Restless Bandits With Cumulative Feedback: Applications in Wireless Networks

Kesav Kaza, Varun Mehta, Rahul Meshram and S. N. Merchant
Department of Electrical Engineering
IIT Bombay, Mumbai INDIA.

***Abstract*—Restless multi-armed bandits(RMAB) with partially observable states have been extensively studied for scheduling in opportunistic communication systems. These RMAB models assume that when the decision maker plays a particular arm, it gathers information about system state through feedback signals. These models allow only one state transition in a single decision interval.**

In this paper, we propose a *cumulative feedback* model, where multiple state transitions occur in a decision interval. We formulate opportunistic scheduling in communication systems and relay selection problem as partially observable RMAB with cumulative feedback. In this model, state of an arm is not observable whether it is played or not. But belief about state is maintained and it is updated at end of each decision interval based on feedback from the played arm. If an arm is not played, then no feedback is available. But belief is updated with natural evolution. In case of large number of channel state transitions in a decision interval for not-played arms, we approximate belief with stationary probability. For this scenario we solve partially observable RMAB using the Whittle index policy. A closed-form expression for Whittle index is obtained for a special case. The efficacy of this policy is illustrated via some numerical examples and it is also compared with other policies.

I. Introduction

Wireless communication systems often operate in uncertain environments such as rapidly varying channel conditions, relative mobility of communicating nodes. Hence, the need for decision making under uncertainty occurs in many applications involving relay selection [1], relay employment in wireless networks [2], channel sensing and scheduling for opportunistic communications [3], [4].

Let us first look at the problem of relay selection in wireless networks. Consider a wireless relay network with a source(S), destination(D) and a set of M relays $R_i, 1 \leq i \leq M$. The links SR_i and R_iD operate on different frequencies. So there are $M+1$ paths or links from source to destination including the direct SD link. The channel quality along each of these paths is varying with time. However, the source is unable to observe the exact channel qualities. The time line is divided into intervals. One relay may be selected for use in each interval. A feedback in form of ACK/NACK is received by the source at the end of each interval signifying success or failure of the message transmission. The source has to plan the sequence of relays to be used such that the expected long term throughput is maximized. Thus, the problem involves sequential decision making, where each decision must take into account the information gathered till that instant in the form of ACKs or NACKs from previous transmissions (decisions).

We now look at the problem of scheduling for opportunistic communication in cognitive radio networks. There are M frequency channels allotted for the use of primary users of the network. A secondary user may use these channels when they are not in use by the primary users. A secondary user can sense one channel at a time; if the sensed channel is free, then it is used for transmission. A feedback in form ACK/NACK is received at the end of each transmission signifying its success or failure respectively. Clearly, the throughput obtained depends on the channel quality. Here, at the beginning of each interval, the secondary user needs to decide which channel to sense in order to maximize probability of successful transmissions. This problem also involves sequential decision making that must take into account information gathered in the form of ACKs/NACKs from previous transmissions.

Often sequential decision problems are modeled using Markov decision processes (MDP) [5], partially observable Markov decision processes (POMDP) [6], [7], and multi-armed bandits (MAB) [8], [9]. In these models, environment/system state transition and decision making occur at discrete time instants, uniformly spaced along the time line. The knowledge of system state at these instants, provides information that is necessary for decision making. This knowledge about the system state depends on the observation or feedback about state transition that occurs as a consequence of the previous decision. All the above models assume that, every state transition is either fully or partially observable by the decision maker. This form of information gathering by the decision maker about the consequence of its actions is imperative for all sequential decision models.

In this work we consider a scenario where the information gathering of the decision maker is not at par with the variation of system state. The instants of decision making are sparse compared to the instants of system state transition. The decision maker does not observe every state transition; instead, observation of the system takes place only when a decision needs to be made. We refer to the information gathered by this form of observation as *cumulative feedback*; it represents the cumulative effect of a series of state transitions.

Recently, restless multi-armed bandits have been used to model the problem of dynamic scheduling of projects/resources in uncertain environments, [8], [9]. Restless multi-armed bandit (RMAB) problem is described

This is a technical report for a paper with the same title submitted to IEEE WCNC 2018.

as follows. It has M independent arms (resources), each arm can be in one of a finite set of states. The play of an arm yields a reward that depends on state of that arm. At every decision epoch, decision maker plays a fixed number of arms simultaneously; the states of arms evolve according to Markov chains and are action dependent. The objective of the decision maker is to select the optimal sequence in which arms should be played such that it maximizes the long term reward function. Finding the optimal sequence that maximizes long term reward for RMAB, i.e., the optimal solution is known to be PSPACE hard, see [10]. In the seminal paper [9], Whittle introduced a hueristic index based policy for RMAB, where state of an arm is mapped to a real valued index and the arms with highest indices are played. The popularity of this index policy is because of its asymptotic optimality, [11]. This index policy now referred to as Whittle index policy.

An approach to obtain index for an arm is to study Lagrangian relaxed version of the stochastic optimization problem. This relaxation reduces the complexity of RMAB problem and it allows to separately solve M restless single-armed bandit (RSAB) problem. Using structural properties of RSAB, one requires to first show that the arm is indexable and later one can compute the index.

The relay selection or employment problem and opportunistic scheduling problem can be modeled using RMAB. Each source-destination link in relay networks corresponds to an arm; each link can be one of finite states which represent the quality of that link. The reward from using a relay link is the throughput that dependent on its state. Also, opportunistic scheduling problem can be modeled as RMAB, where each channel corresponds to an arm, state of an arm describes the channel quality and reward for arm play being the throughput that is state dependent. In these applications, the channel or link qualities are not observed at transmitter, but ACK/NACK feedback is available. Such bandits are referred to as partially observable RMAB.

In this paper we formulate the problem of partially observable RMAB with *cumulative feedback*. In both the applications discussed above, ACK/NACK is a *cumulative feedback* as it represents the impact of multiple channel state transitions in the form of success or failure of a message transmission.

We now discuss some literature on RMAB problem, related in particular to partially observable RMAB. It has been extensively studied for opportunistic communication and resource allocation problems, see [3], [12], [12]–[18]. The work of [3], [12], [13], [15], [19] assumes that state of channel is observed perfectly when channel is used for transmission using ACK/NACK feedback and state is not observed when channel is not used. A variation and extension of this model is proposed in [14], where channel state is also not observed if channel is used for transmission and no ACK received. Further generalization of these two channel models is considered in [17], [18]. In this, channel is not observed for both transmission and no-transmission but ACK/NACK is observed; such bandits are referred to as hidden RMAB. All the above work studied Whittle index based policy. In the work of [3], [15], index policy is also shown to be asymptotically optimal. Alternatively, myopic policy is studied for RMAB, and it is shown to be optimal is special cases, see [4], [19]. Versions of relay selection problem have been analyzed using POMDP in [1], [20].

In all of the above problems, (1) each action generates at most one system state transition till the next decision epoch, (2) the feedback about an action is available for making the next consecutive decision. This implies that the decision maker receives distinct information about every system state transition.

In this paper, we propose *cumulative* feedback based model that relaxes the first assumption of the above work, *i.e.*, multiple state transitions are allowed between consecutive decision epochs. Further, this model does not necessitate receiving distinct information about every state transition.

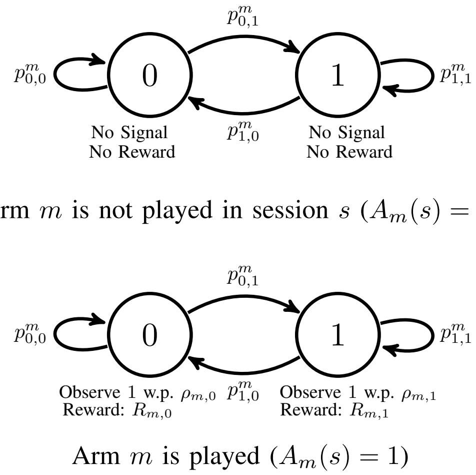
Our contribution and organization of the paper is as follows. We first model each channel using Gilbert-Elliot model, and propose cumulative feedback ACK/NACK model in Section II. Here, we formulate optimization problem using RMAB with partially observable states. To obtain the index, we analyze the single-armed bandit problem in Section III. Here, we also derive structural results and obtain closed-form expression for index in one special case. In Section IV, we illustrate the efficacy of index based policy via few numerical examples. We compare the performance of index policy with myopic and other policies. Finally, we conclude the work and provide some discussion on future work in Section V.

II. Model Description and Preliminaries

Consider a restless multi armed bandit with M independent arms. The time line is divided into sessions that are indexed by s. The arms of the multi-armed bandit represent channels/links in a communication system. We model each channel using Gilbert-Elliot model. In this model each channel has two state, say, good (1) and bad (0). In an arbitrary session, each arm exists in one of two states. $Y_m(s) \in \{0,1\}$ denotes the state of arm m at the beginning of session s. Let $K > 1$ be the number of transitions of channel state of an arm in a given session. The state of each arm evolves according to a Markov chain. $p_{i,j}^{m}$ represents the transition probability of arm m from state i to state j, $i, j \in \{0,1\}$ and the corresponding transition probability matrix (TPM) is denoted by $P_m = [p_{i,j}]$. In a given session s, the decision maker plays one arm out of M arms. $A_m(s)$ denotes the action corresponding to arm m in session s. If arm m is played in session s, then $A_m(s) = 1$ and $A_m(s) = 0$, otherwise. Since only one arm is played in a session, $\sum_{m=1}^{M} A_m(s) = 1$.

At the end of each session a feedback is received by the decision maker in the form of ACK(1) or NACK(0) from the arm that is played. An ACK means a successful session and a NACK means a failed session. $Z_m(s) \in \{0,1\}$ denotes the feedback signal that is obtained at end of session s if arm m is played in session s. This feedback is probabilistic, and we define $\rho_{m,i} := \Pr\{Z_m(s) = 1 \mid A_m(s) = 1, Y_m(s) = i\}$, $i \in \{0,1\}$. We also assume that $\rho_{m,0} < \rho_{m,1}$.

A reward is accrued from arm m, if that arm is played in a session s. It depends on states of the arm, $Y_m(s)$. We denote $R_{m,i}$ as reward from arm m if it is played and it is in state i,



Arm m is not played in session s ($A_m(s) = 0$)

Arm m is played ($A_m(s) = 1$)

Fig. 1. The state transition probabilities, the reward, and the probability of ACK (1) being observed are illustrated above when the arm is not played. Also, the corresponding quantities are illustrated below when the arm is played.

and further $R_{m,0} < R_{m,1}$. No reward is accrued if arm m is not played.

We assume that the exact state of each arm is not observable by the decision maker. The decision maker maintains the belief about the state of each arm. Let $\pi_m(s)$ the probability that arm m is in state 0 at the beginning of session s given the history $H(s)$, where $H(s) = \{A(l), Z(l)\}_{1\leq l<s}$. Thus $\pi_m(s) := \mathsf{Pr}\left(Y_m(s) = 0 \mid H(s)\right)$. The belief $\pi_m(s)$ about arm m, is updated by the decision maker at the end of every session s, based on the action taken $A_m(s)$ and feedback received $Z_m(s)$.

Let $\phi := \{\phi(s)\}_{s\geq 0}$ be the policy, where $\phi(s) : H_s \rightarrow \{1, \cdots, M\}$ maps the history up to session s to action of playing one of the M arms. Let $A^{\phi}_m(s) = 1$, if $\phi(s) = m$, and $A^{\phi}_m(s) = 0$, if $\phi(s) \neq m$. The infinite horizon expected discounted reward under policy ϕ is given by

$$V_\phi(\pi) := \mathrm{E}\Bigg\{ \sum_{s=1}^{\infty} \beta^{s-1} \sum_{m=1}^{M} A^{\phi}_m(s) \left(\pi_m(s) R_{m,0} \right. \\ \left. + (1-\pi_m(s))R_{m,1}\right) \Bigg\}. \quad (1)$$

Here, β is discount parameter, $0 < \beta < 1$ and the initial belief $\pi = [\pi_1, \cdots, \pi_M]$, $\pi_m := \mathsf{Pr}\left(Y_m(1) = 0\right)$. Our objective is to find the policy ϕ that maximizes $V_\phi(\pi)$ for all $\pi \in [0,1]^M$.

In [9], Lagrangian relaxation of this problem is analyzed via introducing subsidy, it is payoff for not playing the arm. The approach to obtain the solution of the relax problem by studying first single-armed restless bandit and this is studied in next section.

III. Single-armed Restless Bandit

We consider a subsidy η is assigned if the arm is not played. We here drop notation m for notational convenience. In the view of subsidy η one can reformulate problem in (1) for single-armed bandit as follows.

$$V_\phi(\pi) := \mathrm{E}\Bigg\{ \sum_{s=1}^{\infty} \beta^{s-1} \Bigg(A^{\phi}(s) \left(\pi(s) R_0 + (1-\pi(s))R_1\right) \\ + \eta(1 - A^{\phi}(s)) \Bigg) \Bigg\}. \quad (2)$$

The goal is to find the policy ϕ that maximizes $V_\phi(\pi)$ for $\pi \in [0,1]$, π is the initial belief.

We now describe the belief update rule and it plays important role in obtaining properties of the value function.

1) If a channel is used for transmission in session s and ACK is received, i.e., $A(s) = 1$ and $Z(s) = 1$, then the belief at the beginning of session $s+1$ is $\pi(s+1) = \gamma_1(\pi(s))$. Here,
$$\gamma_1(\pi(s)) := \frac{(1-\pi(s))\rho_1 p_{1,0} + \pi(s)\rho_0 p_{0,0}}{\rho_1(1-\pi(s)) + \rho_0 \pi(s)}.$$
2) If a channel is used for transmission in session s and NACK is received, i.e., $A(s) = 1$ and $Z(s) = 0$, then the belief at the beginning of session $s+1$ is $\pi(s+1) = \gamma_0(\pi(s))$, where
$$\gamma_0(\pi(s)) := \frac{(1-\pi(s))(1-\rho_1)p_{1,0} + \pi(s)(1-\rho_0)p_{0,0}}{(1-\rho_1)(1-\pi(s)) + (1-\rho_0)\pi(s)}.$$
3) If a channel is not used for transmission, i.e., $A(s) = 0$, then the belief at the beginning of session $s+1$ is $\pi(s+1) = \gamma_2(\pi(s))$, where
$$\gamma_2(\pi(s)) := \left(p_{0,0} - p_{1,0}\right)^K \pi + p_{1,0} \frac{\left(1 - (p_{0,0} - p_{1,0})^K\right)}{1 - (p_{0,0} - p_{1,0})}. \quad (3)$$
This is because the channel is evolving independently, after K transitions of channel state, we obtain belief as given in the expression (3).

Note that whenever $p_{0,0} > p_{1,0}$, the communication channel is called as positively correlated channel. In this paper we study only positively correlated channel model.

Lemma 1: For positively correlated channel, i.e., $p_{0,0} > p_{1,0}$, the belief updates $\gamma_0(\pi)$, $\gamma_1(\pi)$ and $\gamma_2(\pi)$ are increasing in π. Further, $\gamma_1(\pi)$ and $\gamma_0(\pi)$ are convex and concave, respectively. Also, $p_{1,0} \leq \gamma_1(\pi) \leq \gamma_0(\pi) \leq p_{0,0}$.
The proof is straight forward and it can be done by twice differentiating each of the update functions γ_0, γ_1 w.r.t π and looking at their signs.

Remark 1: We can see from the expression of γ_2 in Eqn. (3) that for fixed value of π, as $K \rightarrow \infty$, we get $\gamma_2(\pi) \rightarrow q$, where $q = \frac{p_{1,0}}{1-(p_{0,0}-p_{1,0})}$. The rate of convergence of γ_2 to q depend on $(p_{0,0} - p_{1,0})$. This suggests that for large values of K, we can approximate $\gamma_2(\pi)$ with q. If $|p_{0,0} - p_{1,0}|$ is smaller then k required for this approximation is small. Also, $\gamma_2(\pi) = q$ is good approximation for fast varying channel because K is sufficiently large.

We seek for a stationary deterministic policy. From [21], [22], we know that $\pi(s)$ is a sufficient statistic for constructing such policies and the optimal value function can be determined by solving following dynamic program.

$$\begin{aligned} V_S(\pi) &= R_S(\pi)+\beta\left(\rho(\pi)V(\gamma_1(\pi))+(1-\rho(\pi))V(\gamma_0(\pi))\right)\\ V_{NS}(\pi) &= \eta+\beta V(\gamma_2(\pi))\\ V(\pi) &= \max\{V_S(\pi),V_{NS}(\pi)\}. \qquad (4)\end{aligned}$$

Here $R_S(\pi)=\pi R_0+(1-\pi)R_1$ and $\rho(\pi)=\pi\rho_0+(1-\pi)\rho_1$ We next derive the structural results for value functions.

Lemma 2:

1) For fixed η, $V_S(\pi)$, $V_{NS}(\pi)$ and $V(\pi)$ are convex in π.
2) For fixed π, $V_S(\pi,\eta)$, $V_{NS}(\pi,\eta)$ and $V(\pi,\eta)$ are non-decreasing and convex in η.
3) For fixed subsidy η, $\beta\in(0,1)$, and $p_{0,0}>p_{1,0}$. The value functions $V(\pi)$, $V_S(\pi)$ and $V_{NS}(\pi)$ are decreasing in π.
4) For fixed subsidy η, $\beta\in(0,1)$, and $p_{0,0}>p_{1,0}$. The difference in value function $(V_S(\pi)-V_{NS}(\pi))$ is decreasing in π.

The proofs of parts 1)-4) are given in Appendix B-E respectively.

We define a threshold type policy and we will show that a threshold type optimal policy for single armed bandit for large values of K, i.e., $\gamma_2(\pi)\approx q$.

Definition 1: A policy is called as a threshold type for single armed bandit if there exists $\pi_T\in[0,1]$ such that an optimal action is to play the arm if $\pi\leq\pi_T$ and to not to play the arm if $\pi\geq\pi_T$.

Theorem 1: For fixed subsidy η, $\beta\in(0,1)$, and $p_{0,0}>p_{1,0}$. The optimal policy for single-armed bandit is of a threshold type.

Proof: From the preceding Lemma 2, we know that $(V_s(\pi)-V_{NS}(\pi))$ is a decreasing in π. Further, $V_S(\pi)$ and $V_{NS}(\pi)$ are convex in π. This implies that there exists a either $\pi_T\in[0,1]$ such that $V_S(\pi_T)=V_{NS}(\pi_T)$ or $V_S(\pi)>V_{NS}(\pi)$ for all π, or $V_S(\pi)<V_{NS}(\pi)$ for all π. This leads to desired result. ■

We here define the indexability and will show that a single-armed bandit is indexable. Using exact threshold-type policy result, we define the following.

$$\mathcal{P}_\beta(\eta)=\left\{\pi\in[0,1]:V_S(\pi,\eta)\leq V_{NS}(\pi,\eta)\right\}.$$

It is a set of belief state π for which the optimal action is to not to play the arm, i.e., $A(s)=0$. From [9], we state the definition of indexability.

Definition 2: A single-armed restless bandit is indexable if $\mathcal{P}_\beta(\eta)$ is monotonically increases from $\emptyset$ to entire state space $[0,1]$ as η increases from $-\infty$ to ∞, i.e., $\mathcal{P}_\beta(\eta_1)\backslash\mathcal{P}_\beta(\eta_2)=\emptyset$ whenever $\eta_1\leq\eta_2$.

To show indexability, we require to prove that a threshold π_T as function of η is monotonically increasing. We state the following lemma from [18].

Lemma 3: Let $\pi_T(\eta)=\inf\{\pi\in[0,1]:V_S(\pi,\eta)=V_{NS}(\pi,\eta)\}$, if $\left.\frac{\partial V_S(\pi,\eta)}{\partial\eta}\right|_{\pi=\pi_T(\eta)}<\left.\frac{\partial V_{NS}(\pi,\eta)}{\partial\eta}\right|_{\pi=\pi_T(\eta)}$, then $\pi_T(\eta)$ is monotonically decreasing function of η

Note that the value function may not be differentiable as function of η, in that case it should taken as right partial derivative. It exists due to convexity of value function in η and rewards are bounded.

We now use Definition 2 and Lemma 3 to show that a single-armed restless bandit in our setting is indexable.

Theorem 2: If $\gamma_2(\pi)\approx q$, and $\rho_0<\rho_1$, then a single-armed restless bandit is indexable for $\beta\in(0,1/3)$.

Proof can be found in Appendix A.

Remark 2: We believe that the indexability result is true more generally, where, we do not require any assumption on β. This restriction on β is required here because of difficulty in obtaining closed-form value function expression. But if we assume $\rho_0=0,\rho_1=1$, and $K>1$, we can derive the closed-form expressions of value functions and we can obtain conditions for indexability without any assumption on β.

We now provide definition of the Whittle index from [9].

Definition 3: If an indexable arm is in state π, its Whittle index $W(\pi)$ is

$$W(\pi) = \inf\{\eta\in\mathbb{R}:V_{S,\beta}(\pi,\eta)=V_{NS,\beta}(\pi,\eta)\}. \qquad (5)$$

In order to compute the index, we require to obtain the value function expressions at each threshold π and solve it for subsidy η. Solving these equations, we get desired index for that π. We assume that $p_{0,0}>p_{1,0}$, $\gamma_2(\pi)\approx q$, $R_0=\rho_0=0$, and $0<R_1=\rho_1<1$.

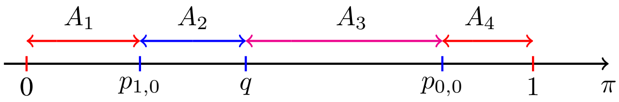
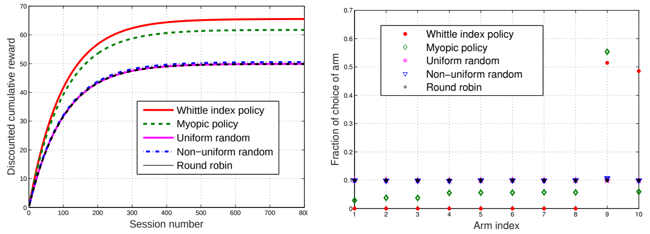


Fig. 2. The different cases to calculate $W(\pi)$.

We consider four intervals, A_1,A_2,A_3, and A_4, this is described in Fig. 2, where we compute the index for each interval separately. We make use of properties of γ_0, γ_1 and γ_2. The index formula for each interval is given as follow.

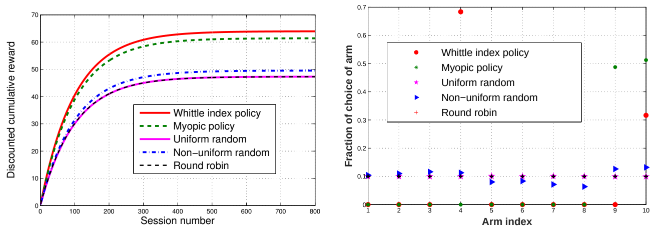
1) For $\pi\in A_1$, the Whittle index $W(\pi)=\rho(\pi)$.
2) For $\pi\in A_2$, we consider following cases.
    a) if $\gamma_0(p_{1,0})\geq\pi$, then Whittle index is
    $$W(\pi)=\frac{\rho(\pi)}{1-\beta(\rho(p_{1,0})-\rho(\pi))}.$$
    b) if $\gamma_0(p_{1,0})<\pi$ but $\gamma_0^2(p_{1,0})\geq\pi$ then Whittle index $W(\pi)=\frac{\rho(\pi)}{C_1}$. Here,
    $$C_1=1-\beta(\rho(p_{1,0})-\rho(\pi))-\beta^2(\rho(\gamma_0(p_{1,0}))-\rho(\pi))+\beta^2\rho(\gamma_0(p_{1,0}))\rho(p_{1,0}).$$
3) For $\pi\in A_3$, obtaining index is tedious, and this has to be computed numerically by value iteration algorithm.
4) For $\pi\in A_4$ the Whittle index is.
    $$W(\pi)=m\pi(1-\beta(p_{0,0}-p_{1,0}))+(1-\beta)c-\beta p_{1,0}m,$$
    $$m=\frac{-\rho_1}{1-\beta(p_{0,0}-p_{1,0})},\quad c=\frac{\rho_1+\frac{-\beta p_{1,0}\rho_1}{1-\beta(p_{0,0}-p_{1,0})}}{1-\beta}.$$

The derivation of Whittle index computation is given in the Appendix.



a) Discounted cumulative reward   b) Arm choice fraction

Fig. 3. We plot a) discounted cumulative rewards as function of sessions for different policies and b) arm choice fraction for each arm with different policies. This is plotted for identical reward for all arms $\rho_1 = 0.9$ and identical $q_m = 0.45$ for $1 \leq m \leq M, m \neq 9$ and $q_9 = 0.4$.



a) Discounted cumulative reward   b) Arm choice fraction

Fig. 4. a) The discounted cumulative reward verses session number for different policies and b) arm choice fraction for each arm with different policies. This is plotted for identical reward for all arms $\rho_1 = 0.9$ but different q_m for each arm.

IV. Numerical Examples

We now present few numerical examples and compare different algorithms that are used to solve partially observable RMAB. The algorithms included in the comparative analysis are 1) Whittle index policy (WI)– plays the arm with highest Whittle index, 2) myopic policy (MP)– plays arm with highest expected immediate reward, 3) uniformly random (UR), 4) non-uniform random (NUR)– plays arm randomly with distribution derived from current belief and 5) round robin (RR)– plays arm in round robin order.

Simulations were performed using MATLAB. In these simulations, the arms start in a random state with a given initial belief about the state of the arm. In each session one arm is played according to the given policy. The reward is accumulated at end of each session from the played arm and it is stored; these rewards are averaged over L iterations.

We will plot and compare the discounted cumulative reward that is obtained from these policies as function of session number. We define arm choice fraction as follows. Let $1_{m,s,l}$ be the indicator variable if arm m is played in session s, and lth iteration. Then $N_{m,l} := \frac{1}{S_{\max}} \sum_{s=1}^{S_{\max}} 1_{m,s,l}$, where $S_{\max}$ number of sessions for which simulations are performed. This is further averaged over L number of iterations. We call this as arm m choice fraction. To gain further insight, we will plot this for all the arms for different policies.

We illustrate three numerical examples for 10 number of arms, i.e., $M = 10$ and we use discount parameter $\beta = 0.99$ and $R_{m,0} = \rho_{m,0} = 0$, $R_{m,1} = \rho_{m,1} = \rho_1$ for $1 \leq m \leq M$. In first two examples, we consider that arms with identical rewards but different transition probabilities. A third example is given for arms with non-identical rewards and transition probabilities.

1) Example-1: In this scenario, all the arms have identical reward from play of that arm. Also, all the arms have same $q_m = 0.45$, except for arm 9, i.e. $q_9 = 0.4$. We use following set of parameters: $\rho_1 = 0.9$ and

$$\begin{aligned} p_{0,0} &= [0.45, 0.5, 0.51, 0.57, 0.63, 0.66, 0.69, 0.75, 0.78, 0.87] \\ p_{1,0} &= [0.45, 0.41, 0.4, 0.35, 0.3, 0.28, 0.25, 0.2, 0.15, 0.1]. \end{aligned}$$

In Fig. 3-a) we plot the discounted cumulative reward as function of number of sessions. It can be seen that the discounted cumulative reward under Whittle index policy (WI) is higher than that of the Myopic policy (MP). We also observe that WI and MP yield higher discounted cumulative reward compared to that of random and round robin policies.

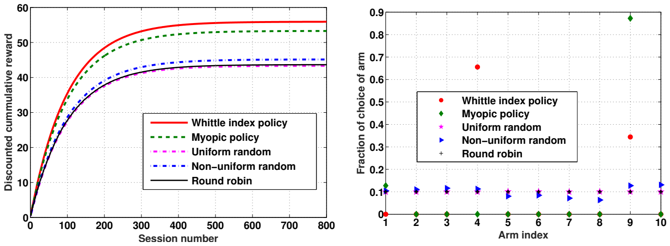
We also plot arm choice fraction in Fig. 3-b). It suggests that Whittle index policy has a tendency to choose the arm from a smaller subset of arms, $\{9, 10\}$ as compared to other policies. This behavior of Whittle index policy might be due to it accounts for future rewards through the action value function. This is also determined by channel characteristics, i.e., $p_{0,0}$, and $p_{1,0}$, where we observe that for arm 9 and 10 difference $(p_{0,0} - p_{1,0})$ is very large compared to other arms. In myopic policy, arm 9 is most frequently played compared to other arms. This is because belief about arm 9 reaches state $q_9 = 0.4$ when that arm is not played, while other arms those are not played reach $q = 0.45$. Note that immediate expected reward from session is state dependent and it is decreasing in state (π). Since rewards are identical for all arms, myopic policy plays arm 9. Another reason for myopic policy to play arm 9 is that $(p_{0,0} - p_{1,0})$ is large. If arm 9 is played and session is successful then state reaches to $p_{1,0} = 0.15$, that means it is more likely to be in good state and hence it will be played again. Whereas for other policies all the arms are played equally often and hence it leads to smaller discounted cumulative reward.

2) Example-2: In this example we consider that all arms has identical reward structure, $\rho_1 = 0.9$ but different values of q_m. We use the following set of parameters:

$$\begin{aligned} p_{0,0} &= [0.5, 0.45, 0.45, 0.78, 0.6, 0.6, 0.7, 0.7, 0.4, 0.45], \\ p_{1,0} &= [0.41, 0.4, 0.35, 0.15, 0.55, 0.5, 0.5, 0.6, 0.3, 0.25], \\ q &= [0.45, 0.42, 0.38, 0.40, 0.57, 0.55, 0.62, 0.66, 0.33, 0.31]. \end{aligned}$$

In Fig. 4-a) we plot the discounted cumulative reward verses sessions. As expected the Whittle index policy yields higher discounted cumulative reward compared to other policies. The myopic policy has better cumulative reward than other policies. We also observe that non-uniform random policy gives better performance compare to round robin and uniform random policy.

In Fig. 4-b), we have plotted arm choice fraction for each arm. It suggests that Whittle index policy has a tendency to choose the arm from a smaller subset of arms, $\{4, 10\}$. Myopic policy chooses the arm from subset of arms, $\{9, 10\}$. Other policies plays all the arms equally. The behavior of Whittle index policy is determined by channel characteristics, i.e., $p_{0,0}$, and $p_{1,0}$, where we observe that for arm 4 has



a) Discounted cumulative reward b) Arm choice fraction

Fig. 5. a) The discounted cumulative reward verses session number for different policies and b) arm choice fraction for each arm with different policies. This is plotted for non-identical reward for all arms.

highest difference in $(p_{0,0} - p_{1,0})$, arm 10 has smallest value of $p_{1,0}$ and $(p_{0,0} - p_{1,0}) = 0.2$. Here index is depend on these parameters and it accounts for future rewards through the action value function. Here, behavior of myopic policy depends on value of q and observe that arm 9 and 10 has least value of q. Thus MP plays arms 9 and 10 most frequently compare to other arms. For other policies all the arms are played equally often.

3) Example-3: In this example arms have non-identical reward structure.

$$\begin{aligned} p_{0,0} &= [0.5, 0.45, 0.45, 0.78, 0.6, 0.6, 0.7, 0.7, 0.4, 0.45], \\ p_{1,0} &= [0.41, 0.4, 0.35, 0.15, 0.55, 0.5, 0.5, 0.6, 0.3, 0.25], \\ q &= [0.45, 0.42, 0.38, 0.40, 0.57, 0.55, 0.62, 0.66, 0.33, 0.31], \\ \rho_1 &= [0.9, 0.8, 0.8, 0.8, 0.9, 0.9, 0.9, 0.9, 0.8, 0.7]. \end{aligned}$$

From Fig. 5-a) we observe that the Whittle index policy and myopic policy yield higher discounted cumulative reward compared to other policies. We notice from Fig. 5-b) that Whittle index policy tend to choose the arm from a smaller subset of arms, $\{4, 9\}$, Myopic policy tend to choose the arm from subset of arms, $\{1, 9\}$. Other policies plays all the arms equally often. This behavior of WI and MP due to channel characteristics and reward structure ρ_1.

V. Concluding Remarks and Discussion

The problem of restless bandits with cumulative feedback has been formulated and solved using Whittle index policy. This cumulative feedback model is applicable in scenarios where rate of system state evolution is faster than the rate of information gathering by the decision maker. To solve the problem Whittle index policy has been studied, a closed form expression for the index obtained for a special case. It's performance was compared with other different policies in numerical simulations. The cumulative feedback model is applicable to problems such as relay selection/employment, opportunistic communication in wireless networks. It minimizes signaling overhead and suitable in applications involving real time multimedia transmission.

A two state Gilbert-Elliot model for communication channel was used in this work. To extend the applicability of this model to multi-state channels, state aggregation can be used. A rigorous analysis of restless bandits with partially observable multiple states(> 2) remains to be done.

References

[1] Y. Li, Y. Hou, Z. Huang, and Y. Wei, "Cooperative relay selection policy using partially observable Markov decision process," in *Proc. of ICNC*, July 2011, vol. 1, pp. 508–512.

[2] K. Kaza, R. Meshram, and S. N. Merchant, "Relay employment problem for unacknowledged transmissions: Myopic policy and structure," in *Proc. of ICC*, May 2017, pp. 1–7.

[3] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Trans. Info. Theory*, vol. 56, no. 11, pp. 5557–5567, November 2010.

[4] K. Wang, L. Chen, and Q. Liu, "On optimality of myopic policy for opportunistic access with nonidentical channels and imperfect sensing," *IEEE Trans. on Veh. Tech.*, vol. 63, no. 5, pp. 2478–2483, June 2014.

[5] Martin L Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, 2014.

[6] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable markov processes over a finite horizon," *Oper. Res.*, vol. 21, no. 5, pp. 1071–1088, 1973.

[7] William S. Lovejoy, "A survey of algorithmic methods for partially observed markov decision processes," *Annals of Oper. Res.*, vol. 28, no. 1, pp. 47–65, 1991.

[8] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed Bandit Allocation Indices*, Wiley, 2011.

[9] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journ. of Appl. Prob.*, vol. 25, no. A, pp. 287–298, 1988.

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

[11] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journ. Appl. Prob.*, vol. 27, no. 3, pp. 637–648, Sept. 1990.

[12] C. P. Li and M. J. Neely, "Network utility maximization over partially observable markovian channels," *Arxiv*, Aug. 2010.

[13] J. Niño-Mora, "An index policy for dynamic fading-channel allocation to heterogeneous mobile users with partial observations," in *Proc. of NGIN*, April 2008, pp. 231–238.

[14] J. Niño-Mora, "A restless bandit marginal productivity index for opportunistic spectrum access with sensing errors," in *Proc. of NET-COOP*, 2009, pp. 60–74.

[15] W. Ouyang, A. Eyrilmaz, and N. Shroff, "Asymptotically optimal downlink scheduling over Markovian fading channels," in *Proc. of INFOCOM*, 2012, pp. 1224–1232.

[16] J. Le Ny, M. Dahleh, and E. Feron, "Multi-uav dynamic routing with partial observations using restless bandit allocation indices," in *Proc. of ACC*, June 2008, pp. 4220–4225.

[17] R. Meshram, D. Manjunath, and A. Gopalan, "A restless bandit with no observable states for recommendation systems and communication link scheduling," in *Proc. of CDC*, 2015.

[18] R. Meshram, D. Manjunath, and A. Gopalan, "On the Whittle index for restless multi-armed hidden Markov bandits," *ArXiv*, March 2016.

[19] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality and performance," *IEEE Trans. on Wire. Comm.*, vol. 7, no. 12, pp. 5431–5440, Dec. 2008.

[20] K. P. Naveen and A. Kumar, "Relay selection with channel probing for geographical forwarding in WSNs," in *Proc. of WiOpt*, May 2012, pp. 246–253.

[21] S. M. Ross, "Quality control under Markovian deterioration," *Management Sci.*, vol. 17, no. 9, pp. 587–596, May 1971.

[22] W. S. Lovejoy, "Some monotonicity results for partially observed Markov decision processes," *Oper. Res.*, vol. 35, no. 5, pp. 736–743, Sept.-Oct. 1987.

[23] KJ Aström, "Optimal control of markov processes with incomplete state information ii: The convexity of the loss function," *J. Math. Ann. Appl*, vol. 26, pp. 403–406, 1969.

[24] Dimitri P Bertsekas, *Dynamic programming and optimal control*, vol. 2, Athena scientific Belmont, MA, 1995.

Appendix

A. Proof of Theorem 2

Using induction technique, one can obtain the following inequalities.

$$\left|\frac{\partial V(\pi,\eta)}{\partial \eta}\right|, \left|\frac{\partial V_S(\pi,\eta)}{\partial \eta}\right|, \left|\frac{\partial V_{NS}(\pi,\eta)}{\partial \eta}\right| \leq \frac{1}{1-\beta}$$

Also,

$$\frac{\partial V_S(\pi,\eta)}{\partial\eta} = \beta\left[\rho(\pi)\frac{\partial V(\gamma_1(\pi),\eta)}{\partial\eta} + (1-\rho(\pi))\frac{\partial V(\gamma_0(\pi),\eta)}{\partial\eta}\right]$$

and

$$\frac{\partial V_{NS}(\pi,\eta)}{\partial\eta} = 1+\beta\frac{\partial V(q,\eta)}{\partial\eta}.$$

Now taking differences

$$\frac{\partial V_{NS}(\pi,\eta)}{\partial\eta} - \frac{\partial V_S(\pi,\eta)}{\partial\eta} = 1+\beta\frac{\partial V(q,\eta)}{\partial\eta} - \beta\left[\rho(\pi)\frac{\partial V(\gamma_1(\pi),\eta)}{\partial\eta} + (1-\rho(\pi))\frac{\partial V(\gamma_0(\pi),\eta)}{\partial\eta}\right]$$

From Lemma 3, we require the above difference to be nonnegative at $\pi_T(\eta)$. This reduces to following expression.

$$\left[\rho(\pi)\frac{\partial V(\gamma_1(\pi),\eta)}{\partial\eta} + (1-\rho(\pi))\frac{\partial V(\gamma_0(\pi),\eta)}{\partial\eta}\right] - \frac{\partial V(q,\eta)}{\partial\eta} < \frac{1}{\beta}. \tag{6}$$

Note that we can provide upper bound on LHS of above expression and it is upper bounded by $2/(1-\beta)$. If $\beta < 1/3$, Eqn. (6) is satisfied. $\pi_T(\eta)$ is decreasing in η. Thus indexability claim follows.

□

B. *Proof of Lemma 2 - Part* 1)

First we prove convexity of the functions $V_S(\pi), V_{NS}(\pi)$ can be proved using induction. It then follows that $V(\pi)$ is convex. Let

$$\begin{aligned} V_{NS,1}(\pi) &= \eta \\ V_{S,1}(\pi) &= R_S(\pi) = \pi R_0 + (1-\pi)R_1 \\ V_1(\pi) &= \max\{V_{S,1}(\pi), V_{NS,1}(\pi)\}. \end{aligned}$$

Clearly, $V_{NS,1}(\pi), V_{S,1}(\pi)$ and in turn $V(\pi)$ are convex in π. Assume this convexity claim holds for $V_{NS,n}(\pi), V_{S,n}(\pi)$. Now,

$$\begin{aligned} V_{S,n+1}(\pi) = R_S(\pi) &+ \beta\rho(\pi)V_n(\gamma_1(\pi)) \\ &+ \beta(1-\rho(\pi))V_n(\gamma_0(\pi)) \end{aligned}$$

$$\begin{aligned} V_{NS,n+1}(\pi) &= \eta + \beta V_n(\gamma_2(\pi)) \\ V_{n+1}(\pi) &= \max\{V_{S,n+1}(\pi), V_{NS,n+1}(\pi)\}. \end{aligned}$$

Define

$$\begin{aligned} b_0 :=& \quad [(1-\pi)(1-\rho_1)p_{10} + \pi(1-\rho_0)p_{00}, \\ & (1-\pi)(1-\rho_1)(1-p_{10}) + \pi(1-\rho_0)(1-p_{00})] \\ b_1 :=& \quad [(1-\pi)\rho_1 p_{10} + \pi\rho_0 p_{00}, \\ & (1-\pi)\rho_1(1-p_{10}) + \pi\rho_0(1-p_{00})] \\ \|\, b_1 \,\|_1 =& \quad \pi\rho_0 + (1-\pi)\rho_1 = \rho(\pi); \\ \|\, b_0 \,\|_1 =& \quad 1 - \pi\rho_0 - (1-\pi)\rho_1 = 1-\rho(\pi); \end{aligned}$$

Now, $V_{S,n+1}(\pi)$ can be rewritten as

$$V_{S,n+1}(\pi) = R_S(\pi) + \beta||b_1||_1 V_n\left(\frac{b_1}{||b_1||_1}\right) + \beta||b_0||_1 V_n\left(\frac{b_0}{||b_0||_1}\right)$$

We know that $V_n(\pi)$ is convex. Using Lemma 2 from [23], $||b_1||_1 V_n\left(\frac{b_1}{||b_1||_1}\right)$ is also convex. This implies that $V_{S,n+1}$ is a sum of convex functions and hence convex.

$$V_{NS,n+1}(\pi) = \eta + \beta V_n(\gamma_2(\pi))$$

Here, $V_n(\pi)$ is convex and $\gamma_2(\pi)$ is linear. Hence, $V_{NS,n+1}(\pi)$ is convex. It follows that $V_{n+1}(\pi)$ is convex. By principle of induction, $V_{S,n}(\pi), V_{NS,n}(\pi)$ and $V_n(\pi)$ are convex for all n. From [24] Chapter 2, as $n \to \infty$, $V_{S,n}(\pi) \to V_S(\pi)$, $V_{NS,n}(\pi) \to V_{NS}(\pi)$ and $V_n(\pi) \to V(\pi)$. This means that the functions V_S, V_{NS}, V are convex in π.

C. *Proof of Lemma 2 - Part* 2)

This result too can be claimed using the induction principle. To emphasize that subsidy η is a variable, value functions are rewritten as $V_S(\pi,\eta), V_{NS}(\pi)$ and $V(\pi,\eta)$. For a fixed π, let

$$\begin{aligned} V_{NS,1}(\pi,\eta) &= \eta \\ V_{S,1}(\pi,\eta) &= R_S(\pi) = \pi R_0 + (1-\pi)R_1 \\ V_1(\pi,\eta) &= \max\{R_S(\pi), \eta\}. \end{aligned}$$

Clearly, all the above functions are convex and non-decreasing in η. Now suppose $V_{S,n}(\pi,\eta), V_{NS,n}(\pi,\eta)$ and in turn $V_n(\pi,\eta)$ are convex.

$$V_{NS,n+1}(\pi,\eta) = \eta + \beta V_n(\gamma_2(\pi),\eta)$$

$$\begin{aligned} V_{S,n+1}(\pi,\eta) = R_S(\pi) + \beta\,(\rho(\pi)&V_n(\gamma_1(\pi),\eta) \\ &+ (1-\rho(\pi))V_n(\gamma_0(\pi),\eta)) \end{aligned}$$

$$V_{n+1}(\pi,\eta) = \max\{V_{S,n+1}(\pi), V_{NS,n+1}(\pi)\}.$$

Here, $V_{NS,n+1}(\pi,\eta)$ is non-decreasing convex in η because it is a sum of two non-decreasing convex functions in η. Further, $V_{S,n+1}(\pi,\eta)$ is sum of a constant function and a convex combination of two non-decreasing convex functions; hence it is convex non-decreaing. By induction $V_{S,n}, V_{NS,n}$ and V_n are non-decreasing convex for any $n \geq 1$. As in part 1) of this lemma, as $n \to \infty$, $V_{S,n}(\pi,\eta) \to V_S(\pi,\eta)$, $V_{NS,n}(\pi,\eta) \to V_{NS}(\pi,\eta)$ and $V_n(\pi,\eta) \to V(\pi,\eta)$. This means that the functions V_S, V_{NS}, V are convex and non-decreasing in η for fixed π. □

D. *Proof of Lemma 2 - Part* 3)

The proof is done by induction technique. Assume that $V_n(\pi)$ is non increasing in π. Let $\pi' > \pi$ and consider playing the arm is optimal. Then

$$V_{n+1}(\pi) \geq R_S(\pi) + \beta\left[\rho(\pi)V_n(\gamma_1(\pi)) + (1-\rho(\pi))V_n(\gamma_0(\pi))\right]$$

Here $R_S(\pi) = \pi R_0 + (1-\pi)R_1$. Note that $R_S(\pi)$ is decreasing in π, i.e. $R_S(\pi') < R_S(\pi)$ whenever $\pi' > \pi$. Hence we get

$$V_{n+1}(\pi) \geq R_S(\pi') + \beta\left[\rho(\pi)V_n(\gamma_1(\pi)) + (1-\rho(\pi))V_n(\gamma_0(\pi))\right].$$

From our assumptions $p_{00} > p_{10}$ and $\rho_1 > \rho_0$ we get stochastic ordering on observation probability, i.e., $[1-\rho(\pi), \rho(\pi)]^T \geq_s [1-\rho(\pi'), \rho(\pi')]^T$. Then, using a property of stochastic ordering [22, Lemma 1.1], we obtain

$$V_{n+1}(\pi) \geq R_S(\pi') + \beta\left[\rho(\pi')V_n(\gamma_1(\pi)) + (1-\rho(\pi'))V_n(\gamma_0(\pi))\right].$$

Now that γ_0, γ_1 are increasing in π and V_n is decreasing in π, we have

$$V_{n+1}(\pi) \geq R_S(\pi')+\beta\left[\rho(\pi')V_n(\gamma_1(\pi')) + (1-\rho(\pi'))V_n(\gamma_0(\pi'))\right].$$

$$V_{n+1}(\pi) \geq V_{n+1}(\pi')$$

This is true for every n. From [21], we know $V_n(\pi) \to V(\pi)$ uniformly. Thus $V(\pi)$ is decreasing in π. Similarly, we can derive proof for $V_S(\pi)$ and $V_{NS}(\pi)$.

□

E. *Proof of Lemma 2 - Part* 4)

Let $d(\pi) := V_S(\pi) - V_{NS}(\pi)$. We want to prove that $d(\pi)$ decreasing in π. This implies that we need to show

$$V_S(\pi) - V_{NS}(\pi) < V_S(\pi') - V_{NS}(\pi'), \text{ whenever } \pi > \pi'. \tag{7}$$

We can rewrite (7) as follows.

$$V_S(\pi) - V_S(\pi') < V_{NS}(\pi) - V_{NS}(\pi'), \tag{8}$$

In our setting $V_{NS}(\pi) - V_{NS}(\pi') = 0$ whenever $\gamma_2(\pi) = q$ and this is true for large values of k. We know that $V_S(\pi) - V_S(\pi') < 0$, as V_S is decreasing in π. Hence Eqn. (8) is true and claim follows.

□

F. *Index Computation*

1) We first compute index for $\pi \in A_1$. We have following expressions.

$$\begin{aligned} V_S(\pi) &= \rho(\pi) + \beta\rho(\pi)V(\gamma_1(\pi)) + \beta(1-\rho(\pi))V(\gamma_0(\pi)) \\ V_{NS}(\pi) &= \eta + \beta V(q) \end{aligned}$$

For $\pi \in A_1$, we also have $V_{S,\beta}(\pi,\eta) = V_{NS,\beta}(\pi,\eta)$ because of a threshold type policy. Solving for η we can obtain index. For this case, we note that $q > \pi$ and hence $V(q) = V_{NS}(q) = \eta + \beta V(q)$, and $V(q) = \frac{\eta}{1-\beta}$. This implies that $V_{NS}(\pi) = \frac{\eta}{1-\beta}$. Also, note that $V(\gamma_1(\pi)) = V_{NS}(\gamma_1(\pi))$, and $V(\gamma_0(\pi)) = V_{NS}(\gamma_0(\pi))$. Hence $V_S(\pi) = \rho(\pi) + \beta\frac{\eta}{1-\beta}$, and $V_{NS}(\pi) = \frac{\eta}{1-\beta}$. By equating these expressions and solving for η we obtain $W(\pi) = \rho(\pi)$.

2) For $\pi \in A_2$ we have $V_{S,\beta}(\pi,\eta) = V_{NS,\beta}(\pi,\eta)$ and solving for η we can obtain index. We also use condition $\rho_0 = 0$. This implies that $\gamma_1(\pi) = p_{1,0}$.
We first show that $V(q) = \frac{\eta}{1-\beta}$ in this case. We observe that $V(q) = \max\{V_S(q), V_{NS}(q)\} = V_{NS}(q)$. After simplification we obtain $V(q) = \frac{\eta}{1-\beta}$.

   a) We first derive index for $\gamma_0(p_{1,0}) \geq \pi$. Since $p_{1,0} < \pi$, we have $V(p_{1,0}) = V_S(p_{1,0})$. Hence we first require to compute $V_S(p_{1,0})$. We also have $\gamma_0(p_{1,0}) \geq \pi$ and hence

$$V(\gamma_0(p_{1,0})) = V_{NS}(\gamma_0(p_{1,0})) = \frac{\eta}{1-\beta}.$$

Then we can write $V_S(p_{1,0})$ as follows.

$$\begin{aligned} V_S(p_{1,0}) = \rho(p_{1,0}) + \beta\rho(p_{1,0})V(p_{1,0}) + \\ \beta(1-\rho(p_{1,0}))V_{NS}(\gamma_0(p_{1,0})). \end{aligned}$$

We further simplify and obtain

$$V_S(p_{1,0}) = \frac{\rho(p_{1,0})}{1-\beta\rho(p_{1,0})} + \frac{\beta(1-\rho(p_{1,0}))}{1-\beta\rho(p_{1,0})}\frac{\eta}{1-\beta}.$$

Then we can rewrite $V_S(\pi)$ in the following way.

$$\begin{aligned} V_S(\pi) &= \rho(\pi) \\ &+ \beta\rho(\pi)\left[\frac{\rho(p_{1,0})}{1-\beta\rho(p_{1,0})} + \frac{\beta(1-\rho(p_{1,0}))}{1-\beta\rho(p_{1,0})}\ \frac{\eta}{1-\beta}\right] \\ &+ \beta(1-\rho(\pi))\frac{\eta}{1-\beta}. \end{aligned}$$

After simplification we get

$$\begin{aligned} V_S(\pi) &= \rho(\pi) + \beta(1-\rho(\pi))\frac{\eta}{1-\beta} \\ &+ \frac{\beta\rho(\pi)\left[\rho(p_{1,0})(1-\beta) + \beta(1-\rho(p_{1,0}))\eta\right]}{(1-\beta\rho(p_{1,0}))(1-\beta)}. \end{aligned}$$

Further,

$$V_{NS}(\pi) = \frac{\eta}{1-\beta}.$$

Equating the preceding two equations, we obtain

$$\begin{aligned} &\rho(\pi) + \frac{\beta\rho(\pi)\rho(p_{1,0})(1-\beta)}{(1-\beta\rho(p_{1,0}))} \\ &= -\beta(1-\rho(p_{1,0}))\eta \times \frac{1}{(1-\beta\rho(p_{1,0}))(1-\beta)} \\ &\quad - \beta(1-\rho(\pi))\frac{\eta}{1-\beta} + \frac{\eta}{1-\beta}. \end{aligned}$$

After simplification, we get

$$\begin{aligned} &\rho(\pi) + \frac{\beta\rho(\pi)\rho(p_{1,0})(1-\beta)}{(1-\beta\rho(p_{1,0}))} \\ &= \frac{\eta}{1-\beta}\left[1 - \frac{\beta(1-\rho(p_{1,0}))}{(1-\beta\rho(p_{1,0}))} - \beta(1-\rho(\pi))\right]. \end{aligned}$$

After further simplification we get

$$W(\pi) = \frac{\rho(\pi)}{1-\beta(\rho(p_{1,0}) - \rho(\pi))}.$$

   b) We now consider the case of $\gamma_0(p_{1,0}) < \pi$ but $\gamma_0^2(p_{1,0}) \geq \pi$. From our assumptions we can obtain the following expressions.

$$\begin{aligned} V(\gamma_0(p_{1,0})) &= V_S(\gamma_0(p_{1,0})), \\ V(\gamma_0^2(p_{1,0})) &= V_{NS}(\gamma_0^2(p_{1,0})) = \frac{\eta}{1-\beta}. \end{aligned}$$

Thus we can write

$$\begin{aligned} V_S(\gamma_0(p_{1,0})) = \rho(\gamma_0(p_{1,0})) + \beta\rho(\gamma_0(p_{1,0}))V(p_{1,0}) \\ + \beta(1-\rho(\gamma_0(p_{1,0})))V(\gamma_0^2(p_{1,0})). \end{aligned}$$

After substituting the value of $V(\gamma_0^2(p_{1,0}))$ we get

$$\begin{aligned} V_S(\gamma_0(p_{1,0})) = \rho(\gamma_0(p_{1,0})) + \beta\rho(\gamma_0(p_{1,0}))V(p_{1,0}) \\ + \beta(1-\rho(\gamma_0(p_{1,0})))\frac{\eta}{1-\beta}. \end{aligned} \tag{9}$$

We can compute $V(p_{1,0}) = V_S(p_{1,0})$ and obtain following expression.

$$\begin{aligned} V_S(p_{1,0}) = \rho(p_{1,0}) + \beta\rho(p_{1,0})V_S(p_{1,0}) + \\ \beta(1-\rho(p_{1,0}))V_S(\gamma_0(p_{1,0})). \end{aligned} \tag{10}$$

After substituting Eqn. (9) in Eqn. (10), we obtain

$$\begin{aligned} V_S(p_{1,0}) &= \rho(p_{1,0}) + \beta\rho(p_{1,0})V_S(p_{1,0}) + \\ \beta(1-\rho(p_{1,0}))&\Big[\rho(\gamma_0(p_{1,0})) + \beta\rho(\gamma_0(p_{1,0}))V(p_{1,0}) \\ &+\beta(1-\rho(\gamma_0(p_{1,0})))\frac{\eta}{1-\beta}\Big]. \end{aligned} \tag{11}$$

After solving Eqn. 11, we can obtain expression of $V_S(p_{1,0})$.

$$V_S(p_{1,0}) = \frac{T_1}{T_2}, \tag{12}$$

where

$$\begin{aligned} T_1 = \rho(p_{1,0}) + \beta(1-\rho(p_{1,0}))\rho(\gamma_0(p_{1,0})) + \\ \beta^2\overline{\rho}(p_{1,0})\overline{\rho}(\gamma_0(p_{1,0}))\frac{\eta}{1-\beta}, \end{aligned}$$

$$\begin{aligned} T_2 &= 1 - \beta\rho(p_{1,0}) - \beta^2\overline{\rho}(p_{1,0})\gamma_0(p_{1,0}), \\ \overline{\rho}(p_{1,0}) &= 1-\rho(p_{1,0}), \\ \overline{\rho}(\gamma_0(p_{1,0})) &= 1-\rho(\gamma_0(p_{1,0})). \end{aligned}$$

We now rewrite $V_S(\pi)$ as follows.

$$\begin{aligned} V_S(\pi) = \rho(\pi) + \beta\rho(\pi)V_S(p_{1,0}) + \\ \beta(1-\rho(\pi))V_{NS}(\gamma_0(\pi)). \end{aligned} \tag{13}$$

Equating $V_S(\pi)$ and $V_{NS}(\pi)$ and solving for η, we obtain required $W(\pi)$.

3) For $\pi \in A_3$, computing index expression is non-trivial because it is tedious to compute the value function expression for $V_S(p_{1,0})$. Hence we use value iteration algorithm.
4) For $\pi \in A_4$ we have to obtain value function expressions for $V_{S,\beta}(\pi,\eta)$ and $V_{NS,\beta}(\pi,\eta)$. Note that to compute $V_{S,\beta}(\pi,\eta)$, we need to compute $V(p_{1,0}) = V_S(p_{1,0})$ and $V(q) = V_S(q)$. In this case the optimal action for the π is not sample the arm once and later sample the arm always. Similarly if the initial action is to sample the arm and later the optimal action is to sample the arm always. This behavior can be observed from the operation $\gamma_0(\pi)$, which is smaller than $p_{0,0}$. Then one can easily show by using induction technique that the $V_S(\pi)$ is linear in π and similarly $V_{NS}(\pi)$ is also linear in π with slope m and intercept c as mentioned earlier. That is,

$$\begin{aligned} V_S(p_{1,0}) &= mp_{1,0} + c \\ V_S(q) &= mq + c. \end{aligned}$$

Therefore,

$$V_{NS}(\pi) = \eta + \beta(mq+c)$$

and

$$V_S(\pi) = \rho(\pi) + \beta\rho(\pi)(mq+c) + \beta(1-\rho(\pi))(m\gamma_0(\pi)+c)$$

$$\begin{aligned} V_S(\pi) = \rho(\pi) + \beta\rho(\pi)(mq+c) + \beta\,\{(m\gamma_0(\pi)+c) \\ -\rho(\pi)(m\gamma_0(\pi)+c))\} \end{aligned}$$

After some simplification

$$V_S(\pi) = \rho(\pi) + \beta\rho(\pi)\left[m(q-\gamma_0(\pi))\right] + \beta(m\gamma_0(\pi)+c)$$

Equating $V_S(\pi)$ and $V_{NS}(\pi)$ and solving for η we get required $W(\pi)$.