Noname manuscript No.
(will be inserted by the editor)

Cooperative Q-learning based channel selection for cognitive radio networks

Feten Slimeni · Zied Chtourou · Bart Scheers · Vincent Le Nir · Rabah Attia

Received: date / Accepted: date

Abstract This paper deals with the jamming attack which may hinder the cognitive radio from efficiently exploiting the spectrum. We model the problem of channel selection as a Markov decision process. We propose a real-time reinforcement learning algorithm based on Q-learning to pro-actively avoid jammed channels. The proposed algorithm is based on wideband spectrum sensing and a greedy policy to learn an efficient real-time strategy. The learning approach is enhanced through cooperation with the receiving CR node based on its sensing results. The algorithm is evaluated through simulations and real measurements with software defined radio equipment. Both simulations and radio measurements reveal that the presented solution achieves a higher packet success rate compared to the classical fixed channel selection and best channel selection without learning. Results are given for various scenarios and diverse jamming strategies.

Keywords Cognitive radio · jammer · Markov decision process · Q-learning · cooperation · channel selection

F. Author
VRIT Lab - Military Academy of Tunisia, Nabeul, Tunisia
E-mail: feten.slimeni@gmail.com

S. Author
VRIT Lab - Military Academy of Tunisia, Nabeul, Tunisia

T. Author
CISS Departement - Royal Military Academy (RMA), Brussels, Belgium

Fourth Author
CISS Departement - Royal Military Academy (RMA), Brussels, Belgium

Fifth Author
SERCOM Lab - EPT University of Carthage, Marsa, Tunisia

1 Introduction

A cognitive radio (CR) refers to a radio system aware of its RF environment and capable of learning and adapting its transmission parameters [1–4]. In addition to the coexistence with incumbents, the system must achieve interferers awareness and avoidance to provide continuous reliable communication wherever and whenever needed. CR anti-jamming techniques have recently attracted research attention since jammers may disturb CR spectral behavior [5–13].

Under the assumption of fixed jamming strategy trying to prevent the CR from an efficient exploitation of the available channels, the CR has to learn how to escape from jammed channels without scarifying a long time. Markov Decision Process (MDP) is a suitable tool to study such problem since it is a stochastic framework modeling an agent decision problem to optimize its outcome. In the CR context of dynamic RF environment and imperfect opponent knowledge, the agent may use reinforcement learning (RL) algorithms to solve the non deterministic MDP and learn the optimal strategy [14]. RL techniques as the Q-learning algorithm are based on the interaction with the environment to update the knowledge and estimate the optimal MDP solution. In [15], a decentralized Q-learning algorithm is proposed to deal with the problem of aggregated interference generated by multiple CRs at passive primary receivers. [16], [17] and [18] study the Q-learning algorithm to solve the CR jamming problem. Differently from works available in literature, we present an on-line Q-learning algorithm based on wideband spectrum sensing and cooperation between two CR nodes to pro-actively avoid jammed channels and overcome hidden jammer problem. Furthermore, we provide both simulation results and real measurements in terms of Packet Success Rate (PSR). The proposed dynamic spectrum access (DSA) algorithm significantly improves the packet success rate compared to both static spectrum access and intelligent spectrum access without learning.

The rest of this paper is organized as follows: Section 2 describes the MDP model. Section 3 presents the proposed Q-learning algorithm. Section 4 discusses the simulation results and section 5 discusses the real measurements performed by software defined radio equipment. Finally, section 6 summarizes the conclusions.

2 Markov decision process

A MDP is a discrete-time stochastic control system that models an agent decision making problem to optimize a final outcome. The agent gets the optimal strategy through solving the MDP. The problem in this paper consists in the jamming attack and the solution consists in the adequate decisions to avoid the jammed channels. The MDP is defined with four components; A finite set of states $\{S_0, \cdots, S_t\}$, where $t = 0, 1, \cdots, N$ represents a sequence of time slots. A finite set of actions $\{a_1, \cdots, a_M\}$. A state transition probability

$P_a(S, S')$ of moving from one state S to another state S' after executing an action a. An immediate reward $R_a(S, S')$ related to the taken decision.

A MDP can be solved through model-based approaches if the transition probability function is known, otherwise model-free approaches are used to solve it based on RL algorithms such as Q-learning [19]. The Q-learning model-free RL algorithm was introduced in [20] as a simple way to learn how to act optimally by successively improving the actions evaluations. This algorithm is able to find a suboptimal good strategy through real time interaction with the environment. The goal is to find a mapping from state/action pairs to Q-values. This result can be represented by a Q-matrix of N rows and M columns. At every time step, the agent measures the feedback of trying an action a in a state S and updates the corresponding $Q(S, a)$ value, using the following expression:

$$Q(S, a) \leftarrow (1 - \alpha)Q(S, a) + \alpha \left[R_a(S, S') + \gamma max_x Q(S', x)\right] \quad (1)$$

where $0 < \alpha \leq 1$ is the learning rate that controls how quickly new estimates are blended into the old ones. $0 \leq \gamma \leq 1$ is the discount factor that controls how much effect future rewards have on the optimal decisions. Equation (1) is repeated for all visited pairs (S, a) until the convergence to almost fixed Q values. The optimal strategy is met when all the different possibilities are infinitely visited during the training period. After this period, the agent starts the exploitation of the solution which corresponds to choosing the action having the maximum Q value in each state: $max_x Q(S, x)$. This standard version of the Q-learning algorithm is said to be asynchronous since at each time step the agent updates a single Q value [21]. It is also called OFF-policy since it allows arbitrary experimentation during the training period [22]. The learning agent applying this algorithm should wait until the convergence to start exploiting the optimal policy which is not suitable in hostile and dynamic environment.

3 Cooperative learning algorithm

We consider a fixed jamming strategy trying to prevent the CR from an efficient exploitation of M available channels. As a defense strategy, the CR has to learn how to escape from jammed channels without scarifying a long training period. The state of the CR is defined by three parameters: $S = \{f_{TX}, n, f_{JX}\}$, where f_{TX} is its current operating frequency and n is the number of successive time slots using this frequency. We opt for mixing spatial and temporal properties in the state space definition to consider the CR staying in the same channel more than one time. To take into consideration the asynchronous jammer behavior, including its random starting time and its unknown current channel, we introduce f_{JX} as the worst (or jammed) frequency to the definition of the state. We consider that at each time slot the CR does wideband spectrum sensing [23] to detect the worst and the best channels. At every state, the CR should choose an action to move to another state. We define its possible actions as a set of M actions, which are the M available channels:

$\{a_1, \cdots, a_M\} = \{f_1, \cdots, f_M\}$. We define a reward function related to the result of the WBSS done every time by the CR node before selecting an action:

$$R_f(S, S') = 1 - \frac{E(f)}{ET}, \tag{2}$$

$E(f)$ is the energy measured over the channel f and ET is the total energy measured over the M channels. Such reward function adapts the CR channel selections to the real time spectrum occupancy, which allows a pro-active collision avoidance.

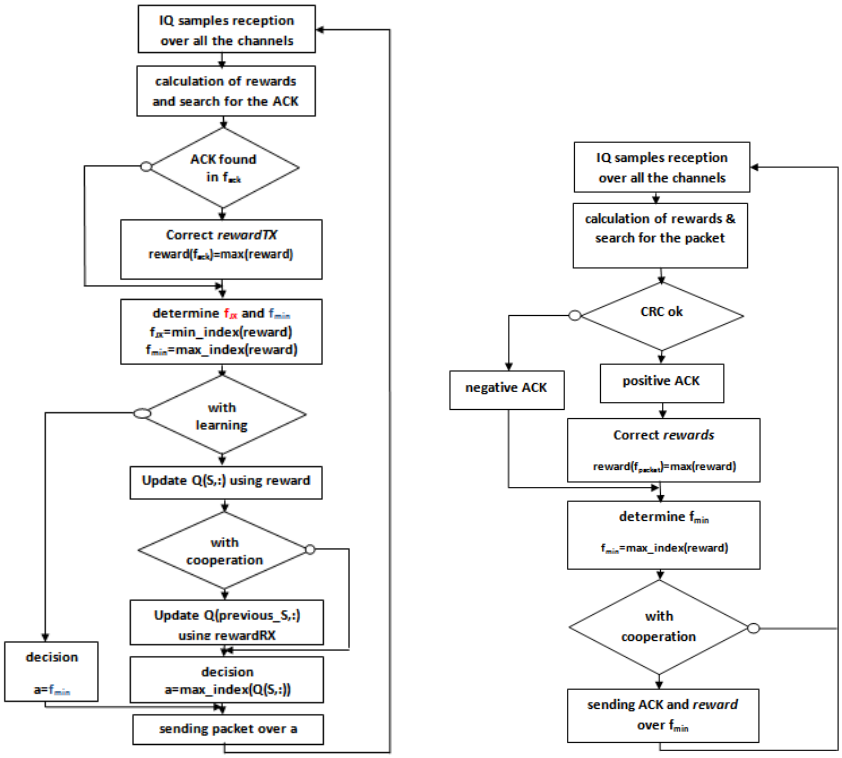
In order to adapt the Q-learning algorithm to the jamming scenario, we extend the on-line algorithm denoted as ON-Policy Synchronous Q-learning (OPSQ-learning) of [18] by adding cooperation between two CR nodes. OPSQ-learning allows the CR to keep learning and choosing the best decisions in real time. It consists in replacing the OFF-policy with ON-policy by selecting the best action instead of trying random actions to minimize the wrong decisions. Furthermore, the CR is able to do a synchronous update of all the Q values related to the current state $Q(S, :)$ by doing wideband spectrum sensing (WBSS) before the action selection. The OPSQ algorithm allows on-line learning during real-time communication without going through a training before an exploitation period. To overcome the problem of hidden jammer that may interfere the transmitted packets without being detected by the learning node, the transmitter may cooperate with the node receiving the packets. This latter transmits the acknowledgment including its own sensing results. The learning node updates the Q values based on both its sensing and the received sensing results which gives more vision about the actual and the previous channels occupancy. The proposed solution is described in algorithm 1, using $R_a^l(S, S')$ to denote the local reward measured by the learning node in the current state S for each possible action a that results in a next state S'. Likewise, $R_a^r(S_p, S_p^{'})$ represents the received reward measured by the cooperative node during the reception of the previous packet. We are considering in this paper one jammed channel, but the proposed learning algorithm, based on wideband energy detection, allows the detection of the jammer even attacking multiple channels.

Algorithm 1 pseudocode for OPSQ-learning

Select a random initial state $S = S_0$
while true **do**
  The learning node does WBSS and checks for acknowledgment reception
  Update all Q values at the current state S based on the local WBSS and the previous state S_p based on the received WBSS results using,$\forall a$:
  $Q(S, a) = (1 - \alpha)Q(S, a) + \alpha(R_a^l(S, S') + \delta max_x Q(S^{'}, x))$
  $Q(S_p, a) = (1 - \alpha)Q(S_p, a) + \alpha(R_a^r(S_p, S_p^{'}) + \delta max_x Q(S_p^{'}, x))$
  Select an action a with max Q value
  Take a and observe next state $S^{'}$
  $S_p = S$
  $S = S^{'}$
end while

Figure 1-(a) details the tasks performed by the learning node. The first step consists in gathering the IQ samples through wideband reception over the considered M channels. Then the rewards associated to all the possible actions are calculated using equation 2 based on energy detection to perform WBSS. During this processing step, the learning node looks blindly for an acknowledgment over the M considered channels without a rendez-vous or a signaling channel. If an acknowledgment is received over a channel f_{ack}, the reward calculated for that channel carrying the acknowledgment should not keep its low value (since it has high energy $E(f_{ack})$) to not falsify the decisions and be considered as a jammed channel. For that, the learning node associates to this channel the maximum reward that he has calculated. The next step consists in deciding which is the jammed channel and which is the best one (having the maximum reward). To evaluate the proposed algorithm, we have compared four channel selection strategies; The first strategy is the classical fixed channel selection that consists in transmitting over the same channel all the time with neither sensing nor learning. The second one is based on sensing without learning. It consists in the selection of the channel having the minimum energy in each time step, it is denoted as the best channel selection. In the third strategy, the learning node applies the proposed OPSQ-learning algorithm but without cooperation, which means updating just the Q values related to the actual state $Q(S,:)$. The action having the maximum Q value, $a = max\_index(Q(S,:))$, is selected to transmit the packets. The last strategy consists in cooperating with the node receiving the packets to have more knowledge. So, the learning node updates the actual state as in the third strategy and he updates also the previous state based on the reward values extracted from the acknowledgment. The received rewards are related to the previous time step when the destination node has received the transmitted packet. If the learning node does not receive the acknowledgment, he considers that the response was jammed or lost and considers null received rewards. Finally, the learning node selects the channel having maximum Q value. For each of the four strategies, the packet is sent over the selected channel.

Figure 1-(b) describes the operations of the CR node receiving the transmitted packets. After a wideband reception of the IQ samples, the channels rewards are calculated based on the detected energies through WBSS. This node looks blindly for the packet over the considered channels and performs the cyclic redundancy check (CRC). If the packet is received correctly over a channel f_{packet}, the CR node decides to send a positive acknowledgment. He also corrects the reward that he calculated for that channel to the maximum reward since it is not a jammed one. If the CRC is false, a negative acknowledgment is sent. In both cases, he selects the best channel having maximum reward to send the ACK sign and the rewards if we have selected the cooperative strategy.



(a) The learning node (b) The cooperative node

Fig. 1: Descriptive diagrams of the learning and cooperative nodes processes

4 Simulation results and discussion

This section concerns MATLAB simulation of the considered four channel selection strategies: (1) the classical fixed one, (2) the sensing based best selection, (3) OPSQ-learning based strategy, (4) cooperative OPSQ-learning based channel selection. We have opt for a high fidelity simulation which provides the flexibility to adjust the CR configurable parameters according to the chosen strategy and to the electromagnetic environment without abstractions of the physical layer [24]. Furthermore, this allows going down to the level of IQ samples and includes signal processing details such as spectrum sensing, frame construction and real modulation & demodulation.

After presenting the simulation model, we will provide the results found considering Additive White Gaussian Noise (AWGN) as statistical channel model. Since this channel includes only the white noise without considering the losses present in a wireless link, we discuss in the last paragraph of this section how the fading could impact the learning process.

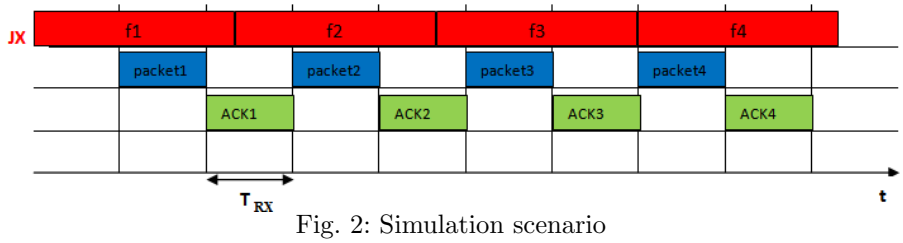


Fig. 2: Simulation scenario

4.1 Simulation model

We are considering four channels ($M = 4$), a learning rate $\alpha = 0.1$ and a discount factor $\gamma = 0.1$. The two CR nodes transmit binary phase shift keying (BPSK) modulated signals of $12KHz$ bandwidth (packets sent by the learning node and acknowledgments sent by the cooperative node) and perform WBSS. The reception period is equal to the transmission period $T_{RX} = T_{packet} = T_{ACK} = 0.98ms$. The node receiving the packets and performing CRC measures the packet success rate (PSR) for the four channel selection strategies as given in table 1. The four strategies corresponds to the four rows of the table. The columns of the table corresponds to two scenarios depending on the visibility of the jammer to the learning node. The PSRs are given for 1000 transmitted packets.

4.2 Simulation results

We started considering a slow sweeping jammer with a dwell time $T_{JX} = 2.28ms$ on each channel, which corresponds to $T_{JX} \approx 2.3T_{packet}$ as represented in figure 2.

In the first scenario corresponding to a jammer detectable by both of the CR nodes (column 1), learning with cooperation (row 4) outperforms learning without cooperation (row 3) since the cooperative node gives more information to the learning node about the jammer that may be not detected during its sensing period but appears during the transmission of the packet. The channel selection based on OPSQ-learning (row 3) is better than selecting the best channel without learning (row 2) since the learning decision is not only based on the actual information but also on the past learned information ($(1-\alpha)Q(S,a)$) and on the future expectation ($\alpha\gamma max_x Q(S',x)$) as given in equation (1) of Q value updates. The best channel selection based just on spectrum sensing (row 2) gives higher success rate than the fixed channel selection (row 1) since this latter is a blind selection staying on the same channel all the time without any information about the channels occupancy.

Choosing the best channel with or without learning (row 2 or row 3) are similar to staying in the same channel (row 1) when the jammer is hidden to the learning node (column 2) since both best channel selections are based only on its sensing result. If the destination node cooperates with the learning

node (row 4), the PSR increases since the cooperative node gives an information about the channel used for the previous packet transmission: packet success implies the jammer absence and packet failure means collision with the jammer.

Figure 3 gives the channels occupancy for each of the learning node and the sweeping jammer over time for both the second and the third strategies. The best channel selection without learning, given in subfigure (a), results in loosing more packets than the strategy based on OPSQ-learning presented in subfigure (b). For example, we consider packet number seven as indicated in the figure. The wideband spectrum sensing gives the following reward vector for both of the strategies: $reward = (0.4145; 0.9982; 0.9981; 0.5892)$, the best channel selection strategy results in the selection of the second channel resulting in collision with the jammer. However, the on-line learning algorithm calculates the $Qvalues = (0.0228; 0.1896; 0.1898; 0.1679)$. Applying the proposed learning algorithm, the third channel having the maximum Q value is selected, as presented in subfigure (b).

According to the presented results, the cooperative OPSQ-learning (row 4) outperforms learning without cooperation. Moreover, a CR applying the proposed OPSQ-learning succeeds better than a CR just sensing the spectrum to select the best channel, if the jammer is detectable. However, these success rates depend on the jammer's period and tactic. **In terms of the jamming period**, we have considered a faster jammer with a dwell time larger than the sensing period but lower then the sensing plus transmission periods of the learning node. The simulation results, given in table 2, give the same conclusions as the results against the slow sweep jammer. The noteworthy difference concerns the best channel selection without learning (row 2) which gives lower PSR than the three other strategies even the fixed channel selection. This is due to the fast sweep jammer which may be detected by the CR node in one channel during the sensing period but moves to another channel during the transmission period. **In terms of the jamming tactic**, we have applied the proposed solution against both a pseudo random jammer and a reactive one, the results are given in tables 3 and 4. Concerning the pseudo random jammer, we have considered a sweep over a sequence of six channels $\{f_1, f_4, f_3, f_3, f_2, f_4\}$. Concerning the reactive jammer, we have considered an intelligent jammer who is capable to do spectrum sensing to jam the detected occupied channel. We assumed that this jammer needs a duration of two time slots before jamming the detected frequency, because it has to do the spectrum sensing, then make the decision and finally hop to the detected frequency. The results of tables 3 and 4 against pseudo random and reactive jammers confirm the same conclusions as the results of table 1 against a sweeping jammer; The channel selection based on the cooperative OPSQ-learning algorithm outperforms the three other considered channel selection strategies for both scenarios of visible and hidden jammer. Furthermore, in the first scenario of detectable jammer, the OPSQ-learning strategy without cooperation outperforms the best channel selection strategy without learning which also outperforms the

	Jammer detectable by the learning node	Jammer hidden to the learning node
Classical fixed channel selection	66.6%	66.6%
Best channel selection without learning	80%	66.6%
Learning without cooperation	82.8%	66.6%
Learning with cooperation	96.8%	84.4%

Table 1: Simulation results: Packet Success Rate against slow sweep jammer

	Jammer detectable by the learning node	Jammer hidden to the learning node
Classical fixed channel selection	73.3%	73.3%
Best channel selection without learning	65.5%	73.3%
Learning without cooperation	77.3%	73.3%
Learning with cooperation	86%	88.7%

Table 2: Simulation results: Packet Success Rate against fast sweep jammer

	Jammer detectable by the learning node	Jammer hidden to the learning node
Classical fixed channel selection	53.7%	53.7%
Best channel selection without learning	77.6%	53.7%
Learning without cooperation	89.5 %	53.7%
Learning with cooperation	99.4 %	74.5 %

Table 3: Simulation results: Packet Success Rate against pseudo random jammer

fixed channel selection. The three strategies gives the same packet success rate if the jammer is hidden to the learning node.

4.3 Discussion of the fading impact

We have presented results found under the assumption of simple AWGN channels. However the received energy could be affected not only by the jamming signal but also by the fading present in wireless channels.

Fading may affect both spectrum sensing and packet transmission, as follows:

	Jammer detectable by the learning node	Jammer hidden to the learning node
Classical fixed channel selection	1%	1%
Best channel selection without learning	96%	1%
Learning without cooperation	97%	1%
Learning with cooperation	97.6%	66.9%

Table 4: Simulation results: Packet Success Rate against reactive jammer

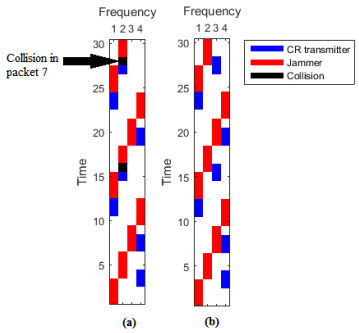


Fig. 3: Best channel selection based on sensing (a) versus channel selection based on learning (b) against a sweeping jammer

– Depending on the coherence time and the spectrum sensing time, fading can influence the rewards measured by the learning node, which affects the Q matrix and may lead to more collisions with the jammer (so decrease of the PSR)
– A lost packet or a drop of the received energy due to fading will falsify the Q matrix in the case of cooperative learning since the learning node updates the Q values based on rewards measured by the receiver node.

5 USRP measurements

We have implemented the physical layer signal processing steps and the four channel selection strategies described previously using Qt Creator/C++ development environment and the Universal Software Peripheral Radio platforms USRP E110 and B205mini. The physical layer is based on BPSK modulation over the four channels: $(432.94; 432.98; 433.02; 433.06)MHz$. Without loss of generality, we have opted for the stop and wait scheme described in figure 4, but the presented study can be applied to any time division multiplexing (TDM) scheme. The learning node does wideband reception of the IQ samples during the reception period T_{RX} detecting acknowledgment (ACK) and jamming signals. The time needed to do blind search of the ACK (for the learning node) or the packet (for the cooperative

node) over the M channels is denoted $T_{process}$. After sending the packet, the learning node waits until the end of the cooperative node processing before returning to the reception step. The cooperative node respects the same doctrine to keep synchronized with the learning node. We call radio period the sum of the reception, transmission and packet/ack processing periods: $T_{radio} = T_{RX}+2*T_{process}+T_{TX}$. Figure 7 describes the alternation between the packet and the acknowledgment transmissions by USRP nodes. The packet success rate (PSR) measured by the CR receiving the packets is given in table 5 for the four considered strategies in both scenarios of a jammer detectable (scenario 1) or hidden (scenario 2) to the learning node. The tests were performed in the Royal Military Academy (RMA) where the USRP platforms were placed in different buildings as described in figures 5 and 6. The jammer was running standalone at start up of USRP E110, the reporting node code was transferred to an Odroid-U3+ connected to one of the two used USRP B205mini, and the learning node was running on a laptop connected to the other USRP B205mini platform.

The real measurements show that the cooperation ameliorates the PSR for both scenarios since the learning node receives the sensing result measured by the cooperative node which helps in learning the jammer's behavior. Without cooperation, the learning node gains in terms of PSR only if he detects the jammer since the proposed learning algorithm is based on the sensing results. Otherwise, the learned strategy has the same PSR as the fixed and the sensing based strategies. Figure 8 gives the best channel selection based on sensing (a) and the channel selection based on learning (b) in the presence of the sweeping jammer. Based just on sensing, the strategy presents wrong decisions due to the asynchronous jammer behavior. This latter may be detected in a channel during the sensing period, but it moves to another channel during the packet transmission period which leads to repeated collisions if this behavior is not learned to proactively avoid the jammed channels. Furthermore, the CR based just on sensing without learning may move from channel to another without avoiding uneeded frequency alteration. However, the learning node ameliorates its behavior over time based on the goodness measures of the available decisions. The Q values are updated based not only on the sensing results but also on the past learned information and the future expectation to take the best decision avoiding collisions.

Tests using real radio equipments were also performed against pseudo random and reactive jammers. The measured results confirm the same conclusions as simulation results; The channel selection based on the proposed cooperative algorithm outperforms learning without cooperation which also outperforms the best channel selection without learning. However, the real USRP measurements are not equal to MATLAB simulation values. This is due to the implemented time division multiplexing scheme that needs a processing time for the blind reception of the packets or the acknowledgments. In MATLAB, the simulation time is different from the real time and

neither the CR nodes nor the jammers need a processing time as presented in figure 2.

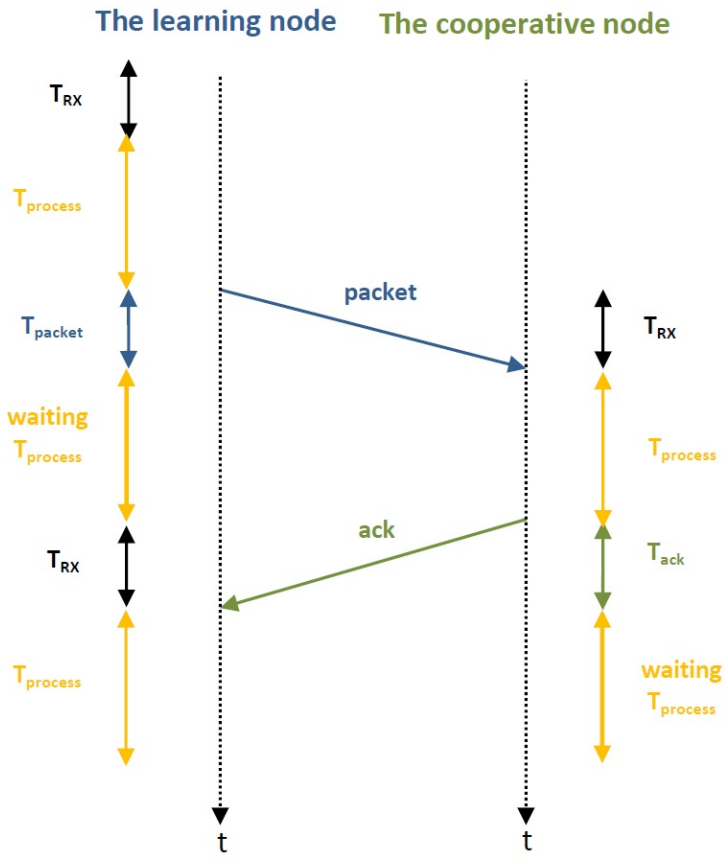


Fig. 4: Cooperation based on stop and wait protocol

	Jammer detectable by the learning node	Jammer hidden to the learning node
Classical fixed channel selection	69%	69%
Best channel selection without learning	76%	69%
Learning without cooperation	87%	69%
Learning with cooperation	94%	82%

Table 5: Implementation results: Packet Success Rate against a sweeping jammer

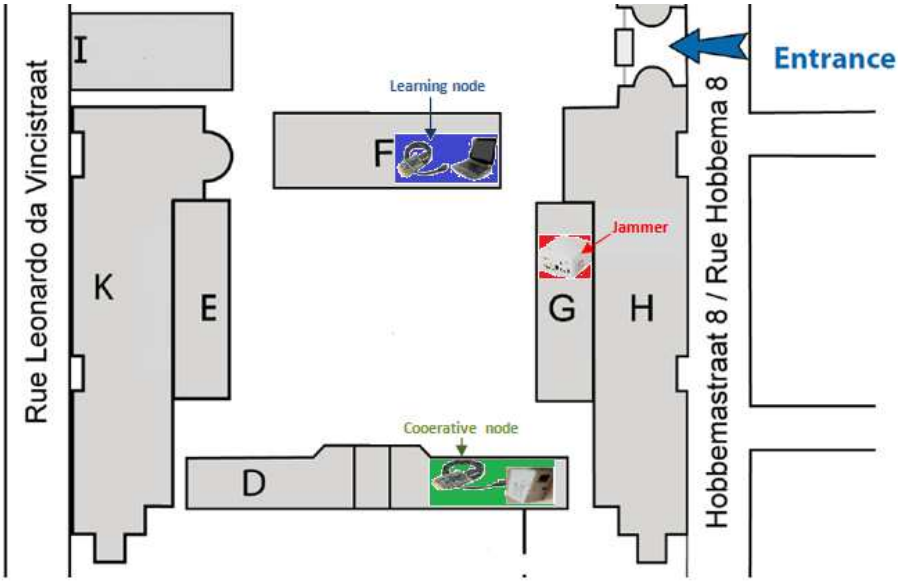


Fig. 5: Scenario1

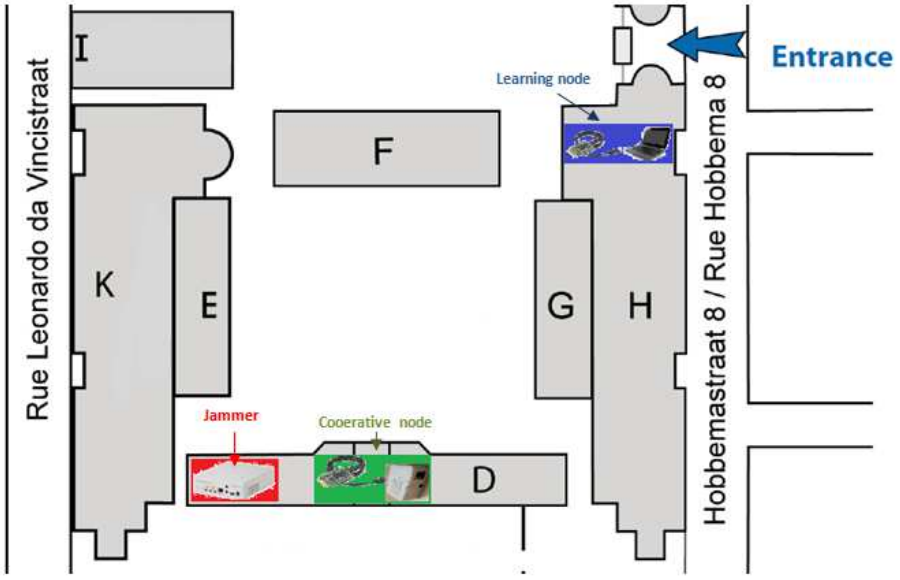


Fig. 6: Scenario2

6 Conclusion

In this paper, we have modeled the cognitive radio jamming attack as a Markov decision process with unknown transition probabilities and rewards. We have proposed an on-policy synchronous Q-learning algorithm

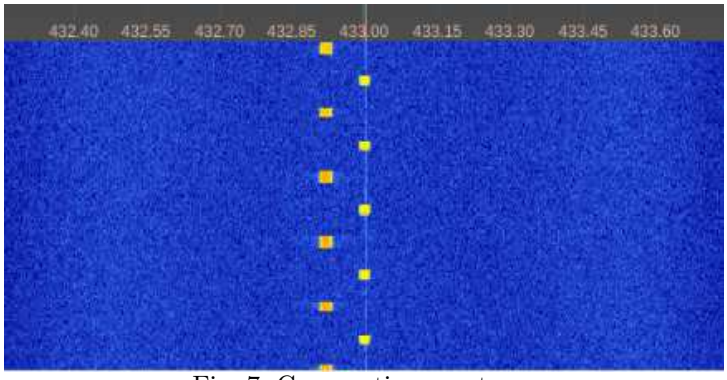
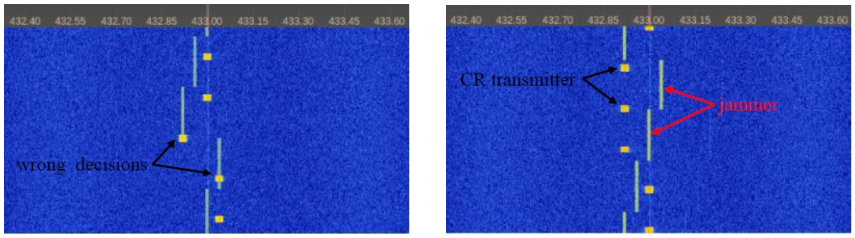


Fig. 7: Cooperation spectrum



(a) (b)

Fig. 8: Best channel selection based on sensing (a) versus channel selection based on learning (b) against a sweeping jammer

based on wideband spectrum sensing and greedy policy to pro-actively avoid the jammed channels. The wideband spectrum sensing speeds up the learning process and the greedy channel selection reduces the packet loss rate. We have proposed an enhancement of the proposed learning algorithm based on the cooperation with the receiving cognitive radio. This latter acknowledges each packet reception and transmits its sensing results to the CR learning node who exploits this information in the update of the Q values.

Simulation results and measurements using real radio equipment are given in terms of packet success rate. We have considered sweeping, pseudo random and reactive jammers. This latter is able to do spectrum sensing in order to detect and interfere the channel carrying the packet. For the real measurements, we have used the universal software defined radio (USRP) platform and Qt Creator/C++ development environment. The results have shown that the channel selection based on the proposed learning algorithm achieves a higher packet success rate than the best channel selection based just on sensing. The results are even better when the learning CR cooperates with the CR receiving the packets to detect the jammer and update the Q values. The proposed solution is applicable not only to avoid ma-

licious interferes and provide continuous reliable communication, but also for the CR coexistence with incumbents.

References

1. Mitola III, J. and G.Q. Maguire Jr, Cognitive radio: making software radios more personal, IEEE Personal Communications Magazine, 6, 13-18, (Aug. 1999)
2. S. Haykin, Cognitive radio: brain-empowered wireless communications, IEEE Journal on Selected Areas in Communications, 23, 201-220 (Feb 2005)
3. Fayaz Akhtar and Mubashir Husain Rehmani and Martin Reisslein, White space: Definitional perspectives and their role in exploiting spectrum opportunities, Telecommunications Policy, 40, 319 - 331 (2016)
4. Yasir Saleem and Farrukh Salim and Mubashir Husain Rehmani, Routing and channel selection from cognitive radio networks perspective: A survey, Computers Electrical Engineering, 42, 117 - 134 (2015)
5. Wang, Wenjing and Bhattacharjee, Shameek and Chatterjee, Mainak and Kwiat, Kevin, Collaborative jamming and collaborative defense in cognitive radio networks, Pervasive and Mobile Computing, 9, 572-587, (2013)
6. Asterjadhi, Alfred and Zorzi, Michele, JENNA: a jamming evasive network-coding neighbor-discovery algorithm for cognitive radio networks, IEEE Wireless Communications, 17, 24-32 (2010)
7. Victor Balogun, Anti-jamming Performance of Hybrid FEC code in the Presence of CRN Random Jammers, International Journal of Novel Research in Engineering and Applied Sciences (IJNREAS), 1 (2014)
8. Suman Bhunia and Xing Su and Shamik Sengupta and Felisa J. Vázquez-Abad, Stochastic Model for Cognitive Radio Networks under Jamming Attacks and Honeypot-Based Prevention, Distributed Computing and Networking - 15th International Conference (ICDCN '14), Coimbatore, India, 438-452 (January 4-7 2014)
9. Wang, Beibei and Wu, Yongle and Liu, K. J. Ray and Clancy, T. Charles, An Anti-Jamming Stochastic Game for Cognitive Radio Networks, IEEE Journal on Selected Areas in Communications (2011)
10. Kresimir Dabcevic and Alejandro Betancourt and Lucio Marcenaro and Carlo S. Regazzoni, A fictitious play-based game-theoretical approach to alleviating jamming attacks for cognitive radios, Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference (2014)
11. Wednel Cadeau, Xiaohua Li, Chengyu Xiong, Markov Model Based Jamming and Anti-Jamming Performance Analysis for Cognitive Radio Networks, Communications and Network (2014)
12. Muhammad Amjad and Fayaz Akhtar and Rehmani, Mubashir Husain and Martin Reisslein and Tariq Umer, Full-Duplex Communication in Cognitive Radio Networks: A Survey, IEEE Communications Surveys and Tutorials,(2017)
13. Jia, Luliang and Yao, Fuqiang and Youming, Sun and Xu, Yuhua and Feng, Shuo and Anpalagan, Alagan, A Hierarchical Learning Solution for Anti-jamming Stackelberg Game with Discrete Power Strategies, IEEE Wireless Communication Letters,(2017)
14. Szepesv´ari, Csaba and Littman, Michael L., Generalized Markov Decision Processes: Dynamic-programming and Reinforcement-learning Algorithms, Brown University, Providence, RI, USA (1996)
15. A. Galindo-Serrano and L. Giupponi, Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks, IEEE Transactions on Vehicular Technology, 59, 1823-1834 (May 2010)
16. Yongle Wu and Beibei Wang and K. J. Ray Liu, Optimal Defense against Jamming Attacks in Cognitive Radio Networks Using the Markov Decision Process Approach, GLOBECOM'10, 1-5 (2010)

17. Chen, Changlong and Song, Min and Xin, Chunsheng and Backens, Jonathan, A game-theoretical anti-jamming scheme for cognitive radio networks, IEEE Network, 27, 22-27 (2013)
18. F. Slimeni and B. Scheers and Z. Chtourou and V. Le Nir, Jamming mitigation in cognitive radio networks using a modified Q-learning algorithm, International Conference on Military Communications and Information Systems (ICMCIS), 1-7 (May 2015)
19. Sutton, Richard S. and Barto, Andrew G., Introduction to Reinforcement Learning, MIT Press, Cambridge, MA, USA (1998)
20. Watkins, Christopher John Cornish Hellaby, Learning from Delayed Rewards, King's College, Cambridge, UK (May 1989)
21. Abounadi, Jinane and Bertsekas, Dimitri P. and Borkar, Vivek, Stochastic Approximation for Nonexpansive Maps: Application to Q-Learning Algorithms, SIAM J. Control Optim., 41, 1-22 (Janv 2002)
22. Even-Dar, Eyal and Mansour, Yishay, Learning Rates for Q-learning, J. Mach. Learn. Res., 5, 1-25 (2004)
23. H. Sun and A. Nallanathan and C. X. Wang and Y. ChenS, Wideband spectrum sensing for cognitive radio networks: a survey, IEEE Wireless Communications, 20, 74-81 (April 2013)
24. V. Le Nir and B. Scheers, Evaluation of open-source software frameworks for high fidelity simulation of cognitive radio networks, International Conference on Military Communications and Information Systems (ICMCIS), 1-6 (May 2015)