Reinforcement Learning Approaches and Evaluation Criteria for Opportunistic Spectrum Access

Clément ROBERT[1,2], Christophe MOY[1], Cheng-Xiang WANG[2]
[1]SUPELEC/IETR, Avenue de la Boulaie, 35576 Cesson-Sévigné, France
[2]Institute of Sensors, Signals and Systems, Heriot-Watt University, Edinburgh EH14 4AS, UK.
christophe.moy@supelec.fr, cheng-xiang.wang@hw.ac.uk

***Abstract*—This paper deals with the learning and decision making issue for cognitive radio (CR). Two reinforcement-learning algorithms proposed in the literature are compared for opportunistic spectrum access (OSA): Upper Confidence Bound (UCB) algorithm and Weight Driven (WD) algorithm. This paper also introduces two new metrics in order to evaluate the machine learning algorithm performance for CR: effective cumulative regret and percentage of successful trials. They provide a fair evaluation means for CR performance.**

***Index Terms*—Cognitive radio, opportunistic spectrum access, machine learning, MAB, UCB**

I. Introduction

The spectrum utilization is becoming sub-optimal due to extensive grants of licenses [1]. Such inefficient and inflexible distribution of the spectrum may not be sustainable in the future as the demand of access to the spectrum increases rapidly with the exponential growth of new high-data-rate applications. An alternate providing more flexibility is to introduce CR systems that are aware of the environment and adapt in order to achieve the best possible use of resources [2]. In this paper, we focus on the problem of OSA allowing Secondary Users (SUs) to transmit in available spectrum left blank by Primary Users (PUs) [3]. The major issue for SUs is then to keep interference to PUs to a minimum. In an OSA context, the key point for SUs is to find the best resources available. In [4], a possible solution to such challenge was proposed based on Reinforcement Learning (RL). RL is a way to tackle learning and decision making for CR, especially when no *a priori* knowledge is available on what to learn about (in OSA, this is the channel vacancy rate), as shown in [5].Such RL algorithms learn from the success and failure of past trials in order to predict which frequency channel(s) is (are) probably free in a given band. Using a RL approach has been motivated as the well-known uniform random walk approach has been proven to be the most inefficient way to explore the statistics of a set of solutions [6]. RL relies on the following principle for OSA: only one channel is sensed at a time; each channel is given a note; a reward is granted to a channel when a trial is successful, e.g., the channel is detected free and an opportunistic transmission can be done. If the channel is detected being occupied by a PU, there is no reward and transmission does not occur in order not to jam the PU. Then the initial problem can now be reduced to the maximization of the reward, which was formalized as the Multi Armed Bandit (MAB) issue in [7][8]. This paper is aimed at comparing two RL algorithms proposed in the literature for OSA: UCB algorithm proposed in [4] and WD algorithm described in [9].

The remainder of this paper is organized as follows. In Section II, UCB and WD algorithms are presented. Section III aims at comparing their performance according to the existing RL metrics. But we show that these metrics are not consistent for CR, so that we propose to define new metrics for CR in Section IV. Finally, conclusions will be drawn in Section V.

II. Reinforcement Learning Algorithms

A. *Cognitive Radio context*

A CR system or equipment must run the cognitive cycle as described in [2] in order to adapt its operation to the changes of the environment, e.g., it must include (simplified here for clarity purposes [10]) sensing, decision/learning, and adaptation processes. A CR can be used for any kind of adaptation, including but not limited to spectrum-oriented adaptation. CR can also contribute to green radio for instance in [11]. This paper focuses on the spectrum-oriented OSA which is a special case of CR.

The advantage of RL algorithms for CR is that in such an approach where only one channel is sensed at a time, the Radio Frequency (RF) and digital processing of the radio do not have to support a larger bandwidth than the bandwidth of one channel which is required for the transmission itself. In other words, there is no need for a wideband RF (and the associated digital processing power) to sense all the channels of interest in parallel. As shown in Fig. 1, the proposed OSA radio equipment can be based on a conventional radio with the addition of sensing, decision making, and learning components. Here, we assume that the OSA system is able to change frequency statically, i.e. select a frequency for communications, but does not change during communications. Therefore, 'adapt' component was present in the conventional radio.

In this paper, we focus on the learning and decision process, which aims at:

1. deciding on which channel to try for a transmission,
2. deciding to transmit or not,
3. updating learning information.

978-1-4799-2003-7/14/$31.00 ©2014 IEEE

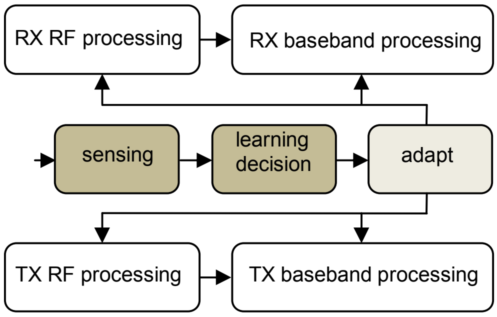

Fig. 1: Cognitive cycle elements to be added to a conventional radio to support the proposed OSA features.

Point 1 uses the results of previous trials which enable to learn the channels' occupancy probabilities. In order to maximize transmission opportunity, it decides to sense the channel with the highest probability of vacancy at next iteration. Point 2 uses the output of the sensing to decide to transmit or not. Whatever the sensing output, point 3 is the learning process that will orient next iteration decision at point 1. Even if no transmission occurs at a stage because the selected channel was not free, the same learning will be preserved. Fig. 2 details how outputs of the sensing block are used for learning and decision making. Outputs of the learning and decision block manage both transmission and reception chains in order to enable the communication at the current iteration. The learning block will also utilize the updated channel knowledge to decide the frequency to be used for next iteration.

B. Model for OSA

In this paper we will consider UCB and WD as RL algorithms. To be fair, both algorithms need to be run using the same model based on the MAB method, which is described as the following. The spectrum is divided in channels denominated by $k \epsilon \{1, 2, \ldots, K\}$, each having the same bandwidth and representing one arm for the RL algorithm. We suppose that time is discrete, slotted in iterations, and only one channel per slot is sensed at each iteration. Moreover we assume the sensing is perfect. The temporal occupancy of every arm follows a Bernoulli distribution θ_k for which the expected value, $\mu_k = E[\theta_k]$, can be set independently. The PU is supposed to be synchronous with SUs. Coordination or cooperation between SUs is not discussed in this study but can be found in [12].

C. UCB algorithm

We define *t* as the discrete time index representing the total number of times that the algorithm has been played. The cumulative number of times that the channel *k* has been chosen in the previous steps is T_k and a_t is the index of the channel chosen at the time index *t* while r_t is the throughput achieved at *t* if *D* bits are transmitted when a channel is free. An independent realization $X_{k,T_k(t)}$ of the statistical distribution θ_k described previously has an empirical sample mean $\overline{X}_{k,T_k(t)}$ and $A_{t,k,T_k(t)}$ is a bias added to the empirical sample mean $\overline{X}_{k,T_k(t)}$ to compute UCB coefficients $B_{t,k,T_k(t)}$.

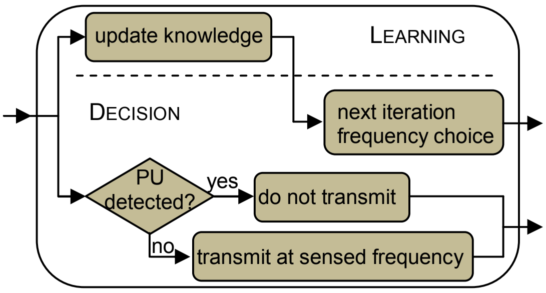

Fig. 2: Learning and decision making processes based on a machine learning approach.

Then UCB algorithm returns the index of the maximum value of $B_{t,k,T_k(t)}$ as indicated in Fig. 3. Indeed $B_{t,k,T_k(t)}$ returns the estimated probability of occupation of all channels, each one upper bounded by its own $A_{t,k,T_k(t)}$ bias. So the SU will choose to transmit at next iteration on the channel having the highest $B_{t,k,T_k(t)}$ index. A consequence is that the best channels are also sensed more than others and the knowledge on their vacancy is closer to the reality.

Adding the bias to the sample mean enables the algorithm to explore the channels that apparently do not provide a lot of communication opportunities (i.e. with a low reward). In fact without bias, the best resource would be picked exclusively while the knowledge on the others would remain uncertain, thus provoking potential divergence if bad luck has artificially lowered first trials on actually "good" channels (and respectively good luck has artificially increased first trials on actually "bad" channels).

Choosing a channel on which the algorithm already has a good knowledge can be related to exploitation. Seeking knowledge on a low potency channel may be assimilated to an exploration process. The particularity of the UCB is that it combines both exploration and exploitation all the time. The amount of each phase can be controlled through the exploration parameter α of the algorithm inside parameter $A_{t,k,T_k(t)}$. In fact, the smaller α is, the more the algorithm relies on past trials. The exploration is therefore reduced and the algorithm is willing to play on well-known channels. The mathematical expression of the bias $A_{t,k,T_k(t)}$, allows us to provide a proof of convergence [8].

```
Parameters:  K, $\alpha$ exploration parameter.
Input:       $\{a_0, r_0, a_1, r_1, \ldots, r_{t-1}, a_{t-1}\}$
Output:      $a_t$
Algorithm:
If
$t < K, a_t = t + 1$
Else
    $T_k(t) \leftarrow \sum_{m=0}^{t-1} \mathbb{1}_{a_m=k}$ ,$\forall k$
    $A_{t,k,T_k(t)} \leftarrow \sqrt{\frac{\alpha \ln(t)}{T_k(t)}}$ ,$\forall k$
    $B_{t,k,T_k(t)} \leftarrow \overline{X}_{k,T_k(t)} + A_{t,k,T_k(t)}$ ,$\forall k$
    Return   $a_t = argmax_k(B_{t,k,T_k(t)})$
```

Fig. 3: UCB algorithm.

If we define the regret of a policy π by $R_t^{\pi} = D\ \mu_k t - W_t^{\pi}$, where D is the number of bits that the cognitive agent can transmit per time slot in one channel, μ_k is the probability that the channel k is free, and W_t^{π} is the cumulated throughput on the time range [0, *t*], we can prove that the regret of UCB is β-consistent (with 0<β<1). This means that UCB algorithm can converge to the optimum set of channels for an infinite time. However, it has been shown by simulations (both with perfect sensing [4] and with sensing errors [13]) and experimentations on real radio signals [14] that convergence to the best channel is very fast (even faster to a set of best channels). This makes UCB accurate for learning into realistic radio conditions even if there are several tens of channels considered [14].

D. *WD algorithm*

The WD algorithm is structured the same way as UCB except from the introduction of a *preferred set*. It also relies on an index given to each channel, called a weight here, based on past trials and measures. Each channel is ranked thanks to this weight which reflects the quality of the resource. The WD algorithm is directly derived for the two stages of RL algorithms proposed in [15]. After each trial, the weight of the channel *k* that has been chosen for transmission is updated as:

$$W_{t+1,k} = W_{t,k} \pm f \qquad (1)$$

where an addition is used if the channel is rewarded and a substraction is used if it is punished. The decision process is based on a statistical distribution based on the weights:

$$P_t(k) = \frac{W_{t,k}}{\sum_{c\epsilon\{1,\dots,K\}} W_{t,c}} \qquad (2)$$

where $P_t(k)$ is the probability that the channel k is chosen. Note that the weight $W_{t,k}$ does not reveal somehow the probability of occupation of the channel as for UCB. Moreover, WD algorithm does not directly select the channel with the highest weight. If the weight of a channel is above a given threshold V_t, then the channel is selected to enter the *preferred set*. When the *preferred set* is full, the choice is restricted to the channels in the set only. In other words, the algorithm moves from exploration to exploitation. The threshold V_t and the size of the *preferred set* allow us to control the trade-off between exploration and exploitation. Here, unlike for UCB algorithm, the two phases are quite separated, as once the *preferred set* is full, exploration only continues on the channels of the set. Another key difference to be pointed out may be that UCB indexes of all channels are updated at each iteration as *t* changes. For WD, the only one change at the iteration *t* is the weight from the chosen channel.

III. Comparison of the performances

To compare the algorithms, two metrics are used. The percentage of choice of the optimal channel is a common method to evaluate the performance of learning algorithms. The cumulative regret is useful because it is directly related to the blocking probability (probability that a channel picked is not free). To guarantee an accurate comparison we analyze the impact of the different control parameters on the performances. We propose to consider a frequency band comprising of 8 channels. The probability of occupancy of each channel used for simulation follows a Bernouilli distribution. Channels are ordered with no loss of generality so that their respective probabilities of vacancy are chosen here as {0.1, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}.

A. *Percentage of choice of the optimal channel*

From the machine learning point of view, a cognitive agent should transmit as much as possible in the optimal channel as it provides the lowest blocking probability for the SU. The quality of a learning algorithm can therefore be evaluated considering the percentage of choice of the optimal channel, which is a usual machine learning criteria.

Fig. 4 compares the percentages of time the most available channel has been chosen within the set of eight channels, as a function of the number of trials, for three UCB algorithms with different values of α (values after UCB in the figure legend) and WD algorithm. Note that α sets the level of exploration. As mathematically proven, UCB converges to the best solution at infinity. However, we can see that after ten thousand trials, all the 4 studied cases play the most available channels more than 70% of time and UCB 1.2 obtains this result after only 1000 to 2000 trials. The higher α, the more the algorithms explore solutions, even if they are not the best. In other words, the lower α, the more confidence UCB makes to its past trials and the less it tries to learn about exotic solutions. That is why the results are converging slower towards the best solution when α is high. However, in terms of optimization, this is a guarantee for UCB to avoid staying in a local minimum and can always converge. Fig. 4 also shows that WD algorithm converges faster than UCB, which will be discussed later.

Fig. 5 is the dual of Fig. 4 for the WD algorithm. Here the parameter influencing the level of exploration is the size of the *preferred set*. In terms of performances, the WD algorithm benefits from the introduction of the preferred set. In the case where the optimal channel and some other good resources are selected for exploitation, the algorithm has more chance to pick the optimal channel as it restricts its exploration range.

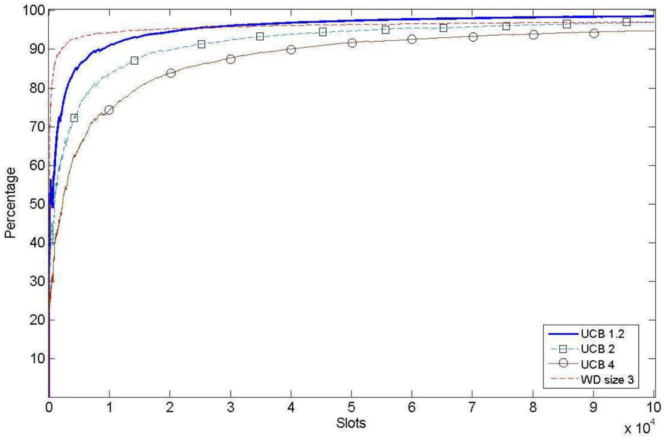

Fig. 4: Percentage of choice of the optimal channel with variable exploration parameter α for UCB algorithm.

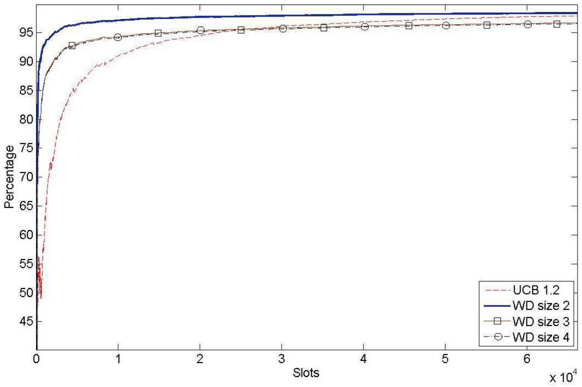

Fig. 5: Percentage of choice of the optimal channel with variable size of the *preferred set* for WD algorithm.

Consequently it overwhelms UCB at the early stage. We remark that for a low value of α and after around 3x10^4 slots, both algorithms have approximately similar performances. Considering a LTE frame iteration period of 1 ms, the matching of the two algorithms would occur after about 30 s.

This does not mean, however, that UCB-based SUs would have reached so much less communication opportunities than WD-based SUs. Selecting another channel than the best one indeed does not mean it is occupied by a PU. In the CR context, selecting the best channel is not the goal, while selecting a good channel may be enough. Although this metric provides a possible criterion to evaluate the performance from a machine learning perspective, it does not make indeed a real sense in the CR context. The most important is to compare the number of opportunities reached for communicating. The introduction of the cumulative regret is a way to take it to the next level. Moreover, it does not convey a clear thought.

B. Cumulative regret

The regret associated with an action is defined in the machine learning community as the difference of the probability of availability of the optimal channel and the probability of availability of the selected one. The summation of the regrets over all the channels and time slots gives the cumulative regret. Such a metric is a better metric for the comparison because it reflects the conflicts with the PU. In fact a high level of regret means that a huge amount of mistakes were made during the learning process compared to the systematic selection of the best channel. This reflects consequently the potential wastes of communication opportunities. We understand now why it is an utmost priority to keep the regret level as low as possible. As in the previous section, the following two figures compare UCB and WD for different parameters α for UCB and size of the *preferred set* for WD. Fig. 6 and Fig. 7 show that the WD algorithm produces a very low regret compared to UCB. Once again, it benefits from the introduction of the *preferred set* that focuses the knowledge on a few good channels. In compensation the WD algorithm cannot provide an equivalent estimation outside of the set. This global lack of knowledge makes it very sensitive to errors in the first few steps when it fills up the set.

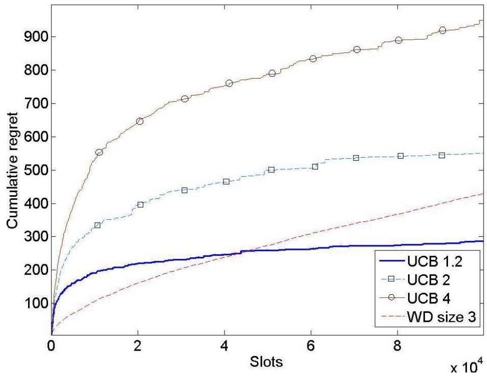

Fig. 6: Cumulative regret with variable exploration parameter for UCB algorithm.

Note that when the UCB's exploration parameter is increased, the cumulative regret is growing since more exploration is done throughout the learning process. Moreover the size of the *preferred set* influences the level of cumulated regret because it bounds the regret when switching to the exploitation phase. In the ideal case for a set of size 3, the best 3 channels are supposed to be selected. Hence the maximum instantaneous regret value is 0.2 (0.9-0.7) while it reaches 0.8 for UCB (0.9-0.1). Note also that WD's cumulative regret is growing faster at the infinite than UCB's. We can explain that by the presence of the set that focuses the exploitation on 3 channels. The algorithm has been structured such that it probes those three channels. On the contrary, UCB is designed to converge towards the optimal channel, which ensures a very low growing rate of the cumulative regret at the infinite.

C. Robustness

The robustness can be defined here as the ability for an algorithm to guaranty a fast convergence towards the best channel. While the UCB algorithm was proven to converge to the optimal channel, WD algorithm is purely empirical, which does not exclude some diverging scenarios. Such cases occur especially when "bad luck" happens at the early stage, while deciding which channels are selected for the *preferred set*. If several channels actually have a high probability to be free, the competition to enter the *preferred set* can provoke some scenarios where the optimal channel remains unpicked.

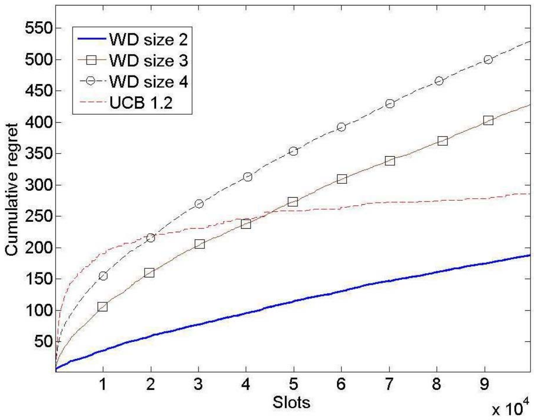

Fig. 7: Cumulative regret with variable size of the *preferred set* for the WD algorithm.

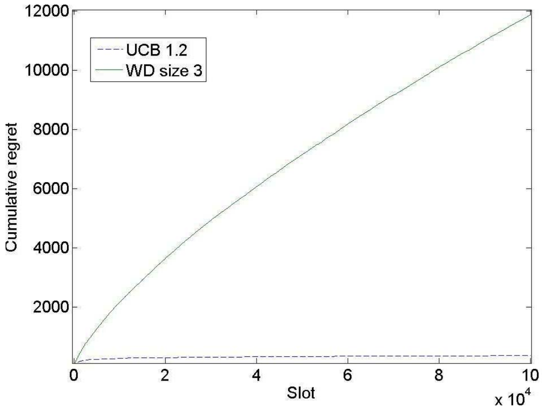

Fig. 8: Cumulative Regret in the case of divergence for the WD algorithm.

The probability of missing the optimal channel increases also when the size of the *preferred set* is reduced. Even if the threshold is increased significantly to extend the exploration phase, the WD algorithm does not explore as efficiently as UCB. Actually, as good resources are granted rapidly of a very high weight value, the probability to explore other channels is very low. Effectively, the more the exploration phase lasts, the less bad resources are probed. WD algorithm is then blocked picking only the good resources he found at the early stages. Fig. 8 shows a case of divergence for the WD algorithm. The temporal vacancy of the channels here follows the following Bernoulli distribution: {0.7, 0.7, 0.7, 0.3, 0.2, 0.6, 0.8, 0.9}. This context is a little bit harder as more than 3 channels (the size of the *preferred set*) have good chances to be free at first trials. So once the algorithm fills its *preferred set*, even another very good channel may be excluded. The cumulative regret of the WD, previously widely below the regret of the UCB is now increasing linearly because the optimal channel was not selected to enter the *preferred set*.

IV. Machine learning analysis in terms of data volume effectively transmitted

The metrics used in the previous section to compare the algorithms are those of the machine-learning community. Yet we showed that they are not suitable to efficiently compare RL algorithms in the context of CR, as the outcomes are different. The most important criterion in CR systems is the amount of data effectively transmitted opportunistically.

A. First new evaluation criteria: effective cumulative regret

The cumulative regret does not exactly reflect truthfully the ability of an algorithm to find transmission opportunities. Not choosing the optimal channel does not necessarily mean that the throughput will be nulled at this stage since other channels might be free. As a result, it would be more relevant to consider another metric that we could call the *effective cumulative regret* or *effective regret*. It is based on the cumulative regret but takes into account the impact of successful trials by nulling the regret when the chosen channel is actually free, even if it is not the optimal resource. The effective cumulative regret R_{eff} is defined as

$$R_{eff} = \sum_t e_{t,a_t}(\theta_{opt} - \theta_{a_t}) \qquad (3)$$

where $e_{t,a_t} = \begin{cases} 0 & \text{if the band } a_t \text{ is free} \\ 1 & \text{if not.} \end{cases}$

This new definition infers that the regret is lowered down when ‘good luck’ happens. It allows us to evaluate more accurately the algorithm’s behavior while not choosing the optimal channel, which occurs at a significant number of times, especially at the early stages. Bringing such probabilistic considerations in the comparison is necessary to evaluate the actual performance of the global CR system.

Fig. 9 shows the simulation results of the effective cumulative regret for both policies. We also displayed on the same plot the classic cumulative regret presented in the previous section. We notice that the growing rate of the effective regret is lowered by the communication opportunities reached out of the optimal channel. The WD algorithm in particular scales very well since it exploits the channels in its *preferred set* more uniformly than UCB. UCB policy tends to pick the optimal channel most of the time and explore the other channels periodically (jumps in the curve). Considering the effective regret will only affect the height of the jumps. We deduce from these results that the gap between UCB and WD at the infinite is finally closing up. If we consider the growing rates between slots 3000 and 10000, we notice that WD’s effective cumulative regret slope has been divided by 4.74 while this same slope is only divided by 3.03 for UCB.

B. Second new evaluation criteria: percentage of successful trials

Following the same idea of using metrics suitable for CR purposes, we suggest the percentage of successful trial. The percentage of choice of the optimal channel is not suitable to fairly estimate one policy’s performances indeed. Choosing the optimal channel does not guarantee that the SU will reach an opportunity of transmission because the channel might be occupied. Moreover picking another channel may also provide a transmission opportunity (imagine the case where several channels are mostly idle).

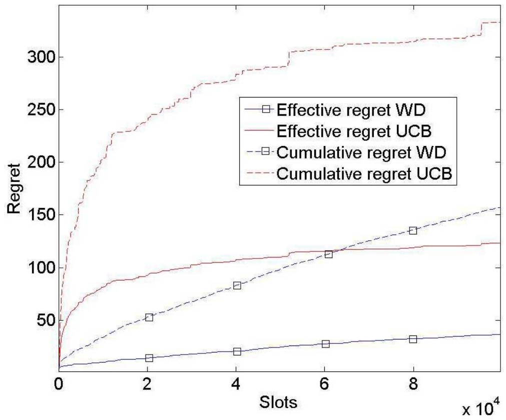

Fig. 9: Effective and classic cumulative regret for both algorithms.

So we consider it is useful to take into account the primary network's usage as part of the metric. It led us to consider the percentage of successful trial that is closely related to the averaged throughput of the system. The metric is constructed as follows:

$$P_{success} = \frac{\sum_t s_t}{t} \qquad (4)$$

with

$$s_t = \begin{cases} 1 & \text{if band } a_t \text{ is free at } t \\ 0 & \text{otherwise.} \end{cases}$$

Fig. 10 shows the percentage of successful trials for both policies. We remark that WD reaches more transmission opportunities at the early stage, which is coherent with the higher percentage of choice of the optimal channel also observed on the same graph. We see that after approximately 3000 slots both policies are equivalent but once again, the WD algorithm is converging faster. The percentage of successful trial is converging towards the maximum channel capacity (although we only consider the percentage of availability here), which is 90%. This metric allows us to analyze both the maximum amount of data that can be transmitted as well as the time that the policy takes to reach it. It is consequently a metric most suitable for a CR perspective. As a conclusion, the new metrics we propose for RL algorithms' performance evaluation mitigate the regret thanks to the effective cumulative regret. They also decrease the non successful trial rate at the beginning of the process while providing a clear idea of communication opportunities. They consequently offer a fair and consistent view on CR matters.

V. Conclusion

The comparison of the WD and UCB algorithms has shown that the main difference of the two RL approaches can be sum-up a trade-off between the ability to have an overall knowledge of the occupancy of the spectrum and the efficiency of convergence. As the WD algorithm prioritizes a fast convergence, it cannot provide the same level of robustness as the UCB algorithm. The more RL algorithms rely on past trials, the faster they can converge to the optimal channel but the risk of divergence is also increased. The choice between the two approaches should mainly be based on a prior estimation of the PU traffic. WD will be preferred if fast convergence is wanted whereas UCB will be picked to provide more robustness in the cases of high uncertainty. We have also defined two new metrics, closer to CR philosophy. We will also propose in the future a mixed solution, combining the advantages of both WD and UCB.

Acknowledgement

Prof. C.-X. Wang gratefully acknowledges the support of his work from the Opening Project of the Key Laboratory of Cognitive Radio and Information Processing (Guilin University of Electronic Technology), Ministry of Education (No. 2013KF01), and the International Science & Technology Cooperation Program of China (No. 2014DFA11640).

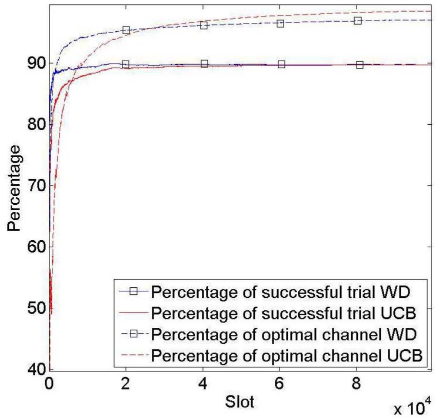

Fig. 10: Percentage of successful trials and percentage of choice of the optimal channel for both algorithms.

References

[1] FCC, "Spectrum policy task force report", Nov. 2002, *http://www.fcc.gov/sptf/files/SEWGFinalReport 1.pdf*

[2] J. Mitola, "Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio," Ph.D. diss., KTH, Sweden, 2000

[3] X. Hong, C.-X. Wang, H.-H. Chen, and Y. Zhang, "Secondary spectrum access networks: recent developments on the spatial models", *IEEE Veh. Tech. Magazine,* vol. 4, no. 2, June 2009.

[4] W. Jouini, D. Ernst, C. Moy, J. Palicot, "Upper confidence bound based decision making strategies and dynamic spectrum access", ICC Conference, Cape Town, South Africa, May 2010.

[5] W. Jouini, C. Moy, and J. Palicot, "Decision making for cognitive radio equipment: analysis of the first 10 years of exploration." *EURASIP Journal on Wireless Communications and Networking,* 2012

[6] S.B. Thrun, "Efficient exploration in reinforcement learning". Technical Report CS-92 – 102, School of Computer Science, Carnegie-Mellon University, 1992.

[7] H. Robbins, "Some aspects of the sequential design of experiments," Bulletin of American Mathematical Society, 58:527–535, 1952

[8] R. Agrawal, "Sample mean based index policies with o(log(n)) regret for the multi-armed bandit problem," Advances in Applied Probability, 27:1054–1078, 1995

[9] T. Jiang, D. Grace, P.D. Mitchell, "Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing", IET Communications, Aug. 2011.

[10] L. Godard, C. Moy, J. Palicot,"An Executable Meta-Model of a Hierarchical and Distributed Architecture Management for the Design of Cognitive Radio Equipments", Annals of Tele-communications, Springer, vol. 64, nb 7-8, Aug. 2009

[11] O. Lazrak, S. Bourbia, C. Moy, D. Le Guennec, P. Leray, K. Grati, A. Gazel, "Management Architecture for Green Cognitive Radio Equipments", Transactions on ETT - special issue on Cognitive Radio, Wiley, to be published end 2013

[12] W. Jouini, M. Di Felice, L. Bononi, C. Moy, "Coordination and Collaboration in Secondary Networks: A Multi-Armed Bandit based Framework", http://arxiv.org/pdf/1204.3005.pdf

[13] W. Jouini, C. Moy, and J. Palicot, "Upper Confidence Bound Algorithm for Opportunistic Spectrum Access with Sensing Errors", CrownCom'11, 1-3 June 2011, Osaka, Japan

[14] C. Moy, "Reinforcement Learning Real Experiments for Opportunistic Spectrum Access", Karlsruhe Workshop on Software Radio, Karlsruhe, Germany, 12-13 March 2014

[15] T. Jiang, D. Grace, Y. Liu, "Two stage reinforcement learning based cognitive radio with exploration control", *IET Commun.*, 2011, 5, (5), pp. 644 – 651.