This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2019.2941717, IEEE Communications Letters

Spectrum Sensing across Multiple Service Providers: A Discounted Thompson Sampling Method

Min Zhou, Tianyu Wang, *Member, IEEE*, and Shaowei Wang, *Senior Member, IEEE*

***Abstract*—Dynamic spectrum access is a potential solution to the long-standing radio spectrum scarcity and usage inefficiency issue, for which spectrum sensing is one of the key challenges. In this paper, we investigate the spectrum sensing order problem in the scenario that the secondary user equipped with limited sensing capability can get access to the idle spectrum across multiple network service providers opportunistically by spectrum sensing, where the probability of the spectrum being idle varies at temporal scale and is not available for the users. We employ an online learning method, referred to as discounted Thompson sampling, to address the formulated optimization task, which can track the changes of the probability of the spectrum availability and yields more spectrum access opportunities compared to other methods.**

***Index Terms*—Online learning, spectrum sensing, Thompson sampling.**

I. Introduction

During the past decades, the rapid growth of mobile data traffic incurs radio spectrum scarcity issue facing mobile service providers (SPs). However, investigations have also shown that the spectrum licensed on different agencies is underutilized. That is, part of the licensed spectrum, which is generally divided into a number of channels in mobile communication systems, is idle at temporal and spatial scales, leading to spectrum usage inefficiency. Dynamic spectrum access is deemed as a potential solution to spectrum scarcity and usage inefficiency problem since the secondary users (SUs) are expected to exploit and get access to the licensed spectrum opportunistically provided that they do not throw too much interference to the licensed primary users [1, 2]. It has attracted much attention in the past two decades.

Spectrum sensing is the prerequisite of dynamic spectrum access. However, spectrum sensing is generally expensive and time-consuming since it is difficult for an SU to sense a wide range of spectrum due to the hardware limitation [3]. As a result, spectrum sensing order is important for the SU in dynamic spectrum access systems. Ideally, an SU should sense the channel with the highest probability of being idle. In [4], a descending channel sensing order is proposed given the prior idle probability of channels. If the activities of primary users are unknown, an SU can sense the channels based on the achievable rates of channels as discussed in [5]. The case of the unknown idle probability of channels is investigated in [6], where a Q-learning method is introduced to improve system performance. A two-stage learning algorithm is proposed in [7], where the SU performs channel selection with reinforcement learning to minimize the time spent on channel sensing.

However, the aforementioned methods cannot cope with the following common scenarios: First, the probability of a given portion of spectrum being idle, e.g., the spectrum licensed to an agency, is usually unavailable; second, even though we could estimate the probability of spectrum being idle by observing it for quite a long time, it could change aperiodically. In other words, we may use outdated statistical information to make decision in this case, resulting in deteriorating system performance. These facts motivate us to set up a reasonable spectrum sensing model and develop efficient algorithm to track the changes of the idle probability of spectrum for dynamic spectrum access systems.

In this paper, we investigate the spectrum sensing problem across multiple SPs, where the probability of spectrum being idle is different for each SP. Moreover, the statistical law of the probability of idle spectrum also varies at temporal scale. We formulate an online learning framework to track the spectrum idle probability and introduce an efficient algorithm, the discounted Thompson sampling (DTS), to address the formulated optimization task. Numerical results show that our proposed method yields spectrum access opportunities almost as many as the best fixed decision in hindsight. The latter is impossible to implement since it needs a prophet. Moreover, our proposed algorithm can track the changes of the idle probability of spectrum more quickly than other methods as discussed in [8–11].

The rest of this paper is organized as follows. In Section II, problem formulation is demonstrated. We give the DTS-based spectrum sensing approach in details in Section III. Section IV presents simulation results and discussions. Conclusions are given in Section V.

Manuscript received July 2, 2019; revised August 4, 2019; accepted September 3, 2019. This work was partially supported by the National Natural Science Foundation of China (61671233, 61801208, 61931023), the Jiangsu Science Foundation (BK20170650), the Postdoctoral Science Foundation of China (BX201700118, 2017M621712), the Jiangsu Postdoctoral Science Foundation (1701118B), and the the open research fund of National Mobile Communications Research Laboratory (2019D02). The associate editor coordinating the review of this letter and approving it for publication was O. Popescu. (*Min Zhou and Tianyu Wang are co-first authors.*) (*Corresponding author: Shaowei Wang.*)

M. Zhou and S. Wang are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: MG1623068@smail.nju.edu.cn; wangsw@nju.edu.cn).

T. Wang is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China, and also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: tianyu.alex.wang@nju.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2019.2941717, IEEE Communications Letters

II. Problem Formulation

Consider spectrum sensing order, also referred to as channels selection across multiple SPs. In practical commercial mobile communication systems, different mobile SPs are assigned different frequency bands for data transmission by radio regulatory agencies. Specifically, during the same time period, the number of active users served by each SP is different in the same region, so the probability of the licensed channels being idle is different for each SP. Moreover, the idle probability of the channels licensed to each SP will vary after a period of time due to the mobility of users. The key challenge for the SU equipped with limited sensing capability is to decide which portion of the spectrum should be sensed to produce more channels free of primary user activities to transmit data by the SU.

Consider that there are $N = 3$ SPs and the spectrum licensed to SP i consists of C non-overlapping channels. Due to the limited spectrum sensing capability of hardware, an SU cannot sense a large range of spectrum. Specifically, at given time slot t, the SU can only keep sensing the channels licensed to one SP which fall into the same frequency band range. We investigate the total data throughput of the SU who utilizes the idle channels by spectrum sensing in T time slots. The probability of spectrum being idle for SP i at time slot t is denoted by P_i, where $i \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$. In practical scenarios, each channel of an SP could have own probability of being idle. Assume that the reward of each channel c of SP i obtained at time slot t, which is denoted as $H_{i,c}$, is Bernoulli with probability P_i, which can be written as $H_{i,c} \sim Bernoulli(P_i)$ for $c \in \{1, 2, ..., C\}$. If channel c is available for the SU, the reward of channel c equals 1; otherwise, it equals 0. The system model is shown in Fig. 1. Recall that P_i cannot be known in advance and would change at temporal scale so that it makes no sense to estimate P_i by observing it for a long time. It is necessary to fulfill spectrum sensing task in an online manner and develop efficient algorithm to keep track of the changes of P_i quickly.

At time slot $t \in \{1, 2, ..., T\}$, an SU chooses SP $i \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$ and senses over the corresponding spectrum. The number of licensed channels being idle obtained by the SU is denoted as $R_{i(t)}(t)$ at time slot t where $i(t)$ is the index of the SP chosen at time slot t. According to the definition of $H_{i,c}$, we have $R_{i(t)}(t) = \sum_{c=1}^{C} H_{i(t),c}$, which is $binomial(C, P_i)$. The SU should dynamically choose the SPs so as to maximize the expected number of idle channels obtained during T time slots. Therefore, our optimization problem can be formulated as follows:

$$\max_{i(t)\in\{\mathrm{I},\mathrm{II},\mathrm{III}\}} \mathbb{E}[\sum_{t=1}^{T} R_{i(t)}(t)]. \tag{1}$$

The optimization task illustrated in (1) requires that the SU should select a sequence of SPs in advance so as to obtain maximum cumulative number of idle channels. The number of idle channels obtained at each time slot is related to the unknown time-varying probability of spectrum being idle, which should be learned in an online manner. The optimization problem defined by (1) can be considered as a multi-armed bandit (MAB) problem [12, 13], where the SP to be selected can be regarded as an arm in the MAB problem. The number of idle channels obtained of selecting SP i at time slot t follows a certain distribution with unknown parameter P_i. The objective is to choose the optimal SP (pull the optimal arm) to sense over its assigned spectrum at time slot t to maximize the expected number of idle channels (reward in the MAB problem) so that the SU has more opportunities to transmit data during T time slots.

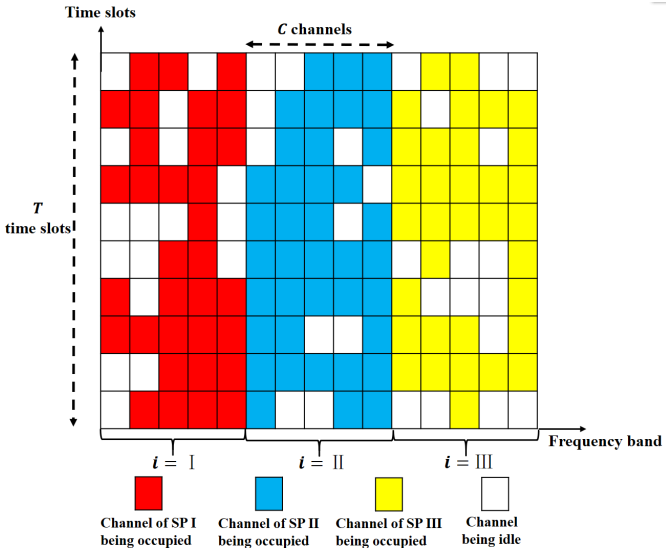

Fig. 1. An illustration of the channels of 3 SPs spanning in frequency and time slots.

III. Discounted Thompson Sampling Algorithm

In the MAB problem, the decision maker needs to select one of the optional arms to pull at each time slot to maximize the expected value of reward. Thompson sampling (TS) is an efficient algorithm to deal with the MAB problem. The TS has access to the history of pulled arms and the reward in the previous time slots and employs this history to select the promising arm at current time slot [14, 15]. Comprehensive theoretical analysis of regret upper bound and performance guarantee of TS can be found in [16]. Since the general TS cannot address our optimization problem with non-stationary arms directly, we introduce a discounted Thompson sampling method.

Binary reward of 0 or 1 is defined for each arm in the classical MAB problem while the reward in our case, i.e., the number of idle channels obtained from SP i at time slot t, is an integer distributed in a bounded interval randomly. We can scale the reward and do a Bernoulli trial with success probability of the scaled reward [14], where r_i is a binary variable of the observed output of Bernoulli trial. However, we cannot know the corresponding parameter of Bernoulli distribution related to P_i. Let n_i represent the total time slots that SP i has been selected by now. S_i denotes the total time slots of $r_i = 1$ and F_i is the total time slots of $r_i = 0$. That is, $S_i + F_i = n_i$. Let H_i represent a vector consisting of all r_i's observed so far when SP i is selected. The parametric

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2019.2941717, IEEE Communications Letters

TABLE I

Algorithm 1 Discounted Thompson sampling algorithm

Parameters: $\gamma \in (0, 1]$
for each provider $i \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$, *Initialize*: $\alpha_i = \beta_i = 1$, $S_i = F_i = 0$;
for $t = 1, 2, ..., T$ **do**
  for $i \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$ **do**
    Update $S_i = \gamma S_i$, $F_i = \gamma F_i$.
    Draw $\theta_i(t) \sim Beta(S_i + \alpha_i, F_i + \beta_i)$.
  end for
  Choose provider $i = \arg\max_i \theta_i$ and observe the reward
  Do Bernoulli trial and observe the output r_i
  if $r_i = 1$ **then**
    $S_i = S_i + 1$;
  else
    $F_i = F_i + 1$;
  end if
end for

likelihood function of H_i is given by

$$p_i(H_i|P_i) = P_i^{S_i}(1 - P_i)^{F_i}, \tag{2}$$

which corresponds to Binomial distribution. Assume that the prior probability of P_i obeys Beta distribution, the posterior probability of P_i, denoted as $p_i(P_i|H_i)$, can be updated with $p_i(H_i|P_i)$, which is deduced in the appendix.

The SU selects an SP and performs spectrum sensing on the corresponding channels at each time slot using **Algorithm 1**, where the hyperparameter γ controls the confidence level of S_i (or F_i) of each SP, to avoid being trapped into always selecting an SP if S_i of the SP is too large. Compared with the Dynamic TS proposed in [11], our DTS method updates S_i (or F_i) of all SPs at each time slot, while the Dynamic TS only updates the parameters of the selected SP. In addition, our method applies the discount factor during all time slots, while the Dynamic TS uses it only if the condition $S_i + F_i > J$ is satisfied for the selected SP, where J is a threshold deciding whether to introduce a discount factor or not.

IV. Simulation Results

Consider 3 SPs licensed different radio spectrum which consists of $C = 20$ channels. P_i of each SP i follows standard uniform distribution, that is, P_i is sampled from uniform distribution bounded in $(0, 1)$. Each channel of SP i yields reward according to Bernoulli distribution with parameter P_i. To show the performance gain of our proposed algorithm, we compare our DTS algorithm with other representative ones: discounted ϵ-greedy, discounted upper confidence bound (Discounted UCB) [9] and dynamic TS [11]. For the UCB algorithm, the SU chooses each SP once at first. Then at any time slot $t > 3$, the SU selects the SP that maximizes the defined upper confidence index which is expressed as a simple function of the empirical distribution of reward. For the ϵ-greedy algorithm, the SU chooses an SP randomly with probability ϵ, or chooses the SP that maximizes the average reward with probability $1 - \epsilon$. For our considered scenarios, $\epsilon = 0.1$. Compared with standard UCB and ϵ-greedy algorithms, their discounted versions (Discounted UCB and Discounted ϵ-greedy) introduce a discount factor to average past rewards in order to reduce the effect of the past observations. For the Discounted ϵ-greedy algorithm, we employ the same discounting method to handle the average reward as the Discounted UCB [9].

Figure 2 illustrates the normalized regret as a function of time slots for different settings of P_i. P_i remain unchanged in $T = 500$ time slots and the hyperparameter γ equals 1. The normalized regret is the ratio of the cumulative reward difference between the oracle decision in hindsight and the mentioned algorithms over the cumulative reward of the oracle decision. All results are averaged by 100 Monte Carlo simulations. In setting 1, 2, 3, the prior P_i of each SP $i \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$ is $\{0.10, 0.28, 0.37\}$, $\{0.27, 0.24, 0.20\}$ and $\{0.10, 0.80, 0.25\}$, respectively. Always choosing the SP with the largest P_i for spectrum sensing all through 500 time slots is the oracle decision if the SU know P_i in advance, which always yields regret of 0. It can be seen from Fig. 2 that the normalized regret of DTS can reach a small value for all three different settings of P_i, indicating that the total number of available channels obtained by DTS is only slightly less than that of the oracle decision even if the difference in P_i among three SPs is small. Recall that the DTS dose not require the prior information of P_i, facilitating its applications in practical dynamic spectrum systems.

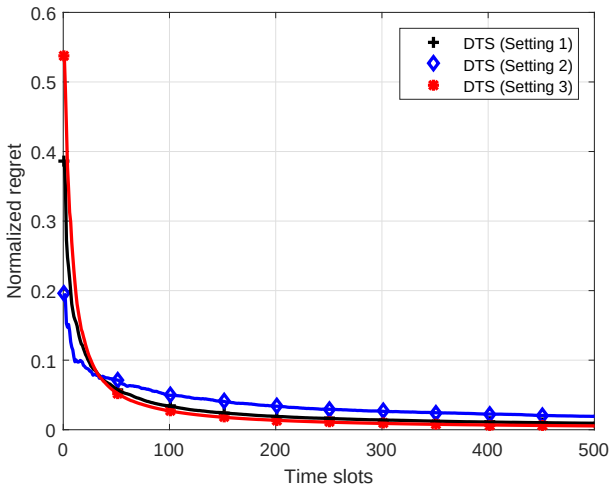

Fig. 2. Normalized regret as a function of the number of time slots for different settings of P_i.

Figure 3 shows the normalized throughput as a function of time slots with fixed P_i for setting 1, where the normalized throughput is defined as the ratio of the total number of idle channels obtained by the mentioned algorithms to that of the oracle decision in hindsight, that is, the SU could always select SP III and the normalized throughput of the oracle decision in hindsight is always 1. In Fig. 3, we can find that the total number of idle channels produced by the four algorithms is close to the oracle decision in hindsight provided that time slots are large enough. Compared with the Discounted UCB and the Discounted ϵ-greedy methods, the DTS algorithm yields higher throughput since it can tend to select the SP with the largest P_i quickly, bringing the SU more opportunities to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2019.2941717, IEEE Communications Letters

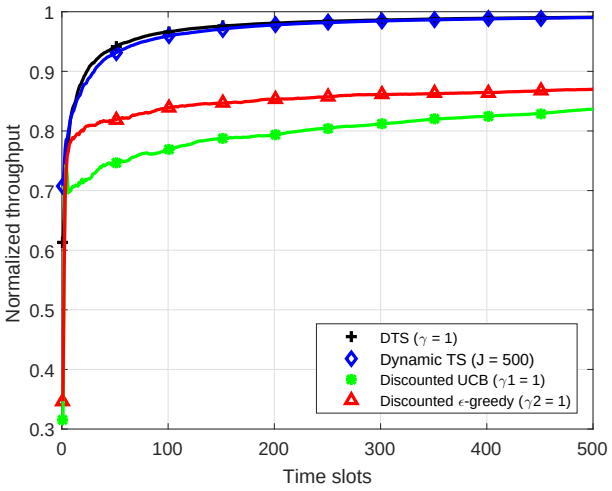

Fig. 3. Normalized throughput as a function of the number of time slots with fixed P_i.

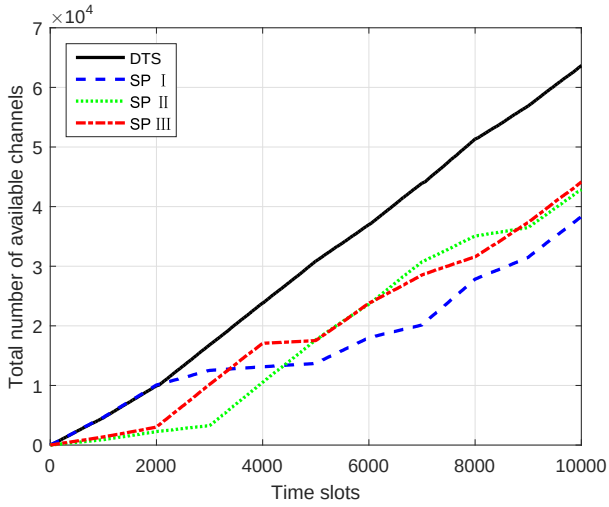

Fig. 4. Total number of obtained idle channels as a function of the number of time slots with time-varying P_i.

get idle channels for data transmission.

TABLE II. Time-varying P_i

P_i \ k	1	2	3	4	5	6	7	8	9	10
P_{I}	0.23	0.27	0.12	0.03	0.03	0.22	0.10	0.39	0.18	0.34
P_{II}	0.05	0.07	0.05	0.36	0.36	0.30	0.36	0.22	0.07	0.31
P_{III}	0.07	0.08	0.36	0.35	0.03	0.32	0.24	0.15	0.30	0.34

Considering the time-varying P_i of SP i, we investigate the total number of idle channels as a function of time slots when P_i changes every 1000 time slots. The hyperparameter γ is set to 0.99 based on a series of numerical experiments. The prior P_i is shown in Table II, where k represents the kth change of P_i. As can be seen in Fig. 4, the gap between the DTS and any scheme of always choosing a given SP becomes larger and larger as the increasing of time slots. The total number of idle channels obtained by our proposed DTS algorithm is at least 52% more than others, indicating that the DTS algorithm can track the changes of P_i fairly quickly.

Figure 5 shows the normalized throughput as a function of time slots with time-varying P_i, which further indicates that our proposed DTS can track the changes of P_i. Moreover, the total number of idle channels produced by the DTS is closer to that of the oracle decision in hindsight compared with others. The reason is that the proposed DTS can find the SP with the largest P_i faster than the Discounted UCB, the Discounted ϵ-greedy and the Dynamic TS without knowing the prior knowledge about P_i.

Figure 6 shows the normalized regret as a function of time slots with time-varying P_i. The normalized regret demonstrates the performance gap between the oracle decision in hindsight and the mentioned four algorithms, and can be regarded as the tracking error of different algorithms. In Fig. 6, it can be seen that the normalized regret of the four algorithms reach a

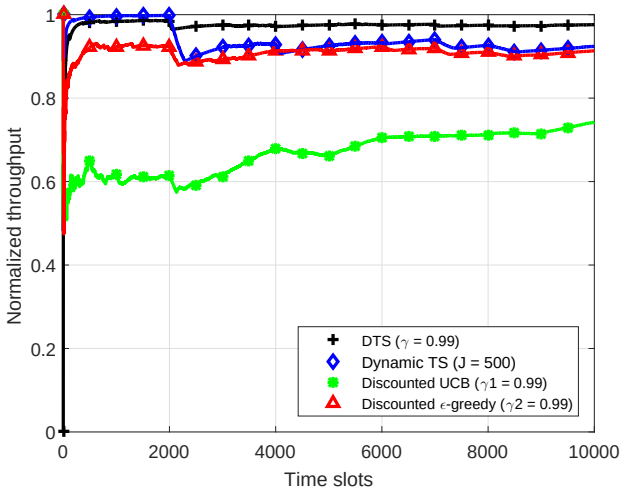

Fig. 5. Normalized throughput as a function of the number of time slots with time-varying P_i.

small and almost constant value provided that the time slots are large enough even though P_i changes at temporal scale. Compared with the Discounted UCB, the Discounted ϵ-greedy and the Dynamic TS, our proposed DTS algorithm produces a lower and stabler average regret.

V. CONCLUSION

We investigated the channels selection problem across multiple commercial mobile communication service providers without prior knowledge about the probability of channels being idle. The optimization task is formulated as an online learning problem and we developed an efficient discounted Thompson sampling algorithm to solve it. Numerical results show that our proposed method can generate spectrum access opportunities almost as many as the oracle decision in hindsight. The idle channels obtained by our proposed method are much more than the discounted upper confidence bound, the discounted ϵ-greedy and the dynamic Thompson sampling

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2019.2941717, IEEE Communications Letters

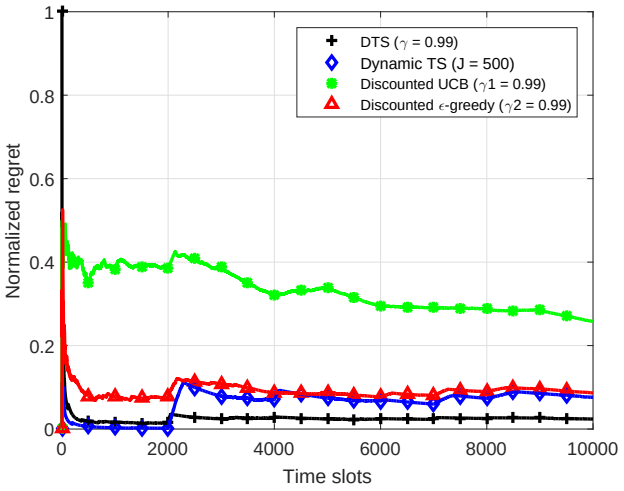

Fig. 6. Normalized regret as a function of the number of time slots with time-varying P_i.

methods. Moreover, our proposed algorithm can adapt to the changes of the probability of channels being idle.

APPENDIX

Since Beta distribution is conjugate prior of Binomial distribution and P_i obeys a prior Beta distribution, the posterior probability of P_i can be written as follows based on Bayes rule [17]:

$$p_i(P_i|H_i) = \frac{p_i(H_i|P_i)\dfrac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}P_i^{\alpha_i-1}(1-P_i)^{\beta_i-1}}{p_i(H_i)}, \quad (3)$$

where

$$\Gamma(\alpha_i) = \int_0^\infty x^{\alpha_i-1}e^{-x}\,dx, \quad (4)$$

and α_i and β_i are two parameters which determines the mean and variance of Beta distribution. Since no prior knowledge about P_i is available at the beginning, we initialize $\alpha_i = \beta_i = 1$ and $Beta(1,1)$ is actually uniform distribution. Substituting (2) in (3), $p_i(P_i|H_i)$ can be rewritten as

$$p_i(P_i|H_i) = \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)p_i(H_i)}P_i^{S_i+\alpha_i-1}(1-P_i)^{F_i+\beta_i-1}. \quad (5)$$

Let $M = \Gamma(\alpha_i+\beta_i)/[\Gamma(\alpha_i)\Gamma(\beta_i)p_i(H_i)]$, we have

$$p_i(P_i|H_i) = MP_i^{S_i+\alpha_i-1}(1-P_i)^{F_i+\beta_i-1}. \quad (6)$$

Notice that $\int p_i(P_i|H_i)\,dP_i = 1$ and $\int x^{\alpha_i-1}(1-x)^{\beta_i-1}\,dx = \Gamma(\alpha_i)\Gamma(\beta_i)/\Gamma(\alpha_i+\beta_i)$, we have

$$p_i(P_i|H_i) = \frac{\Gamma(\alpha_i+\beta_i+n_i)}{\Gamma(\alpha_i+S_i)\Gamma(\beta_i+F_i)}P_i^{S_i+\alpha_i-1}(1-P_i)^{F_i+\beta_i-1}, \quad (7)$$

which also corresponds to Beta distribution with parameters of $S_i+\alpha_i$ and $F_i+\beta_i$. Then the posterior probability of P_i can be updated as follows:

$$p_i(P_i|H_i) = Beta(S_i+\alpha_i, F_i+\beta_i). \quad (8)$$

ACKNOWLEDGEMENT

The authors would like to thank the editors and the anonymous reviewers whose invaluable comments helped substantially improve the presentation of this paper.

REFERENCES

[1] S. Wang, Z. Zhou, M. Ge, and C. Wang, "Resource allocation for heterogeneous cognitive radio networks with imperfect spectrum sensing," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 3, pp. 464–475, Mar. 2013.

[2] Y. Zhang and S. Wang, "Resource allocation for cognitive radio-enabled femtocell networks with imperfect spectrum sensing and channel uncertainty," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7719–7728, Sept. 2016.

[3] A. Ghasemi and E. S. Sousa, "Spectrum sensing in cognitive radio networks: requirements, challenges and design trade-offs," *IEEE Commun. Mag.*, vol. 46, no. 4, pp. 32–39, Apr. 2008.

[4] H. Jiang, L. Lai, R. Fan, and H. V. Poor, "Optimal selection of channel sensing order in cognitive radio," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 297–307, Jan. 2009.

[5] H. T. Cheng and W. Zhuang, "Simple channel sensing order in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 676–688, Apr. 2011.

[6] H. Li, "Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: A two by two case," in *Proc. IEEE SMC*, San Antonio, TX, Oct. 2009, pp. 1893–1898.

[7] V. Raj, I. Dias, T. Tholeti, and S. Kalyani, "Spectrum access in cognitive radio using a two-stage reinforcement learning approach," *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 1, pp. 20–34, Feb. 2018.

[8] R. Vishnu and S. Kalyani, "Taming non-stationary bandits: A Bayesian approach," *arXiv preprint arXiv:1707.09727*, 2017.

[9] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," in *Proc. ALT.*, Oct. 2011, pp. 174–188.

[10] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *Proc. NeurIPS*, Dec. 2014, pp. 199–207.

[11] N. Gupta, O.-C. Granmo, and A. Agrawala, "Thompson sampling for dynamic multi-armed bandits," in *Proc. ICMLA.*, Dec. 2011, pp. 484–489.

[12] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, Dec. 2012.

[13] Y. Lin, T. Wang, and S. Wang, "UAV-assisted emergency communications: An extended multi-armed bandit perspective," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 938–941, May 2019.

[14] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. COLT*, Jun. 2012, pp. 1–26.

[15] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Proc. NeurIPS*, Dec. 2011, pp. 2249–2257.

[16] S. Agrawal and N. Goyal, "Further optimal regret bounds for Thompson sampling," in *Proc. AISTATS*, May. 2013, pp. 99–107.

[17] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen *et al.*, "A tutorial on Thompson sampling," *Found. Trends Mach. Learn.*, vol. 11, no. 1, pp. 1–96, Jul. 2018.