Online Learning for Multi-Channel Opportunistic Access over Unknown Markovian Channels

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*—A fundamental theoretical problem in opportunistic spectrum access is the following: a single secondary user must choose a channel to sense and access at each time, with the availability of each channel (due to primary user behavior) described by a Markov Chain. The problem of maximizing the expected channel usage can be formulated as a restless multi-armed bandit. We present in this paper an online learning algorithm with the best known results to date for this problem in the case when channels are homogeneous and the channel statistics are unknown *a priori*. Specifically, we show that this policy, that we refer to as CSE, achieves a regret (the gap between the rewards accumulated by a model-aware Genie and the policy) that is bounded in finite time by a function that scales as $\mathcal{O}(\log t)$. By explicitly learning the underlying statistics over time, this novel policy outperforms a previously proposed scheme shown to provide near-logarithmic regret.**

***Index Terms*—Restless Multi-Armed Bandit; Logarithmic Regret; Online Learning**

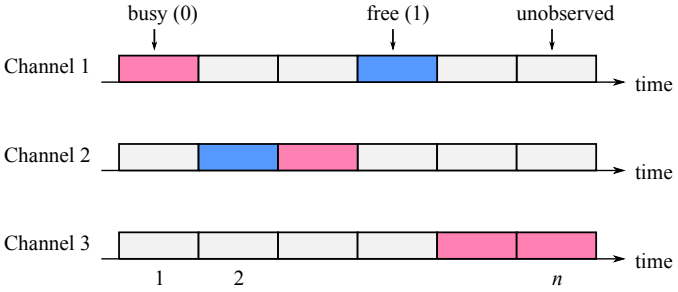


Fig. 1: Sequential selection process: the user selects one channel at a time to sense if it is free. If it is free (blue block), the user may access it for transmissions; otherwise the channel is busy (red block). Only one channel can be sensed each time and other channels are unobserved (grey block).

I. Introduction

One of the fundamental problems in overlay-based opportunistic spectrum access is sequential channel selection by secondary users for sensing and access. In this problem, originally formulated as a Partially Observable Markov Decision Process (POMDP) [1], time is slotted and it is assumed that the secondary user is limited to selecting one channel at a time. It then senses if the channel is free, and if free, may access it for transmissions (see Fig. 1). It is assumed that the primary user behavior on the channels can be modelled as independent two-state Markov Chains (free/busy).

The sequential selection process by the secondary user serves two purposes: on the one hand, it aims to maintain fresh observations of each channel in order to improve its estimates of their condition, and on the other, it tries to select free channels as often as possible to maximize throughput. This problem can be formulated as a restless multi-armed bandit [2], which are highly challenging and known to be computationally intractable in general, specifically they are known to be PSPACE-hard [3].

Bayesian Non-Homogeneous case: If the transition matrices are known, then after each observation any prior distribution of channel availability can be updated to a posterior; for this reason the problem under known transition matrix is referred to as the Bayesian problem. Researchers have also explored the non-homogeneous case (when the transition matrices are different for each channel). For this case, the problem can always be solved using a POMDP solver, however it requires exponential space and time complexity. It has been shown that the famous Whittle's Index [2] can be computed in closed form [4]. This index-based policy, however, is still suboptimal in general, though it can be shown to be asymptotically optimal as the number of channels goes to infinity, under specific regimes (such as scaling the fraction of sampled channels as a constant).

Bayesian Homogeneous case: To gain a deeper theoretical understanding of this problem, researchers have focused on the special case when the channels are homogeneous [5]–[7], i.e. when all channels have identical transition matrices. It was first shown for the case of two homogeneous channels that the tractable Myopic policy (which aims to maximize the immediate rewards at each step) is optimal [7]. Moreover, it was found that the Myopic policy has a simple, semi-universal structure. It suffices to know whether the Markov chains are positively or negatively correlated to implement the Myopic policy; it does not require the full knowledge of the transition matrix beyond this [7]. This has been generalized in subsequent work to the following statement: when the homogeneous channels are positively correlated, then the Myopic policy is optimal for any number of channels. For negatively correlated

This work was supported in part by the U.S. National Science Foundation via grant number AST-1248017.

channels, the Myopic policy is optimal for 2 and 3 channels, but there is a counter-example showing its sub-optimality for 4 channels [5].

Non-Bayesian Non-Homogeneous Case: When the transition matrices are *a priori* unknown, then we have a non-Bayesian formulation, as it is no longer possible to update prior channel distributions to posterior distributions after each observation. In this case, the problem becomes harder as the secondary user must use an online learning policy that periodically explores channels in order the learn their statistics, while continuing to balance exploration and exploitation for keeping fresh observations and getting enough free channel selections to maximize throughput.

It is clear that in the absence of knowledge of the underlying statistics, the secondary user cannot hope to achieve the same performance as a genie that is aware of the underlying statistics (as in the Bayesian case). The only hope is that through learning it is able to achieve the same performance asymptotically. Specifically, the measure for performance of an online-learning policy used is *regret*, defined as the gap in expected reward obtained a genie and the given policy. A sublinear regret is desirable as it implies that the time-averaged regret goes to zero, i.e. asymptotically the policy achieves the same time-averaged reward as the genie under consideration.

Because of the challenging nature of this problem, a number of papers have focused on what has been called *weak regret* [8]–[10], where the genie being compared with is weaker in the sense that it is aware only of the steady state distribution for each channel, and not the full transition matrices. In these papers, the authors have presented policies for which bounds on weak regret can be shown to grow as a logarithmic function of time, i.e. $\mathcal{O}(\log t)$. However, weak regret results are unsatisfactory because the genie is so constrained that it picks the same channel over time, whereas the optimal Bayesian solution would be to switch channels dynamically based on observations. The weaker genie may thus perform quite poorly in practice. In this case, the policies with sublinear weak regret may even potentially outperform the genie, but their performance with respect to the true optimum remains unclear.

Thus far general results on strict regret (comparing the performance of a policy to the genie that knows the probability transition matrices for each channel and can thus perform optimally) have been sparse. Just a few months ago, Ortner *et al.* [11] have provided the first strict-regret result that is applicable to the non-homogeneous non-Bayesian opportunistic spectrum access problem (in fact, their result is aimed more generally at any non-Bayesian restless multi-armed bandit, but they explicitly point out this very problem as a motivating example). They present a policy based on upper-confidence bound methods for reinforcement learning, which is shown to achieve $\mathcal{O}(\sqrt{t})$ regret. Moreover, they show that for general non-Bayesian RMAB, index-based policies are suboptimal.

Non-Bayesian Homogeneous Case: Preceding the recent Ortner *et al.* result, two years ago, we provided the first strict-regret results for case of the non-Bayesian problem with homogeneous channels [12]. Specifically, a meta policy that treats the two structures of the myopic policy for the positive and negatively correlated cases as two arms and learns by exploring and exploiting between these two arms is shown to yield near-logarithmic regret with respect to time whenever the Myopic policy is optimal (this is always true for 2 or 3 homogenous channels, for any number of homogeneous channels when positively correlated, and depending on the transition matrix for more than 3 negatively correlated channels). However, the policy presented in that paper does not attempt to explicitly learn the elements of the transition matrix and therefore there has been room for improvement.

Contribution: We are now in position to describe the nature and contribution of this work:

- We propose an online-learning policy for the non-Bayesian homogeneous channels problem.
- Unlike our prior work on this problem [12], the policy proposed in this paper explicitly estimates (in an iteratively improving fashion) elements of the underlying unknown transition probability matrix.
- We prove that this new policy achieves exactly logarithmic strict regret with respect to time, whenever the Myopic policy is optimal. This is an improvement over the result in [12] which provides near-logarithmic regret.
- Since even the optimal policy must know whether the underlying statistics are positively or negatively correlated, and in the absence of prior knowledge this can only be inferred through statistical observations, we conjecture that the lower bound on regret for this problem is $\Omega(\log t)$. If so, the policy in this paper would be order-optimal, the first such algorithm for this problem.
- Because we restrict our attention to a tractable but important special case of homogeneous channels, we are able to obtain a lower strict-regret of $\mathcal{O}(\log t)$ compared to the $\mathcal{O}(\sqrt{t})$ regret bound for the policy provided by Ortner *et al.* [11][1].

II. System Model

This section introduces the system model for this problem and describes the myopic policy which achieves the optimal performance when the system parameters are known.

A. Two-State RMAB Problem

Consider a time-slotted system with one user and N independent arms. Each arm has two states (0 or 1), evolving as a Markov chain over time. All arms have an identical transition probability matrix $\mathbf{P}$, which is unknown to the user and given by

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

[1]One point bears clarification. The study by Ortner *et al.* [11] explicitly notes that index-based policies are suboptimal for online-learning of restless multi-armed bandits. However the policy proposed in this paper is an index-based policy and is shown to achieve sublinear regret, approaching optimal performance asymptotically over time. The apparent contradiction is resolved by noting that the claim in [11] is for a general restless multi-armed bandit, and clearly does not apply to the special case of homogeneous channels considered here

The arms are assumed to be restless, i.e., the arms' states evolve each time independent of the user's action. At each time slot, the user selects one arm and receives a reward depending on the arm's state. For simplicity, we assume the reward is 1 if the arm is in state 1 and the reward is 0 otherwise. The user can sense only the state of selected arm at each time slot.

The goal is to design an arm selection policy that maximizes the expected total reward over some time horizon. Equivalently, we try to design a policy π that performs well with respect to *regret*, which is the difference between the expected reward obtained by π^* and that obtained π, where π^* is the omniscient optimal policy that knows the transition probability matrix $\mathbf{P}$.

Let $S_i(t) \in \{0,1\}$ denote the state of channel i at time t and $\Omega(t) \triangleq [\omega_1(t), \ldots, \omega_N(t)]$ denote the belief vector, where $\omega_i(t)$ is the probability that $S_i(t) = 1$. The regret obtained by policy π with the initial belief vector $\Omega(1)$ at time n is then given as follows.

$$r^{\pi}(\Omega(1), n) = \mathbb{E}\left\{ \sum_{t=1}^{n} R^{\pi^*}(\Omega(1), t) - \sum_{t=1}^{n} R^{\pi}(\Omega(1), t) \right\}$$

where $R^{\pi^*}(\Omega(1), t)$ and $R^{\pi}(\Omega(1), t)$ denote the reward obtained at time t (provided the initial belief vector is $\Omega(1)$) by applying policy π^* and π, respectively.

B. Myopic Policy

The myopic policy π^{Myopic}, proposed by Zhao *et al* [7], is a solution for the Bayesian RMAB problem. It has a simple structure and depends only on the correlation sign of the transition probability matrix $\mathbf{P}$ (i.e., whether $p_{11} > p_{01}$), therefore can be used as a starting point to design a meta-policy for non-Bayesian restless multi-armed bandit (RMAB) problems.

The structure of the myopic policy uses the concept of *circular order*. For a circular order κ, the starting point is irrelevant, i.e., $\kappa = (n_1, n_2, \cdots, n_N)$ is equivalent to $(n_i, n_{i+1}, \cdots, n_N, n_1, n_2, \cdots, n_{i-1})$. Moreover, let $-\kappa$ denote the reverse circular order of κ. For arm i, let i^+_κ denote its next arm in the circular order κ. With these notations, the structure of the myopic policy is presented as follows.

Let $\kappa(t)$ denote the circular ordering of all arms at time t. The circular order $\kappa(1)$ in time 1 depends on the order of $\Omega(1)$: $\kappa(1) = (n_1, n_2, \cdots, n_N)$ implies that $\omega_{n_1}(1) \le \omega_{n_2}(1) \le \cdots \le \omega_{n_N}(1)$. Let $\hat{a}(t)$ denote the arm selected by the myopic policy in time t. Then $\hat{a}(1) = \arg\max_{i=1,2,\cdots,N} \omega_i(1)$. For $t > 1$, the myopic action $\hat{a}(t)$ is given as follows [7].

- **Policy** π_1 (if $p_{11} > p_{01}$):

$$\hat{a}(t) = \begin{cases} \hat{a}(t-1), & \text{if } S_{\hat{a}(t-1)}(t-1) = 1 \\ \hat{a}(t-1)^+_{\kappa(t)}, & \text{if } S_{\hat{a}(t-1)}(t-1) = 0 \end{cases}$$

  where $\kappa(t) \equiv \kappa(1)$.
- **Policy** π_2 (if $p_{11} \le p_{01}$):

$$\hat{a}(t) = \begin{cases} \hat{a}(t-1), & \text{if } S_{\hat{a}(t-1)}(t-1) = 0 \\ \hat{a}(t-1)^+_{\kappa(t)}, & \text{if } S_{\hat{a}(t-1)}(t-1) = 1 \end{cases}$$

  where $\kappa(t) = \kappa(1)$ when t is odd and $\kappa(t) = -\kappa(1)$ when t is even.

Remark 1: Note that the structure of myopic policy is simple: the only thing the user has to do is to maintain the circular order $\kappa(t)$. Therefore, it requires little computation and memory. Moreover, the myopic policy is optimal in certain scenarios, i.e., $\pi^{\text{Myopic}} = \pi^*$, as shown in the following proposition.

Proposition 1: If $N \le 3$ or $p_{11} \ge p_{01}$, the myopic policy is optimal for all finite time horizon T.

Proof: See [5], [7]. □

Remark 2: Though the myopic policy has many favorable properties, it requires the correlation sign of the system. When the estimation of the correlation sign is incorrect, the myopic policy will result in a large loss of reward. Therefore, it is necessary to design a policy that performs well when no prior parameter knowledge of the system is available.

III. Sensing Policy for Two-State RMAB

This section provides a sensing policy for the non-Bayesian RMAB problem based on the myopic policy.

Time slots are divided into epochs where each epoch contains L slots, in which L can be any positive integer greater than three. During each epoch, the user applies π_1 or π_2 and records the samples of p_{01} and p_{11}; At the end of each epoch, the user makes a decision about which policy (π_1 or π_2) to apply in the next epoch based on previous sample results.

One key question is how to take samples of p_{01} or p_{11} when executing the π_1 and π_2. According to the structure of the myopic policy, one can estimate the system parameter, i.e., p_{11} and p_{01}, by recording the sample mean of p_{01} and p_{11}. Note that when policy π_1 is applied, if the current arm state is 1, the user selects the same arm and can take a sample of p_{11} based on the arm state in the next time slot. Similarly, when policy π_2 is applied, if the current arm state is 1, the user selects the same arm and can take a sample of p_{01} in the next time slot. We next provide an example of taking samples of p_{11}. A system composed of two arms is considered and policy π_1 is applied.

Table I: Example of Sampling p_{11}

Time	1	2	3	4	5	6	7	8	9	$\cdots$
Arm 1	1	0	–	–	–	–	1	1	1	$\cdots$
Arm 2	–	–	1	1	1	0	–	–	–	$\cdots$

Table I shows an example of taking samples of p_{11}. In the table, 1 and 0 represent the arm states and the notation "–" represents the unobserved arm state. In this example, at time 9, the user has six samples of p_{1i} ($i \in \{0,1\}$), where two of them are 0's and four of them are 1's. Hence, the sample mean p_{11} is given by $\widehat{p}_{11} = 4/(2+4) = 2/3$. In this way, the user can obtain samples of p_{11} when applying π_1. Similarly, the user can obtain samples of p_{01} when applying π_2.

Another key question is how to decide the policy (π_1 or π_2) at the end of an epoch. A desirable way is to treat the two polices as arms in a classic non-Bayesian multi-armed bandit problem. In [13], the UCB1 policy is proposed and achieves a logarithmic regret over time. Based on this policy, we provide a decision method with the goal of learning the correlation sign.

Before the details of the sensing policy are provided, some definitions and notations are given as follows. If the current arm state is 1 and the user decides to select the same arm in the next time slot, then the user obtains an *effective sample* of p_{11}. The definition of effective samples for p_{01} is analogous. Let $s_1(t)$ and $s_2(t)$ denote the number of effective samples of p_{11} and p_{01} up to time t, respectively. Moreover, let $\widehat{p}_{11}(s_1(t))$ denote the sample mean of p_{11} with $s_1(t)$ effective samples and $\widehat{p}_{01}(s_2(t))$ denote the sample mean of p_{01} with $s_2(t)$ effective samples. For example, in Table I, when time is 9, $s_1(t) = 6$, $s_2(t) = 0$, $\widehat{p}_{11}(s_1(t)) = 2/3$. Note that if the user applies π_2 for one epoch, $s_2(t)$ can be positive and a meaningful $\widehat{p}_{01}(s_2(t))$ can be obtained. With these notations, the sensing policy is shown in Algorithm 1.

Algorithm 1 Continuous Sampling and Exploitation (CSE)

1: // INITIALIZATION
2: Choose an arbitrary positive integer $L > 3$; $t = 1$;
3: Play policy π_1 until at least one effective sample of p_{11} is obtained, record $s_1(t)$ and $\widehat{p}_{11}(s_1(t))$;
4: Play policy π_2 until at least one effective sample of p_{01} is obtained, record $s_2(t)$ and $\widehat{p}_{01}(s_2(t))$;
5: // MAIN LOOP: TIME ARE DIVIDED INTO EPOCHS
6: **while** 1 **do**
7: At the end of the epoch, time t, decide to apply π_1 or π_2 as follows:

$$\widehat{p}_{11}(s_1(t)) + \sqrt{\frac{2\ln t}{s_1(t)}} \underset{\pi_2}{\overset{\pi_1}{\gtrless}} \widehat{p}_{01}(s_2(t)) + \sqrt{\frac{2\ln t}{s_2(t)}} \quad (1)$$

8: Play π_1 or π_2 based on (1) for L time slots (one epoch);
9: Update $s_1(t)$ or $s_2(t)$ and update the sample mean of $\widehat{p}_{11}(s_1(t))$ or $\widehat{p}_{01}(s_2(t))$;
10: $t \leftarrow t + L$;
11: **end while**

Remark 3: The policy CSE requires no knowledge of the system parameters, which is favorable in practice. Moreover, it achieves a logarithmic regret uniformly over time t, as shown in the next section. The performance of CSE is robust to the choice of L, which is validated in Section V.

IV. Regret Analysis

This section provides an upper bound of the regret that grows logarithmically with time. The main theorem is as follows.

Theorem 1: If $N \leq 3$ or $p_{11} \geq p_{01}$, the regret achieved by CSE in Algorithm 1 at time n is upper bounded by $C_1 \ln n + C_0$, where C_0 and C_1 are constants only dependent on system parameters and k.

Before showing the proof of Theorem 1, we first provide one fact and two lemmas.

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

$$\mathbb{P}\{S_n \geq n\mu + a\} \leq e^{-2a^2/n}$$
$$\mathbb{P}\{S_n \leq n\mu - a\} \leq e^{-2a^2/n}$$

Proof: See [14]. □

Lemma 1: Let $X_1, \cdots, X_n$ be random variables with range $[0, 1]$ and such that $|\mathbb{E}[X_t | X_1, X_2, \cdots, X_{t-1}] - \mu| \leq \epsilon, \forall t$, where ϵ is a constant number such that $0 < \epsilon < \mu$. Let $S_n = X_1 + \cdots + X_n$. Then for all $a \geq 0$,

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

Proof: See [15]. □

Remark 4: Lemma 1 is a generalization of the Chernoff-Hoeffding bound, which allows for bounded differences between the conditional expectations of a sequence of random variables that are revealed sequentially.

The second lemma explores the deviation of the rewards from the steady throughput by the myopic policy. Let $U_i(\Omega(1))$ denote the steady throughput achieved by policy π_i with the initial belief vector $\Omega(1)$, given by

$$U_i(\Omega(1)) \triangleq \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left\{\sum_{t=1}^{T} R^{\pi_i}(\Omega(1), t)\right\}.$$

In [7], it shows that the limit above exists and is independent of the initial belief vector $\Omega(1)$. Therefore, we can rewrite $U_i(\Omega(1))$ as U_i to denote the average expected reward using policy $\pi_i (i = 1, 2)$. Lemma 2 is then provide as follows.

Lemma 2: For any initial belief vector $\Omega(1)$ and any positive integer M,

$$|\mathbb{E}\left\{\sum_{t=1}^{M} R^{\pi_i}(\Omega(1), t)\right\} - M \cdot U_i| < \epsilon_i, \quad i = 1, 2$$

where ϵ_i is a constant depending on $\mathbf{P}$.

Proof: See [15]. □

Remark 5: Lemma 2 states that the expected loss of reward for policy π_i (starting with an arbitrary initial belief vector) compared to the steady throughput U_i is bounded by a constant ϵ_i depending only on the transition probability matrix $\mathbf{P}$.

Now we provide the proof of Theorem 1 given the fact and two lemmas.

Proof: Throughout the proof, policy π refers to the sensing policy CSE. Let $T_i(t)$ denote the number of epochs that using policy π_i up to time t. For simplicity, we prove Theorem 1 for the case $p_{11} \geq p_{01}$ in this paper where the case $p_{11} < p_{01}$ can be proved analogously. In the case $p_{11} \geq p_{01}$,

policy π_1 is the optimal policy, i.e., $\pi_1 = \pi^*$, implying $U_1 > U_2$.

Let $\tilde{r}^{\pi}(\Omega(1), n)$ denote the *equivalent regret*, given by

$$\tilde{r}^{\pi}(\Omega(1), n) := n \cdot U_1 - \mathbb{E}\left\{ \sum_{t=1}^{n} R^{\pi}(\Omega(1), t) \right\}$$

Lemma 2 implies that there exist constant $\epsilon_1 > 0$ such that

$$\mathbb{E}\left\{ \sum_{t=1}^{n} R^{\pi^*}(\Omega(1), t) \right\} < n \cdot U_1 + \epsilon_1.$$

Therefore, to prove Theorem 1, it suffices to show that there exist constants C_1, C_0 such that $\forall n \in \mathbb{N}$,

$$\tilde{r}^{\pi}(\Omega(1), n) \leq C_1 \ln n + C_0 - \epsilon_1$$

Note that the reward loss can be divided into two parts as follows.

$$\begin{aligned} & n \cdot U_1 - \sum_{t=1}^{n} R^{\pi}(\Omega(1), t) \\ = & \underbrace{\sum_{t:\pi_1 \text{ is used}} (U_1 - R^{\pi}(\Omega(1), t))}_{r_1} + \underbrace{\sum_{t:\pi_2 \text{ is used}} (U_1 - R^{\pi}(\Omega(1), t))}_{r_2} \end{aligned}$$

According to Lemma 2, the expected reward loss is at most ϵ_1 each time π_1 is switched to π_2, hence $\mathbb{E}[r_1] \leq \epsilon_1 \cdot \mathbb{E}[T_2(n)]$. On the other hand, let t_{INI} denote the last time slot of initialization process, then $\mathbb{E}[r_2]$ can be bounded as follows.

$$\begin{aligned} \mathbb{E}\{r_2\} &= \mathbb{E}\left\{ \sum_{t:\pi_2 \text{ is used}} (U_1 - R^{\pi}(\Omega(1), t)) \right\} \\ &\leq \mathbb{E}\left\{ \sum_{t:\pi_2 \text{ is used}} (U_2 - R^{\pi}(\Omega(1), t)) \right\} \\ &\quad + (U_1 - U_2) \cdot (L \cdot \mathbb{E}\{T_2(n) + 1\} + \mathbb{E}\{t_{\text{INI}}\}) \\ &\leq \epsilon_2 \cdot \mathbb{E}\{T_2(n)\} + (U_1 - U_2) \cdot \mathbb{E}\{L \cdot T_2(n) + L + t_{\text{INI}}\} \end{aligned}$$

where the first inequality is due to the definition of $T_2(\cdot)$ and the second inequality is due to Lemma 2.

Therefore, in order to prove Theorem 1, we only have to show that $\mathbb{E}\{t_{\text{INI}}\} \sim \mathcal{O}(\ln n)$ and $\mathbb{E}\{T_2(n)\} \sim \mathcal{O}(\ln n)$. We first bound the outage probability of t_{INI} as follows.

Lemma 3: Let $\alpha = \min\{p_{01}, p_{11}, p_{10}, p_{00}\}$, then we have

$$\mathbb{P}\{t_{\text{INI}} \geq 2 \ln n / \ln(1/\alpha)\} \leq 1/n.$$

Proof: If $t_{\text{INI}} \geq 2 \ln n / \ln(1/\alpha)$, then either π_1 or π_2 has been played for at least $\ln n / \ln(1/\alpha)$ time slots without one effective sample. Either event has a probability no greater than $\alpha^{\ln n / \ln(1/\alpha)} = 1/n$. □

With Lemma 3, $\mathbb{E}\{t_{\text{INI}}\}$ can be bounded as follows

$$\begin{aligned} \mathbb{E}\{t_{\text{INI}}\} &= \int_0^{\infty} \mathbb{P}\{t_{\text{INI}} > t\}\, dt = \sum_{t=0}^{\infty} \mathbb{P}\{t_{\text{INI}} > t\} \\ &\leq \sum_{t=0}^{\infty} e \cdot (\sqrt{\alpha})^t = \frac{e}{1 - \sqrt{\alpha}} \end{aligned}$$

where the first equality is another representation of the expected value; the second equality is due to the fact that t_{INI} is an integer and the inequality is due to Lemma 3.

$\mathbb{E}\{T_2(n)\}$ is bounded as follows.

$$\begin{aligned} & \mathbb{E}\{T_2(n)\} \\ &= \mathbb{E}\left\{ T_2(n) | t_{\text{INI}} \geq \frac{2 \ln n}{\ln(1/\alpha)} \right\} \cdot \mathbb{P}\left\{ t_{\text{INI}} \geq \frac{2 \ln n}{\ln(1/\alpha)} \right\} \\ &\quad + \mathbb{E}\left\{ T_2(n) | t_{\text{INI}} \leq \frac{2 \ln n}{\ln(1/\alpha)} \right\} \cdot \mathbb{P}\left\{ t_{\text{INI}} \leq \frac{2 \ln n}{\ln(1/\alpha)} \right\} \\ &\leq n \cdot 1/n + \mathbb{E}\left\{ T_2(n) | t_{\text{INI}} \leq \frac{2 \ln n}{\ln(1/\alpha)} \right\} \end{aligned}$$

We then bound $\mathbb{E}\{T_2(n) | t_{\text{INI}} \leq 2 \ln n / \ln(1/\alpha)\}$. Let $\mathcal{N} := \{n : n \text{ is a time slot at the end of an epoch}\}$. We have

$$\begin{aligned} & \mathbb{E}\{ T_2(n) | t_{\text{INI}} \leq 2 \ln n / \ln(1/\alpha) \} \\ &\leq 2 \ln n / \ln(1/\alpha) + \sum_{t \in \mathcal{N}, t_{\text{INI}} \leq t \leq n} \mathbb{I}\Bigg\{ \widehat{p}_{11}(s_1(t)) + \sqrt{\frac{2 \ln t}{s_1(t)}} \\ &\quad \leq \widehat{p}_{01}(s_2(t)) + \sqrt{\frac{2 \ln t}{s_2(t)}} \Bigg\} \\ &\leq 2 \ln n / \ln(1/\alpha) + l + \sum_{t \in \mathcal{N}, t_{\text{INI}} \leq t \leq n} \mathbb{I}\Bigg\{ \widehat{p}_{11}(s_1(t)) + \sqrt{\frac{2 \ln t}{s_1(t)}} \\ &\quad \leq \widehat{p}_{01}(s_2(t)) + \sqrt{\frac{2 \ln t}{s_2(t)}}, T_2(t) \geq l \Bigg\} \\ &\leq 2 \ln n / \ln(1/\alpha) + l + A + B \end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function and l can be any positive integer, A and B are given by

$$\begin{aligned} A = \sum_{t \in \mathcal{N}, t_{\text{INI}} \leq t \leq n} & \mathbb{I}\Bigg\{ \widehat{p}_{11}(s_1(t)) + \sqrt{\frac{2 \ln t}{s_1(t)}} \leq \widehat{p}_{01}(s_2(t)) \\ & + \sqrt{\frac{2 \ln t}{s_2(t)}}, T_2(t) \geq l, s_2(t) \geq \gamma \cdot l \Bigg\} \\ B = \sum_{t \in \mathcal{N}, t_{\text{INI}} \leq t \leq n} & \mathbb{I}\Bigg\{ \widehat{p}_{11}(s_1(t)) + \sqrt{\frac{2 \ln t}{s_1(t)}} \leq \widehat{p}_{01}(s_2(t)) \\ & + \sqrt{\frac{2 \ln t}{s_2(t)}}, T_2(t) \geq l, s_2(t) < \gamma \cdot l \Bigg\} \end{aligned}$$

in which γ can be any real number in $(0, 1)$. We next bound $\mathbb{E}\{A\}$ and $\mathbb{E}\{B\}$, respectively.

To bound $\mathbb{E}\{A\}$, note that if

$$\mathbb{I}\left\{ \widehat{p}_{11}(s_1(t)) + \sqrt{\frac{2 \ln t}{s_1(t)}} \leq \widehat{p}_{01}(s_2(t)) + \sqrt{\frac{2 \ln t}{s_2(t)}} \right\} = 1$$

then at least one of the following events must hold.

$$\widehat{p}_{11}(s_1(t)) \leq p_{11} - \sqrt{\frac{2\ln t}{s_1(t)}} \tag{2}$$

$$\widehat{p}_{01}(s_2(t)) \geq p_{01} + \sqrt{\frac{2\ln t}{s_2(t)}} \tag{3}$$

$$p_{11} < p_{01} + 2\cdot\sqrt{\frac{2\ln t}{s_2(t)}} \tag{4}$$

Fact 1 implies

$$\mathbb{P}\left\{\widehat{p}_{11}(s_1(t)) \leq p_{11} - \sqrt{\frac{2\ln t}{s_1(t)}}\right\} \leq e^{-4\ln t} = t^{-4}$$

$$\mathbb{P}\left\{\widehat{p}_{01}(s_2(t)) \geq p_{01} + \sqrt{\frac{2\ln t}{s_2(t)}}\right\} \leq e^{-4\ln t} = t^{-4}$$

For $s_2(t) \geq \gamma\cdot l$ and $l \geq \lceil 8\ln n/(\gamma(p_{11}-p_{01})^2)\rceil$,

$$\begin{aligned} p_{11} - p_{01} - 2\cdot\sqrt{\frac{2\ln t}{s_2(t)}} &\geq p_{11} - p_{01} - 2\cdot\sqrt{\frac{2\ln n}{s_2(t)}} \\ &\geq p_{11} - p_{01} - 2\cdot\sqrt{\frac{2\ln n}{\gamma\cdot l}} \geq 0 \end{aligned}$$

Therefore, (4) is false. Hence, if $l \geq \lceil 8\ln n/(\gamma(p_{11}-p_{01})^2)\rceil$, $\mathbb{E}\{A\}$ can be bounded by

$$\begin{aligned} \mathbb{E}\{A\} \leq& \sum_{t=1}^{n}\sum_{s_1=1}^{t}\sum_{s_2=1}^{t}\left(\mathbb{P}\left\{\widehat{p}_{11}(s_1) \leq p_{11} - \sqrt{\frac{2\ln t}{s_1}}\right\}\right. \\ &\left. + \mathbb{P}\left\{\widehat{p}_{01}(s_2) \geq p_{01} + \sqrt{\frac{2\ln t}{s_2}}\right\}\right) \\ \leq& \sum_{t=1}^{n}\sum_{s_1=1}^{t}\sum_{s_2=1}^{t} 2\cdot t^{-4} \leq 2\cdot\sum_{t=1}^{\infty} t^{-2} \leq \pi^2/3 \end{aligned}$$

Next we bound $\mathbb{E}\{B\}$. Note that

$$B \leq \sum_{t=1}^{n}\sum_{s_1=1}^{t}\mathbb{I}\left\{T_2(t) \geq l, s_2(t) < \gamma\cdot l\right\}$$

When the user applies policy π_2, the probability of obtaining an effective sample of p_{01} between $t+1$ and $t+2$ is equal to the probability the arm state at $t+1$ is 0, which is no less than α (recall that $\alpha = \min\{p_{01}, p_{11}, p_{10}, p_{00}\}$). Hence, during an epoch that policy π_2 is applied, the user obtains at least one effective sample of p_{01} during that epoch with probability no less than α, regardless of the arm state at the beginning of the epoch. Let $N_1, N_2, \cdots, N_{T_2(t)}$ denote the number of effective samples of p_{01} obtained in the epochs that applies policy π_2, then $\mathbb{E}\{N_t|N_1, N_2, \cdots, N_{t-1}\} \geq \mathbb{P}\{N_t \geq 1|N_1, N_2, \cdots, N_{t-1}\} \geq \alpha$, $\forall t$. Therefore, for $\gamma = \alpha/2$ and $l \geq (4/\alpha^2)\ln n$, we have

$$\begin{aligned} &\mathbb{P}\left\{T_2(t) \geq l, s_2(t) < \gamma\cdot l\right\} \\ &\leq \mathbb{P}\left\{T_2(t) \geq l, s_2(t) < \gamma\cdot T_2(t)\right\} \\ &\leq \mathbb{P}\left\{\sum_{i=1}^{T_2(t)} N_i < \alpha\cdot T_2(t) - \alpha\cdot T_2(t)/2 | T_2(t) \geq l\right\} \\ &\leq \exp\left\{-\frac{\alpha^2}{2}\cdot l\right\} \leq \frac{1}{n^2} \end{aligned}$$

where the first inequality is due to $T_2(t) \geq l$, the second inequality is because the conditional probability is no smaller than the joint probability and $s_2(t) = N_1 + N_2 + \cdots + N_{T_2(t)}$, the third inequality is due to Lemma 1 and the last inequality is because of the choice of γ and l. Consequently, if $l \geq (4/\alpha^2)\ln n$, we have

$$\begin{aligned} \mathbb{E}\{B\} &\leq \sum_{t=1}^{n}\sum_{s_1=1}^{t}\mathbb{P}\left\{T_2(t) \geq l, s_2(t) < \gamma\cdot l\right\} \\ &\leq \sum_{t=1}^{n}\sum_{s_1=1}^{t}\frac{1}{n^2} = \frac{n(n+1)}{2n^2} < 1 \end{aligned}$$

Finally, with the results above, we can bound the regret $r^{\pi}(\Omega(1), n)$ as follows.

$$r^{\pi}(\Omega(1), n) \leq C_1\ln n + C_0$$

where

$$\begin{aligned} C_0 =& \frac{e|U_1 - U_2|}{1-\sqrt{\alpha}} + (\epsilon_1 + \epsilon_2 + L|U_1 - U_2|)\left(3 + \frac{\pi^2}{3}\right) \\ &+ L|U_1 - U_2| + \max\{\epsilon_1, \epsilon_2\} \\ C_1 =& (\epsilon_1 + \epsilon_2 + L|U_1 - U_2|)\left(\max\left\{\frac{16}{\alpha(p_{11}-p_{01})^2}, \frac{4}{\alpha^2}\right\}\right. \\ &\left. + \frac{2}{\ln(1/\alpha)}\right) \end{aligned}$$

The proof for the case $p_{11} \leq p_{01}$ is analogous and is omitted here. □

Remark 6: Theorem 1 is stated for the cases where the myopic policy is proved to be optimal. In fact, our proof shows an even stronger result: policy CSE achieves the claimed logarithm regret with respect to the myopic policy for any time n. For those cases where the myopic policy is optimal, policy CSE achieves a logarithmic regret with respect to the optimal policy. It is conjectured that the logarithmic regret is the best achievable order for this problem. If the conjecture is true, policy CSE also obtains the lower bound with respect to the order of time.

V. Simulation Results

This section validates the efficiency and the robustness of the proposed algorithm through simulation results. Consider an opportunistic spectrum access problem, shown in Section II-A. The simulations account for two settings, where the correlation sign of $\mathbf{P}$ is positive and negative, respectively. For the former,

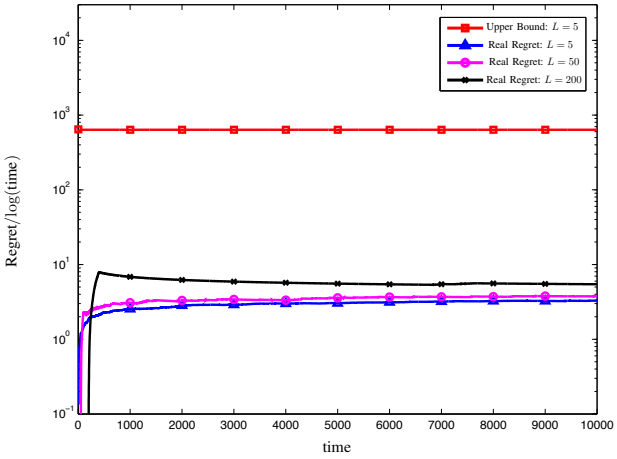
we set the transition probability $p_{11} = 0.8$ and $p_{01} = 0.3$, thus the optimal policy $\pi^* = \pi_1$; for the latter, we set the transition probability $p_{11} = 0.3$ and $p_{01} = 0.8$, thus the optimal policy $\pi^* = \pi_2$.

Fig. 2 compares the regret achieved by CSE with different values of L: $L = 5$, $L = 50$ and $L = 200$. The theoretical bound for $L = 5$ is also computed for comparison. The first observation is that the regret achieved by CSE is not beyond the theoretical bound (for $L = 5$), which agrees with the proof in Section IV. Moreover, the real regret is far less than the theoretical bound, implying there is potential to reduce the constant C_1 in the bound. Though the regret shows a slightly increasing tendency with L, the performance differences among different L's are not significant, showing that CSE is robust to the choice of L. Hence when designing the algorithm in practice, we can choose any integer $L(> 3)$ as the length of the epochs and the performance will not change too much. Another observation is that the proposed algorithm converges quickly to the limit. This property is also favorable in practice, since policy CSE performs well with respect to the regret even if it is played for a short period of time. Note that the observations above stand for both cases when $p_{11} > p_{01}$ and $p_{11} < p_{01}$.

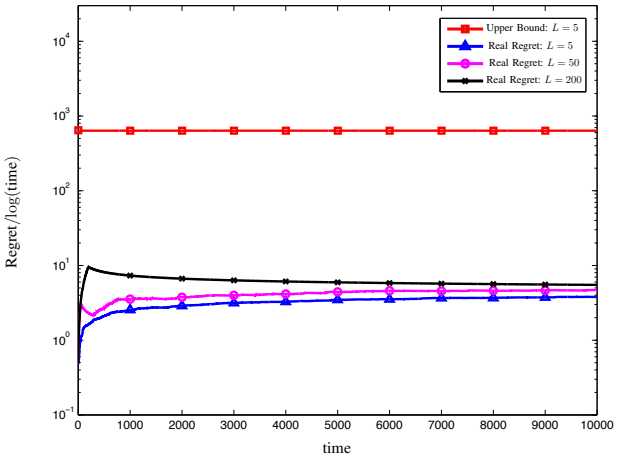
Fig. 3 compares CSE (with fixed $L = 5$) to a previously proposed policy [15], denoted as π_{P}. The design of π_{P} requires a predetermined sequence a_n that monotonically grows to infinity. The regret of π_{P} is on the order of $\mathcal{O}(a_n \log n)$. Consider three choice of a_n: $a_n = 20 + \lceil \log n \rceil^2$, $a_n = 20 + \lceil \sqrt{n} \rceil$ and $a_n = 20 + \lceil \log n \rceil$. The first observation is that the regret of CSE is significantly reduced compared to that of π_{P}. The term $[\text{regret}/\log(\text{time})]$ shows a divergent tendency for π_{P} but converges quickly for CSE. This observation agrees with intuition since the upper bound of π_{P} has an additional term a_n, and therefore grows more quickly than that of policy CSE. The underlying reason is that in policy CSE, p_{01} and p_{11} are estimated based on the samples and these estimated parameter are used for decision making directly. However in π_{P}, the user samples the average reward of π_1 and π_2 (i.e., U_1 and U_2); the time slots are also divided into epochs, but the length of the epoch has to diverge in order to ensure that the sample value of U_1 and U_2 are accurate enough. Policy CSE has another advantage that it does not need a predetermined sequence a_n. For users applying π_{P}, the user is required to make a tradeoff between short-term and long-term performance: if a_n grows fast, π_{P} would perform poorly when played for a long time; on the other hand, if a_n grows too slowly, π_{P} would converge very slowly and have a relatively large regret at the beginning of the operation. This nuisance of selecting a_n is avoided when applying CSE. As mentioned previously, the observation and discussion above are valid for both cases when $p_{11} > p_{01}$ and $p_{11} < p_{01}$.

VI. Conclusion

In this study, we have investigated a non-Bayesian RMAB problem, arising in the context of opportunistic spectrum access. Specifically, in this problem, there are N arms each of



(a)

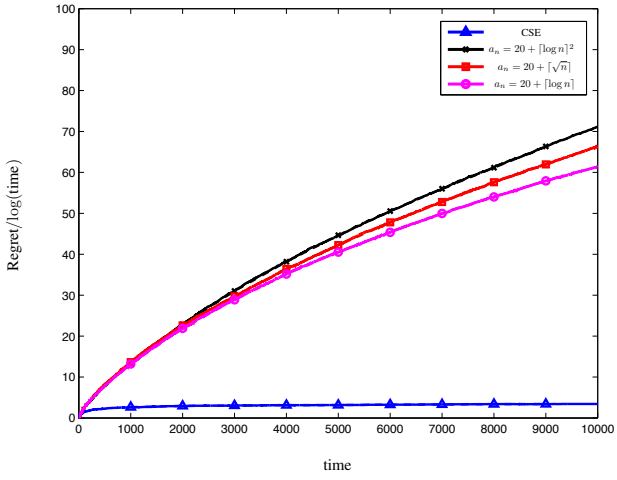


(b)

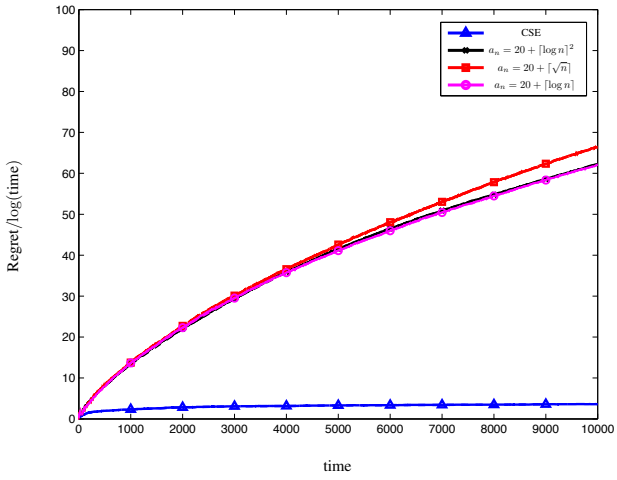
Fig. 2: Regret of policy CSE for different L's. The theoretical bound for $L = 5$ is also plotted for comparison. In (a), the channel is positively related whereas in (b), the channel is negatively related.

which is described by an identical, and independent two-state Markov Chain. For this problem we have shown a policy that yields regret compared to the optimal model-aware genie that is bounded by a logarithmic function of time, whenever the Myopic policy is optimal. This is a significant result because for homogeneous settings the myopic policy is known to be optimal for a wide range of cases: always for 2 and 3 channels, always if the chain is positively correlated, and depending on the transition matrix if it is negatively correlated. The presented policy and its analysis improves over prior results in the literature that proved log weak-regret [8]–[10], and most pertinently our own prior results yielding near-log strict regret [12]. It also improves over the $\mathcal{O}(\sqrt{t})$ regret for the policy recently presented in [11], which, however, applies more generally to other problems as well.

It is conjectured that the logarithmic regret is the best



(a)



(b)

Fig. 3: Comparison between the regret of policy CSE and π_{P}. For π_{P}, three different choices of a_n are considered. In (a), the channel is positively related whereas in (b), the channel is negatively related.

achievable order for the homogeneous non-Bayesian problem. If this is true, our algorithm achieves the lower bound with respect to the order time. Simulation results validate the efficiency and robustness of the proposed algorithm.

Our results not only provide a practical policy for a practical RMAB problem with two states and identical arms, but also suggest a promising approach to investigate other non-Bayesian RMAB problems. It would be of interest to identify other RMAB problems that have a similar structure in terms of the optimal solution structure depending upon a finite number of inequalities with respect to the underlying transitions, and derive similar results accordingly.

REFERENCES

[1] Q. Zhao, L. Tong, A. Swami, and Y. Chen, Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework, IEEE Journal on Selected Areas in Communications: Special Issue on Adaptive, Spectrum Agile and Cognitive Wireles Networks, April 2007

[2] P. Whittle, "Restless bandits: activity allocation in a changing world," *Journal of Applied Probability*, vol. 25A, pp. 287-298, 1988.

[3] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control", *Mathematics of Operations Research*, vol. 24, no. 2, 1999.

[4] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access", *IEEE Transactions on Information Theory*, vol. 56, no. 11, 2010.

[5] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access", *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4040-4050, 2009.

[6] T. Javidi, B. Krishnamachari, Q. Zhao, and M. Liu, "Optimality of Myopic Sensing in Multi-Channel Opportunistic Access", *IEEE International Conference on Communications (ICC)*, May, 2008.

[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*, vol. 7, no. 12, pp. 5431-5440, 2008.

[8] Wenhan Dai, Yi Gai and Bhaskar Krishnamachari, "Efficient Online Learning for Opportunistic Spectrum Access," IEEE INFOCOM Mini-conference, Orlando, FL, USA, March, 2012.

[9] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: a restless bandit approach," *IEEE INFOCOM*, April, 2011.

[10] H. Liu, K. Liu and Q. Zhao, "Logrithmic weak regret of non-bayesian restless multi-armed bandit,"*IEEE ICASSP*, May, 2011.

[11] R. Ortner, D. Ryabko, P. Auer, and R. Munos, "Regret Bounds for Restless Markov Bandits," *the 23th International Conference on Algorithmic Learning Theory (ALT)*, October, 2012.

[12] W. Dai, Y. Gai, B. Krishnamachari and Q. Zhao, "The non-bayesian restless multi-armed bandit: a case of near-logarithmic regret," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May, 2011.

[13] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235-256, 2002.

[14] D. Pollard, *Convergence of Stochastic Processes*. New York, Berlin, Heidelberg, Tokyo: Springer-Verlag, 1984.

[15] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao, "The non-bayesian restless multi-armed bandit: A case of near-logarithmic strict regret," Sep. 2011. [Online]. Available: http://arxiv.org/pdf/1109.1533