Learning from Experts in Cognitive Radio Networks: The Docitive Paradigm

Ana Galindo-Serrano, Lorenza Giupponi, Pol Blasco and Mischa Dohler

***Abstract*—In this paper we introduce the novel paradigm of docition for cognitive radio (CR) networks. We consider that the CRs are intelligent radios implementing a learning process through which they interact with the surrounding environment to make self-adaptive decisions. However, in distributed settings the learning may be complex and slow, due to interactive decision making processes, which results in a non-stationary environment. The docitive paradigm proposes a timely solution based on knowledge sharing, which allows CRs to develop new capacities for selecting actions. We demonstrate that this improves the CRs' learning ability and accuracy, and gives them strategies for action selection in unvisited states. We evaluate the docitive paradigm in the context of a secondary system modeled as a multi-agent system, where the agents are IEEE 802.22 CR base stations, implementing a real-time multi-agent reinforcement learning technique known as decentralized Q-learning. Our goal is to solve the aggregated interference problem generated by multiple CR systems at the receivers of a primary system. We propose three different docitive algorithms and we show their superiority to the well know paradigm of independent learning in terms of speed of convergence and precision.**

***Index Terms*—Cognitive radio, aggregated interference, multi-agent system, decentralized Q-learning, docitive learning.**

I. Introduction

A cognitive radio (CR), as defined in [1], is an intelligent wireless communication system capable of using methodologies of understanding and learning to adapt its internal states and operating parameters to the dynamics of the surrounding environment. The most common application of this technology is of course the exploitation of the spectrum resource for a more efficient usage. Thanks to the advances in multiple disciplines, such as machine learning, this technology is indeed feasible today.

The artificial intelligence community proposes in literature numerous cognitive approaches capable of finding optimal decision policies in dynamical scenarios, characterized by only one decision maker, i.e., single-agent systems. One of them is reinforcement learning (RL), a particularly powerful learning technique that does not require environment models and allows nodes to take actions while they learn. Among reinforcement techniques, Q-learning [2] has been especially well studied, and possesses a firm foundation in the theory of Markov decision processes (MDPs). However, the wireless setting in general and the CR scenario in particular are not always characterized by a node centralizing the radio resource management (RRM) decision process, as it is the case e.g., for traditional cellular networks.

The authors are with the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Parc Mediterrani de la Tecnologia, Av. Carl Friedrich Gauss 7, Barcelona, Spain 08860; phone: +34-93-645-2900, fax: +34-93-645-2901, email: {ana.maria.galindo}@cttc.es.

As a matter of fact, there has been lately a clear trend towards decentralizing RRM functionalities, a representative example being the IEEE P1900.4 standardization effort [3]. Such a decentralized setting can be mapped onto the framework of a multi-agent system, and modeled by means of a stochastic game, of which a MDP would be a special case considering a single agent. Q-learning can still be applied in this case, in the form of the so called decentralized Q-learning. Each node learns here independently from the other nodes, which are assumed to be part of the surrounding environment (i.e., paradigm of independent learning). However, with such an hypothesis, the environment is no longer stationary, since it consists of other agents who are similarly adapting. This may generate oscillating behaviors that not always reach an equilibrium and that are not yet fully understood - even by machine learning experts. The dynamics of learning may be long and complex in terms of required operations and memory, with complexity increasing with an increasing observation space. A possible solution to mitigate this problem, speed up the learning process and create rules for unseen situations, is to propose expert knowledge exchange among learners [4][5].

Even if the process of learning has received a considerable attention from various research communities in the past, the process of knowledge transfer, i.e., teaching, over the wireless medium, however, has received fairly little attention to date. In human society, one learns from others' examples, experiences and activities by taking advice or consulting with each other. In machine learning words, people cooperate to learn. As a result of that, similarly to human beings, CRs are not required to learn everything from their own experiences, but in case of being unable to represent correctly their knowledge or to observe the surrounding environment, they can exchange expert information with peers in order to improve the learning process. We thus aim at introducing an emerging framework referred to as docitive radios, from "docere" = "to teach" in Latin, which relates to radios (or general networking entities) which teach other radios by coordinating with them. This paradigm will be shown to capitalize on the advantages but, most importantly, mitigate major parts of above-mentioned drawbacks of purely CRs.

Whilst applicable to a variety of decentralized problems in communications, in this paper we will focus on managing the aggregated interference at the primary receiver of a digital television (DTV) system generated by multiple IEEE 802.22 wireless regional area networks (WRAN) cells. We will map this scenario onto a multi-agent system, whose objective is that the multiple agents distributively learn an optimal strategy to control the aggregated interference at the primary receivers generated from the secondary systems and to maintain it under

Digital Object Identifier: 10.4108/ICST.CROWNCOM2010.9173
http://dx.doi.org/10.4108/ICST.CROWNCOM2010.9173

a given threshold. We propose a solution to this problem through a form of multi-agent RL, known as distributed Q-learning. We will apply said docitive techniques to demonstrate the superiority of docitive over cognitive techniques already introduced in [6].

The outline of the paper is organized as follows. In Section II, we describe the system model. In Section III we present the distributed Q-learning algorithm where no docition is considered and the nodes follow the well known paradigm of independent learning. Then we modify this algorithm in order to introduce different docitive techniques, which are aimed at improving the precision and stability of the learning process. Section IV describes the simulation scenario and Section V presents relevant simulation results showing the superiority of docition with respect to classical cognition. Finally, we summarize the conclusions in Section VI.

II. System Model

Following the IEEE 802.22 standard specifications, our primary system is characterized by a DTV broadcasting station (hereafter primary base station (BS)) located in the center of a circular cell, and several DTV receivers (hereafter primary receivers) randomly located in the DTV cell coverage area. In the TV bands the primary receivers do not transmit, and only receive signals from the primary BS with an omnidirectional antenna. Therefore the secondary WRAN cells have to be placed far enough from the primary receivers so that they do not cause harmful interference to the primary receivers.

The secondary WRAN cell, with radius r_{SS}, consists of a centralized architecture with point-to-multipoint wireless air interface whereby a secondary BS manages the medium access of the associated M secondary users. Since the IEEE 802.22 standard is orthogonal frequency division multiple access (OFDMA)/ time-division-multiple-access (TDMA) based, we consider the simplifying hypothesis that the WRAN BS assigns at any time all the available OFDMA resource blocks to one secondary user, which results in only one secondary user transmitting at any time in each WRAN cell.

Considering that the secondary system is unaware of the position of the passive primary receivers, the secondary users have to operate far enough from the primary BS in order not to cause harmful interference with the primary system. In [7] a DTV protection contour around the primary BS is defined as a geographical limit where the primary receivers must not receive harmful interference. Based on [7][8], the protection contour is at $R_{PC} = 134.2$ Km from the primary BS. In addition to this, in [7][8], also a *keep-out* region is defined, as a protection region where secondary users are not allowed to opportunistically transmit. This protection region is characterized by different radius, depending on the number of interfering secondary cells and on whether upstream or downstream transmission is considered. Our approach differs from those encountered in literature, in fact, to ease the future deployment of a secondary system, we consider that the radius of the *keep-out* region R_{KO} is the same as the one of the protection contour, so that $R_{KO} = 134.2$ Km. In addition, we propose a smart algorithm for secondary users

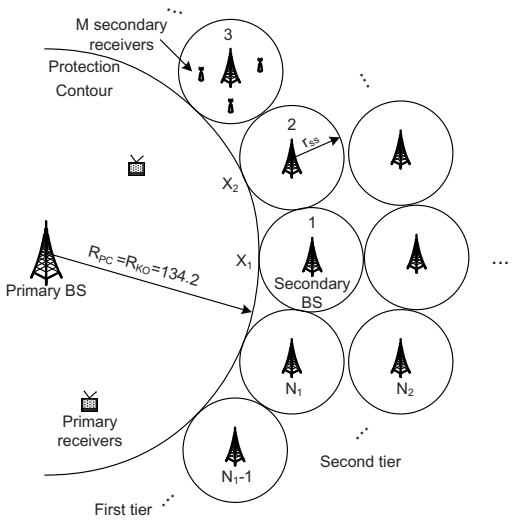

Fig. 1. Docitive scenario composed by a DTV cell and WRAN cells.

power allocation, capable of guaranteeing that the aggregated interference at the primary receivers is maintained below a given threshold. This scenario is depicted in Figure 1, where the secondary system consists of $N = \sum_n N_i$ secondary cells located around the boundary of the the *keep-out* region. In the following we will refer to the tangent point where the i_{th} secondary BS coverage area and the protection contour intersect, as the control point X_i. It is assumed that the i_{th} secondary BS is capable of measuring the interference at its control point X_i.

III. Learning and Docition

The characteristics of the CR network are as follows: (1) the intelligent decisions are made by multiple intelligent and uncoordinated nodes; (2) the nodes partially observes the overall scenario; and (3) their inputs to the intelligent decision process are different from node to node since they come from spatially distributed sources of information. This CR network can be easily mapped onto a multi-agent system, where each secondary BS is an independent intelligent agent. To formulate a mathematical framework for such an environment, we have to account for: (1) a state space that is the product of the individual agents' states; (2) state transitions that are functions of joint actions taken by the agents; (3) revenue to individual agents that depend on joint actions as well. The theoretical framework is found in the so called stochastic games [9] described by the five-tuple $\{\mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}\}$, where:

- $\mathcal{N}$ is the set of agents, i.e., the secondary BSs, indexed $1, 2, \cdots, n$;
- $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$ is the set of possible states;
- $\mathcal{A}$ is the joint action space defined by the product set $\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$, where $\mathcal{A}_i = \{a_1^i, \cdots, a_n^i\}$ is the set of actions available to the i_{th} agents;
- $\mathcal{P}$ is a probabilistic transition function, defining the probability of migrating from one state to another provided the execution of a certain joint action;

Digital Object Identifier: 10.4108/ICST.CROWNCOM2010.9173
http://dx.doi.org/10.4108/ICST.CROWNCOM2010.9173

- $\mathcal{C} = c^1 \times c^2 \times \cdots \times c^n$, where c^i is the cost of the i_{th} agent, which is a function of the joint actions of all n nodes.

The $\mathcal{N}$ agents have thus to distributively learn an optimal policy to achieve the common objective of maintaining the aggregated interference at the primary receivers below a threshold. Known as "multi-agent learning" problem, it can be solved by means of distributed RL approaches, when the probabilistic transition function cannot be deduced. There exist several RL algorithms. For our particular problem, we consider the decentralized Q-learning algorithm as an accurate algorithm to implement the learning process of a CR. However, in this field many problems still remain open. The main challenge is how to ensure that individual decisions of the nodes 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 nodes.

In principle, it is possible to treat the distributed CR network as a centralized one, where each node has complete information about the other nodes and learns the optimal joint policy using standard RL techniques. However, both the state and action spaces scale exponentially with the number of nodes, rendering this approach infeasible for most problems. The Nash-Q learning algorithm proposed by Hu and Wellman [10], extends the Q-learning to multi agent domain, by taking into account the joint actions of the participants, so that the experiences of both the learner itself and all the other nodes have to be uploaded. This creates again scalability problems in the wireless setting, although the algorithm is shown to converge to a Nash equilibrium with probability 1, under some conditions. Alternatively, we can let each node learn its policy independently of the other nodes, 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 [11]. This introduces game-theoretic issues to the learning process, which are not yet fully understood [12].

As a solution to these problems, we propose a distributed approach where nodes share potentially differing amounts of intelligence acquired on the run. This 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 infrastructure-less 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. In the rest of this section, we first describe with details the decentralized Q-learning algorithm following the paradigm of independent learners, according to which each agent learns independently, without taking into account other agents and treating them as part of the environment and not as responsive agents. Then, we propose different modifications of it based on the docitive paradigm.

A. Decentralized Q-Learning: Paradigm of Independent Learning

It is assumed that the environment is a finite-state, discrete-time stochastic dynamical system. The interactions between the multi-agent system and the environment at each time instant t consist of the following sequence.

- The agent i senses the state $s_t^i = s \in \mathcal{S}$.
- Based on s, agent i selects an action $a_t^i = a \in \mathcal{A}_i$.
- As a result, the environment makes a transition to the new state $s_{t+1}^i = v \in \mathcal{S}$.
- The transition to the state v generates a cost $c_t^i = c \in \mathbb{R}$, for agent i.
- The cost c is passed back to the agent and the process is repeated.

The objective of each agent is to find a policy $\pi^*(s) \in \mathcal{A}_i$ for each s, to minimize some cumulative measure of the cost $c_t^i = c(s,a)$ received over time. We define an evaluation function, denoted by $Q(s,a)$, as the expected total discount cost counting over an infinite time. It is given by [13]:

$$Q(s,a) = \mathbb{E}\left\{\sum_{t=0}^{\infty} \gamma^t c(s_t, \pi(s)) | s_0 = s\right\}, \qquad (1)$$

where $\mathbb{E}$ stands for the expectation operator and $0 \leq \gamma < 1$ is a discount factor. If the selected action a at time t following the policy $\pi(s)$ corresponds to the optimal policy $\pi^*(s)$ the Q-function is minimized with respective to the current state.

Let $P_{s,v}(a)$ be the transition probability from state s to next state v, when action a is executed. Then, eq. (1) can be expressed as [13]:

$$\begin{aligned} Q(s,a) &= \mathbb{E}\{\ c(s_0,a_0))|s_0 = s, a_0 = a\} + \\ &\quad \mathbb{E}\left\{\sum_{t=1}^{\infty} \gamma^t c(s_t,a_t)|s_0 = s, a_0 = a\right\} \\ &= \mathbb{E}\{\ c(s,a))\} + \gamma \sum_{v \in \mathcal{S}} P_{s,v}(a) \times \\ &\quad \mathbb{E}\left\{\sum_{t=1}^{\infty} \gamma^{t-1} c(s_t,a_t)|s_1 = v, a_1 = b\right\} \\ &= C(s,a) + \gamma \sum_{v \in \mathcal{S}} P_{s,v}(a) Q(v,b), \qquad (2) \end{aligned}$$

where $C(s,a) = \mathbb{E}\{c(s,a)\}$ denotes the expected value of $c(s,a)$. Eq. (2) indicates that the Q-function of the current state-action pair can be represented in terms of the expected immediate cost of the current state-action pair and the Q-function of the next state-action pairs.

The principle of Bellman's optimality assures that, for single agent environments, there is at least one optimal stationary policy π^*. The optimal value of state s is given by [2]:

$$\begin{aligned} V^*(s) &= V^{\pi^*}(s) \\ &= \min_{a \in \mathcal{A}_i} \left[C(s,a) + \gamma \sum_{v \in \mathcal{S}} P_{s,v}(a) V^*(v) \right]. \qquad (3) \end{aligned}$$

In multi-agent settings where each agent learns independently from the other agents, we approximate the other agents as part of the environment, and we still can apply the Bellman's

Digital Object Identifier: 10.4108/ICST.CROWNCOM2010.9173
http://dx.doi.org/10.4108/ICST.CROWNCOM2010.9173

criterion. In this case, the convergence to optimality proof does not hold strictly, but this independent learning approach has been shown to correctly converge in multiple applications [6][11][14]. Applying the Bellman's criterion, we have to find an intermediate minimal of $Q(s,a)$, denoted by $Q^*(s,a)$, where the intermediate evaluation function for every possible next state-action pair (v,b) is minimized and the optimal action is performed with respect to each next state v. $Q^*(s,a)$ is given by:

$$Q^*(s,a) = C(s,a) + \gamma \sum_{v \in \mathcal{S}} P_{s,v}(a)[\min_{b \in \mathcal{A}_i} Q^*(v,b)]. \quad (4)$$

Then, we can determine the optimal action a^* with respect to the current state s. In other words, we can determine π^*. Therefore, $Q^*(s,a^*)$ is minimal, and can be expressed as:

$$Q^*(s,a^*) = \min_{a \in \mathcal{A}_i} [Q^*(s,a)]\,. \quad (5)$$

The Q-value $Q(s,a)$ represents the expected discounted cost for executing action a at state s and then following policy π thereafter. The task of Q-learning is to determine a π^* without knowing $C(s,a)$ and $P_{s,v}(a)$, which makes it well suited for the learning power allocation in CR systems.

The Q-learning process tries to find $Q^*(s,a)$ in a recursive manner using available information (s,a,v,c), where s and v are the states at time t and $t+1$, respectively; and a and c are the action taken at time t and the immediate cost due to a at s, respectively. The Q-learning rule to update the Q-values relative to agent i is:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[c + \gamma \min_a Q(v,a) - Q(s,a)], \quad (6)$$

where α is the learning rate. For more details about RL and Q-learning the reader is referred to [2][15].

B. Docitive Algorithms

As for teaching approaches, some early contributions in machine learning literature [5][4] suggest that the performances of a decentralized learning system can be improved by using cooperation among learners in a variety of ways. A node e.g., can take advantage of the exchange of information and expert knowledge from other nodes [5], the so-called docitive nodes. Depending on the degree of docition among nodes, we consider in this paper the following cases:

- **Startup Docition.** Docitive radios teach their policies to any newcomers joining the network. In this case, again, each node learns independently; however, when a new node joins the network, instead of learning from scratch how to act in the surrounding environment, it learns the policies already acquired by more expert neighbors. Gains are expected due to a high correlation in the environments of adjacent expert and newcomer nodes. Policies are shared by Q-table exchange.
- **Adaptive Docition.** Docitive radios here share policies, based on performances. The nodes cooperate by exchanging information about the performances of their learning processes, e.g., the variance of the oscillation with respect to the target, the speed of convergence, etc. Based on this information, each node may learn from expert neighbors who are performing better, i.e., are more intelligent. Policies are shared by Q-table exchange.
- **Iterative Docition.** Docitive radios periodically share part of their policies, based on the reliability of their expert knowledge. More expert nodes share their expert knowledge periodically, by exchanging rows of the Q-table, corresponding to states that have been previously visited.

The degree of cooperation, and thus the overhead, augments with an increasing degree of docition. The optimum operating point hence depends on the system architecture, performance requirements, etc.

IV. Scenario

The scenario considered to validate the proposed approach consists of a primary and a secondary system working in the same region and in the same frequency band.

The primary system works at 615 MHz in the ultra high frequency (UHF) band, with a channel bandwidth of $BW = 6$ MHz. The primary BS transmits at power $P_{\mathrm{DTV}} = 1$ MW (90 dBm) effective radiated power (ERP) with an antenna height of 500 m. The primary receivers are placed randomly around the primary BS. According to the federal communication commission (FCC) for DTV, the limit for signal interference noise ratio (SINR) to the primary receivers is $SINR_{Th} = 23$ dB [7].

The secondary system is based on IEEE WRAN 802.22 standard. The secondary BS antenna height is 75 m. The secondary users are located randomly around the secondary BS, which is in charge of allocating power to them. The available power levels are $l = 20$, ranging from -80 dBm and 29.8 dBm ERP. With respect to the propagation models among nodes we consider the international telecommunication union recommendation (ITU-R) P.$1546-1$ [16], where the lognormal shadowing parameter is fixed at 5.54 dB. We consider that the independent WRAN systems switch on randomly during the simulation time. We study the effect of interference generated by the secondary users uplink transmissions onto the primary DTV system.

With respect to the decentralized Q-learning algorithm we consider that it is implemented in each agent with a learning rate $\alpha = 0.5$ and a discount factor of $\gamma = 0.9$. Also, we introduce a probability $\varepsilon = 0.1$ of visiting random states. This parameter is used in the action selection procedure to guarantee that the final policy is a global optimal and not a local one.

The SINR of the primary system $SINR_{X_i}$ at the i_{th} control point X_i is defined as:

$$SINR_{X_i} = \frac{P_{\mathrm{DTV}} h_{X_i}^{\mathrm{DTV}}}{\sigma^2 + \sum_{j=1}^{N} P_{\mathrm{SU}_j} h_{X_i}^{\mathrm{SU}_j}}, \quad (7)$$

where P_{DTV} is the transmitted power of the primary BS, $h_{X_i}^{\mathrm{DTV}}$ is the link gain between the primary BS and the i_{th} control point at the protection contour. P_{SU_j} is the transmission power of secondary user j, $h_{X_i}^{\mathrm{SU}_j}$ is the link gain between secondary user j and the control point X_i. Finally σ^2 is the noise power.

Digital Object Identifier: 10.4108/ICST.CROWNCOM2010.9173
http://dx.doi.org/10.4108/ICST.CROWNCOM2010.9173

In our system, the multiple nodes with distributed learning and docitive capabilities are the secondary BSs. We identify the system state, action, associated cost and the next state, to apply docition in the context of decentralized Q-learning to this scenario:

- **State.** As the system uses a decentralized Q-learning algorithm, the state is defined based on the local views of each WRAN system. The system state of node i at time t is defined as:
$$s_t^i = \left\{ I_t^i, d_t^i \right\} \tag{8}$$
where $i \in \{1, 2, \ldots, N\}$ is the WRAN cell index. $I_t^i \in I$ represents a binary indicator to specify whether the secondary system is generating an aggregated interference above or below the threshold of the primary receivers. This measure is based on the instantaneous SINR value computed or estimated at the control point of the $i-$th WRAN cell. $d_t^i \in D$, where $D = \{1, 2, \ldots, d\}$, indicates an approximate distance between the secondary user and the protection contour. This information is supposed to be reported by the secondary users to the secondary BS together with the spectrum sensing information during the collaborative spectrum sensing procedure [17].
- **Actions.** The set of possible actions is the set P of power levels that the secondary BS can assign to the $m-$th secondry user.
- **Cost.** The cost c_t^i assesses the immediate return incurred due to the assignment of action a at state s. The considered cost function is:
$$c = \left(SINR_t^i - SINR_{Th} \right)^2, \tag{9}$$
where $SINR_t^i$ is the instantaneous SINR in the control point of WRAN cell i. Q-learning aims to minimize this cost, so that the SINR at the control points is $SINR_{Th}$, which guarantees that interference at the primary receivers is below the threshold.
- **Next State.** The state transition from one state to another is determined by the power allocation of the secondary user.

V. Discussion

The starting point for the docitive investigations are the results presented in [6]. Here, the authors analyze a cognitive approach based on independent learners and demonstrate its superiority with respect to other opportunistic approaches. The reason is that the cognitive approach is capable of capturing and adapting to the dynamics of the environment, depending on multiple factors such as shadowing, mobility variable number of active WRAN systems, etc. We now compare the performance of (1) no docition, i.e., independent distributed learning; (2) startup docition; (3) adaptive docition; and (4) iterative docition.

The algorithm of startup docition is implemented in such a way that when a WRAN system switches on, it is able to acquire the whole Q-table of its closest neighbor. The rationale behind this algorithm is that each node does not have to learn its policy on its own, but rather can start learning from its neighbor's policy. The proximity of the two nodes guarantees that the system states of the two nodes are typically correlated and so their decision policies.

On the other hand, the adaptive docition is implemented periodically, so that each secondary system can acquire the Q-table of the neighbor that is experiencing best results. The performance of the learning process of each learning node is analyzed by comparing the variance of the SINR at the corresponding control point with respect to the target of $SINR_{Th} = 23$ dB.

Finally, the iterative docition is also implemented periodically, so that each agent acquires the Q-table rows of the neighbor corresponding to already visited states. The decision about which rows to learn is based on the comparison between the set of Q-values in each row. In particular, the rows with lower Q-values are acquired because those are expected to have been updated more times.

Figure 2 shows the convergence curves of one of the secondary WRAN systems for the four approaches. It can be observed that the docitive paradigms clearly speed up the learning process with respect to the case of independent learners. In particular, for the represented secondary WRAN, if we tolerate a margin of 1 dB in the algorithms' convergence, iterative docition needs 200 learning iterations to achieve convergence, adaptive docition 900, startup docition $10,900$ and independent learning $14,700$ iterations.

As for the performance in terms of precision, i.e., oscillations around the target SINR, Figure 3 depicts the complementary cumulative distribution function (CCDF) of the variance of the average SINR at the control point with respect to the set target of $SINR_{Th} = 23$ dB. It can be observed that due to the distribution of intelligence among interactive learners the paradigm of docition stabilizes the oscillations by reducing the variance of the SINR with respect to the specified target. More precisely, at typical precisions of below 1 %, we observe that the iterative docition outperforms the adaptive and startup docition by about an order of magnitude, and the independent learning algorithm by several orders of magnitude.

Finally, we have utilized the entropy as one possible measure to quantify the "intelligence" of a cognitive algorithm where a cleverer algorithm increases the order at the reference point and hence decreases the entropy measured there. For the four cases, we respectively obtained an entropy of 3.67 (independent learning), 3.01 (startup docition), 2.73 (adaptive docition) and 2.13 (iterative docition). An increase in intelligence – according to the entropy measure – of about 42% is observed between the iterative docition and the independent learning paradigms.

VI. Conclusions

In this paper we have proposed a decentralized Q-learning algorithm to solve the problem of the aggregated interference generated by multiple IEEE 802.22 WRAN systems, to the primary DTV users. Due to the slow convergence behaviors of decentralized learning schemes, we have introduced the paradigm of docition, which facilitates the exchange of expert information among agents in order to improve the learning process. The idea is to make inexpert CRs able to make

Digital Object Identifier: 10.4108/ICST.CROWNCOM2010.9173
http://dx.doi.org/10.4108/ICST.CROWNCOM2010.9173

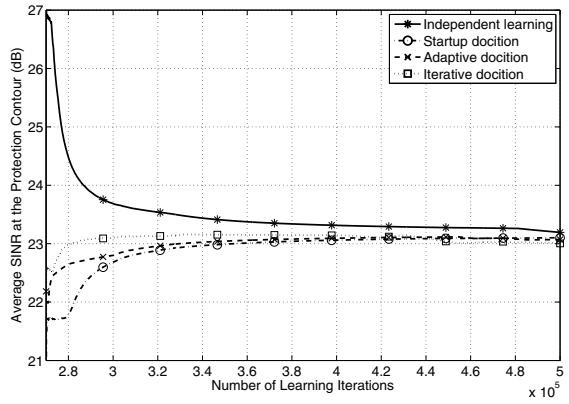

Fig. 2. Convergence of the average SINR at the control point to the specified target, for the four proposed approaches.

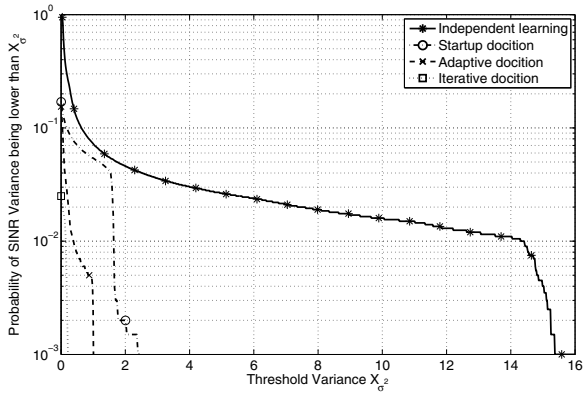

Fig. 3. CCDF of the variance of the average SINR at the control point w.r.t. to the specified $SINR_{Th}$, for the four proposed approaches.

decisions in unvisited states or in situations where they would not be capable of representing their knowledge, or correctly observing the environment. Simulation results show that nodes that take advantage of the knowledge previously acquired by their neighbors speed up their learning process and experience improved precision. Those improvements increase with the amount of the information exchanged among nodes, so that a tradeoff between the performances of the learning process and signalling/complexity required has to be considered and will be studied in future work.

ACKNOWLEDGMENT

This work has been partially supported by the COST Action IC0902, European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (contract n. 216715) and by Generalitat de Catalunya under grant 2009-SGR-940.

REFERENCES

[1] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE J. Select. Areas Commun*, vol. 23, pp. 201–220, Feb. 2005.

[2] J. Hu and M. P. Wellman, "Technical note: Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[3] IEEEP1900 home page. [Online]. Available: http://www.ieeep1900.org

[4] M. N. Ahmadabadi and M. Asadpour, "Expertness based cooperative Q-learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 32, no. 1, pp. 66–76, Feb. 2002.

[5] M. Tan, *Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents*. In M. N. Huhns and M. P. Singh, editors. Morgan Kaufmann, San Francisco, CA, USA., 1993, ch. 26, pp. 451–480.

[6] A. Galindo-Serrano and L. Giupponi, "Decentralized Q-learning for aggregated interference control in completely and parially observable congitive radio networks," in *In Proc. of IEEE Consumer Communications and Networking Conference, IEEE CCNC 2010*, Las Vegas, USA, 9-12 Jan. 2010.

[7] S. Shellhammer, S. Shankar, R. Tandra, and J. Tomcik, "Performance of power detector sensors of DTV signals in IEEE 802.22 WRANs," in *Proc. of First International Workshop on. Technology and Policy for Accessing Spectrum, IEEE ACM TAPAS*, 5 Aug. 2006.

[8] S. Shankar and C. Cordeiro, "Analysis of aggregated interference at DTV receivers in TV bands," *in Proc. of International Conference on Cognitive Radio Oriented Wireless Networks and Communications, CROWNCOM 2008*, May 15-17, 2008, Singapore.

[9] D. Fudenberg and L. D., *The Theory of Learning in Games*. MIT Press, 1998.

[10] J. Watkins and P. Dayan, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, pp. 1039–1069, 2003.

[11] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for aggregated interference control in cognitive radio networks," *IEEE Trans. on Vehicular Technology*, to appear.

[12] P. Hoen and K. Tuyls, "Analyzing multi-agent reinforcement learning using evolutionary dynamics," in *In Proc. of the 15th European Conference on Machine Learning (ECML)*.

[13] Y. Chen, C. Chang, and F. Ren, "Q-learning-based multirate transmission control scheme for RRM in multimedia WCDMA systems," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, Jan. 2004.

[14] L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," *Autonomous Agents and Multi-Agent Systems*, vol. 3, no. 11, pp. 383–434, Nov. 2005.

[15] M. E. Harmon and S. S. Harmon, "Reinforcement learning: A tutorial," 2000. [Online]. Available: http://www.nbu.bg/cogs/events/2000/Readings/Petrov/rltutorial.pdf

[16] (2009) The ITU website. [Online]. Available: http://www.itu.int/rec/R-REC-P.1546-1-200304-S/en

[17] C. Cordeiro, K. Challapali, and D. Birru, "IEEE 802.22: An introduction to the first wireless standard based on cognitive radios," *Journal of Communications*, vol. 1, no. 1, April 2006.

Digital Object Identifier: 10.4108/ICST.CROWNCOM2010.9173
http://dx.doi.org/10.4108/ICST.CROWNCOM2010.9173