
Ahmed AZZOUNA\textsuperscript{a}, Amel GUEZMIL\textsuperscript{a}, Anis SAKLY\textsuperscript{b}, Abdellatif MTIBAA\textsuperscript{a,b}

\textsuperscript{a}CSR GROUP of Electronic and Microelectronic Laboratory, University of Monastir, Tunisia
\textsuperscript{b}National Engineering School of Monastir, University of Monastir, Tunisia

Abstract

Cognitive Radio (CR) is a new generation of wireless communication system that enables unlicensed users to exploit underutilized licensed spectrum to optimize the radio spectrum utilization. The resource allocation is difficult to achieve in a dynamic distributed environment, in which CR users take decisions to select a channel without negotiation, and react to the environmental changes. This paper focuses on using a multi-agent reinforcement-learning (MARL), Q-learning algorithm, on channels selection decision by secondary users in 2×2 and 3×3 cognitive radio system. Numerical results, obtained with MATLAB, demonstrate that resource allocation is realized without any negotiation between secondary and primary users. In this work, the analogy between the numerical and simulated results is also noted.

Keywords: Cognitive Radio; Resource Allocation; Multi-agent Reinforcement Learning; Q-Learning Algorithm

1. Introduction

To address the scarcity of spectrum resources and the wireless networks deploying difficulties, several innovative approaches have been proposed allowing a dynamic and opportunistic use of unused frequency spectrum [1-5], in addition to the current approach of static spectrum allocation.

Cognitive radio (CR) is a concept that responds this challenge. Radio technology is no longer enslaved to a single type of network and a limited number of frequency bands. Based on software radio methods and computing capabilities and analysis, cognitive radio observes, analyzes and adapts to its environment in a very dynamic way. It’s particularly able to choose the frequency bands and radio networks available depending on the desired quality of service (QoS) and rules imposed by the regulator. But beyond the concept developed by Mitola in a founding document [6-7], much work remains to be done: transmission innovative methods, field measurements, network architecture, heterogeneous networks management, detection algorithms, dynamic frequency allocation, networks performance optimization...

Management methods of the frequency spectrum should know considerable changes in the coming years [8]. They are evolving essentially under the influence of two major trends. First, many scientists suppose that currently the spectrum is not enough efficiently used. On the other hand, radio technology...
becoming more and more flexible, allowing the same device to access easily to wide ranges of spectrum and different types of networks (WiFi, WiMAX, 2G, 3G, 3G+ etc.).

Since its founding by Mitola, Cognitive Radio has been the target of several studies in order to find new techniques for spectrum detection [9-10], collaborative spectrum sensing [11-12], spectral efficiency optimization [13-14], radio resource management [15], resource allocation [16-18], and even some contributions to the hardware implementation of some algorithms [19-20].

In contrast to existing systems where the spectrum allocation is static, the cognitive radio terminals can dynamically find the network access frequency by detecting free spectrum bands. The resource allocation is still a major problem addressed by researchers from different angles using several techniques for example dynamic spectrum access [1], learning algorithms [21] such as Q-learning [22-23] and reinforcement-learning [24-25], neurons networks [26-27], inference engine and fuzzy techniques [28] and even meta-heuristic algorithms such as genetic algorithms [29-31] and biological algorithms [32].

This paper focuses on using a multi-agent reinforcement-learning (MARL), Q-learning algorithm, on channels selection decision by secondary users in cognitive radio system with two channels, two SU (2x2) and three canals, three SU (3x3). Multi-agent learning is a promising direction of recent research in the context of intelligent systems. The single-agent case (SARL) has been studied extensively over the past two decades; however, the multi-agent case has been little studied due to its complexity. When multiple autonomous agents learn and operate simultaneously, the environment is strictly unpredictable and all assumptions that are made in the single-agent case, such as stationary and Markov property, are often inapplicable in the multi-agent context.

2. System Model

In multi-user and multi-channel cognitive radio systems, the resource change is very dynamic. Therefore, without negotiation between different SUs, serious problems would occur in data transmission because of the collision the data packet may be lost, see Fig 1.

![Fig. 1. Competition in multi-user and multi-channel cognitive radio systems](image_url)

This paper treats the N×N case systems with N active secondary users (denoted by letters A, B, C,...) and N frequency channels (denoted by numbers 1, 2, 3,...). The channel selection problem becomes a game N×N [33]: the secondary user \( i \), \( i \in \{A, B, C, \ldots\} \), receives a reward \( R_{ij} > 0 \) when transmitting through the free channel \( j \), \( j \in \{1, 2, 3, \ldots\} \), as long as the other SUs do not use this channel and \( R_{ij} = 0 \) if one or more SUs attempt to transmit through this channel.

In this game, there is no communication between SUs and the rewards received are unknown to all users.
3. Q-learning

In cognitive radio systems, each cognitive user is considered as an agent and the wireless network as the external environment. Cognitive radio can be formulated as a system in which communicating agents sense their environment, learn, and adjust their transmission parameters in order to optimize their performance. This formulation represents the reinforcement learning context, see Fig 2.

![Diagram](image)

Fig. 2. Multi-agent reinforcement learning based cognitive radio

Q-learning is an online reinforcement learning algorithm that determines optimal policy without detailed modeling of the system environment [21]. Denote decision epochs by \( t \in \{1, 2, \ldots \} \), a constant epoch duration by \( t_D \), actions by \( a_i \in A \), and delayed rewards by \( r_{t+1}(a_i) \). Each agent \( i \) maintains a Q-table with \( |A| \) entries to keep track of learnt action value or Q-value, \( Q_i(a) \) within an interval of \([0, Q_{\text{max}}]\) for all its possible actions. The Q-value estimates the level of local reward for an action \( a_i \); hence changes in the Q-value will lead to changes in an agent’s action. At each decision epoch \( t \), agent \( i \) chooses an action \( a_i \) and receives a local reward \( r_{t+1}(a_i) \) at time \( t+1 \). The agent \( i \) updates the Q-value of action \( a_i \) at time \( t+1 \) as follows:

\[
Q'_{t+1}(a_i) \leftarrow (1-\alpha)Q_t(a_i) + \alpha r_{t+1}(a_i)
\]

(1)

where \( \alpha \in [0,1] \) is the learning factor.

To assure that all actions will be considered, Boltzmann distribution is used for random exploration, i.e.

\[
P(\text{user } i \text{ choose channel } j) = \frac{e^{Q_i(j)/\gamma}}{e^{Q_i(j)/\gamma} + e^{Q_j(j)/\gamma}}
\]

(2)

where \( \gamma \) is called Boltzmann temperature, which controls the exploration frequency and \( j \) the other users different from user \( i \).

Convergence using geometric argument proposed in [34], provide an intuitive explanation that the updating rule of Q-Value, given in (1), will converge to a stationary equilibrium point. Fig 3 represents...
the dynamics in the Q-learning algorithm for the 2×2 case ($u_\mu = Q_{\mu 1} / Q_{\mu 2}$ versus $u_A = Q_{A1} / Q_{A2}$). As can be seen, the plan is divided into four zones limited by $u_A = 1$ and $u_\mu = 1$. In the first and third regions (I and III), each secondary user prefers selecting a different channel. Meanwhile, in the second and forth regions (II and IV), both secondary users prefer selecting the same channel.

Fig. 3. Dynamics in the Q-learning algorithm

4. Numerical Results

This section demonstrates the analogy between the numerical and simulated results obtained with MATLAB.

4.1. 2×2 case:

The demonstration of Q-learning algorithm convergence is given by Fig 4 and 5. This figures show the dynamics of $u_\mu = Q_{\mu 1} / Q_{\mu 2}$ versus $u_A = Q_{A1} / Q_{A2}$ for different trajectories. For an initial unstable point, where the two SUs collide and choose the same channel (zone I or III in Fig 3), the curves converge to a stable zone (II or IV in Fig 3) which means that SU A selected the channel 1 whilst the SU B selected channel 2 (or vice versa) and the collusion problem is diverted. This result is confirmed by Fig 6 and 7 which represent the SUs probabilities to choose the first channel. For an initial case, where the two SUs choose the first channel ($P_{A1} = P_{B1} = 1$), the Q-learning algorithm has avoided the collision by keeping the second SU choice and switching the first SU to the channel 2 in Fig 6 (and vice versa for Fig 7).

The learning procedure, obtained from 1000 spectrum access periods, is considered as completed when the probabilities of choosing a channel are larger than 95% for one SU and smaller than 5% for the other.

Fig 8 and 9 represent the learning speed delay for different learning factors $\alpha_0$ and different Boltzmann temperatures $\gamma$, respectively. This Figures show that the learning procedure is faster with a higher learning factor and a smaller temperature.

4.2. 3×3 case:

The developed algorithm is still valid for the case of 3 channels and 3 secondary users. Fig 10 shows the evolution of channel’s probability choice for each user. As can be seen, the collision is avoided in a situation where all users initially choose channel 1.
To analyze the impact of channels and SUs number for fixed $\alpha$ and $\gamma$, the learning speed for both cases (2×2 and 3×3) is given in Fig 11. As can be seen, the learning procedure is becoming slower by increasing N. Nevertheless, this minor difference remains insignificant.

![Fig. 4. An example of dynamics of the Q-learning algorithm](image1)

![Fig. 5. An example of dynamics of the Q-learning algorithm](image2)

![Fig. 6. An example of the evolution of channel selection probability (N=2)](image3)

![Fig. 7. An example of the evolution of channel selection probability (N=2)](image4)

![Fig. 8. CDF of learning delay with different learning factor $\alpha_0$](image5)

![Fig. 9. CDF of learning delay with different temperature $\gamma$](image6)
Fig. 10. An example of the evolution of channel selection probability (N=3)

Fig. 11. CDF of learning delay for N=2 and N=3
5. Conclusion

This paper used a multi-agent reinforcement-learning, Q-learning algorithm, on channels selection decision by secondary users in cognitive radio system. For simplicity, 2×2 and 3×3 systems was studied. The learning procedure for spectrum access without negotiation in cognitive radio systems is discussed. During the learning, each SU considered the channel and other secondary users as its environment, updated its Q-values, and took the best action. The learning speed delay was compared for different learning factors α0 and different Boltzmann temperatures γ . Numerical results showed that SU could avoid collusion in both cases 2×2 and 3×3.

References

[11] Hong Li; Ma Junfei; Xu Fangmin; Li ShuRong; Zhou Zheng; "Optimization of Collaborative Spectrum Sensing for Cognitive Radio", IEEE International Conference on Networking, Sensing and Control (ICNSC’08), Sanya, China, April 6-8, 2008.


