selection) is quite large in historical data, the above presented bound can be worse than the one derived in Theorem 5.1.

5.5 Numerical Results and Analysis

We verify the effectiveness of our proposed algorithm with extensive simulations under practical configurations similarly to [89, 123, 141]. We consider a heterogeneous cellular network (HCN) topology consisting of 5 macro and 5 micro BSs in the area of 5×5 Km^2. A snapshot of cell coverage, when all BSs are ON, is plotted in Fig. 5.4. Furthermore, we consider that arrival rate at location x^k is PPP with intensity $\Lambda(x^k, n)$ and MSs call holding time (or file size) is $1/[h(x^k, n)] = 100$ Kbyte. Table 5.1 summarizes all the parameters used for the simulations.

Parameter description	Value
Simulation area	5km × 5km
Maximum transmission power	Macro BS: 20W, Micro BS: 1W
Maximum operational power	Macro BS: 865W, Micro BS: 38W
BS Height	Macro BS: 32m, Micro BS: 12.5m
Intra-cell interference factor	0.01
Channel bandwidth	1.25MHz
Path loss model	COST 231
Arrival rate $\Lambda(x^k)$ in source task	0.05×10^{-4}
Arrival rate $\Lambda(x^k)$ in target task	from 0.05×10^{-4} to 2×10^{-4}
MSs call holding time $1/[h(x^k)]$	100Kbyte
System load threshold ρ^{th}	0.6
Minimum bit rate requirement $\Theta^{\min}$	122kbps
Exploration parameters of RQoS-UCB	$\alpha = 0.25$ and $\beta = 0.32$

TABLE 5.1: Used Simulation Parameters

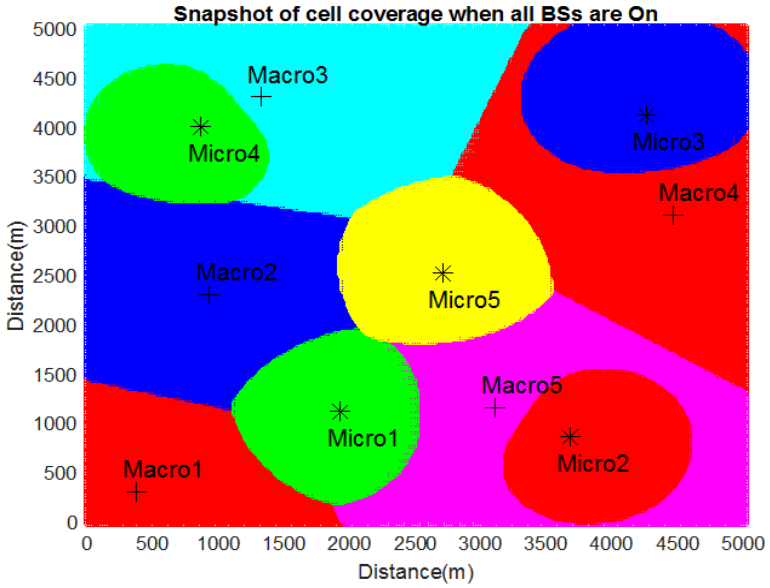


FIGURE 5.4: Snapshot of cell coverage when all BSs (BSs 1-5 are macro BSs, and BSs 6-10 are micro BSs) are ON.

Maximal values of macro and micro BSs transmission power are set to 43 dBm and 30 dBm, respectively. As per the linear relationship between the transmission and maximum operational power consumption for macro and micro BSs [141], the maximum operational (or total) power consumption for macro and micro BS are 865 W and 38 W, respectively[2]. To model the channel propagation environment, we consider the modified COST 231 path loss model with macro BS height as 32 m and micro BS height as 12.5 m, similar to [89, 123, 141]. Other parameters for our numerical analysis follows IEEE 802.16m evaluation methodology document [60, 89, 123, 141]. In order to guarantee the system reliability, the system load threshold $\rho^{th} = 0.6$ is considered for all BSs [123]. To maintain the quality of service (QoS), the target blocking probability is set to 1%, with a minimum bit rate $\Theta^{\min}$ requirement of 122 kbps [5]. In this simulation, we set ϕ to be 0.01 which refers to the intra-cell interference factor. As for the proposed TRQoS-UCB and RQoS-UCB policies, the exploration parameters $\alpha = 0.25$ and $\beta = 0.32$ are set as per the baseline QoS-UCB policy. As per [141], a homogeneous user distribution with intensity $\Lambda = 10^{-4}$ leads to a system load corresponding to about 10% of BSs utilizations in a case where all BSs are turned ON. Therefore, we analyze the effect of the traffic load on the performance of the proposed policy under homogeneous traffic with varying Λ from 0.05×10^{-4} to 2×10^{-4}.

[2]The linear relationship [141] adopted is as follows: $P_T = a \cdot P^{tx} + b$, where $a_{\text{macro}} = 22.6$ and $b_{\text{macro}} = 412.4$ W, $a_{\text{micro}} = 5.5$ and $b_{\text{micro}} = 32$ W. The model and the parameters are derived based on the data sheets of several existing GSM and UMTS BSs which were analyzed at the component level, e.g., power amplifier, antenna, cooling equipment, etc

Learning Algorithms	Operation manner	Computational complexity	Convergence speed
RQoS-UCB	Centralized	$\mathcal{O}(2^{(\lvert\mathcal{Y}\rvert)})$	Fast
TRQoS-UCB	Centralized	$\mathcal{O}(2^{(\lvert\mathcal{Y}\rvert)})$	Very Fast
Decentralized greedy [78, 155]	Decentralized	$\mathcal{O}(\lvert\mathcal{Y}\rvert^2)$	Very Fast
Actor Critic (ACT) [90, 142]	Centralized	$\mathcal{O}(\lvert\mathcal{S}\rvert 2^{(\lvert\mathcal{Y}\rvert)})$	Slow
TACT [89]	Centralized	$\mathcal{O}(\lvert\mathcal{S}\rvert 2^{(\lvert\mathcal{Y}\rvert)})$	Slow
Exhaustive search (ideal policy) [78]	Centralized	$\mathcal{O}(\lvert\mathcal{S}\rvert^2 2^{(\lvert\mathcal{Y}\rvert)})$	–

TABLE 5.2: Algorithms Comparison.

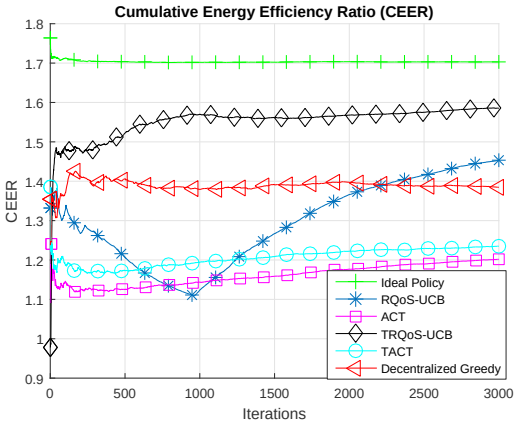
5.5.1 Convergence Analysis

Initially, we analyze and compare, in Fig. 5.5, the convergence behavior of the proposed RQoS-UCB and TRQoS-UCB algorithms with *Actor CriTic (ACT)* [90], *decentralized greedy* [155] and *Transfer Actor CriTic (TACT)* [89] algorithms. We made each algorithm to learn and to find out optimal configuration in one period which means that traffic arrival intensity remains fixed during the simulation. However, the traffic loads in the target task period may be different than that in the source task period, as illustrated in Fig. 5.3. Moreover, Table 5.2 summarizes the complexity and convergence speed of each algorithms. All of the results presented below are obtained by running a simulation over 10^2 runs, each with a duration about 3000 iterations in one period.

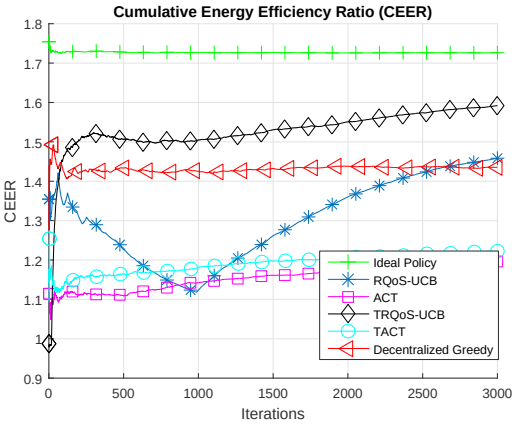
In order to validate the convergence behavior of our proposed RQoS-UCB and TRQoS-UCB algorithms, the global optimal solution achieved by an exhaustive search (ideal policy) is also plotted in Fig. 5.5. First, we present the *cumulative energy efficiency ratio (CEER)* as the performance metric, which is defined as:

$$\text{CEER} = \frac{\text{energy efficiency of BSs when using learning policy}}{\text{energy efficiency of BSs when all BSs are turned ON}}$$

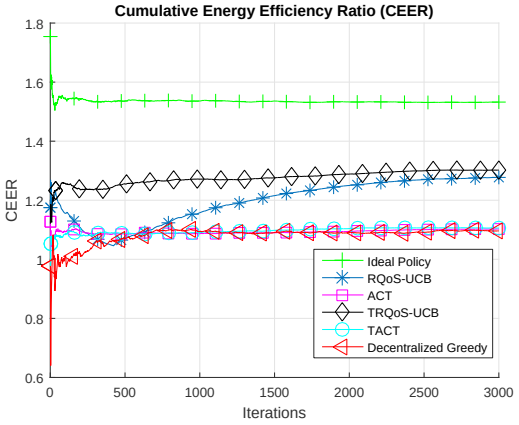
Fig. 5.5 also states that our proposed TRQoS-UCB and RQoS-UCB algorithms are converging towards the ideal policy, while the state-of-the-art ACT, TACT and decentralized greedy algorithms converge to some suboptimal configuration. Furthermore, the performance of ACT and TACT algorithms is worse than the proposed policies, since it has to find the optimal solution among significantly more number of state and action pairs, and thus they requires significantly higher number of iterations to converge to the ideal policy. Moreover, as expected, CEER continues to increase as iterations running, because controller has more observations and stronger confidence bound for the optimal action. Unfortunately, since the proposed learning schemes are performed without the knowledge of traffic loads a priori, their performance are inferior to



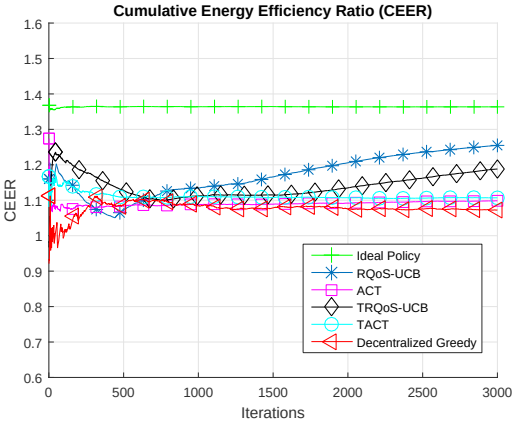
(a) $\Lambda^{target} = 0.05 \times 10^{-4}$, $\Lambda^{source} = 0.05 \times 10^{-4}$



(b) $\Lambda^{target} = 0.5 \times 10^{-4}$, $\Lambda^{source} = 0.05 \times 10^{-4}$



(c) $\Lambda^{target} = 1 \times 10^{-4}$, $\Lambda^{source} = 0.05 \times 10^{-4}$



(d) $\Lambda^{target} = 2 \times 10^{-4}$, $\Lambda^{source} = 0.05 \times 10^{-4}$

FIGURE 5.5: Performance comparison under various homogeneous Poisson point process traffic intensity with transferred knowledge estimated from a source task with traffic intensity $\Lambda^{source} = 0.05 \times 10^{-4}$.

that of the ideal policy, especially at the beginning of simulations. However, we can see that the gap is compensated as learning proceeds or when the TRQoS-UCB scheme is applied with the transferred knowledge.

Furthermore in Fig. 5.5, we also analyze the impact of transferred learning from a fixed source task ($\Lambda^{source} = 0.05 \times 10^{-4}$) to several target tasks with intensities varying from $\Lambda^{target} = 0.05 \times 10^{-4}$ to $\Lambda^{target} = 2 \times 10^{-4}$. As we can see in Fig. 5.5(a), when the traffic arrival intensity of source task and target task are the same, TRQoS-UCB achieves performance jump start in the beginning and quickly converges toward the ideal policy. On the contrary in Fig. 5.5(d), when the traffic arrival intensity of source and target tasks are significantly different, the transferred knowledge impacts the learning in a negative manner and thus TRQoS-UCB achieves similar performance as of classical RQoS-UCB policy. As the dissimilarity between source and target task increases, CEER performance of TRQoS-UCB and RQoS-UCB get closer because the transferred knowledge acquired by TRQoS-UCB policy becomes useless. From this

analysis, we can state that the temporal knowledge transfer improves the convergence speed of classical MAB approaches, but it can have negative effect if traffic loads in source and target environments are significantly different.

5.5.2 Performance under Periodic Traffic Load

We also investigate the effectiveness of the proposed learning framework when traffic loads periodically fluctuates. As stated before, real traffic load follows a periodical pattern and can be approximated by a sinusoidal function as in (5.1). All of the results presented below are obtained by running a simulation over 10^2 runs, each with a duration about 3000 iterations in one period.

Fig. 5.6 compares the network EE for TRQoS-UCB and other state-of-the-art algorithms over fluctuating traffic load during the day. For example, EE for the proposed TRQoS-UCB, RQoS-UCB policies and state-of-the-art ACT and TACT algorithms in peak load period (day time) are nearly the same, however, we can see that BS set selection becomes more crucial as the traffic load decreases (night time) because less number of active BSs are required to serve all MSs, and thus more intelligence is needed for the BS set selection in order to increase the overall network throughput and finally EE of network. Fig. 5.6 depicts that the proposed TRQoS-UCB achieves significantly higher EE compared to other algorithms from literature, and it also confirms that transferred learning always improves the performance than the baseline RQoS-UCB policy. We can also state that the EE gap between TRQoS-UCB and ideal (exhaustive search) policy is very thin which validates its convergence behavior. Fig. 5.6 also presents that our proposed algorithms tracks the temporal variation of the average traffic intensity in an effective way.

Next, Fig. 5.7 depicts the average percentage of energy savings achieved by the proposed TRQoS-UCB policy during one day. As shown in Fig. 5.7, a large amount of energy savings is achieved by the proposed TRQoS-UCB policy, e.g, about 70% during low traffic load period (night time). Moreover, the difference between the exhaustive search (ideal policy) and proposed TRQoS-UCB policy is less than 5%. On the contrary, ACT, TACT and RQoS-UCB algorithms achieve only 60% of energy saving. It is also clear from Figs. 5.7 and 5.6 that decentralized greedy approach achieves higher percentage of energy saving by putting more BSs into sleep mode, but this improvement comes at the cost of user experience and thus achieves comparatively less network energy efficiency.

Fig. 5.8 presents the average number of active BSs along with realistic fluctuating traffic load during the day. It is clear that average number of active BSs during the peak load (in the afternoon) is more compared to night time for all algorithms which depicts that they are able to follow the traffic fluctuation pattern. Furthermore, ACT and TACT algorithms activate comparatively less number of BSs than other algorithms, and hence energy efficiency achieved

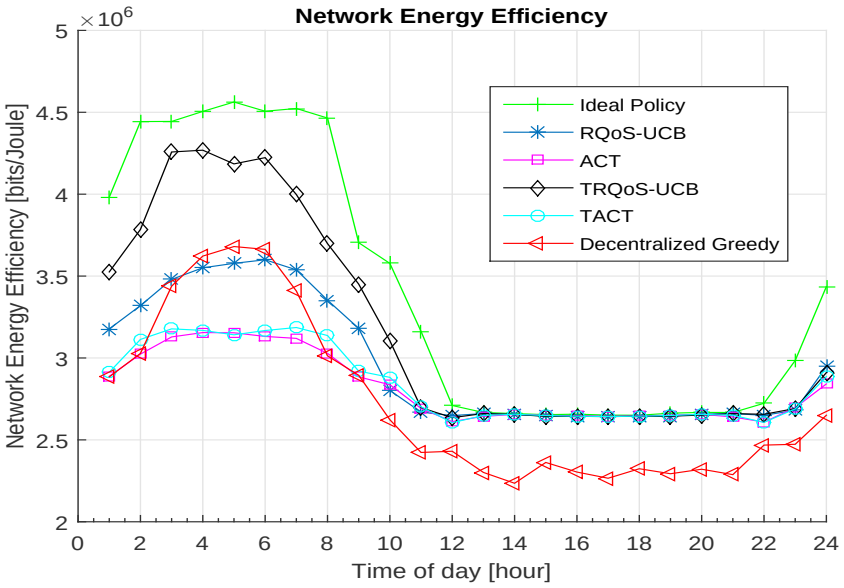


FIGURE 5.6: Network energy efficiency (EE) with several learning algorithms with respect to time of the day

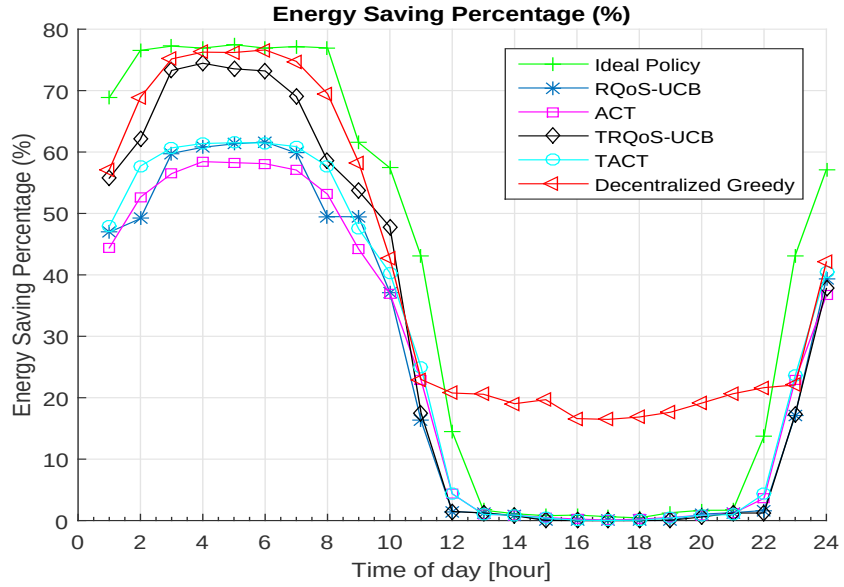


FIGURE 5.7: Average percentage of energy saving achieved by several learning algorithms with respect to time of the day.

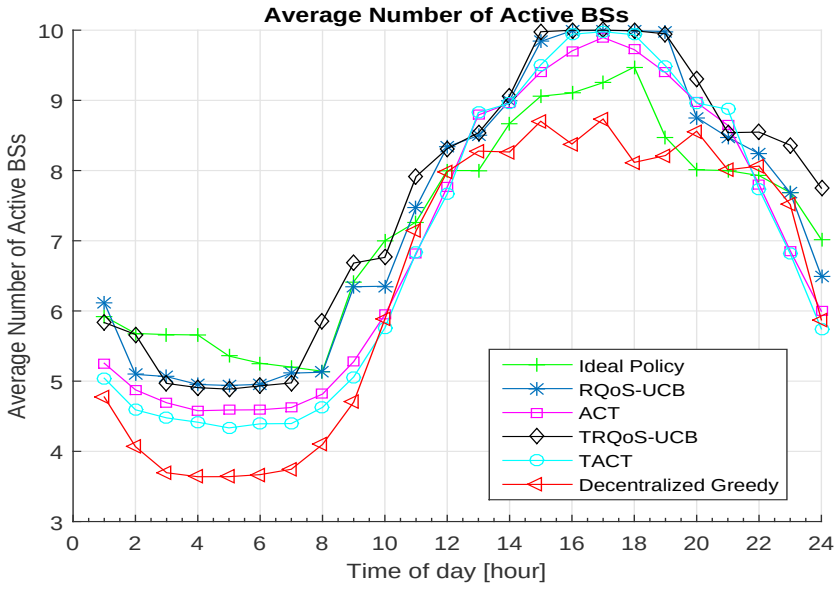


FIGURE 5.8: Average number of active BSs suggested by several learning algorithms with respect to time of the day.

with these algorithms are less compared to others as shown in Fig. 5.6. Besides, the proposed policies, i.e. RQoS-UCB and TRQoS-UCB, activate more micro BSs than macro BSs to cope with the varying traffic load and to save energy, whereas other algorithms reduce the total number of active BSs by putting more micro BSs to sleep. Configuration of BSs operation with TRQoS-UCB policy is better than with other algorithms because those BSs which contribute to the highest interference are switched off hence leveraging higher energy efficiency.

After validating feasibility of the proposed learning framework to maximize EE, we analyze another important performance metric which is per-flow delay measured using the Little's law as defined in [77, 86]:

$$\text{Delay}(n) = \sum_{k \in \mathcal{Y}_n^{on}} \frac{\rho_k(n)}{1 - \rho_k(n)}, \qquad [\text{sec}].$$

Fig. 5.9 depicts the per-flow delay performance with respect to the time of day for several learning algorithms. We can notice that during the day, high traffic load regime, time per-flow delay also increases for all algorithms. We have seen earlier in Fig 5.6 that ideal policy (exhaustive search approach) achieves little gain in energy efficiency compared to proposed policies in low traffic load regime (night time), but it is clear from Fig. 5.9 that ideal policy has higher per-flow delay compared to the proposed policies. Moreover, we can state that all algorithms, i.e.RQoS-UCB, TRQoS-UCB, ACT, TACT, experience a similar level of per-flow delay, whereas decentralized greedy has significantly higher level of per-flow delay compared to others.

Finally, in Fig. 5.10, we analyze the average number of switch that is performed during the day. This analysis is as important as the other analysis because of some practical constraints,

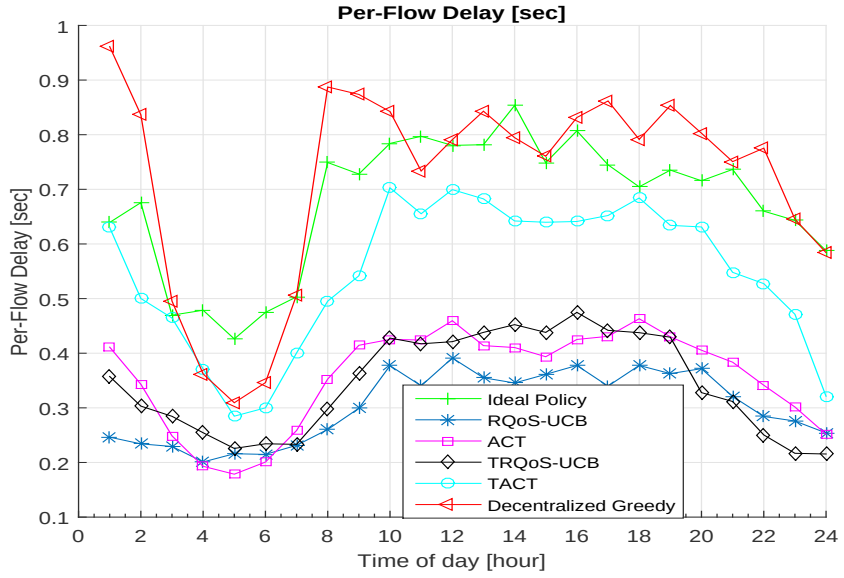


FIGURE 5.9: Per-flow delay is seconds with several learning algorithms with respect to time of the day

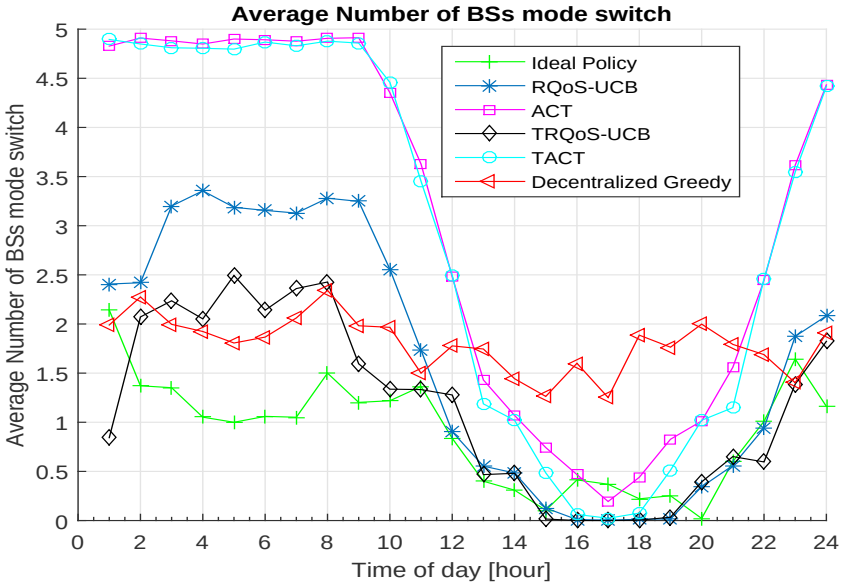


FIGURE 5.10: Average number of BS mode switch required at each iteration during one hour with respect to time of the day.

i.e. time needed to turn ON/OFF the power amplifier, lifetime of power amplifier. If learning requires to switch the mode of more number of BSs at each iteration, then it will significantly reduce the lifetime of power amplifier. As shown in Fig. 5.10, the proposed policy requires only 2 switches at each iteration in low load period (night time in average) which is significantly less compared to 5 switches with ACT and TACT algorithms. The performance of decentralized greedy algorithm remains nearly the same all along the day irrespective of the traffic load.

5.6 Conclusion

In this chapter, we focused on the problem of BS switching operation for EE maximization in heterogeneous wireless cellular network. In particular, we extended our RQoS-UCB algorithm to find the optimal BS switching configuration to increase the global EE of the network. Furthermore, in order to use the temporal dependency of traffic loads, we proposed a transfer RQoS-UCB (TRQoS-UCB) algorithm to improve the baseline RQoS-UCB algorithm by transferring the learned knowledge from historical periods.

Our proposed algorithm has been proved to converge to an optimal solution given certain restrictions that arise during the learning process. The extensive numerical analysis presents effectiveness and robustness of our proposed energy efficiency maximization scheme under periodic practical traffic configuration. We empirically showed that the proposed simple transfer learning algorithm can not only achieve optimal solution but also can achieve significant energy savings up to 70% when traffic load is low.

Future work may be to handle the even more challenging problem of spatial knowledge transfer. Furthermore, future communication systems tend to be hyper-dense with massive small cell base stations serving different types of traffic. Dynamic fully distributed topology management, compared to centralized architecture, would be better to effectively manage such a complex architecture, in order to provide system capacity, reduce power consumption and further reduce cooperation overhead.

Chapter 6

Proof of Concept: Learning for OSA in Real Radio Environment

Contents

6.1 Introduction 101
6.2 Validation of Learning on Real HF Measurements 102
6.2.1 Experimental Results and Analysis 102
6.3 Demonstrators: OSA on Real Radio Spectrum 106
6.3.1 USRP Testbed 107
6.3.2 Experimental Results and Analysis 110
6.4 Conclusion 114

6.1 Introduction

Previous chapters dealt with MAB model and its application to OSA problem. Several low complexity single-player and multi-player learning policies have been also introduced to solve MAB problem. In this chapter, the validation of MAB learning policies in the real environmental conditions is presented. Two scenarios are considered; i) real measurement database of HF band rather than simulated channel occupation patterns is used, ii) testbed has been developed to evaluate various learning policies for OSA in decentralized networks and with real radio signals.

The organization of this chapter is as follows. In Section 6.2, the upper confidence bound (UCB1) algorithm is applied on a scenario where the occupation statistic is drawn from real HF measurements. Then, Section 6.3 presents a proof-of-concept system for OSA with multiple secondary users (SUs). The proposed as well as existing state-of-the-art MAB learning policies have been implemented on the testbed consisting of USRPs from Ettus Research, GNU

radio companion for primary users (PUs) and MATLAB/Simulink for SUs. Finally, Section 6.4 concludes the chapter.

6.2 Validation of Learning on Real HF Measurements

The HF band offers a worldwide coverage since beyond line-of-sight (BLOS) links have been established by using the ionosphere as a passive reflector. This trans-horizon behavior allows the communication between remote sites without using a satellite link, making this band usually considered as an alternative to satellite links in mobile environments. The HF band ranging from 3 to 30 MHz, is mainly used for military communications, but also for aeronautical, maritime, and amateur communications. However, the main limitation is due to the existence of multiple interferences between HF users.

Particularly, in a cognitive radio (CR) context, Dynamic Spectrum Access (DSA) [162] allows HF users to access unoccupied channels by legacy users or PUs at a particular time and area. The main focus of this work is the dynamic access to the HF spectrum by learning from the environment and selecting the channel with better availability to maximize transmission opportunities without colliding with other HF users. In this work, we present the application of the MAB learning policy, UCB1, presented in [2, 15] to provide secondary HF users with a dynamic access to the spectrum as introduced in [67]. Such a mechanism helps HF users to decide which is the best channel in terms of availability at each particular time. It could be used with wideband transceivers or with single-channel transceivers with an automatic frequency tuning. The application of cognitive radio to HF communications is validated in this chapter with the database `HFSA_IDeTIC_F1_V01` [107] of real measurements of the HF band.

We took benefit from a collaboration with IDeTic lab of University of Gran Canaria, Las Palmas, Spain to validate our learning solution based on MAB on real signals, e.g. not an artificially made iid or Markovian primary network. We prove the consistency, efficiency and robustness of UCB class of approaches for spectrum learning and radio spectrum access improvement.

6.2.1 Experimental Results and Analysis

The performance metric used in this chapter is the successful transmission rate achieved by UCB1 algorithm. Moreover, UCB1 is compared to other approaches, same as Section 4.5; 1) random uniform channel selection, 2) best channel selection, 3) worst channel selection, defined as the successful transmission rate achieved by the worst channel in terms of availability, and 4) best opportunistic selection. All of the results presented below are obtained by running a simulation over 10^2 runs.

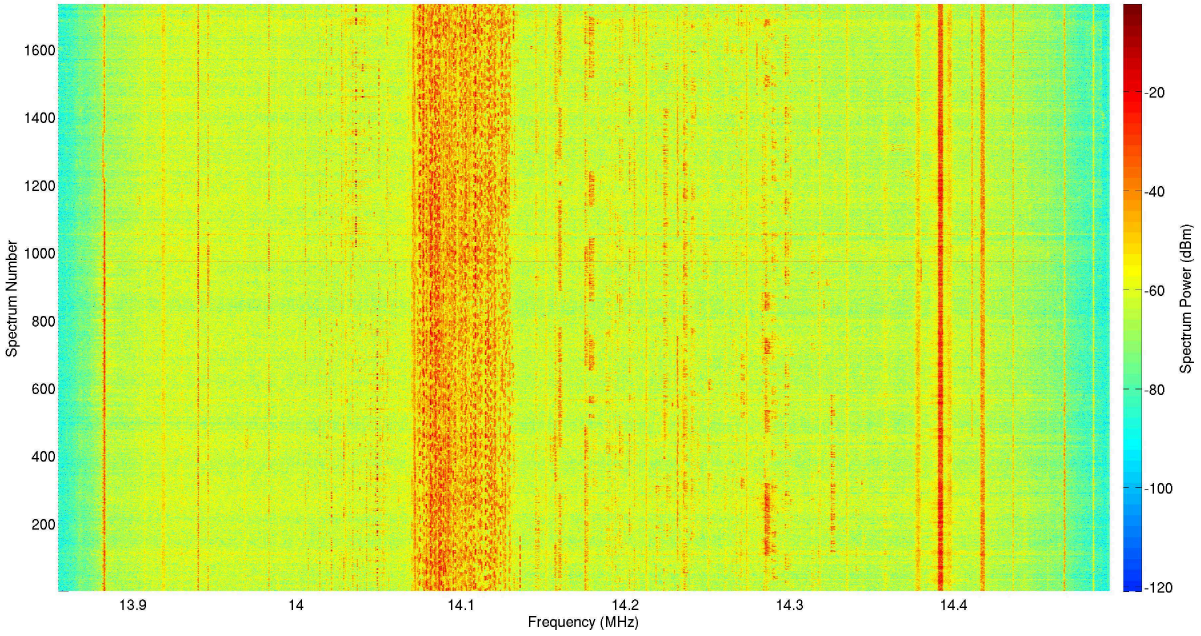


FIGURE 6.1: Example of a measurement of the high activity scenario (acquired at weekends) of the `HFSA_IDeTIC_F1_V01` [107].

The database `HFSA_IDeTIC_F1_V01` [107] consists of 600 kHz bandwidth measurements with a duration of 10 minutes of the 14 MHz band, which covers an amateur band and other frequencies used by non-amateur HF stations. These measurements are segmented in samples that represent the activity in a 3 kHz channel for 2 seconds. Fig. 6.1 illustrates a measurement on the HF band conducted over a weekend and presents a high activity scenario.

For the UCB1 evaluation, we have selected the amateur bands that are heavily occupied during amateur contests at weekends (around 14.1 MHz in Fig. 6.1) to test the algorithm in the worst conditions. In this work, the following model has been used for the OSA: at each time slot, only one channel of a group is sensed; if the selected channel by the algorithm is free, it will be used to transmit and it will get positive reward, i.e. 1. However, if it is occupied by another user, there will not be any reward, i.e. 0, and it will not transmit to avoid a collision with PU's transmission.

Two parameters must be set in advance to evaluate the UCB1 algorithm, the number of channels per group K, and the exploration-exploitation parameter α. When the number of channels per group K is small, there are not many transmission opportunities for SU. Otherwise, if K is larger, the UCB1 needs more time to learn. Since the number of measurements is limited, the value of K is not too large to allow UCB1 to converge. For that reason and in order to find a trade-off between the number of opportunities in a group and the learning time, the number of channels per group K is set to 8 in this chapter. Moreover, the selected value of the exploration-exploitation factor α is 0.4, i.e. UCB1 mainly exploits the previous best channel although there is a short exploration phase.

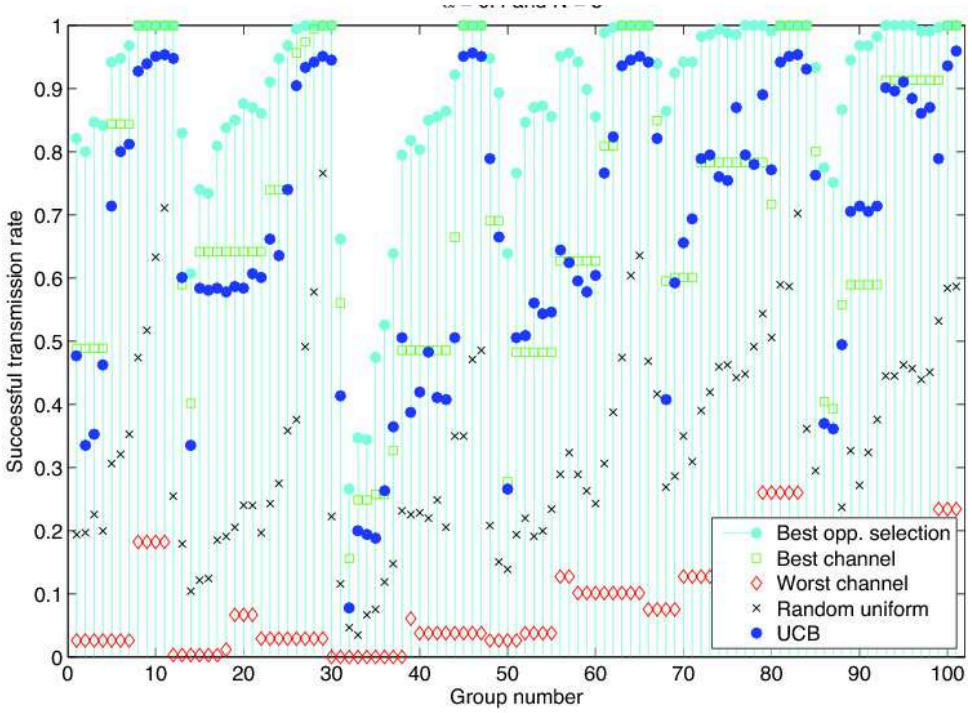
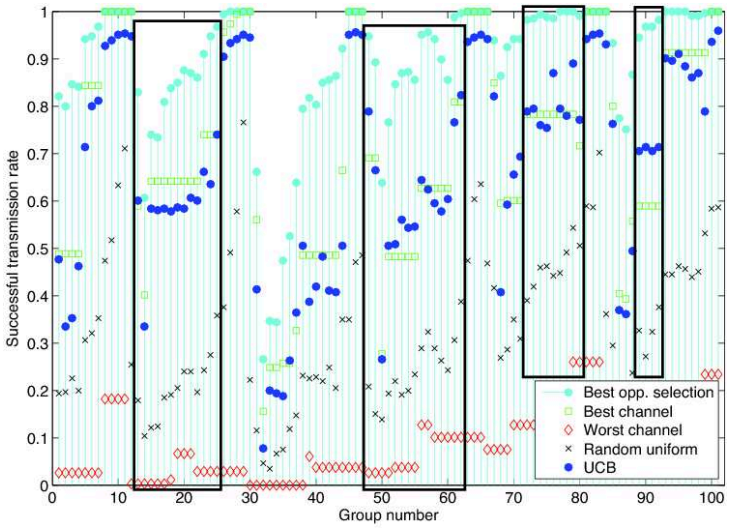


FIGURE 6.2: Comparison of the successful transmission percentage per group with $\alpha = 0.4$ and $K = 8$.

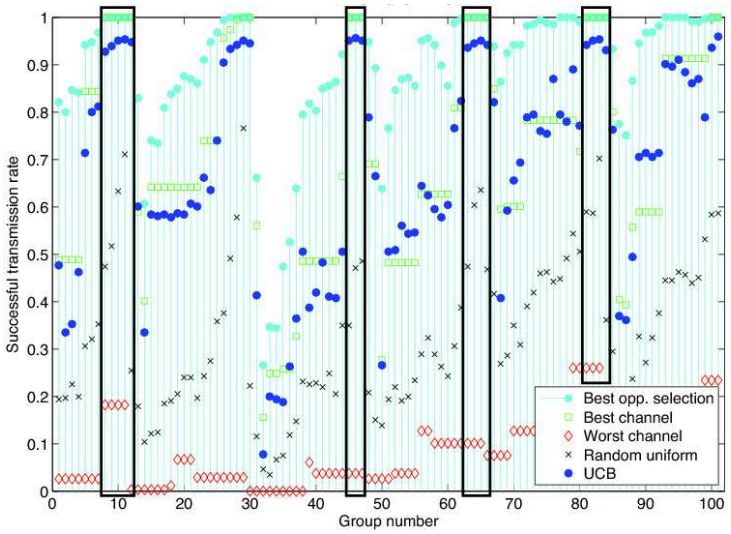
A channel is declared free, if the energy detector output, applied to `HFSA_IDeTIC_F1_V01` database [107], is below a threshold and then the channel is used to transmit. The comparison of the successful transmission percentage achieved by UCB1 and the different approaches is depicted in Fig. 6.2 where the x-axis is the numbering of group each containing $K = 8$ channels. Fig. 6.2 compares the successful transmission percentage achieved by UCB1 with several approaches, such as random uniform channel selection, best channel selection, worst channel selection and best opportunistic selection algorithms. In spite of the fact that the comparison is done over groups of $K = 8$ channels in Fig. 6.2, it is a representative example of the different cases that can be found in this environment.

It is shown in Fig. 6.2 that UCB1 always outperforms a random uniform channel selection. Hence, these results demonstrate that it is worth applying MAB learning approaches to dynamically access the HF band. For groups where the successful transmission rate achieved with the best channel selection is lower than the best opportunistic selection approach, it is shown that there are more transmissions opportunities that SU could take advantage of. However, UCB1 cannot always outperform the best channel selection approach.

For a better understanding of this situation, those groups with this behavior are selected and framed in Fig. 6.3(a). In these groups, the successful transmission rate achieved also depends on the activity pattern of the whole set of channels inside the group that directly affects the performance of the algorithm.



(a) UCB1 outperforms the best channel or it is close to its performance.



(b) UCB1 cannot reach the best opportunistic selection and the best channel performances.

FIGURE 6.3: Detail of the comparative of the successful transmission rate per group with $\alpha = 0.4$ and $K = 8$.

Finally, UCB1 is also compared with the maximum performance that a CR should be able to reach, named the best opportunistic selection, and as expected, UCB1 cannot reach it due to the exploration phase but it is close to it in some groups. These are the groups depicted in Fig. 6.3(b), where there is a channel available for the whole experimentation time. Therefore, the successful transmission rate of the best opportunistic selection is equal to the best channel selection policy and equal to 1 in the comparative in Fig. 6.3(b). In these groups, UCB1 loses the opportunity to transmit due to the exploration phase of the algorithm. Nevertheless, its successful transmission rate remains always higher than random uniform selection.

Besides the graphical description of the performance of UCB1, a numerical evaluation of UCB1

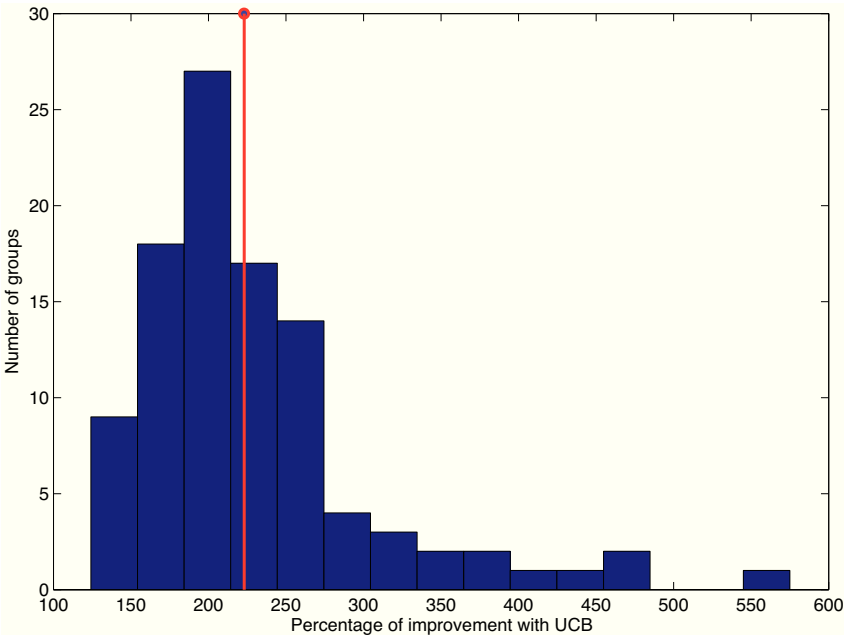


FIGURE 6.4: Histogram of the percentage of improvement with UCB1 respect to random selection with $\alpha = 0.4$ and $K = 8$.

is introduced to compare how good UCB1 is with respect to the different approaches described in Section 6.4. In this work, we quantified the benefits of applying UCB1 by computing the percentage of improvement in terms of the successful transmission rate with respect to a random uniform selection of channels. The histogram of the percentage of improvement of UCB1 w.r.t. a random uniform selection is depicted in Fig. 6.4. It is shown that there is a high variability in the percentage of improvement, from 125% to 500%, due to the fact that it is dependent on the activity pattern of the channels of each group. Nevertheless, we can conclude that in average the percentage of improvement is 228% which reflects that the successful transmission rate achieved with UCB1 is outperforming in three times the successful transmission rate achieved with a random uniform selection. Moreover, as it is depicted in Fig. 6.4, UCB1 improves by 150% the performance of a random uniform selection in terms of successful transmission rate in more than 91% of the evaluated groups. Even though the improvement is lower than the mean improvement, it reflects that the successful transmission rate achieved by UCB1 is outperforming in 2.5 times the successful transmission rate achieved by a random uniform selection.

6.3 Demonstrators: OSA on Real Radio Spectrum

As discussed, one of the most promising DSA approach to overcome the spectrum scarcity is OSA [162]. In this chapter we present a demo which focuses on learning for MAB framework that models OSA in the single and multi-player settings in order to enable SUs to identify and transmits over vacant channels, and keep the channel switching cost (CSC) as minimum as

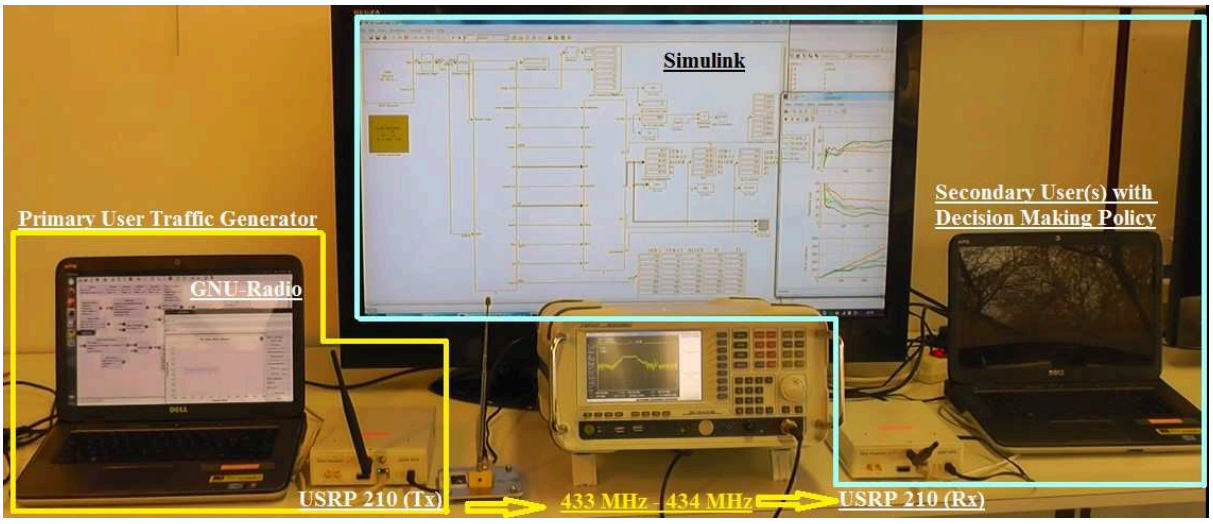


FIGURE 6.5: Left hand side (laptop with GNU Radio + USRP) is generating primary user traffic on 8 channels (Tx). Right hand side (laptop with Simulink + USRP) is a secondary user employing energy detector for channel sensing and online learning algorithm based learning policies (Rx). A spectrum analyzer shows the RF signals.

possible. Here, CSC stands for the total penalty incurred in terms of delay, power, hardware reconfiguration and protocol overhead when SU switches from one frequency channel to another. From the energy efficiency perspective, CSC should be as minimum as possible. Design of such learning policies for CR networks is a challenging task and one of the objective the work presented in this demo. We propose a USRP based testbed for analyzing the performance of learning policies for OSA in CR networks. To the best of our knowledge, the proposed testbed is the first worldwide proof-of-concept which compares the performance of various learning policies using real radio signals.

6.3.1 USRP Testbed

The USRP testbed is shown in Fig. 6.5 and is a significant extension of the testbed in [34, 35, 132]. The extension has been performed during this thesis work and in cooperation with Sumit DARAK, post-doc in SCEE first and now assistant professor in IIIT Delhi. It consists of two units: 1) Left hand side unit is primary user traffic generator, and 2) Right hand side unit acts as SUs. Both the units are discussed in details next.

Primary User Traffic Generator

The chosen design environment for the PU traffic generator is GNU Radio Companion (GRC) and the hardware platform is made of a laptop and an USRP from Ettus Research. The main reason for choosing GRC is the precise control on each parameter of the transmission chain compared to other environments. The detailed design of the proposed PU traffic generator is

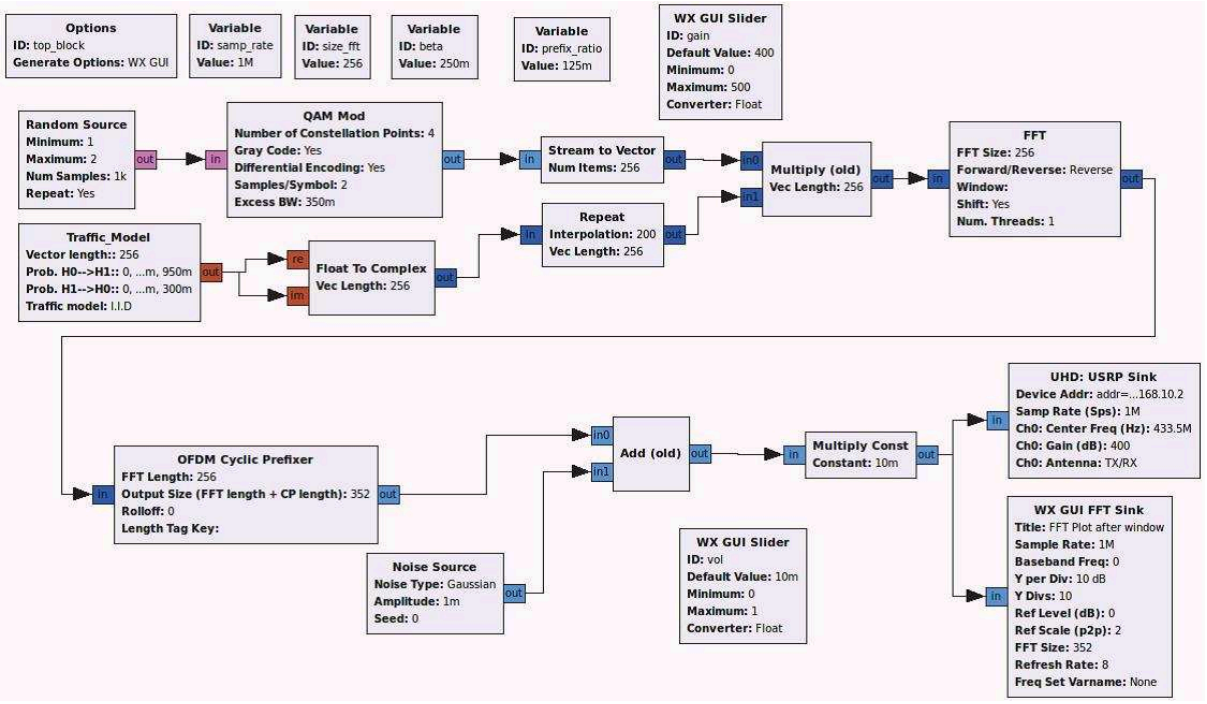


FIGURE 6.6: Detailed block diagram of the proposed primary user traffic generator implemented in GNU Radio Companion.

shown in Fig. 6.6. In the beginning, number of frequency bands, rewards (iid or Markovian) of MAB which models PU traffic and corresponding channel statistics are fixed using the block named Traffic Model in Fig. 6.6. The transmission bandwidth, which is restricted by bandwidth of analog front-end of USRP, is divided into $K = 8$ equal bandwidth frequency bands. In each time slot, masking vector of size K is generated by Traffic Model block based on given frequency band statistics. This masking vector can have 1 or 0 values where 1 and 0 indicate that corresponding band is occupied and vacant, respectively. Next step is mapping data to be transmitted on sub-carriers of occupied bands. The data modulation used is a QPSK modulation. This is followed by sub-carrier mapping using OFDM and transmission via USRP. In the proposed testbed, number of sub-carriers, center frequency and transmission bandwidth are 256, 433.5 MHz and 1 MHz, respectively. The 256 carriers are grouped by 32 in order to make finally 8 channels. For demonstration purpose, each time slot duration is one second so that it can be followed by human eye. However, it can be reduced to the order of milliseconds and will have no direct effect on the performance of learning policies.

Secondary User with Decision Making Policy

The chosen design environment for the SU terminal is MATLAB/Simulink and USRP from Ettus Research as shown in Fig. 6.7. USRP is tuned to receive signal of bandwidth 1 MHz centered at 433.5 MHz. The received signal is then down-sampled, digitized and passed to the learning policy implemented using Simulink. Then, MAB learning policy selects one frequency

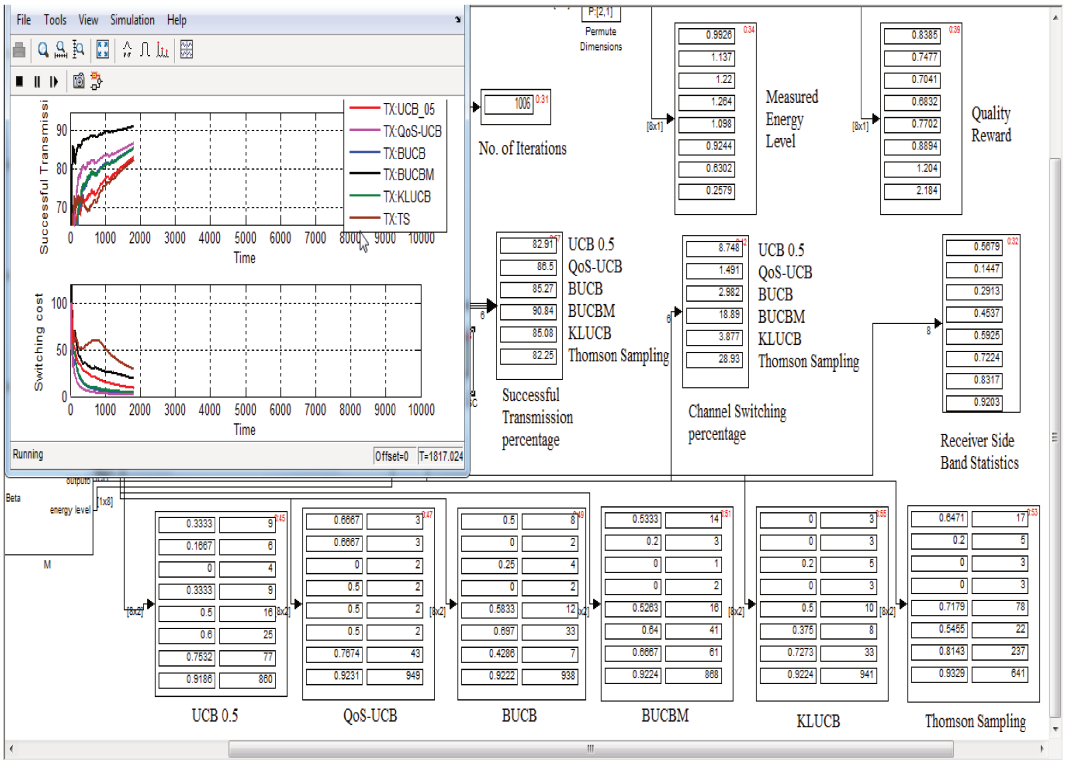


FIGURE 6.7: Block diagram of the Simulink based testbed consisting of SU with decision making policies.

band at each time slot. The chosen channel is sensed using an energy detector. Note that energy detector is not ideal and sensing errors may occur [69]. If the band is declared as vacant, it is assumed that SU transmits over the chosen band. If multiple SUs choose the same frequency band, it is assumed that all users suffer from collision, transmission fails and none receives a reward. In case of multi-player, each SU is independently implemented in Simulink with their respective learning policy. Thus, the proposed testbed with non-ideal energy detectors will enable to study performance of learning policies in the presence of sensing errors. However, performance comparisons of various detectors and their effect on MAB learning policies is not discussed here, and is the future steps of this work.

In the proposed testbed in Fig. 6.7, the receiver, which acts as an SU, consists of MAB learning algorithm for decision making process. We analyze and compare the performance of various MAB learning approaches, such as QoS-UCB ($\alpha = 0.25, \beta = 0.32$), UCB1 ($\alpha = 0.5$), Kullback-Leibler UCB (KL–UCB) [23], Bayesian UCB (BUCB) ([76]), two-stage Bayesian UCB (BUCBM) ([34]) and Thomson-Sampling ([136]) algorithms. In normal setting, SU can sense only one channel in each time slot while in two-stage BUCBM, SU can sense another channel if first channel is sensed as occupied. Intuitively, two-stage BUCBM will have the higher success in the transmission opportunities due to multi-stage sensing and access, however due to multi-stage learning, it will also have significantly higher CSC compared to classical setting. Note that our

proposed QoS-UCB is the only policy which takes into an account availability and quality for decision making.

Synchronization

For demonstration purposes, synchronization has been achieved by switching first channel from occupied to vacant states or vice-versa in each time slot. This enables SUs to detect the transitions between OFDM symbols as well as to synchronize the energy detection phase on an entire OFDM symbol of the primary traffic. In a real OSA scenario, SU should be able to synchronize with PU network via synchronization signals or pilot carriers. Note that the synchronization channel in the proposed approach is not wasted because learning policy does not consider it as synchronization channel and sees it as possible option for data transmission.

6.3.2 Experimental Results and Analysis

In this section, the performance of various MAB learning approaches in terms of number of transmission opportunities, number of collisions and channel switching cost (CSC) is compared on the proposed testbed. We consider $K = 8$ channels in the proposed testbed. Scenario 1 and 2 correspond to the case where the channel occupancy is iid and Bernoulli distributed with vacancy probability (P_{vac}) and scenario 3 and 4 model the channel occupancy as a Markovian process with the transition probabilities (P), as given in Table 6.1.

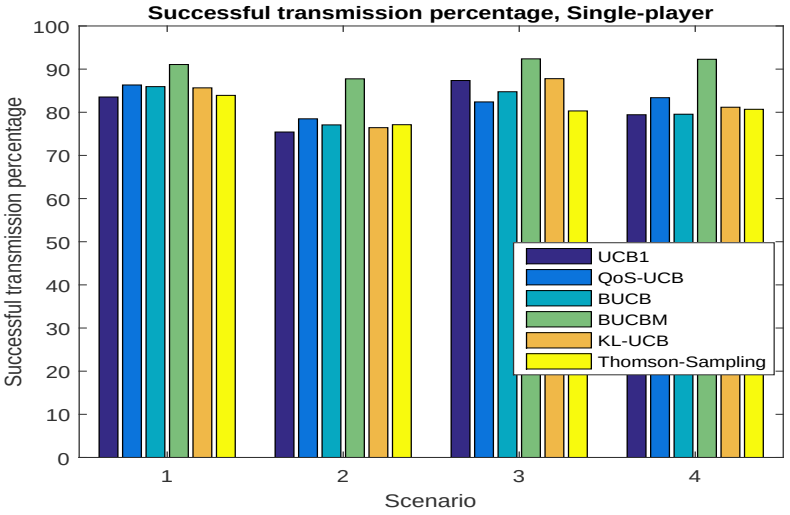
As discussed before in previous chapters, the measured spectrum level is recorded and used as a quality information metric for QoS-UCB policy. Indeed, instead of having only a free or occupied state information at the output of energy detector, we get a soft-metric representing the measured power level which will be used later to rate the band quality. Empirical average of quality information for 8 channels after 1000 time slots and 10 experiment is also presented in Table 6.1.

Each numerical result presented hereafter in this section is the average value achieved by iterating over 10 independent experiments on USRP testbed. Each experiment on USRP testbed consists of a time horizon of 1000 time slots for each SU and one time slot corresponds to one second, however time slot could be considered as 1ms for result analysis in terms of convergence speed. In multi-player case, we assume that all SUs employ the same MAB learning approach but do not exchange any information with other SUs.

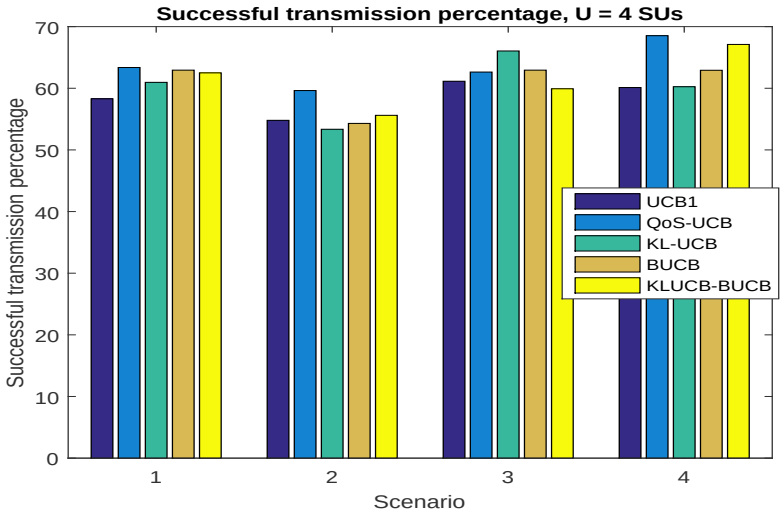
For Scenario 1 to 4, Figs. 6.8(a) and 6.8(b) show percentage of transmissions opportunities exploited by various MAB learning approaches in single-player and multi-player (number of SUs $U = 4$) settings, respectively. It can be observed from Figs. 6.8(a) and 6.8(b) that the proposed QoS-UCB policy exploits similar or higher percentage of transmission opportunities

Scenario	Channels	1	2	3	4	5	6	7	8
Scenario 1	P_{vac}	0.50	0.30	0.40	0.50	0.60	0.70	0.80	0.90
Scenario 2	P_{vac}	0.50	0.90	0.80	0.70	0.60	0.50	0.40	0.30
Scenario 3	P_{01}	0.05	0.03	0.04	0.05	0.06	0.07	0.08	0.09
	P_{10}	0.05	0.07	0.06	0.05	0.04	0.03	0.02	0.01
Scenario 4	P_{01}	0.50	0.74	0.78	0.83	0.85	0.86	0.88	0.90
	P_{10}	0.50	0.26	0.22	0.17	0.15	0.14	0.12	0.10
Empirical quality information	R_1	1.33	1.53	1.41	0.92	1.17	0.70	0.89	1.36

TABLE 6.1: Several scenarios to verify the proposed approach on testbed.



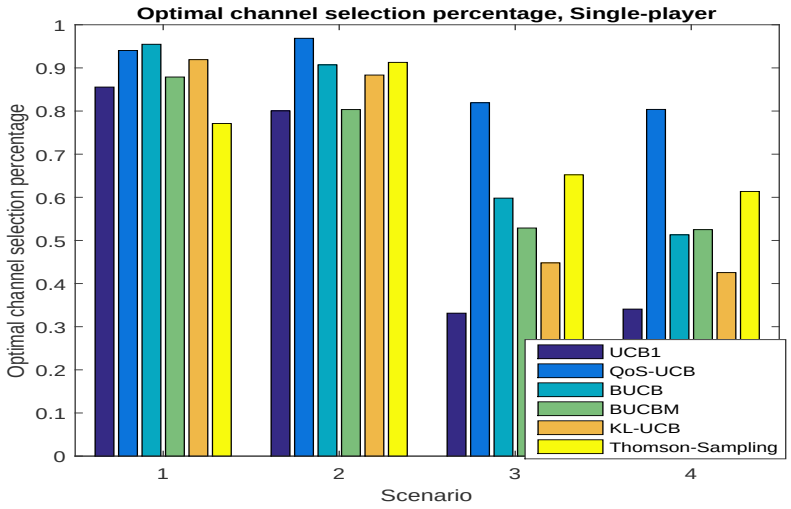
(a) Single-player



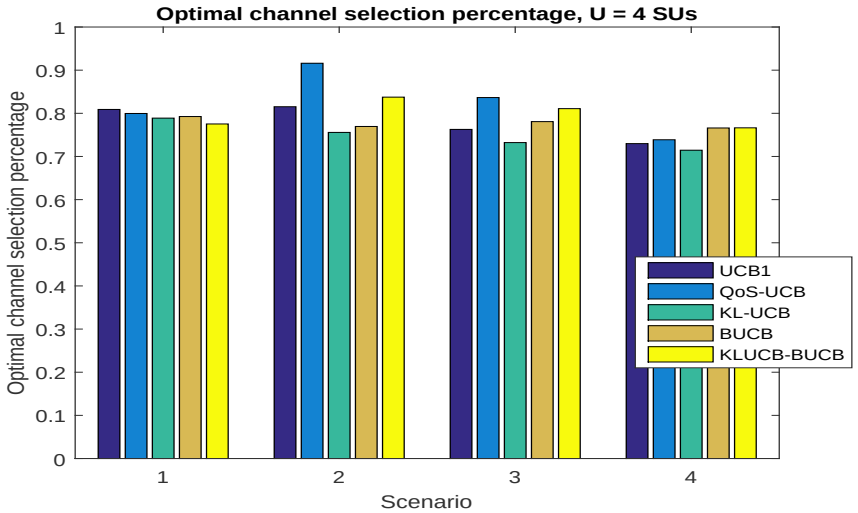
(b) Multi-player

FIGURE 6.8: Comparisons of successful transmission percentage for various P_{vac} distributions from Scenario 1 to 4 of different learning approaches for single-player and multi-player, respectively.

compared to the state-of-the-art decision making approaches for Scenario 1, 2 and 4. This is due to the fact that the proposed QoS-UCB approach learns on dual optimization criteria and hence is able to quickly find an optimal channel. In Scenario 3, we consider slowly varying PUs activity (channel state changes less frequently), and thus it is beneficial for algorithms like UCB1 which is more explorative compared to our proposed QoS-UCB and others. Whereas, less explorative approaches like QoS-UCB and Thomson-Sampling achieve comparatively less percentage of transmission opportunities in Scenario 3. We can also see, in Fig. 6.8(a), that multi-stage BUCB (BUCBM) achieves significantly higher percentage of transmission opportunities compared to our approach at a cost of very high channel switching percentage due to multi-stage decision making.



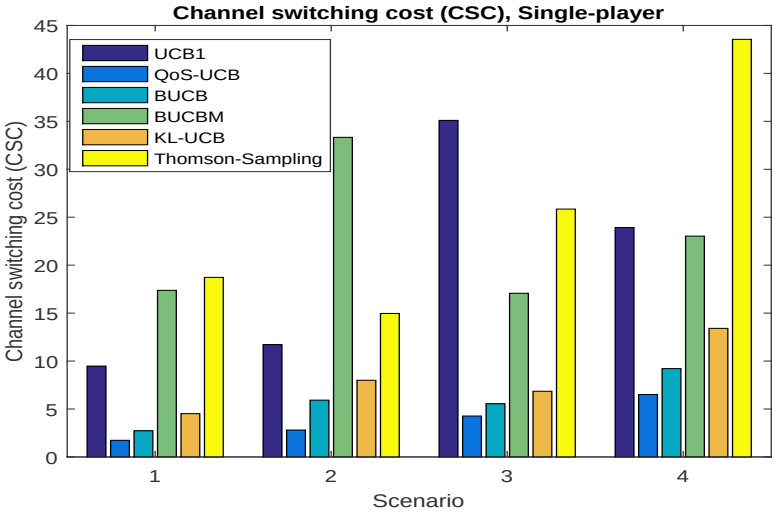
(a) Single-player



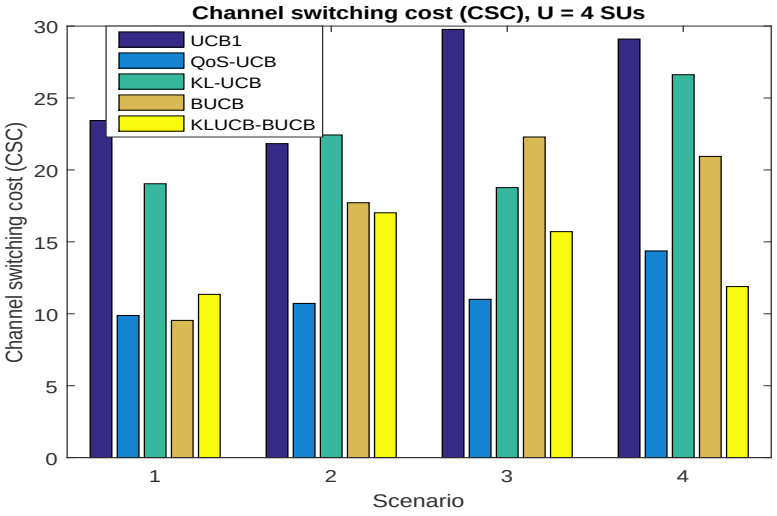
(b) Multi-player

FIGURE 6.9: Comparisons of optimal arm selection percentage of different learning approaches from Scenario 1 to 4 for single-player and multi-player, respectively.

Next, we analyze the optimal arm selection percentage, which is a classical performance measure in machine learning community, of various learning approaches in Figs. 6.9(a) and 6.9(b) for single-player and multi-player ($U = 4$) settings, respectively. We can see that even though state-of-the-art learning approaches achieve similar percentage of transmission opportunities than QoS-UCB policy, this later achieves similar or higher level of optimal arm selection percentage.



(a) Single-player



(b) Multi-player

FIGURE 6.10: Comparisons of channel switching cost of different learning approaches from Scenario 1 to 4 for single-player and multi-player, respectively.

As discussed above, the number of CSC should be as minimum as possible for making SU terminals energy efficient. In Figs. 6.10(a) and 6.10(b), number of CSCs of different MAB learning approaches are compared for several reward distributions in Scenarios 1, 2, 3 and 4 for single-player and multi-player settings, respectively. It is observed that the proposed QoS-UCB policy offers the lowest number of CSC except for $U = 4$ and Scenario 4. Numerically, the

average number of CSC of QoS-UCB policy is significantly lower than that of UCB1, KL-UCB, BUCB and Thomson-Sampling.

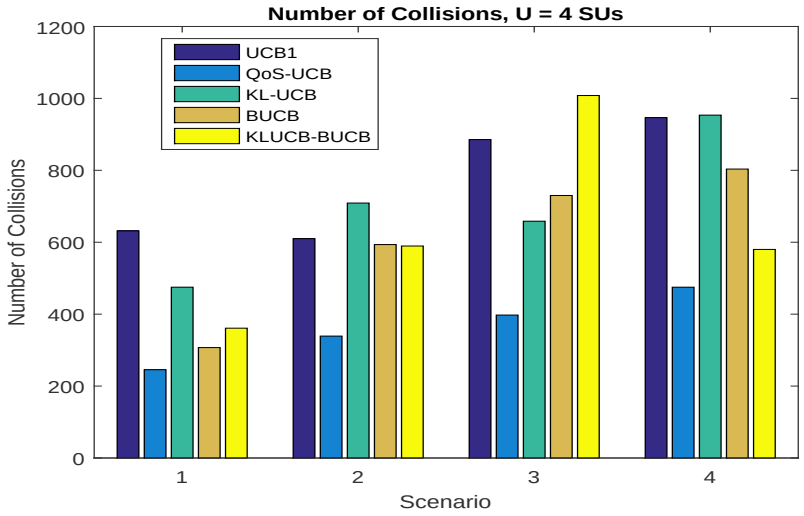


FIGURE 6.11: Comparisons of number of collisions of different learning approaches from Scenario 1 to 4 for multi-player setting.

In additions to CSC, the number of collisions should also be as minimum as possible. This is because, collisions lead to waste of the energy required for data preprocessing and transmission and it may be higher than the energy required for CSC. In Fig. 6.11, the number of collisions suffered by all SUs is compared for several reward distributions in Scenarios 1, 2, 3 and 4 for multi-player. Numerically, SUs with QoS-UCB policy has the lowest number of collisions than SUs with other MAB learning approaches. Thus, these features, low number of CSC and collisions, make QoS-UCB policy comparatively more energy efficient and suitable for battery operated SU terminals. From numerical analysis, we argue that QoS-UCB using dual criterion (e.g. availability and quality) optimization for OSA is not only superior in terms of spectrum utilization but also energy efficient.

6.4 Conclusion

In this work, the validation of the multi-armed bandit (MAB) learning algorithms has been performed on real radio spectrum and also on the HF measurements to dynamically access the spectrum. The obtained results for performance on real HF measurement demonstrate that UCB policy outperform a non-intelligent (i.e. random uniform selection) policy even in the worst conditions in terms of availability. It has been proved that successful transmission percentage achieved with UCB policy is in average three times higher than the successful transmission percentage achieved with a non-intelligent approach. This result shows that UCB is

suitable for the application of cognitive radio principles in the real environment, i.e. HF environment. This work is communicated in IEEE ICC 2016 [106] and IEEE Transactions on Cognitive Communications and Networking [105].

Next, we developed an USRP based testbed for experimentally analyzing the performance of our proposed QoS-UCB and several other learning policies, i.e. UCB1, KLUCB, BUCB and Thomson-Sampling, for OSA in the single-player and decentralized multi-player CR networks. The proposed testbed is the first proof-of-concept which compares the performance of various single-player and decentralized multi-player MAB learning policies using real radio signals. The presented proof-of-concept has received Best Demo (Booth) award in EAI CrownCom 2016 conference [34] and will also support next IEEE ICC 2017 conference tutorial "On-Line Learning for Real-Time Dynamic Spectrum Access: From Theory to Practice". Furthermore, experimental results have been presented to compare MAB learning policies in terms of average spectrum utilization, number of frequency band switchings as well as the number of collisions among SUs.

Chapter 7

General Conclusions and Perspectives

Contents

7.1 Conclusions and Overview 117
7.2 Perspectives and Future Work 119

7.1 Conclusions and Overview

This thesis investigated the fundamental problems arising in applying reinforcement learning to cognitive radio (CR), in general, and to opportunistic spectrum access (OSA) and energy efficiency in particular.

After a first general introduction of CR decision making for green radio in Chapter 1, we positioned the general reinforcement learning we are dealing with in Chapter 2. Moreover, this latter presented several state-of-the-art MAB variants, their detailed classification, and their application to the CR decision making problem. Assuming minimum a priori information about CR equipment's environment, we ventured state-of-the-art literature study on the sequential decision making techniques.

Three main topics were tackled in this thesis: i) decision making policy design for the opportunistic access to the spectrum, ii) decision making policy design for dynamic BSs switching operation to reduce the energy consumption of a network, and iii) proof of concept development of very simple and less complex MAB learning policies on the real radio environment:

- The Chapter 3 presented a new performance evaluation criterium, i.e. the optimal arm identification (OI) factor, to evaluate the *a priori* interest of learning and decision making

policy on a given scenario. Then, the efficiency of OI factor and LZ complexity on the state-of-the-art MAB learning policies applied to the OSA scenario was also verified in Chapter 3. As a result, the work in this chapter was published in EAI CrownCom international conference [113].

- In Chapter 4, a two criteria, i.e. availability and quality, driven algorithm, RQoS-UCB policy, was proposed to properly balance the exploration versus exploitation trade-off seen in restless Markov MAB model. Furthermore, the single-player decision making policy was also extended to the case of multi-player where several non-cooperative unlicensed users (or secondary users) evolve in the cognitive network. In OSA scenario, the proposed fully distributed approach for MAB framework has been demonstrated to largely improve QoS and reduce cooperation overhead. Finally, Chapter 4 presented a convergence study of the proposed single-player and multi-player restless Markov MAB policy which is denoted as a new contribution to the MAB literature and open the way to a large set of machine learning applications. Parts of this chapter were published in IEEE Globecom 2015 [110] and IEEE Transaction on Cognitive Communications and Networking [112]

- Chapter 5 introduced the concept of a flexible network architecture, which enables the base stations to switch between active and sleep modes. Dynamic BSs switching problem has been modeled as a Markov MAB framework, and then the performance of the proposed RQoS-UCB class of algorithms has been illustrated to answer the designed dynamic BSs switching problem. Next, transfer learning (TL) has been investigated to improve the proposed algorithm further by applying knowledge transferred from previous historical period. In heterogeneous cellular networks with a static topology of base stations, it has been demonstrated to largely improve the energy efficiency (EE) of the network. Furthermore in heterogeneous networks with periodic traffic fluctuation, transfer learning has been also shown to significantly increase the convergence speed, achieving a performance jump-start, which reduces QoS fluctuations during environmental changes. Moreover, we also presented analytical convergence analysis of the proposed transfer RQoS-UCB (TRQoS-UCB) policy which is also denoted as a new contribution to the MAB literature for large set of machine learning problems. Work in this chapter is submitted to IET communications journal [116].

- Finally, Chapter 6 presented a validation of the introduced MAB learning approaches on a real radio environment. First, we applied UCB1 on real measurements of HF band instead of artificially generated patterns. This work was published in IEEE ICC 2016 [106] and IEEE Transactions on Cognitive Communications and Networking [105]. Then, a proof-of-concept was developed for OSA in the decentralized network consisting of multiple secondary users. We implemented our proposed QoS-UCB and other MAB learning policies, i.e. UCB1, KLUCB, BUCB and Thomson-Sampling, on the proposed testbed consisting

of USRPs from Ettus Research, GNU radio companion for PUs and Matlab/Simulink for SUs. The experimental results have shown improvements of QoS-UCB in successful transmission percentage, number of SU collisions and the number of channel switchings on real radio testbed. The presented proof-of-concept received Best Demo (Booth) award in EAI CrownCom 2016 conference [34] and will also support next IEEE ICC 2017 conference tutorial "On-Line Learning for Real-Time Dynamic Spectrum Access: From Theory to Practice".

This work has attracted lots of interests, generated several number of publications and contributed to a number of research works on the relevant topics collaboratively with the IDeTIC, Universidad de Las Palmas de Gran Canaria, Spain [104–106] and the Department of Electronics and Communication Engineering, IIIT-Delhi, India [34].

7.2 Perspectives and Future Work

Dynamic Spectrum Access (DSA) and self-organized heterogeneous cellular network will play an important role for ultra-high capacity network in the future 5G communication systems. The MAB learning algorithm proposed in this thesis has been demonstrated as an effective approach to improve the system reliability, QoS/capacity and energy efficiency. The MAB framework based learning algorithms we developed in this work have the potential to be improved further.

- Chapter 3 shows that it is possible to identify, with help of OI factor, the worst scenarios *a priori* and then to predict the performance of MAB learning approach in advance. The results assume that the statistics of the Markov chain, which models each arm's reward process, is known and hence it limits the applicability of evaluations methodology. This opens the way to improve and develop, perhaps even theoretically guaranteed, methods to predict a priori performance of MAB approaches.

- The relationship between the evaluation technique, i.e. Lempel-Ziv complexity, suggested by Macaluso et al. in [100] and our OI factor is not quite clear yet and it is currently under investigation. The aim is to develop an unified performance evaluation methodology for the general reinforcement learning approach.

- The learning policy, RQoS-UCB, proposed in Chapter 4 can be used to learn on many other criteria such as the energy efficiency or actual SINR on the secondary link and will be investigated in future works. Moreover, our model ignores dynamic traffic at the secondary nodes and extension to a queueing-theoretic formulation is left for future work.

- The proposed distributed RQoS-UCB policy in Chapter 4 considers a fixed number of secondary users operating in a network. As a future work, we plan to develop novel

techniques that combine learning with game theory and mechanism design to achieve various performance objectives when the agents are strategic and/or when the set of players are comparatively large.

- Chapter 5 presented a transfer RQoS-UCB algorithm, a reinforcement learning algorithm with temporal transfer learning framework, to solve the problem of BS switching operation for energy efficiency (EE) maximization in heterogeneous wireless cellular network. Future work may handle a transfer learning over spatial dimension.

- Furthermore, future communication systems tend to be hyper-dense networks with massive small cell base stations serving different types of traffic. Dynamic fully distributed topology management, compared to centralized architecture in Chapter 5, would be better to effectively manage such a complex architecture, in order to provide higher system capacity, reduce power consumption and further reduce cooperation overhead.

- Proof-of-concept system developed in Chapter 6 considers energy detector as a spectrum sensing technique. Future work may also study the effect of various spectrum sensing detectors on the performance of learning and decision making policies and realization of actual data transmission on the dynamically chosen frequency band.

Appendix A

List of Publications

Contents

A.1 Patent 121
A.2 Journal Papers 121
A.3 International Conference Papers 122
A.4 National Conference Papers 123
A.5 Demonstration 123

A.1 Patent

1. **Procédé d'accès opportuniste au spectre**
   Navikkumar Modi, Christophe Moy, Philippe Mary.
   French Patent No. 1556916, 2014

A.2 Journal Papers

1. **TRQoS-UCB: A Transfer Restless QoS-UCB policy for Energy Efficient Heterogeneous Cellular Network (to be submitted)**
   Navikkumar Modi, Philippe Mary, Christophe Moy.
   —

2. **QoS driven Channel Selection Algorithm for Cognitive Radio Network: Multi-User Multi-armed Bandit Approach**
   Navikkumar Modi, Philippe Mary, Christophe Moy.
   IEEE Transactions on Cognitive Communications and Networking, 2017

3. **Efficient Learning in Stationary and Non-stationary OSA Scenario with QoS Guaranty**
   Navikkumar Modi, Philippe Mary, Christophe Moy.
   EAI Endorsed Transaction on Wireless Spectrum (TWS), 2017

4. **Hybrid UCB-HMM: A Machine Learning Strategy for Cognitive Radio in HF Band**
   Laura Melián-Gutiérrez, **Navikkumar Modi**, Christophe Moy, Ivan Pérez-Álvarez, Faouzi Bader, Santigo Zazo.
   IEEE Transactions on Cognitive Communications and Networking, 2016

A.3 International Conference Papers

1. **Proof-of-Concept: Spectrum and Energy Efficient Multi-User CR Network via Vacancy and Quality based Channel Selection**
   Navikkumar Modi, Philippe Mary, Christophe Moy, Sumit Darak and Jacques Palicot.
   Accepted in 32nd URSI GASS, Montreal, 19–26 August 2017

2. **A New Evaluation Criteria for Learning Capability in OSA Context**
   Navikkumar Modi, Christophe Moy, Philippe Mary, Jacques Palicot.
   11th EAI International Conference on Cognitive Radio Oriented Wireless Networks 2016 (CROWNCOM 2016), Grenoble, France, May 2016

3. **QoS driven Channel Selection Algorithm for Opportunistic Spectrum Access**
   Navikkumar Modi, Philippe Mary, Christophe Moy.
   IEEE Global Communications (GlobeCom) Workshops (IEEE GlobeCom 2015), San Diego, USA, December 2015

4. **Upper Confidence Bound Learning Approach for real HF Measurements**
   Laura Melián-Gutiérrez, **Navikkumar Modi**, Christophe Moy, Ivan Pérez-Álvarez, Faouzi Bader, Santigo Zazo.
   IEEE International Conference on Communications Workshops (IEEE ICC 2015), London, UK, June, 2015

5. **DSA with Reinforcement Learning in the HF Band (Young Scientist Paper Award)**
   Laura Melián-Gutiérrez, **Navikkumar Modi**, Christophe Moy, Ivan Pérez-Álvarez, Faouzi Bader, Santigo Zazo.
   1st URSI Atlantic Radio Science Conference 2015 (URSI AT-RASC 2015), Gran Canaria, Spain, May, 2015

6. **Experimental Performance Comparison and Analysis for Various MAB Problems under Cognitive Radio Framework**
   Navikkumar Modi, Christophe Moy, Philippe Mary.
   Proceedings of the Wireless Innovation Forum European Conference (WinnComm-Europe), 2014

A.4 National Conference Papers

1. **Machine Learning for Opportunistic Spectrum Access with Energy Consumption Constraint**
   Navikkumar Modi, Christophe Moy, Philippe Mary.
   URSI-France Workshop on Energy and Radio Science, Rennes, France, March 2016

2. **Apprentissage machine orienté QoS pour l'accès opportuniste au spectre**
   Navikkumar Modi, Philippe Mary, Christophe Moy.
   XXVe Colloque GRETSI 2015

A.5 Demonstration

1. **Spectrum Utilization and Reconfiguration Cost Comparison of Various Decision Making Policies for Opportunistic Spectrum Access Using Real Radio Signals (Best Booth (Demo) Award)**
   Sumit Darak, **Navikkumar Modi**, Amor Nafkha, Christophe Moy.
   11th EAI International Conference on Cognitive Radio Oriented Wireless Networks 2016 (CROWNCOM 2016), Grenoble, France, May 2016

Appendix B

Efficient Learning in Non-stationary Scenario

Contents

B.1 Introduction 125

B.2 Non-stationary Markov MAB Problem Formulation 126

B.2.1 Non-stationary Regret 127

B.2.2 Discounted QoS-UCB (DQoS-UCB) Policy 128

B.3 Simulation Results 129

B.3.1 Non-stationary Environment 130

B.4 Conclusion 132

B.1 Introduction

Though the stationary formulation of Markov MAB problem, as in Chapter 4, permits to address exploration versus exploitation dilemma more appropriately, but it may fail to justify a changing environment model where the observed reward distribution undergoes changes in time. As an example, probability of vacancy of each channel is likely to experience changes in time in OSA scenario, which exhibits the limitation of stationary MAB models. In many application domains, abrupt changes in the reward distribution are an intrinsic characteristic of the problem. Standard soft-max and upper confidence bound (UCB) class of policies are not well adapted for abruptly changing environments as it has been stated in [51].

Authors, in [16, 18], introduced soft-max action selection policies, i.e. EXP3, EXP3.S, for non-stationary iid MAB problem, where distribution of reward undergoes abrupt changes in time. Moreover, several policies such as discounted-UCB (D-UCB), sliding-window UCB (SW-UCB),

belonging to the wider family of UCB, are designed to address abruptly changing non-stationary iid MAB problem [51, 87, 137]. These policies are also consistent with more extreme settings, such as the one presented in [138] where reward distribution follows a Brownian motion. Even though theoretically Brownian motion can be seen as a particular Markov process, there is no straightforward links with non-stationary Markov MAB problem. Thus, learning policy for MAB with Markovian reward distribution requires a special attention because of increase in interest to model a cognitive radio learning with Markov MAB framework.

The proposed policy, referred as discounted QoS-UCB (DQoS-UCB), is able to learn on the same characteristics than previous QoS-UCB, i.e. quality and availability, but in non-stationary environments, i.e. when reward distribution evolves with time, allowing to increase the overall network performance. The latter proposed policy is referred as DQoS-UCB because of *discount factor* which weight more recent observations compared to observations acquired in the past.

The rest of this appendix is organized as follows. In Section II; we formulate the non-stationary Markov MAB problem and propose a DQoS-UCB learning policy for it. The numerical results are presented in Section III, verifying the validity and efficiency of the proposed DQoS-UCB policy in non-stationary environment. Finally, Section IV concludes the appendix.

B.2 Non-stationary Markov MAB Problem Formulation

Non-stationary environments are considered in this appendix as shown in Fig. 2.1 where the reward distributions remain constant during the period $a_l, l = 1, \cdots, N_{\text{change}}$ and change at unknown time[1]. Here, $N_{\text{change}} - 1$ is the total number of times the reward distribution of a channel changes up to time n. The position in time of a transition of the reward distribution is unknown for user *a priori* and is assumed to be independent from reward distribution. The $i-$th arm is modeled by an irreducible and aperiodic discrete time Markov chain with finite state space S^i. $P^{i,a_l} = \left\{p_{kl}^{i,a_l}, (k, l \in \{q_0, q_1\}), (q_0, q_1 \in S^i)\right\}$ denotes the state transition probability matrix of the $i-$th arm in period a_l, where q_0 and q_1 are the states, occupied and free respectively. We assume that each arm is mutually independent from others in each period. Let, $\boldsymbol{\pi}_{a_l}^i$ be the stationary distribution of the Markov chain in period a_l defined as $\pi_q^{i,a_l}(n) = \pi_q^{i,a_l}\ \forall n$:

$$\boldsymbol{\pi}^{i,a_l} = [\pi_{q_0}^{i,a_l}, \pi_{q_1}^{i,a_l}] = \left[\frac{p_{q_1q_0}^{i,a_l}}{p_{q_1q_0}^{i,a_l} + p_{q_0q_1}^{i,a_l}}, \frac{p_{q_0q_1}^{i,a_l}}{p_{q_1q_0}^{i,a_l} + p_{q_0q_1}^{i,a_l}}\right]. \tag{B.1}$$

Likewise stationary problem, taking into account the reward associated with quality information and the observed state of each channel i in period a_l, the mean reward is defined as $\mu_{i,a_l}^R =$

[1] However, several other variants, e.g. continuously changing, fixed instance change, etc., of the non-stationary problems are discussed in literature [18, 138].

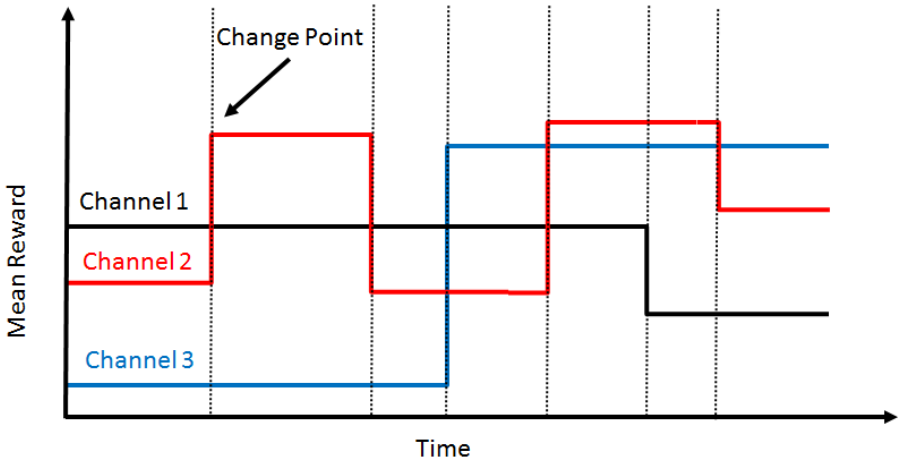


FIGURE 2.1: Abruptly changing non-stationary scenario

$\sum_{q \in S^i} G_q^{i,a_l} r_q^{i,a_l} \pi_q^{i,a_l}$. Let, μ_{1,a_l}^R being the channel with the highest mean reward in period a_l and is referred as an optimal channel for that period.

B.2.1 Non-stationary Regret

The regret of a policy in the non-stationary Markov MAB environment is defined as the difference between the total rewards collected by the optimal policy $\mathcal{A}^*$ (which plays an optimal arm at each time instant) and the total rewards collected by the selected policy. However, it is important to consider that the non-stationary regret is not estimated with respect to the optimal arm on average, but with respect to an optimal policy $\mathcal{A}^*$ following the optimal arm at each time instance. The non-stationary regret $\Phi^{ns}(n)$ defined in this section is similar to the regret for non-stochastic bandit problem presented in [16].

$$\Phi^{ns}(n) = \mathbb{E}^{\mathcal{A}^*}\left[\sum_{t=1}^{n} R_{q_{\mathcal{A}^*(t)}}^{\mathcal{A}^*(t)}(t) r_{q_{\mathcal{A}^*(t)}}^{\mathcal{A}^*(t)}(t)\right] - \mathbb{E}^{\mathcal{A}}\left[\sum_{t=1}^{n} R_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}(t) r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}(t)\right] \tag{B.2}$$

The definition of the regret itself is not straightforward. Indeed, since the underlying process is non-stationary, the proposition depends on time. Let, $\mathcal{A}^*(t)$ and $\mathcal{A}(t)$ being the channel index selected by the optimal policy $\mathcal{A}^*$ and given policy $\mathcal{A}$ at time t, respectively. $q_{\mathcal{A}^*(t)}$ and $q_{\mathcal{A}(t)}$ being the state of the channel $\mathcal{A}^*(t)$ and $\mathcal{A}(t)$ at time t, respectively. Moreover and without any loss of generality, we assume that $\mu_{1,a_l}^R > \mu_{i,a_l}^R > \mu_{K,a_l}^R$, for $i = 2, \cdots, K-1$ and $l = 1, \cdots, N_{\text{change}}$.

B.2.2 Discounted QoS-UCB (DQoS-UCB) Policy

In this section we introduce discounted QoS-UCB (DQoS-UCB) algorithm for the Markov MAB problem. The motivation for the DQoS-UCB policy is to find an optimal channel with less exploration in the case of *abruptly changing environments*. As stated before, the standard QoS-UCB policy is not appropriate for the non-stationary environment, because confidence interval of standard QoS-UCB policy becomes tighter when time goes up. To guaranty the performance of DQoS-UCB policy in non-stationary environment, a discount factor λ is considered for the DQoS-UCB index estimation. The idea behind inclusion of discount factor is to give more weight to recent observations compared to the ones acquired in the past.

The proposed DQoS-UCB policy learns a channel which is optimal in terms of probability of vacancy and quality in each period a_l. Our contribution for non-stationary Markov MAB is stated in Algorithm 7. As with QoS-UCB policy, an SU employing DQoS-UCB policy first

Algorithm 7 DQoS-UCB policy

Input: α, β, $\mathcal{A}(0)$, $T^i(0)=0$, $R^i_{q_1}(0)$ $\forall i \in \mathcal{K}$ and $\lambda < 1$
Output: $a(n+1)$

for $n = 1$ **to** K **do**
  Initialize policy by sensing each channel for at least one time.
end for
while $n > K$ **do**
  $B^i(n, T^i(n)) = \bar{S}^i(T^i(n)) - Q^i(n, T^i(n)) + A^i(n, T^i(n)), \forall i$
  $\mathcal{A}(n) = \arg\max_i B^i(n, T^i(n))$
  Sense channel $i = \mathcal{A}(n)$ and observe current state $S^i(n)$
  if channel $i = \mathcal{A}(n)$ free **then**
    Transmit and observe quality $R^i_{q_1}(T^i(n))$
  else
    Do Not Transmit and $R^i_{q_1}(T^i(n)) = R^i_{q_1}(T^i(n-1))$
  end if
  $T^i(n) = \sum_{t=1}^{n} \mathbf{1}_{\mathcal{A}(t)=i}, \forall i$
  $N^i(\lambda, n) = \sum_{t=0}^{n} \lambda^{n-t} \mathbf{1}_{\mathcal{A}(t)=i}, \forall i$
  $W(\lambda, n) = \sum_{i=1}^{K} N^i(\lambda, n)$
  $\bar{S}^i(T^i(n)) = \frac{1}{N^i(\lambda,n)} \sum_{t=0}^{n} \lambda^{n-t} S^i(t) \mathbf{1}_{\mathcal{A}(t)=i}, \forall i$
  $G^i_{q_1}(T^i(n)) = \frac{1}{N^i(\lambda,n)} \sum_{t=0}^{n} \lambda^{n-t} R^i_{q_1}(t) \mathbf{1}_{\mathcal{A}(t)=i}, \forall i$
  Find $G^{q_1}_{\max}(n) = \max_{i \in \mathcal{K}} G^i_{q_1}(T^i(n))$
  $M^i(n, T^i(n)) = G^{q_1}_{\max}(n) - G^i_{q_1}(T^i(n)), \forall i$
  $Q^i(n, T^i(n)) = \frac{\beta M^i(n, T^i(n)) \ln(W(\lambda,n))}{N^i(\lambda,n)}, \forall i$
  $A^i(n, T^i(n)) = \sqrt{\frac{\alpha \log(W(\lambda,n))}{N^i(\lambda,n)}}, \forall i$
  $n = n + 1$
end while

Channel	$0 < n < 1500$		$1500 \leq n \leq 3000$		$3000 < n < 5000$	
(i)	$P^i_{q_0q_1}$	$P^i_{q_1q_0}$	$P^i_{q_0q_1}$	$P^i_{q_1q_0}$	$P^i_{q_0q_1}$	$P^i_{q_1q_0}$
1	0.40	0.30	0.40	0.30	0.40	0.30
2	0.20	0.60	0.20	0.60	0.20	0.60
3	0.50	0.40	0.50	0.40	0.50	0.40
4	0.50	0.70	0.50	0.70	0.50	0.70
5	0.65	0.80	0.65	0.80	0.65	0.80
6	0.60	0.70	0.60	0.70	0.90	0.10
7	0.40	0.50	0.20	0.50	0.20	0.50
8	0.64	0.80	0.54	0.80	0.54	0.80
9	0.30	0.30	0.90	0.30	0.20	0.30
10	0.90	0.20	0.20	0.20	0.20	0.20

TABLE B.1: State transition probabilities P^i in non-stationary environment

starts to sense all channels at least once initially, and after $n > K$ iterations, it updates the index $B^i(n, T^i(n))$ as in (B.3). However, each term of the index equation is adapted to take into account the non-stationary hypothesis.

$$B^i(n, T^i(n)) \quad = \quad \bar{S}^i(T^i(n)) - Q^i(n, T^i(n)) + A^i(n, T^i(n)), \forall i \tag{B.3}$$

$T^i(n)$ is still the number of times channel i has been sensed by DQoS-UCB policy up to time n. In addition, $N^i(\lambda, n) = \sum_{t=1}^{n} \lambda^{n-t} \mathbf{1}_{\mathcal{A}(t)=i}$ is the discounted number of times channel i has been sensed up to time n, and $W(\lambda, n) = \sum_{i=1}^{K} N^i(\lambda, n)$ is the total discounted time.

Contrary to QoS-UCB policy, empirical mean of observed states $\bar{S}^i(T^i(n))$ and quality $G^i_{q_1}(T^i(n))$ are estimated by taking into account a discount factor $\lambda < 1$ as shown in steps 16 and 17 of Algorithm 7. The coefficients α and β in steps 21 and 20 of Algorithm 7, are the same than in QoS-UCB policy, to weight exploration for vacancy and channel quality, respectively.

B.3 Simulation Results

In this section, DQoS-UCB policy under non-stationary environment is investigated in simulations. We present simulation results focusing on regret, quality information and percentage of optimal channel selection. K is set to 10 channels with two states each, i.e. q_0 (occupied) and q_1 (free) respectively. Simulation results are given by averaging over 100 runs performed using MATLAB.

Channel	$0 < n < 5000$	$0 < n < 1500$	$1500 \leq n \leq 3000$	$3000 < n < 5000$
(i)	$G^i_{q_1}$	μ^R_i	μ^R_i	μ^R_i
1	0.55	0.35	0.35	0.35
2	0.67	0.24	0.24	0.24
3	0.65	0.40	0.40	0.40
4	0.80	0.38	0.38	0.38
5	0.60	0.31	0.31	0.31
6	0.90	0.46	0.46	0.82
7	0.75	0.38	0.28	0.28
8	0.78	0.39	0.36	0.36
9	0.85	0.47	0.66	0.40
10	0.92	0.76	0.51	0.51

TABLE B.2: Observed channel reward and estimated mean reward in non-stationary environment

B.3.1 Non-stationary Environment

The scope of this section is to introduce a setting that allows to analyze the behavior of MAB learning policies in non-stationary abruptly changing environment. Two breakpoints at time $n = 1500$ and $n = 3000$ are introduced in the simulation setting which indicate an abrupt change in the reward distribution. An optimal policy is able to identify an abrupt change in the reward distribution with reduced delay and maximize the opportunistic transmission in non-stationary environment. We compare the performance of UCB1 ($L = 0.5$), QoS-UCB ($\alpha = 0.25$ and $\beta = 0.32$) and DQoS-UCB ($\alpha = 0.25$ and $\beta = 0.32$) policies. Moreover, we introduce a discount factor $\lambda = 0.98$ in the DQoS-UCB policy to cope up with non-stationary scenario.

Table B.1 introduces the transition probability matrices P^i for different time periods where distribution of each channel may change. The mean reward μ^R_i, as detailed in Table B.2, are computed using P^i and G^i_q as detailed in Section B.2. Transition probabilities do not change from channel 1 to 5, but change over time from channel 6 to 10. The reward associated with quality G^i_q remains stationary during the simulation, however, estimated mean reward μ^R_i changes abruptly due to the channel availability variations.

Fig. 2.2 shows the evolution of regret achieved by UCB1, QoS-UCB and DQoS-UCB policy for scenario depicted in Tables B.1 and B.2. As seen in Fig. 2.2, DQoS-UCB achieves significantly lower regret than QoS-UCB and UCB1 policy in non-stationary environment. DQoS-UCB wastes significantly less time than QoS-UCB to identify an abrupt change ($n = 1500$ and $n = 3000$) in the reward distribution, due to inclusion of discount factor λ in the index calculation. On the other hand, past has an higher influence on UCB1 and QoS-UCB preventing them to adapt quickly to changes.

Moreover for further analysis, we compare the average reward associated with quality information observed with DQoS-UCB and QoS-UCB policies. As shown in Fig. 2.3, both algorithms converge to the optimal mean reward in the long run, but DQoS-UCB policy benefits from a less dependency on the past observations. Whereas, QoS-UCB policy takes more time to converge to the optimal mean reward after an abrupt change in the reward distribution, because it considers all past observations to make next decision. Fig. 2.3 states that DQoS-UCB policy is able to track an abrupt change in the reward distribution and achieves higher reward in the long run.

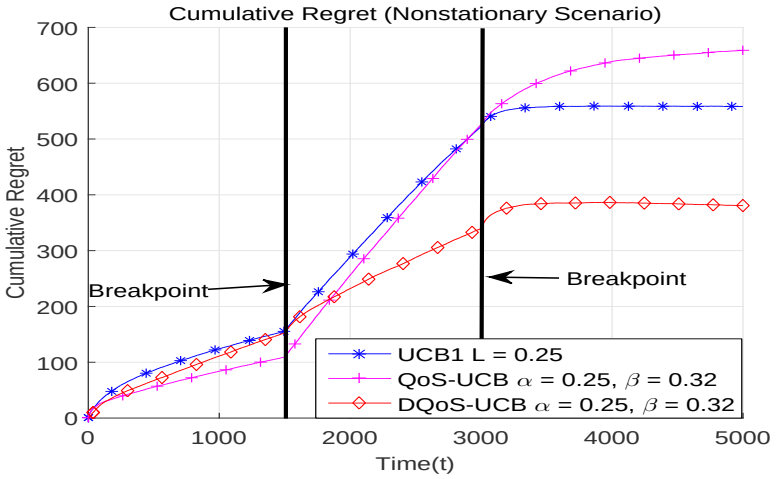


FIGURE 2.2: UCB1, QoS-UCB and DQoS-UCB: Cumulative regret

Fig. 2.4 presents the optimal channel selection percentage for DQoS-UCB and QoS-UCB policies. It is clear from Fig. 2.4 that DQoS-UCB policy finds an optimal channel quickly, and concentrates on it in case of an abrupt change in the reward distribution. Whereas, QoS-UCB takes significantly more time to converge to an optimal channel after an abrupt change, and more specifically when reward distribution changes after sufficient time step. Another important criterion is the number of successful channel access, correlated to the successful opportunistic transmission at the end. We define the successful transmission percentage (STP) as the number of time free slots are detected from the total number of time steps n.

The DQoS-UCB policy achieves higher STP compared to QoS-UCB policy in the long run. However, QoS-UCB policy is able to find an optimal channel in non-stationary environment, but it requires significantly more time to converge to it after an abrupt change in the reward distribution. Thus, DQoS-UCB policy is more appropriate for non-stationary environment compared to QoS-UCB policy.

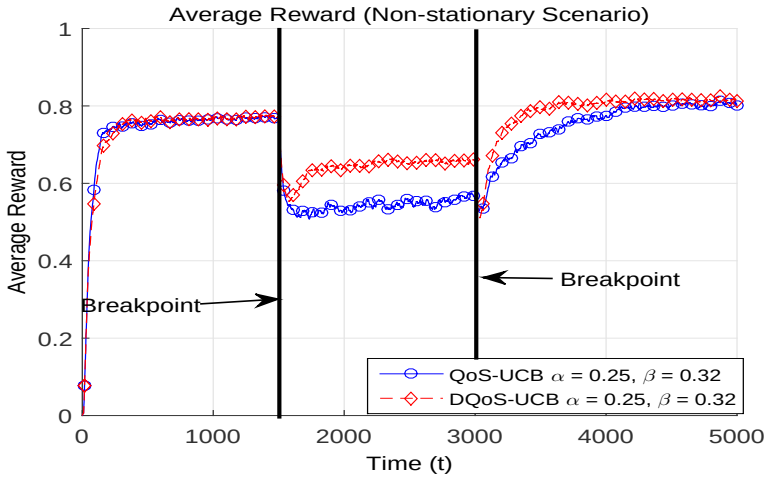


FIGURE 2.3: QoS-UCB and DQoS-UCB: Average reward

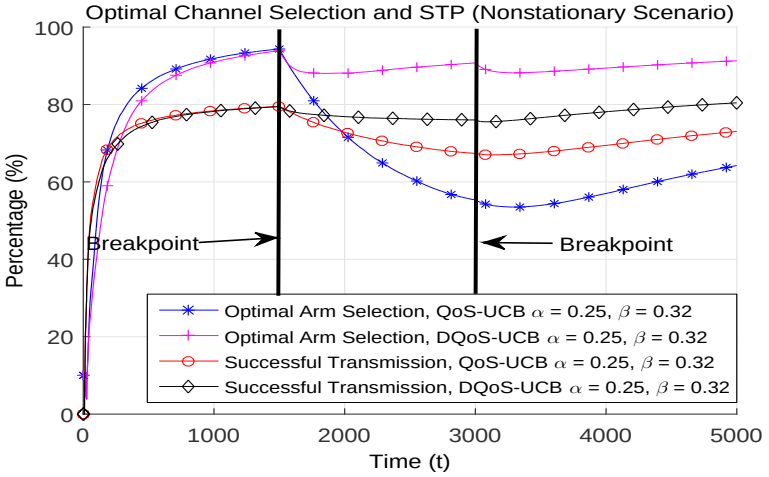


FIGURE 2.4: QoS-UCB and DQoS-UCB: Optimal channel selection percentage and successful transmission percentage (STP) comparison

B.4 Conclusion

This appendix has briefly described the Markov MAB problem taking into account non-stationary reward distribution associated with each arm. An efficient OSA learning algorithm named DQoS-UCB based on channel quality information and availability has been proposed for the non-stationary Markov MAB problem.

This chapter extends original QoS-UCB policy to non-stationary environment by including a discount factor to weaken the confidence interval of QoS-UCB policy which normally becomes tighter when time goes up. Numerical analysis has verified that, in case of non-stationary Markov MAB, both DQoS-UCB and QoS-UCB policies are able to find an optimal channel, but QoS-UCB policy requires longer time to converge to an optimal channel. To design a fully adaptive algorithm which is able to adjust discount factor according to the environment is not

an easy task and remains the subject of future research. Moreover, the theoretical guarantee for convergence in non-stationary environment is challenging and left for future work.

Appendix C

Proof of Theorems and Lemmas

C.1 Proof of Theorem 4.1

In order to bound the regret, we need to bound the expected number of blocks, $\mathbb{E}[F^i(b)]$, for any suboptimal band $i > 1$. Let l being a positive integer and $\mathcal{A}(b)$ the action performed by policy $\mathcal{A}$ in block b. Let us remind that $\mu_i^R = \sum_{q \in S^i} G_q^i r_q^i \pi_q^i$. $n_2(b)$ represents total time spent in SB2 block up to block b. Following the steps as in [145], the number of blocks a band i has been visited up to block b can be expressed as

$$F^i(b) = 1 + \sum_{m=K+1}^{b} \mathbb{1}_{(\mathcal{A}(m)=i)} \tag{C.1}$$

$$F^i(b) = l + \sum_{m=K+1}^{b} \mathbb{1}_{(\mathcal{A}(m)=i, F^i(m-1)\geq l)} \tag{C.2}$$

$$= l + \sum_{m=K+1}^{b} \mathbb{1}_{\left(B^1\left(T_2^1(n_2(m-1)),n_2(m-1)\right)\leq B^i\left(T_2^1(n_2(m-1)),n_2(m-1)\right),F^i(m-1)\geq l\right)} \tag{C.3}$$

$$\leq l + \sum_{m=K+1}^{b} \mathbb{1}_{\left(\exists \omega^i : l \leq \omega^i \leq n_2(m-1), B^i(\omega^i, n_2(m)) > \mu_1^R\right)}$$

$$+ \mathbb{1}_{\left(\exists \omega^1 : 1 \leq \omega^1 \leq n_2(m-1), B^1(\omega^1, n_2(m)) \leq \mu_1^R\right)} \tag{C.4}$$

where (C.2) comes from the fact that each band has been sensed at least l blocks up to block b. (C.3) comes from the reason why suboptimal band i is chosen up to $n_2(m-1)$ time at the end of block m, i.e. the index of an optimal band at block $m-1$, i.e. $B^1\left(T_2^1(n_2(m-1)), n_2(m-1)\right)$, is below the index of the suboptimal band i. Moreover (C.3) is upper bounded by (C.4) because these two conditions are not exclusive. Taking the expectation on both sides and using union

bound we get:

$$\begin{aligned}\mathbb{E}[F^i(b)] &\leq l + \sum_{m=K+1}^{b} \sum_{\omega^i=n_2(l)}^{n_2(m-1)} \mathbb{P}(B^i(\omega^i, n_2(m)) > \mu_1^R) \\ &+ \sum_{m=K+1}^{b} \sum_{\omega^1=1}^{n_2(m-1)} \mathbb{P}(B^1(\omega^1, n_2(m)) \leq \mu_1^R) \\ &\leq l + \sum_{t=1}^{n_2(b)} \sum_{\omega^i=l}^{t-1} \mathbb{P}(B^i(\omega^i, t) > \mu_1^R) + \sum_{t=1}^{n_2(b)} \sum_{\omega^1=1}^{t-1} \mathbb{P}(B^1(\omega^1, t) \leq \mu_1^R) \end{aligned} \tag{C.5}$$

The summation over t starts from 1 instead of $K+1$ because it does not change the validity of the upper bound. Note that channel 1 is optimal in terms of mean reward, μ_1^R, i.e. both in vacancy and quality. G^1 is hence the empirical mean of the quality reward of this channel, it does not mean necessarily that $G^1 = G_{\max}^{q_1}$. Moreover, let's remind that $\Delta\mu_i^R = \mu_1^R - \mu_i^R$. let's choose $l = \left\lceil \frac{4\alpha \ln n}{(\Delta\mu_i^R)^2} \right\rceil$, take expectation on both sides and relaxing the outer sum in (C.5) from $n_2(b)$ to ∞, and proceed from (C.5):

$$\mathbb{E}[F^i(b)] \leq l + \sum_{t=1}^{\infty} \sum_{\omega^i=l}^{t-1} \mathbb{P}(B^i(\omega^i, t) > \mu_1^R) + \sum_{t=1}^{\infty} \sum_{\omega^1=1}^{t-1} \mathbb{P}(B^1(\omega^1, t) \leq \mu_1^R) \tag{C.6}$$

In this proof, we upper bound the (C.6) in order to upper bound the expected number of blocks in suboptimal arms.

Let's start with the first part of (C.6), i.e. $\mathbb{P}(B^i(\omega^i, t) > \mu_1^R)$. By writing $\mu_1^R = \mu_i^R + \Delta\mu_i^R$ and replacing the $B^i(\omega^i, t)$ by its expression, we get

$$\mathbb{P}(B^i(\omega^i, t) > \mu_1^R) = \mathbb{P}\left(\bar{S}^i(\omega^i) - \frac{\beta M^i(\omega^i) \ln(t)}{\omega^i} + \sqrt{\frac{\alpha \ln(t)}{\omega^i}} > \mu_i^R + \Delta\mu_i^R \right) \tag{C.7}$$

For sake of notational simplicity, let's note $D^i(\omega^i, t) = \frac{\beta M^i(\omega^i) \ln(t)}{\omega^i}$. Moreover, using $l = \left\lceil \frac{4\alpha \ln n}{(\Delta\mu_i^R)^2} \right\rceil$ and $\omega^i \geq l$, the third term in (C.7) can be upper-bounded by:

$$\sqrt{\frac{\alpha \ln t}{\omega^i}} \leq \sqrt{\frac{\alpha \ln t}{l}} \leq \sqrt{\frac{\alpha \ln t (\Delta\mu_i^R)^2}{4\alpha \ln t}} = \frac{\Delta\mu_i^R}{2}$$

Substituting this last bound into (C.7) and because all terms are positive we get

$$\mathbb{P}(B^i(\omega^i, t) > \mu_1^R) \leq \mathbb{P}\left(\bar{S}^i(\omega^i) - \mu_i^R > \frac{\Delta\mu_i^R}{2} + D^i(\omega^i, t) \right) \tag{C.8}$$

Let, $O_q^i(t)$ being the number of times reward r_q^i associated with state q of arm i has been observed up to time t, hence $\bar{S}^i(\omega^i) = \frac{1}{\omega^i}\sum_{q\in S^i} r_q^i O_q^i(\omega^i)$. Following from (C.8):

$$\begin{aligned}&\mathbb{P}\left(\bar{S}^i(\omega^i) - \mu_i^R \geq \frac{\Delta\mu_i^R}{2} + D^i(\omega^i,t)\right)\\&= \mathbb{P}\left(\sum_{q\in S^i}\left(-r_q^i O_q^i(\omega^i) + \omega^i G_q^i r_q^i \pi_q^i\right) \leq -\omega^i\left(\frac{\Delta\mu_i^R}{2} + D^i(\omega^i,t)\right)\right)\end{aligned} \tag{C.9}$$

As in [145], let us consider a sample path ι and the both events A and B[1]. If $\iota \notin B$ then $\iota \notin A$ and hence $\mathbb{P}(A) \leq \mathbb{P}(B)$.

$$A = \left\{\iota : \sum_{q\in S^i} -r_q^i O_q^i(\omega^i)(\iota) + \omega^i G_q^i r_q^i \pi_q^i \leq -\omega^i\left(\frac{\Delta\mu_i^R}{2} + D^i(\omega^i,t)\right)\right\}$$

$$B = \bigcup_{q\in S^i}\left\{\iota : -r_q^i O_q^i(\omega^i)(\iota) + \omega^i G_q^i r_q^i \pi_q^i \leq -\omega^i \frac{\frac{\Delta\mu_i^R}{2} + D^i(\omega^i,t)}{|S^i|}\right\}$$

It follows that (C.9) is upper bounded as:

$$\leq \sum_{q\in S^i}\mathbb{P}\left(O_q^i(\omega^i) - \omega^i G_q^i\pi_q^i \geq \omega^i \frac{\frac{\Delta\mu_i^R}{2} + D^i(\omega^i,t)}{r_q^i|S^i|}\right) \tag{C.10}$$

$$= \sum_{q\in S^i}\mathbb{P}\left(\frac{\sum_{k=1}^{\omega^i}\mathbf{1}_{(S^i(k)=q)} - \omega^i G_q^i\pi_q^i}{\omega^i G_q^i\hat{\pi}_q^i} \geq \frac{\frac{\Delta\mu_i^R}{2} + D^i(\omega^i,t)}{r_q^i|S^i|G_q^i\hat{\pi}_q^i}\right)$$

$$\leq \sum_{q\in S^i} N_{\mathbf{h}^i}\exp\left(-\frac{\omega^i\left(\frac{\frac{\Delta\mu_i^R}{2}+D^i(\omega^i,t)}{r_q^i|S^i|G_q^i\hat{\pi}_q^i}\right)^2\gamma^i}{28}\right) \tag{C.11}$$

where $|S^i|$ is the arm i state space cardinality, $\hat{\pi}_q^i = \max\{\pi_q^i, 1-\pi_q^i\}$ and $\hat{\pi}_{\max} = \max_{i\in\mathbb{K}}\hat{\pi}_q^i$. Moreover, (C.11) follows from Lemma 4.1 by considering $n = \omega^i$, $f(X_t^i) = \frac{\mathbf{1}_{(S_t^i=q)} - G_q^i\pi_q^i}{G_q^i\hat{\pi}_q^i}$. The conditions in Lemma 4.1 are fulfilled if $G_q^i \geq \frac{1}{\hat{\pi}_{\max}+\pi_q^i}$. Consider an initial distribution $\mathbf{h}^i$ as defined in [145] and eigenvalue gap γ^i for the ith arm, then

$$N_{\mathbf{h}^i} = \left\|\left(\frac{h_q^i}{\pi_q^i}, q\in S^i\right)\right\|_2 \leq \sum_{q\in S^i}\left\|\frac{h_q^i}{\pi_q^i}\right\|_2 \leq \frac{1}{\pi_{\min}}, \tag{C.12}$$

[1] A sample path of a random process is a particular trajectory in time of a given realization.

In order to lighten the notation, let us redefine the following variables. $G_{\max} \equiv G^{q_1}_{\max}$ but the superscript is dropped and $r_{\max} = \max_{q\in S^i, i\in\mathbb{K}} r^i_q$. Moreover, let's define $M_{\min} = \min_{i\in\mathbb{K}} M^i\left(\omega^i\right)$, $\omega_{\max} = \max_{i\in\mathbb{K}} \omega^i$ and $\omega_{\min} = 1$. From (C.11),

$$
\begin{aligned}
&\mathbb{P}\left(\bar{S}^i(\omega^i) - \mu^R_i \geq \frac{\Delta\mu^R_i}{2} + D^i(\omega^i,t)\right) \\
&\leq \frac{|S^i|}{\pi_{\min}} \exp\left(-\frac{\omega^i\left(\frac{\frac{\Delta\mu^R_i}{2}+D^i(\omega^i,t)}{|S^i|r^i_q G^i_q \hat{\pi}^i_q}\right)^2 \gamma^i}{28}\right) \\
&= \frac{|S^i|}{\pi_{\min}} e^{-\frac{\left(\frac{\Delta\mu^R_i}{2}\right)^2 \gamma^i\omega^i}{28(|S^i|r^i_q G^i_q \hat{\pi}^i_q)^2}} e^{-\frac{\left(\frac{2\Delta\mu^R_i \beta M^i(\omega^i)\ln t}{2\omega^i}\right)\gamma^i\omega^i}{28(|S^i|r^i_q G^i_q \hat{\pi}^i_q)^2}} e^{-\frac{\left(\frac{\beta M^i(\omega^i)\ln t}{\omega^i}\right)^2 \omega^i\gamma^i}{28\left(|S^i|r^i_q G^i_q \hat{\pi}^i_q\right)^2}} \\
&= \frac{|S^i|}{\pi_{\min}} e^{-\frac{\left(\Delta\mu^R_i\right)^2\gamma^i\omega^i}{112(|S^i|r^i_q G^i_q \hat{\pi}^i_q)^2}} t^{-\frac{\Delta\mu^R_i \beta M^i(\omega^i)\gamma^i}{28(|S^i|r^i_q G^i_q \hat{\pi}^i_q)^2}} t^{-\frac{\left(\beta M^i(\omega^i)\right)^2 \ln t\gamma^i}{28\left(|S^i|r^i_q G^i_q \hat{\pi}^i_q\right)^2\omega^i}} \\
&\leq \frac{|S^i|}{\pi_{\min}} t^{-\frac{\Delta\mu^R_i \beta M_{\min}\gamma_{\min}}{28 S^2_{\max} r^2_{\max} G^2_{\max} \hat{\pi}^2_{\max}}}
\end{aligned}
\tag{C.13}
$$

where (C.13) is achieved by noting that $\exp\left(-\frac{\left(\Delta\mu^R_i\right)^2\gamma_{\min}\omega_{\min}}{112 S^2_{\max} r^2_{\max} G^2_{\max} \hat{\pi}^2_{\max}}\right) \geq 0$.

Inserting (C.13) into first part of (C.6), we get

$$
\sum_{t=1}^{\infty}\sum_{\omega^i=l}^{t-1} \mathbb{P}(B^i(\omega^i,t) \geq \mu^R_1) \leq \sum_{t=1}^{\infty}\sum_{\omega^i=1}^{t} \frac{|S^i|}{\pi_{\min}} t^{-\frac{\Delta\mu^R_i \beta M_{\min}\gamma_{\min}}{28 S^2_{\max} r^2_{\max} G^2_{\max} \hat{\pi}^2_{\max}}} \leq \frac{|S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \tag{C.14}
$$

where (C.14) is obtained for of $\beta \geq 84 S^2_{\max} r^2_{\max} G^2_{\max} \hat{\pi}^2_{\max} / \left(\gamma_{\min} \Delta\mu^R_i M_{\min}\right)$.

Similarly, we prove the second part of (C.6):

$$
\mathbb{P}(B^1(\omega^1,t) \leq \mu^R_1) = \mathbb{P}\left(\bar{S}^1(\omega^1) - \frac{\beta M^1(\omega^1)\ln t}{\omega^1} + \sqrt{\frac{\alpha \ln t}{\omega^1}} \leq \mu^R_1\right) \tag{C.15}
$$

Let, $C(\omega^1,t) = \sqrt{\frac{\alpha\ln t}{\omega^1}}$ and $D^1(\omega^1,t) = \frac{\beta M^1(\omega^1)\ln(t)}{\omega^1}$ for notation simplification.

$$
\mathbb{P}(B^1(\omega^1,t) \leq \mu^R_1) = \mathbb{P}\left(\bar{S}^1(\omega^1) - \mu^R_1 \leq -(C(\omega^1,t) - D^1(\omega^1,t))\right) \tag{C.16}
$$

Similarly as shown in (C.11), we obtain

$$
\begin{aligned}
&\mathbb{P}\left(\bar{S}^1(\omega^1) - \mu^R_1 \leq -(C(\omega^1,t) - D^1(\omega^1,t))\right) \\
&= \mathbb{P}\left(\sum_{q\in S^1}(r^1_q O^1_q(\omega^1) - \omega^1 G^1_q r^1_q \pi^1_q) \leq -\omega^1(C(\omega^1,t) - D^1(\omega^1,t))\right)
\end{aligned}
\tag{C.17}
$$

$$\leq \sum_{q\in S^1} \mathbb{P}\left(O_q^1(\omega^1) - \omega^1 G_q^1 \pi_q^1 \leq -\frac{\omega^1(C(\omega^1,t) - D^1(\omega^1,t))}{r_q^1|S^1|}\right), \tag{C.18}$$

$$\leq \sum_{q_1\in S^1} N_{\mathbf{h}^1} \exp\left(-w^1 \frac{\left(C(\omega^1,t) - D^1(\omega^1,t)\right)^2 \gamma^1}{28\left(|S^1| G_q^1 r_q^1 \hat{\pi}_q^1\right)^2}\right) \tag{C.19}$$

where, (C.19) from (C.18) follows same as (C.11), (C.18) from (C.17) follows from the same as achieved in (C.10), and (C.19) follows from Theorem 4.1 with $C(\omega^1,t) - D^1(\omega^1,t)$ can be proved to be positive from a certain time. Indeed, $C(\omega^1,t) - D^1(\omega^1,t) = \sqrt{\frac{\ln t}{w^1}}\left(\sqrt{\alpha} - \beta M^1\left(w^1\right)\sqrt{\frac{\ln t}{w^1}}\right)$ and $\exists A \in \mathbb{N}$, $\exists \epsilon > 0$ such that $\forall t > A$, $\sqrt{\frac{\ln t}{w^1}} < \epsilon$. This can be justified by the fact that $\ln t$ grows always slower than t and so w^1 which can be viewed as a fraction of t. Using inequalities in (C.12) and after replacing $C\left(w^1,t\right)$ and $D^1\left(w^1,t\right)$ by their values and after some calculus we get

$$\mathbb{P}(B^1(\omega^1,t) \leq \mu_1^R) \leq \frac{|S^1|}{\pi_{\min}} t^{-\frac{\gamma_{\min}\left(\alpha - 2\sqrt{\alpha}\beta M^1(\omega^1)\sqrt{\frac{\ln t}{w^1}}\right)}{28(S_{\max}G_{\max}r_{\max}\hat{\pi}_{\max})^2}} t^{-\frac{\gamma_{\min}\left(\beta M^1(\omega^1)\right)^2 \ln t}{28(S_{\max}G_{\max}r_{\max}\hat{\pi}_{\max})^2 \omega^1}} \tag{C.20}$$

Moreover, $\exists A \in \mathbb{N}, \exists \epsilon > 0$, such that $\forall t > A$, $\frac{\ln t}{\omega^1} < \sqrt{\frac{\ln t}{\omega^1}} < \epsilon < 1$. We get

$$\mathbb{P}(B^1(\omega^1,t) \leq \mu_1^R) \leq \frac{|S^1|}{\pi_{\min}} t^{-\frac{\gamma_{\min}\left(\alpha - 2\sqrt{\alpha}\beta M^1(\omega^1)\right)}{28(S_{\max}G_{\max}r_{\max}\hat{\pi}_{\max})^2}} \leq \frac{|S^1|}{\pi_{\min}} t^{-\frac{\gamma_{\min}\left(\alpha - 2\sqrt{\alpha}\beta M_{\max}\right)}{28(S_{\max}G_{\max}r_{\max}\hat{\pi}_{\max})^2}} \tag{C.21}$$

where $M_{\max} = \max_{i\in\mathbb{K}} M^i\left(\omega^i\right)$ and where from (C.20) to (C.21) the second term in t is upper bounded by 1. By choosing α such that $\frac{\gamma_{\min}\left(\alpha - 2\sqrt{\alpha}\beta M_{\max}\right)}{28(S_{\max}G_{\max}r_{\max}\hat{\pi}_{\max})^2} \geq 3$ we obtain

$$\mathbb{P}(B^1(\omega^1,t) \leq \mu_1^R) \leq \frac{|S^1|}{\pi_{\min}} t^{-3} \tag{C.22}$$

Replacing (C.22) into second part of (C.6), we get

$$\sum_{t=1}^{\infty}\sum_{\omega^1=1}^{t-1} \mathbb{P}(B^1\left(\omega^1,t\right) \leq \mu_1^R) \leq \frac{|S^1|}{\pi_{\min}} \sum_{t=1}^{\infty}\sum_{\omega^1=1}^{t} t^{-3} = \frac{|S^1|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \tag{C.23}$$

Then the bound follows from combining (C.14) and (C.23):

$$\mathbb{E}[F^i(b(n))|b(n)=b] \leq \frac{4\alpha \ln n}{(\Delta\mu_i^R)^2} + \frac{|S^1| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \tag{C.24}$$

The SB2 block begins with the state ζ^i and ends with a return to the same state. The total number of plays of sub-optimal arm i at the end of block $b(n)$ is estimated by considering the

observations acquired in: i) the total number of plays of sub-optimal arm i during SB2 sub-block (upper bounded by $\frac{1}{\pi^i_{\min}}$), ii) the total number of plays in SB1 before entering in SB2 (upper bounded by $\Omega^i_{\max}$), and iii) one more play during SB3. Thus, we have

$$\mathbb{E}\left[T^i(n)\right] \leq \left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + 1\right) \mathbb{E}\left[F^i(b(n))\right]$$

Thus,

$$\begin{aligned}\sum_{i\in K}(\mu_1^R - \mu_i^R)\mathbb{E}[T^i(n)] \leq & \sum_{i:\mu_i^R<\mu_1^R}\left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + 1\right)\frac{4\alpha\ln n}{\Delta\mu_i^R} \\ & + \sum_{i:\mu_i^R<\mu_1^R}\left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + 1\right)\Delta\mu_i^R\left[\frac{|S^1|+|S^i|}{\pi_{\min}}\sum_{t=1}^{\infty}t^{-2}\right] \\ & \leq Z_1\ln n + Z_2,\end{aligned}$$

where,

$$\begin{aligned}Z_1 &= \sum_{i:\mu_i^R<\mu_1^R}\left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + 1\right)\frac{4\alpha}{\Delta\mu_i^R} \\ Z_2 &= \sum_{i:\mu_i^R<\mu_1^R}\left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + 1\right)\Delta\mu_i^R\left[\frac{|S^1|+|S^i|}{\pi_{\min}}\sum_{t=1}^{\infty}t^{-2}\right]\end{aligned}$$

Let us now verify that $\exists\alpha$ such that (C.22) is verified, i.e.

$$\alpha - 2\sqrt{\alpha}\beta M_{\max} \geq \frac{84\left(S_{\max}G_{\max}r_{\max}\hat{\pi}_{\max}\right)^2}{\gamma_{\min}}$$

Let note the constant $A = 84\left(S_{\max}G_{\max}r_{\max}\hat{\pi}_{\max}\right)^2/\gamma_{\min}$, $X = \sqrt{\alpha}$. α should verify

$$X^2 - 2X\beta M_{\max} - A \geq 0 \tag{C.25}$$

(C.25) admits two solutions

$$X_1 = \beta M_{\max} - \sqrt{\beta^2 M^2_{\max} + A} \tag{C.26}$$

$$X_2 = \beta M_{\max} + \sqrt{\beta^2 M^2_{\max} + A} \tag{C.27}$$

(C.25) is true if $X < X_1$ or $X > X_2$. If $X_1 > 0$ then $\alpha < \left[\beta M_{\max} - \sqrt{\beta^2 M^2_{\max} + A}\right]^2$ or $\alpha > \left[\beta M_{\max} + \sqrt{\beta^2 M^2_{\max} + A}\right]$. If $X_1 < 0$, then $\alpha > \left[\beta M_{\max} + \sqrt{\beta^2 M^2_{\max} + A}\right]$ and the proof is complete.

C.2 Proof of Theorem 4.2

Assume that the regenerative states are denoted by $\zeta = [\zeta^1, \cdots, \zeta^K]$. The expectation w.r.t. the modified sample path is defined as $\mathbb{E}_\zeta$. Let n^b be the time at the end of the last completed block $b(n)$ for all SUs. By expressing the regret in separate terms, i.e. from 1 to n^b and from $n^b + 1$ to n, we get:

$$\begin{aligned}
\Phi^R(n) &= n\mu_1^R - \mathbb{E}\left[\sum_{t=1}^{n} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}(t) r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}(t)\right] \\
\Phi^R(n) &= \mu_1^R \mathbb{E}_\zeta[n^b] - \mathbb{E}_\zeta\left[\sum_{t=1}^{n^b} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right] + \mu_1^R \mathbb{E}_\zeta[n - n^b] - \mathbb{E}_\zeta\left[\sum_{t=n^b+1}^{n} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right] \\
&= \left\{\mu_1^R \mathbb{E}_\zeta[n^b] - \sum_{i=1}^{K} \mu_i^R \mathbb{E}_\zeta\left[T^i(n)\right]\right\} + \sum_{i=1}^{K} \mu_i^R \mathbb{E}_\zeta\left[T^i(n)\right] - \mathbb{E}_\zeta\left[\sum_{t=1}^{n^b} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right] \\
&\quad + \mu_1^R \mathbb{E}_\zeta\left[n - n^b\right] - \mathbb{E}_\zeta\left[\sum_{t=n^b+1}^{n} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right] \qquad \text{(C.28)}
\end{aligned}$$

First difference in (C.28) is bounded logarithmically with the help of Theorem 4.1 as:

$$\begin{aligned}
\mu_1^R \mathbb{E}_\zeta[n^b] - \sum_{i=1}^{K} \mu_i^R \mathbb{E}_\zeta\left[T^i(n)\right] &\leq \mu_1^R \mathbb{E}_\zeta[n^b] - \mu_1^R \mathbb{E}_\zeta\left[T^1(n)\right] - \sum_{i=2}^{K} \mu_i^R \mathbb{E}_\zeta\left[T^i(n)\right] \\
&\leq \mu_1^R n - \mu_1^R \mathbb{E}_\zeta\left[T^1(n)\right] - \sum_{i=2}^{K} \mu_i^R \mathbb{E}_\zeta\left[T^i(n)\right] \\
&= \mu_1^R \sum_{i=1}^{K} \mathbb{E}_\zeta\left[T^i(n)\right] - \mu_1^R \mathbb{E}_\zeta\left[T^1(n)\right] - \sum_{i=2}^{K} \mu_i^R \mathbb{E}_\zeta\left[T^i(n)\right] \\
&\leq \sum_{i=2}^{K} \left(\mu_1^R - \mu_i^R\right) \mathbb{E}_\zeta\left[T^i(n)\right] \leq Z_1 \ln n + Z_2 \qquad \text{(C.29)}
\end{aligned}$$

We have to bound only the two remaining differences in (C.28). We can bound second difference in (C.28) by following same steps as of Theorem 2 in [146], as:

$$\begin{aligned}
&\sum_{i=1}^{K} \mu_i^R \mathbb{E}_\zeta\left[T^i(n)\right] - \mathbb{E}_\zeta\left[\sum_{t=1}^{n^b} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right] \\
\leq\ &\sum_{i=2}^{K} \mu_i^R (\Omega_{\max}^i + 1) \mathbb{E}_\zeta\left[F^i(b(n))\right] + \mu_1^R \Omega_{\max}^1 \sum_{i=2}^{K} \mathbb{E}_\zeta\left[F^i(b(n))\right] \qquad \text{(C.30)}
\end{aligned}$$

where $\Omega_{\max}^i$ is maximum hitting time starting from an initial state for the ith arm. Finally, the last part in (C.28) is bounded as:

$$\mu_1^R \mathbb{E}_\zeta \left[n - n^b \right] - \mathbb{E}_\zeta \left[\sum_{t=n^b+1}^{n} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)} \right] \leq \mu_1^R \left(\frac{1}{\pi_{\min}} + \max_{i \in \{1,\ldots,K\}} \Omega_{\max}^i + 1 \right) . \tag{C.31}$$

The total number of plays of sub-optimal arm between the last block $b(n)$ end time n^b and total time n is upper bounded by a maximum length of a regenerative block which is estimated with observations acquired in: i) the total number of plays of sub-optimal arm during SB2 sub-block (upper bounded by $\frac{1}{\pi_{\min}}$), ii) the total number of plays in SB1 before entering in SB2 (upper bounded by $\Omega_{\max}$), and iii) one more play during SB3. Thus, we have

From (C.29), (C.30), (C.31) and Theorem 4.1, the upper bound on the regret of RQoS-UCB policy is:

$$\begin{aligned} \Phi^R(n) &\leq Z_1 \ln n + Z_2 + \sum_{i=2}^{K} \mu_i^R (\Omega_{\max}^i + 1) \mathbb{E}_\zeta \left[F^i(b(n)) \right] \\ &+ \mu_1^R \Omega_{\max}^1 \sum_{i=2}^{K} \mathbb{E}_\zeta [F^i(b(n))] + \mu_1^R \left(\frac{1}{\pi_{\min}} + \max_{i \in \{1,\ldots,K\}} \Omega_{\max}^i + 1 \right) \leq Z_3 \ln n + Z_4 \end{aligned}$$

where $Z_1, Z_2, Z_3, Z_4, Z_5, Z_6$ and Z_7 are as stated in Theorems 4.1 and 4.2 and obtained by identification with previous quantities and the proof is complete.

C.3 Proof of Theorem 4.3

The total number of frames $f(n)$ up to time n for which the distributed RQoS-UCB policy suggested suboptimal arms i to sense is bounded in the same way as for the RQoS-UCB policy in a single-user restless Markov MAB setting. Let, $\forall j, i : j \in \{1, \cdots, U\}$ and $i \in \{U+1, \cdots, K\}$ denote the set of optimal and suboptimal channel respectively. For convenience, let $T^i(n) = \sum_{j=1}^{U} T^{i,j}(n)$ and $\sum_{i=1}^{K} T^i(n) = nU$, since each SU selects at least one channel to sense in each slot and there are U SUs. Following the same steps as in [10, 71], the total expected number of blocks $\mathbb{E}\left[F^{i,j}(f(n))\right]$ up to frame $f(n)$ for which SU j implementing distributed RQoS-UCB policy selects suboptimal channels i can be bounded as:

$$\begin{aligned} \mathbb{E}\left[F^{i,j}(f(n))\right] &= \mathbb{P}\left[B^j\left(f^j(n), n\right) \leq B^i\left(f^i(n), n\right)\right], \forall j \in \{1, \cdots, U\} \\ \mathbb{E}\left[F^{i,j}(f(n))\right] &\leq \sum_{j=1}^{U} \mathbb{P}\left[B^j\left(f^j(n), n\right) \leq B^i\left(f^i(n), n\right)\right] \end{aligned}$$

For one user, Theorem 4.1 gives the expected number of blocks used to sense a suboptimal band:

$$\mathbb{E}\left[F^{i,1}(b(n))\right] \leq \frac{4\alpha \ln n}{(\Delta\mu_i^R)^2} + \left[\frac{|S^1| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}\right]$$

Thus, we have for U optimal channels for SU j:

$$\mathbb{E}\left[F^{i,j}(f(n))\right] \leq \sum_{j=1}^{U} \left[\frac{4\alpha \ln n}{(\Delta\mu_{i,j}^R)^2} + \left[\frac{|S^j| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}\right]\right]$$

The total number of time suboptimal arm i at the end of block $b(n)$ is estimated by considering the observations acquired in: i) the total number of plays of sub-optimal arm i during SB2 sub-block (upper bounded by $\frac{1}{\pi_{\min}^i}$), ii) the total number of plays in SB1 before entering in SB2 (upper bounded by $\Omega_{\max}^i$), and iii) after the end of SB3 and before finishing of current frame $f(n)$ within the fixed time slot W. Thus, we have

$$\mathbb{E}\left[T^{i,j}(n)\right] \leq \sum_{j=1}^{U} \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + W\right) \left[\frac{4\alpha \ln n}{(\Delta\mu_{i,j}^R)^2} + \left[\frac{|S^j| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}\right]\right]$$

C.4 Proof of Lemma 4.6

Let assume $U = K = 2$ first. The expected number of frames for which the computed indexes $B^{i,j}(n, T^i(n))$ do not reflect the ideal ordering of μ_i^R, is estimated as:

$$\mathbb{E}[f_w^j(n)] \leq l + \sum_{t=1}^{\infty} \sum_{\omega^{2,j}=l}^{t-1} \mathbb{P}(B^{2,j}(\omega^{2,j}, t) > \mu_1^R) + \sum_{t=1}^{\infty} \sum_{\omega^{1,j}=1}^{t-1} \mathbb{P}(B^{1,j}(\omega^{1,j}, t) \leq \mu_1^R) \tag{C.32}$$

where (C.32) states that the policy index of user 2 is greater than user 1 which indicates wrong ordering of the mean reward. Following the Theorem 4.1, $\mathbb{E}[f_w^j(n)]$ is upper bounded as

$$\mathbb{E}[f_w^j(n)] \leq \frac{4\alpha \ln n}{(\Delta\mu_{1,2}^R)^2} + \frac{|S^1| + |S^2|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \tag{C.33}$$

For the case where $(U, K) > 2$:

$$\mathbb{E}[f_w(n)] \leq U \sum_{a=1}^{U} \sum_{b=1}^{K} \left(l + \sum_{t=K+1}^{n} \mathbf{1}_{\left(B^{a,j}(t-1) \leq B^{b,j}(t-1), T_w^j(t-1) \geq l\right)}\right)$$

where a and b denote channels with a^{th} and b^{th} highest mean rewards. it can be shown:

$$\mathbb{E}[f_w(n)] \leq U \sum_{a=1}^{U} \sum_{b=1}^{K} \left[\frac{4\alpha \ln n}{(\Delta\mu_{a,b}^R)^2} + \frac{|S^a| + |S^b|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}\right]$$

Finally, same as Theorem 4.3, we have

$$\mathbb{E}[T_w(n)] \leq U\sum_{a=1}^{U}\sum_{b=1}^{K}\left(\frac{1}{\pi^b_{\min}}+\Omega^b_{\max}+W\right)\left[\frac{4\alpha\ln n}{(\Delta\mu^R_{a,b})^2}+\frac{|S^a|+|S^b|}{\pi_{\min}}\sum_{t=1}^{\infty}t^{-2}\right]$$

C.5 Proof of Theorem 4.4

Let $\{1,\cdots,U\}$ and $\{U+1,\cdots,K\}$ denote the set of optimal and suboptimal channels, respectively, and n^b is the time at the end of the last completed block $b(n)$ as detailed in Fig. 4.3. Following the same spirit as in the proof of Theorem 4.2, we have

$$\begin{aligned}\Phi^M(n) &= \sum_{j=1}^{U}n\mu^R_j-\sum_{j=1}^{U}\mathbb{E}\left[\sum_{t=1}^{n}G^{\mathcal{A}(t),j}_{q_{\mathcal{A}(t)}}(t)r^{\mathcal{A}(t),j}_{q_{\mathcal{A}(t)}}(t)\right]\\
&=\left\{\mathbb{E}_\zeta[n^b]\sum_{j=1}^{U}\mu^R_j-\sum_{j=1}^{U}\sum_{i=1}^{K}\mu^R_i\mathbb{E}_\zeta\left[V^{i,j}(n)\right]\right\}\\
&+\left\{\sum_{j=1}^{U}\sum_{i=1}^{K}\mu^R_i\mathbb{E}_\zeta\left[V^{i,j}(n)\right]-\sum_{j=1}^{U}\mathbb{E}_\zeta\left[\sum_{t=1}^{n^b}r^{\mathcal{A}(t),j}_{q_{\mathcal{A}(t)}}G^{\mathcal{A}(t),j}_{q_{\mathcal{A}(t)}}\right]\right\}\\
&+\left\{\mathbb{E}_\zeta[n-n^b]\sum_{j=1}^{U}\mu^R_j-\sum_{j=1}^{U}\mathbb{E}_\zeta\left[\sum_{t=n^b+1}^{n}r^{\mathcal{A}(t),j}_{q_{\mathcal{A}(t)}}G^{\mathcal{A}(t),j}_{q_{\mathcal{A}(t)}}\right]\right\}\end{aligned}\tag{C.34}$$

where $V^{i,j}(n)$ is the total number of times where an SU j is the only one to sense and access the channel i up to time n.

Working with the first part of (C.34) we have:

$$\begin{aligned}&\left\{\mathbb{E}_\zeta[n^b]\sum_{j=1}^{U}\mu^R_j-\sum_{j=1}^{U}\sum_{i=1}^{K}\mu^R_i\mathbb{E}_\zeta\left[V^{i,j}(n)\right]\right\}\\
&\leq\sum_{i=1}^{U}\mu^R_i\left(\mathbb{E}_\zeta[n^b]-\mathbb{E}_\zeta\left[V^i(n)\right]\right)\\
&\leq\mu^R_1\left(U\mathbb{E}_\zeta[n^b]-\sum_{i=1}^{U}\mathbb{E}_\zeta\left[V^i(n)\right]\right) &\text{(C.35)}\\
&=\mu^R_1\left(U\mathbb{E}_\zeta[n^b]+\mathbb{E}_\zeta[C(n)]-\sum_{i=1}^{U}\mathbb{E}_\zeta\left[T^i(n)\right]\right)\\
&\leq\mu^R_1\left(\mathbb{E}_\zeta[C(n)]+\sum_{i=U+1}^{K}\mathbb{E}_\zeta\left[T^i(n^b)\right]\right) &\text{(C.36)}\\
&\leq\mu^R_1\left(U(\mathbb{E}[\Upsilon]+1)\,\mathbb{E}[T_w(n)]+\sum_{i=U+1}^{K}\mathbb{E}_\zeta\left[T^i(n)\right]\right) &\text{(C.37)}\end{aligned}$$

where in (C.35), we use the fact that $\mathbb{E}_\zeta\left[V^i(n)\right] = \sum_{j=1}^{U} \mathbb{E}_\zeta\left[V^{i,j}(n)\right]$ because of $V^i(n) < \mathbb{E}_\zeta[n^b]$, since the total number of time a unique SU occupies the channel i is at most $\mathbb{E}_\zeta[n^b]$. In (C.36), we use the fact that the total number of collisions in U optimal channels is defined as $C(n) = \sum_{i=1}^{U}\left(T^i(n) - V^i(n)\right)$. Moreover, in (C.36), we have $U\mathbb{E}_\zeta[n^b] = \left(\sum_{i=1}^{U} T^i(n^b) + \sum_{i=U+1}^{K} T^i(n^b)\right)$, and $\mathbb{E}_\zeta\left[T^i(n)\right] \geq \mathbb{E}_\zeta\left[T^i(n^b)\right]$. Finally (C.37) is achieved by applying Theorem 4.3 and Theorem 3 from [10].

Concerning the second part of (C.34), we have:

$$
\begin{aligned}
&\sum_{j=1}^{U}\sum_{i=1}^{K} \mu_i^R \mathbb{E}_\zeta\left[V^{i,j}(n)\right] - \sum_{j=1}^{U} \mathbb{E}_\zeta\left[\sum_{t=1}^{n^b} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}\right] \\
&\leq \left\{\sum_{j=1}^{U}\sum_{i=1}^{U} \mu_i^R \mathbb{E}_\zeta\left[V^{i,j}(n)\right] - \sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i} r_q^i G_q^i \mathbb{E}_\zeta\left[\sum_{k=1}^{F^{i,j}(f(n))} \sum_{S(t)\in S^{i,j}(k)} \mathbb{1}_{(S(t)=q)}\right]\right\} \\
&+ \left\{\sum_{j=1}^{U}\sum_{i=U+1}^{K} \mu_i^R \mathbb{E}_\zeta\left[V^{i,j}(n)\right] - \sum_{j=1}^{U}\sum_{i=U+1}^{K}\sum_{q\in S^i} r_q^i G_q^i \mathbb{E}_\zeta\left[\sum_{k=1}^{F^{i,j}(f(n))} \sum_{S(t)\in S_2^{i,j}(k)} \mathbb{1}_{(S(t)=q)}\right]\right\} \\
&+ \left\{\sum_{i=1}^{U}\sum_{q\in S^i} r_q^i G_q^i \mathbb{E}_\zeta\left[C(n)\right] + \sum_{j=1}^{K} \mathbb{E}_\zeta\left[C(n)\right] \sum_{i=U+1}^{K}\sum_{q\in S^i} r_q^i G_q^i\right\}
\end{aligned}
\tag{C.38}
$$

where the first and the second part of (C.38) are the rewards collected in optimal and suboptimal arms separately, because $\{1,\cdots,U\}$ and $\{U+1,\cdots,K\}$ refers to the set of optimal and suboptimal arms, respectively. The inequality in (C.38) comes from counting only the rewards obtained during the SB2s $S_2^{i,j}(k)$ for all suboptimal arms. The last part is the reward loss due to the collisions in optimal and suboptimal channels, where $\mathbb{E}_\zeta\left[C(n)\right]$ is the total number of collisions in optimal channels as defined in Lemma 4.7.

Let us start with the first part of (C.38) and following the same reasoning than in [146], we have:

$$
\mathbb{E}_\zeta[\bar{b}^{i,j}(n)] \leq \sum_{k=U+1}^{K} \mathbb{E}_\zeta[F^{k,j}(f(n))], \quad \forall i \in \{1,\cdots,U\}. \tag{C.39}
$$

where, $\{\bar{b}^{i,j}\}$ is the total number of the joined blocks, and is always less than or equal to the total number of discontinuities in the observation of optimal arm. Thus,Each successive combined block $\bar{X}^{i,j}$ can be separated into two sub-blocks: i) $\bar{X}_1^{i,j}$ consisting in the states observed from the beginning of $\bar{X}^{i,j}$ (empty if the first state is $\zeta^{i,j}$) to the state right before observing $\zeta^{i,j}$, and ii) $\bar{X}_2^{i,j}$ consisting in the rest of $\bar{X}^{i,j}$.

Therefore, the first part of (C.38) can upper bound as:

$$\sum_{j=1}^{U}\sum_{i=1}^{U}\mu_i^R\mathbb{E}_\zeta\left[V^{i,j}(n)\right]-\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{F^{i,j}(f(n))}\sum_{S_t^{i,j}\in S^i(k)}\mathbb{1}_{(S_t^{i,j}=q)}\right]$$
$$\leq\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\pi_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}|\bar{X}_2^{i,j}(k)|\right]-\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}\sum_{S(t)\in\bar{S}_2^{i,j}(k)}\mathbb{1}_{(S(t)=q)}\right]$$
$$+\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\pi_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}|\bar{X}_1^{i,j}(k)|\right]-\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}\sum_{S(t)\in\bar{S}_1^{i,j}(k)}\mathbb{1}_{(S(t)=q)}\right] \tag{C.40}$$

$$\leq\sum_{j=1}^{U}\sum_{i=1}^{U}\frac{\mu_i^R}{\pi_\zeta^i}\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]-\sum_{j=1}^{U}\sum_{i=1}^{U}\frac{\mu_i^R}{\pi_\zeta^i}\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]+\sum_{j=1}^{U}\sum_{i=1}^{U}\mu_i^R\Omega_{\max}^i\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]-0 \tag{C.41}$$

$$<\sum_{j=1}^{U}\sum_{i=1}^{U}\mu_i^R\Omega_{\max}^i\sum_{K=U+1}^{K}\mathbb{E}_\zeta[F^{k,j}(f(n))] \tag{C.42}$$

where (C.40) comes from counting the rewards in two different sub-blocks SB1 and SB2. Let $|\bar{X}_2^{i,j}(k)|$ denotes the total number of states in the $k-$th joined SB2 block in which the optimal channel is selected. The inequality in (C.42) is obtained by observing that $\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}|\bar{X}_2^{i,j}(k)|\right]\leq\frac{1}{\pi_\zeta^i}\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]$ in SB2 and $\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}|\bar{X}_1^{i,j}(k)|\right]\leq\Omega_{\max}^i\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]$ in SB1. Since rewards are positive, the last part of (C.40) is larger than 0, and applying Lemma 4.3 to the second part of (C.40) with $\tau=\bar{b}^{i,j}(n)$ and $1/\pi_\zeta^i$, we get $\frac{\pi_q^i}{\pi_\zeta^i}\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]$.

The second part of (C.38) can be upper bound as:

$$\left\{\sum_{j=1}^{U}\sum_{i=U+1}^{K}\mu_i^R\mathbb{E}_\zeta\left[V^{i,j}(n)\right]-\sum_{j=1}^{U}\sum_{i=U+1}^{K}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{F^{i,j}(f(n))}\sum_{S(t)\in S_2^{i,j}(k)}\mathbb{1}_{(S(t)=q)}\right]\right\} \tag{C.43}$$

$$\leq\sum_{j=1}^{U}\sum_{i=U+1}^{K}\mu_i^R\mathbb{E}_\zeta\left[T^{i,j}(n)\right]-\sum_{j=1}^{U}\sum_{i=U+1}^{K}\frac{\mu_i^R}{\pi_\zeta^i}\mathbb{E}_\zeta\left[F^{i,j}(f(n))\right] \tag{C.44}$$

$$=\sum_{j=1}^{U}\sum_{i=U+1}^{K}\mu_i^R\left(\Omega_{\max}^i+W\right)\mathbb{E}_\zeta\left[F^{i,j}(f(n))\right] \tag{C.45}$$

where (C.44) comes from $V^{i,j}(n)\leq T^{i,j}(n)$, and applying Lemma 4.3 to the second part of (C.43). (C.45) is obtained with Theorem 4.3.

Now we bound the last part of (C.38):

$$\begin{aligned}
&\left\{\sum_{i=1}^{U}\sum_{q\in S^i} r_q^i G_q^i \mathbb{E}_\zeta\left[C(n)\right] + \sum_{j=1}^{K}\mathbb{E}_\zeta\left[C(n)\right]\sum_{i=U+1}^{K}\sum_{q\in S^i} r_q^i G_q^i\right\} \\
&\leq \mathbb{E}_\zeta\left[C(n)\right]\left[\sum_{i=1}^{U}\frac{\mu_i^R}{\pi_{\min}^i} + K\sum_{i=U+1}^{K}\frac{\mu_i^R}{\pi_{\min}^i}\right] \leq \mathbb{E}_\zeta\left[C(n)\right]\frac{\mu_1^R}{\pi_{\min}}\left[U + K(K-U)\right] \\
&\leq \left[K^2 - KU + U\right]\frac{\mu_1^R}{\pi_{\min}} U(\mathbb{E}[\Upsilon]+1)\,\mathbb{E}[T_w(n)] \qquad \text{(C.46)}
\end{aligned}$$

where (C.46) is achieved by considering $\mu_i^R = \sum_{q\in S^i} r_q^i G_q^i \pi_q^i$ and using Lemma 4.7.

Combining (C.46), (C.45) and (C.42) into (C.38), we immediately have upper bound on (C.38), as:

$$\begin{aligned}
&\sum_{j=1}^{U}\sum_{i=1}^{K}\mu_i^R\mathbb{E}_\zeta\left[V^{i,j}(n)\right] - \mathbb{E}_\zeta[\sum_{j=1}^{U}\sum_{t=1}^{n^b} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}] \\
&\leq \sum_{j=1}^{U}\sum_{i=1}^{U}\mu_i^R\Omega_{\max}^i\sum_{K=U+1}^{K}\mathbb{E}_\zeta[F^{k,j}(f(n))] + \sum_{j=1}^{U}\sum_{i=U+1}^{K}\mu_i^R\left(\Omega_{\max}^i + W\right)\mathbb{E}_\zeta\left[F^{i,j}(f(n))\right] \\
&+ \left[K^2 - KU + U\right]\frac{\mu_1^R}{\pi_{\min}} U(\mathbb{E}[\Upsilon]+1)\,\mathbb{E}[T_w(n)] \qquad \text{(C.47)}
\end{aligned}$$

Now, in order to upper bound the multi-player sum regret in (C.34), we have to upper bound the last part of (C.34), and is bounded as:

$$\mathbb{E}_\zeta[n-n^b]\sum_{j=1}^{U}\mu_j^R - \mathbb{E}_\zeta[\sum_{j=1}^{U}\sum_{t=n^b+1}^{n} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}] \leq \sum_{j=1}^{U}\mu_j^R\left(\frac{1}{\pi_\zeta} + \Omega_{\max} + W\right) \qquad \text{(C.48)}$$

where, $\left(\frac{1}{\pi_\zeta} + \Omega_{\max} + W\right)$ denotes the maximum length of a block (SB1, SB2 and SB3) in (C.48). The maximum length of frame is estimated by considering the observations acquired in: i) the maximum number of plays of any arm during SB2 sub-block (upper bounded by $\frac{1}{\pi^\zeta}$), ii) the maximum number of plays in SB1 before entering in SB2 (upper bounded by $\Omega_{\max}$), and iii) after the end of SB3 and before finishing of current frame $f(n)$ within the fixed time slot W. Furthermore, $\sum_{j=1}^{U}\mu_j^R$ is the sum of mean reward of optimal arms.

Combining (C.37), (C.47), (C.48) and using Lemma 4.6 and Theorem 4.3, the upper bound on the regret $\Phi^M(n)$ of the multi-user distributed RQoS-UCB policy is obtained as:

$$\begin{aligned}
\Phi^M(n) \leq \mu_1^R \left(U(\mathbb{E}[\Upsilon]+1)\,\mathbb{E}[T_w(n)] + \sum_{i=U+1}^{K} \mathbb{E}_\zeta\left[T^i(n)\right] \right) \\
+ \sum_{j=1}^{U}\sum_{i=1}^{U} \mu_i^R \Omega_{\max}^i \sum_{K=U+1}^{K} \mathbb{E}_\zeta[F^{k,j}(f(n))] \\
+ \sum_{j=1}^{U}\sum_{i=U+1}^{K} \mu_i^R \left(\Omega_{\max}^i + W\right) \mathbb{E}_\zeta\left[F^{i,j}(f(n))\right] \\
+ \left[K^2 - KU + U\right] \frac{\mu_1^R}{\pi_{\min}} U(\mathbb{E}[\Upsilon]+1)\,\mathbb{E}[T_w(n)] + U\mu_1^R \left(\frac{1}{\pi_\zeta} + \Omega_{\max} + W\right) \\
\Phi^M(n) \leq X_3 \ln n + X_4
\end{aligned}$$

where X_i, $i \in \{1, \cdots, 9\}$ are as stated in Theorem 4.4 and the proof is complete.

List of Figures

1.1 Normalized real traffic load during one week that are recorded by cellular operator. The data captures voice call information over one week with a resolution of one second in a metropolitan urban area, and are averaged over 30 minute time-scale [123]. . . . 4

2.1 Cognitive Cycle of [127] . . . 12
2.2 Decision making techniques depending on the two-dimensional design space {*System Complexity, a priori information level*}. . . . 13
2.3 Taxonomy of Machine Learning Algorithms . . . 14
2.4 Reinforcement Learning Framework . . . 15
2.5 Classification of Reinforcement Learning Framework . . . 16
2.6 Taxonomy of transfer learning problem according to three main dimensions: the transfer setting, the transferred knowledge, and the objective [128] . . . 18
2.7 The three main objectives of transfer learning [129]. The red circles highlight the performance improvement in the learning process. . . . 19
2.8 Classification of MAB problem formulation . . . 20

3.1 PUs activity pattern . . . 37
3.2 (a), (b) Probability of success P_{Succ} of MAB policies, i.e. UCB1 and TS, with respect to the average LZ complexity and the probability of free P_{free}. Each point denotes a particular instance of MAB policies applied to $K = 5$ channels. The number of random combinations which we analyzed is 2000. . . . 42
3.3 (a), (b) Probability of success P_{Succ} of MAB policies, i.e. UCB1 and TS, with respect to the OI factor H_{OI} and the probability of free P_{free} applied to same ensemble of channels. . . . 43
3.4 (a), (b) Each point denotes the difference between the probability of success of MAB policies, i.e. UCB1 and TS, and the probability of success of the random channel selection (RCS) approach applied to $K = 5$ channels, respectively. The number of random combinations which we analyzed is 2000. . . . 44
3.5 Average percentage of improvement in the probability of success of MAB policies, i.e. UCB1 and TS, with respect to RCS policy as a function of the OI factor H_{OI} and the probability of free P_{free}. Each point denotes an average percentage of improvement achieved by MAB policies for different H_{OI} and P_{free}. . . . 45

4.1 Interweave cognitive cell scenario . . . 51
4.2 Cognitive radio frame structure with repetitive sensing, learning and transmission slots . . . 51

4.3 Example of block (i.e. SB1, SB2 and SB3 sub-blocks) operation of RQoS-UCB policy. At the end of block 1, RQoS-UCB policy computes the index based on the observations collected in SB2 block, finds a channel having the highest index among the set of channels $\mathcal{K}$, and moves to the channel (for example K) with the highest index for block 2. 57
4.4 Running cycle for 2 different SUs using distributed RQoS-UCB policy. Actions of player 1 and 2 are listed on top and bottom, respectively 59
4.5 The list of state corresponds to a configuration of U SUs in U number of channels. 65
4.6 Percentage of transmission opportunities exploited, achievable throughput and SER for single SUs w.r.t. the number of frames. 71
4.7 Percentage of transmission opportunities exploited after 10000 time steps over 100 iterations for single SU w.r.t. scenarios with different OI factor. 73
4.8 Percentage of transmission opportunities for 2 SUs w.r.t. the number of frames a) and w.r.t. the number of SUs operating in network b). 74
4.9 Average achievable throughput for 2 SUs w.r.t. the number of frames a) and w.r.t. the number of SUs operating in the network b). 75
4.10 Average SER for 2 SUs w.r.t. the number of frames a) and w.r.t. the number of SUs operating in the network b). 77
4.11 Fairness analysis for 4 SUs implementing RQoS-UCB policy w.r.t. the number of frames. 78

5.1 Normalized real traffic load during one week that are recorded by cellular operator. The data captures voice call information over one week with a resolution of one second in a metropolitan urban area, and are averaged over 30 minute time-scale [123]. 84
5.2 Reinforcement learning (RL) framework for BS switching operation. 87
5.3 Transfer learning for Transfer RQoS-UCB (TRQoS-UCB) policy. 89
5.4 Snapshot of cell coverage when all BSs (BSs 1-5 are macro BSs, and BSs 6-10 are micro BSs) are ON. 93
5.5 Performance comparison under various homogeneous Poisson point process traffic intensity with transferred knowledge estimated from a source task with traffic intensity $\Lambda^{source} = 0.05 \times 10^{-4}$. 95
5.6 Network energy efficiency (EE) with several learning algorithms with respect to time of the day 97
5.7 Average percentage of energy saving achieved by several learning algorithms with respect to time of the day. 97
5.8 Average number of active BSs suggested by several learning algorithms with respect to time of the day. 98
5.9 Per-flow delay is seconds with several learning algorithms with respect to time of the day 99
5.10 Average number of BS mode switch required at each iteration during one hour with respect to time of the day. 99

6.1 Example of a measurement of the high activity scenario (acquired at weekends) of the `HFSA_IDeTIC_F1_V01` [107]. 103
6.2 Comparison of the successful transmission percentage per group with $\alpha = 0.4$ and $K = 8$. 104
6.3 Detail of the comparative of the successful transmission rate per group with $\alpha = 0.4$ and $K = 8$. 105

6.4 Histogram of the percentage of improvement with UCB1 respect to random selection with $\alpha = 0.4$ and $K = 8$. 106
6.5 Left hand side (laptop with GNU Radio + USRP) is generating primary user traffic on 8 channels (Tx). Right hand side (laptop with Simulink + USRP) is a secondary user employing energy detector for channel sensing and online learning algorithm based learning policies (Rx). A spectrum analyzer shows the RF signals. 107
6.6 Detailed block diagram of the proposed primary user traffic generator implemented in GNU Radio Companion. 108
6.7 Block diagram of the Simulink based testbed consisting of SU with decision making policies. 109
6.8 Comparisons of successful transmission percentage for various P_{vac} distributions from Scenario 1 to 4 of different learning approaches for single-player and multi-player, respectively. 111
6.9 Comparisons of optimal arm selection percentage of different learning approaches from Scenario 1 to 4 for single-player and multi-player, respectively. 112
6.10 Comparisons of channel switching cost of different learning approaches from Scenario 1 to 4 for single-player and multi-player, respectively. 113
6.11 Comparisons of number of collisions of different learning approaches from Scenario 1 to 4 for multi-player setting. 114

2.1 Abruptly changing non-stationary scenario 127
2.2 UCB1, QoS-UCB and DQoS-UCB: Cumulative regret 131
2.3 QoS-UCB and DQoS-UCB: Average reward 132
2.4 QoS-UCB and DQoS-UCB: Optimal channel selection percentage and successful transmission percentage (STP) comparison 132

List of Tables

2.1 An overview of the single-player policies in the stateless multi-armed bandit framework. The symbols have the following meaning: '↑' ('↑↑') means that the respective property is (strongly) satisfied. In addition, '−' means the property is not known. 26

2.2 An overview of the multi-player policies in the stateless multi-armed bandit framework. The symbols have the following meaning: '↑' ('↑↑') means that the respective property is (strongly) satisfied. In addition, '−' means the property is not known. 27

2.3 An overview of the learning policies in the Markovian multi-armed bandit framework. The symbols have the following meaning: '↑' ('↑↑') means that the respective property is (strongly) satisfied. In addition, '−' means the property is not known. 29

4.1 State transition probabilities, mean availability, empirical mean quality and global mean reward for scenario 1 with $P_{free} = 0.999$ and OI factor $H_{OI} = 0.81$. 69

4.2 Mean availability, empirical mean quality and global mean reward for 3 scenarios with $P_{free} = 0.91$ and different level of OI factor H_{OI}. 69

4.3 Algorithms complexity for K channels and N time slots 78

5.1 Used Simulation Parameters 92

5.2 Algorithms Comparison. 94

6.1 Several scenarios to verify the proposed approach on testbed. 111

B.1 State transition probabilities P^i in non-stationary environment 129

B.2 Observed channel reward and estimated mean reward in non-stationary environment 130

Bibliography

[1] M. Aboy, R. Hornero, D. Abasolo, and D. Alvarez. Interpretation of the lempel-ziv complexity measure in the context of biomedical signal analysis. *IEEE Transactions on Biomedical Engineering*, 53(11):2282–2288, Nov 2006.

[2] R. Agrawal. *Sample mean based index policies with O(log n) regret for the multi-armed bandit problem.*, volume 27, pages 1054–1078. Applied Probability Trust, 1995.

[3] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, June 2012.

[4] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty. Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *COMPUTER NETWORKS JOURNAL (ELSEVIER*, 50:2127–2159, 2006.

[5] A. Alam and L. Dooley. A scalable multimode base station switching model for green cellular networks. In *IEEE Wireless Communications and Networking Conference*, March 2015.

[6] A. S. Alam, L. S. Dooley, and A. S. Poulton. Energy efficient relay-assisted cellular network model using base station switching. In *2012 IEEE Globecom Workshops*, pages 1155–1160, Dec 2012.

[7] A. S. Alam, L. S. Dooley, and A. S. Poulton. Traffic-and-interference aware base station switching for green cellular networks. In *2013 IEEE 18th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 63–67, Sept 2013.

[8] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2nd edition, 2010.

[9] A. Amraoui, B. Benmammar, F. Krief, and F. T. Bendimerad. Intelligent wireless communication system using cognitive radio. *IJDPS International Journal of Distributed and Parallel Systems*, 3(2):91–104, 2012.

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

[11] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, Nov 1987.

[12] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards. *Automatic Control, IEEE Transactions on*, 32(11):977–982, Nov 1987.

[13] J.-Y. Audibert and S. Bubeck. Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory - 2010*, page 13 p., Haifa, Israel, June 2010.

[14] J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19):1876–1902, Apr. 2009.

[15] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, May 2002.

[16] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, Jan. 2003.

[17] P. Auer and R. Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1):55–65, 2010.

[18] O. Besbes, Y. Gur, and A. Zeevi. Non-stationary Stochastic Optimization. *ArXiv e-prints*, July 2013.

[19] T. E. Bogale, L. Vandendorpe, and L. B. Le. Sensing throughput tradeoff for cognitive radio networks with noise variance uncertainty. In *Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), 2014 9th International Conference on*, pages 435–441. IEEE, 2014.

[20] M. Bóna. *A Walk Through Combinatorics: An Introduction to Enumeration and Graph Theory*. World Scientific Publishing Company, 2 edition, Oct. 2006.

[21] P. Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues (Texts in Applied Mathematics)*. Springer, corrected edition, Feb 2008.

[22] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[23] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation. *Annals of Statistics*, 41(3):1516–1541, 2013. Accepted, to appear in Annals of Statistics.

[24] L. A. Celiberto, J. P. Matsuura, R. L. de Mantaras, and R. A. C. Bianchi. Using transfer learning to speed-up reinforcement learning: A cased-based approach. In *Robotics Symposium and Intelligent Robotic Meeting (LARS), 2010 Latin American*, pages 55–60, Oct 2010.

[25] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[26] X. Chen, Z. Zhao, and H. Zhang. Power entangling and matching in cognitive wireless mesh networks by applying conjecture based multi-agent qq-learning approach. In *GLOBECOM Workshops (GC Wkshps), 2010 IEEE*, pages 1124–1129, Dec 2010.

[27] X. Chen, Z. Zhao, and H. Zhang. Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks. *Mobile Computing, IEEE Transactions on*, 12(11):2155–2166, Nov 2013.

[28] X. Chen, Z. Zhao, H. Zhang, and T. Chen. Applying multi-agent q-learning scheme in cognitive wireless mesh networks for green communications. In *Personal, Indoor and Mobile Radio Communications Workshops (PIMRC Workshops), 2010 IEEE 21st International Symposium on*, pages 336–340, Sept 2010.

[29] L. Chiaraviglio, D. Ciullo, M. Meo, and M. Ajmone Marsan. Energy-aware UMTS access networks. *Proc. of WPMC Symposium*, pages 8–11, 2008.

[30] Cisco. Cisco visual networking index: Global mobile data traffic forecast update. 2012–2017.

[31] C. Clancy, J. Hecker, E. Stuntebeck, and T. O'Shea. Applications of machine learning to cognitive radio networks. *IEEE Wireless Communications*, 14(4):47–52, August 2007.

[32] W. Dai, Y. Gai, and B. Krishnamachari. Efficient online learning for opportunistic spectrum access. In *INFOCOM, 2012 Proceedings IEEE*, pages 3086–3090, March 2012.

[33] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao. The non-bayesian restless multi-armed bandit: A case of near-logarithmic regret. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2940–2943, May 2011.

[34] S. Darak, N. Modi, A. Nafkha, and C. Moy. Spectrum utilization and reconfiguration cost comparison of various decision making policies for opportunistic spectrum access using real radio signals. In *11th EAI International Conference on Cognitive Radio Oriented Wireless Networks 2016 (CROWNCOM 2016), Grenoble, France, May 2016*, June 2016.

[35] S. J. Darak, C. Moy, and J. Palicot. Proof-of-concept system for opportunistic spectrum access in multi-user decentralized networks. *EAI Endorsed Transactions on Cognitive Communications*, 16(7), 9 2016.

[36] L. Doyle and T. Forde. *The Wisdom of Crowds: Cognitive Ad Hoc Networks*, pages 203–221. John Wiley & Sons, Ltd, 2007.

[37] E. Even-Dar, S. Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, COLT '02, pages 255–270, London, UK, UK, 2002. Springer-Verlag.

[38] FCC. Notice of proposed rule making and order, December 2003.

[39] Federal Communications Commission. Spectrum policy task force, Nov. 2002.

[40] A. J. Fehske, F. Richter, and G. P. Fettweis. Energy efficiency improvements through micro sites in cellular mobile radio networks. In *2009 IEEE Globecom Workshops*, pages 1–5, Nov 2009.

[41] M. D. Felice, K. R. Chowdhury, and L. Bononi. Learning with the bandit: A cooperative spectrum selection scheme for cognitive radio networks. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pages 1–6, Dec 2011.

[42] G. P. Fettweis and E. Zimmermann. ICT energy consumption-trends and challenges. In *in Proc. WPMC, Lapland, Finland*, page vol. 4, Sep. 2008 2009.

[43] FP7 ABSOLUTE. Aerial Base Stations with Opportunistic Links for Unexpected & Temporary Events, 2015.

[44] FP7 EARTH. Energy efficiency analysis of the reference systems, areas of improvements and target breakdown. Technical report.

[45] FP7 FARAMIR. Enabling Spectrum-Aware Radio Access for Cognitive Radios, 2012.

[46] F. Fu and M. van der Schaar. Learning to compete for resources in wireless stochastic games. *IEEE Transactions on Vehicular Technology*, 58(4):1904–1919, May 2009.

[47] Y. Gai and B. Krishnamachari. Distributed stochastic online learning policies for opportunistic spectrum access. *IEEE Transactions on Signal Processing*, 62(23):6184–6193, Dec 2014.

[48] Y. Gai, B. Krishnamachari, and M. Liu. Online learning for combinatorial network optimization with restless markovian rewards. In *Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2012 9th Annual IEEE Communications Society Conference on*, pages 28–36, June 2012.

[49] A. Galindo-Serrano and L. Giupponi. Distributed q-learning for interference control in ofdma-based femtocell networks. In *Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st*, pages 1–5, May 2010.

[50] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979.

[51] A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In J. Kivinen, C. Szepesvari, E. Ukkonen, and T. Zeugmann, editors, *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*, pages 174–188. Springer Berlin Heidelberg, 2011.

[52] J. Gong, S. Zhou, and Z. Niu. A Dynamic Programming Approach for Base Station Sleeping in Cellular Networks. *IEICE Transactions on Communications*, 95:551–562, 2012.

[53] D. Grace and H. Zhang. *Cognitive Communications: Distributed Artificial Intelligence (DAI), Regulatory Policy and Economics, Implementation*. John Wiley & Sons, 8 2012.

[54] W. Guo and T. O'Farrell. Dynamic cell expansion with self-organizing cooperation. *IEEE Journal on Selected Areas in Communications*, 31(5):851–860, May 2013.

[55] F. Han, Z. Safar, and K. J. R. Liu. Energy-efficient base-station cooperative operation with guaranteed QoS. *IEEE Transactions on Communications*, 61(8):3505–3517, August 2013.

[56] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag. Multi-armed Bandit, Dynamic Environments and Meta-Bandits. In *NIPS-2006 workshop, Online trading between exploration and exploitation*, Whistler, Canada, Nov. 2006.

[57] S. Haykin. Cognitive radio: brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 23(2):201–220, Feb 2005.

[58] E. Hossain, D. Niyato, and Z. Han. *Dynamic Spectrum Access and Management in Cognitive Radio Networks*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[59] M. F. Hossain, K. S. Munasinghe, and A. Jamalipour. Distributed inter-bs cooperation aided energy efficient load balancing for cellular networks. *IEEE Transactions on Wireless Communications*, 12(11):5929–5939, November 2013.

[60] IEEE 802.16 Broadband Wireless Access Working Group. IEEE 802.16 m evaluation methodology document (EMD). July 2008.

[61] IEEE 802.19 Wireless Coexistence Working Group (WG). IEEE 802.19 Wireless Coexistence in the TV White Space, 2014.

[62] IEEE 802.22 Working Group. IEEE 802.22 Wireless Regional Area Networks.

[63] IETF PAWS WG Document. Protocol to Access White-Space (PAWS) Databases: Use Cases and Requirements, 2015.

[64] International Telecommunication Union. *Report ITU-R SM.2152 Definitions of Software Defined Radio (SDR) and Cognitive Radio System (CRS).* 2009.

[65] A. Jain, V. Sharma, and B. Amrutur. Soft real time implementation of a cognitive radio testbed for frequency hopping primary satisfying qos requirements. In *Communications (NCC), 2014 Twentieth National Conference on*, pages 1–6, Feb 2014.

[66] T. Javidi, B. Krishnamachari, Q. Zhao, and M. Liu. Optimality of myopic sensing in multichannel opportunistic access. In *Communications, 2008. ICC '08. IEEE International Conference on*, pages 2107–2112, May 2008.

[67] W. Jouini, D. Ernst, C. Moy, and J. Palicot. Upper confidence bound based decision making strategies and dynamic spectrum access. In *International Conference on Communications, ICC'10*, May 2010.

[68] W. Jouini, C. Moy, and J. Palicot. Decision making for cognitive radio equipment: analysis of the first 10 years of exploration. *EURASIP Journal on Wireless Communications and Networking*, 2012(26), Jan. 2012.

[69] W. Jouini, C. Moy, J. Palicot, et al. Upper confidence bound algorithm for opportunistic spectrum access with sensing errors. *6th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications, Osaka, Japan*, page 17, 2011.

[70] L. P. Kaelbling. *Learning in Embedded Systems.* MIT Press, Cambridge, MA, USA, 1993.

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

[72] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 655–662, New York, NY, USA, 2012. ACM.

[73] S. Kar, H. 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, Dec 2011.

[74] H. Karl. An overview of energy-efficiency techniques for mobile communication systems. *Technical Report, Telecommunication networks Group, Technical University Berlin*, Sept. 2003.

[75] R. M. Karp. *Reducibility among Combinatorial Problems*, chapter Complexity of Computer Computations, pages 85–103. Springer US, Boston, MA, 1972.

[76] E. Kaufmann, O. Cappé, and A. Garivier. On bayesian upper confidence bounds for bandit problems. In N. D. Lawrence and M. A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 592–600, 2012.

[77] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam. Distributed α-optimal user association and cell load balancing in wireless networks. *IEEE/ACM Transactions on Networking*, 20(1):177–190, Feb 2012.

[78] J. Kim, H. W. Lee, and S. Chong. TAES: Traffic-aware energy-saving base station sleeping and clustering in cooperative networks. In *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2015 13th International Symposium on*, pages 259–266, May 2015.

[79] S. Koenig and R. G. Simmons. Complexity analysis of real-time reinforcement learning. In *Proceedings of the 11th National Conference on Artificial Intelligence. Washington, DC, USA, July 11-15, 1993.*, pages 99–107, 1993.

[80] L. Lai, H. Jiang, and H. V. Poor. Medium access in cognitive radio networks: A competitive multi-armed bandit framework. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 98–102, Oct 2008.

[81] L. Lai, Y. Liang, and H. V. Poor. Key agreement over wireless fading channels with an active attacker. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1391–1396, Sept 2010.

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

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

[84] A. Lazaric, M. Restelli, and A. Bonarini. Transfer of samples in batch reinforcement learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 544–551, New York, NY, USA, 2008. ACM.

[85] A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Trans. Inf. Theor.*, 22(1):75–81, Sept. 2006.

[86] A. Leon-Garcia. *Probability, Statistics, and Random Processes for Electrical Engineering*. Pearson/Prentice Hall, third edition, 2008.

[87] K. Levente and S. Csaba. Discounted ucb. In A. P. Danyluk, L. Bottou, and M. L. Littman, editors, *In 2nd PASCAL Challenges Workshop, Venice, Italy*, 2006.

[88] P. Lezaud. Chernoff-type bound for finite markov chains. In *Annals of Applied Probability, Vol. 8 (1998), no. 3, pp. 849–867*, 1998.

[89] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang. TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks. *IEEE Transactions on Wireless Communications*, 13(4):2000–2011, April 2014.

[90] R. Li, Z. Zhao, X. Chen, and H. Zhang. Energy saving through a learning framework in greener cellular radio access networks. In *Global Communications Conference (GLOBECOM), 2012 IEEE*, pages 1556–1561, Dec 2012.

[91] R. Li, Z. Zhao, Y. Wei, X. Zhou, and H. Zhang. Gm-pab: A grid-based energy saving scheme with predicted traffic load guidance for cellular networks. In *2012 IEEE International Conference on Communications (ICC)*, pages 1160–1164, June 2012.

[92] Y.-C. Liang, Y. Zeng, E. C. Peh, and A. T. Hoang. Sensing-throughput tradeoff for cognitive radio networks. *Wireless Communications, IEEE Transactions on*, 7(4):1326–1337, 2008.

[93] H. Liu, K. Liu, and Q. Zhao. Logarithmic weak regret of non-bayesian restless multi-armed bandit. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1968–1971, May 2011.

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

[95] K. Liu and Q. Zhao. A restless bandit formulation of opportunistic access: Indexablity and index policy. In *2008 5th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops*, pages 1–5, June 2008.

[96] K. Liu and Q. Zhao. Cooperative game in dynamic spectrum access with unknown model and imperfect sensing. *IEEE Transactions on Wireless Communications*, 11(4):1596–1604, April 2012.

[97] K. Liu, Q. Zhao, and B. Krishnamachari. Decentralized multi-armed bandit with imperfect observations. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1669–1674, Sept 2010.

[98] K. Liu, Q. Zhao, and B. Krishnamachari. Distributed learning under imperfect sensing in cognitive radio networks. In *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers*, pages 671–675, Nov 2010.

[99] J. Lunden, V. Koivunen, and H. V. Poor. Spectrum exploration and exploitation for cognitive radio: Recent advances. *IEEE Signal Processing Magazine*, 32(3):123–140, May 2015.

[100] I. Macaluso, D. Finn, B. Ozgul, and L. A. DaSilva. Complexity of spectrum activity and benefits of reinforcement learning for dynamic channel selection. *IEEE Journal on Selected Areas in Communications*, 31(11):2237–2248, November 2013.

[101] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo. Optimal energy savings in cellular access networks. In *2009 IEEE International Conference on Communications Workshops*, pages 1–5, June 2009.

[102] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo. Optimal energy savings in cellular access networks. In *2009 IEEE International Conference on Communications Workshops*, pages 1–5, June 2009.

[103] M. A. Marsan and M. Meo. Energy efficient management of two cellular access networks. *SIGMETRICS Perform. Eval. Rev.*, 37(4):69–73, Mar. 2010.

[104] L. Melián-Gutiérrez, N. Modi, C. Moy, I. Pérez-Alvarez, F. Bader, and S. Zazo. DSA with Reinforcement Learning in the HF band. In *URSI Atlantic Radio Science Conference, AT-RASC 2015*, page 2 pages, Gran Canaria, Spain, May 2015.

[105] L. Melián-Gutiérrez, N. Modi, C. Moy, I. Pérez-Alvarez, F. Bader, and S. Zazo. Hybrid ucb-hmm: A machine learning strategy for cognitive radio in hf band. *IEEE Transactions on Cognitive Communications and Networking*, 1(3):347–358, Sept 2015.

[106] L. Melián-Gutiérrez, N. Modi, C. Moy, I. Pérez-Alvarez, F. Bader, and S. Zazo. Upper confidence bound learning approach for real HF measurements. In *IEEE ICC 2015 - Workshop on Advances in Software Defined and Context Aware Cognitive Networks 2015 (IEEE SCAN-2015) (ICC'15 - Workshops 10)*, pages 387–392, London, United Kingdom, June 2015.

[107] L. Melián-Gutiérrez, S. Zazo, J. Blanco-Murillo, I. Pérez-Alvarez, A. García-Rodríguez, and B. Pérez-Díaz. HF spectrum activity prediction model based on HMM for cognitive radio applications. *Physical Communication*, 9:199 – 211, 2013.

[108] J. Mitola. *Cognitive Radio Architecture: The Engineering Foundations of Radio XML:*. John Wiley & Sons, Inc., 2006.

[109] J. Mitola and G. Q. Maguire. Cognitive radio: making software radios more personal. *IEEE Personal Communications*, 6(4):13–18, Aug 1999.

[110] N. Modi, P. Mary, and C. Moy. QoS driven channel selection algorithm for opportunistic spectrum access. In *IEEE Globecom 2015 Workshop on Advances in Software Defined Radio Access Networks and Context-aware Cognitive Networks (IEEE SDRANCAN 2015)*, San Diego, USA, Dec. 2015.

[111] N. Modi, P. Mary, and C. Moy. Efficient learning in stationary and non-stationary osa scenario with qos guaranty. *EAI Endorsed Transactions on Wireless Spectrum*, 17(11), 1 2017.

[112] N. Modi, P. Mary, and C. Moy. Qos driven channel selection algorithm for cognitive radio network: Multi-user multi-armed bandit approach. *IEEE Transactions on Cognitive Communications and Networking*, 2017.

[113] N. Modi, C. Moy, P. Mary, and J. Palicot. *A New Evaluation Criteria for Learning Capability in OSA Context*, pages 3–14. Springer International Publishing, Cham, 2016.

[114] N. Morozs, T. Clarke, D. Grace, and Q. Zhao. Distributed q-learning based dynamic spectrum management in cognitive cellular systems: Choosing the right learning rate. In *2014 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6, June 2014.

[115] R. Munos. From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–130, 2014.

[116] C. M. Navikkumar Modi, Philippe Mary. Trqos-ucb: A transfer restless qos-ucb policy for energy efficient heterogeneous cellular network. *to be submitted to IET Special Issue on Advances in Enabling Technologies for Green Communications and Networking*, 2017.

[117] N. Nayyar, Y. Gai, and B. Krishnamachari. On a restless multi-armed bandit problem with non-identical arms. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 369–376, Sept 2011.

[118] Z. Niu. TANGO: traffic-aware network planning and green operation. *IEEE Wireless Communications*, 18(5):25–29, October 2011.

[119] Z. Niu, Y. Wu, J. Gong, and Z. Yang. Cell zooming for cost-efficient green cellular networks. *IEEE Communications Magazine*, 48(11):74–79, November 2010.

[120] M. NoroozOliaee, B. Hamdaoui, and K. Tumer. Efficient objective functions for coordinated learning in large-scale distributed osa systems. *IEEE Transactions on Mobile Computing*, 12(5):931–944, May 2013.

[121] E. Oh and B. Krishnamachari. Energy savings through dynamic base station switching in cellular wireless access networks. In *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, pages 1–5, Dec 2010.

[122] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu. Toward dynamic energy-efficient operation of cellular network infrastructure. *IEEE Communications Magazine*, 49(6):56–61, June 2011.

[123] E. Oh, K. Son, and B. Krishnamachari. Dynamic base station switching-on/off strategies for green cellular networks. *IEEE Transactions on Wireless Communications*, 12(5):2126–2136, May 2013.

[124] J. Oksanen and V. Koivunen. An order optimal policy for exploiting idle spectrum in cognitive radio networks. *IEEE Transactions on Signal Processing*, 63(5):1214–1227, March 2015.

[125] J. Oksanen, V. Koivunen, and H. V. Poor. A sensing policy based on confidence bounds and a restless multi-armed bandit model. *CoRR*, abs/1211.4384, 2012.

[126] R. Ortner, D. Ryabko, P. Auer, and R. Munos. Regret bounds for restless markov bandits. *To appear in Theoretical Computer Science*, 2014.

[127] J. Palicot. Radio engineering: From software radio to cognitive radio, 2013.

[128] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[129] L. Pat. Transfer of knowledge in cognitive systems. In *Workshop on Structural Knowledge Transfer for Machine Learning at the Twenty-Third International Conference on Machine Learning*, 2006.

[130] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li. Traffic-driven power saving in operational 3G cellular networks. In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, MobiCom '11, pages 121–132, New York, NY, USA, 2011. ACM.

[131] M. H. Rehmani, A. C. Viana, H. Khalife, and S. Fdida. Activity pattern impact of primary radio nodes on channel selection strategies. In *Proceedings of the 4th International Conference on Cognitive Radio and Advanced Spectrum Management*, CogART '11, pages 36:1–36:5, New York, NY, USA, 2011. ACM.

[132] C. Robert, C. Moy, and C.-X. Wang. Reinforcement learning approaches and evaluation criteria for opportunistic spectrum access. In *Communications (ICC), 2014 IEEE International Conference on*, pages 1508–1513, June 2014.

[133] C. Robert, C. Moy, and H. Zhang. Opportunistic Spectrum Access Learning Proof of Concept. In *SDR-WinnComm'14*, page 8 pages, Schaumburg, United States, Mar. 2014.

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

[135] D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *Computing Research Repositor*, 2014.

[136] D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *J. Mach. Learn. Res.*, 17(1):2442–2471, Jan. 2016.

[137] A. Slivkins and E. Upfal. Adapting to a changing environment: the brownian restless bandits. In *21st Conference on Learning Theory (COLT)*, pages 343–354, July 2008.

[138] A. Slivkins and E. Upfal. Adapting to a Changing Environment: the Brownian Restless Bandits. In R. A. Servedio, T. Zhang, R. A. Servedio, and T. Zhang, editors, *COLT*, pages 343–354. Omnipress, 2008.

[139] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin. Energy efficient heterogeneous cellular networks. *IEEE Journal on Selected Areas in Communications*, 31(5):840–850, May 2013.

[140] K. Son, S. Chong, and G. D. Veciana. Dynamic association for load balancing and interference avoidance in multi-cell networks. *IEEE Transactions on Wireless Communications*, 8(7):3566–3576, July 2009.

[141] K. Son, H. Kim, Y. Yi, and B. Krishnamachari. Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks. *IEEE Journal on Selected Areas in Communications*, 29(8):1525–1536, September 2011.

[142] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.

[143] M. E. Taylor, N. K. Jong, and P. Stone. Transferring instances for model-based reinforcement learning. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases - Part II*, ECML PKDD '08, pages 488–505, Berlin, Heidelberg, 2008. Springer-Verlag.

[144] M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *J. Mach. Learn. Res.*, 10:1633–1685, Dec. 2009.

[145] C. Tekin and M. Liu. Online algorithms for the multi-armed bandit problem with markovian rewards. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1675–1682. IEEE, 2010.

[146] C. Tekin and M. Liu. Online learning in opportunistic spectrum access: A restless bandit approach. In *INFOCOM, 2011 Proceedings IEEE*, pages 2462–2470. IEEE, 2011.

[147] C. Tekin and M. Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, Aug 2012.

[148] R. W. Thomas, D. H. Friend, L. A. Dasilva, and A. B. Mackenzie. Cognitive networks: adaptation and learning to achieve end-to-end performance objectives. *IEEE Communications Magazine*, 44(12):51–57, Dec 2006.

[149] L. Torrey and J. Shavlik. Transfer learning. *Handbook of Research on Machine Learning Applications. IGI Global*, 3:17–35, 2009.

[150] O. van den Biggelaar, J.-M. Dricot, P. De Doncker, and F. Horlin. Sensing time and power allocation for cognitive radios using distributed q-learning. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):1–12, 2012.

[151] J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of the 16th European Conference on Machine Learning*, ECML'05, pages 437–448, Berlin, Heidelberg, 2005. Springer-Verlag.

[152] B. Wang and K. J. R. Liu. Advances in cognitive radio networks: A survey. *IEEE Journal of Selected Topics in Signal Processing*, 5(1):5–23, Feb 2011.

[153] Y. Wang, Y. Xu, L. Shen, C. Xu, and Y. Cheng. Two-dimensional pomdp-based opportunistic spectrum access in time-varying environment with fading channels. *Journal of Communications and Networks*, 16(2):217–226, April 2014.

[154] J. C. H. Watkins. Learning from delayed rewards, 1989.

[155] W.-T. Wong, Y.-J. Yu, and A.-C. Pang. Decentralized energy-efficient base station operation for green cellular networks. In *Global Communications Conference (GLOBECOM), 2012 IEEE*, pages 5194–5200, Dec 2012.

[156] C. Wu, K. Chowdhury, M. Di Felice, and W. Meleis. Spectrum management of cognitive radio using multi-agent reinforcement learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Industry Track*, AAMAS '10, pages 1705–1712, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems.

[157] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y. D. Yao. Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution. *IEEE Transactions on Wireless Communications*, 11(4):1380–1391, April 2012.

[158] K. L. A. Yau, P. Komisarczuk, and D. T. Paul. Enhancing network performance in distributed cognitive radio networks using single-agent and multi-agent reinforcement learning. In *Local Computer Networks (LCN), 2010 IEEE 35th Conference on*, pages 152–159, Oct 2010.

[159] K. L. A. Yau, P. Komisarczuk, and P. D. Teal. Applications of reinforcement learning to cognitive radio networks. In *2010 IEEE International Conference on Communications Workshops*, pages 1–6, May 2010.

[160] G. Yuan, R. C. Grammenos, Y. Yang, and W. Wang. Performance analysis of selective opportunistic spectrum access with traffic prediction. *IEEE Transactions on Vehicular Technology*, 59(4):1949–1959, May 2010.

[161] W. Zhang. Performance of real-time and data traffic in heterogeneous overlay wireless networks. In *Proceedings of the 19th International Teletraffic Congress*, September 2005.

[162] Q. Zhao and A. Swami. A survey of dynamic spectrum access: Signal processing and networking perspectives. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–1349–IV–1352, April 2007.

[163] Y. Zhao, J. H. Reed, S. Mao, and K. K. Bae. Overhead Analysis for Radio Environment Map-enabled Cognitive Radio Networks. In *In Proc. of the First IEEE Workshop on Networking Technologies for Software Defined Radio (SDR) Networks*, 2006.

[164] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang. The predictability of cellular networks traffic. In *Communications and Information Technologies (ISCIT), 2012 International Symposium on*, pages 973–978, Oct 2012.