Adaptive Learning of Uncontrolled Restless Bandits with Logarithmic Regret

Cem Tekin, Mingyan Liu

***Abstract*—In this paper we consider the problem of learning the optimal policy for the uncontrolled restless bandit problem. In this problem only the state of the selected arm can be observed, the state transitions are independent of control and the transition law is unknown. We propose a learning algorithm which gives logarithmic regret uniformly over time with respect to the optimal finite horizon policy with known transition law under some assumptions on the transition probabilities of the arms and the structure of the optimal stationary policy for the infinite horizon average reward problem.**

I. Introduction

In an *uncontrolled restless bandit problem* (URBP) there is a set of arms indexed by $1, 2, \ldots, m$ whose state process is discrete and follows a Markov rule independent of each other. The user chooses one arm at each step, gets the reward and observes the current state of that arm. The control action, i.e., the arm selection, does not affect the state transitions. However it is both used to exploit the instantaneous reward and to decrease the uncertainty about the current state of the system by exploring. Thus the optimal policy should balance the tradeoff between exploration and exploitation.

If the structure of the system, i.e., the state transition probabilities and the rewards of the arms are known, then the optimal policy can be found by dynamic programming for any finite horizon problem. In the case of infinite horizon, stationary optimal policies can be found for the discounted problem by using the contraction properties of the dynamic programming operator. For the infinite horizon average reward problem, stationary optimal policies can be found under some assumptions on the transition probabilities [1], [2]. However, knowing the structure of system before using the system is a strong assumption. In most of the systems, the user does not have a perfect model for the system at the beginning but learns the model over time. Therefore, we assume that initially the user does not know the transition probabilities of the arms. The user learns them over time based on its observations. Thus, our goal is to design learning algorithms with fastest convergence rate, i.e., minimum regret where regret of a learning policy at time t is defined as the difference between reward of the optimal policy for the undiscounted t-horizon problem with full information about the system model and the undiscounted reward of the learning policy up to time t.

In this paper we show that under some assumptions on the transition probabilities of the arms and the structure of the optimal policy for the infinite horizon average reward problem, algorithms with logarithmic regret uniformly in time with respect to the optimal policy for the finite time undiscounted problem with known transition probabilities exist. We also claim that logarithmic order is the best achievable order for URBP. To the best of our knowledge this paper is the first attempt to extend the optimal adaptive learning to partially observable Markov decision processes (POMDP).

Related work in optimal adaptive learning started with the paper of Lai and Robbins [3], where the 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 [4] 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. [5], [6] 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. [7] 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 [8], [9] 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 [10]. Decentralized multi-player versions of the i.i.d. multi-armed bandit problem under different collision models were considered in [11], [12], [13]. Other research on adaptive learning focused on Markov Decision Processes (MDP) with finite state and action space. Burnetas and Katehakis [14] 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 [15] 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 [14] 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

C. Tekin and M. Liu are with the Dept. of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48105, {cmtkn, mingyan}@eecs.umich.edu.

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.

The organization of the remainder of this paper is as follows. In Section II we give the problem formulation, notations and some lemmas that will be used throughout the proofs in the paper. In Section III we give sufficient conditions under which the average reward optimality equation has a solution. In Section IV we give an equivalent countable representation of the information state and an assumption under which the regret of a policy can be related to the expected number of times a suboptimal action is taken. Then, we give an upper bound for the regret of an admissable policy in Section V. In Section VI an adaptive learning algorithm is given, and in Section VII an upper bound for the regret of the adaptive learning algorithm is derived. Section VIII concludes the paper.

Due to page limitations, some proofs are not given. They will be included in the full version of the paper.

II. Problem Formulation, Preliminaries and Notation

$\mathbb{N} = \{1, 2, \ldots\}$ is the set of natural numbers, $\mathbb{Z}_+ = \{0, 1, \ldots\}$ is the set of non-negative integers, $(.\bullet.)$ represents the standard inner product, $||.||_1$ represents the l_1 norm for vectors and the induced maximum row sum norm for matrices. For a vector or group of matrices v, (v_{-u}, v') represents the vector or a group of matrices whose uth element is v', while all other elements are the same as the elements of v, e_x is the unit vector whose xth component is one while all other components are zero, and whose dimension will be clear from the context. Unit vectors with dimension $|S_k|$ are represented by e_x^k. Let $\beta = \sum_{t=1}^{\infty} 1/t^2$.

Assume that there are m arms, indexed by the set $M = \{1, 2, \ldots, m\}$. We assume that all arms are independent, irreducible, aperoidoc, discrete time Markov chains $t = 0, 1, \ldots$. Let $x_k \in S_k$ denote a state of arm k where S_k is the state space of arm k. For simplicity x_k also represents the reward from state x_k of arm k and we assume that $S_k \cap S_l = \emptyset$ for $k \neq l$ without loss of generality. Then the state space of the system is $S = S_1 \times \ldots \times S_m$ and $x = (x_1, \ldots, x_m) \in S$ is a state of the system. Let $r_{\max} = \max_{x_k \in S_k, k \in M} x_k$. Let $X_{k,t}$, $X_t = (X_{1,t}, \ldots, X_{m,t})$ be the random variable representing the state of arm k at time t and the state of the system at time t respectively. P_k is the transition probability matrix of arm k, $p_{k,x_k x_k'} = (P_k)_{x_k x_k'} = P(X_{k,t+1} = x_k' | X_{k,t} = x_k)$ and $P = (P_1, \ldots, P_m)$ is the set of transition probability matrices.

There is a user who selects one of the m arms at any t and gets the reward from that arm depending on the state of that arm with the goal of maximizing the undiscounted sum of the rewards for any finite horizon. The user does not know P thus he needs to balance exploration and exploitation in order to maximize his reward. Moreover, the user can only observe the state of the arm he chooses and cannot observe the state of the other arms. Thus, the user should learn the uncontrolled POMDP problem. The action and observation spaces of the user at any time t are $U = \{1, \ldots, m\}$ and $Y = \cup_{k=1}^{m} S_k$ respectively. Then $u_t \in U$, $y_t \in Y$ are the action and the observation of the user at time t respectively and U_t, Y_t are the random variables representing the action and the observation at time t respectively. The history at time t is $z^t = (u_0, y_1, u_1, y_2, \ldots, u_{t-1}, y_t)$.

Let $Q_P(y|u)$ be the substochastic transition probability matrix such that $(Q_P(y|u))_{xx'} = P_P(X_{t+1} = x', Y_{t+1} = y | X_t = x, U_t = u)$. For URBP $Q_P(y|u)$ is the zero matrix for $y \notin S_u$. For $y \in S_u$ only nonzero entries of $Q_P(y|u)$ are the ones for which $x_u = y$.

Finally we give useful definitions and lemmas.

Lemma 1: for $\rho_k, \rho_k' \in [0, 1]$ we have

$$|\rho_1 \ldots \rho_m - \rho_1' \ldots \rho_m'| \leq \sum_{k=1}^{m} |\rho_k - \rho_k'| \tag{1}$$

The norm used in the equations below is the total variation norm. For finite and countable vectors this corresponds to l_1 norm, and the induced matrix norm corresponds to maximum absolute row sum norm.

Definition 1: [16] 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 $x \in \mathcal{S}$,

$$\left\| e_x P^t - \pi \right\| \leq C \rho^t, t \in \mathbb{Z}_+, \tag{2}$$

Lemma 2: ([16] Theorem 3.1.) Let $X = \{X_t, t \in \mathbb{Z}_+\}$ be a uniformly ergodic Markov chain for which (2) 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,

$$\begin{aligned} \left\| \psi_t - \hat{\psi}_t \right\| &\leq \left(\hat{t} + C \frac{\rho^{\hat{t}} - \rho^t}{1 - \rho} \right) \left\| \hat{P} - P \right\| \\ &= C_1(P, t) \left\| \hat{P} - P \right\| \end{aligned} \tag{3}$$

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

III. Solutions of the Average Reward Optimality Equation (AROE)

Assume that the transition probability matrices for the arms are known by the user. Then, URBP turns into an optimization problem (POMDP) rather than a learning problem. In its general form this problem is intractable [17], but heuristics, approximations and exact solutions under different assumptions on the arms are studied by [18], [19], [20] and many others.

One way to represent a POMDP problem is to use the belief space (information state space), i.e., the set of probability distributions over the state space. For URBP with the set of transition probability matrices P the belief space is $\Psi = \{\psi : \psi^T \in \mathbb{R}^{|S|}, \psi_x \geq 0, \forall x \in S, \sum_{x \in S} \psi_x = 1\}$ which is the unit simplex in $\mathbb{R}^{|S|}$. Let ψ_0 denote the initial belief and ψ_t denote the belief at time t. $V_P(\psi, y, u) = \psi Q_P(y|u)\mathbf{1}$ is the probablity that y will be observed given the belief is ψ and action u is taken, $T_P(\psi, y, u) = \psi Q_P(y|u) / V_P(\psi, y, u)$ is the next belief given that action u is taken at belief ψ and y is observed, where $\mathbf{1}$ is the $|S|$ dimensional column vector of 1's. Let Γ be the set of admissable policies, i.e., any policy

for which action at t is a function of ψ_0 and z^t. The AROE is

$$\begin{aligned} & g + h(\psi) \\ = & \max_{u \in U} \{\bar{r}(\psi, u) + \sum_{y \in S_u} V_P(\psi, y, u) h(T_P(\psi, y, u))\}, \end{aligned} \quad (4)$$

where $\bar{r}(\psi, u) = (\psi \bullet r(u)) = \sum_{x_u \in S_u} x_u \phi_{u,x_u}(\psi)$ is the expected reward of action u at belief ψ, $r(u) = (r(x,u))_{x \in S}$ and $r(x,u) = x_u$ is the reward when arm u is chosen at state x. We have the following assumption.

Assumption 1: $p_{k,ij} > 0, \forall k \in M, i, j \in S_k$.

Under this assumption existence of a bounded, convex continuous solution to the AROE is guaranteed.

Lemma 3: Let $h_- = h - \inf_{\psi \in \Psi}(h(\psi))$, $h_+ = h - \sup_{\psi \in \Psi}(h(\psi))$,

$$h_{T,P}(\psi) = \sup_{\gamma \in \Gamma} \left(E^P_{\psi,\gamma} \left[\sum_{t=1}^{T} r^{\gamma}(t) \right] \right)$$

Under Assumption 1 the following holds:
(i) [1] There exists a finite constant g_P and a bounded convex continuous function $h_P : \Psi \to \mathbb{R}$ which is a solution to (4).
(ii) $h_{P-}(\psi) \leq h_{T,P}(\psi) - T g_P \leq h_{P+}(\psi), \forall \psi \in \Psi$.
(iii) $h_{T,P}(\psi) = T g_P + h_P(\psi) + O(1)$ as $T \to \infty$.

IV. Countable Representation of the Information State

We can represent the information state at time t as $(\boldsymbol{s}^t, \boldsymbol{\tau}^t) = ((s^t_1, \ldots, s^t_m), (\tau^t_1, \ldots, \tau^t_m))$, where s^t_k and τ^t_k are the last observed state from arm k and time from the last observation of arm k to t respectively. This representation requires that all arms are sampled at least once, thus we assume that initially the user samples from all the arms once even before the adaptive learning begins. The contribution of this to the regret is at most $m r_{\max}$. Thus, we assume that the user always starts with some initial belief $(\boldsymbol{s}^0, \boldsymbol{\tau}^0)$. This representation of information state will correspond to different points in Ψ under different sets of transition probability matrices. Thus with an abuse of notation we use $\psi_P((\boldsymbol{s}^t, \boldsymbol{\tau}^t)) = \psi_t \in \Psi$ to represent an element of the countable representation of the information state under P at time t. Let $\Psi_C(P)$ be the set of points on Ψ corresponding to the countable representation of the information state under P. Let $O(\psi; P)$, $O((\boldsymbol{s}, \boldsymbol{\tau}); P)$ denote the set of optimal actions at belief ψ, $\psi_P((\boldsymbol{s}, \boldsymbol{\tau}))$ respectively.

Since the user does not know P, at time t he has an estimate $\hat{P}^t = (\hat{P}^t_1, \ldots, \hat{P}^t_m)$ based on his past observations and actions. Then the estimated belief according to $\hat{P}^t$ is $\hat{\psi}_t = \Omega_{\hat{P}^t, (\boldsymbol{s}^0, \boldsymbol{\tau}^0)_0}(u_0, y_1, \ldots, u_{t-1}, y_t) = \bar{\Omega}_{\hat{P}^t}((\boldsymbol{s}^t, \boldsymbol{\tau}^t))$ for appropriate functions Ω and $\bar{\Omega}$. Even when the user knows the optimal policy for the infinite horizon average reward problem, he may not be able to play optimally because he does not know the exact belief ψ^t at time t. In this case, in order for the user to play optimally, there should be an $\epsilon > 0$ such that if $||\psi_t - \hat{\psi}_t||_1 < \epsilon$, the set of actions that are optimal in $\hat{\psi}_t$ should be a subset of the set of actions that are optimal in ψ_t. We will state an assumption under which this propery will hold. We claim that for an arbitrary set P and S this assumption will generally be satisfied, but a charactezation of conditions on P and S for this property to hold is an open problem for the URBP.

Let τ_0 denote a mixing time. Based on τ_0 let $\mathcal{G}(\tau_0, p)$ be the finite partition of $\Psi_C(P)$ into sets $G_{i_1,\ldots,i_m}$ such that $i_k = \tau_0$ or $i_k = (s_k, \tau_k), \tau_k < \tau_0, s_k \in S_k$. Let $\boldsymbol{s'}(G_{i_1,\ldots,i_m}) = \{s_k : i_k \neq \tau_0\}$, $\boldsymbol{\tau'}(G_{i_1,\ldots,i_m}) = \{\tau_k : i_k \neq \tau_0\}$ and

$$\begin{aligned} \mathcal{M}(G_{i_1,\ldots,i_m}) &= \{k : i_k = \tau_0\}, \\ \bar{\mathcal{M}}(G_{i_1,\ldots,i_m}) &= M - \mathcal{M}(G_{i_1,\ldots,i_m}). \end{aligned}$$

Then,

$$G_{i_1,\ldots,i_m} = \{(\boldsymbol{s}, \boldsymbol{\tau}) \in \Psi_C(P) : (\boldsymbol{s}_{\bar{\mathcal{M}}(G_{i_1,\ldots,i_m})} = \boldsymbol{s'},$$

$$\boldsymbol{\tau}_{\bar{\mathcal{M}}(G_{i_1,\ldots,i_m})} = \boldsymbol{\tau'}), s_k \in S_k, \tau_k \geq \tau_0, \forall k \in \mathcal{M}(G_{i_1,\ldots,i_m})\}.$$

We have the following assumption.

Assumption 2: There exists $\tau_0 \in \mathbb{N}$ such that: (i) Every $G \in \mathcal{G}(\tau_0, P)$ which contains infinitely many elements has a suboptimality gap, i.e., the minimum diffirence between the right hand sides of the average cost optimality equation under the optimal action and a suboptimal action, $\delta > 0$, for the information state which is the stationary distribution for G, i.e., $\tau_k = \infty$ for $k \in \mathcal{M}(G)$. (ii) Every $G \in \mathcal{G}(\tau_0, P)$ which contains only one element has a unique optimal action.

For a set $A \in \psi$ let $A(\epsilon)$ be the ϵ *extension* of that set, i.e., $A(\epsilon) = \{\psi \in \Psi : \psi \in A$ or $d_1(\psi, A) < \epsilon\}$, where $d_1(\psi, A)$ is the minimum l_1 distance between ψ and any element of A

Lemma 4: Let τ_0 be the minimum mixing time such that Assumption 2 holds. Let L be the total number of groups under τ_0. Reindex the groups so we have $G_1, \ldots G_L$. Define J_l to be the ϵ' extension of the convex hull of the group G_l. Then $\exists \epsilon' > 0$ such that for all $\psi \in J_l$ a unique action is optimal.

Proof: Let $\xi_{max}(\tau_0)$ be the maximum l_1 distance between any two elements of any group $G \in \mathcal{G}(\tau_0, p)$. We can find a τ_0 such that $\xi_{max}(\tau_0)$ is small enough so by the continuity of the function h_P, the suboptimality gap for any element of any group that contains infinitely many elements is greater than δ', where $0 < \delta' \leq \delta$). Similary by Assumption 2 and using the continuity of h_P for any group $G \in \mathcal{G}(\tau_0, P)$ we can find an $\epsilon' > 0$ such that the suboptimality gap for any belief ψ contained in the ϵ' extension of the convex hull of the points in G has a suboptimality gap δ'' such that $0 < \delta'' \leq \delta$, and these ϵ' extensions will not intersect each other. ■

V. An Upper Bound for Regret

For any admissable policy γ, the regret with respect to the optimal N horizon policy is given by

$$E^P_{\psi_0,\gamma} \left[\sum_{t=1}^{N} r^{\gamma}(t) \right] - \sup_{\gamma' \in \Gamma} \left(E^P_{\psi_0,\gamma'} \left[\sum_{t=1}^{N} r^{\gamma'}(t) \right] \right).$$

First we will derive the regret with respect to the optimal policy as a function of the number of suboptimal plays. Before proceeding we will define expressions to compactly represent the right hand side of the AROE. Let

$$\begin{aligned}\mathcal{L}(\psi,u,h) &= \bar{r}(\psi,u)+<V(\psi,.,u),h>\\ \mathcal{L}^*(\psi,P) &= \max_{u\in U}\mathcal{L}(\psi,u,h_P).\end{aligned}$$

Let

$$\Delta(\psi,u;P)=\mathcal{L}^*(\psi,P)-\mathcal{L}(\psi,u,h) \tag{5}$$

denote the degree of suboptimality of action u at information state ψ and when the set of transition probability matrices is P. From Proposition 1 of [14] we have for all $\gamma\in\Gamma$

$$R_N^\gamma(\psi_0;P)=\sum_{t=0}^{N-1}E^P_{\psi_0,\gamma}[\Delta(\psi_t,U_t;P)] \tag{6}$$

We assume that initially all the arms are sampled once thus the initial belief is $\psi_0=\psi_P((\boldsymbol{s}^0,\boldsymbol{\tau}^0))$. Let ξ be the supremum over ϵ's such that Lemma 4 holds. Let $J_1,\ldots J_L$ be the sets in Lemma 4 formed by ξ.

Thus at any time t, the belief ψ^t will be in one of the sets $J_1,\ldots,J_L$. Let

Let

$$\bar{\Delta}(J_l,u;P)=\sup_{\psi\in J_l}\Delta(\psi_t,u;P)$$

Note that if $U_t\in O(\psi_t;P)$ then $\Delta(\psi_t,U_t;P)$=0, else $U_t\notin O(\psi_t;P)$ then $\Delta(\psi_t,U_t;P)<\bar{\Delta}(J_l,U_t;P)$ w.p.1. Thus we have

$$\begin{aligned}&R_N^\gamma(\psi_0;P)\\ \leq\ &\sum_{t=0}^{N-1}E^P_{\psi_0,\gamma}[\sum_{l=1}^{L}\sum_{u\notin O(J_l;P)}I(\psi_t\in J_l,U_t=u)\bar{\Delta}(J_l,u;P)]\\ =\ &\sum_{l=1}^{L}\sum_{u\notin O(J_l;P)}E^P_{\psi_0,\gamma}[\sum_{t=0}^{N-1}I(\psi_t\in J_l,U_t=u)]\bar{\Delta}(J_l,u;P)\\ =\ &\sum_{l=1}^{L}\sum_{u\notin O(J_l;P)}E^P_{\psi_0,\gamma}[T_N(J_l,u)]\bar{\Delta}(J_l,u;P)\end{aligned} \tag{7}$$

Then we will upper bound $T_N(J_l,u)$ for suboptimal actions by a sum of expressions which we will upper bound individually. Let

$$\begin{aligned}D_{1,1}(N,\epsilon,J_l,u) &= \sum_{t=0}^{N-1}I(\hat{\psi}_t\in J_l,U_t=u,E_t,F_t,\\ &\quad \mathcal{I}(\hat{\psi}_t,u)\geq\mathcal{L}^*(\hat{\psi}_t,P)-2\epsilon)\\ D_{1,2}(N,\epsilon,J_l,u) &= \sum_{t=0}^{N-1}I(\hat{\psi}_t\in J_l,U_t=u,E_t,F_t,\\ &\quad \mathcal{I}(\hat{\psi}_t,u)<\mathcal{L}^*(\hat{\psi}_t,P)-2\epsilon)\\ D_{1,3}(N,\epsilon) &= \sum_{t=0}^{N-1}I(E_t,F_t^C)\\ D_1(N,\epsilon,J_l,u) &= D_{1,1}(N,\epsilon,J_l,u)+D_{1,2}(N,\epsilon,J_l,u)\\ &+ D_{1,3}(N,\epsilon)\end{aligned}$$

$$\begin{aligned}D_{2,1}(N,\epsilon) &= \sum_{t=0}^{N-1}I(||\psi_t-\hat{\psi}_t||_1>\epsilon,E_t)\\ D_{2,2}(N,\epsilon,J_l) &= \sum_{t=0}^{N-1}I(||\psi_t-Bd(J_l)||_1\leq\epsilon,\\ &\quad I(||\psi_t-\hat{\psi}_t||_1\leq\epsilon,\psi_t\in J_l,E_t)\\ D_2(N,\epsilon,J_l) &= D_{2,1}(N,\epsilon)+D_{2,2}(N,\epsilon,J_l),\end{aligned}$$

where $Bd(J_l)$ is the boundary of J_l.

Lemma 5: For any P satisfying Assumption 2

$$\begin{aligned}&E^P_{\psi_0,\gamma}[T_N(J_l,u)]\leq E^P_{\psi_0,\gamma}[D_1(N,\epsilon,J_l,u)]\\ &+E^P_{\psi_0,\gamma}[D_2(N,\epsilon,J_l)]+E^P_{\psi_0,\gamma}[\sum_{t=0}^{N-1}I(E^C(t))]\end{aligned} \tag{8}$$

Proof:

$$\begin{aligned}&T_N(J_l,u)=\sum_{t=0}^{N-1}(I(\psi_t\in J_l,U_t=u,E_t)\\ +\ &I(\psi_t\in J_l,U_t=u,E_t^C))\\ \leq\ &\sum_{t=0}^{N-1}I(\psi_t\in J_l,\hat{\psi}_t\in J_l,U_t=u,E_t)\\ +\ &\sum_{t=0}^{N-1}I(\psi_t\in J_l,\hat{\psi}_t\notin J_l,U_t=u,E_t)+\sum_{t=0}^{N-1}I(E_t^C)\\ \leq\ &\sum_{t=0}^{N-1}I(\hat{\psi}_t\in J_l,U_t=u,E_t)\\ +\ &\sum_{t=0}^{N-1}I(\psi_t\in J_l,\hat{\psi}_t\notin J_l,E_t)+\sum_{t=0}^{N-1}I(E_t^C)\\ \leq\ &D_{1,1}(N,\epsilon,J_l,u)+D_{1,2}(N,\epsilon,J_l,u)+D_{1,3}(N,\epsilon)\\ +\ &D_{2,1}(N,\epsilon)+D_{2,2}(N,\epsilon,J_l)+\sum_{t=0}^{N-1}I(E_t^C)\end{aligned}$$

The result follows from taking the expectation of both sides. ■

VI. An Adaptive Learning Algorithm (ALA)

The Adaptive Learning Algorithm (ALA) given in Figure 1 consists of exploration and exploitation phases. The exploration serves the purpose of accuretely estimating the transition probabilities. To accuretely estimate the transition probability vectors from each state $i\in S_k$ of each arm k, we need to take at least logarithmic number of samples. In order to do this we need to first observe state i of arm k, then observe the next state so we can update the estimated transition probabilities $\bar{p}_{k,ij}$, $j\in S_k$. However, we need the estimates to form a probability distribution. Thus instead of estimates $\bar{p}_{k,ij}$, we use the normalized estimates $\hat{p}_{k,ij}$. If all the states of all the arms are logarithmically sampled by the way descibed above, then ALA will be in the exploitation phase. If ALA is in the exploitation phase at time t, first it computes $\hat{\psi}_t$, the estimated belief at time t, using the set of estimated transition probability

Adaptive Learning Algorithm

Initialize: set $a > 0$, $t = 0$, $N_u = 0, N^k(i,j) = 0, C^k(i) = 0, \forall k \in M, i, j \in S_k$. Then play each arm once so the initial information state can be represented as an element of countable form $(\boldsymbol{s}, \boldsymbol{\tau})_0$.
while $t \geq 0$ **do**
  $\bar{p}_{k,ij} = \frac{1I(N^k(i,j)=0)+N^k(i,j)}{|S_k|I(C^k(i)=0)+C^k(i)}$
  $\hat{p}_{k,ij} = \frac{\bar{p}_{k,ij}}{\sum_{l \in S_k} \bar{p}_{k,il}}$
  $W = \{(k,i), k \in M, i \in S_k : C^k(i) < a \log t\}$.
  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
    Let $\hat{\psi}_t = \Omega_{\hat{P}^t,(\boldsymbol{s},\boldsymbol{\tau})_0}(u_0, y_1, \ldots, u_{t-1}, y_t)$ be the estimate of the information state at time t based on the transition probability estimates $\hat{P}^t$ and history up to time t.
    solve $\hat{g}_t + \hat{h}_t(\psi) = \max_{u \in U}\{\bar{r}(\psi, u) + \sum_{y \in S_u} V(\psi, y, u)\hat{h}_t(T_{\hat{P}^t}(\psi, y, u))\}, \forall \psi \in \Psi$.
    compute the indices of all actions in current information state:
    $\forall u \in U$, $\mathcal{I}_t(\hat{\psi}^t, u) = \sup_{\tilde{P}_u \in \Xi_u}\{\bar{r}(\hat{\psi}_t, u) + \sum_{y \in S_u} V(\hat{\psi}_t, y, u)\hat{h}_t(T_{\hat{P}^t_{-u}, \tilde{P}_u}(\hat{\psi}_t, y, u))\}$ such that $\left\|\hat{P}^t - \tilde{P}\right\|_1 \leq \sqrt{\frac{2 \log t}{N_t(u)}}$.
    Let u^* be the arm with the highest index. (arbitrarily select one if there is more than one arm with the highest index)
    $u(t) = u^*$.
  end if
  $N_{u(t)} = N_{u(t)} + 1$
  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. 1. pseudocode for the Adaptive Learning Algorithm (ALA)

matrices $\hat{P}^t$. Then, it solves the average reward optimality equation using $\hat{P}^t$ for which the solution is given by g_t and $\hat{h}_t$. We assume that the user can compute the solution at every time step, independent of the complexity of the problem. This solution is used to compute the indices $\mathcal{I}_t(\hat{\psi}^t, u)$ for each action $u \in U$ at estimated belief $\hat{\psi}_t$. $\mathcal{I}_t(\hat{\psi}^t, u)$ represents the advantage of choosing action u starting from information state $\hat{\psi}^t$, i.e, the sum of gain and bias, inflated by the uncertainty about the transition probability estimates based on the number of times action u is chosen. After computing the indices for each action, ALA selects the action with the highest index. In case of a tie, ALA arbitrarily selects one of the actions with the highest index. Note that it is possible to update the state transition probabilities even in the exploitation phase given that the arm selected at times $t-1$ and t are the same. Thus even though worst-case exploration rate is logarithmic, in general the number of explorations needed may be less than that. In the next section we will denote the policy corresponding to ALA by γ^A.

VII. Analysis of the Regret of ALA

In this section we will show that when a is sufficiently large, i.e., $a \geq C(P)$, where $C(P)$ is a constant that depends on P, then the regret due to explorations will be logarithmic in time, while the regret due to all other terms are finite, independent of t. Note that since the user does not know P, he cannot know how large he should chose a. For simplicity we assume that the user starts with an a that is large enough without knowing $C(P)$. However, the user can choose $a = a(t)$, a positive increasing function over time such that $\lim_{t \to \infty} a(t) = \infty$, which will guarantee that after some t_0, $a(t) \geq C(P)$ for $t \geq t_0$. In this case it can be shown that the regret at time N is in the order of $a(N) \log N$.

Let E_t be the event that ALA exploits at time t, $F_t = \{\left\|\hat{h}_t - h_P\right\|_\infty \leq \epsilon\}$ and $C_t^k(i)$ be the number of times state i of arm k is observed as the first state in two continuous plays of arm k up to time t. Following lemma will be frequently used in the proofs.

Lemma 6:

$$P\left(|\hat{p}^t_{k,i_k j_k} - p_{k,i_k j_k}| > \epsilon, E_t\right) \leq \frac{1}{(t+1)^2},$$

for all t, $i_k, j_k \in S_k$, $k \in M$, for $a \geq C_P(\epsilon)$.

Proof: By using a Chernoff-Hoeffding bound. ∎

A. Bounding the Expected Number of Explorations

Lemma 7:

$$E^P_{\psi_0,\gamma^A}\left[\sum_{t=0}^{N-1} I(E_t^C)\right] \leq (\sum_{k=1}^{m} |S_k|) a \log N (1 + T_{\max}), \quad (9)$$

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:

$$\sum_{t=0}^{N-1} I(E_t^C) \leq \sum_{k=1}^{m} \sum_{i \in S_k} \sum_{t=0}^{N-1} I(C_t^k(i) \leq a \log t)$$

$$= \sum_{k=1}^{m} \sum_{i \in S_k} \sum_{t=0}^{N-1} I(C_t^k(i) \leq a \log t, C_{t+1}^k(i) \neq C_t^k(i))$$

$$+ \sum_{k=1}^{m} \sum_{i \in S_k} \sum_{t=0}^{N-1} I(C_t^k(i) \leq a \log t, C_{t+1}^k(i) = C_t^k(i))$$

Taking expectation,

$$E^P_{\psi_0,\gamma^A}\left[\sum_{t=0}^{N-1} I(E_t^C)\right] \leq \sum_{k=1}^{m}\sum_{i\in S_k}(a\log N + a\log NT_{\max})$$

■

B. Bounding $E^P_{\psi_0,\gamma^A}[D_2(N,\epsilon,J_l)]$

Lemma 8: for $a \geq C_P(\epsilon/(mS^2_{\max}|S_1|\dots|S_m|C_1(P,\boldsymbol{\tau})))$ we have

$$E^P_{\psi_0,\gamma^A}[D_{2,1}(N,\epsilon)] \leq 2mS^2_{\max}\beta, \tag{10}$$

where $C_1(P) = \max_{\boldsymbol{\tau}} C_1(P,\boldsymbol{\tau})$, $C_1(P,\boldsymbol{\tau}) = \max_{k\in M} C_1(P_k,\tau_k)$ and $C_1(P_k,\tau_k)$ is given in Lemma 2.

Proof:

$$\begin{aligned}
|(\hat{\psi}_t)_x - (\psi_t)_x| &= \left|\prod_{k=1}^{m}\left((\hat{P}_k^t)^{\tau_k}e^k_{s_k}\right)_{x_k} - \prod_{k=1}^{m}\left(P_k^{\tau_k}e^k_{s_k}\right)_{x_k}\right| \\
&\leq \sum_{k=1}^{m}\left|\left((\hat{P}_k^t)^{\tau_k}e^k_{s_k}\right)_{x_k} - \left(P_k^{\tau_k}e^k_{s_k}\right)_{x_k}\right| \\
&\leq \sum_{k=1}^{m}\left\|(\hat{P}_k^t)^{\tau_k}e^k_{s_k} - P_k^{\tau_k}e^k_{s_k}\right\|_1 \\
&\leq C_1(P,\boldsymbol{\tau})\sum_{k=1}^{m}\left\|\hat{P}_k^t - P_k\right\|_1, \qquad (11)
\end{aligned}$$

where last inequality follows from Lemma 2. By (11)

$$\left\|\hat{\psi}_t - \psi_t\right\|_1 \leq |S_1|\dots|S_m|C_1(P,\boldsymbol{\tau})\sum_{k=1}^{m}\left\|\hat{P}_k^t - P_k\right\|_1$$

Thus we have

$$\begin{aligned}
& P\left(\left\|\hat{\psi}_t - \psi_t\right\|_1 > \epsilon, E_t\right) \\
\leq\ & P\left(\sum_{k=1}^{m}\left\|\hat{P}_k^t - P_k\right\|_1 > \epsilon/(|S_1|\dots|S_m|C_1(P,\boldsymbol{\tau})), E_t\right) \\
\leq\ & \sum_{k=1}^{m} P\left(\left\|\hat{P}_k^t - P_k\right\|_1 > \epsilon/(m|S_1|\dots|S_m|C_1(P,\boldsymbol{\tau})), E_t\right) \\
\leq\ & \sum_{k=1}^{m}\sum_{(i_k,j_k)\in S_k\times S_k} P\left(|\hat{p}^t_{k,i_kj_k} - p_{k,i_kj_k}| > \right. \\
& \left. \frac{\epsilon}{(mS^2_{\max}|S_1|\dots|S_m|C_1(P,\boldsymbol{\tau}))}, E_t\right) \\
\leq\ & 2mS^2_{\max}\frac{1}{(t+1)^2},
\end{aligned}$$

where last inequality follows from Lemma 6. Then,

$$\begin{aligned}
E^P_{\psi_0,\gamma^A}[D_{2,1}(N,\epsilon)] &= \sum_{t=0}^{N-1} P_{\psi_0,\gamma^A}\left(\left\|\psi_t - \hat{\psi}_t\right\|_1 > \epsilon, E_t\right) \\
&\leq 2mS^2_{\max}\beta
\end{aligned}$$

■

Next we will bound $E^P_{\psi_0,\gamma^A}[D_{2,2}(N,\epsilon,J_l)]$.

Lemma 9: Let τ_0 be such that Assumption 2 holds. Then for $\epsilon < \xi/2$, $E^P_{\psi_0,\gamma^A}[D_{2,2}(N,\epsilon,J_l)] = 0, l = 1,\dots,L$.

Proof: By Lemma 4, any $\psi_t \in J_l$ is at least ξ away from the boundary of J_l. Thus given $\hat{\psi}_t$ is at most ϵ away from ψ_t, it is at least $\xi/2$ away from the boundary of J_l. ■

C. Bounding $E^P_{\psi_0,\gamma^A}[D_1(N,\epsilon,J_l,u)]$

First we will upper bound $E^P_{\psi_0,\gamma^A}[D_{1,1}(N,\epsilon,J_l,u)]$. Let Ξ_k be the set of $S_k \times S_k$ stochastic matrices, $\Xi = (\Xi_1,\dots,\Xi_m)$. Define the following function:

$$\begin{aligned}
&\text{MakeOpt}\ (\psi,u;P,\epsilon) := \left\{\tilde{P} = (\tilde{P}_1,\dots,\tilde{P}_m) : \tilde{P}_k \in \Xi_k,\right. \\
&\left.\mathcal{L}(\psi,u,h^*_P(T_{\tilde{P}}(.))) \geq \mathcal{L}^*(\psi,P) - \epsilon\right\},
\end{aligned}$$

$$J_{\psi,u}(\hat{P};P,\epsilon) := \inf\{\left\|\hat{P} - \tilde{P}\right\|_1^2 : \tilde{P} \in \ \text{MakeOpt}\ (\psi,u;P,\epsilon)\}.$$

By the definition of MakeOpt, for every action u in every information state ψ there exists a new stochastic matrix group such that action u becomes optimal in ψ under h^*_P.

Lemma 10: $J_{\psi,u}(\hat{P};P,\epsilon)$ is continuous in its first argument. Therefore there exists a function $f_{P,\epsilon}$ such that $f_{P,\epsilon}(\delta) > 0$ for $\delta > 0$ and $\lim_{\delta\longrightarrow 0} f_{P,\epsilon}(\delta) = 0$.

Lemma 11: Let $\delta > 0$ be such that and $\delta < J_{\psi,u}(P;P,3\epsilon)/2, u \notin O(\psi;P), \psi \in \Psi$ and $a \geq C_P(f_{P,3\epsilon}(\delta)/(mS^2_{\max})$. Then

$$E^P_{\psi_0,\gamma^A}[D_{1,1}(N,\epsilon,J_l,u)] \leq (2mS^2_{\max} + 4/\delta)\beta \tag{12}$$

Proof: Since any action can be made optimal at any information state the event $\{\mathcal{I}_t(\hat{\psi}_t,u) \geq \mathcal{L}^*(\hat{\psi}_t,P) - 2\epsilon\}$ is equivalent to

$$\begin{aligned}
&\exists\tilde{P}\in\Xi : \left(\left\|\hat{P}_t - \tilde{P}\right\|_1^2 \leq \frac{2\log t}{N_t(u)}\right), (\bar{r}(\hat{\psi}_t,u) \\
&+ (V(\hat{\psi}_t,.,u) \bullet \hat{h}_t(T_{\tilde{P}}(\hat{\psi}_t,.,u))) \geq \mathcal{L}^*(\hat{\psi}_t,P) - 2\epsilon) \qquad (13)
\end{aligned}$$

On the event F_t we have

$$\left|\sum_{y\in S_u} V(\hat{\psi}_t,y,u)(\hat{h}_t(T_{\tilde{P}}(\hat{\psi}_t,y,u)) - h^*_P(T_{\tilde{P}}(\hat{\psi}_t,y,u))\right| \leq \epsilon,$$

$$\forall u \in U, \tilde{P} \in \Xi. \tag{14}$$

Thus (13) implies

$$\begin{aligned}
&\exists\tilde{P}\in\Xi : \left(\left\|\hat{P}_t - \tilde{P}\right\|_1^2 \leq \frac{2\log t}{N_t(u)}\right), (\bar{r}(\hat{\psi}_t,u) + \\
&(V(\hat{\psi}_t,.,u) \bullet h^*_P(T_{\tilde{P}}(\hat{\psi}_t,.,u))) \geq \mathcal{L}^*(\hat{\psi}_t,P) - 3\epsilon) \qquad (15)
\end{aligned}$$

From the definition of $J_{\psi,u}(\hat{P};P,\epsilon)$ (15) implies

$$J_{\psi,u}(\hat{P}_t;P,3\epsilon) \leq \frac{2\log t}{N_t(u)}.$$

Thus we have

$D_{1,1}(N,\epsilon,J_l,u)$

$$\leq \sum_{t=0}^{N-1} I\left(\hat{\psi}_t \in J_l, U_t = u, J_{\psi,u}(\hat{P}_t; P, 3\epsilon) \leq \frac{2\log t}{N_t(u)}, E_t\right)$$

$$\leq \sum_{t=0}^{N-1} I\left(\hat{\psi}_t \in J_l, U_t = u, E_t, J_{\hat{\psi}_t,u}(P; P, 3\epsilon) \leq \frac{2\log t}{N_t(u)} + \delta\right) \quad (16)$$

$$+ \sum_{t=0}^{N-1} I\left(\hat{\psi}_t \in J_l, U_t = u, E_t, J_{\hat{\psi}_t,u}(P; P, 3\epsilon) > J_{\hat{\psi}_t,u}(\hat{P}_t; P, 3\epsilon) + \delta\right) \quad (17)$$

Note that (16) is less than or equal to

$$\frac{4\log N}{\delta}. \quad (18)$$

By Lemma 10 $J_{\hat{\psi}_t,u}(P; P, 3\epsilon) > J_{\hat{\psi}_t,u}(\hat{P}_t; P, 3\epsilon) + \delta$ implies $\left\|\hat{P}^t - P\right\|_1 > f_{P,3\epsilon}(\delta)$. Thus (17) us upper bounded by

$$\sum_{t=0}^{N-1} I\left(\left\|\hat{P}^t - P\right\|_1 > f_{P,3\epsilon}(\delta), E_t\right)$$

Taking expectation we have

$$\sum_{t=0}^{N-1} P\left(\left\|\hat{P}^t - P\right\|_1 > f_{P,3\epsilon}(\delta), E_t\right)$$

$$\leq \sum_{t=0}^{N-1}\sum_{k=1}^{m}\sum_{(i_k,j_k)\in S_k\times S_k} P\left(|\hat{p}^t_{k,i_kj_k} - p_{k,i_kj_k}| \geq \frac{f_{P,3\epsilon}(\delta)}{mS^2_{\max}}\right)$$

$$\leq mS^2_{\max}\sum_{t=0}^{N-1}\frac{1}{(t+1)^2} \quad (19)$$

Combining (18) and (19) we have

$$E^P_{\psi_0,\gamma^A}[D_{1,1}(N,\epsilon,J_l,u)] \leq (mS^2_{\max} + 4/\delta)\beta$$

■

Lemma 12: For a large enough we have $E^P_{\psi_0,\gamma^A} D_{1,2}(N,\epsilon,J_l,u) \leq 2mS^2_{\max}\beta$

Proof: If suboptimal action u is chosen at information state $\hat{\psi}_t$ this means that for the optimal action $u^* \in O(\hat{\psi}_t; P)$

$$\mathcal{I}_t(\hat{\psi}_t, u^*) \leq \mathcal{I}_t(\hat{\psi}_t, u) < \mathcal{L}^*(\hat{\psi}_t; P) - 2\epsilon$$

This implies

$$\forall \tilde{P} \in \Xi, \left\|\tilde{P} - \hat{P}^t\right\|_1 \leq \sqrt{\frac{2\log t}{N_t(u)}} \Rightarrow$$
$$(V(\hat{\psi}_t, ., u^*) \bullet \hat{h}_t(T_{\tilde{P}}(\hat{\psi}_t, ., u^*)))$$
$$< (V(\hat{\psi}_t, ., u^*) \bullet h^*_P(T_P(\hat{\psi}_t, ., u^*))) - 2\epsilon$$

Since on F_t (14) holds,

$$\{\mathcal{I}_t(\hat{\psi}_t, u^*) \leq \mathcal{L}^*(\hat{\psi}_t; P) - 2\epsilon\}$$

$$\subset \left\{\forall \tilde{P} \in \Xi, \left\|\tilde{P} - \hat{P}^t\right\|_1 \leq \sqrt{\frac{2\log t}{N_t(u)}} \Rightarrow (V(\hat{\psi}_t, ., u^*) \bullet h^*_P(T_{\tilde{P}}(\hat{\psi}_t, ., u^*))) < (V(\hat{\psi}_t, ., u^*) \bullet h^*_P(T_P(\hat{\psi}_t, ., u^*))) - \epsilon\right\} \quad (20)$$

But since h^*_P is continious there exists $\delta_1 > 0$ such that the event in (20) implies

$$\left\|T_{\tilde{P}}(\hat{\psi}_t, y, u^*) - T_P(\hat{\psi}_t, y, u^*)\right\|_1 > \delta_1, \forall y \in S_{u^*}$$

Again since $T(\psi, y, u)$ is continuous in P, there exists $\delta_2 > 0$ such that above equation implies

$$\left\|\tilde{P} - P\right\|_1 > \delta_2 \quad (21)$$

Thus

$$\{\mathcal{I}_t(\hat{\psi}_t, u^*) \leq \mathcal{L}^*(\hat{\psi}_t; P) - 2\epsilon\}$$
$$\subset \left\{\forall \tilde{P} \in \Xi, \left\|\tilde{P} - \hat{P}^t\right\|_1 \leq \sqrt{\frac{2\log t}{N_t(u)}} \Rightarrow \left\|\tilde{P} - P\right\|_1 > \delta_2\right\}$$
$$\subset \left\{\left\|\hat{P}^t - P\right\|_1 > \delta_2\right\}$$

Therefore

$$E^P_{\psi_0,\gamma^A}[D_{1,2}(N,\epsilon,J_l,u)] \leq \sum_{t=0}^{N-1} P\left(\left\|\hat{P}^t - P\right\|_1 > \delta_2, E_t\right)$$

$$\leq \sum_{t=0}^{N-1}\sum_{k=1}^{m}\sum_{(i_k,j_k)\in S_k\times S_k} P\left(|\hat{p}^t_{k,i_kj_k} - p_{k,i_kj_k}| \geq \frac{\delta_2}{mS^2_{\max}}, E_t\right)$$

$$\leq 2mS^2_{\max}\sum_{t=0}^{N-1}\frac{1}{(t+1)^2}$$

for $a \geq C_P(\delta_2/(mS^2_{\max}))$. ■

Let $\mathcal{P}_{P,\gamma}$ be the Markov transition kernel induced on $\Psi_C(P)$ by policy $\gamma \in \Gamma$. Let

$$\Gamma'(P) = \{\gamma \in \Gamma : \mathcal{P}_{P,\gamma} \text{ is a uniformly ergodic transition kernel}\}. \quad (22)$$

For $\gamma \in \Gamma'(P)$ let $\mathcal{P}^*_{P,\gamma}$ be the stochastic kernel of the stationary distribution whose each row is the stationary distribution $\pi^*_{P,\gamma}$.

Assumption 3: There exists $\epsilon > 0$ such that any optimal policy for the infinite horizon average cost MAB problem with transition matrices $\hat{P}$ such that $\left\|\hat{P} - P\right\|_1 < \epsilon$ belongs to $\Gamma'(\hat{P}) \cap \Gamma'(P)$.

Since h_P, $\hat{h}_t$ are unique up to a constant we can set them equal to the bias of the optimal policies $\gamma(P)$ and $\gamma(\hat{P}^t)$ under P and $\hat{P}^t$ respectively. Let $h_{P,\gamma}$ be the bias under policy γ and transition matrices P. Then,

$$h_{P,\gamma}(\psi) = \sum_{t=0}^{\infty} E^P_{\psi_0,\gamma}[r(X_t,U_t) - g_{P,\gamma}] \tag{23}$$

Let $\tilde{r}_{P,\gamma} = (r(\psi,\gamma(\psi)))_{\psi\in\Psi_C(P)}$. We have $g_{P,\gamma} = \mathcal{P}^*_{P,\gamma}\tilde{r}_{P,\gamma}$. Using this we can write $h_{P,\gamma}$ as

$$\begin{aligned} h_{P,\gamma} &= \sum_{t=1}^{N} \mathcal{P}^{t-1}_{P,\gamma}\tilde{r}_{P,\gamma} - N g_{P,\gamma} \\ &+ \sum_{t=N+1}^{\infty} (\mathcal{P}^{t-1}_{P,\gamma} - \mathcal{P}^*_{P,\gamma})\tilde{r}_{P,\gamma} \end{aligned} \tag{24}$$

Lemma 13: There exists $\varsigma > 0$ such that if $\left\|P_k - \hat{P}_k\right\|_1 < \varsigma, \forall k \in M$ then $\left\|h_{P,\gamma} - h_{\hat{P},\gamma}\right\|_\infty < \epsilon$, for any $\gamma \in \Gamma'(P) \cap \Gamma'(\hat{P})$.

Lemma 14: There exists $\varsigma > 0$ such that if $\left\|P_k - \hat{P}_k\right\|_1 < \varsigma, \forall k \in M$ then $\left\|h_P - h_{\hat{P}}\right\|_\infty < \epsilon$.

Lemma 15: Let $\varsigma > 0$ be such that Lemma 14 holds. Then for $a \geq C_P(\varsigma/S^2_{\max})$ we have

$$E^P_{\psi_0,\gamma^A}[D_{1,3}(N,\epsilon)] \leq 2mS^2_{\max}\beta. \tag{25}$$

Proof: We have by Lemma 14,

$$\{\left\|P_k - \hat{P}^t_k\right\|_1 < \varsigma, \forall k \in M\} \subset \{\left\|h_P - h^t\right\|_\infty < \epsilon\}.$$

Thus,

$$\{\left\|P_k - \hat{P}^t_k\right\|_1 \geq \varsigma, \text{ for some } k \in M\} \supset \{\left\|h_P - h^t\right\|_\infty \geq \epsilon\}.$$

Then

$$E^P_{\psi_0,\gamma^A} D_{1,3}(N,\epsilon) = E^P_{\psi_0,\gamma^A}\left[\sum_{t=0}^{N-1} I(E_t, F^C_t)\right]$$

$$\leq \sum_{t=0}^{N-1} P(\left\|P_k - \hat{P}^t_k\right\|_1 \geq \varsigma, \text{ for some } k \in M, E_t)$$

$$\leq \sum_{k=1}^{m} \sum_{(i_k,j_k)\in S_k\times S_k} \sum_{t=0}^{N-1} P\left(|p_{k,i_kj_k} - \hat{p}^t_{k,i_kj_k}| > \frac{\varsigma}{S^2_{\max}}, E_t\right)$$

$$\leq 2mS^2_{\max}\beta$$

■

D. Logarithmic regret upper bound

Theorem 1: Under Assumptions 1, 2, 3, for a sufficiently large, $a \geq C(P)$ for any suboptimal action $u \in U$

$$E^P_{\psi_0,\gamma^*}[T_N(G_l,u)] \leq a\log N(1+T_{\max}) + (8mS^2_{\max} + 4/\delta)\beta.$$

Thus

$$\begin{aligned} R^{\gamma^*}_N(\psi_0;P) &\leq (a\log N(1+T_{\max}) + (8mS^2_{\max} + 4/\delta)\beta) \\ &\times \sum_{l=1}^{L} \sum_{u\notin O(J_l;P)} \bar{\Delta}(J_l,u;P). \end{aligned}$$

Proof: The result follows from Lemmas 7, 8, 9, 11, 12, 15 and (7). ■

VIII. Conclusion

In this paper we proved that given the transition probabilities of the arms are positive for any state and under some assumptions on the structure of the optimal policy for the infinite horizon average reward problem, there exists index policies which gives logarithmic regret with respect to the optimal finite horizon policy uniformly in time. Our future research includes finding the conditions on P such that Assumptions 2, 3 hold.

References

[1] L. K. Platzman, "Optimal infinite-horizon undiscounted control of finite probabilistic systems," *SIAM J. Control Optim.*, vol. 18, pp. 362–380, 1980.

[2] S. P. Hsu, D. M. Chuang, and A. Arapostathis, "On the existence of stationary optimal policies for partially observed mdps under the long-run average cost criterion," *Systems and Control Letters*, vol. 55, pp. 165–173, 2006.

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

[4] 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.

[5] 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.

[6] ——, "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.

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

[8] 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.

[9] ——, "Online learning in opportunistic spectrum access: A restless bandit approach," in *30th IEEE International Conference on Computer Communications (INFOCOM)*, April 2011.

[10] ——, "Online learning of rested and restless bandits," *http://arxiv.org/abs/1102.3508v1*.

[11] ——, "Performance and convergence of multi-user online learning," in *2nd International ICST Conference on Game Theory for Networks (GAMENETS)*, April 2011.

[12] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players, http://arxiv.org/abs/0910.2065."

[13] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple players: Learning under competition," in *Proc. of IEEE INFOCOM*, March 2010.

[14] 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.

[15] 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.

[16] A. Y. Mitrophanov, "Senstivity and convergence of uniformly ergodic markov chains," *J. Appl. Prob.*, vol. 42, pp. 1003–1014, 2005.

[17] 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.

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

[19] S. Guha, K. Mungala, and P. Shi, "Approximation algorithms for restless bandit problems," *20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pp. 28–37, 2009.

[20] 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.