# Approximately Optimal Adaptive Learning in Opportunistic Spectrum Access

Cem Tekin, Mingyan Liu
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, Michigan, 48109-2122
Email: {cmtkn, mingyan}@umich.edu

***Abstract*—In this paper we develop an adaptive learning algorithm which is approximately optimal for an opportunistic spectrum access (OSA) problem with polynomial complexity. In this OSA problem each channel is modeled as a two state discrete time Markov chain with a bad state which yields no reward and a good state which yields reward. This is known as the Gilbert-Elliot channel model and represents variations in the channel condition due to fading, primary user activity, etc. There is a user who can transmit on one channel at a time, and whose goal is to maximize its throughput. Without knowing the transition probabilities and only observing the state of the channel currently selected, the user faces a partially observed Markov decision problem (POMDP) with unknown transition structure. In general, learning the optimal policy in this setting is intractable. We propose a computationally efficient learning algorithm which is approximately optimal for the infinite horizon average reward criterion.**

***Index Terms*—Approximate optimality, online learning, opportunistic spectrum access, restless bandits.**

## I. Introduction

We consider the following opportunistic spectrum access (OSA) problem: There is a set of $m$ channels indexed by $1, 2, \ldots, m$, each modeled as a two-state Markov chain (e.g., the Gilbert-Elliot channel model), with a bad state $b$ that yields no reward and a good state $g$ that yields some reward $r_k > 0$ for channel $k$. The state process of each channel follows a discrete time Markov rule independent of other channels. There is a user whose goal is to maximize its long term average throughput by opportunistically selecting a channel to transmit on at each time step $t = 1, 2, \ldots$. Initially, the user does not know the transition probabilities of the channels and it can only partially observe the system, i.e., at any time $t$ it only knows the state of the channel selected at $t$, but not the states of other channels which continue to evolve. Thus, the user faces a tradeoff between exploration and exploitation. By exploring, the user aims to decrease the uncertainty about the state of the system and the unknown transition probabilities, whereas by exploiting the user aims to maximize its reward. In order to achieve this goal, we would like to develop an adaptive learning algorithm which carefully balances exploration and exploitation. This adaptive learning algorithm should be *admissible*, *computable* and *approximately optimal*. Admissible means that the decision at $t$ should be based on all past decisions and observations and nothing more than the user knows up to this time. Computable means that the number of mathematical operations needed to make the decision at any $t$ should be a polynomial in the number of channels. Approximately optimal means that the infinite horizon average reward of the adaptive learning algorithm should not be worse than a constant factor of the infinite horizon average reward of the optimal policy given the transition probabilities of the channels. Under the assumptions in this paper, learning the optimal policy requires exponential complexity in the number of channels, while we show that approximate optimality can be guaranteed with linear complexity in the number of channels.

When the rewards and transition probabilities of the channels are known by the user, the optimal policy can be found by dynamic programming, and the problem becomes a special case of the restless bandit problem which is known to be intractable in general [1]. However, heuristic, approximately optimal and optimal policies for special cases have been considered by [2], [3], [4] and others. In particular, Guha et. al. [3] proposed the first provably approximately optimal, polynomial complexity policy for the problem outlined above with known channel transition probabilities. The adaptive learning algorithm we develop in this paper is based on a threshold variant ($\epsilon_1$-threshold policy) of Guha's policy which is also approximately optimal.

Specifically, we show that when each channel is ergodic, and given that the user knows that probability of transition from the good state to the bad state is lower bounded by some $\delta$ for all channels, the adaptive learning algorithm based on the $\epsilon_1$-threshold policy achieves the same infinite horizon average reward as the $\epsilon_1$-threshold policy. Moreover, we show that for any finite horizon $N$, the difference between the undiscounted total rewards of our learning algorithm and the $\epsilon_1$-threshold policy is on the order of $\log N$. Since our OSA problem is equivalent to a restless bandit problem we will use the terms channel/arm, and selecting a channel/playing an arm interchangeably.

To summarize, the main contributions of this paper are (1) a threshold variant of Guha's policy (the $\epsilon_1$-threshold policy) which we show to be approximately optimal and computationally simple, and (2) an adaptive learning algorithm based on the $\epsilon_1$-threshold policy which we show to achieve the

This work is supported by NSF grant CIF-0910765 and ARO grant W911NF-11-1-0532.

same infinite horizon average reward as the $\epsilon_1$-threshold policy and logarithmic regret uniform in time in its total reward with respect to the $\epsilon_1$-threshold policy.

The remainder of this paper is organized as follows. Section II presents related work. In Section III we give the problem formulation and preliminaries. We explain Guha's policy in Section IV, and present the threshold variant of Guha's policy in Section V. In Section VI we present the adaptive learning algorithm based on the $\epsilon_1$-threshold policy, and analyze its number of deviations from the $\epsilon_1$-threshold policy in Section VII. Based on this, we derive the infinite horizon average reward of the adaptive learning algorithm, and compare its performance with the $\epsilon_1$-threshold policy for finite time in Section VIII. Discussion and conclusion are given in Sections IX and X respectively.

## II. Related Work

Work in optimal adaptive learning dates back to [5] under a Bayesian setting. Lai and Robbins [6] considered the problem where asymptotically optimal adaptive policies for the multi-armed bandit problem with i.i.d. reward process for each arm were constructed. These are index policies and it is shown that they achieve the optimal regret both in terms of the constant and the order. Later Agrawal [7] considered the i.i.d. problem and provided sample mean based index policies which are easier to compute, order optimal but not optimal in terms of the constant in general. Anantharam et. al. [8], [9] proposed asymptotically optimal policies with multiple plays at each time for i.i.d. and Markovian arms respectively. However, all the above work assumed parametrized distributions for the reward process of the arms. Auer et. al. [10] considered the i.i.d. multi-armed bandit problem and proposed sample mean based index policies with logarithmic regret when reward processes have a bounded support. Their upper bound holds uniformly over time rather than asymptotically but these bounds are not asymptotically optimal. Following this approach Tekin and Liu [11], [12] provided policies with uniformly logarithmic regret bounds with respect to the best single arm policy for restless and rested multi-armed bandit problems and extended the results to multiple plays [13]. Decentralized multi-player versions of the i.i.d. multi-armed bandit problem under different collision models were considered in [14], [15], [16]. Other research on adaptive learning focused on Markov Decision Processes (MDP) with finite state and action spaces. Burnetas and Katehakis [17] proposed index policies with asymptotic logarithmic regret, where the indices are the inflations of right-hand side of the estimated average reward optimality equations based on Kullback Leibler (KL) divergence, and showed that these are asymptotically optimal both in terms of the order and the constant. However, they assumed that the support of the transition probabilities are known. Tewari and Bartlett [18] proposed a learning algorithm that uses $l_1$ distance instead of KL divergence with the same order of regret but a larger constant. Their proof is simpler than the proof in [17] and does not require the support of the transition probabilities to be known. Auer and Ortner proposed another algorithm with logarithmic regret and reduced computation for the MDP problem, which solves the average reward optimality equations only when a confidence interval is halved. In all the above work the MDPs are assumed to be irreducible. Based on the work on MDP, under some assumptions on the transition probabilities and structure of the optimal policy for the infinite horizon average reward problem, [19] proposed a learning algorithm for the restless bandit problem, a special case of the POMDP problem, with logarithmic regret uniformly over time with respect to the optimal undiscounted finite horizon policy given the transition probability matrices.

## III. Problem Formulation and Preliminaries

Let $\mathbb{Z}_+$ denote the set of non-negative integers, and $I(.)$ the indicator function. Assume that there are $m$ arms indexed by the set $M = \{1, 2, \ldots, m\}$. Let $S_k = \{g, b\}$ denote the state space of arm $k$. Let $X_t^k$ denote the random variable representing the state of arm $k$ at time $t$. $P_k$ is the transition probability matrix of arm $k$ where the transition probabilities are $p_{ij}^k = P(X_{t+1}^k = j | X_t^k = i), i, j \in S_k$. We assume that $P_k$ is such that the channels are ergodic. When arm $k$ is played in state $g$ ($b$), it yields reward $r_k > 0$ (0). We assume that the arms are bursty, i.e., $p_{gb}^k + p_{bg}^k < 1, \forall k \in M$. Moreover $p_{gb}^k > \delta > 0, \forall k \in M$. If an arm is played $\tau$ steps ago and the last observed state is $s \in S_k$, let $(s, \tau)$ be the information state for that arm. Let $v_{k,\tau}$ ($u_{k,\tau}$) be the probability that arm $k$ will be in state $g$ given that it is observed $\tau$ steps ago in state $b$ ($g$). We have

$$v_{k,\tau} = \frac{p_{bg}^k}{p_{bg}^k + p_{gb}^k}(1 - (1 - p_{bg}^k - p_{gb}^k)^\tau),$$

$$u_{k,\tau} = \frac{p_{bg}^k}{p_{bg}^k + p_{gb}^k} + \frac{p_{gb}^k}{p_{bg}^k + p_{gb}^k}(1 - p_{bg}^k - p_{gb}^k)^\tau,$$

and $v_{k,\tau}$, $1 - u_{k,\tau}$ are monotonically increasing concave functions by the burstiness assumption.

There exists a user whose goal is to maximize the infinite horizon average reward by only playing one of the arms at each time step. We assume that there is a dummy arm which yields no reward and the user has the option to select this arm, i.e., not play at each time step. The user does not know the transition matrices $P_k, k \in M$, but knows the bound $\delta$ on $p_{gb}^k$, and can only observe the reward of the arm it plays at time $t$. We note that the user knows that the reward of a bad state is 0, thus observing the reward of an arm is equivalent to observing the state of the arm from the user's perspective. Without loss of generality we assume that the user knows the rewards of the good states, since this information can be acquired by initially sampling each arm until a good state is observed. Let $\gamma$ be an admissible algorithm for the user. We represent the expectation with respect $\gamma$ when the transition matrices are $P = (P_1, \ldots, P_k)$ and initial state is $\psi_0$ by $E_{\psi_0,\gamma}^P[.]$. Many subsequent expressions depend on the algorithm $\gamma$ used by the user, but we will explicitly state this dependence only when it is not clear from the context.

Let $u(t)$ denote the arm selected by the user at time $t$. We define a *continuous play* of arm $k$ starting at time $t$ with state $s$ as a pair of plays in which arm $k$ is selected at times $t$ and $t+1$ and state $s$ is observed at time $t$. Let

$$\begin{aligned} & N_n^k(s,s') \\ = & \sum_{t=1}^{n-1} I(u(t)=u(t+1)=k, X_t^k=s, X_{t+1}^k=s') \end{aligned}$$

be the number of times transition from $s$ to $s'$ is observed in continuous plays of arm $k$ up to time $n$. Let

$$C_n^k(s) = \sum_{s' \in \{g,b\}} N_n^k(s,s')$$

be the number of continuous plays of arm $k$ starting with state $s$ up to time $n$. These quantities will be used to estimate the state transition probabilities. Below, we give a definition and a lemma that will be used in the proofs. The norm used is the total variation norm.

*Definition 1:* [20] A Markov chain $X = \{X_t, t \in \mathbb{Z}_+\}$ on a measurable space $(\mathcal{S}, \mathcal{B})$, with transition kernel $P(x, \mathcal{G})$ is uniformly ergodic if there exists constants $\rho < 1, C < \infty$ such that for all $x \in \mathcal{S}$,

$$\left\| e_x P^t - \pi \right\| \leq C\rho^t, t \in \mathbb{Z}_+ \ , \tag{1}$$

where $e_x$ is the $|S|$-dimensional unit row vector whose $x$-th component is one while all other components are zero and $\pi$ is the row vector representing the stationary distribution of the Markov chain.

*Lemma 1:* ([20] Theorem 3.1.) Let $X = \{X_t, t \in \mathbb{Z}_+\}$ be a uniformly ergodic Markov chain for which (1) holds. Let $\hat{X} = \{\hat{X}_t, t \in \mathbb{Z}_+\}$ be the perturbed chain with transition kernel $\hat{P}$. Given the two chains have the same initial distribution let $\psi_t, \hat{\psi}_t$ be the distribution of $X, \hat{X}$ at time $t$ respectively. Then,

$$\begin{aligned} \left\| \psi_t - \hat{\psi}_t \right\| &\leq \left( \hat{t} + C\frac{\rho^{\hat{t}} - \rho^t}{1-\rho} \right) \left\| \hat{P} - P \right\| \\ &= C_1(P,t) \left\| \hat{P} - P \right\| \end{aligned}$$

where $\hat{t} = \left\lceil \log_\rho C^{-1} \right\rceil$.

Clearly for a finite state Markov chain uniform ergodicity is equivalent to ergodicity, and the total variation norm is the $l_1$ norm for vectors, and the induced norm is the maximum row sum norm for matrices.

## IV. Guha's Policy

For the optimization version of the problem we consider, where $P_k$'s are known by the user, Guha et. al. [3] proposed a $(2+\epsilon)$ approximate policy for the infinite horizon average reward problem. Under this approach, Whittle's LP relaxation was first used, where the constraint that exactly one arm is played at each time step is replaced by an average constraint that on average one arm is played at a time. Let $OPT$ be the optimal value of Whittle's LP. Guha et al. showed that $OPT$ is at least the value of the optimal policy in the original problem. The arms are then decoupled by considering the Lagrangian of Whittle's LP. Thus instead of solving the original problem which has a size exponential in $m$, $m$ individual optimization problems are solved, one for each arm. The Lagrange multiplier $\lambda > 0$ is treated as penalty per play and it was shown that the optimal single arm policy has the structure of the policy $\mathcal{P}_k(\tau)$ given in Figure 1: whenever an arm is played and a good state is observed, it will also be played in the next time; if a bad state is observed then the user will wait $\tau - 1$ time steps before playing that arm again. Thus, $\tau$ is called the *waiting time*. Let $R_{k,\tau}$ and $Q_{k,\tau}$ be the average reward and rate of play for policy $\mathcal{P}_k(\tau)$ respectively. $Q_{k,\tau}$ is defined as the average number of times arm $k$ will be played under a single arm policy with waiting time $\tau$. Then from Lemma A.1 of [3] we know that

$$\begin{aligned} R_{k,\tau} &= \frac{r_k v_{k,\tau}}{v_{k,\tau} + \tau p_{gb}^k}, \\ Q_{k,\tau} &= \frac{v_{k,\tau} + p_{gb}^k}{v_{k,\tau} + \tau p_{gb}^k}. \end{aligned}$$

Then, if playing arm $k$ is penalized by $\lambda$, the gain of $\mathcal{P}_k(\tau)$ will be

$$F_{k,\lambda,\tau} = R_{k,\tau} - \lambda Q_{k,\tau}.$$

The optimal single arm policy for arm $k$ under penalty $\lambda$ is thus $\mathcal{P}_k(\tau_k(\lambda))$, where

$$\tau_k(\lambda) = \min \arg \max_{\tau \geq 1} F_{k,\lambda,\tau},$$

and the optimal gain is

$$H_{k,\lambda} = \max_{\tau \geq 1} F_{k,\lambda,\tau}.$$

$H_{k,\lambda}$ is a non-increasing function of $\lambda$ by Lemma 2.6 of [3]. Let $G_\lambda = \sum_{k=1}^m H_{k,\lambda}$. Guha et. al. proposed the algorithm in Figure 2, and showed that the infinite horizon average reward of this algorithm is at least $OPT/(2+\epsilon)$, where $\epsilon > 0$ is the performance parameter given as an input by the user which we will refer to as the *stepsize*. The instantaneous and the long term average reward are balanced by viewing $\lambda$ as an amortized reward per play and $H_{k,\lambda}$ as the per step reward. This balancing procedure is given in Figure 3. After computing the balanced $\lambda$, the optimal single arm policy for this $\lambda$ is combined with the priority scheme in Figure 2 so that at all times at most one arm is played. Denote the gain and the waiting time for the optimal arm $k$ policy at the balanced $\lambda$ by $H_k$ and $\tau_k$.

Note that it is required that at any $t$ one and only one arm must be in good state in Guha's policy. This can be satisfied by initially sampling from $m-1$ arms until a bad state is observed and sampling from the last arm until a good state is observed. Such an initialization will not change the infinite horizon average reward, and in this paper we always assume that such an initialization is completed before the play begins.

At time $t$:
1. If arm $k$ is just observed in state $g$, also play arm $k$ at $t+1$.
2. If arm $k$ is just observed in state $b$, wait $\tau-1$ steps, and then play arm $k$.

Fig. 1. Policy $\mathcal{P}_k(\tau)$

Choose a balanced $\lambda$ by the procedure in Figure 3. Let $S=\{k: H_{k,\lambda}>0\}$, $\tau_k=\tau_k(\lambda)$.
Only play the arms in $S$ according to the following priority scheme:
At time $t$:
1. Exploit: If $\exists k\in S$ in state $(g,1)$, play arm $k$.
2. Explore: If $\exists k\in S$ in state $(b,\tau):\tau\geq\tau_k$, play arm $k$.
3. Idle: If 1 and 2 do not hold do not play any arm.

Fig. 2. Guha's Policy

## V. A Threshold Policy

In this section we consider a threshold variant of Guha's policy, called the $\epsilon_1$-threshold policy. The difference between the two is in balancing the Lagrange multiplier $\lambda$. The complete policy is shown in Figure 4. Let $\tilde{H}_{k,\lambda}$, $\tilde{\tau}_{k,\lambda}$ denote the optimal gain and the optimal waiting time for arm $k$ calculated by the $\epsilon_1$-threshold policy when the penalty per play is $\lambda$. For any $\lambda$ if the optimal single arm policy for arm $k$ has gain $H_{k,\lambda}<\epsilon_1$, that arm is considered not worth playing and $\tilde{H}_{k,\lambda}=0,\tilde{\tau}_{k,\lambda}=\infty$. For any $\lambda$ and any arm $k$ with the optimal gain greater than or equal to $\epsilon_1$, the optimal waiting time after a bad state and the optimal gain are the same as Guha's policy.

Note that at any $\lambda$, any arm $k$ which will be played by the $\epsilon_1$-threshold policy will also be played by Guha's policy with $\tau_{k,\lambda}=\tilde{\tau}_{k,\lambda}$. Arm $k$ with $H_{k,\lambda}<\epsilon_1$ in Guha's policy will not be played by the $\epsilon_1$-threshold policy. The following Lemma states that the average reward of an $\epsilon_1$-threshold policy cannot be much less than $OPT/2$.

*Lemma 2:* Consider the $\epsilon_1$-threshold policy shown in Figure 4 with step size $\epsilon_2$. The average reward of this policy is at least

$$\frac{OPT}{2(1+\epsilon_2)}-m\epsilon_1\ .$$

*Proof:* Let $\lambda^*$ be the balanced Lagrange multiplier computed by Guha's policy with an input of $\epsilon_2$. Then from Figure 3 we have,

$$\lambda^*<\sum_{k=1}^{m}H_{k,\lambda^*}\leq(1+\epsilon_2)\lambda^*$$

For any $\lambda$ we have

$$\sum_{k=1}^{m}H_{k,\lambda}-m\epsilon_1\leq\sum_{k=1}^{m}\tilde{H}_{k,\lambda}\leq\sum_{k=1}^{m}H_{k,\lambda}\ . \tag{2}$$

We consider two cases:
Case 1: $\lambda^*<\sum_{k=1}^{m}\tilde{H}_{k,\lambda^*}$. Then, $\lambda^*$ is also the balanced Lagrange multiplier computed by the $\epsilon_1$-threshold policy.
Case 2: $\lambda^*\geq\sum_{k=1}^{m}\tilde{H}_{k,\lambda^*}$. Then, $\epsilon_1$-threshold policy will continue the process of decreasing $\lambda$ and recomputing $\tilde{G}_\lambda$ until it reaches some $\lambda'$ such that

$$\lambda'<\sum_{k=1}^{m}\tilde{H}_{k,\lambda'}\leq(1+\epsilon_2)\lambda'\ .$$

Since $\tilde{H}_{k,\lambda}$ is non-increasing in $\lambda$ we have

$$\sum_{k=1}^{m}\tilde{H}_{k,\lambda'}\geq\sum_{k=1}^{m}\tilde{H}_{k,\lambda^*}$$

Thus by (2),

$$(1+\epsilon_2)\lambda'\geq\sum_{k=1}^{m}\tilde{H}_{k,\lambda^*}\geq\sum_{k=1}^{m}H_{k,\lambda^*}-m\epsilon_1.$$

By Guha's policy $\sum_{k=1}^{m}H_{k,\lambda^*}\geq OPT/2$. Therefore,

$$\begin{aligned}\sum_{k=1}^{m}\tilde{H}_{k,\lambda'} &\geq OPT/2-m\epsilon_1,\\ \lambda' &\geq OPT/(2(1+\epsilon_2))-m\epsilon_1\end{aligned}$$

The result follows from Theorem 2.7 of [3]. ∎

The following lemma shows that computing $\tilde{\tau}_k$ for the $\epsilon_1$-threshold policy can be done by considering waiting times in a finite window.

Input: $\epsilon$. Perform binary search to find the balanced $\lambda=\lambda(\epsilon)$:
1. Start with $\lambda=\sum_{k=1}^{m}r_k$, Calculate $G_\lambda=\sum_{k=1}^{m}H_{k,\lambda}$ .
2. While $\lambda>G_\lambda$
2.1 $\lambda=\lambda/(1+\epsilon)$,
2.2 Calculate $G_\lambda$.
3. Output $\lambda,\tau_k,k\in M$

Fig. 3. Procedure for the balanced choice of $\lambda$

$\epsilon_1$-threshold policy
1: Input: $\epsilon_1,\epsilon_2$
2: Initialize: $\lambda=\sum_{k=1}^{m}r_k$.
3: Compute $H_{k,\lambda},\tau_{k,\lambda},\forall k\in M$.
4: **for** $k=1,2,\ldots,m$ **do**
5: **if** $H_{k,\lambda}<\epsilon_1$ **then**
6: Set $\tilde{H}_{k,\lambda}=0,\tilde{\tau}_{k,\lambda}=\infty$
7: **else**
8: Set $\tilde{H}_{k,\lambda}=H_{k,\lambda},\tilde{\tau}_{k,\lambda}=\tau_{k,\lambda}$,
9: **end if**
10: **end for**
11: $\tilde{G}_\lambda=\sum_{k=1}^{m}\tilde{H}_{k,\lambda}$.
12: **if** $\lambda<\tilde{G}_\lambda$ **then**
13: Play Guha's policy with $\tau_1=\tilde{\tau}_{1,\lambda},\ldots,\tau_m=\tilde{\tau}_{m,\lambda}$.
14: **else**
15: $\lambda=\lambda/(1+\epsilon_2)$. Return to Step 3
16: **end if**

Fig. 4. pseudocode for the $\epsilon_1$-threshold policy

*Lemma 3:* For any $\lambda$, in order to compute $\tilde{\tau}_k, k \in M$, the $\epsilon_1$-threshold policy only requires to evaluate $F_{k,\lambda,\tau}$ for $\tau \in [1, \tau^*(\epsilon_1)]$, where $\tau^*(\epsilon_1) = \lceil r_{\max}/(\delta\epsilon_1) \rceil$.

*Proof:* For any $\lambda$, $F_{k,\lambda,\tau} \leq R_{k,\tau}$. For $\tau \geq \tau^*(\epsilon_1)$,

$$R_{k,\tau} = r_k \frac{v_{k,\tau}}{v_{k,\tau} + \tau p^k_{gb}} \leq \frac{r_{\max}}{\tau p^k_{gb}} \leq \frac{r_{\max}}{\delta\tau}.$$

■

The following lemma shows that the procedure of decreasing $\lambda$ can only repeat a finite number of times.

*Lemma 4:* Assume that there exists an arm $k$ such that for some $\lambda > 0$, $\tilde{H}_{k,\lambda} \geq \epsilon_1$. Otherwise, no arm will be played by the $\epsilon_1$-threshold policy. Let

$$\hat{\lambda} = \sup\{\lambda : \tilde{H}_{k,\lambda} \geq \epsilon_1\},$$

$\lambda^* = \min\{\hat{\lambda}, \epsilon_1\}$. Let $z(\epsilon_2)$ be the number of cycles, i.e., the number of times $\lambda$ is decreased until the computation of $\tilde{\tau}_k, k \in M$ is completed. We have

$$z(\epsilon_2) \leq \min\left\{ z' \in \mathbb{Z}_+ \text{ such that } (1+\epsilon_2)^{z'} \geq \sum_{k=1}^m r_k/\lambda^* \right\}.$$

*Proof:* Since $\tilde{H}_{k,\lambda}$ is non-increasing in $\lambda$, $\tilde{H}_{k,\lambda^*} \geq \lambda^*$. The result follows from this. ■

Let $\Theta(\epsilon_2) = \{\sum_{k=1}^m r_k, \sum_{k=1}^m r_k/(1+\epsilon_2), \ldots, \sum_{k=1}^m r_k/(1+\epsilon_2)^{z(\epsilon_2)}\}$ be the set of values $\lambda$ takes in $z(\epsilon_2)$ cycles, and

$$\begin{aligned} T_k(\lambda) &= \arg\max_{\tau \geq 1} R_{k,\tau} - \lambda Q_{k,t}, \\ T'_k(\lambda) &= \arg\max_{\tau \geq 1, \tau \notin T_k(\lambda)} R_{k,\tau} - \lambda Q_{k,t}, \end{aligned}$$

be the set of optimal waiting times, and best suboptimal waiting times under penalty $\lambda$ respectively. Let

$$\begin{aligned} \delta(k,\lambda) = \quad & (R_{k,\tau_k} - \lambda Q_{k,\tau_k}) - (R_{k,\tau'_k} - \lambda Q_{k,\tau'_k}), \\ & \tau_k \in T_k(\lambda), \tau'_k \in T'_k(\lambda), \end{aligned}$$

and $\delta_2 = \min_{k \in M, \lambda \in \Theta(\epsilon_2)} \delta(k,\lambda)$.

Consider a different set of transition probabilities $\hat{P} = (\hat{P}_1, \ldots, \hat{P}_m)$. Let $\hat{R}_{k,\tau}$, $\hat{Q}_{k,\tau}$ and $\hat{\tilde{\tau}}_k$ denote the average reward, average number of plays and the optimal waiting time for arm $k$ under $\epsilon_1$-threshold policy and $\hat{P}$ respectively.

*Lemma 5:* For $\epsilon_3 = \delta_2/\left(2(1+\sum_{k=1}^m r_k)\right)$, the event

$$\begin{aligned} \Big\{ & |R_{k,\tau} - \hat{R}_{k,\tau}| < \epsilon_3, |Q_{k,\tau} - \hat{Q}_{k,\tau}| < \epsilon_3, \\ & \forall \tau \in [1, \tau^*(\epsilon_1)] \} \end{aligned} \tag{3}$$

implies the event $\{\tilde{\tau}_k = \hat{\tilde{\tau}}_k, \forall k \in M\}$.

*Proof:* By (3), for any $\lambda \in \Theta$, $\tau \in [1, \tau^*(\epsilon_1)]$,

$$\begin{aligned} & |(R_{k,\tau} - \lambda Q_{k,\tau}) - (\hat{R}_{k,\tau} - \lambda \hat{Q}_{k,\tau})| \\ \leq \; & |R_{k,\tau} - \hat{R}_{k,\tau}| + \lambda |Q_{k,\tau} - \hat{Q}_{k,\tau}| \\ \leq \; & |R_{k,\tau} - \hat{R}_{k,\tau}| + \sum_{k=1}^m r_k |Q_{k,\tau} - \hat{Q}_{k,\tau}| \\ < \; & (1 + \sum_{k=1}^m r_k)\epsilon_3 = \frac{\delta_2}{2}. \end{aligned}$$

Thus, $\hat{F}_{k,\lambda,\tilde{\tau}_k}$ can be at most $\delta_2/2$ smaller than $F_{k,\lambda,\tilde{\tau}_k}$, while for any other $\tau \neq \tilde{\tau}_k$, $\hat{F}_{k,\lambda,\tau}$ can be at most $\delta_2/2$ larger than $F_{k,\lambda,\tau}$ for any $\lambda$. Thus the maximizers are the same for all $\lambda$ and the result follows. ■

The following lemma shows that $\tilde{\tau}_1, \ldots, \tilde{\tau}_m$ for the $\epsilon_1$-threshold policy can be efficiently computed. We define a mathematical operation to be the computation of $R_{k,\tau} - \lambda Q_{k,\tau}$. We do not count other operations such as additions and multiplications.

*Lemma 6:* Finding the balanced $\lambda$ and $\tilde{\tau}_1, \ldots, \tilde{\tau}_m$ requires at most $m \lceil \log(z(\epsilon_2)) \rceil \tau^*(\epsilon_1)$ mathematical operations.

*Proof:* Since $\tilde{G}_\lambda = \sum_{k=1}^m \tilde{H}_{k,\lambda}$ is decreasing in $\lambda$, the balanced $\lambda$ can be computed by binary search. By Lemma 4 the number of cycles required to find the optimal $\lambda$ by binary search is $\lceil \log(z(\epsilon_2)) \rceil$. For each $\lambda$ and each arm $k$, $\tilde{H}_{k,\lambda}$ and $\tau_k(\lambda)$ can be calculated by at most $\tau^*(\epsilon_1)$ mathematical operations. ■

## VI. The Adaptive Balance Algorithm (ABA)

We propose the Adaptive Balance Algorithm (ABA) given in Figure 5 as a learning algorithm which is based on the $\epsilon_1$-threshold policy instead of Guha's policy. This choice has several reasons. The first concerns the union bound we will use to relate the probability that the adaptive algorithm deviates from the $\epsilon_1$-threshold policy with the probability of accurately calculating the average reward and the rate of play for the single arm policies given the estimated transition probabilities. In order to have finite number of terms in the union bound, we need to evaluate the gains $F_{k,\lambda,\tau}$ at finite number of waiting times $\tau$. We do this by the choice of a finite time window $[1, \tau^*]$, for which we can bound our loss in terms of the average reward. Thus, the optimal single arm waiting times are computed by comparing $F_{k,\lambda,\tau}$'s in $[1, \tau^*]$. The second is due to the non-monotonic behaviour of the gain $F_{k,\lambda,\tau}$ with respect to the waiting time $\tau$. For example, there exists transition probabilities satisfying the burstiness assumption such that the maximum of $F_{k,\lambda,\tau}$ occurs at $\tau > \tau^*$, while the second maximum is at $\tau = 1$. Then, by considering the time window $[1, \tau^*]$, it will not be possible to play with the same waiting times as in Guha's policy independent of how much we explore. The third is that for any $OPT/(1+\epsilon)$ optimal Guha's policy, there exists $\epsilon_1$ and $\epsilon_2$ such that the $\epsilon_1$-threshold policy is $OPT/(1+\epsilon)$ optimal. Thus, any average reward that can be achieved by Guha's policy can also be achieved by the $\epsilon_1$-threshold policy.

Let $\hat{p}^k_{bg}(t), \hat{p}^k_{gb}(t), k \in M$, and $\hat{P}(t) = (\hat{P}_1(t), \ldots, \hat{P}_k(t))$ be the estimated transition probabilities and the estimated transition probability matrices at time $t$ respectively. We will use $\hat{\cdot}$ to represent the quantities computed according to $\hat{P}(t)$.

ABA consists of exploration and exploitation phases. Exploration serves the purpose of estimating the transition probabilities. If at time $t$ the number of samples used to estimate the transition probability from state $g$ or $b$ of any arm is less than $a \log t$, ABA explores to increase the accuracy of the estimated transition probabilities. We call $a$ the *exploration constant*. In general it should be chosen large enough (depending on

$P, r_1, \ldots, r_m$) so that our results will hold. We will describe an alternative way to choose $a$ (independent of $P, r_1, \ldots, r_m$) in Section IX. If all the transition probabilities are accurately estimated, then ABA exploits by using these probabilities in the $\epsilon_1$-threshold policy to select an arm. Note that the transition probability estimates can also be updated after an exploitation step, depending on whether a continuous play of an arm occurred or not. We denote ABA by $\gamma^A$.

In the next section, we will show that the expected number of times in which ABA deviates from the $\epsilon_1$-threshold policy given $P$ is logarithmic in time.

Adaptive Balance Algorithm

Input: $\epsilon_1$, $\epsilon_2$ $\tau^*(\epsilon_1)$, $a > 0$.
Initialize: Set $t = 1$, $N^k(i,j) = 0, C^k(i) = 0, \forall k \in M, i, j \in S_k$. Play each arm once so the initial information state can be represented as an element of countable form $(s_1, \tau_1), \ldots, (s_m, \tau_m)$, where only one arm is observed in state $g$ one step ago while all other arms are observed in state $b$, $\tau_k > 1$ steps ago.
**while** $t \geq 1$ **do**
  $\hat{p}^k_{gb} = \frac{1I(N^k(g,b)=0)+N^k(g,b)}{2I(C^k(g)=0)+C^k(g)}$,
  $\hat{p}^k_{bg} = \frac{1I(N^k(b,g)=0)+N^k(b,g)}{2I(C^k(b)=0)+C^k(b)}$,
  $W = \{(k,i), k \in M, i \in S_k : C^k(i) < a \log t \}$.
  **if** $W \neq \emptyset$ **then**
    EXPLORE
    **if** $u(t-1) \in W$ **then**
      $u(t) = u(t-1)$
    **else**
      select $u(t) \in W$ arbitrarily.
    **end if**
  **else**
    EXPLOIT
    Start with $\lambda = \sum_{k=1}^m r_k$.
    Run the procedure for the balanced choice $\lambda$ given by the $\epsilon_1$-threshold policy with step size $\epsilon_2$ and transition matrices $\hat{P}(t)$.
    Obtain $\hat{\tau}_1, \ldots, \hat{\tau}_m$.
    Play according to Guha's Policy with parameters $\hat{\tau}_1, \ldots, \hat{\tau}_m$ for only one step.
  **end if**
  **if** $u(t-1) = u(t)$ **then**
    **for** $i, j \in S_{u(t)}$ **do**
      **if** State $j$ is observed at $t$, state $i$ is observed at $t-1$ **then**
        $N^{u(t)}(i,j) = N^{u(t)}(i,j) + 1$, $C^{u(t)}(i) = C^{u(t)}(i) + 1$.
      **end if**
    **end for**
  **end if**
  $t := t + 1$
**end while**

Fig. 5. pseudocode for the Adaptive Balance Algorithm (ABA)

## VII. Number of Deviations of ABA from the $\epsilon_1$-threshold policy

Let $\gamma^{\epsilon_1, P}$ be the rule determined by the $\epsilon_1$-threshold policy given $\epsilon_2$ and $P = (P_1, \ldots, P_k)$, and $\tilde{\tau}_1, \ldots, \tilde{\tau}_m$ be the waiting times after a bad state for $\gamma^{\epsilon_1, P}$. Let $T_N$ be the number of times $\gamma^{\epsilon_1, P}$ is not played up to $N$. Let $E_t$ be the event that ABA exploits at time $t$. Then,

$$
\begin{aligned}
T_N &\leq \sum_{t=1}^{N} I(\hat{\tau}_k(t) \neq \tilde{\tau}_k \text{ for some } k \in M) \\
&\leq \sum_{t=1}^{N} I(\hat{\tau}_k(t) \neq \tilde{\tau}_k \text{ for some } k \in M, E_t) + \sum_{t=1}^{N} I(E_t^C) \\
&\leq \sum_{k=1}^{m} \sum_{t=1}^{N} I(\hat{\tau}_k(t) \neq \tilde{\tau}_k, E_t) + \sum_{t=1}^{N} I(E_t^C) \\
&\leq \sum_{k=1}^{m} \sum_{t=1}^{N} I(|R_{k,\tau} - \hat{R}_{k,\tau}(t)| \geq \epsilon_3 \text{ or } |Q_{k,\tau} - \hat{Q}_{k,\tau}| \geq \epsilon_3 \\
&\quad \text{for some } \tau \in [1, \tau^*(\epsilon_1)], E_t) + \sum_{t=1}^{N} I(E_t^C) \\
&\leq \sum_{k=1}^{m} \sum_{t=1}^{N} \sum_{\tau=1}^{\tau^*(\epsilon_1)} \Big( I(|R_{k,\tau} - \hat{R}_{k,\tau}(t)| \geq \epsilon_3, E_t) \\
&\quad + I(|Q_{k,\tau} - \hat{Q}_{k,\tau}| \geq \epsilon_3, E_t) \Big) + \sum_{t=1}^{N} I(E_t^C) \qquad (4)
\end{aligned}
$$

We first bound the regret due to explorations.

*Lemma 7:*

$$
E^P_{\psi_0, \gamma^A} \left[ \sum_{t=0}^{N-1} I(E_t^C) \right] \leq 2ma \log N (1 + T_{\max}),
$$

where $T_{\max} = \max_{k \in M, i, j \in S_k} E[T_{k,ij}] + 1$, $T_{k,ij}$ is the hitting time of state $j$ of arm $k$ starting from state $i$ of arm $k$. Since all arms are ergodic $E[T_{k,ij}]$ is finite for all $k, i, j$.

*Proof:* The number of transition probability updates that results from explorations up to time $N-1$ is at most $\sum_{k=1}^m \sum_{i \in S_k} a \log N$. The expected time spent in exploration during a single update is at most $(1 + T_{\max})$. ■

The next two lemmas bound the probability of deviation of $\hat{R}_{k,\tau}(t)$ and $\hat{Q}_{k,\tau}(t)$ from $R_{k,\tau}$ and $Q_{k,\tau}$ respectively. Let $C_1(P_k, \tau_k), k \in M, \tau_k \in [1, \tau^*(\epsilon_1)]$ be the constant given in Lemma 1, $C_1(P) = \max_{k \in M, \tau_k \in [1, \tau^*(\epsilon_1)]} C_1(P_k, \tau_k)$.

*Lemma 8:* When ABA is used, for

$$
a \geq \frac{3}{2(\min\{\frac{C_1(P)\epsilon\delta^2}{4r_{\max}}, \frac{\epsilon\delta^2}{2r_{\max}}\})^2},
$$

we have on the event $E_t$ (here we only consider deviations in exploitation steps)

$$
P(|R_{k,\tau} - \hat{R}_{k,\tau}(t)| \geq \epsilon) \leq \frac{18}{t^2}.
$$

*Proof:*

$$\begin{aligned}
& P(|R_{k,\tau} - \hat{R}_{k,\tau}(t)| \geq \epsilon) \\
= \; & P\left(\left|\frac{r_k v_{k,\tau}}{v_{k,\tau} + \tau p^k_{gb}} - \frac{r_k \hat{v}_{k,\tau}(t)}{\hat{v}_{k,\tau}(t) + \tau \hat{p}^k_{gb}(t)}\right|\right) \\
= \; & P\left(\tau r_k |v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \right. \\
& \left. \epsilon |v_{k,\tau} + \tau p^k_{gb}||\hat{v}_{k,\tau}(t) + \tau \hat{p}^k_{gb}(t)|\right) \\
\leq \; & P\left(\tau r_k |v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \epsilon \tau^2 \delta^2\right) \\
\leq \; & P\left(|v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \frac{\epsilon \delta^2}{r_{\max}}\right) \\
\leq \; & \frac{18}{t^2},
\end{aligned}$$

where the last inequality follows form Lemma 12 since $a \geq 3/(2(\min\{\frac{C_1(P)\epsilon\delta^2}{4r_{\max}}, \frac{\epsilon\delta^2}{2r_{\max}}\})^2)$. ■

*Lemma 9:* When ABA is used, for

$$a \geq \frac{3}{2(\min\{\frac{\epsilon\delta^2 C_1(P)}{4}, \frac{\epsilon\delta^2}{2}\})^2},$$

we have on the event $E_t$

$$P(|Q_{k,\tau} - \hat{Q}_{k,\tau}(t)| \geq \epsilon) \leq \frac{18}{t^2}.$$

*Proof:*

$$\begin{aligned}
& P(|Q_{k,\tau} - \hat{Q}_{k,\tau}(t)| \geq \epsilon) \\
= \; & P\left(\left|\frac{v_{k,\tau} + p^k_{gb}}{v_{k,\tau} + \tau p^k_{gb}} - \frac{\hat{v}_{k,\tau}(t) + \hat{p}^k_{gb}(t)}{\hat{v}_{k,\tau}(t) + \tau \hat{p}^k_{gb}(t)}\right|\right) \\
= \; & P\left((\tau-1)|v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \right. \\
& \left. \epsilon |v_{k,\tau} + \tau p^k_{gb}||\hat{v}_{k,\tau}(t) + \tau \hat{p}^k_{gb}(t)|\right) \\
\leq \; & P\left((\tau-1)|v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \epsilon \tau^2 \delta^2\right) \\
\leq \; & P\left(|v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \epsilon \delta^2\right) \\
\leq \; & \frac{18}{t^2},
\end{aligned}$$

where the last inequality follows form Lemma 12 since $a \geq 3/(2(\min\{\epsilon\delta^2 C_1(P)/4, (\epsilon\delta^2)/2\})^2)$. ■

The lower bound on the exploration constant $a$ in Lemmas 8 and 9 is sufficient to make the estimated transition probabilities at an exploitation step close enough to the true transition probabilities to guarantee that the estimated waiting time is equal to the exact waiting time with very high probability, i.e., the probability of error at any time $t$ is $O(1/t^2)$. The following theorem bounds the expected number of times ABA differs from $\gamma^{\epsilon_1,P}$.

*Theorem 1:*

$$E[T_N] \leq 36m\tau^*(\epsilon_1)\beta + 2ma\log N(1 + T_{\max}),$$

for

$$a \geq \frac{3}{2\min\{\frac{C_1(P)\epsilon_3\delta^2}{4r_{\max}}, \frac{\epsilon_3\delta^2}{2r_{\max}}, \frac{C_1(P)\epsilon_3\delta^2}{4}, \frac{\epsilon_3\delta^2}{2}\}}, \tag{5}$$

where

$$\beta = \sum_{t=1}^{\infty} \frac{1}{t^2}.$$

*Proof:* Taking the expectation of (4) and using Lemma 7

$$\begin{aligned}
E[T_N] \leq & \sum_{k=1}^{m}\sum_{t=1}^{N}\sum_{\tau=1}^{\tau^*(\epsilon_1)} \Big(P(|R_{k,\tau} - \hat{R}_{k,\tau}(t)| \geq \epsilon_3, E_t) \\
& + P(|Q_{k,\tau} - \hat{Q}_{k,\tau}| \geq \epsilon_3, E_t)\Big) + 2ma\log N(1 + T_{\max}).
\end{aligned}$$

Then, by results of Lemmas 8, 9,

$$\begin{aligned}
E[T_N] & \leq \sum_{k=1}^{m}\sum_{t=1}^{N}\sum_{\tau=1}^{\tau^*(\epsilon_1)} \frac{20}{t^2} + 2ma\log N(1 + T_{\max}) \\
& \leq 36m\tau^*(\epsilon_1)\beta + 2ma\log N(1 + T_{\max}).
\end{aligned}$$

■

## VIII. Performance of ABA

In this section we consider the performance of ABA. First we show that the performance of ABA is at most $\epsilon$ worse than $OPT/2$. Since each arm is an ergodic Markov chain, the $\epsilon_1$-threshold policy is ergodic. Thus, after a single deviation from the $\epsilon_1$-threshold policy only a finite difference in reward from the $\epsilon_1$-threshold policy can occur.

*Theorem 2:* Given $\delta, \epsilon_1, \epsilon_2$ and $a$ as in (5), the infinite horizon average reward of ABA is at least

$$\frac{OPT}{2(1+\epsilon_2)} - m\epsilon_1 = \frac{OPT}{2} - \epsilon,$$

for

$$\epsilon = \frac{\epsilon_2 OPT}{2(1+\epsilon_2)} + m\epsilon_1.$$

Moreover, the number of mathematical operations required to select an arm at any time is at most

$$m\,\lceil \log(z(\epsilon_2)) \rceil\, \tau^*(\epsilon_1).$$

*Proof:* Since, after each deviation from the $\epsilon_1$-threshold policy only a finite difference in reward from the $\epsilon_1$-threshold policy can occur and the expected number of deviations of ABA is logarithmic (even sublinear is sufficient), ABA and the $\epsilon_1$-threshold policy have the same infinite horizon average reward. Computational complexity follows from Lemma 6. ■

ABA has the fastest rate of convergence (logarithmic in time) to the $\epsilon_1$-threshold policy given $P$, i.e., $\gamma^{\epsilon_1,P}$. This follows from the large deviation bounds where in order to logarithmically upper bound the number of errors in exploitations, at least logarithmic number of explorations is required. Although finite time performance of Guha's policy and $\gamma^{\epsilon_1,P}$ is not investigated, minimizing the number of deviations will keep the performance of ABA very close to $\gamma^{\epsilon_1,P}$ for any finite time. We define the regret of ABA with respect to $\gamma^{\epsilon_1,P}$ at time $N$ as the difference between the expected total reward of $\gamma^{\epsilon_1,P}$ and ABA at time $N$. Next, we will show that this regret is logarithmic, uniformly over time.

*Theorem 3:* Let $r^{\gamma}(t)$ be the reward obtained at $t$ by policy $\gamma$. Given $\delta, \epsilon_1, \epsilon_2$ and $a$ as in (5),

$$\begin{aligned} & \left| E_{P,\psi_0}^{\gamma^A}\left[\sum_{t=1}^{N} r^{\gamma^A}(t)\right] - E_{P,\psi_0}^{\gamma^{\epsilon_1,P}}\left[\sum_{t=1}^{N} r^{\gamma^{\epsilon_1,P}}(t)\right]\right| \\ \leq \quad & K(36m\tau^*(\epsilon_1)\beta + 2ma\log N(1+T_{\max})), \end{aligned}$$

where $K$ is the maximum difference in expected reward resulting from a single deviation from $\gamma^{\epsilon_1,P}$.

*Proof:* A single deviation from $\gamma^{\epsilon_1,P}$ results in a difference at most $K$. The expected number of deviations from $\gamma^{\epsilon_1,P}$ is at most $(20m\tau^*(\epsilon_1)\beta + 2ma\log N(1+T_{\max}))$ from Theorem 1. ∎

## IX. Discussion

We first comment on the choice of the exploration constant $a$. Note that in computing the lower bound for $a$ given by (5), $\epsilon_3$ and $C_1(P)$ are not known by the user. One way to overcome this is to increase $a$ over time. Let $a^*$ be the value of the lower bound. Thus, instead of exploring when $C_t^k(s) < a\log t$ for some $k \in M, s \in S_k$, ABA will explore when $C_t^k(s) < a(t)\log t$ for some $k \in M, s \in S_k$, where $a$ is an increasing function such that $a(1) = 1$, $\lim_{t\to\infty} a(t) = \infty$. Then after some $t_0$, we will have $a(t) > a^*, t \geq t_0$ so our proofs for the number of deviations from the $\epsilon_1$-threshold policy in exploitation steps will hold. Clearly, the number of explorations will be in the order $a(t)\log t$ rather than $\log t$. Given that $a(t)\log t$ is sublinear in $t$, Theorem 2 will still hold. The performance difference given in Theorem 3 will be bounded by $a(N)\log N$ instead of $\log N$.

Secondly, we note that our results hold under the burstiness assumption, i.e., $p_{gb}^k + p_{bg}^k < 1, \forall k \in M$. This is a sufficient condition for the approximate optimality of Guha's policy and the ABA. It is an open problem to find approximately optimal algorithms under weaker assumptions on the transition probabilities.

Thirdly, we will compare the results obtained in the previous sections with the results in [12] and [19]. The algorithm in [12], i.e., the *regenerative cycle algorithm* (RCA) assigns an index to each channel which is based on the sample mean of the rewards from that channel plus an exploration term that depends on how many times that channel is selected. Indices in RCA can be computed recursively since they depend on the sample mean, and the computation may not be necessary at every $t$ since RCA operates in blocks. Thus, RCA is computationally simpler than ABA. It is shown that for any $t$ the regret of RCA with respect to the best single-channel policy (policy which always selects the channel with the highest mean reward) is logarithmic in time. This result holds for general finite state channels. However, the best single-channel policy may have linear regret with respect to the optimal policy which is allowed to switch channels at every time [21]. Another algorithm is the *adaptive learning algorithm* (ALA) proposed in [19]. ALA assigns an index to each channel based on an inflation of the right hand side of the estimated average reward optimality equation. At any time if the transition probability estimates are accurate, ALA exploits by choosing the channel with the highest index. Otherwise, it explores to estimate the transition probabilities. Thus, at each exploitation phase ALA needs to solve the estimated average reward optimality equations for a POMDP which is intractable. However, under some assumptions on the structure of the optimal policy for the infinite horizon average reward problem, ALA is shown to achieve logarithmic regret with respect to the optimal policy for the finite horizon undiscounted problem. Thus, we can say that ABA lies in between the two algorithms discussed above. It is both efficient in terms of computation and performance.

Finally, we note that the adaptive learning approach we used here can be generalized for learning different policies, whenever the computation of actions are related to the transition probability estimates in such a way that it is possible to exploit some large deviation bound. As an example, we can develop a similar adaptive algorithm with logarithmic regret with respect to the myopic policy. Although myopic policy is in general not optimal for the restless bandit problem it is computationally simple and its optimality is shown under some special cases [4].

## X. Conclusion

In this paper we proposed an adaptive learning algorithm for the OSA problem which is approximately optimal and poly-time computable. Our algorithm is based on learning a threshold-variant of Guha's policy which is proved to be approximately optimal when the transition probabilities of channels are known by the user. To the best of our knowledge this is the first result in OSA showing that approximate optimality can be achieved with a computationally efficient algorithm.

## References


[1] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Math. Oper. Res.*, vol. 24, no. 2, pp. 293–305, 1999.

[2] P. Whitlle, "Restless bandits," *J. Appl. Prob.*, pp. 301–313, 1988.

[3] S. Guha, K. Mungala, and P. Shi, "Approximation algorithms for restless bandit problems," *Journal of the ACM*, vol. 58, December 2010.

[4] S. H. A. 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, September 2009.

[5] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 55, pp. 527–535, 1952.

[6] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[7] R. Agrawal, "Sample mean based index policies with o(log n) regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, December 1995.

[8] V. Anantharam, P. Varaiya, and J. . Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards," *IEEE Trans. Automat. Contr.*, pp. 968–975, November 1987.

[9] ——, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards," *IEEE Trans. Automat. Contr.*, pp. 977–982, November 1987.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, p. 235256, 2002.

[11] C. Tekin and M. Liu, "Online algorithms for the multi-armed bandit problem with markovian rewards," in *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computation*, September.
[12] ——, "Online learning in opportunistic spectrum access: A restless bandit approach," in *30th IEEE International Conference on Computer Communications (INFOCOM)*, April 2011.
[13] ——, "Online learning of rested and restless bandits," *submitted to IEEE Transactions on Information Theory, under revision*.
[14] ——, "Performance and convergence of multi-user online learning," in *2nd International ICST Conference on Game Theory for Networks (GAMENETS)*, April 2011.
[15] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, pp. 5667 – 5681, November 2010.
[16] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple players: Learning under competition," in *Proc. of IEEE INFOCOM*, March 2010.
[17] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for markov decision processes," *Mathematics of Operations Research*, vol. 22, no. 1, pp. 222–255, 1997.
[18] A. Tewari and P. Bartlett, "Optimistic linear programming gives logarithmic regret for irreducible mdps," *Advances in Neural Information Processing Systems*, vol. 20, pp. 1505–1512, 2008.
[19] C. Tekin and M. Liu, "Adaptive learning of uncontrolled restless bandits with logarithmic regret," in *Proc. Forty-Ninth Annual Allerton Conference on Communication, Control, and Computing*, September 2011.
[20] A. Y. Mitrophanov, "Senstivity and convergence of uniformly ergodic markov chains," *J. Appl. Prob.*, vol. 42, pp. 1003–1014, 2005.
[21] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, pp. 48–77, 2002.

## APPENDIX

The following lemma, which is a large deviation bound, will be frequently used in the proofs.

*Lemma 10:* (Chernoff-Hoeffding Bound) Let $X_1, \ldots, X_n$ be random variables with common range [0,1], such that $E[X_t|X_{t-1}, \ldots, X_1] = \mu$. Let $S_n = X_1 + \ldots + X_n$. Then for all $\epsilon \geq 0$

$$P(|S_n - n\mu| \geq \epsilon) \leq 2e^{\frac{-2\epsilon^2}{n}}$$

Using Lemma 10, we will show that the probability that an estimated transition probability is significantly different from the true transition probability given ABA is in an exploitation phase is very small.

*Lemma 11:*

$$P\left(|\hat{p}^k_{ss'}(t) - p^k_{ss'}| > \epsilon, E_t\right) \leq \frac{2}{t^2},$$

for all $t$, $s, s' \in S_k$, $k \in M$, for $a \geq 3/(2\epsilon^2)$.

*Proof:* Let $t(l)$ be the time $C^k_{t(l)}(s) = l$. We have,

$$\begin{aligned}\hat{p}^k_{ss'}(t) &= \frac{N^k_t(s,s')}{C^k_t(s)} \\ &= \frac{\sum_{l=1}^{C^k_t(s)} I(X^k_{t(l)-1} = s, X^k_{t(l)} = s')}{C^k_t(s)}.\end{aligned}$$

Note that $I(X^k_{t(l)-1} = s, X^k_{t(l)} = s'), l = 1, 2, \ldots, C^k_t(s)$ are i.i.d. random variables with mean $p^k_{ss'}$. Then

$$\begin{aligned}&P\left(|\hat{p}^k_{ss'}(t) - p^k_{ss'}| > \epsilon, E_t\right) \\ &= P\left(\left|\frac{\sum_{l=1}^{C^k_t(s)} I(X^k_{t(l)-1} = s, X^k_{t(l)} = s')}{C^k_t(s)} - p^k_{ss'}\right| \geq \epsilon, E_t\right) \\ &= \sum_{b=1}^{t} P\left(\left|\sum_{l=1}^{C^k_t(s)} I(X^k_{t(l)-1} = s, X^k_{t(l)} = s') \right.\right. \\ &\quad \left.\left. -C^k_t(s)p^k_{ss'}\right| \geq C^k_t(s)\epsilon, C^k_t(s) = b, E_t\right) \\ &\leq \sum_{b=1}^{t} 2e^{\frac{-2(a\log t\epsilon)^2}{a\log t}} = \sum_{b=1}^{t} e^{-2a\log t(\epsilon)^2} \\ &= 2\sum_{b=1}^{t}\frac{1}{t^{2a\epsilon^2}} = \frac{1}{t^{2a\epsilon^2-1}} \leq \frac{2}{t^2},\end{aligned}$$

where we used Lemma 10 and the fact that $C^k_t(s) \geq a\log t$ w.p.1. in the event $E_t$. ∎

The following Lemma which is an intermediate step in proving that if time $t$ is an exploitation phase then the difference between $R_{k,\tau}$, $\hat{R}_{k,\tau}$ and $Q_{k,\tau}$,$\hat{Q}_{k,\tau}$ will be small with high probability, is proved using Lemma 11.

*Lemma 12:* When ABA is used, we have for $a \geq 3/(2(\min\{\epsilon C_1(P)/4, \epsilon/2\})^2)$,

$$P(|v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \epsilon, E_t) \leq \frac{18}{t^2}.$$

*Proof:*

$$\begin{aligned}&P(|v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \epsilon, E_t) \\ \leq\ & P(|v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \epsilon, \\ & |p^k_{gb} - \hat{p}^k_{gb}(t)| < \eta, E_t) \\ +\ & P(|p^k_{gb} - \hat{p}^k_{gb}(t)| \geq \eta, E_t),\end{aligned}$$

for any $\eta$. Letting $\eta = \epsilon/2$ and using Lemma 11 we have

$$\begin{aligned}&P(|v_{k,\tau}\hat{p}^k_{gb}(t) - p^k_{gb}\hat{v}_{k,\tau}(t)| \geq \epsilon, E_t) \\ \leq\ & 4\left(P\left(|p^k_{gb} - \hat{p}^k_{gb}(t)| \geq \frac{\epsilon}{4C_1(P)}, E_t\right)\right. \\ & \left. +P\left(|p^k_{bg} - \hat{p}^k_{bg}(t)| \geq \frac{\epsilon}{4C_1(P)}, E_t\right)\right) \\ +\ & P(|p^k_{gb} - \hat{p}^k_{gb}(t)| \geq \epsilon/2, E_t) \leq \frac{18}{t^2}.\end{aligned}$$

∎