A New Evaluation Criteria for Learning Capability in OSA Context

Navikkumar Modi[1(✉)], Christophe Moy[1], Philippe Mary[2], and Jacques Palicot[1]

[1] CentraleSupelec/IETR, Avenue de la Boulaie, 35576 Cesson Sevigne, France
{navikkumar.modi,christophe.moy,jacques.palicot}@centralesupelec.fr
[2] INSA de Rennes, IETR, UMR CNRS 6164, 35043 Rennes, France
philippe.mary@insa-rennes.fr

Abstract. The activity pattern of different primary users (PUs) in the spectrum bands has a severe effect on the ability of the multi-armed bandit (MAB) policies to exploit spectrum opportunities. In order to apply MAB paradigm to opportunistic spectrum access (OSA), we must find out first whether the target channel set contains sufficient structure, over an appropriate time scale, to be identified by MAB policies. In this paper, we propose a criteria for analyzing suitability of MAB learning policies for OSA scenario. We propose a new criteria to evaluate the structure of random samples measured over time and referred as Optimal Arm Identification (OI) factor. OI factor refers to the difficulty associated with the identification of the optimal channel for opportunistic access. We found in particular that the ability of a secondary user to learn the activity of PUs spectrum is highly correlated to the OI factor but not really to the well known LZ complexity measure. Moreover, in case of very high OI factor, MAB policies achieve very little percentage of improvement compared to random channel selection (RCS) approach.

Keywords: Cognitive radio · Opportunistic spectrum access · Reinforcement learning · Multi-armed bandit · Lempel-Ziv (LZ) complexity · Optimal Arm Identification (OI) factor

1 Introduction

Spectrum learning and decision making is a core part of the cognitive radio (CR) to get access to the underutilized spectrum when not occupied by a licensed or primary users (PUs). In particular, we deal with multi-armed bandit (MAB) paradigm, which allows unlicensed or secondary user (SU) to make action to select a free channel to transmit when no PUs are using it, and finally learns about the *optimal channel*, i.e. channel with the highest probability of being vacant, in the long run [1–3].

The PU activity pattern, i.e. presence or absence of PU signal in the spectrum band, can be modeled as a 2-state Markov Process [3–5]. In case of SU trying to learn about the probability for a channel to be vacant, the success of

D. Noguet et al. (Eds.): CROWNCOM 2016, LNICST 172, pp. 3–14, 2016.
DOI: 10.1007/978-3-319-40352-6_1

MAB policy is affected by the amount of structure obtained in the PUs activity pattern in these channels [6]. There exist several pieces of work on opportunistic spectrum access (OSA), by means of learning and predicting an opportunity, which deals with CR to find out the PUs activity pattern. In this paper, we address the following fundamental questions, affecting MAB performance: (i) when is it advantageous to apply MAB learning framework to address the problem of opportunistic access for SU? (ii) Is the use of MAB policies for OSA is justified over a simple non-intelligent approach?

In almost all cases, the performance of MAB learning policies has been studied with respect to PUs activity level, i.e. probability that PUs occupy radio channels [2,3]. In some cases, spectrum utilization is modeled as an independent and identically distributed (i.i.d) process [1], and it does not take into account the likely sequential activity patterns of PUs in the spectrum. To address the first question raised above, the Lempel-Ziv (LZ) complexity was introduced in [6] to characterize the PU activity pattern for general reinforcement learning (RL) problem. However, MAB paradigm is a special kind of RL game where SU maximizes its long term reward by making action to learn about the optimal channel, opposed to general RL framework where SU is interacting with a system by making actions and learns about the underlying structure of the system. Moreover, in this paper, we propose the *Optimal Arm Identification (OI) factor* to identify the difficulty associated with prediction of an optimal channel having highest probability of being vacant from the set of channels. Finally, the last question raised above is answered by comparing the performance of MAB policies against the random channel selection (RCS) approach (a non-intelligent approach).

We found out that, for several spectrum utilization patterns, MAB policies can be beneficial compared to non-intelligent approach, but the percentage of improvement is highly correlated with the level of OI factor and very little affected by the level of LZ complexity. This result does not just emphasize aphorism that performance of MAB policies in OSA framework depends extremely on the OI factor associated with the selected channel set. The work presented in this paper can be the answer to the question raised in several papers about the effectiveness of MAB framework for OSA scenario. The remainder of this paper is organized as follows. In Sect. 2, we introduce MAB framework and RL policies which are used to verify the effect of PUs activity pattern on spectrum learning performance. Section 3 and 4 contain our main contributions where in Sect. 3, LZ complexity is revisited and a new criteria measuring the structure of spectrum utilization pattern, named OI, is introduced and in Sect. 4, numerical results giving the efficiency of MAB policies w.r.t. the output of LZ and OI criteria are presented. Finally, Sect. 5 concludes the paper.

2 System Model and RL Policies

We consider a network with a single[1] secondary transceiver pair (Tx-Rx) and a set of channel $\mathbb{K} = \{1, \cdots, K\}$. SU can access one of the K channels if it is not

[1] The presented analysis can also be justified for multiple SUs scenario, where each SU tries to find optimal channel following underlying activity pattern of PUs.

occupied by PUs. The i-th channel is modeled by an irreducible and aperiodic discrete time Markov chain with finite state space S^i. $P^i = \left\{p^i_{kl}, (k,l \in \{0,1\})\right\}$ denotes the state transition probability matrix of the i-th channel, where 0 and 1 are the Markov states, i.e. occupied and free respectively. Let, $\boldsymbol{\pi}^i$ be the stationary distribution of the Markov chain defined as:

$$\boldsymbol{\pi}^i = [\pi^i_0, \pi^i_1] = \left[\frac{p^i_{10}}{p^i_{10} + p^i_{01}}, \frac{p^i_{01}}{p^i_{10} + p^i_{01}}\right]. \tag{1}$$

$S^i(t)$ being the state of the channel i at time t and $r^i(t) \in \mathbb{R}$ is the reward associated to the band i. Without loss of generality we can assume, that $r^i(t) = S^i(t)$, i.e. $S^i(t) = 1$ if sensed free and $S^i(t) = 0$ if sensed occupied. The stationary mean reward μ^i of the i-th channel under stationary distribution π^i is given by: $\mu^i = \pi^i$. A channel is said optimal when it has the highest mean reward μ^{i^*}, such that $\mu^{i^*} > \mu^i$ and $i^* \neq i, i \in \{1, \cdots, K\}$, i.e. a channel with the highest probability to be vacant. The mean reward optimality gap is defined as $\Delta_i = \mu^{i^*} - \mu^i$. The regret $R(t)$ of a MAB policy up to time t, is defined as the reward loss due to selecting sub-optimal channel μ^i:

$$R(t) = t\mu^{i^*} - \sum_{m=0}^{t} r(m), \tag{2}$$

2.1 Reinforcement Learning (RL) Approaches

We consider two different reinforcement learning (RL) strategies, i.e. UCB1 and Thomson-Sampling (TS), in order to evaluate the learning efficiency of MAB policies on channel set containing different PUs activity pattern. These policies are based on RL algorithms introduced in [7–9] as an approach to solve MAB problem and they attempt to identify the most vacant channel in order to maximize their long term reward. Figure 1 illustrates a realization of the random process: 'occupancy of spectrum bands by PUs'. In this figure, all channels do not have the same occupancy ratio and it seems intuitively clear that the more different the channel occupations are, the easier the learning.

Upper Confidence Bound (UCB) Policy. It has been shown previously in [1] that UCB1 allows spectrum learning and decision making in OSA context in order to maximize the transmission opportunities. UCB1 is a RL based policy, learning about the optimal channel from previously observed rewards starting from scratch, i.e. without any *a priori* knowledge on the activity within the set of channels. For each time t, UCB1 policy updates indices named as $B_{t,i,T_i(t)}$, where $T_i(t)$ is the number of times the i-th channel has been sensed up to time t, and returns the channel index $a_t = i$ of the maximum UCB1 index. UCB1 is detailed in Algorithm 1 where α is the exploration-exploitation coefficient. If α increases, the bias $A_{t,i,T_i(t)}$ dominates and UCB1 policy explores new channels. Otherwise, if α decreases, the index computation is governed by $\bar{X}_{i,T_i(t)}$ and the policy tends to exploit the previously observed optimal channel.

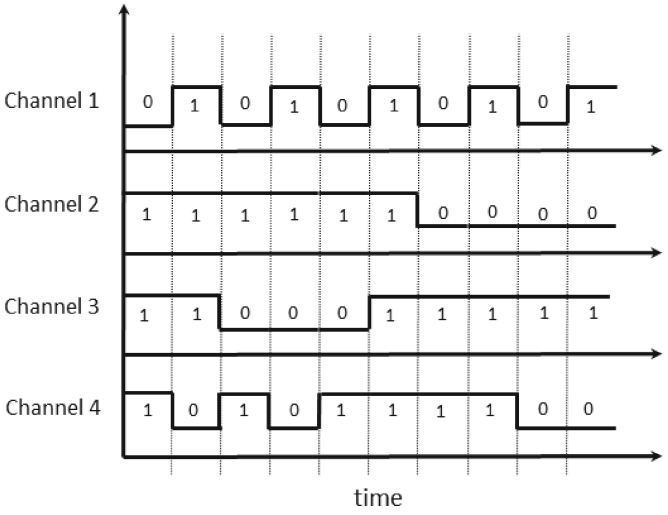

Fig. 1. PUs activity pattern

Algorithm 1. UCB1 policy

Input: K, α
Output: a_t
for $t = 1$ **to** n **do**
  if $t \leq K$ **then**
    $a_t = t + 1$
  else
    $T_i(t) = \sum_{m=0}^{t-1} \mathbb{1}_{a_m=i}, \forall i$
    $\bar{X}_{i,T_i(t)} = \frac{\sum_{m=0}^{t-1} S^i(m)\mathbb{1}_{a_m=i}}{T_i(t)}, \quad A_{t,i,T_i(t)} = \sqrt{\frac{\alpha \ln t}{T_i(t)}}, \forall i$
    $B_{t,i,T_i(t)} = \bar{X}_{i,T_i(t)} + A_{t,i,T_i(t)}, \forall i$
    $a_t = \arg\max_i (B_{t,i,T_i(t)})$
  end if
end for

Thomson-Sampling (TS) Policy. Introduced in [8,9] and detailed in Algorithm 2, TS selects a channel having the highest $J_{t,i,T_i(t)}$ index, computed with the β function w.r.t. two arguments, i.e. $G_{i,T^i(t)} = \sum_{m=0}^{t-1} S^i(m)\mathbb{1}_{a_m=i}$ and $F_{i,T^i(t)} = T^i(t) - G_{i,T^i(t)}$, where $T^i(t)$ has the same meaning than previously. The former argument is the total number of free state observed up to time t for channel i and the second is the total number of occupied state. For start, no prior knowledge on the mean reward of each channel is assumed i.e. uniform distribution and hence the index for all channels is set to $\beta(1, 1)$. TS policy updates the distribution on mean reward μ^i as $\beta\left(G_{i,T^i(t)} + 1, F_{i,T^i(t)} + 1\right)$.

Algorithm 2. Thomson-Sampling (TS) policy

Input: K, $G_{i,1} = 0$, $F_{i,1} = 0$
Output: a_t
1: **for** $t = 1$ **to** n **do**
2: $J_{t,i,T_i(t)} = \beta(G_{i,T^i(t)} + 1, F_{i,T^i(t)} + 1)$
3: Sense channel $a_t = \arg\max_i \left(J_{t,i,T_i(t)}\right)$
4: Observe state $S^i(t)$
5: $T_i(t) = \sum_{m=0}^{t-1} \mathbb{1}_{a_m=i}, \forall i, \qquad G_{i,T_i(t)} = \sum_{m=0}^{t-1} S^i(m)\mathbb{1}_{a_m=i}, \forall i$
6: $F_{i,T_i(t)} = T_i(t) - G_{i,T_i(t)}, \forall i$
7: **end for**

3 PUs Activity Pattern vs Difficulties of Prediction

In general, multi-armed bandit (MAB) algorithms are evaluated with the PUs traffic load which characterizes the occupancy of the spectrum band. Intuitively, the higher occupancy of the channel by PUs, the more difficult the opportunistic access for SU will be. However, traffic load of the PUs is not sufficient for evaluating the efficiency of MAB policies. In fact, performance of MAB policies leverages on the structure of the PUs activity pattern and also on the difficulties associated with identification of the optimal channel, i.e. channel with optimal mean reward distribution μ^{i^*}. The ON/OFF PUs activity model approximates the spectrum usage pattern as depicted in Fig. 1. Moreover, if the separation between the mean reward distribution of the optimal and a sub-optimal channel is large, SU should be able to converge to the optimal channel faster, and thus achieves a higher number of opportunistic accesses. Therefore, estimating the amount of structure present in the PUs activity pattern is of essential interest for applying machine learning strategies to OSA.

3.1 Lempel-Ziv (LZ) Complexity

Lempel-Ziv (LZ) complexity was proposed in [11] as a measure for characterizing randomness of sequences. It has been widely adopted in several research areas such as biomedical signal analysis, data compression and pattern recognition. Lempel and Ziv, in [11], have associated to every sequence a complexity c which is estimated by looking at the sequence and incrementing c every time a new substring of consecutive symbols is available. Then c is normalized via the asymptotic limit $n/\log_2(n)$, where n is the length of the sequence. LZ complexity is a property of individual sequences and it can be estimated regardless of any assumptions about the underlying process that generated the data. In [6], the authors have applied the LZ definition to the production rate of new patterns in Markovian processes. This is of particular interest when PUs activity is modeled as Markov process to evaluate the efficiency of MAB policies. For an ergodic source, LZ complexity equals the entropy rate of the source, which for a Markov chain S is given by [6]:

$$h(S) = -\sum_{k,l} \pi_k p_{k,l} \log p_{k,l}, \qquad k, l \in \{0, 1\}, \tag{3}$$

where $p_{k,l}$ is the transition probability between state k and l. System with LZ complexity equal to 1 implies very high rate of new patterns production and thus it could make difficult for the learning policy to predict the next sequence. For example in Fig. 1, channels 1 to 4 have different PUs activity pattern characterized by normalized LZ complexity of 0.05, 0.30, 0.60 and 0.66, respectively. It is clear that prediction of next vacancy is an easy task in case of channel 1 which has lower LZ complexity, whereas it becomes more and more difficult to predict next vacancy in channel 4 which has higher LZ complexity.

3.2 Optimal Arm Identification (OI) Factor

As stated before, performance of MAB policy applied to OSA context also leverages on the separation between optimal and sub-optimal channels mean reward distribution. Here, we define another criteria to characterize the difficulty for a MAB to learn the PUs spectrum occupancy. The MAB policy learns, based on past observations, which channel is optimal in term of mean reward distribution in the long run. The optimal arm identification for MAB framework has been studied since the 1950s under the name 'ranking and identification problems' [12,13].

In recent advances in MAB context, an important focus was set on a different perspective, in which each observation is considered as a reward: the user tries to maximize his cumulative reward. Equivalently, its goal is to minimize the expected regret $R(t)$, as defined in (2), up to time $t > 1$. As stated in [7,14], regret $R(t)$, defined as the reward loss due to the selection of sub-optimal channels, up to time t is upper bounded uniformly by a logarithmic function:

$$R(t) \leq a \sum_{i:\mu^i < \mu^{i^*}} \frac{\ln t}{(\mu^{i^*} - \mu^i)} + b \sum_{i:\mu^i < \mu^{i^*}} (\mu^{i^*} - \mu^i), \tag{4}$$

where a and b are constants independent from channel parameters and time t. As stated in (4), upper bound on regret of MAB policy is scaled by the change in mean reward optimality gap $\Delta_i = (\mu^{i^*} - \mu^i)$. Intuitively, decreasing Δ_i makes the upper bound looser and thus increases the uncertainty on MAB policies performance. In this paper, we propose the OI factor H_1 as a measure of difficulty associated with finding an optimal channel among several other channels:

$$H_1 = 1 - \sum_{i=1}^{K} \frac{(\mu^{i^*} - \mu^i)}{K}, \tag{5}$$

where, μ^i and μ^{i^*} are the mean reward distribution of sub-optimal and optimal channels respectively, K is the number of channels and i^* is the index of optimal channel. H_1 measures how close the mean reward of all sub-optimal channels are from the mean reward of the optimal channel. If H_1 is close to 1 then all channels have very closely distributed mean reward, thus it becomes almost impossible for learning policy to identify the optimal channel from the set of channels.

4 Impact of LZ Complexity and OI Factor on Prediction Accuracy

In this section, two MAB policies, i.e. UCB1 and Thomson-Sampling (TS), are investigated and the performance they achieve are put in correlation with the information given by LZ complexity and OI factor H_1. Markov chains with several levels of stationary distribution $\boldsymbol{\pi} = [\pi_0, \pi_1]$, LZ complexity and H_1 factor are generated for further numerical analysis. For simulation convenience, some parameters need to be set. Indeed, (1) is an undetermined system with two unknowns p_{01} and p_{10}. Therefore, as a side step, we considered 9 different levels of π_1, i.e. probability of being vacant, as $0.1, 0.2, \cdots, 0.9$. For these values of π_1, we obtained 45 different transition probability matrices P, each corresponding to different LZ complexity. A total of $\binom{45}{5}$ combinations are obtained by considering 45 different transition probability matrices and $K = 5$ channels, and those correspond to various H_1 factor. Finally, MAB policies are applied to the randomly selected 2000 combinations from a total of $\binom{45}{5}$ combinations. Every point in each figure corresponds to one realization of MAB policies. For each realization, policy is executed over 10^2 iterations of 10^4 time slots each. Moreover, the exploitation-exploration coefficient in UCB1 is set to $\alpha = 0.5$ which is proved to be efficient for maintaining a good tradeoff between exploration and exploitation [10].

4.1 Probability of Success

Probability of success P_{Succ} is computed by considering the number of times vacant channel is explored over the number of iterations. The success probability depends on the probability P_f that there exists at least one free channel from the set of channels K. Considering that the channel occupation is independent from one channel to another, we have [6]:

$$P_f = 1 - \prod_{i=1}^{K} \pi_0^i, \tag{6}$$

where π_0^i is the probability that the i-th channel is occupied.

Figure 2(a) and (b) depict the probability of success P_{Succ} of UCB1 and TS policies, i.e. the probability that these policies access to a free channel, according to the probability that at least one channel is free, i.e. P_f and LZ complexity. In both figures, success probability increases with P_f for a given level of LZ complexity. However, in Fig. 2(a) and (b), for a given P_f, several values of LZ complexity lead to the same level of performance for UCB1 and TS algorithms. This reveals that LZ complexity is not really related to the ability of UCB1 and TS policies to learn the scenario. For instance in Fig. 2(a) and (b), SU is able to achieve more than 90 % of probability of success on a channel set with LZ complexity of 0.2 and $P_f = 0.98$, whereas it only achieve 75 % of probability of success on a channel set with LZ complexity of 0.6 and $P_f = 0.98$. In that case,

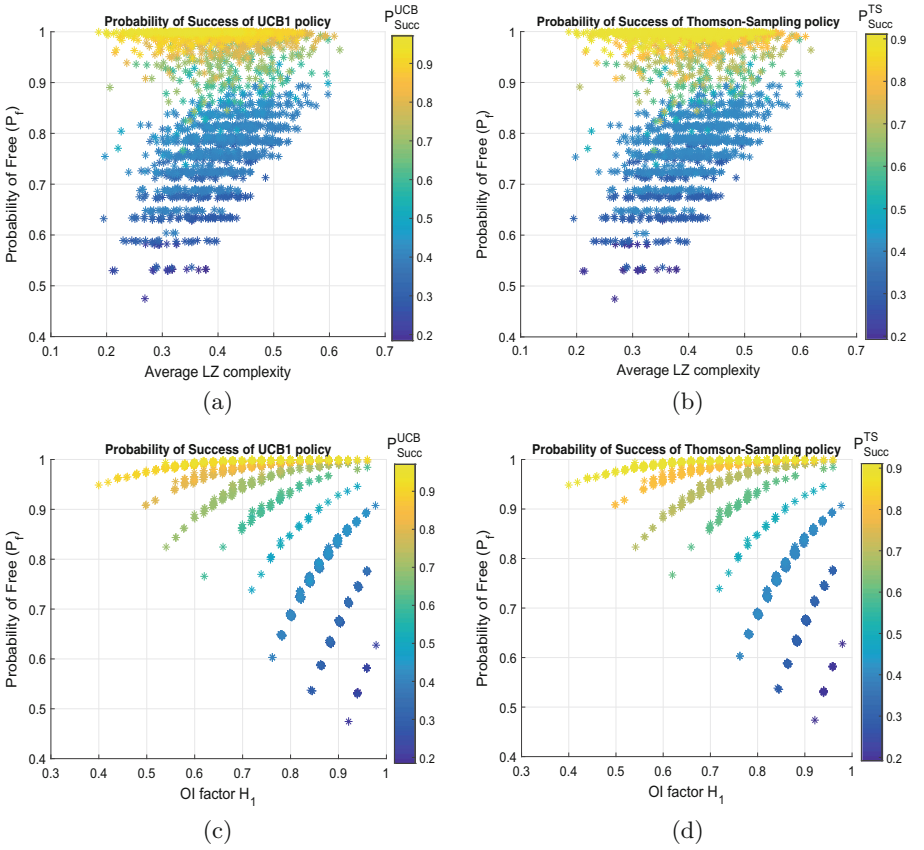

Fig. 2. (a), (b) Probability of success P_{Succ} of MAB policies, i.e. UCB1 and TS, with respect to the average LZ complexity and the probability of free P_f. Each point denotes a particular realization of MAB policies applied to $K = 5$ channels. The number of random combinations which we analyzed is 2000. (c), (d) Probability of success P_{Succ} of MAB policies, i.e. UCB1 and TS, with respect to the OI factor H_1 and the probability of free P_f applied to same set of channels.

the variation in the probability of success is up to 15 % however the variation of P_{Succ} along the x-axis can be even less important for lower values of P_f.

On the other hand, Fig. 2(c) and (d) show the probability of success of UCB1 and TS policies according to H_1 and the probability of free P_f. As we can see that H_1 is highly correlated to UCB1 and TS policies performance on a given scenario. In order to achieve very high level of P_{Succ}, H_1 required to be low. For instance in Fig. 2(c) and (d), SU is able to achieve more than 90 % of P_{Succ} on a channel set when H_1 is 0.4 and $P_f = 0.95$, whereas it only achieves 50 % of P_{Succ} on a channel set when $H_1 = 0.95$ and $P_f = 0.95$. Thus, we can state that P^{UCB}_{Succ} varies up to 40 % according to the changes in H_1, along x-axis, for certain values of P_f.

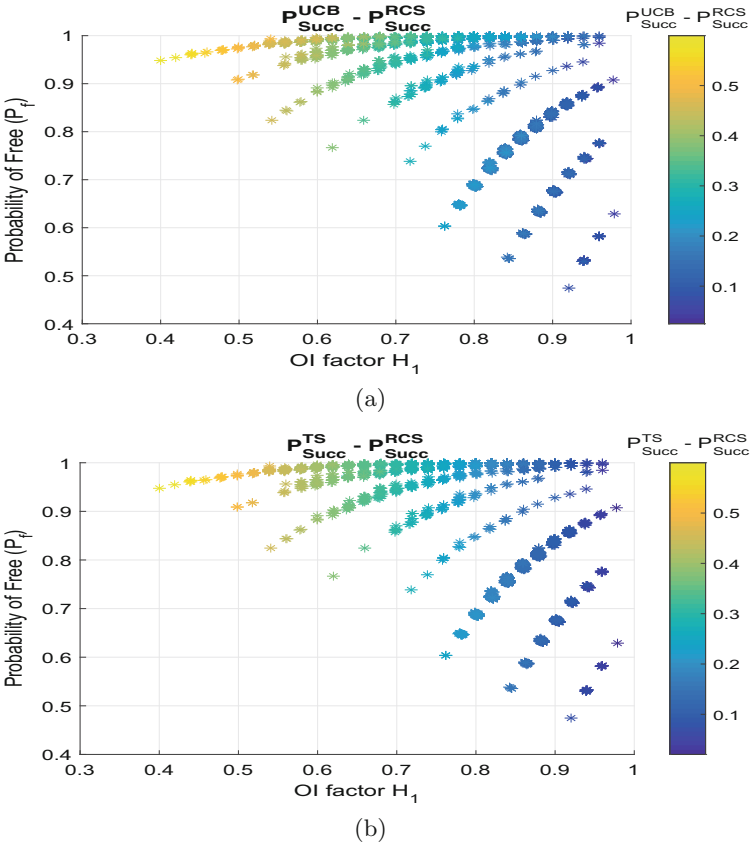

Fig. 3. (a), (b) Each point denotes the difference between the probability of success of MAB policies, i.e. UCB1 and TS, and the probability of success of the random channel selection (RCS) approach applied to $K = 5$ channels, respectively. 2000 random combinations have been analyzed.

4.2 Comparison with Random Channel Selection Policy

Figure 3(a) and (b) show difference between probability of success of MAB policies, i.e. UCB1 and TS, and random channel selection (RCS) approach. As expected, MAB policies outperform RCS approach in general, but difference becomes negligible for very high H_1 regime, i.e. the mean rewards of sub-optimal and optimal channels become equivalent. For a given P_f, performance of MAB policies decreases when H_1 increases. For instance in Fig. 3(a) and (b), we can notice that $P_{Succ}^{UCB} - P_{Succ}^{RCS}$ and $P_{Succ}^{TS} - P_{Succ}^{RCS}$ vary up to 50 % along the x-axis, i.e. H_1.

Figure 4 shows the average percentage of improvement in the probability of success achieved by MAB policies, i.e. UCB1 and TS, with respect to RCS approach under various PUs activity pattern. As we stated before, percentage

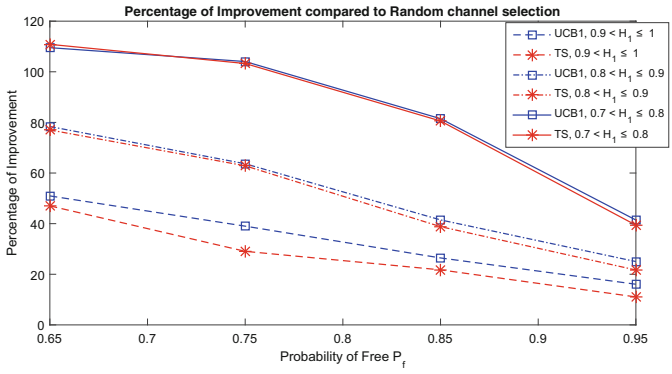

Fig. 4. Average percentage of improvement in the probability of success of MAB policies, i.e. UCB1 and TS, with respect to RCS policy as a function of the OI factor H_1 and the probability of free P_f. Each point denotes an average percentage of improvement achieved by MAB policies applied to several combinations of H_1 and P_f.

of improvement of MAB policies compared to RCS approach decreases when P_f increases, because RCS approach is able to find more opportunities in high P_f regime. On the contrary, average percentage of improvement of MAB policies also decreases when P_f decreases after certain limit. It is due to the fact that there are not many opportunities available to exploit for MAB policies in low P_f regime. As stated in Fig. 4, combinations with low H_1, i.e. $0.7 < H_1 \leq 0.8$, increases the percentage of improvement of MAB policies compared to RCS approach. Even for high H_1, i.e. $0.9 < H_1 \leq 1$, the relative improvement of learning policies is still noticeable, i.e. more than 15 %. It also reveals that all MAB policies, i.e. UCB1 and TS, achieve nearly same level of percentage of improvement for low H_1, i.e. $0.7 < H_1 \leq 0.8$, whereas in case of high H_1, i.e. $0.9 < H_1 \leq 1$, UCB1 policy significantly outperforms TS policy. Figures 3 and 4 prove that OI factor H_1 is rather well suitable compared to LZ complexity to analyze learning capability of MAB policies in OSA context.

5 Conclusions

While MAB policies, e.g. UCB1 and TS, are often assumed to be beneficial for OSA context, the problem of characterizing the scenarios where they are effective is barely studied. In this paper, we propose a new criteria, named OI factor, to characterize the situations where MAB policies will be good a priori. We evaluate the performance of UCB1 and TS on various scenarios, and correlate this to the output of OI factor and LZ complexity. Our findings show that LZ complexity does not give sufficient insights on how MAB policies behave on learning scenarios. On the other hand, OI factor is well connected to the percentage of success of MAB policies. Hence, we suggest to use OI factor in order to know if learning compared to random channel selection is beneficial for a given scenario

or not. Moreover, MAB learning can achieve more than 50 % of improvement in the probability of success compared to the non-intelligent approach in scenarios presenting low OI factors.

Acknowledgments. This work has received a French government support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference No. ANR-10-LABX-07-01. The authors would also like to thank the Region Bretagne, France, for its support of this work. Authors would like to thank Hamed Ahmadi, from University College of Dublin, Ireland and CONNECT research center, for introducing Lempel-Ziv complexity to us. Authors would also like to thank Sumit J. Darak, from IIIT-Delhi, India, for introducing Thomson-Sampling policy to us.

References

1. Jouini, W., Ernst, D., Moy, C., Palicot, J.: Upper confidence bound based decision making strategies and dynamic spectrum access. In: proceedings of IEEE International Conference on Communications (ICC), pp. 1–5 (2010)
2. Liu, H., Liu, K., Zhao, Q.: Learning in a changing world: restless multiarmed bandit with unknown dynamics. IEEE Trans. Inf. Theor. **59**(3), 1902–1916 (2013)
3. Modi, N., Mary, P., Moy, C.: QoS driven channel selection algorithm for opportunistic spectrum access. In: Proceedings of IEEE GC 2015 Workshop on Advances in Software Defined Radio Access Networks and Context-aware Cognitive Networks, San Diego, USA (2015)
4. Rehmani, M.H., Viana, A.C., Khalife, H., Fdida, S.: Activity pattern impact of primary radio nodes on channel selection strategies. In: Proceedings of the 4th International Conference on Cognitive Radio and Advanced Spectrum Management (CogART 2011), Barcelona, Spain (2011)
5. Yuan, G., Grammenos, R.C., Yang, Y., Wang, W.: Performance analysis of selective opportunistic spectrum access with traffic prediction. IEEE Trans. Veh. Technol. **59**(4), 1949–1959 (2010)
6. Macaluso, I., Finn, D., Ozgul, B., DaSilva, L.A.: Complexity of spectrum activity and benefits of reinforcement learning for dynamic channel selection. IEEE J. Sel. Areas Commun. **31**(11), 2237–2248 (2013)
7. Auer, P., Cesa-Bianchi, N., Paul, F.: Finite-time analysis of the multiarmed bandit problem. J. Mach. Learn. **47**(2–3), 235–256 (2002)
8. Russo, D., Van Roy, B.: An information-theoretic analysis of Thompson sampling. Computing Research Repositor (2014)
9. Agrawal, S., Goyal, N.: Analysis of Thompson sampling for the multi-armed bandit problem. In: Proceedings of the 25th Annual Conference on Learning Theory (COLT) (2012)
10. Melián-Gutiérrez, L., Modi, N., Moy, C., Pérez-lvarez, I., Bader, F., Zazo, S.: Upper confidence bound learning approach for real HF measurements. In: proceedings of IEEE ICC 2015-Workshop on Advances in Software Defined and Context Aware Cognitive Networks, London, UK, pp. 387–392 (2015)
11. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Trans. Inf. Theor. **22**(1), 75–81 (1976)

12. Audibert, J., Bubeck, S., Munos, R.: Best arm identification in multi-armed bandits. In: Proceedings of the 23th Annual Conference on Learning Theory (COLT), Haifa, Israel, pp. 41–53 (2010)
13. Kalyanakrishnan, S., Tewari, A., Auer, P., Stone, P.: PAC subset selection in stochastic multi-armed bandits. In: Proceedings of the 29th International Conference on Machine Learning ICML 2012, Edinburgh, Scotland, UK (2012)
14. Tekin, C., Liu, M.: Online algorithms for the multi-armed bandit problem with Markovian rewards. In: Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton) (2010)