Improving Reinforcement Learning Algorithms for Dynamic Spectrum Allocation in Cognitive Sensor Networks

[1]Leonardo Roveda Faganello, [1]Rafael Kunst, [1,2]Cristiano Bonato Both
[1]Lisandro Zambenedetti Granville, [1]Juergen Rochol
[1]Federal University of Rio Grande do Sul (UFRGS)
[2]University of Santa Cruz do Sul (UNISC)
{lrfaganello, rkunst, cbboth, granville, juergen}@inf.ufrgs.br

***Abstract*—Cognitive Radio Networks enable a higher number of users to access the spectrum of frequency simultaneously. This access is possible due to the implementation of dynamic spectrum allocation algorithms. In this context, one of the main algorithms found in the literature is the reinforcement learning based approach called Q-Learning. Although been widely applied, this algorithm does not take into account accurate information about the behavior of users neither the channel propagation conditions. In this sense, we propose three improvements to the dynamic spectrum allocation algorithms based on reinforcement learning for cognitive sensor networks. Simulation results show that all the proposed algorithms allow allocating channels with up to 6dB better quality and 4% higher efficiency than Q-Learning.**

I. Introduction

Cognitive Radio (CR) networks have been proposed to deal with the scarcity of frequency spectrum. This scarcity is caused mainly because of increasing demands for network resources, especially from emerging wireless applications, such as smart grid, broadband cellular access, and industrial sensor systems [1]. CR networks allow the coexistence of two types of users. The first one, called primary, holds the rights of transmission over a well-specified frequency. The second one, known as secondary, accesses the spectrum opportunistically, *i.e.*, taking dynamic advantage of underutilized frequencies [2]. The coexistence of primary and secondary users enables a higher number of users accessing the spectrum of frequencies simultaneously and, consequently, the scarce spectrum is better occupied. Still, to properly realize CR networks, it is fundamental to design and implement optimized allocation algorithms to organize the transmissions among the myriad of multiple simultaneous users.

In the process of deploying CR networks, spectrum allocation policies and regulations are being proposed and enforced by the governments worldwide. For example, in the United States, the release of the National Broadband Plan, the publication of final rules for TV white spaces, and the ongoing procedures for secondary use of the 2360 - 2400 MHz band for medical body area networks will allow more flexible and efficient use of spectrum [1]. Other possible applications of CR include the implementation of opportunistic spectrum sharing among wireless sensors, such as in IEEE 802.15.4 industrial networks [3] and health care applications [4].

Key research efforts have been carried out to handle spectrum allocation in CR networks. Zhou *et al.* [5] propose a probabilistic resource allocation approach to exploit the flexibility offered by CR networks. Their algorithm, however, only considers the channel availability, which is computed by sensing the spectrum. No mention is made about the accuracy of the spectrum sensing function. In addition, details of the channel behavior model by the users are not defined. Zhao *et al.* [6], on their research, propose to allocate spectrum for secondary users traffic based on genetic algorithms. These algorithms are fed by a channel availability matrix that, unfortunately, does not take into account neither the Signal to Interference plus Noise Ratio (SINR) for the conditions of the wireless channel, nor the available channel capacity. Other efforts [7] [8] [9] have also been proposed, but they also have limitations because: (i) the models for representing the behavior of users over the channels are not properly defined or (ii) the channel conditions are not considered.

In this paper, we initially introduce an accurate model for the behavior of the wireless channel users. Afterwards, we review Q-Learning [10], which is one of the main wireless channels allocation algorithms found in the literature. Taking Q-Learning as a starting point, we then propose three new approaches using the historical behavior of users and the channel conditions. First, we introduce *Q-Learning+*, a Q-Learning adaptation to use the accurate information about the channel availability. Second, we present a new reinforcement learning-based allocation algorithm that considers SINR of different transmission cells, called *Q-Noise*. Finally, we propose *Q-Noise+*, which takes into account both the information about the behavior of users and the channel conditions to select the best available channel. Results obtained through simulations show that our proposed algorithms present better performance in terms of allocation than the traditional Q-Learning, selecting the best channel, *i.e*, the one with the highest relation between signal quality and underutilization.

The remainder of this paper is organized as follows. In Section II, we present background aspects on CR, a system model

978-1-4673-5939-9/13/$31.00 ©2013 IEEE

to spectrum allocation, and a reinforcement learning algorithm for spectrum decision. In Section III, we present and discuss three improvements upon the previous algorithm. In Section IV, we introduce and discuss the obtained results, comparing the traditional Q-Learning, with the proposed algorithms. Finally, the paper is closed in Section V, where conclusions and directions for future investigations are presented.

II. Background on Cognitive Radio

The demand for wireless services is increasing because the number of wireless network users has been growing in the recent past. WirelessHART networks based on the Highway Addressable Remote Transducer (HART) Protocol using IEEE 802.15.4 standard are examples of applications that require large portions of spectrum frequencies on industrial scenario [3]. Current spectrum concession policy also contributes to the critical problem of spectrum scarcity. Often, the spectrum allocation is defined by governmental regulatory agencies, such as the Federal Communications Commission (FCC) in the United States, that statically defines two types of accesses. The users of the first type are called licensed and they hold the rights of using a predetermined portion of the frequency spectrum. The second type of users is only able to transmit in Industrial, Scientific, and Medical (ISM) frequency ranges, which are free to be used by any kind of wireless device in a license-exempt manner [2]. These last users are called secondary.

Even though most of the licensed spectrum frequencies are currently allocated, in practice they are sporadically used by primary users. This leads to the rather obvious underutilization of the frequency spectrum. A possible approach to deal with this problem is to implement the so-called Dynamic Spectrum Access (DSA) techniques. CR is an enabling technology of DSA by allowing secondary network users to opportunistically access licensed spectrum frequencies, through spectrum sharing techniques [11].

The shared nature of the channel, however, demands the coordination of transmission attempts between primary and secondary users. To define whether to attempt the transmission in a given channel, secondary users must implement spectrum decision, an important function to allow secondary users to decide which is the best channel among the available ones. Moreover, this decision must lead to minimum interference in adjacent channels. In this context, the spectrum decision must take into account other aspects, such as the activities of other users in the CR network and channel conditions [2]. In the next subsections, we present background aspects on both the utilized system model and the spectrum decision algorithm.

A. System Model

In this paper, we use an industrial scenario model where different devices share the same frequency range, as depicted in Fig. 1. In these scenarios, wireless-enabled devices, such as laptops, tablets, smart phones, and WiFi access points may share the spectrum with sensors as temperature gauge. Although each of these devices transmits in a different channel, interference may occur because of the shared nature of the spectrum.

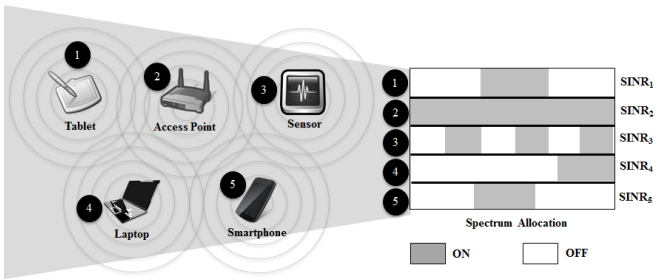

Fig. 1. Industrial Scenario

Another important aspect is the occasional underutilization of some channels because the transmissions are typically not continuous. The underutilized frequency ranges can be opportunistically accessed by other devices (secondary users) using DSA techniques. In this case, the spectrum occupancy model must generate accurate information on temporal and frequency behavior of primary users. Ghosh *et al.* [12] propose a statistical wireless spectrum occupancy modeling, where the utilization (ON) and the idle (OFF) period of channels are governed by two independent Poisson processes.

To model an industrial scenario, we argue that, besides considering the behavior of the primary users, it is fundamental to consider the interference among channels. To consider interference, we model the SINR that is caused by simultaneous transmissions. SINR is defined as described in Equation 1.

$$SINR = \frac{P}{I + N} \tag{1}$$

where $P = (\Psi_s / PL_{s,r})$ is the received signal power, $I = \sum_{i \neq s} (\Psi_s / PL_{i,r})$ is the power of all simultaneous interference, and N is the environmental noise considering the occurrence of Additive White Gaussian Noise (AWGN). Moreover, Ψ_s is the transmission power of device s, and $PL_{s,r}$ represents the Path Loss for a fixed range. Typically, I has a fixed positive value.

B. Spectrum Decision Algorithm

Spectrum decision gained a lot of attention in the research community in the recent past. Centralized and distributed spectrum decision approaches have been proposed, considering both cooperative and non-cooperative techniques. Many types of algorithms are able to deal with the diverse spectrum decision approaches. Important examples include genetic algorithms, game theory, pricing and auction, local bargaining, and graph coloring [6]. All these approaches use artificial intelligence techniques to decide how to implement spectrum decision.

Another important type of artificial intelligence algorithm used in spectrum decision is reinforcement learning [13]. Q-Learning, originally proposed by Watkins [10], is one of the most important algorithms based on reinforcement learning.

Q-Learning also considers a set of states $X = x_1, x_2, ..., x_n$ that can be correlated with a set of actions $A = a_1, a_2, ..., a_n$. Based on the current state x_t ($x_t \in X$), a learner chooses an action a_t to perform ($a_t \in A$). This action causes a state transition according to a probability $P_{xy}(a)$. The new state $x_{t+1} = y \in X$ allows the learner to calculate a reward. This process is repeated until the reward is maximized considering the goal of the learner, *i.e.*, the metric to be optimized.

In DSA, the main goal of the secondary users is to maximize the quantity of transmissions during a given time interval. A typical approach towards this goal is to select a frequency band that has a high probability of being underutilized during this period of time. Q-Learning can be applied in this context to select such spectrum band aiming to minimize the need for changing the transmission band and consequently to avoid spectrum hand-offs. Q-Learning decision is based on a set of channels C with identical bandwidth. Although in practice the effective throughput of a channel is also affected by impairments, such as noise and interference, Q-Learning considers only the probability of collisions as a decision metric. In other words, a transmission attempt is considered successful if no collision between secondary and primary user is detected.

Decisions taken by the Q-Learning algorithm considers cumulative rewards calculated based on a set of actions and rules that cause transitions in a set of environment states. The goal of the algorithm is to maximize the obtained reward r_t in the state x, after executing action a in the instant of time t. This reward can be calculated for different performance metrics, such as throughput, delay, and jitter. The total time of transmission is divided into T epochs of transmission, with duration t_D, where $t_D = 1, 2, ..., n$ and corresponds to the number of transmission attempts in the epoch T. The reward obtained in an epoch T is only calculated after all t_D transmission are tried. The number of successful transmission is represented by N_D, where $0 \leq N_D \leq t_D$ for $t_D \neq 0$. Thus, r_t obtained when a channel C_i is selected in the instant of time t is given by Equation 2.

$$r_t = \frac{N_D}{t_D} \tag{2}$$

The selection of a channel C_i corresponds to an action taken in the instant of time t (a_t) that lasts one epoch, *i.e.*, it receives a reward $r_t(a_t)$ at the instant of time $t + 1$. The action to be taken is selected based on Q-value, which is stored in a structure called Q-table. This value is updated at the end of each epoch according to Equation 3.

$$Q_{t+1(a_t)} = (1 - \alpha)Q_t(a_t) + \alpha r_t(a_t) \tag{3}$$

where $0 \leq \alpha \leq 1$ is the learning rate of the algorithm. In practice, the higher α, more weight is given for the reward obtained in the last epoch. On the other hand, a lower value of α gives more importance to the information about the channel occupancy in detriment to the relevance of the last action.

III. Improving Reinforcement Learning Algorithms in DSA Context

Q-Learning approach considers the information about the historical channel occupancy, as explained before. However, in the original implementation of the algorithm there are two limitations that can lead to imprecise decisions. The first one is related to the nonexistence of a limit of epochs to be considered as accurate information. As the occupancy pattern of a given channel may change over time, a limit of epochs to be considered can improve the decisions of Q-Learning. To deal with this problem, we propose an adaptation called Q-Learning+. This improvement is presented in Subsection III-A. The second one is the impossibility of considering the channel conditions to take a decision. Considering the channel conditions is important because an available channel may be noisy and therefore not useful for transmissions. In order to consider the occurrence of noises and interference, we introduce a new algorithm called Q-Noise, which is presented in Subsection III-B. Moreover, we propose the integration between Q-Learning+ and Q-Noise to take into account both a precise information about the channel occupancy and the SINR of different channels. We call this algorithm of Q-Noise+, which is presented in Subsection III-C.

A. Q-Learning+

Q-Learning algorithm applies a reward-based approach which takes into account two criteria: (i) the transmission successful rate in the last epoch, and (ii) the sum of the successful rates of all past epochs. The importance of the accurate information is calculated as a complement of the learning rate, *i.e.*, $1 - \alpha$. However, except the most recent epoch, the remaining precise information is considered to have the same weight in the decision process. Such approach can lead to imprecise decisions, especially in scenarios where a high number of epochs are used, as for example, industrial sensor networks. In these scenarios, the occupancy rate of a channel may suffer considerable variations, thus we claim that not the whole information should be considered, but only the epochs which are relevant to reflect the current occupancy rate.

In order to allow such approach, we propose an alteration to Q-Learning, which we call Q-Learning+, that allows the algorithm to consider a finite amount of past epochs. Moreover, our proposed solution is able to decrease the weight of a specific epoch as it gets older. On the other hand, the newer the epoch, the higher is its weight. This is possible because we define a look back value (l) that indicates the amount of epochs to be considered. An one-dimensional matrix with length l is also defined to store the weight of each analyzed epoch. Considering this information, the Q-value at the instant of time $t + 1$ is calculated as expressed in Equation 4.

$$Q_{t+1}(a_t) = (1 - \alpha) \sum_{i=1}^{l} [w_{t-i} r_{t-i}](a_t) + \alpha r_t(a_t) \tag{4}$$

where w_i represents the weight of the last l instants of time and r_i is the reward calculated based on $l + 1$ actions. It is

important to point that the sum of all the elements of the matrix is equal to 1, *i.e.* $\sum_{i=1}^{l} w_i = 1$.

B. Q-Noise

Although Q-Learning+ is designed to improve the efficiency of the original approach, it keeps only considering the amount of successful transmissions as a decision metric, ignoring the propagation conditions of the channel. In order to consider also the channel conditions, we propose a new algorithm called Q-Noise. This algorithm considers the transmission quality as a secondary metric, which is calculated according to the SINR measured in a given channel. Q-Noise approach tries to avoid the selection of a channel when it is available but noisy. In this case, both Q-Learning and Q-Learning+ would elect this channel as a transmission candidate, but the transmission quality should be unacceptable.

Q-Noise deals with two criteria to take the decision. The first one is exactly the same used by Q-Learning, *i.e.*, the learning rate considering the reward obtained in an epoch T. The second one is the novelty of Q-Noise, since a quality criteria is included to consider both the SINR level of the channel, and the importance of the SINR for a given transmission. In other words, this new approach is able to calculate a weighted reward in the last epoch with respect to the trade-off between channel availability and transmission quality. The Q-value considering both criteria is calculated as in Equation 5.

$$Q_{t+1} = (1-\alpha)Q_t + \alpha r_t(a_t) + (S_W * \eta) \quad (5)$$

In this equation, two new terms are included, S_W and η. S_W ($0 \geq S_W \geq 1$) represents the weight of the SINR in the calculated reward. This weight is a parameter that defines the importance given to the quality of transmission, *i.e*, the higher the S_W, the higher the impact of the SINR on the Q-value of the channel. In addition, case S_W is defined as 0, the Q-Noise returns to the traditional Q-Learning, because the condition of the channel is not considered. Furthermore, η is a value which corresponds to the intervals of SINR, as described in Table I. The values defined for η have been chosen to change the Q-value according to the channel conditions. In good propagation conditions, η will be higher, increasing the Q-value of the channel. On the other hand, as the channel conditions get worse, η decreases, unchanging the Q-value.

C. Q-Noise+

After the Q-Learning+ and Q-Noise proposals, we introduce the integration between both algorithms, called Q-Noise+. The aim of this integration is to consider both the accurate information about the historical behavior of users and the channel conditions to select the best available channel. This is possible because to calculate the new Q-value we take into account the historic of channel occupancy and the SINR level as expressed in Equation 6.

$$Q_{t+1}(a_t) = (1-\alpha)\sum_{i=1}^{l}[w_{t-i}r_{t-i}](a_t) + \alpha r_t(a_t) + (S_W * \eta) \quad (6)$$

We define η according to the approach presented by Rappaport [14], where, for example, the SINR value of a channel may vary between 15dB and 25dB, with a mean of 18dB. In our methodology, the initial SINR of each channel is randomly selected inside the referred range of values. Based on the SINR value, a AWGN model is applied to each channel. After applying AWGN, the result is normalized to define η, according to the correspondences presented in Table I.

TABLE I
NOISE LEVEL CORRESPONDENCE

SINR value	**Corresponding η**
SINR $<$ 15dB	0
15dB $\leq$ SINR $<$ 17dB	0.25
17dB $\leq$ SINR $<$ 20dB	0.5
20dB $\leq$ SINR $<$ 25dB	0.75
SINR $\geq$ 25dB	1

Case S_W is defined as 0, the Q-Noise+ changes to Q-Learning+ because it only considers information about the historic occupancy of the channel, without considering the condition of the channel. In the next section, we define a simulation scenario and present results comparing Q-Learning with the proposed algorithms.

IV. PERFORMANCE EVALUATION

The main focus of this section is to discuss the improvements provided by our approaches to allow DSA in the context of wireless cognitive sensors to industrial scenarios. This discussion is based on a performance evaluation of the three proposed algorithms in comparison to Q-Learning. The section is organized as it follows. First we present the evaluation methodology in Subsection IV-A. Second we present and discuss the results in Subsection IV-B.

A. Evaluation Methodology

Our methodology is organized in two phases. First, the statistical wireless spectrum occupancy modeling proposed by Ghosh *et al.* [12] was used to present an industrial scenario. The model proposed by Ghosh considers ON/OFF periods defined by two independent Poisson processes with arrival rates extracted from real data measurements, using the Universal Software Radio Peripheral (USRP). Second, the default values for the improving reinforcement algorithms were defined as presented in Table II.

The α parameter is defined with 0.6 as its default value to analyze 5 channels during 100 transmission attempts. In addition, for each 5 transmissions, which represent the duration of an epoch, the Q-value of the selected channel is updated. The three proposed algorithms use different additional parameters to calculate its Q-values. The Q-Learning+ and the Q-Noise+ use 3 and [0.7 0.2 0.1] as default values for l and w, respectively. In addition, the Q-Noise and Q-Noise+ use 0.7 as default value for S_w. After updating the Q-value, there is an ϵ coefficient of 25% to explore a random channel regardless

TABLE II
DEFAULT VALUES FOR SIMULATIONS

Parameter	Default value
Learning rate (α)	0.6
Number of channels (n)	5
Transmission attempts (t)	100
Exploration coefficient (ϵ)	0.25
Epoch duration (e)	5
Threshold between Q-values to perform spectrum hand-off (β)	0.1
Look back - Q-Learning+ and Q-Noise+ (l)	3
Historic weight - Q-Learning+ and Q-Noise+ (w)	[0.7 0.2 0.1]
SINR weight - Q-Noise+ and Q-Noise (S_w)	0.7
Confidence interval	95%

of having a low or high Q-value. Moreover, the reinforced learning algorithms use a threshold of 0.1 as default value to evaluate the possibility of spectrum hand-off. Furthermore, all the results were obtained with a confidence interval of 95%.

B. Results

In this article we analyze four metrics to evaluate the performance of the proposed solutions in comparison with the traditional Q-Learning. First, we analyze the allocation efficiency of the proposed solutions when the number of transmission varies, as can be observed in Fig. 2. The graph shows that as the number of transmission increase, the algorithms efficiency is improved. All four algorithms have similar performance after 10^4 transmissions. However, Q-Learning+ and Q-Noise+ attain up 4% higher efficiency in the start of transmissions. This behavior shows that our proposals are able to learn faster than the traditional Q-Learning because in our approach the accurate information about the historic of the channel usage is previously known. Thus Q-value calculation obtains more precise results than Q-Learning, leading to better performance.

The second approach discussed in this paper is related to the allocation efficiency of the algorithms when the number of available channels is changed. The graph shows that the performance of the algorithms is very sensible to this metric. As the number of available channels increases, the algorithms are able to find more underutilized portions of the channels during transmissions. Q-Learning+ and Q-Noise+ present the best results, with a gain of about 3%, because in these approaches the amount of information allows a more precise analysis of the behavior of the users to perform spectrum allocation.

The third aspect discussed is the probability of successful transmission without any prior knowledge about Q-value (ε), as showed in Fig. 4. In this scenario, the proposed algorithms that consider information about the historic behavior present

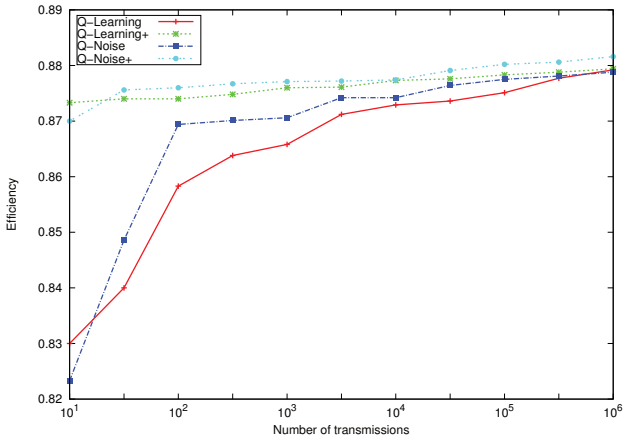

Fig. 2. Number of transmissions x allocation efficiency

worst results if compared to the traditional Q-Learning and Q-Noise. This behavior happens because the main approach for these two proposals is to consider the accurate information collected during previous transmissions in a given channel. Differently from Q-Learning and Q-Noise, these algorithms do not take advantage in randomly exploring the available channels. This design decision leads to no change in performance in situations where there is no previous knowledge about the behavior of the users in the channels. On the other hand, Q-Noise improves its performance overcoming the traditional Q-Learning in approximately 2%. In addition, the Q-Noise selects the channel with the best wireless conditions, enabling a higher quality of transmission.

The final metric evaluated in this paper is channel quality and its influence in the decision of the algorithms. In this graph the results related to Q-Learning+ were not plotted because they would be the same as those obtained using Q-Learning, since none of these algorithms consider the channel propagation conditions to calculate Q-value. Fig. 5 shows that Q-Noise and Q-Noise+ reach an average SINR which

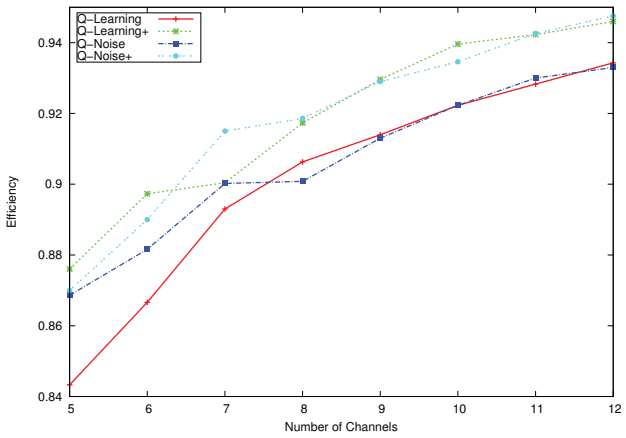

Fig. 3. Number of channels x allocation efficiency

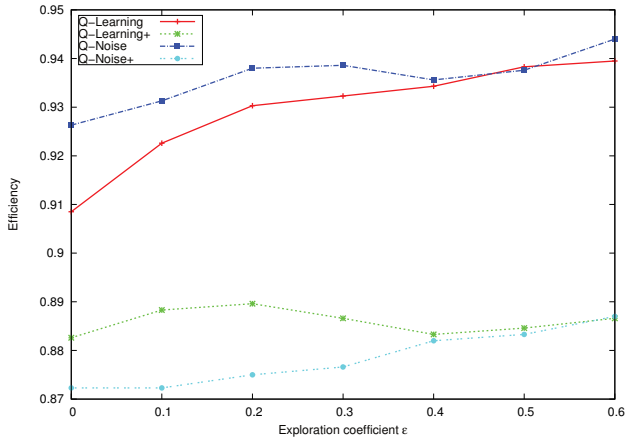

Fig. 4. Exploration coefficient (ε) x allocation efficiency

is higher than that obtained by Q-Learning. We can conclude that the probability of successful transmissions is higher in our proposals than in the traditional Q-Learning, allowing better quality of the overall transmissions in the CR networks. The gains obtained by Q-noise are around 6dB in the best case.

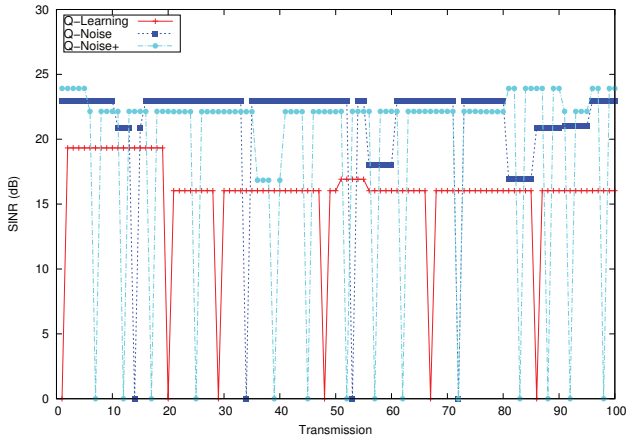

Fig. 5. SINR during transmissions using Q-Learning, Q-Noise and Q-Noise+

V. Conclusion

In this article we introduced a model that describes the behavior of wireless channel users in the context of CR sensor networks. Using this model and Q-Learning as starting points, we proposed three new approaches to improve the reinforcement learning algorithms to provide DSA among wireless sensors. The first approach, Q-Learning+ considers accurate information about the wireless channel users behavior to take decisions. The second one, named Q-Noise considers the channel propagation conditions by analyzing the SINR, and finally the third proposal, called Q-Noise+ considers the integration between the historic of the behavior of the users and SINR of the channels to allow more precise decisions.

Regarding the performance of the proposals, we can conclude that our approaches are able to improve the results obtained by reinforcement learning algorithms. In the context of CR sensor for industrial networks, the improvement in the performance is related to the accurate information of the historic behavior of the users and the wireless channel conditions for the algorithms to take more precise decisions. The results show that our proposals are able to learn faster than the traditional Q-Learning approach, obtaining better performance even in situations where a few number of transmissions is available. As the number of transmissions increases, the difference in the performance in comparison with Q-Learning diminishes, but always remains better.

Directions for future investigations include cooperative spectrum decision with the goal of improving even more the DSA. This approach may allow more precise decisions, however it may introduce network overhead caused by the message exchange necessary to share information among the sensors. The trade-off between more precise decisions and network overhead must also be analyzed. Another possible future investigation involves aspects related to the quality of service provided for CR networks applications. For example, solutions related to medical sensors are sensitive to delay and jitter, thus it is important to allow quality of service enabled solutions in this context.

References

[1] J. Wang, M. Ghosh, and K. Challapali, "Emerging cognitive radio applications: A survey," *IEEE Communications Magazine*, vol. 49, no. 3, pp. 74–81, March 2011.

[2] I. Akyildiz, W.-Y. Lee, M. Vuran, and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 40–48, April 2008.

[3] I. Muller, J. C. Netto, and C. E. Perreira, "WirelessHART field devices," *IEEE Instrumentation & Measurement Magazine*, vol. 14, no. 6, pp. 20–25, December 2011.

[4] P. Phunchongharn, E. Hossain, D. Niyato, and S. Camorlinga, "A cognitive radio system for e-health applications in a hospital environment," *IEEE Wireless Communic.*, vol. 17, no. 1, pp. 20–28, February 2010.

[5] X. Zhou, G. Y. Li, D. Li, D. Wang, and A. C. K. Soong, "Probabilistic Resource Allocation for Opportunistic Spectrum Access," *IEEE Transactions on Wireless Communications*, vol. 9, no. 9, pp. 2870–2879, September 2010.

[6] Z. Zhao, Z. Peng, S. Zheng, and J. Shang, "Cognitive radio spectrum allocation using evolutionary algorithms," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4421–4425, September 2009.

[7] G. Ko, A. Franklin, S.-j. You, J.-s. Pak, M.-s. Song, and C.-j. Kim, "Channel management in IEEE 802.22 WRAN systems," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 88–94, September 2010.

[8] Y. Zhang and C. Leung, "Resource allocation in an OFDM-based cognitive radio system," *IEEE Transactions on Communications*, vol. 57, no. 7, pp. 1928–1931, July 2009.

[9] R. Zhang, Y.-C. Liang, and S. Cui, "Dynamic Resource Allocation in Cognitive Radio Networks," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 102–114, May 2010.

[10] C. Watkins and P. Dayan, "Q-learning," *Journal Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.

[11] J. Mitola and G. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13–18, 1999.

[12] C. Ghosh, S. Pagadarai, P. Agrawal, and M. Wyglinski, "A framework for statistical wireless spectrum occupancy modeling," *IEEE Transactions on Wireless Communications*, vol. 9, no. 1, pp. 38–44, January 2010.

[13] T. Jiang, D. Grace, and P. Mitchell, "Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing," *IET Communications*, vol. 5, no. 10, pp. 1309–1317, August 2011.

[14] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.