

# Indexability and Rollout Policy for Multi-State Partially Observable Restless Bandits

Rahul Meshram

Deptt. of Elect. Comm. Engg.  
IIT Allahabad  
INDIA

Kesav Kaza

Deptt. of Elecl. Engg.  
Polytechnique Montreal  
CANADA

**Abstract**—Restless multi-armed bandits with partially observable states has applications in communication systems, age of information and recommendation systems. In this paper, we study multi-state partially observable restless bandit models. We consider three different models based on information observable to decision maker—1) no information is observable from actions of a bandit 2) perfect information from bandit is observable only for one action on bandit, there is a fixed restart state, i.e., transition occurs from all other states to that state 3) perfect state information is available to decision maker for both actions on a bandit and there are two restart state for two actions. We develop the structural properties. We also show a threshold type policy and indexability for model 2 and 3. We present Monte Carlo (MC) rollout policy. We use it for Whittle index computation in case of model 2. We obtain the concentration bound on value function in terms of horizon length and number of trajectories for MC rollout policy. We derive explicit index formula for model 3. We finally describe Monte Carlo rollout policy for model 1 when it is difficult to show indexability. We demonstrate the numerical examples using myopic policy, Monte Carlo rollout policy and Whittle index policy. We observe that Monte Carlo rollout policy is good competitive policy to myopic.

## I. INTRODUCTION

Restless multi-armed bandits with partially observable states have been recently found applications in online recommendation systems [1], opportunistic communication systems [2]–[4], machine maintenance [5], age of information, [6]. Restless multi-armed bandits (RMABs) are class of sequential decision problem with multiple independent Markov processes which are coupled via number of independent process that are activated simultaneously, [7]. In a partially observable model, states of Markov chains are not observable at time of decision making but signals are observable. The solution of RMAB are computationally challenging and known to be PSPACE Hard problem, [8]. In fact a popular heuristic Whittle index based policy have been studied and it has shown to be asymptotically optimal, [9]. The essential idea of index policy is to decouple the independent Markov processes (arms) by solving relaxed constrained problem with Lagrangian method. Later one need to show indexability for each processes and has to provide computational method for index which maps the state of each process to a real number. The process (arm) with the highest index is played at each time instant.

Most of RMAB problems with partially observable states are studied for two state model with various assumptions on transition probabilities, reward structure and observation

probabilities, [1]–[4], [10]–[12]. Much less attention is given to more than two state model. Multi-state partially observable RMAB has been studied in [6], [13]–[15]. In [13], [14], the optimality of myopic policy is shown under specific model assumption for identical communication channels. In [6], authors have proposed and analyzed greedy policy for age of information problem. In [15], authors have studied a pilot allocation problem in wireless networks over partially observable fading channel with approximation on multi state model. Further, they analyzed index policy and asymptotic optimality is proved. To derive obtain indexability, one require to study a single armed bandit model and it is partially observable Markov decision process (POMDP). The properties of POMDP are derived in [16].

In this paper, we study partially observable RMAB with more than two state model. We consider three different models based on information observable to decision maker. In first model we study with no state is observable for any actions. In second model the decision maker can observe the perfect state for one of the actions. In third model we assume that decision maker observes perfect state for both actions. We obtain structural properties and discuss about indexability for these models. We discuss simulation based MC rollout policy. In first model, indexability is very difficult to obtain and hence use rollout policy. In second model, we show indexability but difficult to derive index, this motivated rollout policy based index computation method. We obtain the concentration bound for rollout policy with threshold type structure. In third model, we show indexability and derive explicit index formula. Finally we illustrate performance of proposed policy using numerical examples.

The paper is organized as follows. We present model descriptions and preliminaries in Section II. The structural properties and indexability are developed in Section III. Monte Carlo rollout policy is discussed in Section IV. Numerical examples and discussion are presented in Section V.

## II. MODEL DESCRIPTION

Consider  $N$  partially observable restless multi-armed bandits, where  $\mathbf{M}_i = \{\mathcal{S}_i, \mathcal{A}_i, \mathcal{P}_i, \mathcal{R}_i, \mathcal{O}_i, \mathcal{Q}_i, \beta\}$ ,  $i = 1, 2, \dots, N$ . Let  $\mathcal{S}_i$  be the state space,  $\mathcal{S}_i = \{1, 2, \dots, n\}$ ,  $\mathcal{A}_i = \{0, 1\}$  is action space,  $\mathcal{P}_i = \{[p_{jk}^a]\}_{\{a \in \mathcal{A}\}}$  is the transition probability matrix and  $p_{jk}^a$  is the transition probability from state  $j$  to  $k$  when action  $a$  is applied. The decision maker (DM) does not observe the state of

systems but makes his decisions based on the information obtained via evolution of states. Based on this observed information, the decision maker selects action  $a_{t,i} \in \mathcal{A}_i$  at time  $t = 1, 2, \dots$ . The state of system  $i$  at time  $t$  is denoted by  $s_{t,i} \in \mathcal{S}$ . A DM receives a real valued reward  $r_i(j, a)$  if  $a_{t,i} = a$  and  $s_{t,i} = j$ . The system  $i$  make transition to state  $s_{t+1,i}$ , and  $p_{jk}^a = \Pr(s_{t+1,i} = j \mid s_{t,i} = i, a_{t,i} = a)$ . A DM perceives one of finite number of messages. Assume that  $\mathcal{O} = \{1, 2, 3, \dots, K\}$  represents the set of messages<sup>1</sup>. If the message  $k \in \mathcal{O}$  is observed with known probability from state  $j$  under action  $a$  for system  $i$  and this is denoted by  $q_{i,jk}^a = \Pr(k \mid s_{t,i} = j, a_{t,i} = a)$ . Thus  $\mathcal{Q}_i = [[q_{i,jk}^a]]_{\{a \in \mathcal{A}\}}$ . The discount parameter is denoted by  $\beta$ . Each bandit evolves in discrete time steps.

An infinite horizon discounted problem with a policy  $\phi$  is given as follows.

$$V_\phi(s) = \mathbb{E}_\phi \left( \sum_{t=0}^{\infty} \sum_{i=1}^N \beta^t r_i(s_{t,i}, a_{t,i}) \right). \quad (1)$$

There is an activation constrained on bandits, i.e.,  $\sum_{i=1}^N a_{t,i} = 1$ . The policy  $\phi : H_t \rightarrow \{1, 2, 3, \dots, N\}$ , where  $H_t$  denotes the history upto time  $t$ , and  $H_t := \{a_1, o_1, \dots, a_{t-1}, o_{t-1}\}$ . The Markov stationary deterministic policy is studied.  $V_\phi(s)$  is the value function for given initial state  $s$ . DM's goal is to choose the strategy  $\phi$  to optimize  $V_\phi(s)$ , subject to constraint  $\sum_{i=1}^N a_{t,i} = 1$ . Thus, the optimal value function is denoted by  $V^*$ . The discounted relaxed constrained problem using Lagrangian method is written as follows.

$$\begin{aligned} V_\phi(s) &= \mathbb{E}_\phi \left( \sum_{i=1}^N \sum_{t=0}^{\infty} \beta^t [r_i(s_{t,i}, a_{t,i}) + W(1 - a_{t,i})] \right). \\ V^*(s) &= \max_{\phi \in \Phi} V_\phi(s). \end{aligned} \quad (2)$$

Here,  $\Phi$  is the space of all Markov stationary deterministic policies.

### A. A single-armed restless bandit and preliminaries

In this section, a single armed bandit with partially observable state is discussed and we remove dependence of arm on  $i$  for notation simplicity. A single armed restless bandit is a special case of partially observable Markov decision processes (POMDPs). We can rewrite problem in (2) for partially observable with belief  $\pi$ . The DM only observes messages (signals) but no state information. The DM maintains initial belief as prior  $\pi \in \Pi(\mathcal{S})$ , where  $\Pi(\mathcal{S}) = \{\pi = (\pi(1), \pi(2), \dots, \pi(n)) \mid \sum_{j=1}^n \pi(j) = 1, 0 \leq \pi(j) \leq 1, \text{ for all } j \in \mathcal{S}\}$  is belief space and  $\pi(j)$  is probability of state being  $j$ , i.e.,  $s = j$ . Based on initial belief  $\pi$ , the value function under policy  $\phi$  is

$$V_\phi(\pi) = \mathbb{E}_\phi \left( \sum_{t=0}^{\infty} \beta^t \left[ \sum_{j=1}^n r(s_t = j, a_t) \pi(j) + W(1 - a_{t,i}) \right] \right).$$

<sup>1</sup>Example is a google news recommendation system, where different messages correspond to actions of a user—like, dislike, watch later etc. The user takes different actions with some probability based on user interest state. This generates reward to RS based on user behavior

The DM optimizes the value function and it is given by

$$V^*(\pi) = \max_{\phi \in \Phi} V_\phi(\pi). \quad (3)$$

From [17], [18], we know that the information observed in the history  $H_t$  is captured in form of belief  $\pi_t \in \Pi(\mathcal{S})$ ,  $\pi_t$  is the Bayesian posterior over states given history

$$\begin{aligned} \pi_t(j) &= \Pr(s_t = j \mid H_t) \\ \pi_t(j) &= \Pr(s_t = j \mid \pi_{t-1}, o_t = k, a_t = a) \\ \pi_t &= (\pi_t(1), \pi_t(2), \dots, \pi_t(n)). \end{aligned}$$

This is shown to be sufficient information which captures all history upto  $t$ . Note that there are two actions are available to a single armed bandit—play or not play. Corresponding to this, there are actions dependent transition probabilities. We study the following models for a single armed bandit based on transition probabilities and information observed from each action.

1) *Model 1*: In this model, a decision maker does not observe state from both actions. This is an example of two action POMDP, where action  $a = 1$  provides a signals and other action  $a = 0$  provides no information to decision maker. For action  $a = 1$ , DM observes a signal  $k$  and the posterior belief is computed and the computations are as follows. Let  $\xi(j, k \mid \pi, a)$  be the probability that the message  $k$  is received from state  $j$  given prior  $\pi_t$  and action  $a$ , and  $\xi(j, k \mid \pi, a) = \sum_{i \in \mathcal{S}} \pi_t(i) p_{i,j}^a q_{i,k}^a$ . Define  $\sigma(k \mid \pi, a)$  is the probability of observing message  $k$  given prior  $\pi_t$  and action  $a$ . It is given by

$$\begin{aligned} \sigma(k \mid \pi_t, a) &= \sum_{j=1}^n \xi(j, k \mid \pi_t, a) \\ &= \sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{S}} \pi_t(i) p_{i,j}^a q_{i,k}^a. \end{aligned}$$

The Bayesian posterior given prior  $\pi_t$  and action  $a$  and signal  $k$  is denoted by  $\Gamma(\pi_t, a, k)$ , and  $\Gamma_j(\pi, a, k) = \frac{\xi(j, k \mid \pi, a)}{\sigma(k \mid \pi, a)}$ ,  $\Gamma(\pi_t, a, k) = (\Gamma_1(\pi_t, a, k), \dots, \Gamma_n(\pi_t, a, k)) \in \Pi(\mathcal{S})$ . Then

$$\pi_{t+1}(l) = \Gamma_l(\pi_t, a_t, o_t = k) = \frac{\sum_{i \in \mathcal{S}} \pi_t(i) p_{i,l}^a q_{i,k}^a}{\sum_{j \in \mathcal{S}} \sum_{i \in \mathcal{S}} \pi_t(i) p_{i,j}^a q_{i,k}^a}.$$

When action  $a = 0$ , no signal is observed and hence the posterior belief  $\pi_{t+1} = \pi_t P^0$ .

Let  $B(\mathcal{S})$  be the set of bounded real valued functions on  $\Pi(\mathcal{S})$ . Define function  $g : \Pi(\mathcal{S}) \times \mathcal{A} \times B(\mathcal{S}) \rightarrow \mathcal{R}$  and we can write  $g(\pi, a, V)$  as function of immediate reward and future value function, thus

$$\begin{aligned} g(\pi, a = 1, V) &= \sum_{j=1}^n \pi(j) r(j, a = 1) + \beta \sum_{k \in \mathcal{O}} \sigma(k \mid \pi, a) \\ &\quad \times V(\Gamma(\pi, a, k)) \\ g(\pi, a = 0, V) &= \sum_{j=1}^n \pi(j) r(j, a = 0) + W + \beta V(\pi P^0) \end{aligned}$$

for  $\pi \in \Pi(\mathcal{S})$ ,  $a \in \mathcal{A}$  and  $V \in B(\mathcal{S})$ .  $P^0$  is the transition probability for not playing arm.

An optimal dynamic programming algorithm is given as follows.

$$V^*(\pi) = \max_{a \in \mathcal{A}} g(\pi, a, V^*). \quad (4)$$

It is difficult to claim indexability for this model and apply index policy. Hence we study MC rollout policy in next section.

2) *Model 2*: In this model, a decision maker takes action  $a = 1$ , it just provide signals but does not provide any perfect information about state. The action  $a = 0$  gives perfect state information. Moreover, the transition occurs to a fixed state  $m$  which is restart state. Then the dynamic program is given as follows.

$$\begin{aligned} g(\pi, a = 1, V) &= \sum_{j=1}^n \pi(j)r(j, a = 1) + \beta \sum_{k \in \mathcal{O}} \sigma(k | \pi, a) \\ &\quad \times V(\Gamma(\pi, a, k)) \\ g(\pi, a = 0, V) &= \sum_{j=1}^n \pi(j)r(j, a = 0) + W + \beta V(e_m) \end{aligned}$$

where  $e_m = [0, 0, \dots, 1, 0, \dots, 0]^T$ , 1 is for state  $m$ . The transition probability matrix of not playing action is  $P^0$ , and  $m$ th column of it is a unit vector, i.e., all elements are 1 and remaining columns are zero vectors. An optimal dynamic programming algorithm is given by

$$V^*(\pi) = \max_{a \in \mathcal{A}} g(\pi, a, V^*). \quad (5)$$

In next section, we show that a bandit is indexable and but it is difficult to obtain closed form expression of Whittle index.

3) *Model 3*: In this model, we further relax assumptions stated in previous models. We assume that state is perfectly observable for both actions. Moreover for action  $a = 1$ , transition from state  $i$  to fixed state  $m_1 \in \mathcal{S}$  occurs with probability 1,  $i = 1, 2, \dots, n$ . Similarly, for action  $a = 0$  a state transition from state  $i$  to a fixed state  $m_2 \in \mathcal{S}$  occurs with probability 1. The dynamic program is

$$\begin{aligned} g(\pi, a = 1, V) &= \sum_{j=1}^n \pi(j)r(j, a = 1) + \beta V(e_{m_1}) \\ g(\pi, a = 0, V) &= \sum_{j=1}^n \pi(j)r(j, a = 0) + W + \beta V(e_{m_2}) \end{aligned}$$

where  $e_{m_1} = [0, 0, \dots, 1, 0, \dots, 0]^T$ , 1 is at position  $m_1$ .  $e_{m_2} = [0, 0, \dots, 1, 0, \dots, 0]^T$ , 1 is at position  $m_2$ . An optimal dynamic programming algorithm is

$$V^*(\pi) = \max_{a \in \mathcal{A}} g(\pi, a, V^*). \quad (6)$$

We will show that a bandit is indexable and even obtain the closed form expression of Whittle index.

### III. STRUCTURAL RESULTS AND INDEXABILITY

In this section we provide structural results, indexability of a restless bandits and derive index formula. We derive two key results—monotonicity of optimal value functions and threshold type policy.

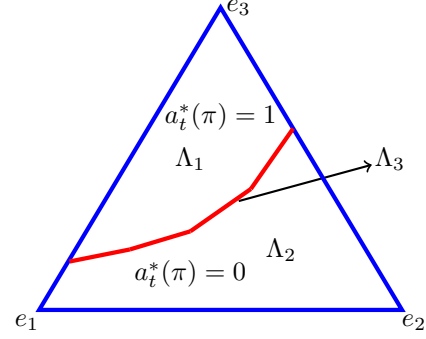


Fig. 1. Threshold type policy illustration

#### A. Structural Properties

*Lemma 1 (Convexity of value function)*: For infinite horizon problem, the optimal value function  $V^*(\pi)$  is convex in  $\pi$  for  $\pi \in \Pi(S)$ .

Proof of this result using induction method, and it uses [19, Lemma 2] to prove convexity of value function. Proof is along lines of [12, Lemma 2]. We use maximum likelihood ratio (MLR) order for comparison of belief  $\pi$ 's MLR order is denoted as  $\pi \geq_r \tilde{\pi}$ . Totally positive order 2 (TP<sub>2</sub>) for comparison of transition probability matrices.

*Lemma 2 (Monotonicity of value function)*: [16]: The optimal value function  $V^*(\pi)$  is monotone in belief  $\pi$ , that is,  $V^*(\pi) \geq V^*(\tilde{\pi})$  whenever  $\pi \geq_r \tilde{\pi}$  for  $\pi, \tilde{\pi} \in \Pi(S)$  under following assumptions.

- reward  $r(j, a)$  is non decreasing in  $j \in \mathcal{S}$  for fixed  $a$ .
- transition probability matrices  $P^1$  and  $P^0$  are TP<sub>2</sub> ordered.
- the observation row vector for arm  $q(j) \geq q(k)$  for  $j \geq i$  and  $i, j \in \mathcal{S}$ .

We sketch the proof. The assumptions stated here preserves monotonicity in belief  $\Gamma$ , and  $\sigma$  whenever there is ordering in prior  $\pi$ , action  $a$  and observation  $k$ . This preserves the ordering in value functions in belief  $\pi$ . Using induction method on dynamic program and monotonicity of value functions in belief, we get the desired result.

We note that the Lemma 1 and 2 holds for all models under different assumptions on model. But threshold policy and indexability holds true only for model 2 and 3.

A threshold type policy provides partition of belief state space  $\Pi(S)$  into three disjoint regions,  $\Lambda_1, \Lambda_2, \Lambda_3 \subseteq \Pi(S)$ , where  $\Lambda_1 = \{\pi \in \Pi(S) : a_t^*(\pi) = 1\}$ ,  $\Lambda_2 = \{\pi \in \Pi(S) : a_t^*(\pi) = 0\}$  and  $\Lambda_3 = \{\pi \in \Pi(S) : a_t^*(\pi) = 1 \text{ and } 0\}$ .  $a_t^*(\pi) \in \{0, 1\}$  is the optimal action for belief  $\pi$  at time step  $t$ . Illustration of this is given in Fig. 1.

*Definition 1 (Threshold type policy)*: The optimal policy is called a threshold type if one of the following holds true.

- 1) The optimal action  $a_t^*(\pi) = 1$  for all  $t$  and all  $\pi \in \Pi(S)$ , that is  $\Lambda_1 = \Pi(S)$  and  $\Lambda_2 = \Lambda_3 = \emptyset$ .
- 2) The optimal action  $a_t^*(\pi) = 0$  for all  $t$  and all  $\pi \in \Pi(S)$ , that is  $\Lambda_2 = \Pi(S)$  and  $\Lambda_1 = \Lambda_3 = \emptyset$ .
- 3) The optimal action  $a_t^*(\pi) = 1$  for all  $\pi \in \Lambda_1$ ,  $a_t^*(\pi) =$

0 for all  $\pi \in \Lambda_2$ , and  $a_i^*(\pi) = 1$  and 0 for all  $\pi \in \Lambda_3$ , that is,  $\Lambda_1, \Lambda_2, \Lambda_3 \neq \emptyset$ . Also  $\Lambda_1 \cap \Lambda_2 \cap \Lambda_3 = \emptyset$ .

We next show a threshold policy result and indexability for model 2 and 3. We make use of same assumption as stated in previous Lemma 2.

*Lemma 3 (Threshold type policy):* In Model 2 and Model 3, the optimal policy is of threshold type.

We provide a sketch of the proof. Define  $f(\pi, V^*) = g(\pi, a = 1, V^*) - g(\pi, a = 0, V^*)$ . We show that  $f(\pi, V^*)$  is non decreasing in  $\pi$ . In these model, not playing action, i.e.,  $a = 0$  implies restart state where transition occurs to a fixed state. Thus the future value function for action  $a = 0$  is constant. From definition of  $f(\pi, V^*)$ , that term gets canceled, hence using Lemma 2, we show that  $f(\pi, V^*)$  is non decreasing in  $\pi$ . This is sufficient for threshold type policy. Detailed proof is given in Appendix.

### B. Indexability and Whittle index

From threshold policy result in Lemma 3, we define

$$U_1(W) := \{\pi \in \Pi(S) : V(\pi, a = 1, W) > V(\pi, a = 0, W)\}$$

$$U_0(W) := \{\pi \in \Pi(S) : V(\pi, a = 1, W) \leq V(\pi, a = 0, W)\}$$

Hence  $U_0(W) = \Lambda_2 \cup \Lambda_3$ .

*Definition 2 (Indexability [7]):* As subsidy  $W$  increases from  $-\infty$  to  $+\infty$ ,  $U_0(W)$  increases from  $\emptyset$  to full set  $\Pi(S)$ .

To show the indexability we require that whenever  $W_2 > W_1$  implies  $U_0(W_1) \subseteq U_0(W_2)$ . We use the following result for indexability.

*Lemma 4:* For  $\pi \in \Pi(S)$  if

$$\left. \frac{\partial V(\pi, 1, W)}{\partial W} \right|_{\pi=\pi_T(W)} < \left. \frac{\partial V(\pi, 0, W)}{\partial W} \right|_{\pi=\pi_T(W)}, \quad (7)$$

and  $\pi_T(W) \in \Lambda_3$ , then  $U_0(W)$  is a monotonically increasing function of  $W$ .

Proof of this lemma is analogous to [12, Lemma 4]. We now present main result.

*Theorem 1 (Indexable):* The single-armed restless hidden Markov bandit is indexable for  $0 < \beta < 1$  and  $W_a \leq W \leq W_b$ .

*Proof:* From Definition 2, we need to show that  $U_0(W_1) \subseteq U_0(W_2)$  whenever  $W_2 > W_1$ . Note that  $V(\pi, a = 1, W) - V(\pi, a = 0, W)$  is decreasing in  $W^2$  for fixed  $\pi, \beta$ . Therefore, equation (7) holds true. Using Lemma 4,  $U_0(W_1) \subseteq U_0(W_2)$  whenever  $W_2 > W_1$  and  $W_1, W_2 \in [W_a, W_b]$ . This completes the proof. ■

We next define the Whittle index.

*Definition 3 (Whittle index [7]):* If an arm is indexable and is in state  $\pi \in \Pi(S)$ , then its Whittle index,  $W(\pi)$ , is  $W(\pi) := \inf_W \{W : V(\pi, 1, W) = V(\pi, 0, W)\}$ .

$W(\pi)$  is a minimum subsidy  $W$  such that the optimal action is not to play the arm at given  $\pi$ . The Whittle index formula requires explicit expression of  $V(\pi, 1, W)$

<sup>2</sup>By induction method, one can show that  $V(\pi, a = 1, W)$  is non decreasing  $W$  and  $V(\pi, a = 0, W)$  is strictly increasing in  $W$  for fixed  $\beta$  and  $\pi$

and  $V(\pi, 0, W)$ . Then we have to equate and solve this for  $W$ . For Model 2, the index formula is not feasible but we will provide approximate index computation algorithm. For Model 3, we obtain closed form expression of index and this is given in next lemma.

*Lemma 5 (Whittle index formula for model 3):* Whittle index for given belief  $\pi$  is computed based on region of  $e_{m_1}$  and  $e_{m_2}$ . We assume that  $m_1 > m_2$ .

- if  $e_{m_1} \in U_1(W)$  and  $e_{m_2} \in U_0(W)$ , then

$$W(\pi) = (1 - \beta) \left( \sum_{j=1}^n [r(j, 1) - r(j, 0)] \pi(j) \right) + \beta [r(m_1, 1) - r(m_2, 0)]$$

- if  $e_{m_1}, e_{m_2} \in U_1(W)$  then

$$W(\pi) = \sum_{j=1}^n [r(j, 1) - r(j, 0)] \pi(j) + \beta [r(m_1, 1) - r(m_2, 1)]$$

- if  $e_{m_1}, e_{m_2} \in U_0(W)$  then

$$W(\pi) = \sum_{j=1}^n [r(j, 1) - r(j, 0)] \pi(j) + \beta [r(m_1, 0) - r(m_2, 0)].$$

Proof of this is given in Appendix. When there is no reward from not playing except subsidy  $W$ , we can have  $r(j, 0) = 0$  for all  $j \in S$ .

## IV. MONTE CARLO ROLLOUT POLICY

We now discuss Monte Carlo rollout policy algorithm for a single-armed bandit in case of Model 2. Algorithm is based on simulations, where initial belief state  $\pi$  and a subsidy  $W$  is given as input. We run multiple-trajectories, and each trajectory consists of (belief state  $\pi$ , action  $a$ , and observed reward  $r$ ) Thus the information obtained from a single trajectory upto horizon length  $H$  is  $\{\pi_{1,l}, a_{1,l}, r_{1,l}, \pi_{2,l}, a_{2,l}, r_{2,l}, \dots, \pi_{H,l}, a_{H,l}, r_{H,l}\}$  under policy  $\phi$ . Here,  $l$  denotes a trajectory. The value estimate of  $k$ th trajectory starting from belief state  $\pi$ , action  $a = 1$  and action  $a = 0$  is

$$Q_{H,l}^\phi(\pi, a, W) = \sum_{h=1}^H \beta^{h-1} r_{h,l}^\phi$$

$$= \sum_{h=1}^H \beta^{h-1} r(\pi_{h,l}, a_{h,l}).$$

Then value estimate for state  $\pi$  and action  $a$  over  $L$  trajectories under policy  $\phi$  is

$$\tilde{Q}_{H,L}^\phi(\pi, a, W) = \frac{1}{L} \sum_{l=1}^L Q_{H,l}^\phi(\pi, a, W).$$

The output of Monte Carlo algorithm is  $\tilde{V}_{\phi,H,L}(\pi, a = 1, W)$  and  $\tilde{V}_{\phi,H,L}(\pi, a = 0, W)$ .

$$\tilde{V}_{\phi,H,L}(\pi, a = 1, W) = r(\pi, a = 1) + \tilde{Q}_{H,L}^\phi(\pi, a = 1, W)$$

$$\begin{aligned}\tilde{V}_{\phi,H,L}(\pi, a = 0, W) &= W + r(\pi, a = 0) + \\ &\quad \tilde{Q}_{H,L}^{\phi}(\pi, a = 1, W)\end{aligned}$$

### A. Index computation for Model 2

We present algorithm for Whittle index computation using Monte Carlo rollout policy. It is described in Algorithm 1. Input is state  $\pi$ , and initialize value  $W$ . We run Monte Carlo rollout policy under threshold policy  $\phi$  for  $W$  and state  $\pi$ . We obtain approximate value functions  $\tilde{V}_{\phi,H,L}(\pi, a = 1, W)$  and  $\tilde{V}_{\phi,H,L}(\pi, a = 0, W)$ . If the difference between these approximate value functions is higher than  $\epsilon > 9$ , then we change  $W$  to new value of  $W$ ; otherwise exit an algorithm with output index =  $W$ . The convergence of this algorithm follows from two-timescales stochastic approximation algorithms, [20, Chapter 6]. In our setting, Monte Carlo rollout policy algorithm runs on faster timescale and the subsidy  $W$  is updated on slower timescale. We use  $\gamma$  as learning rate for  $W$ .

---

#### Algorithm 1: Whittle index computation algorithm for an arm

---

**Input:** State of arm  $\pi$

**Initialize**  $W_{old} = W$ ,  $\epsilon = 0.05$   $\Delta = 1$ , and Stepsize  $\gamma$

**Define:**  $W_{new} = W_{old}$

**While** ( $\Delta > \epsilon$ )

**1. Use Monte Carlo rollout policy**

**Compute:**  $\tilde{V}_{\phi,H,L}(\pi, a = 1, W_{new})$  and  $\tilde{V}_{\phi,H,L}(\pi, a = 0, W_{new})$

**2. Define**

$$\Delta(\pi, W_{new}) = \tilde{V}_{\phi,H,L}(\pi, a = 1, W_{new}) - \tilde{V}_{\phi,H,L}(\pi, a = 0, W_{new})$$

$$W_{old} = W_{new}$$

$$W_{new} = W_{old} + \gamma \Delta(\pi, W_{new})$$

**End**

**3. Output:**  $W(\pi)$

---

We derive following result with Monte Carlo rollout policy assuming there optimal policy exists and it of threshold type, say,  $\phi$ .

*Theorem 2:* We assume that  $r(s, a) \in [0, 1]$ ,  $0 \leq W \leq 1$ . For sufficiently large horizon length  $H$ , there exist number  $\tilde{L}$  such that for all  $L > \tilde{L}$  we have with probability  $1 - \frac{2}{H^2}$

$$\left| V_{\phi}(\pi, a, W) - \tilde{V}_{\phi,H,L}(\pi, a, W) \right| \leq \sqrt{\frac{z^2 \log(H)}{L}}$$

for  $a \in \{0, 1\}$ . Here,  $z = \frac{(1-\beta^H)}{1-\beta}$ .

We discuss the proof idea. We simulate  $L$  number of trajectories which are independent and cumulative reward collected along each trajectory is random. Trajectories are generated using a fixed policy  $\phi$ . We use Hoeffding inequality [21]. The probability of deviation between the infinite horizon discounted value function under policy  $\phi$  and estimated value function obtained using over  $L$  number of

simulated trajectories greater than confidence bound decays exponentially fast. After simplifications we obtain desired result. Detail steps are given in Appendix.

### B. Monte Carlo rollout policy for Model 1

As discussed in earlier section index policy is not applicable to Model 1, however we can use Monte Carlo rollout policy. Here, arm is selected based on state-action value estimate obtained using fixed Rollout policy  $\phi$  that selects an arm at each time step. Note that we are directly applying this policy to RMAB.

Detail of rollout policy is as follows. There are  $L$  trajectories simulated for a fixed horizon length  $H$  using a known transition and reward model. Along each trajectory, a fixed policy  $\phi$  is employed according to which one arm is played at each time step from  $N$  arms. The information obtained from a single trajectory upto horizon length  $H$  is

$$\{\pi_{t,j,l}, a_{t,j,l}, r_{t,j,l}^{\phi}\}_{j=1, t=1}^{N,H} \quad (8)$$

under policy  $\phi$ . Here,  $l$  denotes a trajectory, the belief for arm  $j$  is  $\pi_{t,j,l} \in \Pi(\mathcal{S})$ , action of arm  $j$  is  $a_{t,j,l} \in \mathcal{A}$ , moreover it has constraint  $\sum_{j=1}^N a_{t,j,l} = 1$ .  $r_{t,j,l}^{\phi}$  is reward from arm  $j$  under policy  $\phi$ . The value estimate of trajectory  $l$  starting from belief state  $\pi = (\pi_1, \dots, \pi_N)$ , and  $\pi_j \in \Pi(\mathcal{S})$  for  $N$  arms and initial action  $\alpha \in \{1, 2, \dots, N\}$  and is  $Q_{H,l}^{\phi}(\pi, \alpha) = \sum_{h=1}^H \beta^{h-1} r_{h,l}^{\phi} = \sum_{h=1}^H \beta^{h-1} r(\pi_{h,l}, \alpha_{h,l}, \phi)$ . Then, the value estimate for state  $\pi$  and action  $a$  over  $L$  trajectories under policy  $\phi$  is

$$\tilde{Q}_{H,L}^{\phi}(\pi, \alpha) = \frac{1}{L} \sum_{l=1}^L Q_{H,l}^{\phi}(\pi, \alpha).$$

We use myopic (greedy) policy as base policy  $\phi$  that is implemented for a trajectory. One step policy improvement is performed, and the optimal action is selected according follow rule.

$$j^*(\pi) = \arg \max_{1 \leq j \leq N} \left[ r(\pi, \alpha = j) + \beta \tilde{Q}_{H,L}^{\phi}(\pi, \alpha = j) \right]. \quad (9)$$

In each time step, an arm is played based on the above rule. Detailed discussion on rollout policy for multi-action RMAB and fully observable state is given in [22]. In next section we present numerical examples using Monte Carlo rollout policy.

## V. NUMERICAL RESULTS AND DISCUSSION

We describe three numerical examples that demonstrate the performance of index policy, myopic policy and Monte Carlo rollout policy. In the myopic policy, the arm with highest immediate expected payoff is played at each time step. In index policy, the arm with highest index is played.

We present first numerical example for model 1. We use following parameters. The number of arms  $N = 15$ , number of states  $n = 4$ , discount parameter  $\beta = 0.95$ , number of message  $K = 2$  and binary reward is considered for each state. Assume that the transition probabilities and observation probabilities are know. As the states are not observable at

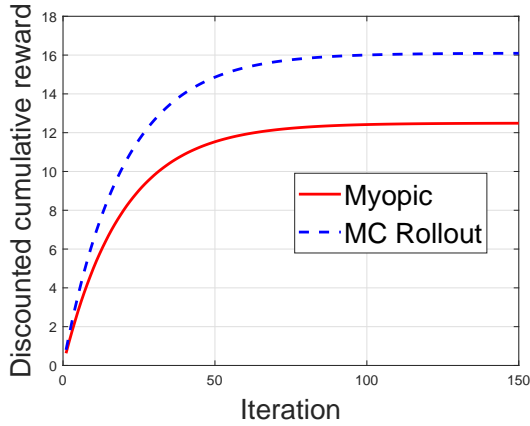


Fig. 2. Model 1 : Myopic vs Monte Carlo rollout policy

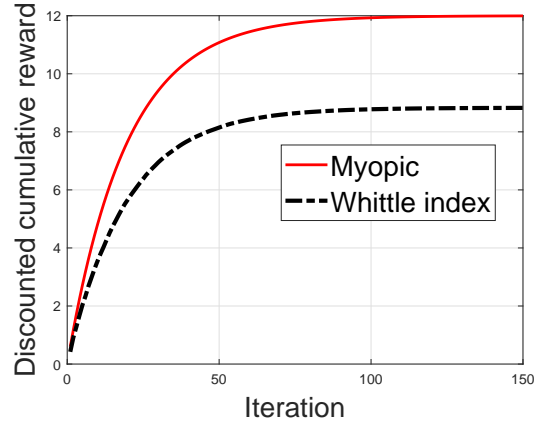


Fig. 4. Model 3 : Myopic vs Whittle index policy

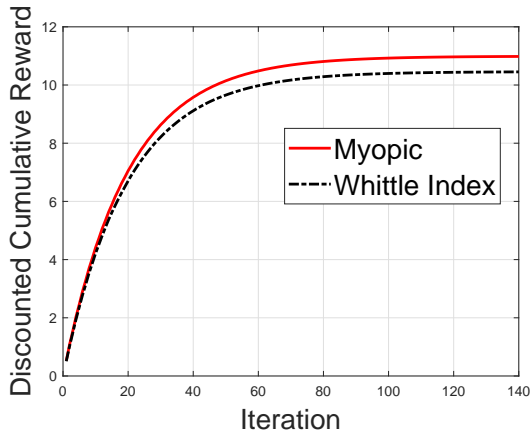


Fig. 3. Model 2 : Myopic vs Approximate index policy

all in this model, we do not make assumption on transition probabilities, i.e.  $TP_2$  order. We compare Monte Carlo rollout policy and myopic policy. We use number of horizon  $H = 5$  and number of trajectories  $L = 100$ . We plot iteration vs discounted cumulative reward. We observe from Fig. 2 that Monte Carlo rollout policy performs better than myopic policy up to 25%. Though rollout policy is computationally expensive it has advantages in terms of higher cumulative reward.

In our second example is for model 2, where we consider number of arms  $N = 5$ , number of states  $n = 4$ ,  $\beta = 0.95$ ,  $K = 2$  and binary reward is obtained from each state after play of arm and no reward is obtained after not playing of arm. In this example we compare index policy and myopic policy. We note that index computation is performed using Monte Carlo rollout policy, where we use  $H = 5$ . We observe from Fig. 3 that myopic policy performs better than approximate index policy based algorithm. Myopic performs better by 5%. This difference is due to approximation in index computation.

In our third example, we present numerical example for model 3. Here, Whittle index formula is explicitly available.

We compare myopic policy and Whittle index policy for  $N = 15$ ,  $n = 4$  and discount parameter  $\beta = 0.95$ . We observe from Fig. 4 that Whittle index policy performs poor than myopic policy. This is due to myopic policy plays only a fixed arm, 3 for all times whereas Whittle index policy plays more than one arm more frequently based on index. In this example it suggest Whittle index policy is not optimal but it is fair and plays other arms as well.

## VI. CONCLUDING REMARKS

In this paper we studied partially observable restless multi-armed bandits. We considered three different models based on information observable to decision maker.

From numerical examples, it suggests that application of directly Monte Carlo rollout policy on restless multi-armed bandits can have advantages over myopic policy. In general, an index policy for multi state partially observable models need not be optimal. We observed that Whittle index policy need not be optimal even we have index formula. A simple rollout policy is competitive to myopic policy when no index formula is available.

This opens interesting future direction of work on MC rollout policy for other partially observable models when indexability and index computations are infeasible.

## REFERENCES

- [1] R. Meshram, D. Manjunath, and A. Gopalan, "A restless bandit with no observable states for recommendation systems and communication link scheduling," in *Proc. IEEE CDC*, 2015.
- [2] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance," *IEEE Transactions on Wireless Communication*, vol. 7, no. 12, pp. 5431–5440, December 2008.
- [3] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589–600, April 2007.
- [4] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Transactions Information Theory*, vol. 56, no. 11, pp. 5557–5567, November 2010.
- [5] A. Abbou and V. Makis, "Group maintenance: A restless bandits approach," *INFORMS Journal of Computing*, pp. 1–13, 2019.

- [6] Y. Shao, Q.Cao, S. C. Liew, and H. Chen, "Partially observable minimum-age scheduling: The greedy policy," *Arxiv*, pp. 1–16, 2020.
- [7] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, no. A, pp. 287–298, 1988.
- [8] C. H. Papadimitriou and J. H. Tsitsiklis, "The complexity of optimal queueing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, May 1999.
- [9] W. Ouyang, A. Eyrlmaz, and N. Shroff, "Asymptotically optimal downlink scheduling over Markovian fading channels," in *Proceedings of IEEE INFOCOM*, 2012, pp. 1224–1232.
- [10] J. L. Ny, M. Dahleh, and E. Feron, "Multi-UAV dynamic routing with partial observations using restless bandit allocation indices," in *Proceedings of American Control Conference (ACC 2008)*, 2008, pp. 4220–4225.
- [11] S. H. A. Ahmad, M. Liu, T. Javidi, and Q. Zhao, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4040–4050, September 2009.
- [12] R. Meshram, D. Manjunath, and A. Gopalan, "On the Whittle index for restless multi-armed hidden markov bandits," *IEEE Transactions on Automatic Control*, vol. 69, pp. 3046–3053, 2018.
- [13] Y. Ouyang and D. Teneketzis, "On the optimality of myopic sensing in multi-state channels," *IEEE Transactions on Information Theory*, vol. 60, pp. 681–696, 2014.
- [14] K. Wang, L. Chen, and Q. Liu, "On optimality of myopic sensing policy with imperfect sensing in multi-channel opportunistic access," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3854–3862, September 2013.
- [15] M. Larranga, M. Assaad, A. Destounis, and G. S. Paschos, "Asymptotically optimal pilot allocation over markovian fading channels," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5395–5418, 2018.
- [16] W. S. Lovejoy, "Some monotonicity results for partially observed Markov decision processes," *Operations Research*, vol. 35, no. 5, pp. 736–743, October 1987.
- [17] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1, Athena Scientific, Belmont, Massachusetts, 1st edition, 1995.
- [18] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 2, Athena Scientific, Belmont, Massachusetts, 1st edition, 1995.
- [19] K. J. Astrom, "Optimal control of Markov processes with incomplete state information II. The convexity of loss function," *Mathematical Analysis and Applications*, vol. 26, no. 2, pp. 403–406, May 1969.
- [20] V. S. Borkar, *Stochastic Approximation: A Dynamical System Viewpoint*, Cambridge University Press, 2008.
- [21] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963.
- [22] R. Meshram and K. Kaza, "Simulation based algorithms for Markov decision processes and multi-action restless bandits," *Arxiv*, 2020.

## APPENDIX

### A. Proof of Lemma 3

We now define  $f$  for finite horizon as follows.

$$f(\pi, V_t^*) = g(\pi, a, V_t^*) - g(\pi, a', V_t^*) \quad a \geq a', \quad a, a' \in \mathcal{A} \quad (10)$$

We next show a threshold-type policy result and to claim this result, we require to show that  $f(\pi, V_t^*)$  is nondecreasing in  $\pi \in \Pi(S)$ . This property also referred to as submodularity of function. Even though optimal value function  $V_t^*(\pi)$  is monotone in  $\pi$ , we can not say about this difference for model 1. To see this, we substitute value of  $g$  in Eqn. (10), then

$$\begin{aligned} f(\pi, V_t^*) &= \sum_{i=1}^n \pi_i (r(i, a) - r(i, a')) + \\ &\beta \sum_{k \in O} \sigma(k | \pi, a) V_t^*(\Gamma(\pi, a, k)) - \\ &\beta \sum_{k \in O} \sigma(k | \pi, a') V_t^*(\Gamma(\pi, a', k)) \end{aligned} \quad (11)$$

Note that monotonicity of value function, we can say that term 1 and term 2 in Eqn. (11) is monotone but third term has negative sign, which introduces difficulty for threshold policy behavior.

But in case of model 2 and 3, we can claim threshold policy result. Under structural assumption on model, i.e.,  $\Gamma(\pi, a', k) = e_i$  where  $e_i$  is the unit vector of dimension  $n$  with 1 at  $i$ th position and zero at remaining position. This simplifies the Eqn, (11) as follows.

$$\begin{aligned} f(\pi, V_t^*) &= \sum_{i=1}^n \pi_i (r(i, a) - r(i, a')) + \\ &\beta \sum_{k \in O} \sigma(k | \pi, a) V_t^*(\Gamma(\pi, a, k)) - \\ &\beta V_t^*(e_i) \end{aligned} \quad (12)$$

Now observe that third term is just constant and hence we can now claim the monotonicity of  $f(\pi, V_t^*)$  in  $\pi$  under assumptions in Lemma 2. This proves the threshold policy result.  $\square$

### B. Proof of Lemma 5

- We first derive index for  $e_{m_1} \in U_1(W)$  and  $e_{m_2} \in U_0(W)$ . We define the action value function  $V_1(\pi) = V(\pi, a = 1)$  and  $V_0(\pi) = V(\pi, a = 0)$ .

$$\begin{aligned} V(e_{m_1}) &= V_1(e_{m_1}) \\ V_1(e_{m_1}) &= r(m_1, 1) + \beta V_1(e_{m_1}). \end{aligned}$$

Thus

$$V(e_{m_1}) = \frac{r(m_1, 1)}{1 - \beta}.$$

The action value function for action 1 with belief  $\pi$  is

$$\begin{aligned} V_1(\pi) &= \sum_{j=1}^n r(j, 1) + \beta V(e_{m_1}) \\ &= \sum_{j=1}^n r(j, 1) + \beta \frac{r(m_1, 1)}{1 - \beta}. \end{aligned}$$

We now obtain the  $V_0(\pi)$ .

$$\begin{aligned} V_0(e_{m_2}) &= W + r(m_2, 0) + \beta V(e_{m_2}) \\ &= \frac{W + r(m_2, 0)}{1 - \beta}. \end{aligned}$$

$$\begin{aligned} V_0(\pi) &= W + \sum_{j=1}^n r(j, 0) \pi(j) + \beta V(e_{m_2}) \\ &= W + \sum_{j=1}^n r(j, 0) \pi(j) + \beta \left( \frac{W + r(m_2, 0)}{1 - \beta} \right) \end{aligned}$$

From the threshold policy we know that at  $\pi$  we have  $V_1(\pi) = V_0(\pi)$ . After equating and solving we get

$$\begin{aligned} W(\pi) &= (1 - \beta) \left( \sum_{j=1}^n [r(j, 1) - r(j, 0)] \pi(j) \right) + \\ &\beta [r(m_1, 1) - r(m_2, 0)]. \end{aligned}$$

This is an index formula.

- We now derive the index when  $e_{m_1}, e_{m_2} \in U_1(W)$ . We obtain value function expression first.

$$\begin{aligned} V(e_{m_1}) &= V_1(e_{m_1}) \\ V_1(e_{m_1}) &= r(m_1, 1) + \beta V(e_{m_1}) \\ V(e_{m_1}) &= \frac{r(m_1, 1)}{1 - \beta}, \end{aligned}$$

and

$$\begin{aligned} V(e_{m_2}) &= r(m_2, 1) + \beta V(e_{m_1}) \\ V(e_{m_1}) &= r(m_2, 1) + \beta \frac{r(m_1, 1)}{1 - \beta}. \end{aligned}$$

Then

$$\begin{aligned} V_1(\pi) &= \sum_{j=1}^n r(j, 1)\pi(j) + \beta V(e_{m_1}) \\ &= \sum_{j=1}^n r(j, 1)\pi(j) + \beta \frac{r(m_1, 1)}{1 - \beta} \end{aligned}$$

and

$$\begin{aligned} V_0(\pi) &= W + \sum_{j=1}^n r(j, 0)\pi(j) + \beta V(e_{m_2}) \\ &= W + \sum_{j=1}^n r(j, 0)\pi(j) + \beta \left[ r(m_2, 1) + \beta \frac{r(m_1, 1)}{1 - \beta} \right] \end{aligned}$$

After equating  $V_1(\pi)$  and  $V_0(\pi)$  and solving for  $W$ , we have

$$\begin{aligned} W(\pi) &= \sum_{j=1}^n [r(j, 1) - r(j, 0)] \pi(j) + \\ &\quad \beta [r(m_1, 1) - r(m_2, 1)]. \end{aligned}$$

- We now derive index formula when  $e_{m_1}, e_{m_2} \in U_0(W)$ . We obtain

$$V(e_{m_2}) = V_0(e_{m_2}) \quad (13)$$

$$= W + r(m_2, 0) + \beta V(e_{m_2}) \quad (14)$$

$$= \frac{W + r(m_2, 0)}{1 - \beta}. \quad (15)$$

$$V(e_{m_1}) = W + r(m_1, 0) + \beta V(e_{m_2}).$$

Then

$$V_1(\pi) = \sum_{j=1}^n r(j, 1)\pi(j) + \beta V(e_{m_1})$$

$$V_2(\pi) = W + \sum_{j=1}^n r(j, 1)\pi(j) + \beta V(e_{m_2})$$

After equating and solving these equations for  $W$ , we obtain

$$\begin{aligned} W(\pi) &= \sum_{j=1}^n [r(j, 1) - r(j, 0)] \pi(j) + \\ &\quad \beta [r(m_1, 0) - r(m_2, 0)]. \end{aligned}$$

□

### C. Proof of Theorem 2

Initial belief is  $\pi_0 = \pi$ . The immediate expected reward at time  $t$  for action  $a = 1$  is  $r(\pi_t, a = 1) = \sum_{i \in S} \pi_t(i) r(i, a = 1)$ . We have assumed  $0 < r(i, a = 1) \leq 1$ . Then immediate expected reward for action  $a = 1$  is bounded, and  $0 < r(\pi_t, a = 1) \leq 1$  and here  $R_{\max} = 1$  and  $R_{\min} = 0$ . Similarly the immediate expected reward for action  $a = 0$  is  $0 < r(\pi_t, a = 0) + W < 1$  for any  $\pi_t$ . We assume that  $0 \leq W \leq 1$ .

We suppose that  $V_\phi(\pi, a, W)$  is the value function for an arm under policy  $\phi$ , with initial state  $\pi$ , action  $a$  and subsidy  $W$ .

Note that  $\{Q_{h,l}^\phi(\pi, a, W)\}_{l=1}^L$  are independent random trajectories generated using policy  $\phi$  for horizon length  $H$  starting from state  $\pi$ , action  $a$  and subsidy  $W$ . Thus, for each trajectory  $l$ , we have  $Q_{l,H}^\phi(\pi, a, W) \in \left[0, \frac{(1-\beta^H)}{1-\beta}\right]$ . This is due to reward is bounded in each steps by  $R_{\max} = 1$ . Let  $z = \frac{(1-\beta^H)}{1-\beta}$ .

Define the action value function under policy  $\phi$  is  $Q^\phi(\pi, a, W)$  for starting belief  $\pi$  and action  $a$ . This is discounted cumulative expected reward for infinite horizon problem. Thus we utilize the Hoeffding inequality [21] for independent random bounded random variables. We have following inequality.

$$\Pr \left( \left| Q^\phi(\pi, a, W) - \frac{1}{L} \sum_{l=1}^L Q_{H,l}^\phi(\pi, a, W) \right| > \epsilon \right) \leq 2 \exp \left( -\frac{2L^2 \epsilon^2}{Lz^2} \right)$$

Thus RHS of preceding term is

$$2 \exp \left( -\frac{2L^2 \epsilon^2}{Lz^2} \right) = 2 \exp \left( -\frac{2L \epsilon^2}{z^2} \right)$$

We want this term to  $\delta$  Hence

$$2 \exp \left( -\frac{2L \epsilon^2}{z^2} \right) = \delta$$

After rearranging terms, we have

$$\epsilon = \sqrt{\frac{z^2}{2L} \log(2/\delta)}$$

Setting  $\delta = \frac{2}{H^2}$  we get following inequality with probability  $1 - \frac{2}{H^2}$

$$\left| Q^\phi(\pi, a, W) - \frac{1}{L} \sum_{l=1}^L Q_{H,l}^\phi(\pi, a, W) \right| \leq \sqrt{\frac{z^2 \log H}{L}}$$

We know that

$$V_\phi(\pi, W) = \max_{a \in \{0,1\}} Q^\phi(\pi, a, W)$$

Thus we can have following inequality with high probability  $1 - \frac{2}{H^2}$  for sufficiently large horizon  $H$  and  $L > \tilde{L}$

$$\left| V_\phi(\pi, a, W) - \tilde{V}_{\phi,H,L}(\pi, a, W) \right| \leq \sqrt{\frac{z^2 \log H}{L}}$$

for  $a \in \{0, 1\}$ . This completes the proof. □

□