2013 IEEE 24th International Symposium on Personal, Indoor and Mobile Radio Communications: Workshop on Cognitive Radio Medium Access Control and Network Solutions

Introducing strategic measure actions in multi-armed bandits

Stefano Boldrini*†, *Student Member, IEEE*, Jocelyn Fiorina†, *Member, IEEE*, and Maria-Gabriella Di Benedetto*, *Senior Member, IEEE*
* Department of Information Engineering, Electronics and Telecommunications (DIET)
Sapienza University of Rome, Rome, Italy
E-mail: {boldrini, dibenedetto}@newyork.ing.uniroma1.it
† Telecommunications Department, Supélec, Gif-sur-Yvette, France
E-mail: {stefano.boldrini, jocelyn.fiorina}@supelec.fr

***Abstract*—Multi-armed bandits may be used for modelling the process of selecting one among different wireless networks, given a set of system constraints typically formed by user-perceived network quality indicators. This work proposes a novel multi-armed bandit, that is made appropriate to the above context by introducing a distinction between two actions, to *measure* and to *use*, in order to better reflect real communication application scenarios. The impact of this introduction is analysed through simulations by comparing a traditional multi-armed bandit algorithm against methods that integrate the new concept of measuring vs. using. Results show that performance in terms of regret can be significantly improved using the proposed algorithms if the period needed for measuring is at least 3 times shorter than the one for the using action. The classical method would require a significantly shorter measuring period to reach the same regret, i.e. much stricter constraints on the allowed measure action duration.**

***Index Terms*—Multi-armed bandit, exploration, exploitation, regret, learning, UCB, wireless network selection.**

I. Introduction

In cognitive radio, awareness of the surrounding radio environment is key to enabling cognitive devices to react, adapt, and eventually optimize resource usage and performance, as a function of radio conditions.

The above concept can be extended for modelling the actions involved in selecting one among different available wireless networks, possibly made of different technologies, that may be available in a given geographical area at a given instant in time, based on a criterion of optimality. When the only a priori knowledge, made available at the cognitive device, is formed by a survey of available networks, prediction in performance estimation for each of the available networks may drive the selection decision towards maximum reward.

The above problem falls in a category of classical optimization problems that has been named "Multi-Armed Bandit (MAB)" [1], [2].

In classical MAB, only one action is modelled: the selection action. In real application network selection scenarios, we need, however, to represent at least two steps in the selection process, that is, performance prediction by measuring vs. effective use of the resource. This introduces an additional complexity to the selection action, that must be integrated in the optimization process.

In order to solve the above realistic network selection problem, we propose, in this paper, a modified MAB model, and related algorithms, that incorporates two possible actions at a given decision time: *measuring* vs. *using*. Results obtained by applying the proposed method vs. using classical MAB solutions are compared; in particular, we analyse one of the most extensively-used MAB algorithms of the "Upper Confidence Bound (UCB)" family, named UCB1, since this has proved to produce the minimum regret under given boundary conditions, when the "using" action only is foreseen [3].

The paper is organized as follows: Section II makes a brief overview of MAB problems and the algorithms available in literature for its solution; Section III introduces the proposed model, while Section IV describes the new algorithms; Section V presents simulation results and contains a discussion; conclusion and future work are reported in Section VI.

II. Multi-armed bandit problems

Multi-armed bandit is a learning-theory, well-known, resource allocation problem. The classical model includes 1 player and K arms; arms provide mutually independent stochastic rewards, characterised by unknown average values.

At each step, the player selects one arm and obtains the corresponding reward realization as a feedback. Since no a priori knowledge is available to the player, the selection in the first step is random. Typically, in succeeding steps the player cycles all possible arms in order to form a reference record of reward values for all possible given choices.

After this "initialization", the decision problem consists in estimating which of the arms may contribute to produce, in a given time horizon, the maximum cumulative reward, that is defined as the cumulative reward obtained when always choosing the arm with the highest reward average value.

A problem that arises is the *exploration vs. exploitation* trade-off. *Exploration* indicates that the player chooses an arm which is not known to be the best one, i.e. the one with the highest average reward, just to improve the knowledge on its reward, while *exploitation* indicates that prior observations should be exploited to select the arm that is thought to be the best one, the one that can offer the highest average reward.

978-1-4577-1348-4/13/$31.00 ©2013 IEEE

In [1], *regret*, i.e. the difference between the above maximum cumulative reward and the cumulative reward obtained by the actually selected arms, was proposed as an evaluation parameter for measuring algorithms performance. It was also shown that the asymptotic best achievable performance is a regret that grows logarithmically with time. In [4], the problem was extended to M multiple plays, and later on a switching cost was also introduced [5]. Reference [6] proposed easy-to-compute index-type algorithms, and, in 2002, the UCB1 algorithm was introduced [3], where it was shown that the best achievable performance is a regret that grows logarithmically uniformly over time, and not only asymptotically. UCB1 was later widely used in literature [7], [8]. In the recent past, several variants of the above algorithm were proposed, such as MUCB [9] and LLR [10], to cite a few.

In the context of wireless network selection, the arms might represent the different networks, and the rewards might be, for example, the quality of communication experience they offer.

III. The new proposed model

The proposed model is introduced in the following. Time is divided into steps. There are 1 player and K arms, $\mathcal{K} = \{1, \dots, K\}$. A reward is related to each arm: $\forall\, k \in \mathcal{K}$, reward $\{W_k(n) \colon n \in \mathbb{N}\}$ is a stationary ergodic random process related to arm k; its statistics are not known a priori. Given a time step n, $W_k(n)$ is a random variable that can assume a value in the real positive numbers set $\mathbb{R}^+$; Probability Density Function (PDF) of $W_k(n)$ is not known a priori; μ_k is the mean value of $W_k(n)$, associated to arm k: $\mu_k = E(W_k(n))\ \forall\, k \in \mathcal{K}$.

There are two distinct actions: to *measure* ("*m*") and to *use* ("*u*"); at the beginning of time step n, the player can choose to apply action a to arm k: $c_n = (a_n\,,\, k_n)\,,\ a \in \{m, u\}\,,\ k \in \mathcal{K}$. Every choice c_n obtains a feedback $f(c_n)$. *Measure* and *use* actions have durations T_M and T_U respectively; $T_U = NT_M\,,\ N \in \mathbb{N}$.

Feedback $f(c_n)$ is a pair, composed by: 1) a realization of $W_k(n)$ at time step n, $w_k(n)$: it is the current reward value associated to arm k; 2) a gain $g(c_n)$; therefore: $f(c_n) = (w_k(n)\,,\, g(c_n))$.

Gain $g(c_n)$ is a function of the chosen action and of $W_k(n)$; it is always equal to zero when *measure* action is chosen and it assumes the value of the realization of $W_k(n)$ at time step n, $w_k(n)$, when arm k is *used* at time step n:

$$g(c_n) = \begin{cases} 0 \quad \forall\, k & \text{if } a_n = m \\ w_k(n) & \text{if } a_n = u \end{cases} \quad . \tag{1}$$

Performance of an algorithm can be expressed by the *regret* of not always using the arm with the highest reward mean value: $k^* = \arg\max_{k \in \mathcal{K}} \mu_k$.

Regret at time step n is defined as:

$$R(n) = G_{MAX}(n) - E(G(n)), \tag{2}$$

where $G(n) = \sum_{i=1}^{n} g(c_i)$, and $G_{MAX}(n)$ is the maximum possible cumulative mean gain at time step n, obtained by always *using* the arm k^* (and never *measuring*): $G_{MAX}(n) = E(G(n))\colon c_n = (u\,,\, k^*) \quad \forall\, n$.

The goal is to find an algorithm that minimizes regret evolution in time.

Note that *measure* action gets a feedback in T_M that is usually shorter than T_U (i.e. $N > 1$); this advantage, however, is paid through the cost of having a null gain. In other words, if at a certain time step the player chooses to measure an arm in order to have more information about its reward (and estimate the potential gain it can obtain if in a future step it chooses to use it) in a shorter time T_M, it "pays" this decision by receiving a null gain.

In this model the classical *exploration vs. exploitation* trade-off is slightly modified. *Exploration* is performed while measuring, i.e. by acquiring information about other arms being conscious that the prize for it is not the gain of a sub-optimal arm (like in the classical model), but a null gain. *Exploitation*, instead, is performed while using an arm, in order to obtain the gain it can offer.

IV. Exploitation of measure and use actions: new algorithms

Two new algorithms are proposed. The first is a modified version of UCB1. In UCB1, the selected arm is the one with highest index, that is composed by the sum of two terms: the estimated reward mean value and a bias, that is a logarithmic function of time and the number of times the arm has been selected until now [3]. The goal of the bias is to raise the index value of an arm that has not been selected since long time, and therefore to introduce exploration. In analogy with this behaviour, *modified UCB1*, introduced here, performs a *measure* action when the bias has an effect on the arm selection, i.e. when it permits to select an arm that has not the highest estimated reward mean value. In other words, when an exploration would be performed in the classical UCB1, this is converted into a *measure* action in this modified version. All the other times the action performed is always *use*.

The second proposed algorithm is specifically thought to exploit the difference between measuring and using. It is divided into two phases: in the first it performs only measures, in the second one it mostly uses the arm with the highest estimated average value, but also measures the other arms from time to time. The two phases are better described in the following.

Phase 1 (initialization): during this phase the selected action is always *measure*, and all the arms are chosen according to a round robin schedule. Every arm is measured d_1 times, and therefore this phase duration is $d_1 T_M$; d_1 is a parameter that can be decided on the fly and adjusted. The goal of this phase is to have a reliable estimate of arms reward average value. The estimates $\hat{\mu}_k$ are therefore:

$$\hat{\mu}_k = \frac{1}{d_1} \sum_{i=0}^{d_1 - 1} w_k(k + iK) \quad \forall\, k \in \mathcal{K}. \tag{3}$$

Phase 2: during this phase the player starts performing *use* actions: based on the estimates obtained thanks to the first phase, it chooses to use the arm with the highest estimated mean reward value: at time step n the choice is therefore $c_n =$

$(u\,,\,\bar{k})$, where $\bar{k} = \arg\max_{k\in\mathcal{K}} \hat{\mu}_k$. As feedback it obtains $f(c_n) = (w_{\bar{k}}(n)\,,\, g(c_n) = w_{\bar{k}}(n))$: it updates therefore the estimate $\hat{\mu}_{\bar{k}}$ and obtains the arm's reward realization at time step n as gain. At the next step where the player can make a choice, i.e. after a period T_U, the chosen arm will be the one with the highest estimated mean reward value, and so on.

However, in this phase also some measure actions are provided. The first measure is performed after d_2 uses, i.e. after a period $d_2 T_U$. After that, intervals between measures grow logarithmically with time. To be more precise, if t_i indicates the instant in which the measure should start (and, given that time is divided into steps, the measure will effectively start in the first time step n_i that begins right after t_i), $t_i = t_{i-1} + \log t_{i-1}$, with $i \geq 1$, $i_0 = d_2 T_U > 1$. At each measure action, the arm chosen for being measured is the one with the oldest estimate, i.e. the one whose estimate is the less updated.

Algorithm pseudocode
initialization: measure each arm d_1 times with round-robin schedule
loop
  if n: $t < t_{\text{next}}$ **then**
    $\bar{k} = \arg\max_k \hat{\mu}_k \quad \rightarrow \quad c_n = (u\,,\,\bar{k})$
    $f(c_n) = (w_{\bar{k}}(n)\,,\, g(c_n) = w_{\bar{k}}(n))$
    Estimate $\hat{\mu}_{\bar{k}}$ update
    $t = t + T_U$
  else
    $\bar{k} = k$ with less updated $\hat{\mu}_k \quad \rightarrow \quad c_n = (m\,,\,\bar{k})$
    $f(c_n) = (w_{\bar{k}}(n)\,,\, g(c_n) = 0)$
    Estimate $\hat{\mu}_{\bar{k}}$ update
    $t = t + T_M \quad \rightarrow \quad t_{\text{next}} = t + \log t$
  end if
end loop

The ideas behind the proposal of such an algorithm are the following: with the first phase it can collect an estimate in the shortest possible time; it should be "reliable enough" for taking next decision (i.e. which arm to use), and that depends on d_1 value and on the arms reward distributions (unknown). A null gain throughout all this phase is accepted with the idea of having "stronger" estimates for future decisions (exploration).

Based on these estimates, decisions are taken in the second phase (exploitation). Anyway the estimate of the used arm is continuously updated, and the periodic measures permit to update all the other arms estimates. The rule of logarithmic-growing intervals between subsequent measures was inspired by results in literature [1], [3]. In fact, a regret that grows logarithmically with time is the best performance it can be obtained. By inserting measure actions with logarithmic-growing intervals, regret's logarithmic growth is not perturbed, performance does not get worse for its effect.

During the first phase, where only measures are performed and the obtained gain is equal to zero, the regret in time can be expressed by a straight line with a slope $\beta_M = \frac{\mu^*}{T_M}$.

During the second phase the regret can still be expressed by a straight line with a slope β that may vary. When a measure is performed $\beta = \beta_M = \frac{\mu^*}{T_M}$. When a use is performed, the slope depends on which arm is being used. In mean, if arm k is being used, $\beta_k = \frac{\mu^* - \mu_k}{T_U}$. Considering all the arms, the slope is

$$\beta_U = \sum_{k=1}^{K} p_k \beta_k, \tag{4}$$

where p_k is the probability of using arm k, that depends on rewards distribution and the chosen algorithm.

V. Experimentation

A. Simulations

Regret obtained by classical UCB1 and the two algorithms introduced in Section IV was analysed.

Simulations were performed with different values of T_M/T_U ratio, that correspond to systems able to provide a measure in a time that is a certain percentage shorter than using period. In particular, recalling that $T_U = NT_M\,,\, N \in \mathbb{N}$, simulations were performed with $1 \leq N \leq 7$.

Other simulations details are the following: there are 5 arms: $K = 5$; reward values are binary, i.e. $W_k(n) \in \{0, 1\}$; PDF of $W_k(n)$ follows a Bernoulli distribution with success probabilities fixed to these values: $\mu_1 = 0.6\,,\, \mu_2 = 0.8\,,\, \mu_3 = 0.1\,,\, \mu_4 = 0.3\,,\, \mu_5 = 0.7$.

Moreover, the proposed algorithm was used twice, with two different values of the number of times each arm is measured in the first phase: $d_1 = 1$ and $d_1 = 5$; the initial interval coefficient value between measuring instants was set to $d_2 = 5$. All results are obtained through the mean of 500 runs and are reported in Section V-B.

B. Experimental results

In the shown examples, T_M/T_U ratio starts from value of 1 (Figure 1) and then decreases: in Figure 2 $T_M/T_U = 1/3$ and in Figure 3 $T_M/T_U = 1/6$.

Regrets obtained with both versions of UCB1 show a trend that is logarithmic with time. Modified UCB1's regret reaches much higher values compared to classical UCB1's ones when $T_M = T_U$: this was expected because the latter always *use* an arm, obtaining therefore a gain, and there is in fact no "cost" for doing this since the duration of the two actions are the same. As T_M/T_U ratio decreases, however, this "cost" becomes considerable, and therefore the gap between the performance of the two UCB1 algorithms becomes smaller. This is due to the fact that when modified UCB1 performs an exploration, i.e. measures, it "wastes" less time.

When $T_M/T_U = 1/6$, shown in Figure 3, UCB1's modified version shows a regret that is always lower than the classical version. Therefore $N = 6$ is the value that permits to have a significant performance improvement even with the same algorithm, slightly modified to better adapt to the proposed model. This means that measure becomes interesting when the system is able to provide it with a duration 6 time inferior to the use duration.

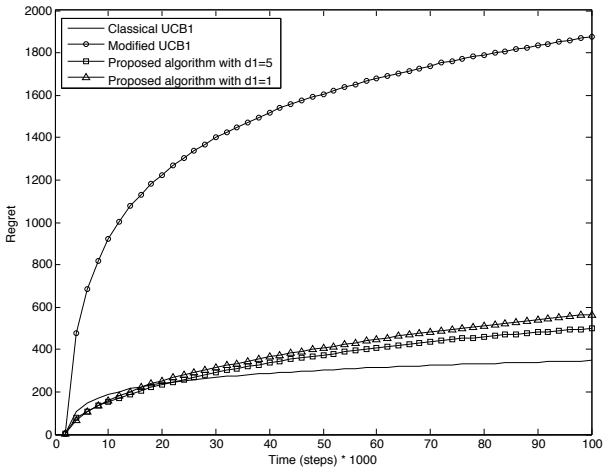

Fig. 1. Comparison among regret obtained with classical UCB1, modified UCB1 and the proposed algorithm (with $d_1 = 1$ and $d_1 = 5$) when $T_M = T_U$.

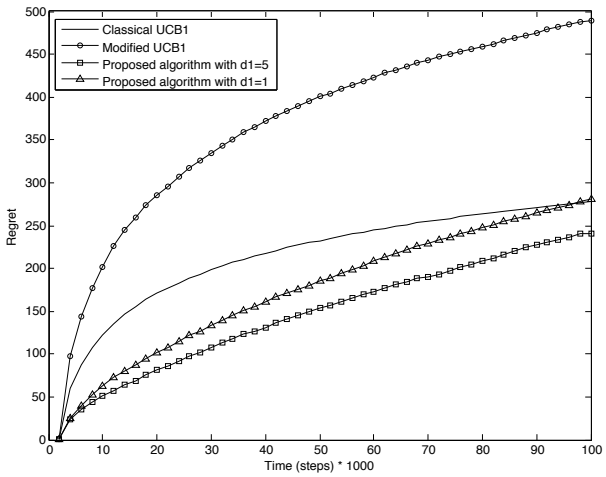

Fig. 2. Comparison among regret obtained with classical UCB1, modified UCB1 and the proposed algorithm (with $d_1 = 1$ and $d_1 = 5$) when $T_M/T_U = 1/3$.

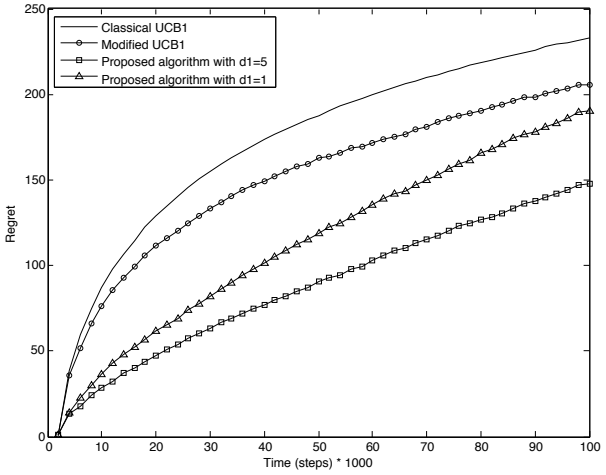

Fig. 3. Comparison among regret obtained with classical UCB1, modified UCB1 and the proposed algorithm (with $d_1 = 1$ and $d_1 = 5$) when $T_M/T_U = 1/6$.

By analysing performance in terms of regret obtained with the new proposed algorithm, even in this case as T_M/T_U decreases performance of the new algorithm increases, i.e. its regret decreases. This was expected since it means that the *measure* action presents a lower "cost" in terms of time spent for it.

When $T_M = T_U$ the new algorithm overcomes UCB1's performance until step $2 \cdot 10^4$ in both cases $d_1 = 1$ and $d_1 = 5$ (except for the very first steps, as better explained later), because the initialization phase is more efficient, but presents a higher regret in the next steps, because there is no "cost" for using an arm and obtain its gain.

As it can be seen in Figure 2, it obtains significantly better performance as $T_M \leq \frac{1}{3} T_U$, with an always-lower regret (at least until step 10^5, time horizon used in these simulations). Therefore by better exploiting the possibilities that the proposed model offers, even with a very simple algorithm, it suffices to have a ratio $T_M/T_U \leq 1/3$ to obtain significantly lower regret values.

Another consideration should be done on the very first steps. Since in the first phase the proposed algorithm performs only measures, and therefore obtains a null gain, its regret is always higher than the one obtained through an algorithm that *uses* an arm. This cannot be avoided, given the model, if not skipping the only-measures phase. The worst case, i.e. when $T_M = T_U$, is shown in Figure 4. Here it can be seen that for the first 100 steps (case $d_1 = 1$) and 350 steps (case $d_1 = 5$) new algorithm's regret is higher than UCB1's one.

Figure 5 shows the average number of time steps needed to new algorithm's regret to get a lower value compared to UCB1's regret as T_U/T_M ratio increases. As it can be seen, it becomes lower with an increasing T_U/T_M value; it means that fewer steps are necessary to "win" over UCB1 if T_M becomes smaller respect to T_U.

Time required to "win" over UCB1 depends on T_U/T_M and on the number times d_1 each arm is measured in the first phase of the algorithm.

In some practical situations it is more desirable to obtain a lower (than UCB1) regret as soon as possible, even if this will be "paid" with worse performance in the following steps. This trade-off, based on T_U/T_M ratio and the chosen value for d_1, strongly depends on scenario parameters, number of arms and rewards distribution.

VI. Conclusion and Future Work

In this work a new model for multi-armed bandit problems was proposed. Its main feature is the introduction of two distinct possible actions the player can perform: to *measure* and to *use*.

This new model was introduced in order to better reflect real practical scenarios. As already mentioned, an example of such a scenario could be a device that must choose between different wireless networks based on the performance they can

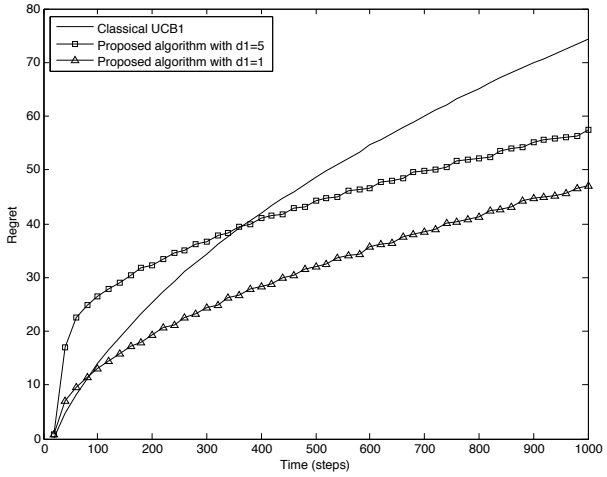

Fig. 4. Comparison among regret obtained with classical UCB1 and the proposed algorithm (with $d_1 = 1$ and $d_1 = 5$) when $T_M = T_U$, zoom on the first 1000 steps.

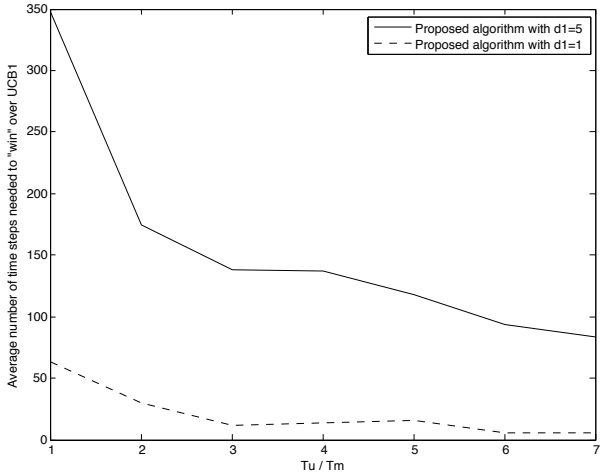

Fig. 5. Average number of time steps needed to new algorithm's regret to get a lower value compared to UCB1's regret, with an increasing T_U/T_M ratio.

offer, where the performance can be expressed as SNR, the average delay, the jitter or any other parameter of interest in the data exchange. With this model multi-armed bandit problems can potentially be closer to reality.

The impact of the introduction of such a model was analysed and discussed through simulations, in which the performance in terms of regret (a classical performance evaluation parameter often used in MAB problems) of a modified version of UCB1 algorithm and a new proposed algorithm, that exploits the introduced model novelties, is evaluated and compared to the one obtained by classical UCB1.

Results obtained from the performed simulations show that, as the ratio between T_M and T_U decreases, i.e. the measuring period duration gets smaller and smaller respect to the use period one, performance of both tested algorithms increases: regret grows slower and reaches lower values. In fact, the same measure action is performed in a smaller period, in this sense the measure is more "powerful".

It can be noted that a ratio $T_M/T_U \leq 1/6$ is needed for the modified UCB1 in order to obtain a regret that is lower than classical UCB1's one; otherwise measure is not "powerful" enough and the choice to perform such an action is more a disadvantage than an advantage. Less restrictive constraints, i.e. a ratio $T_M/T_U \leq 1/3$, are sufficient to obtain significantly better performance (always compared to classical UCB1 algorithm) with the proposed algorithm.

Moreover, the initial "loss" duration, i.e. the initial period where new algorithm's regret grows more than UCB1's one, gets lower as T_M/T_U ratio decreases. In other words, the time needed to reach UCB1's regret decreases. This is significant considering real scenarios because, given a T_M/T_U ratio, it can affect the decision of the measuring phase duration Kd_1T_M.

Future work could cope with deeper investigation on the trade-off between measure and use: how many measures vs. how many uses need to be performed in function of the scenario parameters. Moreover, other algorithms that can better exploit the new proposed model and obtain therefore better performance in terms of regret can be found and tested.

ACKNOWLEDGMENT

This work was partly supported by COST Action IC0902 "Cognitive Radio and Networking for Cooperative Coexistence of Heterogeneous Wireless Networks", funded by the European Science Foundation, and partly by European Commission Network of Excellence ACROPOLIS "Advanced coexistence technologies for radio optimisation in licensed and unlicensed spectrum".

REFERENCES

[1] T. L. Lai, and H. Robbins, *Asymptotically efficient adaptive allocation rules*, Advances in applied mathematics, No. 6, 1985.

[2] A. Mahajan, and D. Teneketzis, *Multi-armed bandit problems*, Foundations and Applications of Sensor Management, Springer US, 2008.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, *Finite-time analysis of the multiarmed bandit problem*, Machine Learning, No. 47, 2002, Kluwer Academic Publisher.

[4] V. Anantharam, P. Varaiya, and J. Walrand, *Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays – Part I: IID rewards*, IEEE Transactions on Automatic Control, Vol. 32, No. 11, 1987.

[5] R. Agrawal, M. Hegde, and D. Teneketzis, *Multi-armed bandit problems with multiple plays and switching cost*, Stochastics and Stochastic Reports, Vol. 29, 1990, Gordon and Breach Science Publishers.

[6] R. Agrawal, *Sample mean based index policies with O(log n) regret for the multi-armed bandit problem*, Advances in Applied Probability, Vol. 27, 1995.

[7] L. Lai, H. El Gamal, H. Jiang, and H. V. Poor, *Cognitive medium access: exploration, exploitation, and competition*, IEEE Transactions on Mobile Computing, Vol. 10, No. 2, 2011.

[8] D. Kalathil, N. Nayyar, and R. Jain, *Decentralized learning for multiplayer multi-armed bandits*, 51st IEEE Conference on Decision and Control, December, 10–13, 2012, Maui, Hawaii, USA.

[9] W. Jouini, *Contribution to learning and decision making under uncertainty for Cognitive Radio*, Ph.D. thesis, Supélec, 2012.

[10] Y. Gai, B. Krishnamachari, and R. Jain, *Combinatorial network optimization with unknown variables: multi-armed bandits with linear rewards and individual observations*, IEEE/ACM Transactions on Networking, Vol. 20, No. 5, 2012.

[11] J. Vermorel, and M. Mohri, *Multi-armed bandit algorithms and empirical evaluation*, Machine Learning: ECML 2005, Springer, 2005.