Learning spectrum opportunities in non-stationary radio environments

Jan Oksanen, Visa Koivunen
Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University, Finland
jan.oksanen@aalto.fi, visa.koivunen@aalto.fi

***Abstract*—Learning-based sensing policies for multi-band flexible spectrum use, in particular cognitive radios operating in non-stationary radio environments are proposed. The proposed policies stem from the stochastic non-stationary restless multi-armed bandit formulation of opportunistic spectrum access. The non-stationary radio environment assumed in this paper is an appropriate model for a realistic cognitive radio systems, where the obtainable data rates depend on many unknown time-varying factors. These are e.g. mobility, fading and primary user activity. The developed policies are index policies, where the index of a frequency band depends on the discounted average reward of the band and a recency-based exploration bonus. The exploration bonus encourages sensing frequency bands that have not been explored for a long time. However, there is a maximum number of time instances when any band can remain unexplored. These index policies are computationally simple making them attractive for mobile cognitive radios. In our simulation examples, we demonstrate that the proposed policies can often provide higher cumulative data rate than other existing state-of-the-art policies.**

***Index Terms*—Flexible spectrum use, cognitive radio, opportunistic spectrum access, multi-armed bandit, dynamic propagation environment**

I. Introduction

Flexible spectrum use and cognitive radios (CRs) address the important problem of radio spectrum scarcity. The apparent lack of usable spectrum is in fact caused by rigid allocation and regulation policies instead of spectrum being fully in use. Radio spectrum is a time-frequency-location varying resource and the current regulation policies do not exploit that property. In CR, secondary users (SUs) are allowed to access licensed spectrum provided that the interference caused to primary users (PUs) is negligible. In order to identify such spectrum opportunities, the SUs need to sense the spectrum. The spectrum of interest may consist of multiple non-contiguous frequency bands, which the SUs may need to sense one at a time. Selecting the frequency band to be sensed is the task of a spectrum *sensing policy*, that is the scope of this paper. The SUs would like the sensing policy to select bands that are expected to consistently produce high data rate. However, the expected data rate from a given band depends on a number of unknown factors associated with channel quality including fading, path loss and interference as well as PUs activity.

In this paper, we develop reinforcement learning based methods for online learning of the optimal frequency bands for secondary use based on the past spectrum sensing results by an SU. A particularly suitable formulation for this problem is the restless multi armed bandit (RMAB) problem [1], [2], [3], [4]. In the RMAB formulation frequency bands are seen as slot machines that produce random rewards when played. For the CR these rewards come in the form of data rate or throughput. The expected rewards from the machines are unknown to a player whose task is to maximize the overall cumulative reward. Commonly the reward statistics are assumed to be stationary [1], [2], [3], [4], which may not be a valid assumption in practice. Motivated by more realistic CR scenarios where the statistics of the data rates may be volatile due to mobility, fading, and hourly/daily/weekly variations in the PU networks traffic load, the rewards in this paper are non-stationary. Using the non-stationary bandit problem formulation we propose two recency based index policies, that are based on the idea of promoting sensing such frequency bands that have not been sensed for a long time. The policies in this paper extend our earlier work in [4] where stationary rewards were assumed. The non-stationary model considered in this paper is more practical than the one in [4], since in reality the state of the radio spectrum and wireless channel conditions are non-stationary. The main difference in the policies proposed in this paper and the ones in [4] is the use of a discounted average reward estimate and a recency-based exploration bonus that allows regulating the maximum time for any frequency band to remain unexplored.

The structure of this paper is as follows. Section II describes the system model and the mathematical problem formulation. Section III gives a short literature review of non-stationary bandit problems. The proposed recency-based sensing policies for non-stationary environments are presented in Section IV. Simulation results of the proposed policy and other state-of-the art policies from the literature are presented in Section V. Concluding remarks and discussion on future research directions are given in Section VI.

II. The System Model

We assume a multi-band CR where one or multiple SUs are sensing K frequency bands in the hope of identifying spectrum opportunities. If at time n an SU senses band k and decides to access it, the SU obtains a random bounded data rate with an unknown expected value $\mu_k(n)$. The obtained average data rate from a given frequency band depends on bandwidth, channel quality as well as the spectrum usage patterns of the PU, for example. The data rates are non-stationary, i.e., the $\mu_k(n)$'s may change over time. We call a bandit problem

piece-wise stationary, if the expected data rates may change only at abrupt and unknown change points.

The performance of a policy π solving a non-stationary bandit problem can be measured by its regret. Regret is defined as the expected difference between the total achieved reward by a learning policy and the total reward achieved by the optimal genie aided policy that always chooses the optimal band. In this sense, regret can be seen as a measure for the cost of learning. By denoting the expected rate of band k at time n as $\mu_k(n)$, the regret of a policy π over time horizon T may be expressed as [5]

$$R^{\pi}(T) = \mathbb{E}\left[\sum_{n=1}^{T} \sum_{k:\mu_k(n)<\mu_{k_n^*}(n)} \left(\mu_{k_n^*}(n) - \mu_k(n)\right) \mathbb{1}_{\{k_n^{\pi}=k\}}\right],$$

where k_n^* denotes the index of the band with the highest average reward at time instant n, indicator function $\mathbb{1}_u$ obtains value 1 if u is true and otherwise 0, and k_n^{π} denotes the index of the band chosen by policy π at time instant n.

III. Related work

In the following we briefly describe other state-of-the-art policies used for comparison in the simulation examples. In [5] the piecewise stationary bandit problem, with known number of abrupt changes in the reward statistics during a fixed time horizon T was analysed. It was shown that for any policy and a known the number of change points, the achievable average regret is at least of order $O(\sqrt{T})$. Since typically the movements of wireless terminals are not known beforehand, assuming the number of change points to be known a priori is often not well justified.

a) Upper confidence bound based exploration: Policies for non-stationary bandit problens using upper confidence bounds (UCBs) have been proposed in [6], [5], [7]. In [6] a discounted UCB (D-UCB) policy was proposed. In [5] it was shown that the D-UCB achieves sublinear regret when the number of change points is known. In [5] a sliding window UCB (SW-UCB) policy was proposed and shown to achieve sublinear regret with known number of change points.

b) Randomized exploration: Perhaps the simplest policy suitable for non-stationary bandit problems is the ϵ-greedy policy [8]. The ϵ-greedy policy selects a (uniformly) random band with probability ϵ, while with probability $1-\epsilon$ the policy selects the band with the highest empirical discounted mean reward. Since for any band the minimum probability of being sensed is fixed, the regret of the ϵ-greedy policy is linear.

In adversarial bandit problems [9] no statistical assumptions are made about the rewards. This makes some of the adversarial policies also applicable to non-stationary settings. In [10] it was shown, that the randomised EXP3.S policy [9] achieves sublinear regret in non-stationary bandit problems that is of the same order as the regret of D-UCB and SW-UCB. However, sublinear regret is achieved with optimal parameter tuning, that requires the number of change points to be known in advance.

c) Meta-bandits and change point detection: In [11], [12], [13] and [14] meta-bandit policies were proposed for non-stationary bandit problems. The core idea in meta-bandit policies is to use stationary bandit policies together with change point detection schemes that identify time instances when the average rewards have changed.

IV. Proposed Recency-Based Exploration Policies

In this paper we propose two recency-based index policies for non-stationary scenarios. These policies extend our earlier work [4] to cases where the state of the radio spectrum and consequently the rewards are non-stationary. In order to understand the policies proposed in this paper, we need to describe the policy for stationary scenarios first. The recency-based exploration (RBE) policy in [4] senses first each band once, after which the band with the highest index is always selected. The index of channel k at time n is defined as

$$I(n) = \bar{x}_k(n) + g(n/\tau_k(n)), \tag{1}$$

where $\bar{x}_k(n)$ denotes the average observed reward from band k, $\tau_k(n)$ the time instant when band k was sensed last time and $g(n/\tau_k(n))$ the exploration bonus. The exploration bonus $g(n/\tau_k(n))$ is an increasing, concave and unbounded function when $\tau_k(n) < n$, and $g(n/\tau_k(n)) = 0$ when $\tau_k(n) = n$. An example of such function would be

$$g(n/\tau_k(n)) = \sqrt{\log(n/\tau_k(n))}.$$

The idea of the exploration bonus is to promote sensing bands that have not been sensed for a long time, however, such that asymptotically the suboptimal bands will be sensed at an exponentially decreases rate. As the time interval between two consecutive sensings of a suboptimal band grows exponentially, the regret of the policy grows logarithmically.

In this paper, we extend the RBE policy of [4] for scenarios where the rewards are non-stationary. We call the first proposed policy in this paper the non-stationary RBE (NRBE). In a non-stationary setting, there are two technical problems that need to be addressed. Firstly, the reward statistics are time-varying and need to be tracked. This is achieved by replacing the sample averages employed to estimate the expectations by discounted averages. In particular, we employ the discounted average reward $\bar{x}_k^{\gamma}(n)$ computed as

$$\bar{x}_k^{\gamma}(n) = \frac{\sum_{t=1}^{n} \gamma^{n-t} x_k(t) \mathbb{1}_{\{k_t^{\pi}=k\}}}{\sum_{t=1}^{n} \gamma^{n-t} \mathbb{1}_{\{k_t^{\pi}=k\}}}, \tag{2}$$

where $0 < \gamma < 1$ denotes the discounting factor, $x_k(t)$ is the observed reward from band k at time instant t and k_n^{π} denotes the index of the band chosen by policy π at time instant n. The indicator function $\mathbb{1}_u$ obtains value 1 if u is true and otherwise 0. Typically $0.9 \leq \gamma < 1$. The discounting factor γ decreases the weight of the past observations at an exponential rate, so that changes in the mean rewards can be tracked.

Secondly, for the non-stationary problem the exploration bonus needs to be modified such that it guarantees exploration every once in a while. To this end, we notice that the

Initialization:
- Sense each band once.

Loop: $n = K+1, K+2, \ldots$
- Sense band k that maximizes

$$I_k(n) = \bar{x}_k^{\gamma}(n) + g\left(\frac{\tilde{n}}{\tilde{n} - \delta_k(n)}\right).$$

Fig. 1. The proposed NRBE policy. Variable $\bar{x}_k^{\gamma}(n)$ is the discounted average reward up to time n at band k and $g(\cdot)$ is the exploration bonus.

Initialization:
- Sense each band once.

Loop: $n = K+1, K+2, \ldots$
- Sense band k that maximizes

$$I_k(n) = \bar{x}_k^{W}(n) + g\left(\frac{\min(n,W)}{\min(n,W) - \delta_k(n)}\right).$$

Fig. 2. The proposed WRBE policy. Variable $\bar{x}_k^{W}(n)$ is the sample average of the last W rewards at time n at band k and $g(\cdot)$ the exploration bonus.

exploration bonus in (1) can be expressed as

$$g\left(\frac{n}{\tau_k(n)}\right) = g\left(\frac{n}{n - \delta_k(n)}\right),$$

where $\delta_k(n)$ is the number of time instances passed since band k was sensed last time. In stationary bandit problems the term n in the exploration bonus had the desired effect of asymptotically increasing the time interval between two consecutive sensings of any suboptimal band. However, since in the non-stationary problems currently suboptimal band may later become the optimal band, such asymptotic convergence property is no longer needed. The method proposed in this paper replaces the time index n in the exploration bonus by a constant value $\tilde{n} > 1$ that represents the maximum allowed time that any band can remain unexplored. The resulting NRBE policy is shown in Fig. 1. Since the $g(y)$ is an unbounded increasing function in y, it can be noticed that when $\delta_k(n) = \tilde{n}$,

$$g\left(\frac{\tilde{n}}{\tilde{n} - \delta_k(n)}\right) = \infty$$

and band k will be selected for sensing. Consequently, the smaller the value of $\tilde{n}$, the more eager to explore the other bands the policy will be. When possible, the value of $\tilde{n}$ should be set as close as possible to the expected time interval that the optimal arm remains constant.

We also propose a windowed RBE (WRBE) policy shown in Fig. 2. The WRBE policy operates like the stationary RBE policy in [4], except that a time window of W most recent rewards are considered. The index of band k is defined as

$$I_k(n) = \bar{x}_k^{W}(n) + g\left(\frac{\min(n,W)}{\min(n,W) - \delta_k(n)}\right),$$

where $\bar{x}_k^{W}(n)$ denotes the sample average of the W most recent rewards. In exploration bonus time variable n has been replaced by $\min(n,W)$, since in the WRBE time counter is considered to start from the beginning of the observation window. Hence, the "current time instance" in the WRBE is n if $n < W$ and W if $n \geq W$. Similarly as the value of $\tilde{n}$ in NRBE-policy, the value of W should be chosen to match the expected duration of the best arm remaining unchanged.

V. Simulation examples

In the simulations we consider three scenarios: a piece-wise stationary reward scenario, a scenario with continuous reward mean fluctuations and for the sake of comparison a stationary scenario to learn how the proposed polices converge. In the first scenario there are 3 frequency bands where the mean reward of one of the bands changes abruptly 4 times. The mean rewards of the bands are $\mu_1(n) = 0.5$, $\mu_2(n) = 0.4$ and

$$\mu_3(n) = \begin{cases} 0.9, & \text{if } 3000 \leq n < 5000 \text{ or } 10000 \leq n < 12000 \\ 0.3, & \text{otherwise.} \end{cases}$$

Such abrupt changes could take place when the CR moves from one radio environment to another, such as from an office room to a corridor, around a street corner or during hand-off to another base station. Similar abrupt changes could occur when the PU or a source of interference changes it's transmit power of spectrum usage rate.

In the second scenario there are 10 frequency bands with expected data rates continuously varying according to a random walk. In the second scenario the mean reward evolution is such that initially all bands are identical $\mu_k(1) = 0.5$, $\forall k$, and at time n ($n \geq 2$) the mean rewards change randomly as

$$\mu_k(n) = B\big(\mu_k(n-1) + 0.02 \cdot u(n)\big),$$

where $u(n) \sim \mathrm{U}\left[-\frac{1}{2}, \frac{1}{2}\right]$ is a uniformly distributed random variable, and $B(y) = y$ for $y \in [0,1]$, $B(y) = 0$ for $y < 0$ and $B(y) = 1$ for $y > 1$. The second scenario resembles a situation where the channel condition between the CR transmitter and receiver gradually changes for example due to distance dependent path loss.

In the third scenario we compare the policies' in a stationary situation. This is done in order to see the performance loss caused by the non-stationarity assumption when the environment is actually stationary as well as the convergence of the non-stationary policies. In this scenario there are three bands with mean rewards $\mu_1(n) = 0.5$, $\mu_2(n) = 0.7$ and $\mu_3(n) = 0.9$.

In all scenarios the achieved rewards are assumed to be Bernoulli distributed with success probability $\mu_k(n)$. The simulations are run for $2 \cdot 10^4$ time instances and the resulting regret curves are averages of 1000 independent Monte Carlo realizations. In the second scenario the random fluctuations of the mean rewards are different for each Monte Carlo run.

We compare the proposed NRBE and WRBE policies against three other non-stationary policies from the literature: SW-UCB [5], D-UCB [5] and EXP3.S [9]. For comparison we also include simulations of the stationary UCB policy [15] using the confidence bound $\sqrt{\ln(n)/(2m_k(n))}$, where $m_k(n)$ denotes the number of times band k has been sensed until time n. The parameter values for D-UCB (using the notation in [5]) are $\gamma = 0.9982$ (discount factor) and $\xi = 0.15$. For SW-UCB the parameters are $\tau = 1780$ (sliding window size) and $\xi = 0.6$. Using the notation in [9] the parameters for the EXP3.S policy

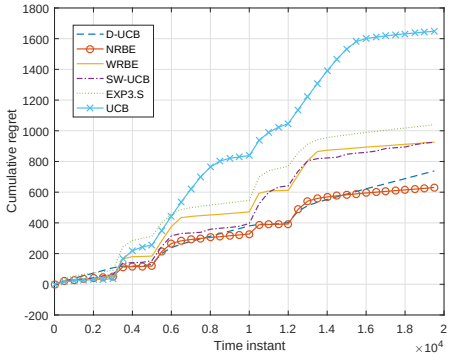

Fig. 3. Cumulative regret for the first scenario with piece-wise stationary rewards. The lowest cumulative regret is achieved by the proposed NRBE policy and the D-UCB policy. The stationary UCB policy suffers a clear performance loss in this non-stationary environment.

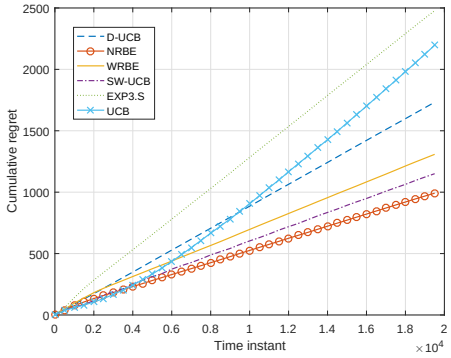

Fig. 4. Cumulative regret for the second scenario where the expected rewards fluctuate continuously. The best overall performance is achieved again by the NRBE policy, while the performance of the SW-UCB policy is not far.

are $\alpha = 5 \cdot 10^{-5}$ and $\gamma = 0.0658$. These values have been selected according to the parameter optimization derived in [5] and [9] for scenario 1. For the sake of comparison, the same parameter values are used for the NRBE and WRBE policies as the D-UCB and SW-UCB policies, i.e., $\gamma = 0.9982$, $\tilde{n} = W = 1780$. These parameter values are not optimal for the proposed policies. In our informal experiments in scenario 1 it was found that the NRBE policy achieves the lowest regret with $\gamma = 0.98$ and $\tilde{n} = 1700$. For the WRBE policy the best performance in scenario 1 was achieved with $W = 800$. However, it was found that the WRBE policy is very sensitive to the size W of the time window. On the other hand, it was noticed that out of all the simulated policies the NRBE is the least sensitive to its parameter values γ and $\tilde{n}$. The exploration bonus used in both NRBE and WRBE is

$$g\left(\frac{\tilde{n}}{\tilde{n}-\delta_k(n)}\right) = \sqrt{\log\left(\frac{\tilde{n}}{\tilde{n}-\delta_k(n)}\right)},$$

where the in the case of WRBE $\tilde{n}$ is replaced by $\min(n, W)$.

Fig. 3 shows the regret of the policies in the first scenario with piece-wise stationary rewards. It can be seen that the performances of the NRBE and D-UCB policies are on par, the NRBE providing slightly better performance. The performance of the WRBE policy is close to that of the SW-UCB policy. The performance of the stationary UCB policy in a non-stationary environment is poor.

In Fig. 4 the cumulative regret for the second scenario where the mean rewards fluctuate continuously are shown. The proposed NRBE achieves again the lowest overall regret while the regret of the SW-UCB comes fairly close.

The cumulative average regrets of the policies in the third stationary scenario are shown in Fig. 5. As expected the stationary UCB policy achieves the best performance, while the NRBE and WRBE policies achieve notably lowest regrets among the non-stationary policies.

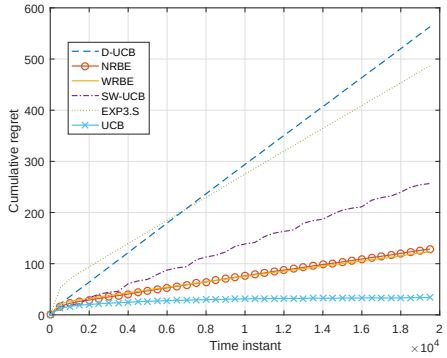

Fig. 5. Cumulative regret in a stationary scenario. The non-stationary policies suffer a performance loss compared tot he stationary UCB-policy. The smallest performance loss is achieved by the proposed WRBE and NRBE policies.

VI. Conclusions

In this paper we proposed two learning-based sensing policies for opportunistic spectrum access, where the spectrum opportunities and the expected data rates may be non-stationary. Using the stochastic non-stationary restless multi-armed bandit formulation we derived two policies, namely the NRBE and WRBE policies, for learning the optimal frequency bands in multi-band cognitive radio. The non-stationarity assumption is applicable for realistic scenarios where the state of the spectrum and consequently the rewards are often non-stationary. Since the proposed policies in this paper are index policies, they are extremely simple to implement. We have defined the index of a frequency band to be the sum of its discounted average observed data rate and a recency-based exploration bonus. The exploration bonus has been designed such that it encourages sensing frequency bands that have not been sensed for a long time, however such that the number of time instances that any band can stay unexplored is limited. The simulation examples have demonstrated that the proposed policies, in particular the NRBE policy, often provide better performance than other state-of-the-art policies. In our future work, we intend to provide rigorous analysis of the performance of the NRBE and WRBE policies. Such analysis may give insight into policies' sensitivity to the parameter values as well as suggest optimal parameter tuning when a priori knowledge about number of change points is available.

REFERENCES

[1] Q. Zhao, S. Geirhofer, L. Tong, and B. M. Sadler, "Opportunistic spectrum access via periodic channel sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 785–796, Feb. 2008.

[2] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multi-armed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, March 2013.

[3] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5588–5611, August 2012.

[4] J. Oksanen and V. Koivunen, "An order optimal policy for exploiting idle spectrum in cognitive radio networks," *Signal Processing, IEEE Transactions on*, 2015.

[5] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Algorithmic Learning Theory*, ser. Lecture Notes in Computer Science. Springer, 2011, vol. 6925, pp. 174–188.

[6] L. Kocsis and C. Szepesvári, "Discounted UCB," in *2nd PASCAL Challenges Workshop*, 2006.

[7] A. Alaya-Feki, E. Moulines, and A. LeCornec, "Dynamic spectrum access with non-stationary multi-armed bandit," in *IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, July 2008, pp. 416–420.

[8] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, 2nd ed. Cambridge, MA: MIT Press, 1998, 325 pages.

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[10] S. Vakili, Q. Zhao, and Y. Zhou, "Time-varying stochastic multi-armed bandit problems," in *48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 2103–2107.

[11] C. Hartland, N. Baskiotis, S. Gelly, O. Teytaud, and M. Sebag, "Multi-armed bandit, dynamic environments and meta-bandits," in *Online Trading of Exploration and Exploitation Workshop, NIPS*, December 2006.

[12] A. Alaya-Feki, B. Sayrac, E. Moulines, and A. L. Cornec, "Opportunistic spectrum access: Online search of optimality," in *IEEE Global Telecommunications Conference*, Nov 2008, pp. 1–5.

[13] J. Mellor and J. Shapiro, "Thompson sampling in switching environments with Bayesian online change detection," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 2013, pp. 442–450.

[14] A. Robin and R. Feraud, "Exp3 with drift detection for the switching bandit problem," in *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics*, 2015, pp. xx–yy.

[15] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.