Docitive Networks – A Step Beyond Cognition –

by
Pol Blasco Moreno

Master thesis advisor
Dr. Mischa Dohler

January 2011

A thesis submitted to the Departament de Teoria del Senyal i Comunicacions of the Universitat Politècnica de Catalunya for the degree Master of Science

Intelligent Energy Area
Centre Tecnològic de Telecomunicacions de Catalunya
Castelldefels, Barcelona

Contents

List of Acronyms

ABS	Access Base Station
BS	Base Station
BuNGee	Beyond Next Generation Mobile Broadband
BW	Band Width
CR	Cognitive Radio
EDGE	Enhanced Data Rates for GSM Evolution
EGT	Evolutionary Game Theory
ESS	Evolutionary Stable Strategy
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GT	Game Theory
HBS	Hub Base Station
HDxPa	High Data Packet Access
LOS	Line of Sight
LTE	Long Term Evolution
LTE-A	LTE-Advance
MAC	Medium Access Control
MAS	Multi Agent System
MDEG	Markov Decision Evolutionary Game
MDP	Markov Decision Process
MIMO	Multiple-Input-Multiple-Output
ML	Machine Learning
MS	Mobile Subscriber
NLOS	non-Line of Sight
OFDM	Orthogonal Frequency Division Multiplexing
OSA	Opportunistic Spectrum Access

PBL	Problem Based Learning
PHY	Physical Layer
PL	Path Loss
POMDP	Partially Observable Markov Decision Process
QoS	Quality of Service
RD	Replicator Dynamics
RF	Radio Frequency
RL	Reinforcement Learning
RRM	Radio Resource Management
RS	Relay Station
SINR	Signal-to-Interference and Noise Ratio
SNR	Signal-to-Noise Ratio
SON	Self-organized Networks
SU	Secondary User
TDMA	Time Division Multiple Access
UMTS	Universal Mobile Telecommunications System
W-LAN	Wireless LAN

Chapter 1

Introduction

The contents of this thesis deals with an entirely novel concept, i.e. docitive networks. Docition, from the Latin word "docere" meaning "to teach" and pronounced [dozishen], is a communication paradigm which encourages base stations (BSs) and/or mobile stations (MSs) to teach each other of their respective prior actions. It is shown to yield the much needed benefits to make future wireless communication systems succeed. The approach goes well beyond any current best practices which use adaptive algorithms as well as any forward looking approaches which typically advocate cognitive approaches. The docitive communication paradigm is thus going beyond any forward looking state-of-the-art.

This introductory chapter gives the technical and strategic rational for chosen approach and thus paves the way for subsequent technical chapters. Notably, Section 1.1 will briefly overview the evolution of cellular systems, whilst Section 1.2 highlights the main drivers for today's designs. This allows us to understand the problems of current cellular systems, as outlined in Section 1.3. This, in turn, allows us to propose a set of solutions given in Section 1.4. Finally, the chapter is summarized and linked to subsequent developments in this thesis in Section 1.5.

1.1 Overview of Cellular Systems

Cellular systems have probably been one of the major technological breakthroughs in recent years which − along with the Internet − has allowed ubiquitous voice and data coverage. We briefly review the notion of cellular systems for completeness, before dealing in more details with 3GPP and WiMAX systems

1.1.1 Quick Introduction

The notion of cellular networks is composed of two concepts:

- **Cellular.** The concept of using cells is exemplified in Figure 1.1. The cellular rollout facilitates location independent communications across a wide geographic coverage

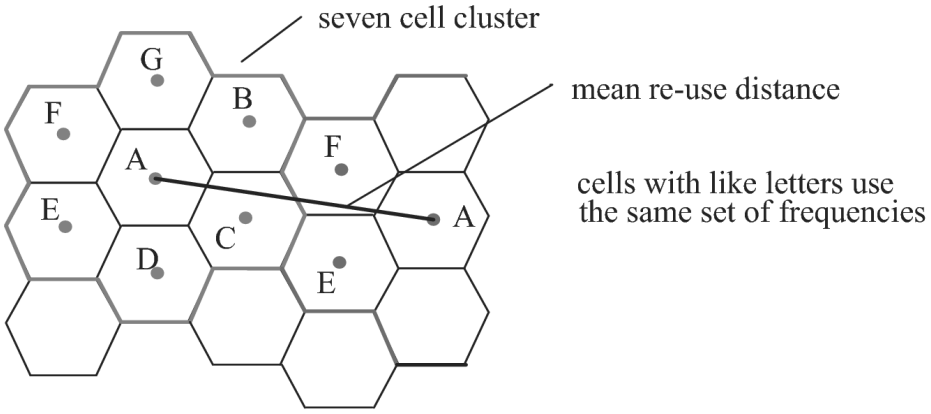


Figure 1.1: Exemplifying the concept of using cells to facilitate universal geographical coverage.

(range in order of km). The coverage area is divided into cells, allowing for lower transmission powers, better coverage and higher capacity.

- **Network.** As exemplified in Figure 1.2, the principle element is the access network allowing the MS and BS to communicate with each other. A further important element is the backhaul network which bridges the access network with the proprietary core network. The latter then typically feeds into other proprietary networks or the Internet at large.

Given these general elements, various industrial initiatives have defined and deployed viable cellular systems, the most important ones being 3GPP and WiMAX. Their timeline, importance and interactions are briefly reviewed below.

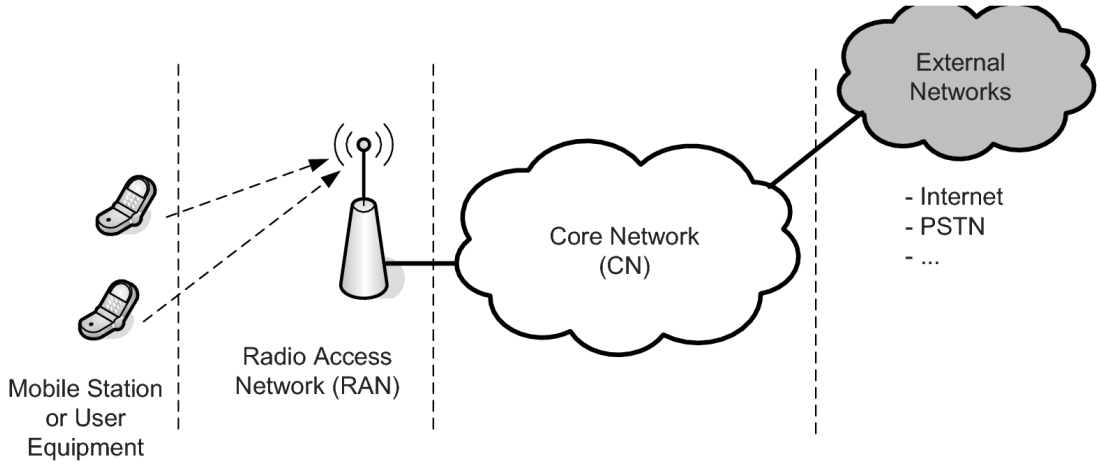


Figure 1.2: Exemplifying the most important networking elements, i.e. access, backhaul and core networks.

1.1.2 Cellular Letter Salad

The first digital system of importance has been GSM which has evolved ever since:

- **2G Networks.** GSM (Global System for Mobile Communications) was initiated in Europe in 1990 and is today a system used truly worldwide. IS95 (Interim Standard 95) has been initiated in the US and is as a standard currently discontinued. Both systems are mainly voice systems, although data transport is possible at extremely low rate.

- **2.5G Networks.** The main constituent is GPRS (General Packet Radio System), an evolution of GSM and thus also a worldwide system in use today. The main focus of this system is data transmission.

- **3G Networks.** The 3G evolution from GSM is EDGE (Enhanced Data Rates for GSM Evolution). The 3GPP contribution however commences truly here with UMTS (Universal Mobile Telecommunication System). A US driven approach is CDMA2000 (based on 2G CDMA Technology) and was standardized by 3GPP2, which has been discontinued in 2008. WiMAX, driven by IEEE 802.16 technology, has also commenced to appear at this point and has been labeled by ITU as a 3G technology.

- **3.5G Networks.** These are mainly 3GPP evolutions having lead to HDxPA (High Data Packet Access) networks which are high data rate systems.

- **3.9G Networks.** These are LTE (Long Term Evolution) systems and are mainly a UMTS evolution/revolution to be used worldwide. Note that both LTE and WiMAX are regarded as beyond 3G (B3G) systems but are strictly speaking not 4G since not fulfilling the requirements set out by the ITU for 4G next generation mobile networks (NGMN). NGMN requires downlink rates of 100 Mbps for mobile and 1 Gbps for fixed-nomadic users at bandwidths of around 100 MHz which is the prime design target of LTE Advanced and WiMAX II. Therefore, even though LTE is (somehow wrongly but understandably) marketed as 4G, it is not and we still need to wait for LTE-A.

- **4G Networks.** LTE-A (LTE Advanced) is an LTE evolution/revolution to be used worldwide. WiMAX hopes to get a 4G label from ITU through its WiMAX II system which is based on IEEE 802.16j/m high capacity designs.

The timeline until today has been summarized in Figure 1.3.

1.1.3 3GPP versus WiMAX

Both initiatives have evolved in parallel with initially different goals:

- **3GPP.** It is a consortium of companies designing cellular systems which are then typically officially standardized by the ITU. The 3GPP consortium is more driven by companies which pay a lot of attention to and thus have a great experience in designs ensuring low power consumption at the MS side, mobility support at the BS and billing capabilities at the architectural side. In addition, 3GPP typically works on a complete architecture, involving all OSI layers and all networking elements.

- **WiMAX.** WiMAX is IEEE driven which typically only standardizes PHY and MAC layers. The IEEE is support by industries which are good in designs yielding high data rates, however, for rather static short-range systems with private owners. Therefore, IEEE systems typically do not support mobility and are not power efficient. This however is changing lately, notably with the IEEE 802.16 and other standards. It is thus expected that future 3GPP and IEEE design will resemble at least in the PHY and MAC layers.

Due to the likely resemblance of future releases from both initiatives, the developments of this thesis − incidentally performed for WiMAX type systems − are equally applicable to 3GPP type systems.

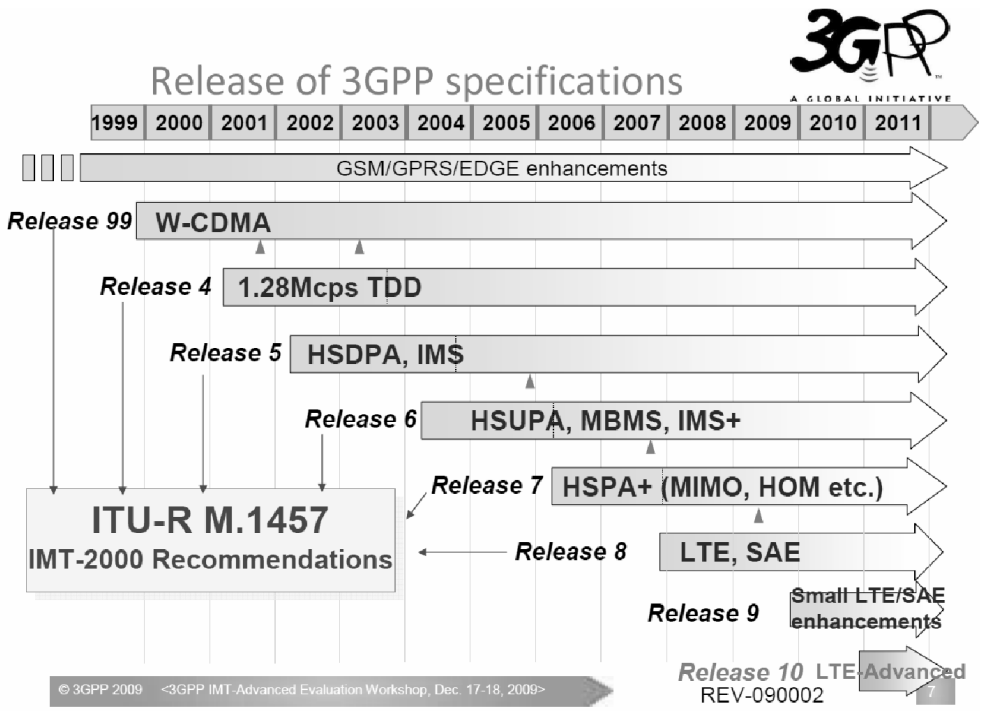


Figure 1.3: The 3GPP timeline of the various standards document releases.

1.2 Main Cellular Design Drivers

Above mentioned systems currently evolve where the main design drivers are capacity and cost.

1.2.1 General Requirements

Beyond next generation mobile broadband promises ubiquitous coverage, access for any fixed and mobile device, support of bandwidth-intensive applications such as video/audio streaming or mobile TV, among others, anywhere in the cell. Subscribers expect networks to combine WLAN's data rates and cellular's extensive coverage and mobility support. This triggered the European Commission, endorsed by the European Council in December 2008 to invest 1 billion Euros in upcoming years on Internet broadband infrastructure [3]. These developments have been mirrored in the UK [2] and Germany [31] and show that broadband is seen as the core of Europe's business and technology pulse.

Market requirements for mobile broadband operators include the following:

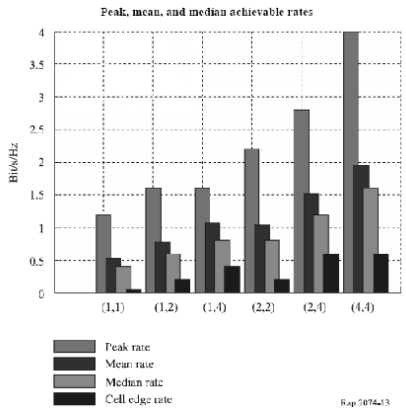
- **High Traffic Levels from Mobile Subscribers.** Subscribers will expect to pay a flat monthly fee and be able to access the network as they do with their current work or home broadband connection. In some cases, the future network will in fact replace this connection. Even nomadic and mobile users will often be heavier users than current cellular data users, as they will have data-centric and consumer electronic devices (e.g. multimedia players or game consoles) that will typically run applications that generate heavy network traffic. Furthermore, these subscribers may even use the beyond next generation network at home or in the office where they have an alternative wireline broadband connection, simply because it is more convenient.

- **Bandwidth-Intensive Applications.** Increasingly, mobile subscribers will use their devices for a wide range of broadband applications that for the first time will be accessible truly everywhere. Some of these (video streaming or music downloads) will require operators to increase their network capacity. Uplink demands will grow as well, as subscribers generate the content (e.g. photographs, movies) they upload to their favorite websites.

From the above it becomes clear that ubiquitous coverage and high throughput are central to the expectations of potential beyond next generation networking users across markets worldwide.

1.2.2 Capacity

Exemplifying the needs, a typical beyond next generation mobile networking BT in dense urban area has a cell radius of 500-600 meters and covers an area of approximately 1km2. The total aggregated capacity of a single BT having 40MHz spectrum in a full cellular deployment presently (and in the near future) does not exceed 100Mbps. Therefore,



City	Average People/Km^2
Athens	5,400
Madrid	5,200
London	5,100
Barcelona	4,850
Warsaw	4,300
Naples	4,100
Berlin	3,750
Paris	3,550
Vienna	3,400

Figure 1.4: Peak, mean and median achievable rates for different IMT MIMO systems (left). Typical population densities in European cities (right). [1]

the current maximum capacity density for present day networks is not able to exceed 100Mbps/Km2. Other techniques might be available to boost capacity further; however, none of them is known to be a viable and cost efficient solution [1].

This dire capacity situation is essentially reflected by the minimum requirements for compliance with IMT-Advanced. The ITU-R target for compliance with IMT-Advanced, as described in ITU-R M.2133 "Requirements, evaluation criteria and submission templates for the development of IMTAdvanced", is 2.2b/s/Hz for downlink and 1.4b/s/Hz for uplink for BTS coverage in urban deployment. Figure 1.4 (left), provided in ITU-R Report M.2074 "Radio aspects for the terrestrial component of IMT-2000 and systems beyond IMT-2000", shows the target spectral efficiencies for IMT-Advanced systems. There, different MIMO configurations are used which give the same order of magnitude in performance of the mean spectral efficiency.

The current capacity density may be adequate in less densely populated areas but is clearly insufficient in urban areas where the number of users per square kilometer is expected to be much higher. In the 10 most densely populated cities in the world the population density ranges from 13,400 people/Km2 (Shanghai, China) to 29,650 people/Km2 (Mumbai, India). In Europe, however, the urban population density is much lower as exemplified in Figure 1.4(right). Importantly, the numbers in the table are averages. Clearly, in the city commercial areas the population density is much higher during business hours, and can easily reach 8,000 people/Km2. These are also the areas and the times where the demand for broadband wireless access will be the highest. As for the capacity requirements, a simple calculation shows that the capacity density of current mobile networks is far from answering the market demand in dense urban areas during business hours.

Assume a density of 8,000 people/Km2, of which only 10% subscribe to the broadband access service. Of these subscribers, assume only 20% require access at the same time (in peak traffic hours). We can also safely assume that in several years from now every subscriber will expect a transmission rate of 5Mbps. Therefore, the required capacity

density in the near future should be:

$$\text{Capacity} = 8,000 \times 10\% \times 20\% \times 5Mbps = 800Mbps/Km^2$$

This required capacity density is **an order of magnitude higher than the forward looking current state of the art**[1]. It is essentially the need and inspiration of the developments in this thesis.

1.2.3 Cost

Whilst not focus of this thesis, cost plays an integral role in the success of the uptake of these high capacity technologies. It is therefore utmost important to ensure designs which are cost effective in that they maximize the number of bits transmitted per Euro spent in developing, deploying and maintaining the infrastructure.

1.3 Problems of Modern Cellular Systems

The first digital system, GSM, and its capabilities has evolved significantly, and so have the problems, in particularly in the light of above design drivers. We briefly summarize the most important ones, which essentially justify developments into docitive systems.

1.3.1 Interference

The required high capacity densities can only be supported by an increasingly wide spectrum and dense network of BSs. This is a trend already well observed in LTE and LTE-A. With the decrease of distance between BSs as well as MSs, the amount of interference generated by and into the communicating links increases. This poses serious problems on the design as the interference needs to be catered for to ensure availability and promised capacities.

1.3.2 Radio Resource Management (RRM)

Another problem is the significant increase of complexity of resource scheduling algorithms, mainly due to the wider bandwidth and more users available. Resources are typically scheduled in time, frequency and power. To establish stable RRM algorithms which fulfill the capacity and QoS requirements of the users is one of the hardest issues in RRM design.

1.3.3 Self-Organizing Networking (SON) Capabilities

Finally, the sheer complexity of the system in terms of the large amount of degrees of freedom, dynamics, interactions, failures, etc, etc, require the ability of the system to self-organize, self-configure, self-heal, etc. Manual configuration is clearly not possible, whilst automated solutions are very complex as of today. SON remains one of the unsolved core problems in telecommunications of the 21st century.

1.4 Solutions to Design Problems

The community is well aware of above problems and is thus striving for a viable solution. One such option is to enhance the existing deployment paradigm of rooftop located network base stations by significantly increasing the capacity of the base stations. An alternative approach is that of making the deployment grid above rooftops denser; it is problematic as access to rooftops is costly and becoming challenging due to residents' objections (radiation related concerns). We will henceforth discuss a possible set of solutions, not claiming to be the only way forward.

1.4.1 Design Highlights

From an architectural point of view, the following unprecedented approach is proposed which is currently being examined in BuNGee [46] as well as ETSI BRAN [45]:

- to have a much denser base station grid below the rooftops (e.g., on utility poles) and thereby bringing the backhaul network below rooftop;
- to exercise aggressive reuse combined with high spectrum efficiency, by using novel antenna, RF (radio frequency), base-band and network techniques;
- to undertake a joint design of backhaul and access networks, using heterogeneous radio elements, licensed and licensed-exempt spectrum, a cognitive radio approach, among others, aimed at achieving a maximum system capacity and QoS (quality of service);
- to design a data and control plane protocol suite that facilitates autonomous operation by means of a complete self-organizing networking paradigm.

Having a denser base station grid coupled with **aggressive reuse of resources** allows us to significantly decrease the transmission powers and thereby the electromagnetic exposure in urban environments. Initial studies have shown that this would allow us to operate at power levels ten times lower than to best practise today [1]. This would guarantee an evenly distributed and much lower exposure level throughout the urban environment.

The corner-stone of the architecture, however, is the tightly coupled **joint design of access and backhaul** network which is facilitated and driven by the fact that **both use the same bands** and are becoming spatially very close. Whilst WiMAX was initially conceived as a point-to-multipoint system for fixed access, including backhauling, the IEEE 802.16e release has allowed it to be used as a high-capacity BWA access network too - a joint design is therefore outstandingly promising and the inspiration for said approach.

In the context of 3GPP, such a joint design paradigm is also vital if some of the spectral chunks are to lie within license-exempt bands which are used by a backhauling system. We are certain that only such an approach allows beyond next generation capacity density needs to be met at a reasonable cost.

1.4.2 Architectural Solutions

The novel heterogeneous architecture, combined with the deployment approach and integrated usage of licensed and un-licensed spectrum, allows a significant increase of available capacity to all the users in any point of the deployment [35] . Referring to Figure 1.5, and currently specified in BuNGee, the architecture includes:

- Hub BS (HBS), connected to the operator back-haul;
- Access BS (ABS), connected to the HBS by a self-backhaul link on one side and to the mobile user terminals on the other side;
- Relay stations (RS) to be used for extending coverage in some (rare) situations;
- Femto-cells, using preferably the un-licensed spectrum but at the same radio technology as used in licensed bands. The Femto BS, especially in below 1GHz license-exempt spectrum, have a larger coverage in up-link;
- MS, connected to HBS, ABS, Relay and FBS. An MS can receive traffic from multiple BSs (MS1 and MS2) or can be used as an up-link repeater (MS4 is using MS2) or can participate in the up-link cooperative MIMO, for which at a 2nd stage both MS2 and MS4 transmit to ABS2 in the same time.

The user traffic has different QoS requirements: we exploit the availability of the un-licensed spectrum in both self-backhauling and the access part of the network to serve users with traffic having less stringent QoS requirements. Such traffic can be provided by the Hub BS, Access BS and Femto BS via un-licensed spectrum [35]. The more demanding traffic will be provided over the licensed spectrum.

1.4.3 Algorithmic Solutions

From an algorithmic point of view, to facilitate the goal of a **capacity density of 1 Gbps/km^2** to be achieved, it is proposed to utilize the same spectral bands for the access as well as backhaul links. Here, the backhaul links are formed between a HBS and several below-rooftop ABSs which in turn serve associated MSs. WiMAX naturally lends itself to such an design for the following reasons:

1. In its original standards formulation, i.e., IEEE 802.16-2004 or IEEE 802.16d, WiMAX has been designed for high-capacity wireless links.
2. In a subsequent standards edition, i.e., IEEE 802.16e-2005 or simply IEEE 802.16e, mobility support has been introduced thus allowing for true mobile cellular access provision.
3. In a recent amendment of the standard, i.e., IEEE 802.16j-2009, multihop relaying functionalities are facilitated which are fundamentally required when relaying data traffic from the HBS to the MS via the ABS or a set of ABSs.

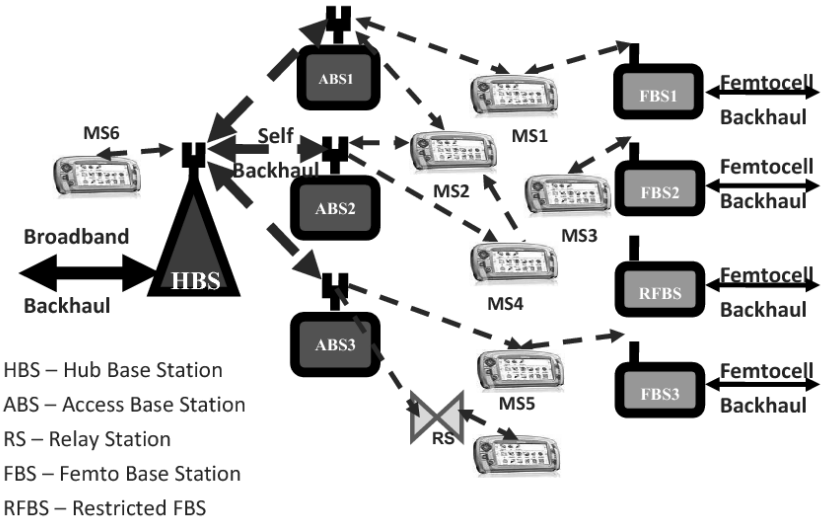


Figure 1.5: High capacity system architecture, forming the basis of BuNGee and ETSI developments.

However, whilst the building blocks are available, some serious challenges remain to be addressed and thus constitute the prime focus of our design:

1. The multihop option of IEEE 802.16j-2009 allows only time-division relaying, i.e., the HBS first needs to communicate to the ABSs and only then can the ABSs communicate to the MSs. The spectral efficiency is thus roughly halved. A more aggressive approach would be to allow **both backhaul as well as access links to communicate simultaneously**. This, however, constitutes a serious challenge in interference management, i.e., interference avoidance, mitigation and suppression.

2. The complexity of the complete system at hand is very large. Notably, the system to be optimized will be composed of at least one HBS, several decentralized ABSs and a fairly large amount of MSs. In addition, the optimization scope will include the operation over license and license-exempt bands, presenting different interference conditions. If the optimization problem can be formalized, it is likely to be NP-hard and/or non-convex and thus a solution eludes the majority of tools available to date dealing with system optimization.

3. The system as a whole is **highly dynamic**, likely to yield non-stationary effects in both observation as well as actions to be taken by the involved parties. This means that the system should be **sufficiently adaptive and self-organizing** in the sense that changes in the operational conditions should be handled well by the system. Another implication is that most theoretical toolboxes break down and more **computerized solving methods**, such as machine learning, have to be invoked to yield viable results.

4. The memory and coherence of the system is not negligible in that a specific action taken by the system (such as instructing a specific ABS to transmit at a given power level) is correlated with the action taken at a subsequent time instant as well as with the action taken under similarly occurring circumstances. This implies that using fixed or simple (memory-less) opportunistic strategies are inherently sub-optimal. We thus concentrate on truly cognitive approaches which capitalize on the peculiarities of the system under consideration. The prime problem with cognitive approaches, however, is the poor convergence in time and to the set target.

In summary, we assume that a) the system uses the same time and frequency band for the wireless backhaul and access links; b) decisions are taken in a distributed fashion; and c) the system obeys as much as possible the WiMAX specs. The aim is to formulate the problem in such a way that a) a solution in form of decisions can be found, even if only iteratively and numerically; b) these decision yield clear instructions on radio resource management functionalities to all involved parties; and c) these decisions are based on truly intelligent algorithms with elements of memory, learning, teaching and intelligent decision taking.

1.5 Summary and Organization of the Master Thesis

In summary, we have established that modern cellular systems offer many opportunities but also suffer from serious short comings. Notably, driven by unprecedented capacity requirements and pressure on cost, the problem of interference, resource management and the system's self organization is posing serious challenges on the viability of any solution. The proposed high capacity architectural solution, stipulated by BuNGee and ETSI BRAN, allows overcoming these shortcomings. The required algorithms, however, need to be intelligent as otherwise the high degree of complexity cannot be managed. This, in essence, is the rational for using truly cognitive and docitive algorithms, as outlined in subsequent chapters.

The rest of this document is organized as follows. Chapter 2 introduces the concept of Cognitive Radio (CR) and outlines several cognitive algorithms. The absence of cognitive algorithms positively able to deal with the high degree of complexity of the architectural solution proposed motivates Chapter 3. In Chapter 3 the concept of docition is presented as a clear candidate to outperform the cognitive algorithms. In Chapter 4 we discuss a set of simulations on pertinent cognitive and docitive algorithms. Finally Chapter 5 concludes the thesis and paves the way for future work.

Chapter 2

Cognitive Networks – The Intelligent Learning Paradigm

2.1 Introduction to Cognitive Networks

Cognitive Radio (CR) has been in the spotlight of the research efforts carried out by the wireless research community during the past decade. The term *Cognitive Radio* was first coined by J. Mitola [28] in 1999. He claimed that CR could outperform fixed etiquette communications technologies (like GSM, by that year the leading mobile wireless technology) in terms of radio resource efficient usage. The CR has knowledge of its own internal structure and of its environment (i.e. the radio spectra), with this knowledge at a hand is able to draw intelligent decisions adapting to the statistical variations of the environment. CR ought to learn environment dynamics by taking into account past and present values while sensing the spectra and learning from their own policies; CR should make intelligent decisions, using long term benefit estimations. A high level cycle of CR is depicted in Figure 2.1. A software-defined radio, such as CR, may be able to tune a large number of communication parameters i.e modulation scheme, bit rate and many more (e.g. in [18] a radio receiver is design to adapt most of its configuration parameters). However a consid-

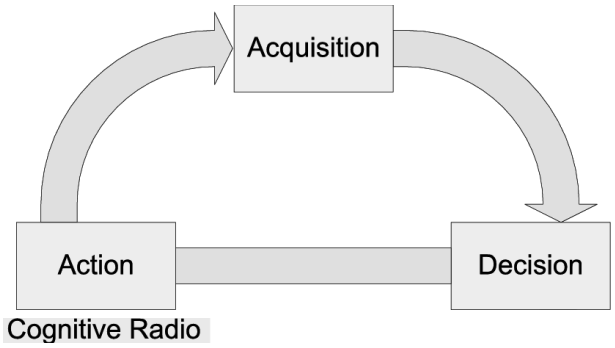


Figure 2.1: Cognitive cycle

erable part of the state-of-the-art cognitive technique literature is focus on transmit-power control and spectrum management.

Radio resource management (RRM) usually deals with a general scenario with the presence of a primary user and several secondary users (SU). One of the main challenges relays on the spectrum sharing techniques/policies that the SU may use in order to maximize the spectrum utilization. Nowadays most of the networks use fixed spectrum access leading to a misuse of large frequency bands, dynamics spectrum sharing techniques are viewed as a novel approach to increase the spectrum utilization. Several authors have purposed to use Opportunistic spectrum access (OSA) policies in which nodes are allowed to access a particular frequency band as long as their transmissions do not cause harmful interference to the primary (protected) users of that band. Whilst OSA techniques take myopic decisions based only on present information and fixed rules (i.e. threshold comparison) cognitive approaches learn from the past actions to forecast action's benefits. Cognitive approaches advantage OSA avoiding non intelligent decisions that bring to sub-optimal solutions.

Example 1 *Two users are receiving at the very same frequency band, both users are using the same OSA algorithm to guarantee that its SINR is above a certain threshold th. A plausible OSA technique is to increase the transmit power if the SINR is below th and in the contrary do nothing. Imagine the situation where user* 1 *has a large interference and decides to increase the transmit power. Immediately user* 1 *will measure that is SINR has increased, however sooner user* 2 *will notice that its SINR has decreased below th and immediately user* 2 *will increase its transmit power. Promptly the SINR of user* 1 *will decrease, eventually after a certain number of iterations both users will be transmitting at maximum power.*

A general scenario in CR and RRM consist of multiple intelligent autonomous agents, sharing a common resource (usually the radio spectra), the state of the system depends on the independent decisions made by multiple agents and on random fluctuations; there is no central entity in charge of providing global control; the individual decisions of each agents have to be self-adaptive depending on the decisions made by the other agents. As a result, the common objective of the multiple agents is to distributively learn an optimal strategy to optimize a certain measure of performance (i.e. SINR or a spectrum access policy).

A number of authors have proposed Markov decision process (MDP) as a viable model for representing and finding the optimal access policy of RRM. MDPs provide a mathematical framework for modeling decision-making in situations where outcomes are partly random and partly under the control of a decision maker. More precisely, a MDP is a discrete time stochastic control process. Consider an agent in the mentioned scenario, choosing one of a finite collection of actions $\mathbf{a}$ (to transmit or not to transmit, change transmission power) every certain time Δ_t. At time t, the agent is able to measure the state s_t of the scenario (i.e. spectrum occupied, SINR level), and can decide for its action a_t accordingly. Then the SU receives a cost whose mean depends on the state s_t and the action a_t), finally the state s of the scenario changes to $s_{t+\Delta_t}$ partially influenced by the agent action a_t. When the probability of transition from one state s_t to the next state

$s_{t+\Delta_t}$ does not depend on past states $s_{t-k\cdot\Delta_t}$ the system is said to have the Markov property and hence is a MDP. The task facing the SU is that of determining an optimal policy, one that maximizes/minimizes total discounted expected cost associated with a measure of performance. One way to find an optimal policy π^* for a MDP is using machine learning (ML) techniques.

2.2 Game Theory – Knowing the Rules

2.2.1 Core Idea

Game Theory (GT) attempts to mathematically capture behavior in strategic situations, or games, in which an individual's success in making choices depends on the choices of others. While initially developed to analyze competitions in which one individual does better at another's expense (zero sum games), it has been expanded to treat a wide class of interactions.

In general GT focuses on the *interaction of players*: players care whom are they playing with, players are assumed to be rational playing, according to **a priori** known rules. GT, thus is, in some way, based on the fact that the rules that drive the players are a priori known. Traditional GT is concern with the statement of the problem and with proof of existence of an *equilibrium*. However, few attention has been paid to the utility of the equilibrium and the way to reach it. Traditional applications of GT attempt to find equilibria in these games. In an equilibrium, each player of the game has adopted a strategy that they are unlikely to change. Many equilibrium concepts have been developed (most famously the Nash equilibrium) in an attempt to capture this idea. These equilibrium concepts are motivated differently depending on the field of application, although they often overlap or coincide. This methodology is not without criticism, and debates continue over the appropriateness of particular equilibrium concepts. In many situations the game equilibria lead the players to play strategies which are not intelligent, however the players get trapped in those strategies since any change in their strategies may incur an immediate decrease in performance (see Example 1).

2.2.2 Taxonomy

There is a vast array of game types, in the following we present a very brief taxonomy centralized in games, that we believe most, related to CR.

- **Relation between players:**
  - A game is **cooperative** if the players are able to form binding commitments.
  - In **noncooperative** games this is not possible. Often it is assumed that communication among players is allowed in cooperative games, but not in noncooperative ones. Of the two types of games, noncooperative games are able to

model situations to the finest details, producing accurate results. Cooperative games focus on the game at large. Considerable efforts have been made to link the two approaches.

  - **Hybrid games** contain cooperative and non-cooperative elements. For instance, coalitions of players are formed in a cooperative game, but these play in a non-cooperative fashion.

- **Relation between players benefits**

  - A **symmetric game** is a game where the payoffs for playing a particular strategy depend only on the other strategies employed, not on who is playing them. If the identities of the players can be changed without changing the payoff to the strategies, then a game is symmetric which in general is not the case in CR.
  - In **zero-sum games** choices by players can neither increase nor decrease the available resources. In zero-sum games the total benefit to all players in the game, for every combination of strategies, always adds to zero (more informally, a player benefits only at the equal expense of others).

- **Games with memory: Dynamic games** are of special interest since the players are assumed to have memory, in there later players have some knowledge about earlier actions. This need not be perfect information about every action of earlier players; it might be very little knowledge. An important subset of dynamic games consists of games of perfect information. A game is one of perfect information if all players know the moves previously made by all other players. Thus, only sequential games can be games of perfect information, since in simultaneous games not every player knows the actions of the others. Perfect information is often confused with complete information, which is a similar concept. Complete information requires that every player know the strategies and payoffs of the other players but not necessarily the actions.

  - A game of special interest due to its relation with the Markov Decision Processes is the **stochastic games**, it belongs to the family of dynamic games, has probabilistic transitions and is played by one or more players.
  - **Evolutionary games**, belonging tho both non-cooperative games and dynamic games, imagines that the games is played *repeatedly* by individuals who are *randomly* drawn from *large populations*, on the contrary to the majority of games where it is assumed that the game is played by a single player which is fully rational and is aware of the other player preferences over game outcomes.

- **Continues and discrete spaces:** Much of game theory is concerned with finite, **discrete games**, that have a finite number of players, moves, events, outcomes, etc. Many concepts can be extended, however. **Continuous games** allow players to choose a strategy from a continuous strategy set.

2.2.3 Problems

GT usually assumes hard constraints in the description of the game, which limits the applicability of the game to very specific scenarios. Most common constraints are assumptions on the player knowledge, the nature of the utility incurred by each action played or the number of interacting players. Unfortunately the wireless scenario is very complex, some of the effects that a game has to take in to account to properly model a CR problem are: simultaneous interaction of many players (probably more than two), the randomness of the utility (i.e. due to the random fluctuations in the communications channel) and the limited knowledge of the players in decentralized scenarios.

2.3 Machine Learning – Learning the Rules

2.3.1 Core Idea

Machine Learning (ML) is concerned with the development of algorithms that automatically learn the properties of the environment and adapt their behavior to them. ML algorithms learn from the experience and increase its performance through the analysis of incoming data. Put more precisely

An algorithm is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [27].

The task T may involve any function kind (i.e. SINR and SNR maximization, power allocation, spectrum holes detection etc). Meanwhile wireless research community has paid little attention to ML.

In general ML, contrary to GT, focuses in the agent, it provides the agent with methods to achieve their targets, without caring whom is playing with nor does the agent know any rules **a priori**. Learning the, a priori known, rules and hence find the optimal policy is the core of ML algorithms.

ML algorithms have been widely used in problems such as speech recognition, outperforming all approaches that have been attempted to date, additionally in data mining applications machine learning algorithms are being used to discover hidden features in massive data bases. Latter interest in CR has increased the demand of intelligent algorithms, ML offers a vast array of learning algorithms and the formal mathematical framework to embed them.

2.3.2 Taxonomy

Based on the type of learning paradigm, ML algorithms may be divided in several groups.

- **Supervised learning:** We begin by considering **supervised learning**, which is illustrated in Figure 2.2. Suppose that the teacher has certain knowledge about the

environment being represented by value pairs of environment-state and the optimal-action to take $(\mathbf{S}, \mathbf{a}^*)$. The learning algorithm is trained by the training vector, containing environment-state values $\mathbf{S}$, the outcome of the learning algorithm (action $\mathbf{a}$) is compared with the optimal action $\mathbf{a}^*$, known by the teacher, the error (i.e. the action taken was or was not the optimal) is then fed back to the learning algorithm. The learning algorithm adjust itself autonomously under the influence of the training sequence and the error signal. This adjustment is usually done repeatedly until the learning algorithm emulates the teacher. A common off line training procedure is to divide the training sequence in three non-overlapping parts: often call the training part, validation part and test part. The algorithm is trained and adjusted with the training part, after that its performance is checked with the validation part, if the performance is good enough the training ends if not the algorithm is trained again with the training part until it reaches the desired performance. Once the training has ended, the algorithm is tested with the test part to definitively validate its performance. The type of training experience available can have a significant impact on success or failure of the ML algorithm.

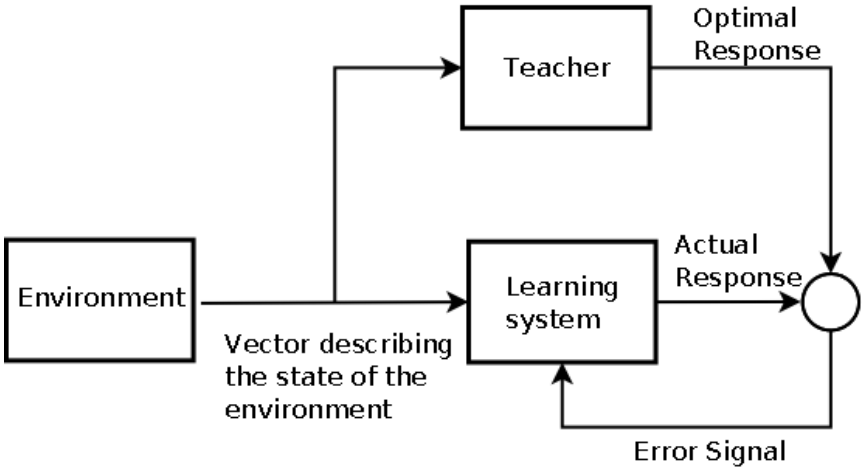


Figure 2.2: Block diagram of learning with teacher from [20]

- **Unsupervised Learning:** seeks to determine how the data is organized, without the external help of any expert entity (see Figure 2.3. Several methods of this discipline are being used in data mining and neuroscience. Typical examples of unsupervised learning algorithms are based on independent component analysis (ICA) for blind source separation (i.e. cocktail party problems). Among neural network models, the Self-organizing map (SOM) and Adaptive resonance theory (ART) are commonly used by unsupervised learning algorithms. Those algorithms perform an statistical analysis of the data in order to decide for the optimal actions to take but thereafter never receive any feed back from the environment about its actions/decisions optimality.

- **Reinforcement learning (RL):** The technique of RL [38] is concerned with the problem of finding suitable actions to take in a given situation in order to maximize a reward. The RL algorithms achieve learning by an online repeated process of interaction with the environment, as shown in Figure 2.4. The RL receive feedback of the environment for each action it takes. Notice that the difference between RL and supervised learning relays on the fact that whereas supervises learning receives information about which was the optimal action to be taken, RL receives partial information about the performance of the action taken. Said feedback is used in RL to update the action-decision rule, as a consequence of that if one action has repeatedly bad feedbacks it will become less likely to be taken. One of the advantages of the RL is that due to the fact that the learning is done online is able to adapt to the temporal dynamics of the data. Some authors have proposed a decentralized RL scheme as a method to solve MDP. The power of RL lies in its ability to solve the MDP without computing the transition probabilities. Q-learning [44] is a RL algorithm that does not need a model of its environment and can be used on-line. Therefore, it is very suited for scenarios with multiple unknown users. In the following chapter we will go deep insight Q-learning. For a complete introduction to RL the reader is refereed to the referenced literature but specially to [38].

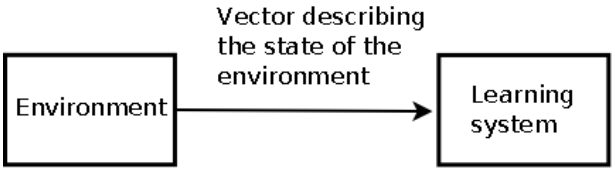


Figure 2.3: Block diagram of unsupervised Learning

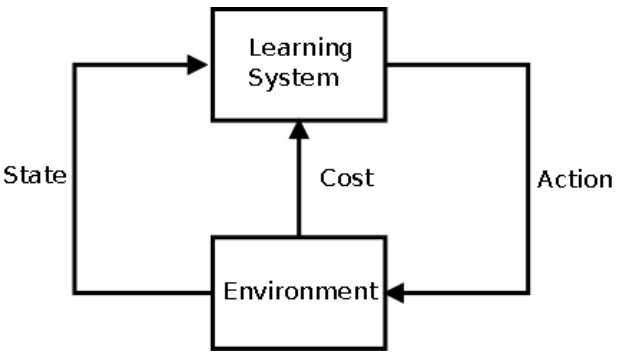


Figure 2.4: Block diagram of Reinforcement Learning

2.3.3 Q-learning

Reinforcement Learning and thus Q-learning are algorithms whose ability to find optimal action polices in MDP problems makes them really suitable for Cognitive Radio.

In the following we extend the basic notions on MDP previously introduced. Formally a MDP is defined in terms of a discrete-time stochastic dynamic system with finite state set $\mathcal{S} = \{s_1^r, s_2^r, \ldots, s_l^r\}$. Time is represented with a sequence of discrete time steps, $t = 0, 1, \ldots$. At each time step t, an agent observes the system's current state s and selects an action a. Let's define s_t as the observed state and assume that the action a_t is selected from a finite set of possible actions $\mathcal{A} = \{a(1), ..., a(m)\}$ at time t. When the agent executes action $a(k) \in \mathcal{A}$, the system's state at the next step changes from s_t to s_{t+1}. Given any state and action, s and a, the probability of each possible next state, s', is

$$P_{ss'}^a = Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$$

The model is Markov if the state transitions $P_{ss'}^a$ are independent of any previous environment states or agent actions. We further assume that the application of action a_t in state s_t incurs an immediate cost c_t ($c_t(s_t, a_t)$). The agent goal is to minimize[1] the expected discount cost C_t

$$C_t = \sum_{k=0}^{\infty} \gamma^k c_{t+k+1}$$

where γ is a parameter, $0 < \gamma < 1$, called the *discount factor*. Given any current state s and action a, together with any next state s', the expected value of the next cost is

$$C_{ss'}^a = E\{c_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}$$

Almost all RL algorithms are based on estimating value-functions of states that estimate *how good* is for a given agent to perform a given action in a given sate. Recall that a policy, π, is the decision function that links states to the actions to take, thus the probability of taking action a in state s is $\pi(s, a)$. The notion of *good* is defined in terms of future cost that can be achieved, or, to be precise, in terms of expected cost. The expected cost $V^\pi(s)$ when starting in state s and following a policy π thereafter is

$$V^\pi(s) = E_\pi\{C_t | s_t = s\} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k c_{t+k+1} | s_t = s \right\}$$

The function V^π is usually known as the *state-value function for policy* π. In a similar way we define the value of taking action a (*how good* is taking a) in state s under policy π, as the expected cost from starting form state s, taking action a, an thereafter following policy π

$$Q^\pi(s, a) = E_\pi \{C_t | s_r = s, a_t = a\} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k c_{t+k+1} | s_t = s, a_t = a \right\}$$

[1]It is noteworthy that the word return is used when *good* actions provide hight returns and the word cost is used when *good* actions provide low returns. For this reason when we speak about returns the best action/policy maximizes the expected return whilst when talking about costs the best action/policy minimizes the expected cost.

Q^π is usually known as the *action-value function for a policy* π. If $Q^*(s,a)$ is known then is also known which is the action with smaller long term cost in each state, that is not the case with $V^*(s)$'s because there is no information about the actions *goodness*, however one can look one step ahead and compare the state-value functions V^* of the previous state s and the new state s'.
RL is concerned with the estimation of those value functions from the experience. A fundamental property of value functions used throughout RL and dynamics programming is that they satisfy particular recursive relationships. For any policy π and any state s, the following equation links the value of s and the value of its possible successor states s'

$$V^\pi(s) = \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} \left[C^a_{ss'} + \gamma V^\pi(s')\right] \tag{2.1}$$

which is known as the Bellman equation for V^π [8] and it forms the basis of a number of methods to compute, approximate and learn V^π. The best policy is the policy with minimum expected cost for all states. In other words $\pi > \pi'$ if and only if $V^\pi(s) < V^{\pi'}(s)\ \forall s$. The best policy, the policy that does better or equal than any other, is the *optimal policy* denoted by π^*. The optimal state-value function, denoted V^*, is defined as

$$V^*(s) = \min_\pi V^\pi(s)$$

The optimal action-value function, denoted Q^*, is defined as

$$Q^*(s,a) = \min_\pi Q^\pi(s,a)$$

For the state-action pair (s,a), $Q^*(s,a)$ gives the expected return of taking action a in state s and following the optimal policy π^* thereafter. Thus we can rewrite the previous equation

$$Q^*(s,a) = E\{c_{t+1} + \gamma V^*(s_{t+1}|s_t = s, a_t = a)\}$$

V^* must satisfy the Bellman Equation 2.1, intuitively the Bellman optimality equations express the fact that the value of one state s under an optimal policy must equal the expected return for the best action from that state s.

$$\begin{aligned} V^*(s) &= \min_{a \in \mathcal{A}(s)} Q^{\pi^*}(s,a) \\ &= \min_a \sum_{s'} P^a_{ss'} [C^a_{ss'} + \gamma V^*(s')] \end{aligned} \tag{2.2}$$

a very similar Bellman optimality equation exist for Q^*

$$\begin{aligned} Q^*(s,a) &= E\left\{C_{t+1} + \gamma \min_{a'} Q^*(s_{t+1,a'}|s_t = s, a_t = a)\right\} \\ &= \sum_{s'} P^a_{ss'} [C^a_{ss'} + \gamma \min_{a'} Q^*(s', a')] \end{aligned} \tag{2.3}$$

The reader interested in the complete development of Equations 2.2 and 2.3 is refereed to [44] or [38]. The Bellman optimality equation (Equation 2.2) defines a system of equations, one for each state, so if there are N states, then there are N equations in N unknowns. If the dynamics of the environment are known ($C^a_{ss'}$ and $P^a_{ss'}$), then in principle one can solve the system of equations for V^* using one of the number of methods that exist. However this solution relies on three assumptions that are never completely true in practice: 1) the environment's dynamics are perfectly known , 2) the computational resources available are large enough to solve the equations system, and 3) the system dynamics fulfill the Markov property. For the CR type of tasks in which we are interested, one is generally not able to implement this solution exactly because various combinations of these assumptions doesn't hold.

The elegance of V^* relays on the fact that if one uses it to evaluate the short-term consequences of actions, specifically, the one-step consequences, then a myopic policy is actually optimal in the long-term sense in which we are interested, because V^* already takes into account the cost consequences of all possible future behavior. By means of V^*, the optimal expected long-term cost is locally and immediately available for each state. Hence, a one-step ahead search yields to long-term optimal actions. Having Q^* makes choosing optimal actions still easier. With Q^*, the agent does not even have to do a one-step ahead search: for any state s, it can simply find any action which minimizes $Q^*(s, a)$. The action-value function effectively caches the results of all one-step ahead searches. It provides the optimal expected long-term cost as a value that is locally and immediately available for each state-action pair. Hence, at the cost of representing a function of state-action pairs, instead of just of states, the optimal action value function allows optimal actions to be selected without having to know anything about possible successor states and their values, that is, without having to know anything about the environment's dynamics[38].

Estimation of $Q^*(s, a)$

To estimate the action-value functions and discover optimal policies RL requires only experience, state-action-reward values, acquired through continuous interaction with the environment. Learning from the experience is appealing because it does not require any prior knowledge of system dynamics. The most simple way to estimate the action-value functions $Q^*(s, a)$ is to average the returns after taking action a in state s. By the law of large numbers the sequence of averages must converge to their expected values, however there are a number of alternative methods to estimate the action-value function.

Q-Learning (Watkins 1989 in [44]) is a RL algorithm that uses the following updating rule to estimate $Q^*(s, a)$

$$Q(s, a) \leftarrow Q(s, a) + \alpha[c + \gamma \min_a Q(s', a) - Q(s, a)] \tag{2.4}$$

where α is the learning rate and γ the discount rate. The influence of both parameters in the performance of Q-learning may be very large, actually the parameter tuning task may

become cumbersome in some cases, in [33] some hints and rules for the setting of those parameters are shown. In Q-learning (Equation 2.4) the action-value function estimation is updated by a weighted combination of the immediate cost obtained and the *goodness* $Q(s', a^*)$ of the next state s'. Image the situation where an action has an immediate low cost value, however the actions drives the system to a state with very high estimate action-value cost, Q-learning will hence penalize the action taken.

RL learning presents several advantages compared to Dynamic Programming or Monte Carlo Methods. RL estimate the action-value functions on the contrary to Dynamic Programming which needs all the environment dynamics information in order to find the optimal policy π^*. Monte Carlo methods can also learn from raw experience, however the learning in a episode-wise fashion whilst in RL is action-wise.

Exploration and exploitation

To estimate the action-value function, the Q-learning has to visit all the sates several times in order to find the action with smaller estimated action-value in for all the states. For this reason the Q-learning learning process has to *explore* the system taking actions which are no the optimal according to the estimation of $Q^*(s, a)$ and hence explore new unvisited states, avoiding being trapped in a small group of states and so in local minima of the problem solution. Additional the learning process has also to *exploit*, that is when the learning algorithm takes the actions with best expected cost. Is in the decision making policy $\pi(s, a)$ where the *exploration* and the *exploration* of the system is ensured, we present the two most common decision methods. The hard decision policy known as ϵ**-greedy** explores with probability ϵ and is greedy (takes the actions with lower Q-values) the rest of the time. Alternatively the Boltzman distribution is used to avoid hard decisions policies, Q-values are then used to build up a Boltzman distribution, the probability of taking an action is thus

$$\pi(s, a) = \frac{e^{-Q(s,a)\tau}}{\sum_{a'} e^{-Q(s,a')\tau}}$$

Representation Mechanisms

In order to learn from the past Q-learning has to store the Q-values ($Q(s, a)$) in a representation mechanism. For problems with a small number of state-actions pairs the most usual mechanism is a look-up table. However, if the number of states-actions pairs increases or the input variables (states or actions) are defined as continuous variables instead of discrete variables, the memory requirements may become unfeasible, so that there is a necessity for more efficient representation mechanism. The representation mechanisms are often call function approximation because they takes samples from the desired function and attempt to, from them, construct an approximation of the entire function. In [38] several function approximation mechanism are presented, such as gradient-descent methods and several linear methods. Other function approximation methods, more related with the

ML community, are Supervised Learning methods like the Neural Networks or Support Vector Machines. In [41] a Neural Network is used represent the action-value function in a multi-agent Q-learning problem, in as similar way Support Vector Machines are used in [25].

Noisy state observations

In many real world problems, it is not possible for the agent to have perfect knowledge and complete perception of the state of the environment. As a result it makes sense to consider situations in which the agents observe the state of the environment, but this observations are noisy. Partially Observable Markov Decision Process (POMDP)[30], based on state estimator, extend the normal MDP. In [49] a POMDP is used to model a decentralized MAC protocol. A very interesting paper can be found in [13] where Q-learning is used along Neural Networks representation mechanism to solve an aggregate-interference problem on a cognitive radio system modeled by a POMDP.

Decentralized Q-Learning

When dealing with multiple agents constantly interacting we incur to a field known as multi agent systems (MAS). In this field many problems still remain open - even for machine learning experts. The main challenge is how to ensure that individual decisions of the agents result in jointly optimal decisions for the group, considering that the standard convergence proof for Q-learning does not hold in this case as the transition model depends on the unknown policy of the other learning agents. The Q-learning version for decentralized systems is call *decentralized Q-learning*. In principle, it is possible to treat the distributed cognitive radio network as a centralized one, where each agent has complete information about the other agents and learns the optimal joint policy π using standard RL techniques. However, both the state and action spaces scale exponentially with the number of agents, rendering this approach infeasible for most problems. Alternatively, we can let each agent learn its policy independently of the other agents, but then the transition model depends on the policy of the other learning nodes, which may result in oscillatory behaviors and in slow speed of convergence to prior set targets [13]. This introduces game theoretic issues to the learning process, which are not yet fully understood [21].

2.3.4 Evolutionary Games

Game Theory usually assumes that the game is played by a single player which is fully rational and is aware of the other player preferences over game outcomes. Evolutionary Game theory (EGT), instead, imagines that the games is played *repeatedly* by individuals who are *randomly* drawn from *large populations*. We believe that EGT, besides of classically being classified into the field GT, is an hybrid between ML and GT due to its temporal learning component and its ability to adapt to the environment dynamics (by adapting the

populations share). In the following the the basic notions of EGT are explained, special attention should be paid to the replicator dynamics which links RL with EGT.

Specifically each individual is preprogrammed to some *strategy*[2] $\mathbf{x}$ and it is assumed that some evolutionary selection process operates over time and strategies distribution [47]. Most of the literature on EGT is focused on the *pairwise* interaction of individuals. The expected payoff (reward) for one individual playing strategy $\mathbf{x}$ against one individual playing strategy $\mathbf{y}$ is $u(\mathbf{x}, \mathbf{y})$. The payoff is usually call fitness, strategies with larger fitness are expected to propagate faster in a population. $u(\mathbf{x}, \mathbf{y})$ is linear in $\mathbf{x}, \mathbf{y}$ such that u($\mathbf{x}$,$\mathbf{y}$)= $\mathbf{x}A\mathbf{y}$ where A is known as the payoff matrix.

It is usually assumed that each individual plays a pure strategy, the population shares[3] between individuals playing pure strategies define the mixed strategy $\mathbf{x}$ of the whole population. The payoff received by the individuals against other individuals of the same population determine the offspring and hence the dynamics of the population share that determine the mixed strategy (see Equation 2.6).

A common way to treat the population in *multiplayer* games is to assume that are close entities playing a mixed strategy and varying its own strategy according to its performance.

In *multipopulation* problems it is assumed that there are many populations of individuals and the interaction is done pairwise between individuals of two random populations. Alternatively *multiplayer* games focus on the case where the interactions between individuals are no pairwise any more, since the payoff of one individual depends on its own strategy and the strategy of the randomly selected individuals of the other populations. It is very common to treat the population of individuals as a single *player*, since they behave similar. A logical extension of the multipopulation and multiplayer cases is the *multiplayer game between many different populations* in this case one individual is randomly selected from each population, the interaction is done between many individuals and the utility received by each individual will determine the offspring of individuals of the same specie in the same population. The population share of the different populations or more precisely the strategy $\mathbf{x}_i$ of the different players varies according to the reward received. *Multiagent* learning problem are equivalent to multiplayer game between many different populations.

Evolutionary Stable Strategy (ESS)

One of the main definitions in EGT is the *evolutionary stability criteria* which defines the robustness of a whole population playing a mixed strategy $\mathbf{x}$ against other strategy $\mathbf{y}$ called mutation or invasion strategy. The rational goes as follows: suppose that as stated above two individuals are repeatedly drawn from at random to play a certain pure or mixed strategy $\mathbf{x}$. Then inject a small population share of player who are likewise programmed to play an other strategy $\mathbf{y}$. The incumbent strategy is said to be *evolutionary stable* if,

[2]Here we use the term strategy to refer to mixed strategy (contains several actions) and to pure strategies (contain just one action). Hence the mixed strategy set of player i is $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iN}]$ where N is the number of actions available to player i and x_{ih} are the population shares of player i or the probability for player i of doing action h.

[3]Ratio of individuals playing a certain action $\mathbf{a}$ to the total number of individuals in the population.

for each such mutant strategy, there is an invasion barrier ϵ, such that is the amount of mutants falls bellow this barrier, then the incumbent strategy earns higher payoff than the mutant strategy. It is noteworthy that this approach if focused on symmetric pairwise interactions within a single large population. Actually, most of the analysis from [47] are almost exclusively focused on symmetric two-player games. The criterion of evolutionary stability refers implicitly to a close connection between the payoffs in the game and the spreading of a strategy in a population. The theory does not describe how to reach an evolutionary stable strategy. Instead, it asks whether, once reached, a strategy is robust to evolutionary pressures.

Definition 1 *A strategy $\boldsymbol{x}$ is said to be ESS if for every $\boldsymbol{x} \neq \boldsymbol{y}$ there exist some $\bar{\epsilon_y}$ such that inequality 2.5 holds for all $\epsilon \in (0, \bar{\epsilon_y})$.*

$$u[\mathbf{x}, \epsilon\mathbf{y} + (1-\epsilon)\mathbf{x}] > u[\mathbf{y}, \epsilon\mathbf{y} + (1-\epsilon)\mathbf{x}] \tag{2.5}$$

The concept of ESS is extended to multi population[4] problems in Chapter 5 of [47] defining weak and strict evolutionary criteria. In general evolutionary stability does not imply that the average population fitness $u(\mathbf{x}, \mathbf{x})$ is maximized. However, as stated in [47], even weak criteria of evolutionary stability in multipopulation problems reject all but strict Nash equilibria.

Proposition 1 *being $\Theta$the polyhedron of mixed strategy profiles, $\boldsymbol{x} \in \Theta$ is evolutionary stable if and only if $\boldsymbol{x}$ is a strict Nash equilibrium.*

In general the ESS characterize the property of robustness against invaders (mutation)

- If an ESS is reached, then the proportions of each population remain constant in time (the average rewards of the population will remain constant, since no mutant strategy may induce any population to change strategy)
- The notion of ESS is stronger than Nash equilibrium in which it is only requested that a single user would not benefit by a change (mutation) of its behavior.

The Replicator Dynamics (RD)

Along with the ESS the RD form the basis of the EGT. RD combines two basic elements, *the selection* where individuals, or strategies, with higher fitness have more offsprings than others with lower fitness and *the mutation* that controls the turn up of new strategies of the player. In Chapter 3 of [47] the RD is formalized as a system of ordinary differential equations that do not include any mutation mechanism at all, there after, in Chapter 4 [47], some mutation mechanism are introduced and in Chapter 6 [47] the RD equation for multipopulation are stated.

[4] In general multi population models study the interactions of n-players games. We can image large populations of individuals, one such population for each player in the game, one individual in each population is randomly selected to play the game.

The RD the population share of strategy $\mathbf{x}$ in the population grows at a rate proportional to the difference between the payoff of that strategy and the average payoff of the population

$$\dot{x}_{ih} = [u_i(e_i^h, x_{-i}) - u_i(x)]x_{ih} \tag{2.6}$$

where x_{ih} is the population share playing strategy h of player i and $u_i(e_i^h, x_{-i})$ is the average fitness for player i when playing strategy h. An equivalent expression for two player games is the matrix form

$$\begin{aligned} \dot{x}_h &= [e^h \cdot A\mathbf{y} - \mathbf{x} \cdot A\mathbf{y}]x_h \\ \dot{y}_h &= [e^h \cdot B^T\mathbf{x} - \mathbf{y} \cdot B^T\mathbf{x}]x_h \end{aligned} \tag{2.7}$$

where A and B are the payoff matrices.

Replicator Dynamics and reinforcement learning In [10] is shown a direct relation between the continuous time RL equations and the *selection* and *mutation* mechanism on the RD.

Starting with a group of general RL agents[5] and following the work in [36] we derive the coupled replicator dynamics of collective learning in multiagent systems. Let's assume that there are two agents X and Y that at each time step take one out of N actions. Let the probabilities for X to chose action $i = 1, ..., N$ at time t be $x_i(n)$ and $y_i(n)$ for Y. Let R_{ij}^X and R_{ij}^Y denote the rewards for X taking action i and Y taking action j. Define $Q_i^X(n)$ and $Q_i^Y(n)$ as the memories of agents X' and Y' respectively, strong information of past benefits from the past actions, the update the memories is ruled by

$$\begin{aligned} Q_i^X(n+1) - Q_i^X(n) &= R_{ij}^X - \alpha_X Q_i^X(n) \\ Q_i^Y(n+1) - Q_i^Y(n) &= R_{ij}^Y - \alpha_Y Q_i^Y(n) \end{aligned} \tag{2.8}$$

where α_X and α_Y control each agent memory. The agents choose actions according to the Bolztmann distribution

$$x_i(n) = \frac{e^{\tau_X Q_i^X(n)}}{\sum_j e^{\tau_X Q_j^X(n)}} \tag{2.9}$$

with a similar equation for $y_i(n)$. We are interested in how the strategies ($\mathbf{x}(n)$ and $\mathbf{y}(n)$) change in time, the dynamics of the strategies are strongly linked to the dynamics of the rewards or Q values. Making the time limit derivative of the strategy $x_i(n)$ in Equation 2.9 we obtain

$$\dot{x}_i = \Big[\tau_X\Big[R_i^X - \sum_j x_j R_j^X\Big] + \sum_j x_j \log(x_i/x_j)\Big]x_i \tag{2.10}$$

[5]While the term agents comes from the world of ML, in the GT world is more used the term player, however in this document both have the same meaning.

If we further assume a fixed reward for action pairs (i, j) and that x and y are independently distributed, the continuous time dynamic becomes

$$\dot{x}_i = \tau_X \underbrace{\Big[(A\mathbf{y})_i - \mathbf{x}A\mathbf{y}\Big]x_i}_{selection} + \underbrace{x_i \sum_j x_j \log(x_i/x_j)}_{mutation} \tag{2.11}$$

a similar equation can be obtained for $y_i(n)$. The above equation is composed of two main terms: *the selection*, the same as the replicator dynamics (Equation 2.6), where strategies with higher payoff than the average payoff have more offsprings (positive derivative) than others, the second term is the *mutation* which we will study more accurately in the next sections. We have shown that the RD (Equation 2.6) plus a mutation mechanism (see Chapter 4 in [47] for more details in mutation) is equivalent to the continuous time dynamics of a general RL algorithm. Similar work has been done in [43] but focused specifically in Q-leading with the same results, the same authors purpose in [23] and extension of the Q-learning that fits better the RD model.

Mutation and Entropy The mutation term in Equation 2.11 can be expressed in terms of the $\mathbf{x}$ entropy

$$\begin{aligned}
x_i \sum_j x_j \log(x_i/x_j) &= x_i \left[\sum_j x_j \left(\log(x_j) - \log(x_i) \right) \right] \\
&= x_i \left[\sum_j \left(-x_j \log\left(\frac{1}{x_j}\right) - x_j \log(x_i) \right) \right] \\
&= x_i \left[-\sum_j x_j \log\left(\frac{1}{x_j}\right) - \sum_j x_j \log(x_i) \right] \\
&= -x_i \left[H(\mathbf{x}) + \log(x_i) \right] \\
&= x_i \left[\log\left(\frac{1}{x_i}\right) - H(\mathbf{x}) \right] \\
&= x_i \left[I(x_i) - E\{I(\mathbf{x})\} \right]
\end{aligned}$$

where $H(\mathbf{x})$ is the entropy of the strategy $\mathbf{x}$ and $I(x_i)$ is the information given by action x_i. The mutation depends only on the specific probability distribution of the action (the selection mechanism, see Equation 2.9) not on the instantaneous rewards of other players. Actually the mutation mechanism smooths the information contained along all the actions: it increases the information given by the actions that contain less information and does the other way around with the actions that contain more information. This tends to produce flatter the probability density function $\mathbf{x}$. It is noteworthy that the sum off all mutations

in a specific strategy sums up to 0, since

$$\begin{aligned}\sum_i \Big[x_i\Big[H(\mathbf{x}) + \log(x_i)\Big]\Big] &= \sum_i [x_i H(\mathbf{x})] + \sum_i [x_i \log(x_i)] \\ &= H(\mathbf{x}) \sum_i x_i - \sum_i \left[x_i \log\left(\frac{1}{x_i}\right)\right] \\ &= H(\mathbf{x}) - H(\mathbf{x}) = 0\end{aligned}$$

More work on how the distribution of $\mathbf{x}$ affects the mutation should be done, notice that different RL algorithms may have different mutation mechanism.

2.3.5 Problems

Most of the state of the art RL algorithms are black box models. This makes it difficult to gain detailed insight into the learning process and parameter tuning becomes a troublesome task [43]. There is a lack of mathematical tools to predict the performance of ML algorithms. ML algorithm configuration (i.e. define states, set cost functions) usually ends up with heuristic approaches far more based on trial and error than on analytical proofs (see [33] from some hints on the parameter tuning). Additionally, decentralized Q-learning, which is very suitable for cognitive radio algorithms, on the contrary to Q-learning has no prove of convergence, and it may take a huge numbers of iterations before to converge to the optimal policy (if it does)[16]. EGT and hence RD have the same scenario restrictions than GT for this reason are not suitable to be used directly in CR.

2.4 Summary

We have seen that one of the most suitable intelligent learning paradigm for CR is RL from ML. RL is able to learn the environment dynamics and get advanced to the systems changes. Q-learning is one of the most promising RL algorithms since is able to find optimal policies in MDP systems (remember that CR scenarios are usually modeled as MDP) and additionally, besides its convergence is not proof, is able to work in a fully decentralized fashion. The major drawbacks are the low convergence speed of the decentralized version and the almost imperative necessity of using heuristic trial and error methods to set up the algorithm (tune learning parameters, decide the state and action space and design a valuable cost function). RD may help in the future to tune the parameters of Q-learning. Additionally it has been shown the relation between the mutation term of RD (equivalent to exploration in RL) with the entropy of the agent actions, which may help to better understand its behavior. However the gap between the RD assumptions and the real CR scenarios is still to large. Finally, due to the before mentioned reasons next Chapter introduces a novel and largely unexploited concept of docitive networks with the prime aim of speed up the convergence and enhance in general the RL performance.

Chapter 3

Docitive Networks – The Intelligent Teaching Paradigm

3.1 Introduction to Docitive Networks

The emerging and largely unexploited concept of docitive networks was first introduced in the position paper [11] and lately extended in [16]. This Chapter is thus heavily based on those two seminal papers. Cognition, from cognoscere = to know in Latin, is typically defined as "a process involved in gaining knowledge and comprehension, including thinking, knowing, remembering, judging, and problem solving" [19]. CR learn from the environment by sensing the surroundings and draw intelligent decisions (see Section 2.1 for a basic introduction to CR). This learning process is often a lengthy and complex process in itself, with complexity increasing with an increasing observation space. It is however needed to truly realize a cognitive radio as otherwise only opportunistic access is guaranteed at best. Whilst cognition and learning have received a considerable attention from various communities in the past, the process of knowledge transfer, i.e. teaching, over the wireless medium however has received fairly little attention to date. The position paper [11] aims at introducing an emerging framework referred to as docitive radios, from docere = to teach in Latin, which relates to radios (or general entities) which teach other radios. These radios are not (only) supposed to teach them the end-result (e.g. in form of "the spectrum is occupied"), but rather elements of the methods of getting there. This concept mimics well our societydriven pupil-teacher paradigm and is expected to yield significant benefits for cognitive algorithms and thus more efficient communications. Important and unprecedented questions arise in this context, which will be elaborated on throughout this Chapter.

3.2 Docition - Teaching the Rules

3.2.1 Core Idea

A typical approach models a cognitive radio system as a multiagent system, where the radios learn through the paradigm of multiagent learning. When it comes to MAS, Q-learning can be adapted to this setting, by implementing decentralized Q-learning (see Section 2.3.3). In this case, each node has to build a state-action space where it needs to learn the optimal policy for taking actions in each state. With an increasing dimension of the state-action space required to viably support future wireless communications, the training process may be extremely time consuming and complex. However, if nodes are instructed to learn some disjoint or random parts of the state-action space, then they can share the acquired knowledge with their neighboring nodes. This facilitates learning but does not yield the end result per se. Contributions in literature [39] suggest that the performances of such a MAS can be improved by using cooperation among learners in a variety of ways. Depending on the degree of cooperation among CR, in [11] is proposed to consider the following cases for future studies:

- **Independent Learners:** The agents do not cooperate, ignore the actions and rewards of the other agents in the system and learn their strategies independently.
- **Cooperative Learners Sharing State Information:** The agents follow the paradigm of independent learning, but can share instantaneous information about their state. It is expected that sharing state information is beneficial in case that it is relevant and sufficient for learning.
- **Cooperative Learners Share Policies or Episodes:** The agents follow the paradigm of independent learning, but can share information about sequences of state, action and reward and learned decision policies corresponding to specific states. It is expected that such cooperative agents can speed up learning, measured by the average number of learning iterations, and reduce the time for exploration.
- **Cooperative Learners Performing Joint Tasks:** Agents can share all the information required to cooperatively carry out a certain task. In this case the learning process may be longer, since the state-action space is bigger, but oscillatory behaviors are expected to be reduced.
- **Team Learners:** The multi-agent system can be regarded as a single agent in which each joint action is represented as a single action. The optimal Q-values for the joint actions can be learned using standard single-agent Q-learning. No communication is needed between the nodes but they all have to observe the joint action and all individual rewards.

From the above we incur that the concept of joint learning has received attention in recent years in the ML and artificial intelligence community; however, its application to CR

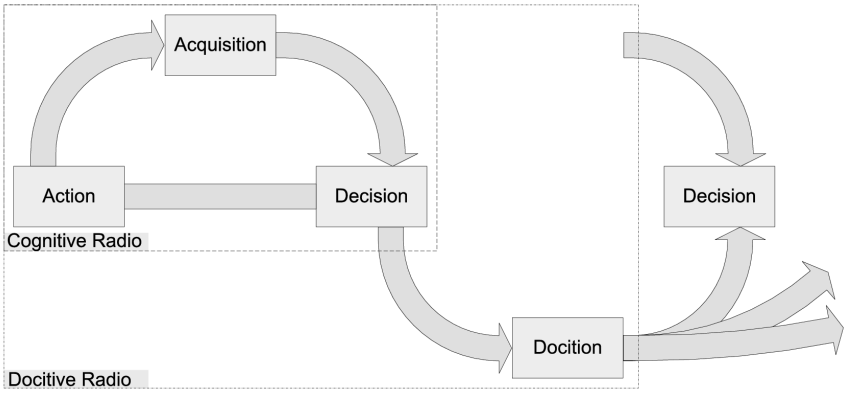


Figure 3.1: Docitive cycle which extends the cognitive cycle through the teaching element [11]

operating primarily over a wireless broadcast channel has received few attention up to today and, coupled with the potential gains, essentially inspired the concept of docitive radios. As illustrated in Figure 3.1, the canonical cognitive cycle is advantageously extended by the following element: Docition. It is realized by means of an entity which facilitates knowledge dissemination and propagation with the non-trivial aim to facilitate learning. It is inspired by the so-far-successful problem based learning (PBL) concept used at schools in our society. In PBL, teachers are encouraged to be coaches and not information givers with the aim to have pupils work as a team using critical thinking to synthesize and apply knowledge; apprehend through dialogue, questioning, reciprocal teaching, and mentoring. Translated back to the wireless setting, this implies a distributed approach where agents share potentially differing amounts of intelligence acquired on the run. This, in turn, is expected to sharpen and speed up the learning process. Any achieved gains, however, need to be gauged against the overhead incurred due to the exchange of docitive information. The range of application scenarios is vast, including infrastructureless CR networks, novel cellular systems such as femtocells, etc. The distributed learning and teaching paradigm applied to these novel networking architectures, however, raises unprecedented questions, where we first concentrate on learning and subsequently on teaching issues.

Multiple agents have thus to distributively learn an optimal policy to achieve a common objective. Known as "multi-agent learning" problem, it can be solved by means of distributed RL approaches such as distributed Q-learning. As for teaching approaches, some early contributions in literature [39] suggest that the performances of a decentralized learning system can be improved by using cooperation among learners in a variety of ways. A *learning agent* e.g. can take advantage of the exchange of information and expert knowledge from other agents [39], the so-called *docitive agents*.

3.2.2 Taxonomy

In [11] and [16] different degrees of docition are considered, in this chapter we extend the previous work by formalizing and arranging the docition essence. We consider the docition paradigm in a MAS but without relaying in any specific learning algorithm, and thus with no loss of generality. Essentially, the docition paradigm is divided in three parts, *When*, *Who* and *What to teach/learn?* Its answers combination leads to different degrees of docition.

The *When to learn?* question refers to the fact that an agent may receive docition several times during the entire period of the learning process. How often should the docition be done and whether the docition is better at the beginning or at later stages of the learning process are still open questions. An agent may get benefit from being taught a *low level* instruction (i.e. *in state* s_1 *use action* a_2 *a* 30% *of the time*), however, in the other hand *high level* advices may be the most advantageous (i.e. *in states* s_1, s_2, s_3 *try to use lower powers*). *What* should be taught? entire policies, low level instructions or high level advices? Two basic elements are fundamental for the docition paradigm, the *docitive agent* and the *learning agent*. The docition direction goes from the most experts to the novice agents, however agents sharing the same environment may be in different conditions with utterly opposed optimal policies. Thus a critical trade-off exist between *expertness* and *conditions correlation* (between docitive and learning agent). *Who* should be the teacher? The expert agent or an agent learning in similar conditions? Additionally, in the state-of-the-art, there are no general measures of expertness neither of condition correlation and most of them are designed ad-hoc. Several questions concerning *When*, *Who*, and *What to teach/learn?* still remain open. In the following the different docition paradigms are carefully explained and depicted.

- *When?* Docition is expected to sharpen and speed up the learning process. Any achieved gains, however, need to be gauged against the overhead incurred due to the exchange of docitive information [16]. The overuse of docition may lead also to an increase of the oscillations in the system. Depending on the docition time-line, one can consider the following docition paradigms
  - **Single Learning:** The learning agent here receives docition just once. The docition may be done at the very beginning of the learning stage (i.e Start-up Docition in [9]). Alternatively, the docition may be triggered by other signals (i.e. learning agent performance)
  - **Iterative learning:** The learning agent receives docition multiple times. The docition interval may be fixed or driven by external signals (i.e. system global performance)
- *What?* The information shared by the docitive agents may be *low level* information directly extracted from the core of the learning algorithm or interpreted and transformed *high level* information. The information may concern to specific states, *state-based* docition, or to the whole state-space, *system-based* docition. Depending

on the its nature the shared information here we propose the following cases for future use.

  - **Advices:** The docitive agent shares information in form of advice. Advice is a set of *high level* instructions about a task solution that may not be complete or perfectly correct. Advice-taking algorithms allow advice to be followed, refined, or ignored according to its value [42].
  - **Rules:** The docitive agent teaches very reliable information in form of rules. Rules are a set of *high level* instructions about a task solution that are presumably perfectly correct. In contrast to advices the rules must always be followed.
  - **State information:** In this setting *low level* information about the internal state (i.e. learning algorithm parameters, current state) of the docitive agent is shared with the learning agent.
  - **Policies or Episodes:** The docitive entity shares is own current policy π (i.e. the Q-values in Q-learning [14]).

- *Who?* To learn incorrect information may incur to an irregular behavior of the learning agent and hence destabilize the whole system. Thus the learning agent has to estimate the feasibility of learning from each docitive agents. The feasibility may be measured in terms of *expertness* of the docitive agent or in terms of the environment conditions correlation between the learning and the docitive agent. A learning agent may weight docitive information received from multiple docitive sources in the same docitive instance [4], called *weighted learning*.

  - **Performance based Expertness:** A simple metric for expertness is to quantify the instantaneous, windowed or accumulated difference between the actual performance of a cognitive algorithm with the targeted one. It ignores agent-internal processes related to learning, such as building the Q-table, but values the ability of a agent to meet a given performance target [14].
  - **Reward based Expertness:** A more sophisticated approach capitalizes on the fact that typical cognitive approaches, such RL algorithms, issue a reward for every action taken. A quantification of the agents's expertness can hence be taken from the instantaneous, windowed or accumulated rewards acquired by the docitive node during its lifetime. A variety of promising metrics, such as norm, absolute, positive, negative and average move, have been proposed and discussed in [4].
  - **Updates based Expertness:** This expertness metric assumes that the docitive agent increases its expertness at each learning iteration [4]. Hence the agents that have updated more times its learning algorithm and thus probably the agents that have learned longer, are assumed to be the most experts.
  - **Action Entropy based Expertness:** EGT is closely related to RL (see section 2.3.4). In GT the equilibriums are found to be pure strategies. However, since

the action space is not continuous, some mixed strategy equilibriums may arise. Here we propose to make docition based on the entropy of the actions that the agent will take in a certain state. The entropy is a measure of the *pureness* of a strategy. The docition should be done from agent-states with low entropy (almost pure strategy) to high entropy (almost random action) agents-states.

- **Differential Entropy based Expertness:** The previous methods only indirectly consider the environment, which can however be taken into account in an explicit manner by utilizing the stimulus difference in input and output entropy. Notably, the intelligence of a cognitive engine can be quantified by the reduction of high-entropy input disorder into low-entropy output order conditioned/ centered on the performance target(s) [17].
- **Scenario conditions based:** This docition paradigm capitalizes on the fact that agents facing similar conditions have similar optimal policies. Few methods are found in the literature to measure the similarity or the correlation of the conditions faced by the learning agents. Here we cite some as a example, in [16] the learning agents receive dociton from the closest docitive agent. Alternatively [15] the similarity of the impact that the agents actions may have on the environment is used.

Every docition method should contain any combination of the three main groups (*What*, *Who* and *When*) of docitive characteristics. Every combination leads to different degrees of docition.

Additionally we can consider two extreme cases of docition

- **No Docition:** The cognitive decisions engines do not generally cooperate with other agents an thus each agent learns its strategies independently. The advantage is that there is no overhead due to cooperation: however, performance is clearly expected to be worst.
- **Perfect Docition:** The multi-user system can be regarded as an intelligent system in which each joint action is represented as a single action. In order to apply this approach, a central controller should model the MDP and communicate to each agent its individual actions. Alternatively, all agents should model the complete MDP separately and select their individual actions; in this case, no communication is needed between the agents but they all have to observe the joint action and all individual rewards. Although this approach leads to the optimal solution, it is infeasible for problems with many nodes since the joint action space, which is exponential in the number of agents, becomes intractable. This is why it will not be analyzed in following Chapters.

3.2.3 Problems

Whilst the exchange of end-results among cooperatively sensing agents has been explored in the wireless communication domain and the joint learning via exchange of states has

been known in the ML community, no viable framework is available to date which quantifies the gains of a docitive system operating in a wireless setting. Numerous problems hence remain, one of the core problems is how to quantify the degree of intelligence[1] of a cognitive algorithm. With this information at hand, intelligence gradients can be established where docition should primarily happen along the strongest gradient. This would also allow one to quantify the tradeoff between providing docitive information versus the cost to deliver it via the wireless interface. Some other pertinent questions are how much information should be taught, can it be encoded such that learning radios with differing degrees of intelligence can profit from a single multicast transmission, how much feedback is needed, how often should be taught, etc?

3.3 State of the Art

3.3.1 Machine Learning Community

The exchange of acquired knowledge between learning agents has been studied in the ML community. *Inductive Learning* focuses on storing knowledge gained while solving one problem and applying it to a different but related problem [48]. Inductive learning may be applied when the different learning scenarios have similar or correlated conditions (i.e. the abilities acquired while learning to walk presumably apply when one learns to run, and knowledge gained while learning to recognize cars could apply when recognizing trucks). This area of research bears some relation to the long history of psychological literature on transfer of learning, although formal ties between the two fields are limited. Closely related to inductive learning the *Transfer Learning* [32] deals with the transfer of knowledge: extract the knowledge from the data or a learning algorithm, transfer the knowledge and finally how does the learning agent assimilate the new incoming knowledge. In [42] RL employs transfer learning techniques to increase its performance and speed-up the convergence. In there, the techniques used to extract the knowledge are related to inductive logic programming [27] [29], additionally the kernel-based representation mechanism [26] [37] used enforces the transfer and acquisition of knowledge [24] [25]. The docition literature related to wireless (see Section 3.3.2) focus on *low level* information transfer in model free algorithms (such as Q-learning) in the contrary the transfer learning deals also with *high level* advices or rules even in model-based RL [40].

A Few methods, apart from the Inductive Learning and Transfer Learning methods have focused on transfer learning for RL algorithms. Two papers are stressed, in [4] learning robots share knowledge a simple way: exchange their Q-Tables, a vast array of promising expertness measures are introduced and compared. In [34] a RL algorithm is modified to embed the acquired docitive knowledge in a more sophisticated way.

[1] The degree of intelligence is closely related to the concept of expert that we have previously introduced.

3.3.2 Wireless Community

Docition is a new and largely unexploited paradigm, thus the wireless community has paid little attention to it. Research in Docition is mainly driven by Dohler [11], Giupponi [16] and Galindo-Serrano [15]. Besides the fact that docition is a general concept that may be applied to a large variety learning algorithms the docition paradigm has been implemented always with Q-learning. Thus the docition is basically done by exchanging values of the Q-table between docitive agents and learning agents (policy share docition). Value exchange is only possible when the $Q(s, a)$ is represented as a look-up table (see Section 2.3.3). If other representation mechanism are used a different docitive method than the Q-values exchange must be used. Earlier contributions to docition are [11] and [16] there the basics of the docition paradigm are introduced and the basic concepts of ML are reviewed in order to position the emerging docitive with known cognitive approaches. In [11] the cognitive cycle is extended to the docitive cycle. Additionally the first docition degree taxonomy is introduced in [16]. Both papers use Q-learning and a Digital Television (DTV) Scenario to capitalize the docition benefits. [14] extends the previous work on DTV in [11] and [16] by introducing new docition degrees and proffering the simulations. Other publication focus on interference management in Femtocell scenarios through learning algorithm capitalizing docition [15] and [17]. The energy benefits of cognition in beyond next generation wireless communications is studied in [9]. Accordingly to the before refereed literature the docition actually speeds up the learning process by approximately one order of magnitude.

3.4 Summary

Docitive radios and networks emphasize on the teaching mechanisms and capabilities of cognitive networks, and are understood to be a general framework encompassing prior and emerging mechanisms in this domain. In docitive networks the learning agents share information to speed up its convergence and draw better and more reliable decisions. Simulation results in the literature confirm that RL performance can be actually enhanced using the docition paradigm. In the next Chapter we will show, with simulated results, the benefits of docitive networks.

Chapter 4

Application to 1Gbps/km^2 Architecture

4.1 System Model

We consider an urban area with one HBS, providing service and coexisting with N ABSs. Each ABS provides service to a single associated MSs. We consider orthogonal frequency division multiplexing (OFDM) symbols grouped into one channels. Both HBS and ABS systems operate in the same frequency band, which allows to increase the spectral efficiency per area through spatial frequency re-use. We assume multiple antennas in the HBS, with a total of 24 beams and an overlap of 3-4 beams, as a consequence we can assume that $N_b = 6 - 8$ independent beams coexist in the HBS area. In each beam there is an ABS distribution (consisting of N_{ABS} ABS), the HBS talks through all beams at the very same time, however the communication between the HBS and the ABSs of a single beam is TDMA. In Figure 4.1 is shown as an Example a deployment consisting on $N_b = 6$ distributions of $N_{ABS} = 4$ ABSs, making a total of $N = 24$ ABSs. Figure 4.1 shows in detail the structure of a single ABS distribution, corresponding to a single beam, from the Example in Figure 4.1. In this sample example the distance between the HBS an the center of the distribution is 350m, the separation between ABS is 100m and the MSs are located within a 75 m radius coverage of each ABSs.

Both HBS and ABS systems operate in the same frequency band and have the same amount R of available sub-channels, which allows to increase the spectral efficiency per area through spatial frequency re-use. We focus only on the downlink operation.

We denote by $\mathbf{p}^{i,A} = (p_1^{i,A}, \ldots, p_R^{i,A})$ and $\mathbf{p}^H = (p_1^H, \ldots, p_R^H)$ the transmission power vector of ABS i and the HBS with $p_r^{i,A}$ and p_r^H denoting the downlink transmission power of ABS and HBS respectively in sub-channel r.

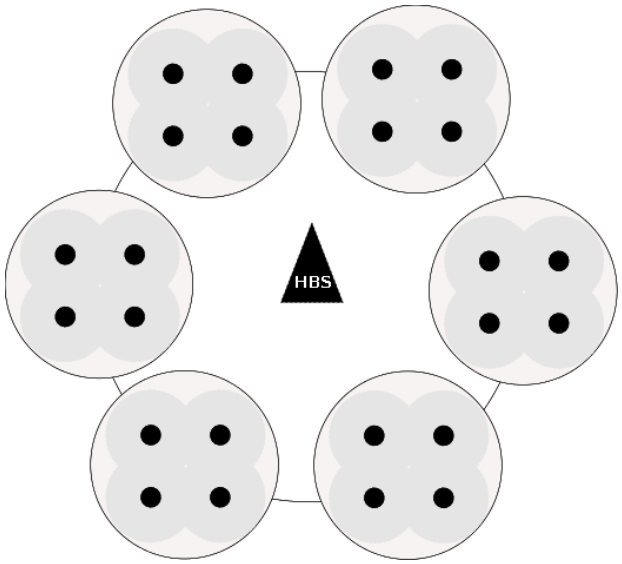


Figure 4.1: HBS system consisting on 6 ABS distributions of 4 ABSs each.

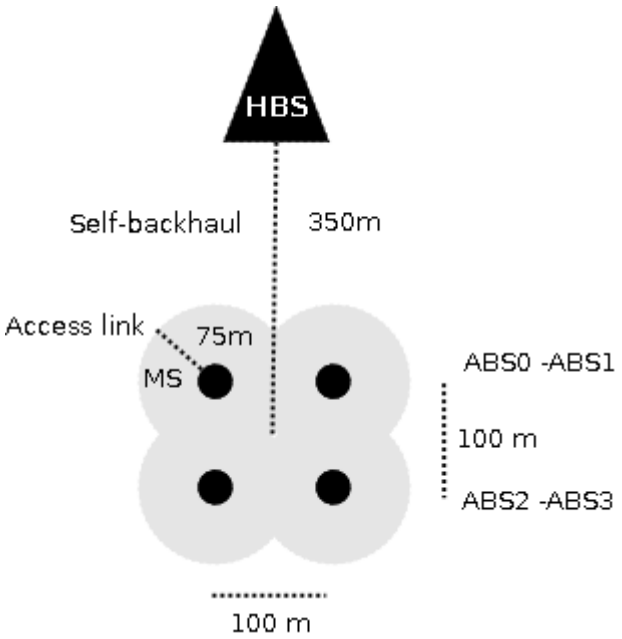


Figure 4.2: ABSs distribution representing a beam or a grouping of beams of the HBS

4.1.1 Signal-to-Noise-and-Interference Ratios

We analyze the system performance in terms of signal interference noise ratio (SINR) and capacity given in (bits/s). The SINR at MS $q \in Q$ allocated in sub-channel r of ABS i is:

$$\gamma_r^{i,u} = \frac{p_r^{i,A} h_{ii,r}^{Au}}{p_r^H h_{i,r}^{Hu} + \sum_{j=1, j\neq i}^{N} p_r^{j,A} h_{ji,r}^{Au} + \sigma^2} \tag{4.1}$$

with $i = 1, \ldots, N$. Here $h_{ii,r}^{Au}$ indicates the link gain between the transmitting ABS i and its MS u; $h_{i,r}^{H,u}$ indicates the link gain between the interfering HBS and MS u in ABS i; $h_{ji,r}^{Au}$ indicates the link gain between the interfering ABS j and MS u of ABS i, finally σ^2 is the noise power. The SINR of ABS i in sub-channel r is:

$$\gamma_r^{i,A} = \frac{p_r^H h_{i,r}^{HA}}{\sum_{j=1, j\neq i}^{N} p_r^{j,A} h_{ji,r}^{AA} + \sigma^2} \tag{4.2}$$

with $i = 1, \ldots, N$. Here, $h_{i,r}^{HA}$ indicates the link gain between the transmitting HBS and the receiving ABS i, in sub-channel r and $h_{ji,r}^{AA}$ indicates the link gain between interfering ABS j and ABS i.

4.1.2 Link Capacities

As for the resulting capacities, the capacity of ABS i used for its Q MSs is:

$$\widehat{C^{i,A}} = \sum_{r=0}^{R} \frac{BW}{R} \log_2(1 + \gamma_r^{i,u}) \tag{4.3}$$

with $i = 1, \ldots, N$. The capacity of the self-backhaul to ABS i is

$$C^{i,A} = \sum_{r=0}^{R} \frac{BW}{R} \log_2(1 + \gamma_r^{i,A}). \tag{4.4}$$

The capacity of the HBS and the entire self-backhaul is:

$$C^H = \sum_{i=0}^{N} C^{i,A} \tag{4.5}$$

The capacity of the self-backhaul and the sum of the capacities of the ABSs have to satisfy the following constraints:

$$C^{i,A} \geq \widehat{C^{i,A}} \tag{4.6}$$

$$C^H \geq \sum_{i=0}^{N} \widehat{C^{i,A}} \tag{4.7}$$

which is to guarantee that there is no buffer overflow in the ABSs.

Assume that the before mentioned constrains are fulfilled, as a consequence the bottle neck of the system is in the access link and hence the total capacity of beam is the sum of all the access capacities of ABS i to its MS

$$\widehat{C^A} = \sum_{i=1}^{N_{ABS}} BW \log_2(1+\gamma^i) \tag{4.8}$$

$$= \sum_{i=1}^{N_{ABS}} BW \log_2(1+\frac{p^{i,A}h_{ii}^A}{p^H h_i^H + \sum_{j=1,j\neq i}^{N_{ABS}} p^{j,A}h_{ji}^A + \sigma^2}) \tag{4.9}$$

4.1.3 Bandwidth Usage

The aim of BuNGee is to deliver $1Gbps/Km^2$ any where in the cell. To achieve such density of capacity, the bandwidth required is

$$BW = \frac{10^9 \cdot Sup}{\sum_{i=1}^{R} \log_2(1+\gamma^i)} \tag{4.10}$$

where *sup* is the area covered by an antenna beam in Km^2.

4.1.4 Channel Model

A suitable model for a Beyond-3G short range urban scenario is the COST-WI model [7]. This model has also been accepted by the ITU-R and was selected as Urban/Alternative Flat Suburban path-loss model in the IEEE 802.16 standard for fixed wireless access [12]. The model distinguishes between LOS (Line of Sight) and NLOS (Non Line of Sight) situations. Here, we assume that the channel gains $h_{i,r}^{HA}$ between the transmitting HBS and the receiving ABS i, in sub-channel r and $h_{ii,r}^{Aq}$ between the transmitting ABS i and its user q are LOS with short-range path-loss (PL) model $PL(dB) = 30.18 + 26.0\log_{10}(d)$. Alternatively, it is assumed that the channel gains $h_{i,r}^{H,q}$ between the interfering HBS and MS q in ABS i, $h_{ji,r}^{Aq}$ between the interfering ABS j and user q of ABS i and $h_{ji,r}^{AA}$ between interfering ABS j and ABS i are NLOS with short-range path-loss model $PL(dB) = 11.14 + 38.0\log_{10}(d)$. Here, d is the distance (in meters) between the transmitter and the receiver.

4.2 General Learning Process

In this section we describe the system learning process for the self-adaptation of the transmission power associated with the ABSs. We propose a decentralized RL scheme, where

the ABSs are multiple agents aiming at learning an optimal control policy by repeatedly interacting with the controlled environment in such a way that their performances, evaluated by scalar costs, are minimized [38]. There exist several RL algorithms. For our particular problem and for the reasons before mentioned in Section 2.4, we consider the decentralized Q-learning as an accurate algorithm to implement interference control in the proposed scenario. Additionally we consider the docition paradigm (see Chapter 3), where expert ABSs teach newly ABSs, as a complement to the decentralized Q-learning. We consider that each agent is characterized by as many learning processes as available sub-channels. The Q-Learning has already been introduced in Section 2.3.3, in the following we describe in details the power allocation scheme proposed in this paper for each sub-channel.

In our system, the multiple agents with learning capabilities are the ABSs, so that for each sub-channel they are in charge of identifying the current environment state, select the action based on the Q-learning methodology and execute it. In the following, for each agent $i = 1, 2, \ldots, N$ and sub-channel $r = 1, 2, \ldots, R$ we define system state, action, associated cost, next state and docition. To simplify the notation we will refer in the following to a system with only one sub-channel.

1. **State.** The system state for agent i consists of two parts: the individual state of agent i and the global state of the self-backhaul capacity, to achieve both the capacity constraints (4.6). In particular, at time t the state for ABS i is defined in the Table 4.1.

   To parameterize the global state of the self-backhaul, C^H is normalized with respect to the ideal maximum upper bound capacity C^H_{ub}. C^H_{ub} is computed equivalently to C^H but assuming the interference terms to be negligible. As a result, we define the normalized capacity of the system's self-backhaul as:

   $$\widehat{C}^H = \frac{C^H}{C^H_{ub}} \tag{4.11}$$

   We will consider that this normalized parameter can take four possible values: "High" i.e., above 0.7; "Mid-High" i.e., between 0.7 and 0.5; "Mid-low" i.e., between 0.5 and 0.25 and "Low" i.e., for values below 0.25. In addition, we consider two different situations: the self-backhaul capacity to $ABS\ i$ is above or below the capacity of the $ABS\ i$ to the MSs. However, even if the local constraint is fulfilled and the capacity of $ABS\ i$ to its users $\widehat{C^{i,A}}$ is below the self-backhaul capacity to $ABS\ i$, $C^{i,A}$, there may be a waste of self-backhaul capacity, and ideally $\widehat{C}^{i,A}$ should be equal to $C^{i,A}$. For this reason, a third situation is considered, i.e., $C^{i,A}$ is 10% above $\widehat{C^{i,A}}$, which we will consider as the most efficient state.

2. **Actions.** The set of possible actions consists of: *maintain*, *increase* or *decrease* $p_r^{i,A}$ by one dBm.

3. **Cost.** The cost assesses the immediate return incurred due to the assignment of action a at state s. The aim of the learning algorithm is to find the optimal $p_r^{i,A}$

Table 4.1: Definition of states.

	High $\hat{C}^H$	*Mid-high* $\hat{C}^H$	*Mid-low* $\hat{C}^H$	*Low* $\hat{C}^H$
$\widehat{C^{i,A}} > C^{i,A}$	State 1	State 2	State 3	State 4
$\widehat{C^{i,A}} \simeq< C^{i,A}$	State 5	State 6	State 7	State 8
$\widehat{C^{i,A}} < C^{i,A}$	State 9	State 10	State 11	State 12

that maximizes the capacity C^H at the same time that (4.6) is satisfied. The cost function may be of the form:

$$c(s,a) = \frac{|C^{i,A} - \widehat{C^{i,A}}|^2}{C^H}. \tag{4.12}$$

The cost decreases while capacity C^H increases and $C^{i,A}$ gets closer to $\widehat{C^{i,A}}$. Nevertheless it may happen that $\widehat{C^{i,A}}$ remains infinitely close and above $C^{i,A}$, so that the constraint is not fulfilled, but the cost C is infinitely small. For this reason, we introduce some modifications in the expression of the cost. There are notably some states that are more desirable than others. In particular, states 1 to 4 are not convenient since here the system does not satisfy the capacity constraint. On the other hand, states 5 to 8 are more convenient since here the constraint is satisfied and $C^{i,A}$ is efficiently utilized by the MSs. As a result, we modify the cost function as follows:

$$C = \begin{cases} |100(C^{i,A} - \widehat{C^{i,A}})|^2\gamma_{ec} + \alpha_{ec} & \text{states 1 to 4} \\ \frac{|100(C^{i,A} - \widehat{C^{i,A}})|^2 - \alpha_d}{\gamma_d(1+\hat{C}^H)} & \text{states 5 to 8} \\ \frac{|100(C^{i,A} - \widehat{C^{i,A}})|^2}{1+\hat{C}^H} & \text{states 9 to 12} \end{cases} \tag{4.13}$$

Here, γ_{ec} and α_{ec} increase the cost of states 1 to 4, and γ_d and α_d decrease the cost of states 5 to 8.

4. **Next State.** The state transition from s to v is determined by the power allocation policy.

5. **Docition.** The docition implemented is one of the most simple. When an ABS is turned-on, it receives docition from and expert ABS that has already learnt the problem. In this case the information shared corresponds to the Q-values of the Q-table from the expert ABS. As a measure of expertness we have used a condition similarity metric. The metric of similarity is the distance between the ABS and the MU. Accordingly to the taxonomy presented in Section 3.2.2 the Docition implemented is *Single Learning* (because the docition is done once at the beginning), *system based* because the shared information contains data of all the states, since the information shared are the Q-values the diction is sharing *low level* information in form of *policies*

finally the docition is *Scenario conditions based* since is done between ABSs in similar conditions. Notice that we do not use docition in all the examples, when used it will be clearly mentioned. To maintain coherence with the literature, we further use the term Start-up Docition to name the before described docition paradigm.

With the Q-learning configuration and the Start-up Docition description at hand, we proceed to the discussion of the simulation results.

4.3 Simulation Results

4.3.1 Description and Motivation of Simulations

In the following sections we show the simulation results for different experiments. In Section 4.3.2 we analyze the architecture and cognitive algorithm solution proposed. Simulation results show the total capacity achieved by the joint access-backhaul link design proposed. Further, in Section 4.3.3 we focus in the cognitive algorithm proposed (decentralized Q-learning) and analyze its convergence and performance. Section 4.3.4 zooms in to a single agent, shown details of the cognitive algorithm dynamics and internal states. Docition is also implemented in this Section and compared to normal cognitive algorithm. Comparison is two fold, first in terms of convergence speed and secondly we pay attention to sustainability through the energy consumption of the learning and teaching process. The energy gains are translated to an interference reduction to primary users and to a low consumption of battery (in case of MSs). Both measures, speed of convergence and energy efficiency, are of capital importance in modern cellular systems.

4.3.2 Hight Capacity Achievements

We focused on a single beam, that consists of one HBS (shared by all the beams) and $N_{ABS} = 4$ ABS, placed in a 2 by 2 matrix form. The separation between ABSs is 100m, and the HBS is located 350 m away from the ABS distribution center. We consider one MS for each ABSs, which is randomly located within a 75 m radius coverage of each ABSs. For the whole system we consider that the different ABSs distributions are located equidistantly from the HBS forming a circumference. The distance between the center of a distribution and the two adjacent ones is 350 m.The important parameters configuring the learning algorithm are the number of iterations, i.e., 200000, $\alpha = 0.5$, $\gamma = 0.9$ and the initial values of the Q-table i.e, 10000.The extra cost parameters are: $\gamma_{ec} = 1.4(1 + \frac{r_{1,4}}{r_T})$ and $\alpha_{ec} = 1 + \frac{r_{1,4}}{r_T}$, where $\frac{r_{1,4}}{r_T}$ is the ratio of number of visits to states 1 to 4 to the total number of visits to all the states. The discount parameters are: $\gamma_d = 1.43$ and $\alpha_d = -0.57$.

We further assume that the coverage area of an ABS distribution is circular with radius 145m and hence a surface of $0.67km^2$ which is a bit larger than the real one.

The simulation results for 50 trials show that on average the bandwidth required to achieve $1Gbps/Km^2$ (see Section 1.4)is $15.1MHz$ (see Figure 4.4), with this bandwidth the total capacity achieved by the HBS is 4 Gbps (all beams). The mean capacity of the sum

access link (ABS to the MSs) is $5.1 bps/hz$ and the mean backhaul capacity is $5.4 bps/hz$. Figure 4.3 shows the capacity in bps/hz of the access link and the backhaul link. In Figure 4.5 the total capacity for a HBS is shown when using a bandwidth of $40 Mhz$. Likewise Figure 4.6 shows the capacity density in terms of $Gbps/Km^2$ when using a bandwidth of 40 Mhz.

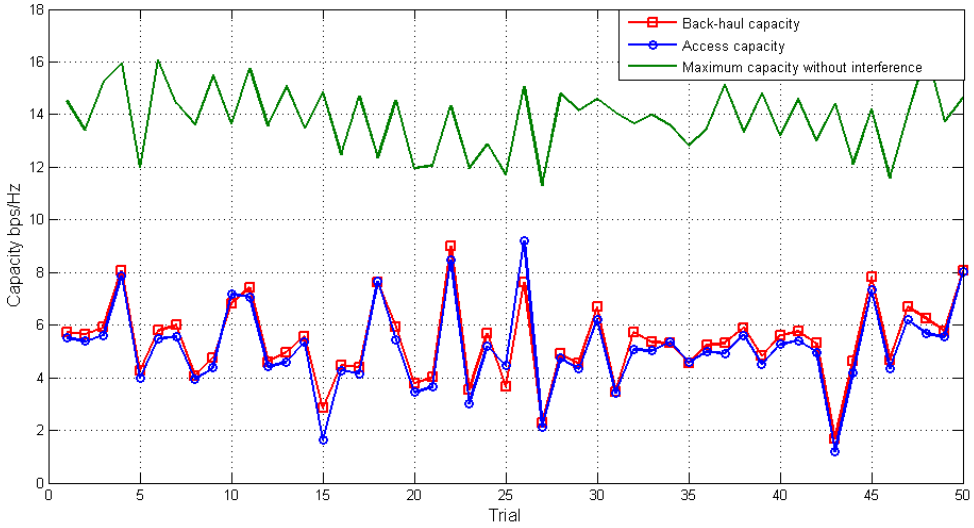


Figure 4.3: Capacity of the System in Gbps/Hz

4.3.3 Convergence Results

The scenario considered for evaluation consists of one HBS and $N_{ABS} = 16$ ABSs in one beam, placed in a 4×4 matrix form. The separation between ABSs is 350 m and the HBS is located at 1350 m in the y direction from the center of the ABSs distribution. We consider one MS q for each ABS, which is randomly located within a 200 m radius coverage area of each ABS. The downlink transmission power p_r^H of the HBS is fixed at 46 dBm; additionally the downlink transmission powers $p_r^{i,A}$ of the ABSs can be adaptively fixed between 1 dBm and 31 dBm at a 1 dBm step.

The considered Q-learning passes through 3 different stages: high exploration stage (20% random decisions) during the first third of iterations; a low exploration stage (10% random decisions) during the second third; and finally an exploitation stage with no random decisions. The other important parameters of the Q-learning configuration are the same as in Section 4.3.2.

Figure 4.7 shows the number of ABSs satisfying the constraint (4.6). After a sufficient number of iterations, 15 and later 16 ABSs have learned the optimal decision policy.

Figure 4.8 compares the total access capacity of the MSs allocated by the N_{ABS} ABSs, normalized with respect to the maximum upper bound capacity of the self-backhaul C_{ub}^H,

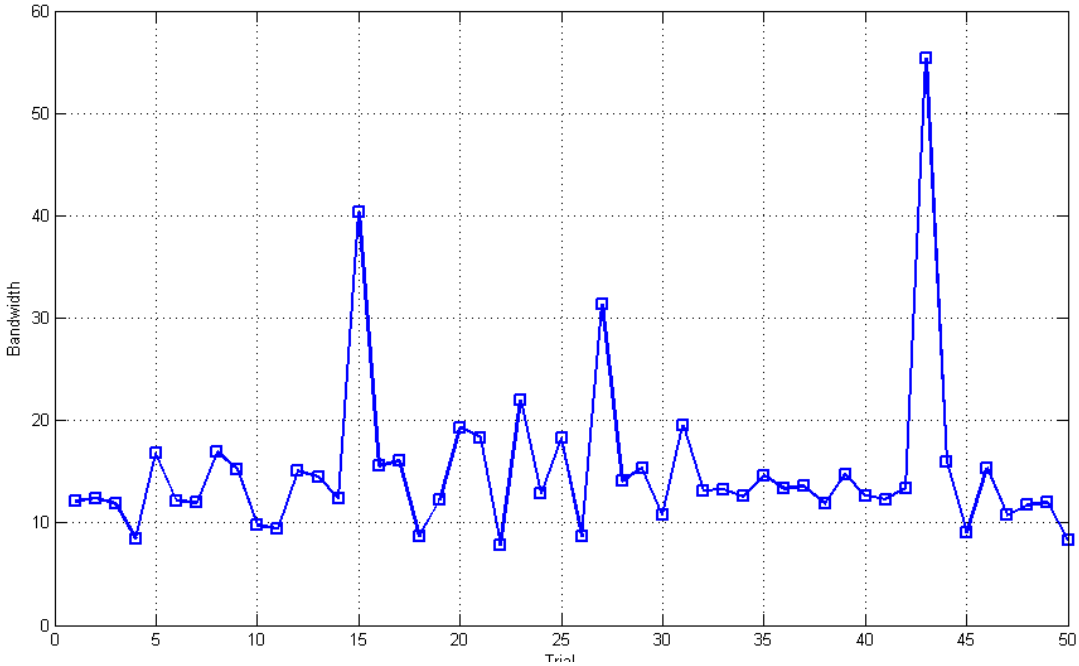


Figure 4.4: Bandwidth MHz required to achieve $1Gbps/Km^2$

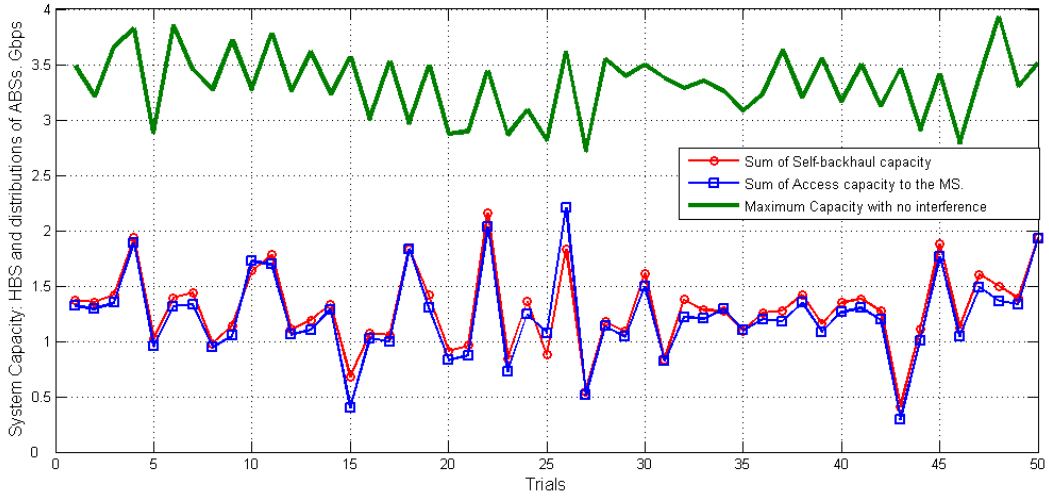


Figure 4.5: System Capacity assuming a bandwidth of 40Mhz.

i.e., $\frac{1}{C_{ub}^{H}} \sum_{i=1}^{N} \widehat{C^{i,A}}$, to the total normalized capacity of the self-backhaul link $\widehat{C^{H}}$. It can be observed that the self-backhaul link is about 6 % larger than the access link capacity, which indicates that the designed decentralized system operates effectively.

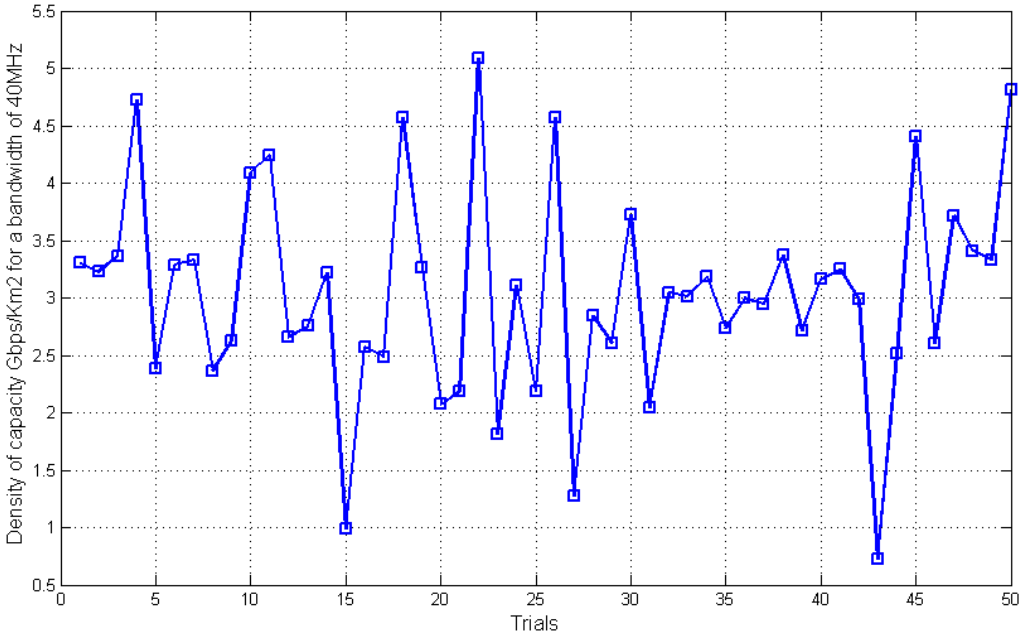


Figure 4.6: Total capacity density ($Gbps/Km^2$) of the system for 40Mhz Bandwidth.

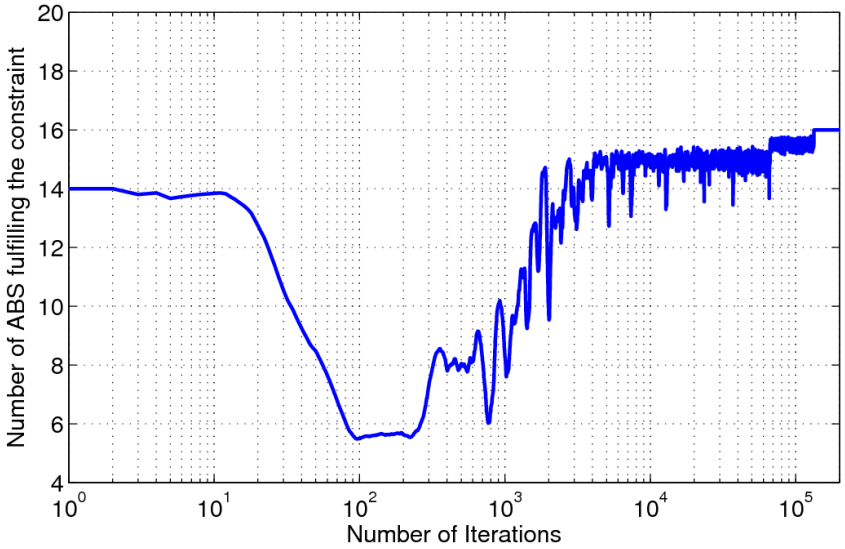


Figure 4.7: Number of ABSs satisfying the constraints (4.6).

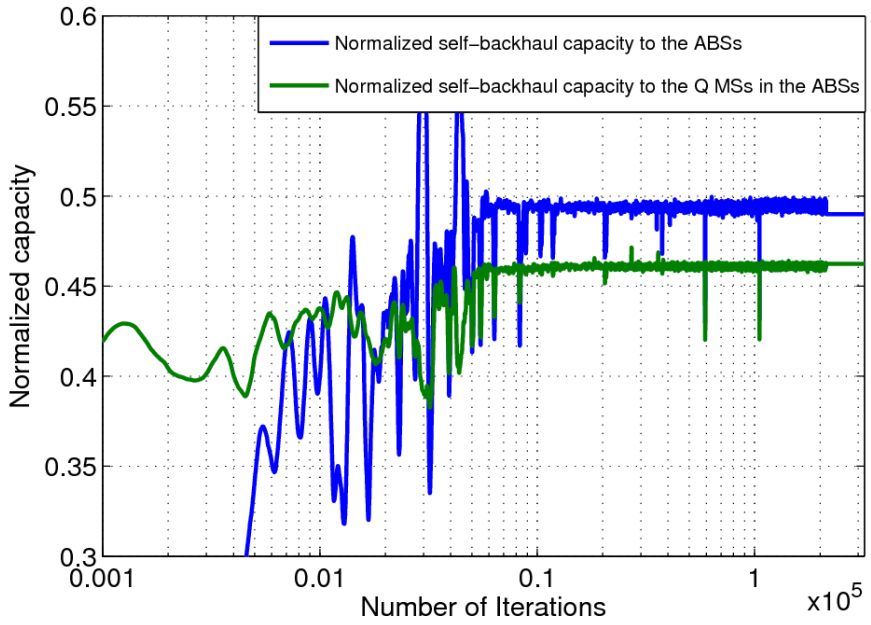


Figure 4.8: Normalized total self-backhaul capacity $\hat{C}^H$ compared to the sum of the normalized capacity used for the Q MS allocated in all the ABSs.

4.3.4 Single Agent Results

Cognitive algorithm

The structure of the ABSs distribution and the Q-learning parameters are the same as in Section 4.3.3. In Figure 4.9, we show the normalized error between $\widehat{C^{i,A}}$ and $C^{i,A}$, defined as:

$$\epsilon = \frac{C^{i,A} - \widehat{C^{i,A}}}{C^{i,A}}. \tag{4.14}$$

It can be observed that the normalized error is reduced by the learning algorithm, and the oscillations around zero are reduced whilst increasing the number of iterations. Finally, some iterations after the beginning of the exploitation stage, the error remains constant.

Figure 4.10 shows the instantaneous cost value of one ABS, e.g., ABS 4. The cost is reduced whilst the iterations increase to finally remain level off. There is a clear correspondence between the cost plot and the normalized error plot in Figure 4.9. For instance, at iteration 1000 a negative peak can be observed in Figure 4.9, which translates into a stage with a large cost in Figure 4.10.

In Figure 4.11, the average values of the cost $c(s, a)$ are shown by states. Here, we can easily distinguish among three groups of states: states 6 and 7 with a small (negative) average cost, states 10 and 11 with an intermediate average cost, and states $2, 3$ and 9 with large average costs. This is consistent with the cost function definition in 4.13.

Finally, Figure 4.12 shows the final Q-values $Q(s, a)$. Notice that the states with small-

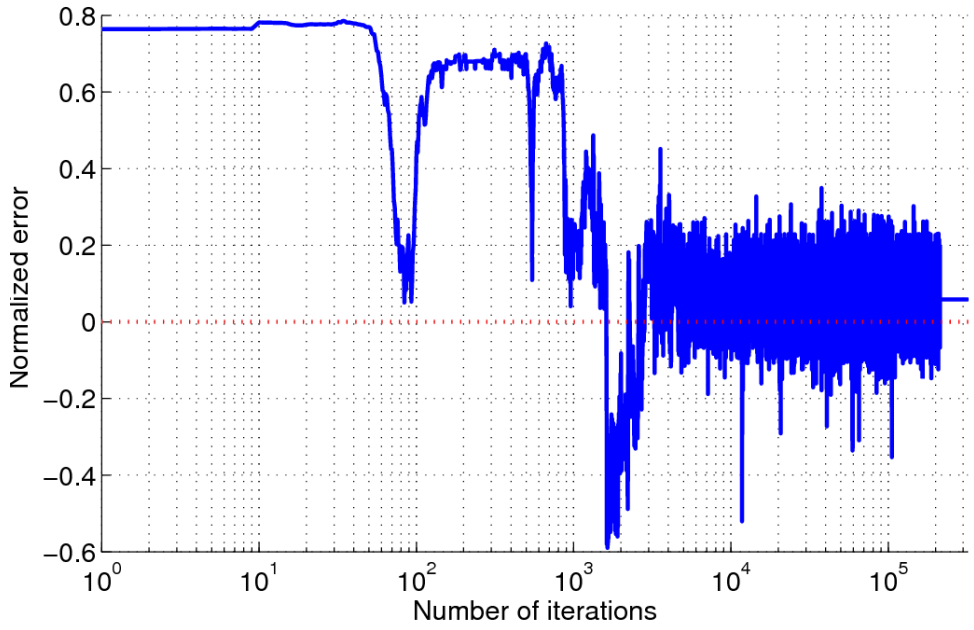


Figure 4.9: Normalized error for ABS 4.

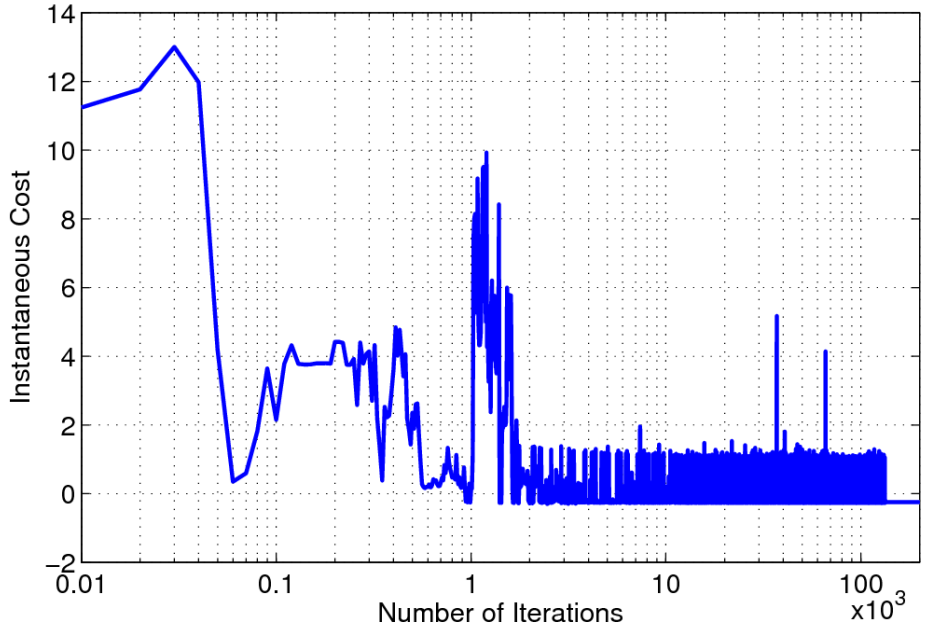


Figure 4.10: Instantaneous Cost for ABS 4.

est Q-values are states 6, 7, 10 and 11. These correspond to the states where the capacity constraints are accomplished and the $\widehat{C}^H$ achieved is between 0.7 and 0.25 (see Table 4.1).

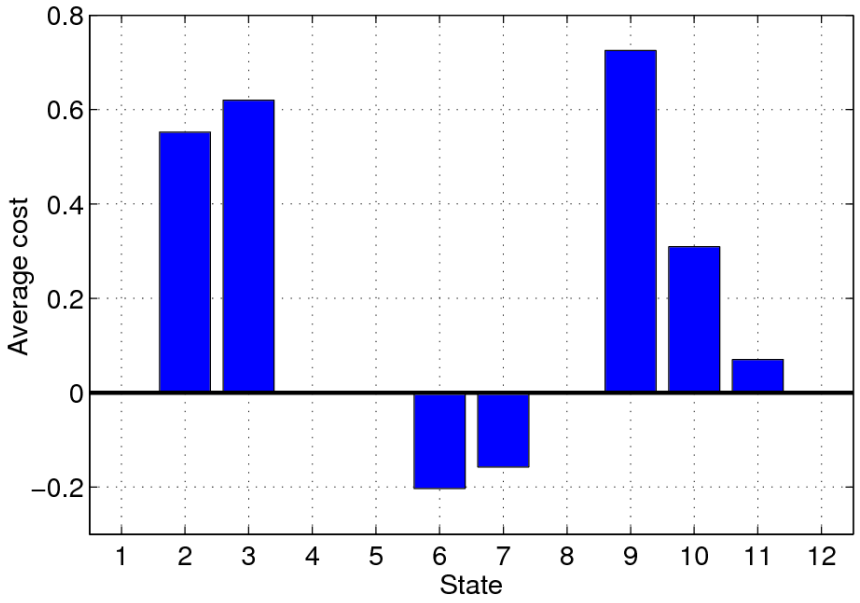


Figure 4.11: Average cost value for each state, for ABS 4.

What is noteworthy is that the two state-action pairs with very small Q-values correspond to states 6 and 7 with the *maintain* power action. This is consistent with the fact that once the ABS has reached a stationary behavior and found the optimal policy, the action most likely to maintain the system in the same conditions is to *maintain* the power at the selected level.

Benefits of Docition

The structure of the ABSs distribution is very similar to the one studied in Section 4.3.3, in this case the number of ABSs N_{ABS} is 4 placed in a 2x2 matrix form.

We assume that 3 ABSs learns and operate in the scenario according to the no docition paradigm. After a certain number of iterations, a new and inexpert ABS switches on, hence the whole systems has to autonomously and distributively reconfigure itself. Since the knowledge of the expert ABSs is large the adaptation time is short, however the new ABS is not expert and has to learn the optimal power allocation policy from scratch. When following the No Docition case the new ABS learns the power allocation policy without support of more expert nodes. On the other hand, when considering the Startup Docition scheme, the new ABS is taught the policy that a more expert ABS has already learnt. In particular, the Q-table of an expert node is sent to the new ABS.

Since in the No Docition case the new ABS has to learn alone the exploration rate (probability of random action selection) is set to 10% in the Startup Docition case the exploration rate is set to 1%. The important parameters configuring the learning algorithm

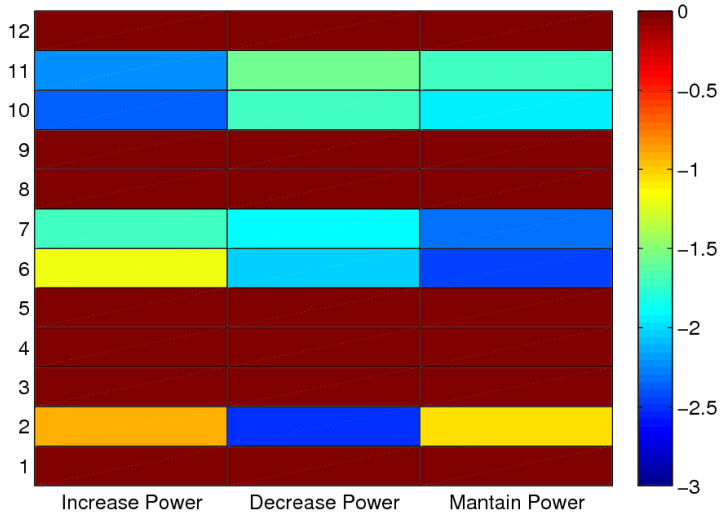


Figure 4.12: Final Q-values for ABS 4.

are the number of iterations, i.e., 200000, new ABS switch on time, i.e., iteration 100000, $\alpha = 0.5$, $\gamma = 0.9$ and initial Q-values i.e, 10000. The extra cost parameters are: $\gamma_{ec} = 1.2(1 + \frac{r_{1,4}}{r_T})$ and $\alpha_{ec} = 10^{-4}(1 + \frac{r_{1,4}}{r_T})$, where $\frac{r_{1,4}}{r_T}$ is the ratio of visiting states 1 to 4 and the total number of visited states. The discount parameters are: $\alpha_d = 1.14 \cdot 10^{-4}$ and $\gamma_d = 1.14$.

In Figure 4.13, we show the normalized error (see Equation 4.14). It can be observed that the normalized error is reduced by the learning algorithm and the oscillations are reduced whilst increasing the number of iterations. In the Startup Docition case, the normalized error decreases rapidly at the same time as the ABS state goes to the desired ones (the two black doted horizontal lines mark the error threshold for the desired states). The oscillations in the performance are clearly smaller in the case of Startup Docition. We defined the convergence time t_{conv} as the first time that the ABS is in the desired states in 80% of the time (iterations). We observe that t_{conv} for No Docition is 20200 and for the Startup Docition 4300; this thus yields a gain of almost an order of magnitude.

In Figure 4.14 the energy consumption of a new ABS for downloading files of different sizes is shown. The energy consumption in the Startup Docition case is smaller than in the No Docition case. The benefits in the case of docition are significantly higher for small file sizes since the No Docition ABS is not able to learn the power allocation policy in the short downloading time. However, when the file size becomes larger and thus the downloading time, the energy consumption of both gets closer since the No Docition ABS has already learned part of the optimal power allocation policy.

Docition yields better performance of the learning process at the expense of cooperation and thus an overhead. This overhead is however negligible, since it only consists in the transmission of a Q-table (e.g. 288 bytes) when a new learning node switches on.

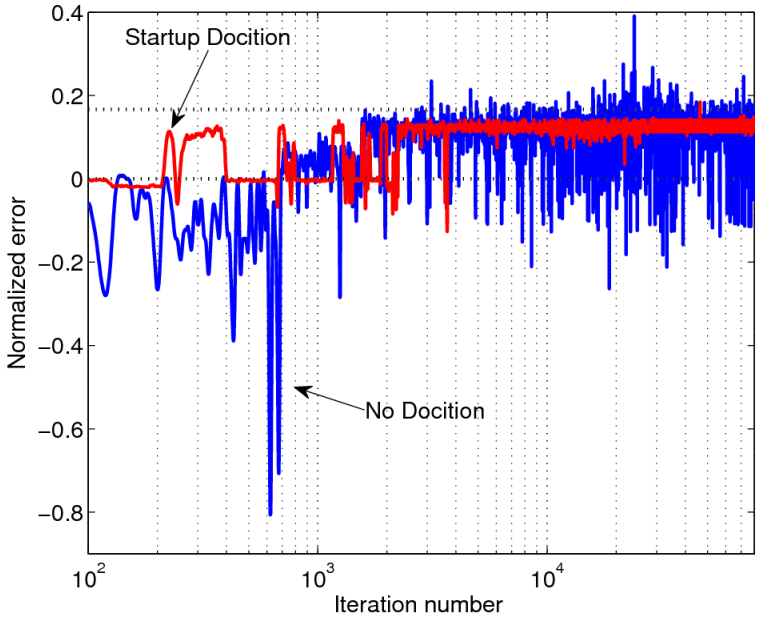


Figure 4.13: Cognitive and docitive normalized error comparison.

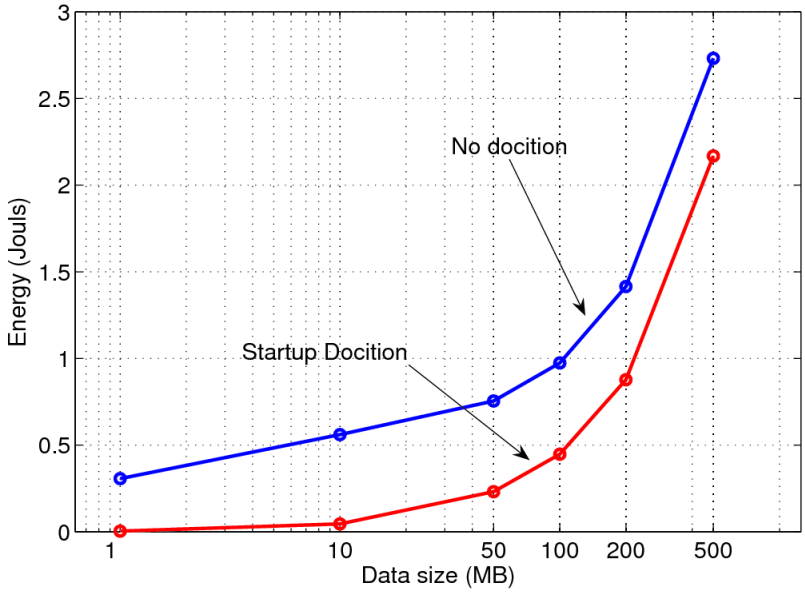


Figure 4.14: Energy comparison of cognitive versus docitive file downloads.

Chapter 5

Conclusions and Outlook

5.1 Conclusions

The prime of this Master Thesis was to prove that a joint access-backhaul link design, where access and backhaul utilize the very same time-frequency resources, combined with docitive-cognitive intelligence is a possible solution for decentralized and self-organizing networks. This is seen as a facilitator for Gbps/km^2 as the number of ABSs can be easily scaled with the capacity needs without jeopardizing operational conditions. Docition has been introduced and at the same time the different docition paradigms have been classified concerning on its teaching essences, leading the way for future research on this topic. Additionally we have focused on the capability of self-organization of the ABS nodes, which implement a state of the art RL scheme, the decentralized Q-learning, and the novel introduced paradigm of docition. Based on this algorithm, the learning ABSs set their transmission parameters in order to efficiently operate in the same frequency band as the other access nodes and the self-backhaul link. We have actually shown that in the proposed architecture, the expert ABSs reached the 1Gbps/km^2 occupying 15MHz of spectrum. The training process required to operate in an efficient manner is however very long, which is a typical limitation of all distributed learning schemes, thus making truly cognitive approaches not sustainable in a real wireless setting. We showed that if the more expert nodes are willing to share their expert knowledge with other nodes, at the expense of marginal signalling overhead transmitted over the air, the energy consumption of the access base stations in the training phase can be reduced but even more important the speed of convergence is dramatically increased. Docition thus emerges as a capital paradigm for the design of decentralized and self-organizing networks since is able to speed-up the inescapable learning process.

5.2 Future Work

In the following we present three different future work proposals for the docitive-cognitive networks.

5.2.1 Cognitive-Docitive Algorithms for Channel Allocation

The architectural design proposed in this thesis, uses the same frequency and time for the access and backhaul and thus the proposed cognitive interference control algorithm was absolutely necessary. Cognitive-Docitive algorithms may be also applied to other aspects of the cellular systems i.e. to find out optimal channel allocation policies in OFDM systems. There are plenty of opportunities for intelligent algorithms and thus also for docition in the future cellular communications.

5.2.2 Mathematical Analysis of Docition and Cognition

A general mathematical framework is necessary for cognition and docition. Then one will be able gain detail insight the learning dynamics, understand more about its, apparently random, fluctuations, quantify its performance, expertness and intelligence. With this information at hand, it will be possible to better understand the hidden mechanism of docition and cognition. For example intelligence gradients can be established where docition should primarily happen along the strongest gradient. Additionally, as another Example, analytical tools can also be used to proof that certain types of functions are optimal for certain tasks (i.e. optimal cost function, action selection function). Various disciplines such as Optimal Control, Functional Analysis and Dynamics Systems are related to the formal analysis of Cognition and Docition.

5.2.3 Replicator Dynamics and Decentralized RL in Wireless

RL is able to learn the environment dynamics and get advanced to the systems changes. Q-learning is one of the most promising RL algorithms and additionally, besides its convergence is not proof, is able to work in a fully decentralized fashion. The major drawbacks are the low convergence speed of the decentralized version and the almost imperative necessity of using heuristic trial and error methods to set up the algorithm and the absence of any analytical tool to gain deep insight into its dynamics. In [43],[36] and [10] is found that in the continuous time limit the RL equations are the same as the replicator dynamics equations plus a mutation term. Hence EGT may be used to predict the behavior of the RL and to enhance its performance by means of more efficient parameter tuning. However multiagent wireless scenario is more complex than the scenarios considered in the nowadays literature since in wireless there are usually several numbers of players interacting in a dynamic environment.

In the flowing there are some facts that have to be considered before start working in the problem

- The RL algorithm usually work with states so a Markov Decision Evolutionary Game MDEG [6] should be used. The states may not be related to the battery level [5] but to the general state of the network and to the specific state of the player channel.

- The most suitable evolutionary game frame work is multiplayer interactions from several different populations.
- Each player usually has more than two actions.
- It may be worth to work with a scenario description similar to the one in [5].

The solution to the differential equation system of the Replicator dynamics gives an accurate idea of the RL performance, stability of the solution and the path to reach it. In [23] the divergences between the replicator dynamics model and the RL performance are studied. It is also needed to study what properties are desired in the RD equations: do we need a unique equilibrium? or it may be better to have a hight performance at expenses of not finding an equilibrium? We also have to check if we really have or not a potential in the game. There are other dynamics that replicator dynamics, may be there are some that better fits the wireless problem (see [22]).

5.3 Published Work

- Journal and Magazines
  - L. Giupponi, A. Galindo, **P. Blasco**, M. Dohler, Docitive Networks -An Emerging Paradigm for Dynamic Spectrum Management, *IEEE Wireless Communications Magazine*, Vol. 17, No. 4, pp. 47-54, August 2010.
  - M. Dohler, L. Giupponi, A. Galindo, **P. Blasco**, Docitive Networks: A Novel Framework Beyond Cognition, *IEEE Communications Society, Multimdia Communications TC, E-Letter*, January 2010.
- Int'l conferences
  - **P. Blasco**, L. Giupponi, A. Galindo, M. Dohler, Energy Benefits of Cooperative Docitive over Cognitive Networks, *in Proceedings of the 3rd European Wireless Technology Conference 2010 in the European Microwave Week*, Sept 26 - Oct. 1, Paris, France.
  - **P. Blasco**, L. Giupponi, A. Galindo, M. Dohler, Aggressive Joint Access & Backhaul Design For Distributed-Cognition 1Gbps/km2 System Architecture, *in Proceedings of 8th International Conference on Wired/Wireless Internet Communications* (WWIC 2010), 1-3 June, 2010, Lulea (Sweden).
  - A. Galindo, L. Giupponi, **P. Blasco**, M. Dohler, Learning from Experts in Cognitive Radio Networks: The Docitive Paradigm, *in Proceedings of 5th International Conference on Cognitive Radio Oriented Wireless Networks and Communications* (CrownCom 2010), 9-11 June 2010, Cannes (France).

Bibliography

[1] Beyond next generation mobile broadband -description of work. Technical report, FP7-ICT-2009-4-248267, 2009.

[2] Berr, digital britain: The interim report. January 2009.

[3] [online] http://europa.eu/rapid/pressreleasesaction.do?reference=ip/09/142;, last accessed 24 March 2009.

[4] Majid Nili Ahmadabadi and Masoud Asadpour. Expertness based cooperative q-learning. *Transactions on Systems, Man and Cybernetics*, 32, 2002.

[5] Eitan Altman, Rachif El-Azouzi, Yezekael Hayel, and Hamidou Tembine. Evolutionary power control games in wireless networks.

[6] Eitan Altman and Yezekael Hayel. Markov decision evolutionary games. *Proceedings of the 13th Symposium on Dynamic Games and Applications*, 2008.

[7] Baum, Hansen, Del Galdo, Miljoevic, Salo, and Kyösti. An interim channel model for beyond-3G systems. *IEEE*, 2005.

[8] Richard E. Bellman and Stuart E. Dreyfus. *Applied Dynamic Programming*. Princeton University Press, 1962.

[9] Pol Blasco, Lorenza Giupponi, Ana Galindo-Serrano, and Mischa Dohler. Energy benefits of cooperative docitive over cognitive networks. *Proceedings of the 3rd European Wireless Technology Conference in the European Microwave Week*, 2010.

[10] T. Börgers and R. Sarin. Reinforcement and replicator dynamics. *J. Econ. Theory*, 77:1–14, 1997.

[11] Mischa Dohler, Lorenza Giupponi, Ana Galindo-Serrano, and Pol Blasco. Docitive netwroks: A novel framework beyond cognition. *IEEE Communications Society, Multimedia Communications TC, E-Letter*, 2010.

[12] V. Erceg, K. V. S. Hari, M. S. Smith, and D. S. Baum. Channel models for fixed wireless applications. *IEEE Broad Band Wireless Working Group, Tech. Rep.*, 21:139–150, 2001.

[13] Ana Galindo-Serrano and Lorenza Giupponi. Distributed q-learning for aggregated interference control in cognitive radio networks. *IEEE Transactions on Vehicular Technology*, 59:1823 – 1834, 2010.

[14] Ana Galindo-Serrano, Lorenza Giupponi, Pol Blasco, and Mischa Dohler. Learning from experts in cognitive radio networks: The docitive paradigm. *Proceedings of 5th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom 2010)*, 2010.

[15] Ana Galindo-Serrano, Lorenza Giupponi, and Mischa Dohler. Cognition and docition in ofdma-based femtocell networks. *in Proceedings of IEEE Global Communications Conference*, 2010.

[16] Lorenza Giupponi, Ana Maria Galindo, Pol Blasco, and Mischa Dohler. Docitive network- an emerging parading for dynamic spectrum management. *IEEE Wireless Communications Magazine*, 17:47–54, 2010.

[17] Lorenza Giupponi, Ana Galindo-Serrano, and Mischa Dohler. From cognition to docition: The teaching radio paradigm for distributed & autonomous deployments. *Computer Communications*, 2010.

[18] Jon Hamkins and Marvin K. Simon. *Autonomos Software-Defined Radio Recievers for Deep Space Applications*. Jet Propulsion Laboratory, 2006.

[19] S. Haykin. Cognitive radio: brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 23:201–220, 2005.

[20] Simon Haykin. *Neural Networks: A comprehensive foundation*. Tom Robbins, 1931.

[21] P. Hoen and karl Tuyls. Analyizing multi-agent reinforcement learning using evolutionary dynamics. *Procedings of the 15th European Conference on Machine Learning (ECML)*.

[22] Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40:479–519, 2003.

[23] Michael Kaisers and Karls Tuyls. Frequency adjusted multi-agent q-learning. *Conference on Autonomous Agents ans Multi-Agent Systems*, 10-14:309–315, 2010.

[24] Richard Maclin, Jude Shavlik, Lisa Torrey, Trevor Walker, and Edward Wild. Giving advice about prefered actions to reinforcement learners via knowledg-based kernel regression. *Procedings of theIAAA*, 2005.

[25] Richard Maclin, Jude Shavlik, Trevor Walker, and Lisa Torrey. Knowledge-based support-vector regresion for reinforcement learning. *Proceedings of the IJCAI'05 workshop on Reasoning, Representation and Learning in Computer Games. Edimburg Scotland,*, 2005.

[26] Olvi L. Mangasarian, Jude W. Shavlik, and Edward W. Wild. Knowledg-based kernel approximation. *Journal of Machine Learning Research*, 2004.

[27] Tom M. Mitchell. *Machine Learning.* Mc Graw-Hill, 1997.

[28] J. Mitola and Gerald Q. Maguire. Cognitive radio: Making software radios more personal. *IEEE Pcrsonal Communications*, 1999.

[29] Stephen muggleton and luc de raedt. Inductive logic programming: theory and methods. *Journal of Logix Programming*, 1994.

[30] K. P. Murphy. A survey of pomdp solution techniques. Technical report, U.C. Berkley, 2000.

[31] [Online:]. https://www.communicationsdirectnews.com//do.php/110/34476?199. last accessed 24 March 2009.

[32] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE transactions on knowledge and data engineering*, 2010.

[33] A. Potapov and M. k. Ali. Convergence of reinforcement learning algorithms and accelartaion of learning. *Physical Rewiev*, 67, 2003.

[34] Carlos H.c. Ribero. Embedding a priori knowledge in reinforcement learing. *Journal of Intelligent and Robotics Systems*, 21:51–71, 1998.

[35] Ze'ev ROTH, Mariana Goldhamer, Naftali Chayat, Alister Burr, Mischa Dohler, Nikolaos Bartzoudis, Chris Walker, Yigal Liebe, Claude Oestges adn Miroslaw Brzozowy, and Isabelle Bucaille. Vision and architecture supporting wireless gbit/sec/km^2 capacity density deployements.

[36] Yuzuru Sato and James P. Crutchfield. Coupled replicator equations for the dynamics of learing in multiagent systems. 2003.

[37] Alex J. Smola and Berhard Scholkopf. A tutorial on suppport vector regresion. 2003.

[38] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learing: An introducition.* MIT Press, Cambridge, MA, 1998.

[39] M. Tan. *Multi-Agent Reinforcement Learing: Independent vs. Cooperative Agents.* 1993.

[40] Matthew E. Taylor, Nicholas K. Jong, and Peter Stone. Transfering instances for model-based reinforcement learing. *Procedings of the ALAMAS+ALAG workshop at AAMAS*, 2008.

[41] Gerald Tesauro. Pricing in agent economies using neural networks and multi-agent q-learning.

[42] Lisa Torrey. *Relational Transfer in Reinforcement Learing.* PhD thesis, University of Wisconsin-Madison, 2009.

[43] Karl Tuyls, Pieter Jan'T Hoen, and Bram Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents ans Multi-Agent Systems*, 12:115–153, 2006.

[44] Christopher J.C.H. Watkins. *Learning from delayed rewards.* PhD thesis, University of Cambridge, Psychology Department., 1989.

[45] The ETS web page (2011). http://www.etsi.org.

[46] The BuNGee website (2011). http://www.ict-bungee.eu.

[47] W. Weibull, Jorgen. *Evolutionary Game Theory.* The MIT Press, Cambridge, Massachusetts, London, England., 1995.

[48] West, Jeremy, Dan Ventura, and Sean Warnick. Spring research presentation: A theoretical foundation for inductive transfer. *Brigham Young University, College of Physical and Mathematical Sciences*, 2007.

[49] Qing Zhao, Lang Tong, Anantharm Swami, and Yunxia Chen. Decentralized cognitive mac for opportunistic spectrum access in ad hoc networks: A pomdp framework. *IEEE Journal On Selected Areas in Communications*, 25, 2007.