Efficient Online Learning for Opportunistic Spectrum Access

Wenhan Dai[†], Yi Gai[♯] and Bhaskar Krishnamachari[♯]
[†]Massachusetts Institute of Technology, Cambridge, MA, USA
[♯]University of Southern California, Los Angeles, CA, USA
Email: whdai@mit.edu, {ygai, bkrishna}@usc.edu

***Abstract*—The problem of opportunistic spectrum access in cognitive radio networks has been recently formulated as a non-Bayesian restless multi-armed bandit problem. In this problem, there are N arms (corresponding to channels) and one player (corresponding to a secondary user). The state of each arm evolves as a finite-state Markov chain with unknown parameters. At each time slot, the player can select $K < N$ arms to play and receives state-dependent rewards (corresponding to the throughput obtained given the activity of primary users). The objective is to maximize the expected total rewards (i.e., total throughput) obtained over multiple plays. The performance of an algorithm for such a multi-armed bandit problem is measured in terms of regret, defined as the difference in expected reward compared to a model-aware genie who always plays the best K arms. In this paper, we propose a new continuous exploration and exploitation (CEE) algorithm for this problem. When no information is available about the dynamics of the arms, CEE is the first algorithm to guarantee near-logarithmic regret uniformly over time. When some bounds corresponding to the stationary state distributions and the state-dependent rewards are known, we show that CEE can be easily modified to achieve logarithmic regret over time. In contrast, prior algorithms require additional information concerning bounds on the second eigenvalues of the transition matrices in order to guarantee logarithmic regret. Finally, we show through numerical simulations that CEE is more efficient than prior algorithms.**

I. Introduction

Multi-arm bandit (MAB) problems are widely used to make optimal decisions in dynamic environments. In the classic MAB problem, there are N independent arms and one player. At every time slot, the player selects $K(\geq 1)$ arms to sense and receives a certain amount of rewards. In the classic non-Bayesian formulation, the reward of each arm evolves in i.i.d. over time and is unknown to the player. The player seeks to design a policy which can maximize the expected total reward.

One variant of multi-armed bandits is the restless multi-arm bandit problem (RMAB). In this case, all the arms, whether selected (activated) or not, evolve as a Markov chain at every time slot. When one arm is played, its transition matrix may be different from that when it is not played. Even if the player knows the parameters of the model, referred as the Bayesian RMAB since the beliefs on each arm can be updated at each time based on the observations, the design of the optimal policy turns to be a PSPACE hard optimization problem [2].

In this paper, we consider the more challenging non-Bayesian RMAB problems, in which parameters of the model are unknown to the player. The objective is to minimize *regret*, defined as the gap between the expected reward that can be achieved by a suitably defined genie that knows the parameters and that obtained by the given policy. As stated before, finding the optimal policy, which is in general non-stationary, is P-SPACE hard even if the parameters are known. So we use instead a weaker notion of regret, where the genie always selects the K most rewarding arms that have highest stationary rewards when activated.

We propose a sample mean-based index policy without information about the system. We prove that this algorithm achieves regret arbitrarily close to logarithmic uniformly over time horizon. Specifically, the regret can be bound by $Z_1G(n)\ln n + Z_2\ln n + Z_3G(n) + Z_4$, where n is time, $Z_i, i = 1, 2, 3, 4$ are constants and $G(n)$ can be any divergent non-decreasing sequence of positive integers. The significance of such a sub-linear time regret bound is that the time-averaged regret tends to zero (or possibly even negative since the genie we compare with is not using a globally optimal policy), implying the time-averaged rewards of the policy will approach or even possibly exceed those obtained by the stationary policy adopted by the model-aware genie.

If some bounds corresponding to the stationary state distributions and the state-dependent rewards are known, we show that the algorithm can be easily modified and achieves logarithmic regret over time. Compared to prior work, our algorithm requires the least information about the system. Moreover, our simulation results show that our algorithm obtains the lowest regret compared to previously proposed algorithms when the parameters just satisfy the theoretical boundaries.

Research in restless multi-arm bandit problems has a lot of applications. For instance, it has been applied to dynamic spectrum sensing for opportunistic spectrum access in cognitive radio networks, where a secondary user must select K of N channels to sense at each time to maximize its expected reward from transmission opportunities. If the primary user occupancy on each channel is modeled as a Markov chain with unknown parameters, then we obtain an RMAB problem. We conduct our simulation-based evaluations in the context of this

This research was sponsored in part by the U.S. Army Research Laboratory under the Network Science Collaborative Technology Alliance, Agreement Number W911NF-09-2-0053, and by the U.S. National Science Foundation under award number CNS-1049541.

particular problem of opportunistic spectrum access.

The remainder of this paper is organized as follows: in Section II, we briefly review related work on MAB problems. In Section III, we formulate the general RMAB problem. In Section IV, we introduce a sample mean based policy and provide a proof for the regret upper bound for single channel selection cases. In Section V, we evaluate our algorithm and compare it via simulations with two previous proposed algorithm. We conclude the paper in Section VI.

II. Related Work

In 1985, Lai and Robbins proved that the minimum regret grows with time in a logarithmic order [8]. They also proposed the first policy that achieved the optimal logarithmic regret for multi-armed bandit problems in which the rewards are i.i.d. over time. Auer *et al.* developed UCB1 policy in 2002, applying to i.i.d. reward distributions with finite support, achieving logarithmic regret over time [9]. Their policy is based on the sample mean of the observed data, and has a rather simple index selection method.

One important variant of classic multi-armed bandit problem is the Bayesian MAB. In this case, *a priori* probabilistic knowledge about the problem and system is required. Gittins and Jones presented a simple approach for the rested bandit problem, in which one arm is activated at each time and only the activated arm changes state as a known Markov process [6]. The optimal policy is to play the arm with highest Gittins' index. The *restless bandit problem* was posed by Whittle in 1988 [1], in which all the arms can change state. The optimal solution for this problem has been shown to be PSPACE-hard by Papadimitriou and Tsitsiklis [2]. The restless bandit problem has no general solution though it may be solved in special cases. For instance, when each channel is modeled as identical two-state Markov chain, the myopic policy is proved to be optimal if the channel number is no more than 3 or is positively correlated [7].

There have been a few recent attempts to solve the restless multi-arm bandit problem under unknown models. In [11], Tekin and Liu use a weaker definition of regret and propose a policy (RCA) that achieves logarithmic regret when certain knowledge about the system is known. The algorithm only exploits part of observing data and leaves space to improve performances. In [5], Haoyang Liu *et al.* proposed a policy, referred to as RUCB, achieving a logarithmic regret over time when certain system parameters are known. The regret they adopt is the same as in [11]. They also extend the RUCB policy to achieve a near-logarithmic regret over time when no knowledge about the system is available. However, they only give the upper bound of regret at the end of a certain time point referred as *epoch*. When no *a priori* information about the system is known, their analysis of regret gives the upper bound over time only asymptotically, not uniformly.

In our previous work [4], we adopted a stronger definition of regret, which is defined as the reward loss with the optimal policy. Our policy achieve a near-logarithmic regret without *a prior* of the system. It applies to special cases of the RMAB, in particular the same scenario as in [7].

III. Problem Formulation

We consider a time-slotted system with one player and N independent arms. At each time slot, the player selects $K(< N)$ arms and gets a certain amount of rewards according to the current state of the arm. Each arm is modeled as a independent discrete-time, irreducible and aperiodic Markov chain with finite state space. Generally, the transition matrices in the activated model and the passive model are not necessarily identical. The player can only see the state of the sensed arm and does not know the transitions of the arms. The player aims to maximize its expected total reward (throughput) over some time horizon by choosing judiciously a sensing policy (algorithm) ϕ that governs the channel selection in each slot based on observing history.

Let S^i denote the state space of arm i. Denote r_x^i the reward obtained from state x of arm i, $x \in S^i$. Without loss of generality, we assume $r_x^i \leq 1, \forall x \in S^i, \forall i$. Let P_j denote the active transition matrix of arm j and Q_j denote the passive transition matrix. Let $\pi^i = \{\pi_x^i, x \in S^i\}$ denote the stationary distribution of arm i in the active model, where π_x^i is the stationary probability of arm i being in state x (under P_i). The stationary mean reward of arm i, denoted by μ^i, is the expected reward of arm i under its stationary distribution:

$$\mu^i = \sum_{x \in S^i} r_x^i \pi_x^i \tag{1}$$

Consider the permutation of $\{1, \cdots, N\}$ denoted as σ, such that $\mu^{\sigma(1)} > \mu^{\sigma(2)} > \mu^{\sigma(3)} > \cdots \mu^{\sigma(N)}$. We are interested in designing policies that perform well with respect to $regret$, which is defined as the difference between the expected reward that is obtained by using the policy selecting K best arms and that obtained by the given policy. The best arm obtains the highest stationary mean reward.

Let $Y^{\Phi}(t)$ denote the reward obtained at time t with policy Φ. The total reward achieved by policy Φ is given by

$$R^{\Phi}(t) = \sum_{j=1}^{t} Y^{\Phi}(j) \tag{2}$$

and the regret $r^{\Phi}(t)$ achieved by policy Φ is given by

$$r^{\Phi}(t) = t \sum_{j=1}^{K} \mu^{\sigma(j)} - \mathbb{E}(R^{\Phi}(t)) \tag{3}$$

The objective is to minimize the growth rate of the regret.

IV. Analysis for Single Arm Selection

In this section, we focus on the situation when $K = 1$. In this case, the player selects one arm each time. We first show an algorithm called *Continuous Exploration and Exploitation* (CEE) and then prove that our algorithm achieves a near-logarithmic regret with time.

A. The CEE Algorithm for non-Bayesian RMAB

Our CEE algorithm (see Algorithm 1) works as follows. We first process the initialization by selecting each arm for certain time slots (we call these time slots *step*), then iterate the arm selection by searching the index that maximizes the equation shown in line 8 in Algorithm 1 and operating this arm for one *step*. A key issue is how long to operate each arm at each step. It turns out from the analysis we present in the next subsection that it is desirable to slowly increase the duration of each step using any (arbitrarily slowly) divergent non-decreasing sequence of positive integers $\{B_i\}_{i=1}^{\infty}$.

Algorithm 1 Continuous Exploration and Exploitation(CEE): Single Arm Selection

1: // INITIALIZATION
2: Play arm i for B_i time slots, denote $\hat{A}_i(1)$ as the sample mean of these B_i rewards, $i = 1, 2, \cdots, N$
3: $\hat{X}_i = \hat{A}_i(1)$, $i = 1, 2, \cdots, N$
4: $n = \sum_{i=1}^{N} B_i$
5: $i = N + 1$, $i_j = 1$, $j = 1, 2, \cdots, N$
6: // MAIN LOOP
7: **while** 1 **do**
8: Find j such that $j = \arg\max \frac{\hat{X}_j}{i_j} + \sqrt{\frac{L \ln n}{i_j}}$(L can be any constant greater than 2)
9: $i_j = i_j + 1$
10: Play arm j for B_i slots, let $\hat{A}_j(i_j)$ record the sample mean of these B_i rewards
11: $\hat{X}_j = \hat{X}_j + \hat{A}_j(i_j), i = i + 1, n = n + B_i$;
12: **end while**

B. Regret Analysis

We first define the discrete function $G(n)$, which represents the value of B_i, at the n^{th} time step in Algorithm 1:

$$G(n) = \min_{I} B_I \ s.t. \sum_{i=1}^{I} B_i \geq n \tag{4}$$

Since $B_i \geq 1$, it is obvious that $G(n) \leq B_n, \forall n$. Note that since B_i can be any arbitrarily slow non-decreasing diverging sequence, $G(n)$ can also grow arbitrarily slowly.

In this subsection, we show that the regret achieved by our algorithm has a near-logarithmic order. This is given in the following Theorem 1.

Theorem 1: Assume all arms are modeled as finite state, irreducible, aperiodic and reversible Markov chains. All the states (rewards) are positive. The expected regret with Algorithm 1 after n time slots is at most $Z_1 G(n) \ln n + Z_2 \ln n + Z_3 G(n) + Z_4$, where Z_1, Z_2, Z_3, Z_4 are constants only related to $P_i, i = 1, 2, \cdots, N$, explicit expressions are at the end of proof for Theorem 1.

The proof of Theorem 1 uses the following fact and two lemmas that we present next.

Fact 1: (Chernoff-Hoeffding bound) Let $X_1, \cdots, X_n$ be random variables with common range $[0, 1]$ and such that $\mathbb{E}[X_t | X_1, \cdots, X_{t-1}] = \mu$. Let $S_n = X_1 + \cdots + X_n$. Then for all $a \geq 0$, $\mathbb{P}\{S_n \geq n\mu + a\} \leq e^{-2a^2/n}$ and $\mathbb{P}\{S_n \leq n\mu - a\} \leq e^{-2a^2/n}$.

The first lemma is a non-trivial variant of the Chernoff-Hoeffding bound, first introduced in our recent work [4], that allows for bounded differences between the conditional expectations of sequence of random variables that we revealed sequentially:

Lemma 1: [4] Let $X_1, \cdots, X_n$ be random variables with range $[0, b]$ and such that $|\mathbb{E}[X_t | X_1, \cdots, X_{t-1}] - \mu| \leq C$. C is a constant number such that $0 < C < \mu$. Let $S_n = X_1 + \cdots + X_n$. Then for all $a \geq 0$,

$$\mathbb{P}\{S_n \geq n(\mu + C) + a\} \leq e^{-2(\frac{a(\mu - C)}{b(\mu + C)})^2 / n}$$

and

$$\mathbb{P}\{S_n \leq n(\mu - C) - a\} \leq e^{-2(a/b)^2 / n}$$

Proof: See [10]. ■

Lemma 2: [3] Consider an irreducible, aperiodic Markov chain with state space S, matrix of transition probabilities P, an initial distribution $\vec{q}$ which is positive in all states, and stationary distribution $\vec{\pi}(\pi_s$ is the stationary probability of state s). The state (reward) at time t is denoted by $s(t)$. Let μ denote the mean reward. If we play the chain for an arbitrary time T, then there exists a value $A_P \leq (\min_{s \in S} \pi_s)^{-1} \sum_{s \in S} s$ such that $\mathbb{E}[\sum_{t=1}^{T} s(t) - \mu T] \leq A_P$.

Lemma 2 shows that if a player keeps selecting the optimal arm, the difference between the expected reward and the highest stationary reward is bounded by a constant.

Based on these two lemmas, we can give the proof of Theorem 1 show as below.

Proof: Below is a sketch of the proof. A detailed proof can be found in [10].

$\sigma^{(1)}$ is the index of the optimal arm. The regret comes from two parts: the regret when selecting an arm other than arm $\sigma^{(1)}$; the difference between $\mu^{\sigma(1)}$ and $\mathbb{E}(Y^{\Phi}(t))$ when selecting arm $\sigma^{(1)}$. At most we lose a constant value from the second part of the regret by Lemma 2. Next we will show the number of selections of one arm other than $\sigma^{(1)}$ in line 8 is bounded by $O(\ln n)$, then the first part of regret can be bounded by $O(G(n) \ln n)$ and the total regret can be bounded by $O(G(n) \ln n)$.

For ease of exposition, we discuss the time slots n such that $G||n$, where $G||n$ denotes the time n is the end of certain step. We define q as the smallest index such that

$$B_q \geq \lceil \max\{\frac{2C_P}{\mu^{\sigma(1)} - \mu^{\sigma(2)}}, \frac{C_P}{\mu^{\sigma(l)}}, l = 1, 2, \cdots, N\} \rceil$$

where $C_P = \max\{(\min_{x \in S^i} \pi_x^i)^{-1} \sum_{s \in S^i} s, 1 \leq i \leq N\}$.

We denote the following intermedium variables for ease of expression: $c_{t,s} = \sqrt{(L \ln t)/s}$, $w^* = q(\mu^{\sigma(1)} - \frac{C_P}{B_q})$ and $w^i = q\frac{\mu^{\sigma(i)} - C_P/B_q}{\mu^{\sigma(i)} + C_P/B_q}(\mu^{\sigma(i)} + \frac{C_P}{B_q} - 1)$.

We have following propositions:

If arm $\sigma(1)$ is selected for $s(> \alpha^*)$ steps, then

$$\exp(-2(w^* - sc_{t,s})^2/(s - q)) \leq t^{-4}. \tag{5}$$

, where $\alpha^* = 1 + \lceil \max\{q, [w^*/(\sqrt{L} - \sqrt{2})]^2\} \rceil$

Similarly, if arm $\sigma(i)$ is selected for $s(>\alpha^i)$ steps,

$$\exp(\frac{-2(w^i+sc_{t,s})^2}{s-q}) \leq t^{-4} \quad (6)$$

, where $\alpha^i = 1+\lceil \max\{q, [w^i/(\sqrt{L}-\sqrt{2})]^2\}\rceil$.

Moreover, there exists

$$\gamma = \lceil \max\{(N-1)(4\alpha^*+1)+\alpha^*, (N-1)e^{4\alpha^*/L}+\alpha^*, \max_{2\leq i\leq N}\{(N-1)(4\alpha^i+1)+\alpha^i, (N-1)e^{4\alpha^i/L}+\alpha^i\}\}\rceil$$

such that for the time n, if $G(n) > B_\gamma$, then arm $\sigma(1)$ is selected at least α^* times and arm $\sigma(i)$ is selected at least α^i times. Next we will bound the number of times we fail to choose the optimal arm by logarithmic order.

Denote $T_j(n)$ as the number of times we select arm $\sigma(j)$ up to time n. Then, for any positive integer l, we have

$$\begin{aligned} T_j(n) &= \\ &1+\sum_{t\geq\sum_{i=1}^N B_i, G||t}^{n} \mathbb{I}\{\frac{\hat{X}_{\sigma(1)}(t)}{i_{\sigma(1)}(t)}+c_{t,i_{\sigma(1)}} < \frac{\hat{X}_{\sigma(j)}(t)}{i_{\sigma(j)}(t)}+c_{t,i_j}\} \\ &\leq \sum_{t=B_1+\cdots+B_\gamma, G||t}^{n} \sum_{s_1=\alpha^*}^{\alpha(t),t=B_1+\cdots+B_{\alpha(t)}} \sum_{s_j=\max(\alpha^j,l)}^{\beta(t),t=B_1+\cdots+B_{\beta(t)}} \\ &\mathbb{I}\{\frac{\hat{X}_{\sigma(1),s_1}}{s_1}+c_{t,s_1} \leq \frac{\hat{X}_{\sigma(j),s_j}}{s_j}+c_{t,s_j}\}+l+\gamma \end{aligned}$$

where $\mathbb{I}\{x\}$ is the indicate function; $i_{\sigma(j)}(t)$ is the number of times we select arm $\sigma(j)$ when up to time $t, \forall j=2,\cdots,N$; $\hat{X}_{\sigma(j)}(t)$ is the sum of every sample mean of arm $\sigma(j)$ for $i_{\sigma(j)}(t)$ plays up to time t; $\hat{X}_{\sigma(j),s_j}$ is the sum of every sample mean for s_j times selecting arm $\sigma(j)$.

The condition $\{\frac{\hat{X}_{\sigma(1),s_1}}{s_1}+c_{t,s_1} \leq \frac{\hat{X}_{\sigma(j),s_j}}{s_j}+c_{t,s_j}\}$ implies that at least one of the following must hold:

$$\frac{\hat{X}_{\sigma(1),s_1}}{s_1} \leq \mu^{\sigma(1)} - \frac{C_P}{B_q} - c_{t,s_1} \quad (7)$$

$$\frac{\hat{X}_{\sigma(j),s_j}}{s_j} \geq \mu^{\sigma(j)} + \frac{C_P}{B_q} + \frac{\mu^{\sigma(j)}+C_P/B_q}{\mu^{\sigma(j)}-C_P/B_q}c_{t,s_j} \quad (8)$$

$$\mu^{\sigma(1)} - \frac{C_P}{B_q} < \mu^{\sigma(j)} + \frac{C_P}{B_q} + (1+\frac{\mu^{\sigma(j)}+C_P/B_q}{\mu^{\sigma(j)}-C_P/B_q})c_{t,s_j} \quad (9)$$

Applying Lemma 1 and (5) and (6), we have:

$$\mathbb{P}(\frac{\hat{X}_{\sigma(1),s_1}}{s_1} \leq \mu^{\sigma(1)} - \frac{C_P}{B_q} - c_{t,s_1}) \leq t^{-4}$$

$$\mathbb{P}(\frac{\hat{X}_{\sigma(j),s_j}}{s_j} \geq \mu^{\sigma(j)} + \frac{C_P}{B_q} + \frac{\mu^{\sigma(j)}+C_P/B_q}{\mu^{\sigma(j)}-C_P/B_q}c_{t,s_j}) \leq t^{-4}$$

Denote $\lambda_j(n)$ as

$$\lambda_j(n) = \lceil (L(1+\frac{\mu^{\sigma(j)}+C_P/B_q}{\mu^{\sigma(j)}-C_P/B_q})^2 \ln n)/(\mu^{\sigma(1)}-\mu^{\sigma(j)} - \frac{2C_P}{B_q})^2\rceil$$

For $l \geq \lambda_j(n)$, (9) is false. So we get:

$$\mathbb{E}(T_j(n)) \leq \lambda_j(n)+\gamma+\sum_{t=1}^{\infty}\sum_{s_1=1}^{t}\sum_{s_j=1}^{t} 2t^{-4} \leq \lambda_j(n)+\gamma+\frac{\pi^2}{3}.$$

The first part of the regret is bounded by

$$\sum_{j=2}^{N} \mathbb{E}[T_j(n)](G(n)(\mu^{\sigma(1)}-\mu^{\sigma(j)})+2C_P)$$

and the second part is bounded by $C_P\sum_{j=2}^N \mathbb{E}(T_j(n)$.

Therefore, we have:

$$r^\Phi(n) \leq G(n)+\sum_{j=2}^{N}(G(n)(\mu^{\sigma(1)}-\mu^{\sigma(j)})+3C_P)(\lambda_j(n)+\gamma+\frac{\pi^2}{3})$$

This inequality can be readily translated to the simplified form of the bound given in the statement of Theorem 1, where:

$$\begin{aligned} Z_1 &= \sum_{j=2}^{N}(\mu^{\sigma(1)}-\mu^{\sigma(j)})\lceil \frac{L(1+\frac{\mu^{\sigma(j)}+C_P/B_q}{\mu^{\sigma(j)}-C_P/B_q})^2}{(\mu^{\sigma(1)}-\mu^{\sigma(j)}-\frac{2C_P}{B_q})^2}\rceil \\ Z_2 &= 3C_P\sum_{j=2}^{N}\lceil \frac{L(1+\frac{\mu^{\sigma(j)}+C_P/B_q}{\mu^{\sigma(j)}-C_P/B_q})^2}{(\mu^{\sigma(1)}-\mu^{\sigma(j)}-\frac{2C_P}{B_q})^2}\rceil \\ Z_3 &= (\gamma+\frac{\pi^2}{3})\sum_{j=2}^{N}(\mu^{\sigma(1)}-\mu^{\sigma(j)})+1 \\ Z_4 &= 3(N-1)C_P(\gamma+\frac{\pi^2}{3}) \end{aligned}$$

■

Remark 1: : Algorithm 1 and Theorem 1 can be easily extended to multi-arm case, where K is a known positive integer. A near-logarithmic regret with time is also achieved. Due to space constraints, we omit this part and details are in [10].

C. Corollary

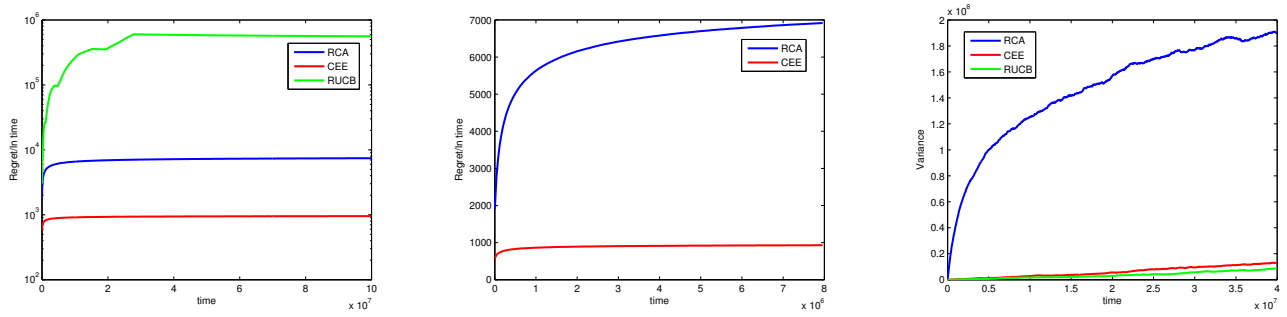
From the analysis above, we see that if sequence $\{B_i\}_{i=1}^{\infty}$ is constant and $B_i \geq \lceil \max\{\frac{2C_P}{\mu^{\sigma(1)}-\mu^{\sigma(2)}}, \frac{C_P}{\mu^{\sigma(l)}}, l=1,2,\cdots,N\}\rceil$, then Algorithm 1 achieves logarithmic regret over time. Specifically, we have the following corollary:

Corollary 1: The system model is the same as that in Theorem 1. In Algorithm 1, if

$$B_i \equiv \lceil \max\{\frac{2C_P}{\mu^{\sigma(1)}-\mu^{\sigma(2)}}, \frac{C_P}{\mu^{\sigma(l)}}, l=1,2,\cdots,N\}\rceil \forall i \in \mathbb{N}$$

the expected regret after n time slots is at most $Z_1'B_1\ln n + Z_2'\ln n + Z_3'B_1 + Z_4'$, where coefficients are obtained by putting $q=1$ in previous derivation.

Remark 2: : This corollary is a special case for Theorem 1, and reveals that we can design an algorithm achieving logarithmic regret over time if certain knowledge of the system is available.



(a) Regret/ln time for RCA, CEE and RUCB (b) Regret/ln time for RCA and CEE (c) Reward variance for RCA, CEE and RUCB

Fig. 1. Regret and variance performance for RCA, CEE and RUCB

V. Numerical Results

In this section, we compare our algorithm with two previously proposed policies in the context of opportunistic spectrum access, RCA [11] and RUCB [5]. We focus on two properties of the algorithms: regret and variance, which show the efficiency and stability of the algorithms respectively. The state of each channel evolves as an irreducible, aperiodic Markov chain. We consider $N = 5$ channels with two states, 0 or 1. At each time slot, the player activates 1 channel. The active and passive transition matrix for each channel are the same, i.e. $P_j = Q_j, 1 \leq j \leq N$. We set the non-decreasing sequence $\{B_i\}_{i=1}^{\infty}$ in Algorithm 1 a constant sequence. The transition probabilities and rewards are shown in table I.

S	p_{01}, p_{10}	r_0, r_1
ch.1	0.3, 0.9	0.1,1
ch.2	0.8, 0.7	0.1,1
ch.3	0.5, 0.1	0.1,1
ch.4	0.2, 0.4	0.1,1
ch.5	0.1, 0.5	0.1,1

TABLE I
Transition Probabilities and Rewards

For fairness, we set parameters for all three algorithms just passing the theoretical bound. We set $L = 415$ in RCA, $L = 3126$ and $D = 171520$ in the RUCB algorithm. In CEE Algorithm 1, we set L to be 2.1 and B_i to be 49.

In Figure 1(a), we present the regret of RCA, CEE and RUCB over 10 runs for 100 million time slots. In Figure 1(b), we show the first 8 million time slots of regret to compare the converging speed between RCA and CEE. In order to access the stability of each algorithm, we also present variances of rewards over 100 runs in Figure 1(c).

It is observed that CEE shows substantially better regret performance than both RCA and RUCB. Besides, $\text{regret}/\ln \text{time}$ converges much more quickly in CEE than in RCA and RUCB. It is reasonable because in CEE, the selection of arm depends on the whole observing history, thus uses data much more efficiently. We also observe RUCB and CEE outperform RCA significantly in stability. The reward variances of RCA are much higher than CEE and RUCB. One possible reason is in RCA, the time interval between two selection is a random variable which reduces stability.

VI. Conclusion

In this paper, we consider the non-Bayesian restless multi-arm bandit problem which is important for opportunistic spectrum access in cognitive radio networks. We adopt a weak notion of regret, defined as the gap of expected reward compared to a genie who always plays the K best arms. We propose an algorithm achieving a near-logarithmic regret over time when no *a prior* information about the system is available. We present another policy to achieve exact logarithmic regret if some bounds pertaining to the stationary state distribution and corresponding rewards are known. Compared with prior work, this algorithm requires the least information. We also present numerical results and analysis showing that CEE significantly outperforms both of the two previously prosed algorithms, in terms of regret and stability.

References

[1] P. Whittle, "Restless Bandits: Activity Allocation in a Changing World," Journal of Applied Probability, Vol. 25, 1988.

[2] C. H. Papadimitriou and J. N. Tsitsiklis, "The Complexity Of Optimal Queueing Network Control," *Mathematics of Operations Research*, Vol. 24, 1994.

[3] V. Anantharam, P. Varaiya, J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part II: Markovian Rewards," *IEEE Transaction on Automatic Control*, Vol. AC-32 ,No.11 ,pp. 977-982, Nov., 1987.

[4] W. Dai, Y. Gai, B. Krishnamachari, Q. Zhao, "The Non-Bayesian Restless Multi-armed Bandit: A Case Of Near-Logarithmic Regret," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May, 2011

[5] H. Liu, K. Liu, and Q. Zhao, "Logarithmic Weak Regret of Non-Bayesian Restless Multi-Armed Bandit," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May, 2011

[6] J. C. Gittins and D. M. Jones, A dynamic allocation index for sequential design of experiments," *Progress in Statistics, Euro. Meet. Statis.*, vol. 1, pp. 241-266, 1972.

[7] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance," *IEEE Transactions on Wireless Communications*, 2008.

[8] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, 1985.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, 47(2-3), 2002.

[10] W. Dai, Y. Gai, B. Krishnamachari, "Efficient Online Learning for Opportunistic Spectrum Access," Arxiv pre-print http://arxiv.org/abs/1109.1552, September 2011

[11] C. Tekin and M. Liu, "Online Learning in Opportunistic Spectrum Access: A Restless Bandit Approach," Arxiv pre-print http://arxiv.org/abs/1010.0056, October 2010.