2015 IEEE 54th Annual Conference on Decision and Control (CDC)
December 15-18, 2015. Osaka, Japan

A Restless Bandit With No Observable States for Recommendation Systems and Communication Link Scheduling

Rahul Meshram and D. Manjunath
Deptt. of Elecl. Engg.
IIT Bombay, Mumbai INDIA

Aditya Gopalan
Deptt. of Elecl. Commun. Engg.
Indian Inst. of Science, Bangalore INDIA.

***Abstract*— A restless bandit is used to model a user's interest in a topic or item. The interest evolves as a Markov chain whose transition probabilities depend on the action (display the ad or desist) in a time step. A unit reward is obtained if the ad is displayed and if the user clicks on the ad. If no ad is displayed then a fixed reward is assumed. The probability of click-through is determined by the state of the Markov chain. The recommender never gets to observe the state but in each time step it has a belief, denoted by π_t, about the state of the Markov chain. π_t evolves as a function of the action and the signal from each state. For the one-armed restless bandit with two states, we characterize the policy that maximizes the infinite horizon discounted reward. We first characterize the value function as a function of the system parameters and then characterize the optimal policies for different ranges of the parameters. We will see that the Gilbert-Elliot channel in which the two states have different success probabilities becomes a special case. For one special case, we argue that the optimal policy is of the threshold type with one threshold; extensive numerical results indicate that this may be true in general.**

I. Introduction

Consider an item that is available to be recommended to a user, or an ad that is is available to be displayed. Different users react differently to the frequency with which a given item is shown to them. Some users may be annoyed if something is offered repeatedly, which in turn may make it less likely that they will click-through if they were shown this itnem in the recent past. Alternatively, it could be that the user is at first disinterested and develops a curiosity, possibly turning into a propensity to click-through, if the object is recommended frequently. Other users may be more random and their behavior may not depend on the history of the showing of the item. One way to capture the effect of the history of showing an ad would be to assume that the user behavior is represented by a Markov chain with each state corresponding to the user's 'taste' and hence a different click-through probability. The state transitions occur at the beginning of each session in which a fresh item is shown. In each session, the recommendation system (RS) has two options—display or desist. The Markov chain changes state according to a transition probability matrix that depends whether the item was shown or was not shown. The values of these transition probabilities determine the 'type' of the user. If the RS has more than one ad or item that it can show, and wants to use a policy that maximizes a suitably defined reward, a reasonable model for the system is that of a restless multiarmed bandit. Further, the only signal available to the RS from each display of the ad is whether there was click-through event or not. The actual state of the user is not observable by the RS. In this paper, we describe a restless one-armed bandit model that captures the behavior of the preceding system in which there is just one item. This is the first step in analyzing the multiarmed system.

For a second motivating application, consider a set of communication links that are abstracted by the Gilbert-Elliot model [1], [2]. Here the link is in one of two states, say 0 and 1. Let $r_i > 0$ be the probability of a successful transmission when the link is in state i, $i = 0, 1$. Assume that there is an alternate link that has a fixed probability of success, say r_2. The state of the first link evolves according to a Markov chain and in each slot the transmitter needs to decide which of the two channels it is going to use to optimize a suitable objective function, e.g., the number of bits transmitted through the channel. Note that in practical scenarios, the transmitter obtains only the result of the transmission and cannot measure the state. Since the channel state is evolving independent of the transmissions, the restless bandit is the appropriate model. Further, the transmitter only knows the result of its transmission and cannot observe the state of the channel. Thus this may be viewed as a special case of the recommendation system, in that the transition probabilities between states of a channel do not depend on the action taken, i.e., to transmit or not to transmit.

Motivated by the preceding examples, we analyze a one-armed restless bandit when the state of the bandit is never observed by the sampling process or the controller. This is the first step towards building and analyzing the performance of a general restless multiarmed bandit model.

A. Related Work

Restless bandits with unobserved states are a special case of partially observed Markov decision processes (POMDPs) and have been studied to model several systems. Early work is in the context of machine repair problems. In [3], a machine is modeled as a two-state Markov chain with three actions and it is shown that the optimal policy is of the threshold type with three thresholds. In [4], a similar model is considered and formulas for the optimal costs and policy are obtained. This and some additional models are considered in [5] and, once again, several structural results are obtained. Also see [6] for more models in this regard.

The work of Rahul Meshram and D. Manjunath was carried out in the Bharti Centre for Communications at IIT Bombay. D. Manjunath is also supported by grants from CEFIPRA and DST.

Recent interest in restless bandit models has been motivated by scheduling in communication systems. A Gilbert-Elliot channel in which the two states can sustain different rates is considered in [7]. Here, in each slot, the transmitter has to choose from three actions—transmit at a low rate, at high rate, or spend some time and probe to obtain accurate state information before transmitting. They characterize the optimal policies and evaluate the objective function. For a similar system, the conditions under which myopic policies are optimal is investigated in [8]; similar results for multiarmed bandit problems are obtained in [9]. In [10], scheduling in a dynamic spectrum access system is modeled as a restless multiarmed bandit problem. Each channel is represented by an arm of the bandit and the channels are assumed to be Gilbert-Elliot channels with the bad state yielding a loss and the good state yielding a success each with probability one. Further, it is assumed that the transition probabilities between the states are known. The authors show that the single armed bandit has a threshold policy and then proceed to analyze the multiarmed system. See [11], [12] for variants of this model.

Multiarmed bandit models for recommendation systems and for online advertising have been modeled as contextual bandits, e.g., [13], [14], [15] and the user interests are assumed to be independent of the recommendation history. To the best of our knowledge, models in which the probability of a success depends on the state of the user have not been studied in this literature.

B. Our Contribution

The key modeling assumption in this paper is that the actual state is never observed, only the signal from the state is observed when the arm is sampled. Like in the literature discussed above, we also consider an infinite horizon, discounted reward problem.

We analyze the restless one-armed bandit model in which the arm can be in one two states and the state is never observed. Instead, a belief π about the state can be constructed based on past observations, which serves as a sufficient statistic. In each time step, one of two actions can be taken and the belief evolves based on the action taken and the reward observed. Our interest is in maximization of the expected discounted reward over the infinite horizon. We obtain structural properties of the value functions in terms of the parameters and also discuss the structure of the optimal policy for several special cases of the parameter values. For one special case, we argue that the optimal policy is of threshold type and that there is a single threshold. We also present some numerical results that suggest that a threshold type optimal policy is more generally true. The next section describes the model. In Section III we describe the general results that are applicable for all values of the parameters. In Section IV we refine some of the results for some special cases of the parameters. For a specific case, we argue the possible optimality of the threshold policy. We conclude with some numerical results and a discussion on directions for future work in Section V.

II. Model Description and Preliminaries

A one arm bandit is represented as a two-state Markov chain. X_t denotes the state at the beginning of time slot t with $X_t \in \{0,1\}$. $a_t \in \{0,1\}$ is the action in slot t with the following interpretation.

$$a_t = \begin{cases} 1 & \text{Markov chain is sampled} \\ 0 & \text{Markov chain is not sampled} \end{cases}$$

Sampling corresponds to displaying the item or transmitting the packet, and not sampling corresponds to not displaying the item or using the second channel that has deterministic characteristics. The Markov chain changes state at the end of each slot according to transition probabilities that can depend on a_t. $P_{ij}(a)$ is the transition probability from state i to state j under action a. Define the parameters λ_i and μ_i, $i = 0, 1$ as

$$P_{00}(0) = \lambda_0, \quad P_{10}(0) = \lambda_1,$$
$$P_{00}(1) = \mu_0, \quad P_{10}(1) = \mu_1.$$

$R_t(i,a)$ is the reward in slot t when the Markov chain is in state i and the action is a. We let

$$R_t(i,1) = r_i \quad R_t(i,0) = r_2.$$

We will assume $0 \leq r_0 < r_1 \leq 1$ and not place any restriction on r_2.

Let π_t denote the belief that the state of the Markov chain is in state 0 at the beginning of slot t, i.e., $\pi_t = \mathsf{Pr}\left(X_t = 0\right)$. π_{t+1} is a function of π_t, a_t and R_t. From Bayes' theorem, the following can be derived. If $a_t = 1$, i.e., the arm is sampled, and $R_t = 0$ then

$$\pi_{t+1} = \gamma_0(\pi_t) := \frac{\pi_t(1-r_0)\mu_0 + (1-\pi_t)(1-r_1)\mu_1}{\pi_t(1-r_0) + (1-\pi_t)(1-r_1)}$$

If $a_t = 1$ and $R_t = 1$ then

$$\pi_{t+1} = \gamma_1(\pi_t) := \frac{\pi_t r_0\mu_0 + (1-\pi_t)r_1\mu_1}{\pi_t r_0 + (1-\pi_t)r_1}$$

Finally, if $a_t = 0$, i.e., the arm is not sampled at t, then

$$\pi_{t+1} = \gamma_2(\pi_t) := \pi_t\lambda_0 + (1-\pi_t)\lambda_1$$

Using first and second derivatives the following properties of $\gamma_i(\pi)$ are straightforward.

Property 1:
1) If $\lambda_0 < \lambda_1$ then $\gamma_2(\pi)$ is linear decreasing in π. Further, $\lambda_0 \leq \gamma_2(\pi) \leq \lambda_1$.
2) If $\lambda_0 > \lambda_1$ then $\gamma_2(\pi)$ is linear increasing in π. Further, $\lambda_1 \leq \gamma_2(\pi) \leq \lambda_0$.
3) If $\mu_0 > \mu_1$ then $\gamma_1(\pi)$ is convex increasing in π. Further, $\mu_1 \leq \gamma_1(\pi) \leq \mu_0$.
4) If $\mu_0 > \mu_1$ then $\gamma_0(\pi)$ is concave increasing in π. Further, $\mu_1 \leq \gamma_0(\pi) \leq \mu_0$.
5) $\gamma_0(0) = \gamma_1(0) = \mu_1$ and $\gamma_0(1) = \gamma_1(1) = \mu_0$. Furthe, if $\mu_0 > \mu_1$ then $\gamma_1(\pi) < \gamma_0(\pi)$ for $0 < \pi < 1$.

Let H_t denote the history of actions and rewards up to time t. Let $\phi_t : H_t \to \{0,1\}$ be the strategy that determines the action at time t. For a strategy ϕ, a discount factor β, $0 < \beta < 1$, and an initial belief π at time $t = 1$ (i.e.,

$\mathsf{Pr}(X_1 = 0) = \pi$), the expected infinite horizon discounted reward is

$$V_\phi(\pi) := E\left\{\sum_{t=1}^{\infty} \beta^{t-1} R_t(X_t, \phi(H_t))\right\}.$$

where E denotes the expectation operator. Our interest is in a strategy that maximizes $V_\phi(\pi)$ for all π, $0 \leq \pi \leq 1$. The following facts are well-known from the theory of partially observable Markov decision processes [3], [16], [17]. π_t captures the information in H_t, control strategies can be restricted to stationary Markov policies, and the optimum objective function, $V(\pi)$, solves the following dynamic program

$$\begin{aligned} V(\pi) &= \max\{\rho(\pi) + \beta\left(\rho(\pi)V(\gamma_1(\pi)) + (1-\rho(\pi))\times \right. \\ &\qquad \left. V(\gamma_0(\pi))\right), \quad r_2 + \beta V(\gamma_2(\pi))\}, \end{aligned} \tag{1}$$

where we denote $\rho(\pi) = \pi r_0 + (1-\pi) r_1$. Let $V_S(\pi)$ be the optimal value of the objective function if $a_1 = 1$, i.e., if the Markov chain is sampled, and $V_{NS}(\pi)$ be the optimal value for $a_1 = 0$, for not sampling it. In other words, we can write the following.

$$\begin{aligned} V_S(\pi) &= \rho(\pi) + \beta\left(\rho(\pi)V(\gamma_1(\pi)) + (1-\rho(\pi))V(\gamma_0(\pi))\right) \\ V_{NS}(\pi) &= r_2 + \beta V(\gamma_2(\pi)) \\ V(\pi) &= \max\{V_S(\pi), V_{NS}(\pi)\} \end{aligned} \tag{2}$$

The following background lemma from [18] will be useful in our analysis.

Lemma 1 ([18]): If $f : \Re_+^n \to \Re_+$ is a convex function then for $x \in \Re_+^n$, $g(x) := ||x||_1 f\left(\frac{x}{||x||_1}\right)$ is also a convex function.

III. General Results

In this section we present some results that are applicable for all values of the parameters. We begin by showing that $V(\pi)$, $V_S(\pi)$, and $V_{NS}(\pi)$ are all convex. Proof is by induction and is omitted.

Property 2: (Convexity of value functions over the belief state) $V(\pi)$, $V_{NS}(\pi)$ and $V_S(\pi)$ are all convex functions of $\pi \in [0,1]$.

Although the above may be anticipated because of similar results in the literature, the non observability of the state and the consequent structure of the γs necessitates the proof. Next we characterize $V(\pi)$ as a function of r_2.

Property 3: (Convexity of value functions over passive reward) For a fixed π, $V(\pi, r_2)$, $V_S(\pi, r_2)$, and $V_{NS}(\pi, r_2)$ are non decreasing and convex in r_2.

Once again, proof is by induction and is omitted. r_2 plays a central role in the development of the Whittle's index in a multiarmed bandit model where each arm is modeled as in this paper.

The following property characterizes the optimal policy as a function of r_2 and π.

Property 4: (Optimal policy for extremal values of passive reward)

1) If $r_2 = \infty$ then not sampling is the optimal action for all π. Further, for this case

$$V_{NS}(\pi) \;=\; V(\pi) \;=\; \frac{r_2}{1-\beta} \tag{3}$$

2) If $r_2 = -\infty$ then sampling is the optimal action for all π. Further, in this case

$$\begin{aligned} V(\pi, -\infty) &= V_S(\pi, -\infty) = m\pi + c \\ m &= \frac{r_0 - r_1}{1 - \beta(\mu_0 - \mu_1)} \\ c &= \frac{r_1 + \frac{\beta\mu_1(r_0 - r_1)}{1-\beta(\mu_0-\mu_1)}}{1-\beta}. \end{aligned} \tag{4}$$

The optimal action for both these cases is straightforward. (3) is obtained by simplifying (2) while the linear form of the value function in (4) is derived in the appendix. (3) is also the optimal value function for all r_2 for which $a_1 = 0$ is the optimal action for all π. In this case we also have the value function when $a_1 = 1$ to be

$$V_S(\pi) = \pi r_0 + (1-\pi) r_1 + \frac{\beta r_2}{1-\beta}. \tag{5}$$

Observe that in the preceding, $V_{NS}(\pi)$ is constant and $V_S(\pi)$ is decreasing in π. Similarly, (5) is the optimal value function for all r_2 for which $a_1 = 1$ is the optimal action for all π. Also, in this case, the value function when $a_1 = 0$ is

$$\begin{aligned} V_{NS}(\pi) &= r_2 + \beta V(\gamma_2(\pi)) \\ &= \left[\beta m(\lambda_0 - \lambda_1)\right]\pi + \left[r_2 + \beta(c + \lambda_1 m)\right]. \end{aligned} \tag{6}$$

We see that $V_{NS}(\pi)$ is also linear in π.

Remark 1: For r_2 a large negative value $V_S(\pi, r_2) > V_{NS}(\pi, r_2)$ for all π. And, for r_2 a large positive value, $V_S(\pi, r_2) < V_{NS}(\pi, r_2)$. Thus, for a fixed π, $V_S(\pi, r_2)$ and $V_{NS}(\pi, r_2)$ intersect at least once.

We first define

$$\tilde{r}_2(\pi) \;:=\; \{r_2 : V_{NS}(\pi, r_2) = V_S(\pi, r_2)\},$$

$$\begin{aligned} r_L &:= \min_{\pi\in[0,1]} \tilde{r}_2(\pi), \qquad r_H := \max_{\pi\in[0,1]} \tilde{r}_2(\pi), \\ \pi_H &:= \arg\max_{\pi\in[0,1]} \tilde{r}_2(\pi), \qquad \pi_L := \arg\min_{\pi\in[0,1]} \tilde{r}_2(\pi). \end{aligned}$$

From the definition, for $r_2 > r_H$, not sampling is the optimal policy and for $r_2 < r_L$, sampling is optimal for all $\pi \in [0,1]$. If $r_2 = r_H$ and $\pi = \pi_H$, or if $r_2 = r_L$ and $\pi = \pi_L$, then both actions are optimal. From this we see that for a fixed π, $V_S(\pi, r_2)$ and $V_{NS}(\pi, r_2)$ do not intersect for $r_2 \in [-\infty, r_L)$ or for $r_2 \in (r_H, \infty]$ for any $\pi \in [0,1]$. This in turn means that $V_S(\pi, r_2)$ and $V_{NS}(\pi, r_2)$ intersect for $r_2 \in [r_L, r_H]$ for every $\pi \in [0,1]$. The following obtains r_H, r_L, π_H and π_L.

Property 5: (Range of the passive reward resulting in constant optimal policies)

$$\begin{aligned} r_H &= r_1 \qquad & r_L &= r_1 + \frac{(r_0 - r_1)\left[1 - \beta(\lambda_0 - \mu_1)\right]}{1 - \beta(\mu_0 - \mu_1)} \\ \pi_H &= 0 \qquad & \pi_L &= 1 \end{aligned}$$

Proof: We first obtain r_H and π_H by equating the RHS of (3) and (5). This gives us $r_2 = \pi r_0 + (1-\pi)r_1$ which achieves its maximum at $\pi = 0$ corresponding to $r_H = r_1$.

To obtain r_L and π_L, we equate the RHS of (6) and (4) and obtain

$$\begin{aligned} r_2 &= \frac{(r_0 - r_1)(1-\beta(\lambda_0-\lambda_1))}{1-\beta(\mu_0-\mu_1)}\pi + \\ & \quad r_1 + \frac{\beta(r_0-r_1)(\mu_1-\lambda_1)}{1-\beta(\mu_0-\mu_1)}. \end{aligned}$$

Since $0 < \lambda_0, \lambda_1 < 1$, $|\lambda_0 - \lambda_1| < 1$. Similarly, $|\mu_0 - \mu_1| < 1$. Hence, from our assumption that $r_0 < r_1$, the coefficient of π is always negative for sufficiently small β. Thus the minimum value of the RHS of the preceding equation is achieved at $\pi = 1$ corresponding to $\pi_L = 1$ and r_H as in the statement of the property. ∎

The following lemma follows from the preceding results.

Lemma 2: (Range of the passive reward resulting in constant optimal policies)

1) If $r_2 < r_L$ then sampling is the optimal policy for all π.
2) If $r_2 > r_H$ then not sampling is the optimal policy for all π.

IV. Special Cases

We begin by characterizing the value function for the case when $\mu_0 > \mu_1$ and $\lambda_0 > \lambda_1$ in the following lemma.

Lemma 3: If $\mu_0 > \mu_1$ and $\lambda_0 > \lambda_1$, then

1) $V(\pi, r_2)$ and $V_{NS}(\pi, r_2)$ are non increasing functions of π for a fixed r_2.
2) $V_S(\pi, r_2)$ is a decreasing function of π for a fixed r_2.

Next, consider the case when $\mu_0 \geq \lambda_0, \lambda_1 \geq \mu_1$. The following lemma refines the policy characterization of Lemma 2.

Lemma 4: For $\mu_0 \geq \mu_1$, and $\lambda_1, \lambda_2 \in [\mu_1, \mu_0]$, if $r_2 > \rho(\mu_1)$ and $\mu_1 \leq \pi \leq 1$, then the optimal policy is to not sample.

Proof: First, we see that $r_2 > \rho(\pi)$ for every $\mu_1 \leq \pi \leq 1$. The proof is by induction over the n-step value function, followed by a limiting argument. Let $V_{S,1}(\pi) = \rho(\pi)$, $V_{NS,1}(\pi) = r_2$. Then $V_1(\pi) = \max\{V_{S,1}(\pi), V_{NS,1}(\pi)\} = V_{NS,1}(\pi) = r_2$. Writing

$$\begin{aligned} V_{n+1}(\pi) &= \max\{V_{S,n+1}(\pi), V_{NS,n+1}(\pi)\}, \\ V_{S,n+1}(\pi) &= \rho(\pi) + \beta\rho(\pi)V_n(\gamma_1(\pi)), \\ &\quad +\beta(1-\rho(\pi))V_n(\gamma_0(\pi)) \\ V_{NS,n+1}(\pi) &= r_2 + \beta V_n(\gamma_2(\pi)), \end{aligned} \tag{7}$$

making the induction hypothesis

$$V_n(\pi) = V_{NS,n}(\pi) = r_2 + \beta V_{n-1}(\gamma_2(\pi)),$$

and recursing, we obtain

$$V_n(\pi) = \frac{1-\beta^{n+1}}{1-\beta} r_2.$$

Substituting this $V_n(\pi)$ in (7), we get

$$\begin{aligned} V_{S,n+1}(\pi) &= \rho(\pi) + \beta\left(\frac{1-\beta^{n+1}}{1-\beta} r_2\right) \\ V_{NS,n+1}(\pi) &= r_2 + \beta\left(\frac{1-\beta^{n+1}}{1-\beta} r_2\right). \end{aligned}$$

Since $r_2 > \rho(\pi)$, $V_{S,n+1}(\pi) < V_{NS,n+1}(\pi)$ and hence $V_{n+1}(\pi) = V_{NS,n+1}(\pi) = \frac{1-\beta^{n+2}}{1-\beta} r_2$. Thus by induction, $V_n(\pi) = V_{NS,n}(\pi)$ for all $n \geq 1$. From [3], [16], [17], we know by the contractivity of the Bellman operator (5), that

$$\begin{aligned} \lim_{n\to\infty} V_n(\pi) &= V(\pi) \\ \lim_{n\to\infty} V_{NS,n}(\pi) &= V_{NS}(\pi) \\ \lim_{n\to\infty} V_n(\pi) &= V(\pi) = V_{NS}(\pi) = \frac{r_2}{1-\beta}. \end{aligned}$$

∎

Remark 2: Observe that under the conditions of the preceding lemma, for all $0 \leq \pi \leq 1$, $\gamma_0(\pi)$, $\gamma_1(\pi)$ and $\gamma_2(\pi)$ are all in $[\mu_1, 1]$. Thus irrespective of the initial belief, π_1, the belief in the second slot, π_2, will be in $[\mu_1, 1]$. Thus in the second slot, we enter the regime of the lemma. Thus in this case the optimal policy will sample at most once which will be in the first slot.

We now consider the consider the case when $\mu_0 = \lambda_0$ and $\mu_1 = \lambda_1$. Recall that this corresponds to the scheduling problem over the generalized Gilbert-Elliot channel.

Lemma 5: For $\lambda_0 = \mu_0 > \mu_1 = \lambda_1$, if $r_2 < \rho(\mu_0)$ and $0 \leq \pi \leq \mu_0$, the optimal policy is to sample.

Proof: Proof is along the same lines as the preceding lemma. From our assumption on μ_0 and μ_1, we see that $r_2 < \rho(\pi)$ for all $\pi \in [0, \mu_0]$. Writing $V_{S,1}(\pi)$, and $V_{NS,1}(\pi)$ as before, we now see that $V_1(\pi) = V_{S,1}(\pi)$. Also, $\mu_1 \leq \gamma_0(\pi), \gamma_1(\pi), \gamma_2(\pi) \leq \mu_0$ for all $\pi \in [0,1]$. Writing the recursion,

$$\begin{aligned} V_{n+1}(\pi) &= \max\{V_{S,n+1}(\pi), V_{NS,n+1}(\pi)\} \\ V_{S,n+1}(\pi) &= \rho(\pi) + \beta\rho(\pi)V_n(\gamma_1(\pi)) + \\ &\quad \beta(1-\rho(\pi))V_n(\gamma_0(\pi)) \\ V_{NS,n+1}(\pi) &= r_2 + \beta V_n(\gamma_2(\pi)), \end{aligned} \tag{8}$$

and, as before, making the induction hypothesis that

$$V_n(\pi) = V_{S,n}(\pi),$$

we can show that both $V_{S,n+1}(\pi)$ and $V_{NS,n+1}(\pi)$ are linear in π. Hence $V_{S,n}$ and $V_{NS,n}(\pi)$ will be

$$\begin{aligned} V_{S,n+1}(\pi) &= m\pi + c, \\ V_{NS,n+1}(\pi) &= \beta m(\lambda_0 - \lambda_1)\pi + r_2 + \beta c + \beta\lambda_1 m, \end{aligned}$$

where,

$$\begin{aligned} m &= \frac{r_0 - r_1}{1-\beta(\mu_0-\mu_1)}, \\ c &= \frac{r_1 + \frac{\beta\mu_1(r_0-r_1)}{1-\beta(\mu_0-\mu_1)}}{1-\beta}. \end{aligned}$$

We want to show that $V_{S,n+1}(\pi) > V_{NS,n+1}(\pi)$, i.e.,

$$m\pi + c > \beta m(\lambda_0 - \lambda_1)\pi + r_2 + \beta c + \beta\lambda_1 m$$

After simplifying this requirement reduces to $\rho(\pi) > r_2$ for $\pi \in [0, \mu_0]$ which is true. We thus have $V_{S,n+1}(\pi) > V_{NS,n+1}(\pi)$ for all $\pi \in [0, \mu_0]$. Hence $V_{n+1}(\pi) = V_{S,n+1}(\pi)$. Thus by induction $V_n(\pi) = V_{S,n}(\pi)$ for all $n \geq 1$. taking limits as $n \to \infty$, we know that $V_n(\pi) \to V(\pi)$, $V_{S,n}(\pi) \to V_S(\pi)$ and hence

$$V(\pi) = V_S(\pi) = m\pi + c.$$

■

Remark 3: Like in Lemma 4, observe that under the conditions of Lemma 5, for $0 \leq \pi \leq 1$, $\gamma_0(\pi)$, $\gamma_1(\pi)$, and $\gamma_2(\pi)$ are all in $[0, \mu_0]$. Hence, irrespective of π_1, the initial belief, the optimal policy will sample always (except possibly in the first slot).

Lemma 6: For $\lambda_0 = \mu_0 > \mu_1 = \lambda_1$, and $r_L < r_2 < r_H$, the inequality $V(0, r_2) > V(1, r_2)$ is satisfied.

Proof: The lemma is proved if we show the following.

$$V(0,r_2) = \begin{cases} V_S(0,r_2) & \text{for } r_2 \leq r_L \\ V_S(0,r_2) & \text{for } r_L \leq r_2 \leq r_H \\ V_{NS}(0,r_2) & \text{for } r_2 \geq r_H \end{cases}$$

$$V(1,r_2) = \begin{cases} V_S(1,r_2) & \text{for } r_2 \leq r_L \\ V_{NS}(1,r_2) & \text{for } r_L \leq r_2 \leq r_H \\ V_{NS}(1,r_2) & \text{for } r_2 \geq r_H \end{cases}$$

For $r_2 \geq r_H$ we know that not sampling is optimal for all π. Thus, in this range of r_2,

$$V(0, r_2) = V(1, r_2) = V_{NS}(0, r_2) = \frac{r_2}{1-\beta}$$

Similarly, for $r_2 \leq r_L$, we know that sampling is optimal for all π. Thus, in this range of r_2, $V(0, r_2) = V_S(0, r_2)$ and $V(1, r_2) = V_S(1, r_2)$. Further, from Property 4,

$$V(0,r_2) = \frac{r_1 + \frac{\beta\mu_1(r_0 - r_1)}{1-\beta(\mu_0-\mu_1)}}{1-\beta}.$$

$$V(1,r_2) = \frac{r_0 - r_1}{1-\beta(\mu_0-\mu_1)} + \frac{r_1 + \frac{\beta\mu_1(r_0-r_1)}{1-\beta(\mu_0-\mu_1)}}{1-\beta}.$$

Also observe that in this range of r_2, $V(0, r_2) > V(1, r_2)$.

Next, for $r_L < r_2 < r_H$, we have

$$\begin{aligned} V_S(0,r_2) &= r_1 + \beta V(\mu_1). \\ V_S(1,r_2) &= r_0 + \beta V(\mu_0). \\ V_{NS}(0,r_2) &= r_2 + \beta V(\lambda_1). \\ V_{NS}(1,r_2) &= r_2 + \beta V(\lambda_0). \end{aligned}$$

By our assumption that $\mu_0 = \lambda_0$ and $\mu_1 = \lambda_1$, and from Property 5, we obtain $r_L = r_0$. Also, $V_S(0, r_2) > V_{NS}(0, r_2)$ and $V_S(1, r_2) < V_{NS}(1, r_2)$. Thus $V(0, r_2) = V_S(0, r_2)$ and $V(1, r_2) = V_{NS}(1, r_2)$. This completes the proof. ■

Note that that the above does not prove that there is one threshold; there could be non contiguous ranges of π for which sampling will be the optimal policy. We now argue that there is possibly just a single threshold when $\lambda_0 = \mu_0 \geq \mu_1 = \lambda_1$.

From Lemma 6, we know that there is at least one threshold. Define

$$\pi_T = \arg \min_{0 \leq \pi \leq 1} V_S(\pi) = V_{NS}(\pi).$$

Observe that $\gamma_0(0) = \gamma_1(0) = \gamma_2(0) = \mu_1$ and $\gamma_0(1) = \gamma_1(1) = \gamma_2(1) = \mu_0$. Also, from Property 1, $\gamma_0(\pi)$ is concave, $\gamma_1(\pi)$ is convex and $\gamma_2(\pi)$ is linear and all are increasing in π. Thus $\gamma_0(\pi) > \gamma_2(\pi) > \gamma_1(\pi)$. This means that from Lemma 3, $V(\gamma_0(\pi)) \leq V(\gamma_2(\pi)) \leq V(\gamma_1(\pi))$.

Next, writing $V_{NS}(\pi) - V_S(\pi)$ as

$$\begin{aligned} V_{NS}(\pi) - V_S(\pi) = {} & [r_2 - \rho(\pi)] - \\ & \beta\left[\rho(\pi)\left[V(\gamma_1(\pi)) - V(\gamma_2(\pi))\right]\right. \\ & \left. +(1-\rho(\pi))\left[V(\gamma_0(\pi)) - V(\gamma_2(\pi))\right]\right], \end{aligned} \quad (9)$$

we observe that the first term increases with π. Further, from the ordering on the γs, $[V(\gamma_1(\pi)) - V(\gamma_2(\pi))] \geq 0$ and $[V(\gamma_0(\pi)) - V(\gamma_2(\pi))] \leq 0$ but the weight on the latter is increasing with π. Further, since $V(\pi)$ is bounded, the term multiplying β in (9) is bounded and β can be made small enough to make the right hand side of (9) to be an increasing function of π for all $\pi > \pi_T$. Thus there is a $\beta_1 \in (0, 1)$ such that for all $\beta < \beta_1$ there is a single threshold.

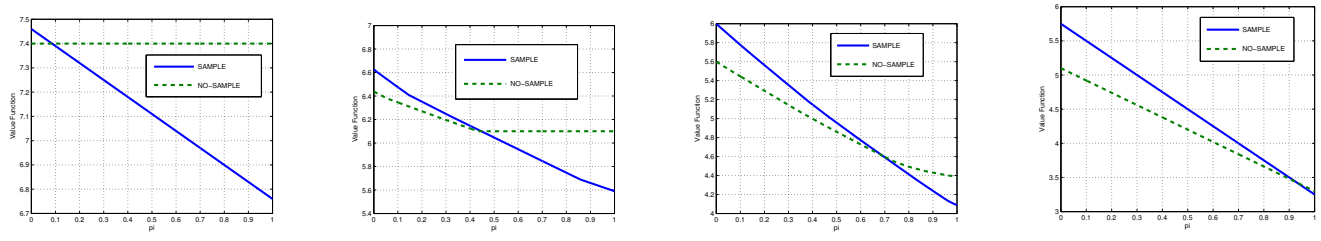
V. Numerical Results and Discussion

From Remark 1, when $r_L \leq r_2 \leq r_H$, $V_S(\pi)$ and $V_{NS}(\pi)$ intersect at least once. Thus the optimal policy would be of the threshold type possibly many thresholds. We believe that in most cases there is just one threshold, i.e., there is a π_T such that for $0 \leq \pi \leq \pi_T$, the optimal policy would be to sample and for $\pi_T \leq \pi \leq 1$, the optimal policy would be to not sample for most other reasonable values of λ_i and μ_i. This is also borne out by our extensive numerical study, a sample of which is discussed below.

In Fig. 1 we have assumed that $\mu_0 = \lambda_0$ and $\mu_1 = \lambda_1$ and we plot $V_{NS}(\pi)$ and $V_S(\pi)$ for four different values of r_2. This also corresponds to the Gilbert-Elliot channel. Observe the lone threshold and also the decreasing of the threshold with increasing r_2. Also observe that when r_2 is large, $V_S(\pi)$ is linear and $V_{NS}(\pi)$ is a constant as expected and when r_2 is small both $V_S(\pi)$ and $V_{NS}(\pi)$ are linear. In Fig. 2 we plot the same for the case when $\mu_0 \neq \lambda_0$ and $\mu_1 \neq \lambda_1$, Once again observe the single threshold. Here we also see that the value functions are piecewise linear. We also observe that $V_S(\pi)$ is decreasing in π.

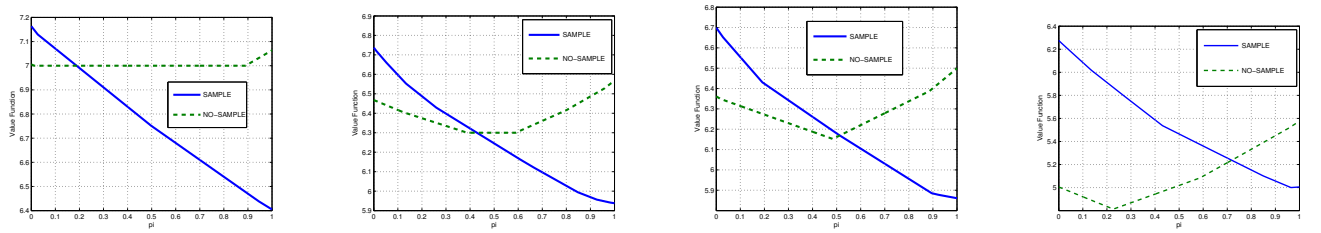
We also see that in all the cases presented, the threshold decreases as r_2 is increased. For the restless multiarmed bandit, in the language of [19], r_2 can be interpreted as the subsidy given to not sampling the arm. Also from [19], we know that if π_T is decreasing with increasing r_2, then the multiarmed case is indexable, i.e., the Whittle's index can be used to select the optimum arm in each slot.

The formal proofs for the optimality of the threshold policy in the one armed case and the indexability in the multiarmed case remain the subject of future work.



$r_2 = 0.74$ $r_2 = 0.65$ $r_2 = 0.5$ $r_2 = 0.15$

Fig. 1. $V_{NS}(\pi)$ and $V_S(\pi)$ plotted for the case when $\mu_0 = \lambda_0 = 0.9$, $\mu_1 = \lambda_1 = 0.1$, $r_0 = 0.1$, $r_1 = 0.8$, and $\beta = 0.9$.



$r_2 = 0.70$ $r_2 = 0.63$ $r_2 = 0.60$ $r_2 = 0.10$

Fig. 2. $V_{NS}(\pi)$ and $V_S(\pi)$ plotted for the case when $\mu_0 = 0.9$, $\mu_1 = 0.1$, $\lambda_0 = 0.1$, $\lambda_1 = 0.9$, $r_0 = 0.1$, $r_1 = 0.8$, and $\beta = 0.9$.

REFERENCES

[1] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Systems Technical Journal*, vol. 39, 1960.

[2] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Systems Technical Journal*, vol. 42, pp. 1977–1997, 1963.

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

[4] E. L. Sernik and S. I. Marcus, "Optimal cost and policy for a Markovian replacement problem," *Journal of Optimization Theory and Applications*, vol. 71, no. 1, pp. 403–406, Oct. 1991.

[5] E. L. Sernik and S. I. Marcus, "On the computation of optimal cost function for discrete time Markov models with partial observations," *Annals of Operations Research*, vol. 29, pp. 471–512, 1991.

[6] J. S. Hughes, "A note on quality control under Markovian deterioration," *Operations Research*, vol. 28, no. 2, pp. 421–424, March-April. 1980.

[7] A. Laourine and L. Tong, "Betting on Gilbert-Elliot channels," *IEEE Transactions on Wireless Communication*, vol. 9, no. 2, pp. 723–732, Feb. 2010.

[8] 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.

[9] 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.

[10] 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, Nov. 2010.

[11] W. Ouyang, S. Murugesan, A. Eyrilmaz, and N. Shroff, "Exploiting channel memory for joint estimation and scheduling in downlink networks," in *Proceedings of IEEE INFOCOM*, 2011.

[12] C. Li and M. J. Neely, "Network utility maximization over partially observable markovian channels," *Performance Evaluation*, vol. 70, no. 7–8, pp. 528–548, July 2013.

[13] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Advances in Neural Information Processing Systems*. NIPS, Dec. 2007, pp. 1–8.

[14] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat, "Leveraging side observations in stochastic bandits," *Arxiv*, 2012.

[15] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Conference on World Wide Web*. ACM, April 2010, pp. 661–670.

[16] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1, Athena Scientific, Belmont, Massachusetts, 1st edition, 1995.

[17] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 2, Athena Scientific, Belmont, Massachusetts, 1st edition, 1995.

[18] 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.

[19] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, no. A, pp. 287–298, 1988.