This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2018.2799521, IEEE Transactions on Automatic Control

On the Whittle Index for Restless Multi-armed Hidden Markov Bandits

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*—We consider a restless multi-armed bandit in which each arm can be in one of two states. When an arm is sampled, the state of the arm is not available to the sampler. Instead, a binary signal with a known randomness that depends on the state of the arm is available. No signal is available if the arm is not sampled. An arm-dependent reward is accrued from each sampling. In each time step, each arm changes state according to known transition probabilities which in turn depend on whether the arm is sampled or not sampled. Since the state of the arm is never visible and has to be inferred from the current belief and a possible binary signal, we call this the hidden Markov bandit. Our interest is in a policy to select the arm(s) in each time step to maximize the infinite horizon discounted reward. Specifically, we seek the use of Whittle's index in selecting the arms.**

We first analyze the single-armed bandit and show that in general, it admits an approximate threshold-type optimal policy when there is a positive reward for the 'no-sample' action. We also identify several special cases for which the threshold policy is indeed the optimal policy. Next, we show that such a single-armed bandit also satisfies an approximate-indexability property. For the case when the single-armed bandit admits a threshold-type optimal policy, we perform the calculation of the Whittle index for each arm. Numerical examples illustrate the analytical results.

I. Introduction

Restless multi-armed bandit problems are a generalization of the classical multi-armed bandit (MAB) problem. In the MAB, the sampler chooses one of N arms in each time-step and receives a reward. Each arm can be in one of M states and the reward is dependent on the state of the arm. The sampled arm changes state according to a known law while the other arms are frozen. In the RMAB, all the arms change their state at each time-step, i.e., the arms are restless. The law that governs the change of state could depend on whether the arm was sampled or not sampled. In this paper we introduce a class of RMAB problems where the player never gets to observe the state of the arm. The objective in both MAB and RMAB is to choose the sequence of arms to sample so as to maximize a long term reward function. We begin with two motivating examples for the models that we introduce in this paper.

A. Motivation

Opportunistic access in time-slotted multi-channel communication systems for Gilbert-Elliot channels [1] is being extensively studied. In the typical model there are N channels and each channel can be in one of two states—a good state and a bad state. Each channel independently evolves between these two states according to a two-state Markov chain. The sender can transmit on one of these N channels in each time slot. If the selected channel is in the good state, then the transmission is successful, and if it is in the bad state, it is unsuccessful. The sender receives instantaneous error-free feedback about the result of the transmission in both these cases. If the sender knows the transition probabilities of the channels, then using the feedback, it can calculate a 'belief' for the state of each channel in a slot. This belief may be used to select the channel in each slot to optimize a suitable reward function. This system and its myriad variations have been studied as restless multi-armed bandit (RMAB) problems.

Consider a system as above except that now the probability of success in the good state and of failure in the bad state are both less than one and the sender knows these probabilities. This generalization of the Gilbert-Elliot channel means that the sender does not get perfect information about the state of the channel from the feedback. However, it can update its a posteriori belief about the state of the channel based on the feedback, and use this updated belief in the subsequent slot.

As a second motivating example, consider an advertisement (ad) placement system (APS) for a user in a web browsing session. Assume that the APS has to place one ad from M candidate ads each of which has a known click-through probability and an expected reward determined from the user profile. It is conceivable that the click-through probabilities for ads in a session depend on the history of the ads shown; users often react differently depending upon the frequency with which an ad is shown. Some users may, due to annoyance, respond negatively to repeated displays of an ad, which has the effect of lowering the click-through probability if they were shown this ad in the past. Others may convert disinterest to curiosity if an ad is repeated, thereby increasing the click-through probability. Yet other users may be more random or oblivious to what has been shown, and may behave independently of the history.

The effect of recommendation history on a user's interest can be modeled as follows. A state is associated with each candidate ad and the state changes at the end of each session (the state intuitively signifies the interest level of the user in the ad). The transition probabilities for this change of state depend on whether the ad is shown or not shown to the user in the session. Assume that the state change behavior is independent of the past and of the state change of the other ads. Each

Rahul Meshram and D. Manjunath are also affiliated with the Bharti Centre for Communication at IIT Bombay. Their work was supported in part by grants from DST and CEFIPRA. The work of Aditya Gopalan was supported in part by DST INSPIRE faculty grant IFA13-ENG-69.

state is associated with a value of click-through probability and expected revenue. The state transition and the click-through probabilities determine the 'type' or profile of the user. In each session the APS only observes a 'signal' or outcome (click or no-click) for the ad that it displayed and no signal for those that are not displayed. The action and the outcome is used to update its belief about the current state of the user for each ad. The objective of the APS would be to choose the ad in each session that optimizes a long term objective. Clearly, this is also a RMAB with the added generalization that the transition probabilities for the arms depend on the action in that stage.

In this paper we analyze this generalization of the restless multi-armed bandit—the states are never explicitly observed and the transition probabilities depend in general on the action chosen. To the best of our knowledge, such systems have not been considered in the literature.

B. Literature Overview

Restless multi-armed bandits (RMAB) are a special class of partially observed Markov decisions processes (POMDPs) and are in general PSPACE-hard [2], but many special cases have been studied. An important recent application of RMABs is in dynamic spectrum access systems, e.g., [3], [4], [5], [6]. A common channel shared by many heterogeneous users, each of whom see the channel as an independent Gilbert-Elliott channel is considered in [3] where an index-based policy to maximize the discounted infinite-horizon throughput minus the transmission costs is derived. In [4], the occupancy of channels by primary users is modeled as a two-state Markov chain. The secondary users (SUs) sense the channel using error-prone spectrum sensors before transmitting. Again, an index policy to maximize the infinite-horizon discounted throughput is derived. In [5], the objective is similar to that of [3] and it is shown that a Whittle's index based policy is optimal. In [6] multiple service classes are considered and the objective is to maximize a utility function based on the queue occupancies. Conditions for a myopic policy, based on instantaneous reward, to be optimal are derived. Myopic policies are also the subject of interest in several other recent works, including [7], [8], [9]. Utility functions are used in [10] that considers a system similar to that of [5]. Opportunistic spectrum access as POMDPs are also studied in [11], [12], [13].

In much of the restless multi-armed bandit literature, including the references in the preceding, the solution method is to seek an 'index-based' policy where the state of each arm is mapped to an index and at each step the arms with the highest index values are sampled. Whittle's index, first proposed in [14], is based on a Lagrangian relaxation and decomposition and is a popular one; see e.g., [15], [16], [5], [17], [18], [19]. An alternative indexing scheme is based on partial and generalized conservation laws [20] and on marginal productivity [4]; in this paper, we will concentrate on the Whittle index. The first step in determining if an index-based policy can be used is to prove indexability. Whittle indexability is shown by analyzing the one armed bandit as a POMDP, the analyses of which borrows significantly from early work on POMDPs that model machine repair problems like in [21], [22], [23]. These are described next.

In [21], 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 [23], a similar model is considered and the formulas for the optimal costs and the policy are obtained. This and some additional models are considered in [22] and, once again, several structural results are obtained. Also see [24] for more such models.

The key features in the single-arm problems considered in the preceding are as follows. One or more of the actions provides the sampler with *exact* information about the state of the Markov chain. Furthermore, the transition probability of the state of the arms does *not* depend on the action. These are also the features of each of the arms of the RMAB models discussed earlier. In this paper we consider a model that drops both these restrictions. Since the state is never observed but only estimated from the signals when the arm is sampled, our model can be called a 'hidden Markov restless multi-armed bandit.' A *rested* hidden Markov bandit has been studied in [25], where the state of an arm does not change if it is not sampled. The (arguably simpler) information structure in a hidden rested bandit admits an analytical solution via Gittins indices.

A further simplification that is often made in showing indexability is to *assume,* without a formal proof, the existence of a threshold-type optimal policy for the single-arm case, i.e., it is optimal to play the arm if the state is higher than the threshold and optimal to not play if the state is below the threshold as in, e.g., [3]. Under this simplification, in many cases, the state of the arm can be mapped to an index without actually calculating the threshold. In Section V we describe a method to do this.

C. Summary of the Contributions

We now summarize the key contributions of this paper. We consider restless multi-armed bandits in which the transition probabilities of the arms depends on whether the arm is sampled or not sampled. Although the applications for this model appear to be many, to the best of our knowledge, this is not a well-studied problem. In addition, the states of the arms are never observed and only a belief about the state of the arm can be computed using prior belief and the conditional probabilities of the observation from a play of the arm. Once again, we believe such a system has not been studied. The preceding features make the system hard to analyze using well known techniques. Hence we develop the notion of an *approximately threshold* type optimal policy and prove that in general the single armed bandit that we consider admits such an optimal policy. For some special cases of the system parameters we also show that the single armed bandit in fact admits a threshold-type optimal policy. We then define *approximate-indexability* and show that the arms defined by our model also satisfy this property. This justifies the use of Whittle's index based policy for the restless multi-armed hidden Markov bandits. For the case when a threshold type policy is indeed the optimal policy, we outline the procedure to compute the Whittle's index. Numerical examples illustrate the theory.

The model details are described in the next section.

II. Model Description and Preliminaries

We consider the following restless, multi-armed bandit problem with N arms. Time is slotted and indexed by t. Each arm has two states, 0 and 1. Let $X_n(t) \in \{0,1\}$ be the state of arm n at the beginning of time t. Let $A_n(t) \in \{0,1\}$ denote the action in slot t for arm n, i.e.,

$$A_n(t) = \begin{cases} 1 & \text{arm } n \text{ is sampled in slot } t, \\ 0 & \text{arm } n \text{ is not sampled in slot } t. \end{cases}$$

We will assume that $\sum_{n=1}^{N} A_n(t) = 1$ for all t, exactly one arm is sampled in each slot. Arm n changes state at the end of each slot according to transition probabilities that depend on $A_n(t)$. Define the following transition probabilities:

$$\begin{aligned} \mathsf{Pr}\left(X_n(t+1)=0 | X_n(t)=0, A_n(t)=0\right) &= \lambda_{n,0}, \\ \mathsf{Pr}\left(X_n(t+1)=0 | X_n(t)=1, A_n(t)=0\right) &= \lambda_{n,1}, \\ \mathsf{Pr}\left(X_n(t+1)=0 | X_n(t)=0, A_n(t)=1\right) &= \mu_{n,0}, \\ \mathsf{Pr}\left(X_n(t+1)=0 | X_n(t)=1, A_n(t)=1\right) &= \mu_{n,1}. \end{aligned}$$

In slot t, if arm n is in state i and it is sampled, then a binary signal $Z_n(t)$ is observed and a reward $R_{n,i}(t,1)$ is accrued. If the arm is not sampled, then a reward $R_{n,i}(t,0)$ is accrued and no signal is observed. Let

$$\mathsf{Pr}\left(Z_n(t)=1 \mid X_n(t)=i, A_n(t)=1\right) = \rho_{n,i}$$

and denote

$$R_{n,i}(t,1) = \eta_{n,i} \qquad R_{n,i}(t,0) = \eta_{n,2}.$$

Fig. 1 illustrates the model and the parameters.

In most applications, $Z_n(t) = 1$ would correspond to a 'good' or favorable output e.g., a successful transmission or click-through in the motivating examples. Hence, we will make the reasonable assumption that $\rho_{n,0} < \rho_{n,1}$ and $\eta_{n,0} < \eta_{n,1}$ for all n.

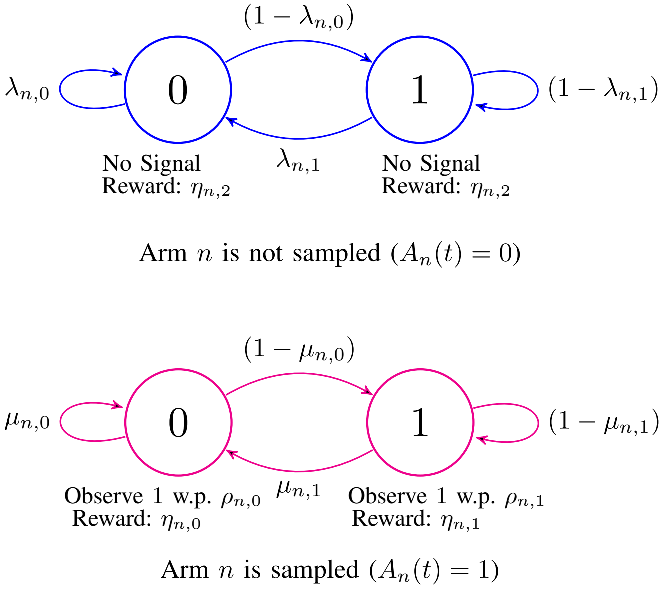
Remark 1:

- In the communication system example that maximizes throughput, no reward is accrued if there is no transmission. Also, in the APS example, no revenue is accrued if there is no ad displayed. Thus in both these cases, $\eta_{n,2} = 0$ is reasonable.
- Further, for communication over Gilbert-Elliot channels, $\lambda_{n,i} = \mu_{n,i}$ for $i = 0, 1$.

We assume that $\lambda_{n,i}$, $\mu_{n,i}$, and $\rho_{n,i}$ are known. The sampler cannot directly observe the state of the arm, and hence does not know the state of the arms at the beginning of each time slot. Instead, it can maintain the posterior or belief distribution $\pi_n(t)$ that arm n is in state 0 given all past actions and observations, i.e., $\pi_n(t) = \mathsf{Pr}\left(X_n(t)=0 \mid (A_n(s), Z_n(s))_{s=1}^{t-1}\right)$, and is assumed known at the beginning of slot t. Thus the expected reward from sampling arm n is

$$\pi_n(t)\eta_{n,0} + (1-\pi_n(t))\eta_{n,1}$$

and that from not sampling the arm is $\eta_{n,2}$.



Arm n is not sampled ($A_n(t) = 0$)

Arm n is sampled ($A_n(t) = 1$)

Fig. 1. Top: State transition probabilities, the expected reward, and the probability of binary signal 1 being observed when the arm is not sampled. Bottom: The corresponding quantities when the arm is sampled

Define the vector $\pi(t) = [\pi_1(t), \ldots, \pi_N(t)] \in [0,1]^N$. Let H_t denote the history of actions and observed signals up to the beginning of time slot t, i.e., $H_t \equiv (A_n(s), Z_n(s))_{1 \le n \le N, 1 \le s < t}$. In each slot, exactly one arm is to be sampled and let $\phi = \{\phi(t)\}_{t>0}$ be the sampling strategy with $\phi(t)$ defined as follows. $\phi(t) : H_t \to \{1, \ldots, N\}$ maps the history up to time slot t to the action of sampling one of the N arms at time slot t. Let

$$A_n^{\phi}(t) = \begin{cases} 1 & \text{if } \phi(t) = n, \\ 0 & \text{if } \phi(t) \neq n. \end{cases}$$

The infinite horizon expected discounted reward under sampling policy ϕ is given by

$$\begin{aligned} V_\phi(\pi) := E\Bigg\{\sum_{t=1}^{\infty} \beta^{t-1} \Bigg(\sum_{n=1}^{N} A_n^{\phi}(t)\ (\pi_n(t)\ \eta_{n,0} \\ +(1-\pi_n(t))\ \eta_{n,1}) + \left(1 - A_n^{\phi}(t)\right)\ \eta_{n,2}\Bigg)\Bigg\}. \end{aligned} \tag{1}$$

Here β, $0 < \beta < 1$, is the discount factor and the initial belief is π, i.e.,$\mathsf{Pr}\left(X_n(1) = 0\right) = \pi_n$. Our interest is in a strategy that maximizes $V_\phi(\pi)$ for all $\pi \in [0,1]^N$

We begin by analyzing the single arm bandit in the next section. Before proceeding we state the following background lemma derived from [26] that will be useful. See the appendix in [29] for the proof.

Lemma 1 ([26]): 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.

Notation. For sets A and B, $A \setminus B$ is used to denote all the elements in A which are not in B.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2018.2799521, IEEE Transactions on Automatic Control

III. Approximate Threshold Policy for the Restless Single Armed Bandit with Hidden States

For notational convenience we will drop the subscript n in the notation of the previous section. Further, we will assume that $\eta_0 = \rho_0$ and $\eta_1 = \rho_1$. Thus η_0 and η_1 will be in $(0,1)$ while there will be no restrictions on the range of η_2. Extending the results to the case of arbitrary η_0 and η_1 is straightforward.

Recall that $\pi(t) = \mathsf{Pr}\left(X(t) = 0 \mid H_t\right)$ and we can use Bayes' theorem to obtain $\pi(t+1)$ from $\pi(t)$, $A(t)$ and $Z(t)$ as follows.

1) If $A(t) = 1$, i.e., the arm is sampled, and $Z(t) = 0$ then

$$\begin{aligned}\pi(t+1) &= \gamma_0(\pi(t)) \\ &:= \frac{\pi(t)(1-\rho_0)\mu_0 + (1-\pi(t))(1-\rho_1)\mu_1}{\pi(t)(1-\rho_0) + (1-\pi(t))(1-\rho_1)}.\end{aligned}$$

2) If $A(t) = 1$ and $Z(t) = 1$ then

$$\pi(t+1) = \gamma_1(\pi(t)) := \frac{\pi(t)\rho_0\mu_0 + (1-\pi(t))\rho_1\mu_1}{\pi(t)\rho_0 + (1-\pi(t))\rho_1}.$$

3) Finally, if $A(t) = 0$, i.e., the arm is not sampled at t, then

$$\pi(t+1) = \gamma_2(\pi(t)) := \pi(t)\lambda_0 + (1-\pi(t))\lambda_1.$$

Recall that the policy is denoted by $\phi(t) : H_t \to \{0,1\}$ and it maps the history up to time t to one of two actions with 1 indicating sampling the arm and 0 indicating not sampling the arm. The following is well known [21], [27], [28]: (1) $\pi(t)$ captures the information in H_t, in the sense that it is a sufficient statistic for constructing policies depending on the history, (2) Optimal strategies can be restricted to stationary Markov policies, and (3) The optimum objective or value function, $V(\pi)$, is determined by solving the following dynamic program

$$\begin{aligned}V(\pi) &= \max\{\rho(\pi) + \beta\left(\rho(\pi)V(\gamma_1(\pi)) + (1-\rho(\pi))\times\right. \\ &\qquad \left. V(\gamma_0(\pi))\right),\ \ \eta_2 + \beta V(\gamma_2(\pi))\}, \qquad (2)\end{aligned}$$

where $\rho(\pi) = \pi\rho_0 + (1-\pi)\rho_1$.

Let π be the belief at the beginning of time slot $t = 1$. Let $V_S(\pi)$ be the optimal value of the objective function if $A(1) = 1$, i.e., if the arm is sampled, and $V_{NS}(\pi)$ be the optimal value if $A(1) = 0$, i.e., if the arm is not sampled. We can now write the following:

$$\begin{aligned}V_S(\pi) &= \rho(\pi) + \beta\left(\rho(\pi)V(\gamma_1(\pi))\right. \\ &\qquad \left.+(1-\rho(\pi))V(\gamma_0(\pi))\right), \qquad (3) \\ V_{NS}(\pi) &= \eta_2 + \beta V(\gamma_2(\pi)), \\ V(\pi) &= \max\{V_S(\pi), V_{NS}(\pi)\}. \qquad (4)\end{aligned}$$

Our first objective is to describe the structure of the value function of the single arm system as a function of two variables—π (the belief) and η_2 (the reward for not sampling). We begin by analyzing the structure of $V(\pi,\eta_2)$, $V_S(\pi,\eta_2)$, and $V_{NS}(\pi,\eta_2)$ when one of π or η_2 is fixed. To keep the notation simple, when the dependence on η_2 is not made explicit it is fixed. The following is proved in the appendix in [29].

Lemma 2:

1) (Convexity of value functions over the belief state) For fixed η_2, $V(\pi)$, $V_{NS}(\pi)$ and $V_S(\pi)$ are all convex functions of π.
2) (Convexity and monotonicity of value functions over passive reward) For a fixed π, $V(\pi,\eta_2)$, $V_S(\pi,\eta_2)$, and $V_{NS}(\pi,\eta_2)$ are non-decreasing and convex in η_2.

□

We are now ready to state the first main result of this paper.

Theorem 1 (Approximately threshold-type optimal policies): For a restless single-armed hidden Markov bandit with two states, $0 < \rho_0 < \rho_1 < 1$ and a given η_2, there exists $\beta_1 \in (0,1)$ such that for all $\beta \leq \beta_1$, one of the following statements is true.

1) A *threshold-type* optimal policy exists, i.e., there exists $\pi_T \in [0,1]$ for which it is optimal to sample at $\pi \in [0,\pi_T]$ and to not sample at $\pi \in (\pi_T, 0]$.
2) An *approximately threshold-type* optimal policy exists, i.e., there exist $\epsilon > 0$ and $\pi_T, \pi^\circ \in [0,1]$ with $\rho(\pi^\circ) = \eta_2$ such that an optimal policy samples at $\pi \in [0,\pi_T] \setminus (\pi^\circ - \epsilon, \pi^\circ + \epsilon)$ and does not sample at $\pi \in (\pi_T, 1] \setminus (\pi^\circ - \epsilon, \pi^\circ + \epsilon)$.

Remark 2: The result essentially states that, under a suitable discount factor $0 < \beta < \beta_1$, an optimal policy has a threshold-structure at all belief states $[0,1]$, except possibly within a small neighbourhood of radius ϵ around the belief state π°.

Proof: Define the intervals S_1 and S_2 as follows.

$$\begin{aligned}S_1 &= \{\pi : \pi \in [0,1] :\ \ \eta_2 < \rho(\pi)\} \\ S_2 &= \{\pi : \pi \in [0,1] :\ \ \eta_2 \geq \rho(\pi)\}\end{aligned}$$

In the following we will use the subscript β to make the dependence of V_S, V_{NS}, and V on β explicit. For notational convenience, let us define

$$\begin{aligned}V_{a,\beta}(\pi,\eta_2) &:= \left[\rho(\pi)V_\beta(\gamma_1(\pi),\eta_2) + \right. \\ &\qquad \left.(1-\rho(\pi))V_\beta(\gamma_0(\pi),\eta_2)\right].\end{aligned}$$

From (3), we see that $\beta V_{a,\beta}(\pi,\eta_2)$ is the second term for the expression for $V_{S,\beta}(\pi,\eta_2)$. For a fixed β, $V_\beta(\pi,\eta_2)$ and $V_{a,\beta}(\pi,\eta_2)$ are bounded for all $\pi \in [0,1]$; this follows from ρ_0, ρ_1, and η_2 being bounded and $0 < \beta < 1$. Further, we can show that for fixed π and η_2, $V_\beta(\pi,\eta_2)$ is an increasing function of β; see appendix in [29].

For each belief state $\pi \in [0,1]$ satisfying $\eta_2 \neq \rho(\pi) = \pi\rho_0 + (1-\pi)\rho_1$, let us define[1] $\beta_1(\pi)$ as

$$\begin{aligned}\beta_1(\pi) &:= \sup\left\{\beta \in (0,1)\ :\right. \\ &\qquad \left.\frac{|\eta_2 - \rho(\pi)|}{\beta} > |V_\beta(\gamma_2(\pi)) - V_{a,\beta}(\pi)|\right\}. \qquad (5)\end{aligned}$$

Such a $\beta_1(\pi)$ exists in $(0,1]$ because, as we have argued previously, the difference between V and V_a is bounded, and moreover, $|\eta_2 - \rho(\pi)| > 0$. Now define, for any $\epsilon \geq 0$, the set

$$C_\epsilon := \{\pi \in [0,1] : |\rho(\pi) - \eta_2| \geq \epsilon\},$$

[1] We follow the standard convention that $\sup\{x : x \in \emptyset\} = -\infty$ (resp. $\inf\{x : x \in \emptyset\} = +\infty$), where $\emptyset$ denotes the empty set, and in this case we say that the supremum (resp. infimum) does not exist or is not finite.

and the quantity

$$\beta_{1,\epsilon} := \inf\left\{\beta_1(\pi) : \pi \in C_\epsilon\right\}.$$

It follows that $\beta_{1,\epsilon}$ is finite (i.e., the set C_ϵ is nonempty) whenever either

1) $\eta_2 \notin \{\rho(\pi) : \pi \in [0,1]\}$. In this case we will have a (perfect) threshold-type optimal policy by taking $\epsilon = 0 \Rightarrow C_\epsilon = [0,1]$ as will follow below.
2) $\eta_2 \in \{\rho(\pi) : \pi \in [0,1]\}$ and $\epsilon < \max\{\pi^\circ, 1-\pi^\circ\}$ with $\rho(\pi^\circ) = \eta_2$. Note that in this case, $S_1 = [0,\pi^\circ)$ and $S_2 = [\pi^\circ, 1]$. Here, by taking any $0 < \epsilon < \max\{\pi^\circ, 1-\pi^\circ\}$, we will have an approximate threshold-type optimal policy as will follow below. We remark that in this case, for any ϵ as above, it can be argued that $\beta_{1,\epsilon}$ is positive as follows. Given the expected reward parameters ρ_0, ρ_1 and η_2, let $u := \max\{\rho_0, \rho_1, \eta\}$, so that $|V_\beta(\cdot)| \leq \frac{u}{1-\beta}$ uniformly, implying that $|V_\beta(\gamma_2(\pi)) - V_{a,\beta}(\pi)| \leq \frac{2u}{1-\beta}$ for all π. Now, for any $\pi \in C_\epsilon$, we have
$$\begin{aligned} &\delta := \frac{\epsilon}{2u+\epsilon} \\ &\Rightarrow \frac{2u}{1-\delta} = \frac{\epsilon}{\delta} \leq \frac{|\rho(\pi)-\eta_2|}{\delta} \\ &\Rightarrow |V_\delta(\gamma_2(\pi)) - V_{a,\delta}(\pi)| \leq \frac{|\rho(\pi)-\eta_2|}{\delta} \\ &\Rightarrow \delta \in \beta_1(\pi), \end{aligned}$$
and so the infimum of all such numbers must satisfy $\beta_{1,\epsilon} \geq \delta = \frac{\epsilon}{2u+\epsilon} > 0$.

We now claim that for any ϵ for which $\beta_{1,\epsilon}$ is finite, and for any $\beta < \beta_{1,\epsilon}$, the optimal policy chooses to sample whenever the belief state is in the region $S_1 \cap C_\epsilon$, and to not sample in the region $S_2 \cap C_\epsilon$.

First, for $\pi \in S_1 \cap C_\epsilon$, $V_{S,\beta}(\pi,\eta_2) > V_{NS,\beta}(\pi,\eta_2)$. To see this, write

$$\begin{aligned} V_{S,\beta}(\pi,\eta_2) - V_{NS,\beta}(\pi,\eta_2) \;&=\; (\rho(\pi) - \eta_2) \\ &\quad -\beta\left(V_\beta(\gamma_2(\pi),\eta_2) - V_{a,\beta}(\pi,\eta_2)\right). \end{aligned}$$

For $\pi \in S_1$, the term in the first parentheses in the right hand side (RHS) above is positive. We now consider two cases. If the term in the second parentheses is negative, then the RHS is positive and the claim holds. On the other hand, if the term is positive, then from the definition of $\beta_{1,\epsilon}$, for all $\beta < \beta_1$, the second term is less than the first and for this case too the claim follows.

On the other hand, for $\pi \in S_2 \cap C_\epsilon$, the claim follows by observing that

$$V_{a,\beta}(\pi,\eta_2) - V_\beta(\gamma_2(\pi),\eta_2) < \frac{\eta_2 - \rho(\pi)}{\beta}.$$

whenever $\beta < \beta_1(\pi)$. Hence $V_S(\pi) < V_{NS}(\pi)$ for $\beta < \beta_{1,\epsilon}$. This completes the proof. ■

This theorem states that if $\eta_2 \in [\rho_0, \rho_1]$, then there is at least an approximate threshold policy. Of course if $\eta_2 < \rho_0$, then the policy is to always sample corresponding to a threshold policy with $\pi_T = 1$. Similarly, $\eta_2 > \rho_1$ corresponds to a threshold policy with $\pi_T = 0$.

A. Special case: Existence of a threshold-type optimal policy

In Theorem 1, we have introduced two approximations—a restriction on the range of β, and also a 'hole' in the range of π, the state of the arm, for which we do not know the optimal policy. We now consider a special case where we do not need to use these approximations, i.e., the optimal policy is always of the threshold type. The key idea behind these is to use Lemma 2 and Lemma 3 (below) and argue that the difference between the value functions from sampling and not sampling, $(V_S(\pi) - V_{NS}(\pi))$, which we call the *sampling advantage,* is monotonic in π under these special cases of λs and μs.

Assume $\eta_0 = \rho_0$ and $\eta_1 = \rho_1$. We will need the following lemma that shows that for a suitable range of parameter values, $(V_S(\pi) - V_{NS}(\pi))$ is monotonic.

Lemma 3: (Monotonicity of the sampling advantage) For a fixed η_2 and $\beta \in (0,1]$, $(V_S(\pi) - V_{NS}(\pi))$ is a decreasing function in π for the following cases.

1) $0 \leq \mu_0 - \mu_1 \leq \frac{1}{5}$ and $|\lambda_0 - \lambda_1| \leq \frac{1}{5}$.
2) $0 \leq \mu_1 - \mu_0 \leq \frac{1}{3}$ $|\lambda_0 - \lambda_1| \leq \frac{1}{3}$.

The proof is available in the appendix in [29]. This now enables us to state the following result.

Theorem 2 (Exact threshold-type optimal policies): For a restless single-armed hidden Markov bandit with two states, $0 < \rho_0 = \eta_0 < \rho_1 = \eta_1 < 1$ and given η_2, for all $\beta \in (0,1]$, a *threshold-type* optimal policy exists, i.e., there exists $\pi_T \in [0,1]$ for which it is optimal to sample at $\pi \in [0,\pi_T]$ and to not sample at $\pi \in (\pi_T, 0]$, whenever

1) $0 \leq \mu_0 - \mu_1 \leq \frac{1}{5}$ and $|\lambda_0 - \lambda_1| \leq \frac{1}{5}$, or
2) $0 \leq \mu_1 - \mu_0 \leq \frac{1}{3}$ and $|\lambda_0 - \lambda_1| \leq \frac{1}{3}$.

Proof: For a fixed β and η_2, from Lemma 3, we also know that $(V_S(\pi) - V_{NS}(\pi))$ is decreasing in π. Also $V_S(\pi)$ and $V_{NS}(\pi)$ are convex in π. This implies that there is at most one point in $(0,1)$ at which $V_S(\pi)$ and $V_{NS}(\pi)$ intersect. This completes the proof. ■

Remark 3: Note that we do not make any assumption on the ordering of λ_0 and λ_1 except that the absolute difference is bounded by $\frac{1}{5}$ or by $\frac{1}{3}$ which in turn depends on the ordering of μ_0 and μ_1.

B. Numerical Examples

Theorem 1 introduces two approximations—an upper bound on the discount factor, and a 'hole' in $[0,1]$ where we do not know the optimal policy. We believe that this is just an artifact of the proof technique and that the restriction on β and the hole need not actually exist. To see this we conducted an extensive numerical experiments in which the value functions were evaluated numerically using value iteration. Some of the plots for $V_S(\pi)$ and $V_{NS}(\pi)$ for a sample set of μ_i, λ_i, and ρ_i for different values of the discount factor β and η_2 are available in [29]. All our results indicated that there is just one threshold even when β is very large and even close to 1. This leads us to believe that both the approximations may not be needed, and to state the following conjecture.

Conjecture 1 (Existence of threshold-type optimal policies): For a restless single-armed hidden Markov bandit with two states with $0 < \rho_0 < \rho_1 < 1$, a threshold-type optimal policy exists, i.e., there exists $\pi_T \in [0,1]$ for which it is optimal to sample at $\pi \in [0, \pi_T]$ and to not sample at $\pi \in (\pi_T, 0]$.

IV. Approximate Indexability of the Restless Multi-armed Bandit with Hidden States

We are now ready to analyze the general case of the *multi-armed* bandit setting. As we have discussed in the introduction, finding the optimal policy is, in general, a hard problem. A heuristic that is widely used in optimally selecting the arm at each time step is due to Whittle [14]. This heuristic is in general suboptimal but has a good empirical performance and a large class of practical problems use this policy because of its simplicity. In some cases, it can also be shown to be optimal, e.g., [5]. The arm selection in each time slot proceeds as follows. The belief vector $\pi(t)$ is used to calculate the Whittle's index (defined below) for each arm and the arm with the highest index is sampled. To be able to compute such an index for each arm, we first need to determine if the arm is indexable. Toward determining indexability, let us first define,

$$\mathcal{P}_\beta(\eta_2) \quad := \quad \{\pi \in [0,1] : V_{S,\beta}(\pi,\eta_2) \leq V_{NS,\beta}(\pi,\eta_2)\}.$$

In other words, for a given β, $\mathcal{P}_\beta(\eta_2)$ is the set of all belief states π for which not sampling is the optimal action. From [14], indexability of an arm is defined as follows.

Definition 1 (Indexability): An arm in the single-armed bandit process is indexable if $\mathcal{P}_\beta(\eta_2)$ monotonically increases from $\emptyset$ to the entire state space $[0,1]$ as η_2 increases from $-\infty$ to ∞, i.e., $\mathcal{P}_\beta(\eta_2^{(a)}) \setminus \mathcal{P}_\beta(\eta_2^{(b)}) = \emptyset$ whenever $\eta_2^{(a)} \leq \eta_2^{(b)}$. A restless multi-armed bandit problem is indexable if every arm is indexable.

Definition 2 (Approximate or ϵ-indexability): For $\epsilon \geq 0$, an arm is said to be *ϵ-indexable* for the single-armed bandit process if, for $\eta_2^{(a)} < \eta_2^{(b)}$, we have $\mathcal{P}_\beta(\eta_2^{(a)}) \setminus \mathcal{P}_\beta(\eta_2^{(b)}) \subseteq [\tilde{\pi} - \epsilon, \tilde{\pi} + \epsilon]$ for some $\tilde{\pi} \in [0,1]$.

Next we define the Whittle index for an arm in state π.

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

$$W(\pi) \quad = \quad \inf\{\eta_2 \in \mathbb{R} : V_{S,\beta}(\pi,\eta_2) = V_{NS,\beta}(\pi,\eta_2)\}. \tag{6}$$

In other words, $W(\pi)$ is the minimum value of the *no-sampling subsidy* η_2 such that the optimal action at belief state π is to not sample the arm. Our next objective is to show that the arms in our problem are all indexable. Showing indexability, at a high level, requires us to show that the set $\mathcal{P}_\beta(\eta_2)$ increases monotonically as η_2 increases. We now prove the second key result of the paper, on the approximate-indexability of an arm.

Theorem 3: (ϵ-Indexability of the single-armed bandit) For a restless single-armed hidden Markov bandit with two states, $0 < \rho_0 < \rho_1 < 1$, there exists a β_2, $0 < \beta_2 < 1$, and $\epsilon \geq 0$ such that for all $\beta < \beta_2$, the arm is ϵ-indexable.

Proof: First, we make the intuitive claim that there exist finite η_L, η_H, such that $\mathcal{P}_\beta(\eta_2) = \emptyset$ (resp. $\mathcal{P}_\beta(\eta_2) = [0,1]$) when η_2 is less than (resp. greater than) η_L (resp. η_H). This is because the rewards are finite and the objective function is a discounted reward.

Lemma 4: If for each $\eta_2 \in [\eta_L, \eta_H]$,

$$\left.\frac{\partial V_S(\pi,\eta_2)}{\partial \eta_2}\right|_{\pi=\pi_T(\eta_2)} < \left.\frac{\partial V_{NS}(\pi,\eta_2)}{\partial \eta_2}\right|_{\pi=\pi_T(\eta_2)}, \tag{7}$$

then $\pi_T(\eta_2)$ is a monotonically decreasing function of η_2 in $[\eta_L, \eta_H]$. Here, $\frac{\partial V_S(\pi,\eta_2)}{\partial \eta_2}$ denotes the right partial derivative of $V_S(\pi, \cdot)$.

Henceforth, we assume that $\eta_2 \in [\eta_L, \eta_H]$.

Taking the partial derivative of $V_{S,\beta}(\pi,\eta_2)$, and $V_{NS,\beta}(\pi,\eta_2)$ with respect to η_2 we obtain

$$\begin{aligned} \frac{\partial V_{S,\beta}(\pi,\eta_2)}{\partial \eta_2} &= \beta \left[\rho(\pi) \frac{\partial V_\beta\left(\gamma_1(\pi), \eta_2\right)}{\partial \eta_2} \right. \\ &\qquad \left. + (1-\rho(\pi)) \frac{\partial V_\beta\left(\gamma_0(\pi), \eta_2\right)}{\partial \eta_2} \right], \quad (8) \\ \frac{\partial V_{NS,\beta}(\pi,\eta_2)}{\partial \eta_2} &= 1 + \beta \frac{\partial V_\beta\left(\gamma_2(\pi), \eta_2\right)}{\partial \eta_2}. \quad (9) \end{aligned}$$

Taking (9) - (8), we obtain

$$\begin{aligned} &\frac{\partial V_{NS,\beta}(\pi,\eta_2)}{\partial \eta_2} - \frac{\partial V_{S,\beta}(\pi,\eta_2)}{\partial \eta_2} = \\ &1 + \beta \frac{\partial V_\beta\left(\gamma_2(\pi), \eta_2\right)}{\partial \eta_2} - \beta \left[r(\pi) \frac{\partial V_\beta\left(\gamma_1(\pi), \eta_2\right)}{\partial \eta_2} \right. \\ &\qquad \left. + (1-r(\pi)) \frac{\partial V_\beta\left(\gamma_0(\pi), \eta_2\right)}{\partial \eta_2} \right]. \end{aligned}$$

We now show that the above is greater than 0 at $\pi = \pi_T(\eta_2)$. After rearranging the terms this requirement reduces to requiring that

$$\begin{aligned} \frac{1}{\beta} > \Bigg\{ & \left[\rho(\pi) \frac{\partial V_\beta\left(\gamma_1(\pi), \eta_2\right)}{\partial \eta_2} \right. \\ & \left. + (1-\rho(\pi)) \frac{\partial V_\beta\left(\gamma_0(\pi), \eta_2\right)}{\partial \eta_2} \right]_{\pi=\pi_T(\eta_2)} \\ & - \left[\frac{\partial V_\beta\left(\gamma_2(\pi), \eta_2\right)}{\partial \eta_2} \right]_{\pi=\pi_T(\eta_2)} \Bigg\}. \end{aligned} \tag{10}$$

Since $V_\beta(\pi,\eta_2)$ is a bounded function for fixed β, $0 < \beta < 1$, finite η_2, and $\pi \in [0,1]$, the partial (right) derivative of $V_\beta(\pi,\eta_2)$ with respect to η_2 is also bounded. This means that we can find β_2 such that for all $0 < \beta < \beta_2$, the conclusion (7) of Lemma 4 holds. We will also require β to be in $(0, \beta_1)$ with β_1 from the conclusion of Theorem 1.

Thus, letting $\beta_3 = \min\{\beta_1, \beta_2\}$, we get the first crossing point $\pi_T(\eta_2)$ to be monotone non-decreasing with η_2.

To complete the proof, note that the only other states $\pi > \pi_T(\eta_2)$ at which the optimal action *may* play the no-sampling action must lie within an ϵ-radius hole around π°, as shown in Theorem 1. This establishes the conclusion of the theorem. ■

Under the conditions of Theorem 2, we can do away with the approximations of Theorem 3 and explicitly characterize a bound on the discount β required for indexability. Specifically, we state the following.

Theorem 4: For a restless single-armed hidden Markov bandit with two states, $0 < \rho_0 < \rho_1 < 1$, and finite η_2 if either

1) $0 \leq \mu_0 - \mu_1 \leq \frac{1}{5}$ and $|\lambda_0 - \lambda_1| \leq \frac{1}{5}$, or
2) $0 \leq \mu_1 - \mu_0 \leq \frac{1}{3}$ and $|\lambda_0 - \lambda_1| \leq \frac{1}{3}$

is true then for all $\beta \in (0, 1/3)$, the arm is indexable.

Proof: We know from Theorem 2 that the optimal policies are threshold type with single threshold, i.e., $\pi_T(\eta_2)$ is unique for given η_2. Further, we can obtain the following inequalities using induction techniques as in, for example, Lemma 2

$$\left|\frac{\partial V(\pi, \eta_2)}{\partial \eta_2}\right|, \left|\frac{\partial V_S(\pi, \eta_2)}{\partial \eta_2}\right|, \left|\frac{\partial V_{NS}(\pi, \eta_2)}{\partial \eta_2}\right| \leq \frac{1}{1-\beta} \tag{11}$$

show that (10) is true for range of the parameters that we consider here. This is done by using (11), and upper bounding the RHS of (10) as follows.

$$\begin{aligned} RHS &\leq \rho(\pi)\frac{1}{1-\beta} + (1-\rho(\pi))\frac{1}{1-\beta} + \frac{1}{1-\beta}, \\ &= \frac{2}{1-\beta}. \end{aligned}$$

If $\beta < 1/3$, then $\frac{2}{1-\beta} < \frac{1}{\beta}$ implying (10) to complete the proof. ■

Remark 4: Theorem 3 tells us that the restless multi-armed bandit with hidden states is approximately indexable. Like in Theorem 1, we believe that the approximation is just an artifact of the proof technique and the result is possibly more generally true and also without the restriction on β. This is also borne out by extensive numerical study that we conducted. In Fig. 2 we show a sample plot of $\pi_T(\eta_2)$, the threshold belief as a function of the passive subsidy η_2 for different β. We see that π_T increases with η_2 leading us to believe that indexability is more generally true.

V. Explicit Calculation of the Whittle Index for the Class of Threshold Policies

Recall Conjecture 1 on a threshold policy for the single-arm hidden Markov bandit. For the cases when the conjecture is true, we can use the definition of the Whittle index for an arm and explicitly evaluate it. The calculations though are tedious and require us to exercise care in enumerating the various cases. This is because the monotonicity and convexity properties of the γs depend on the ordering of μs and λs. In the following we will consider, for the sake of an example, one case $\lambda_0 = \mu_0 > \mu_1 = \lambda_1$. The other cases have similar calculations and will be omitted here. We will also continue to assume that $0 < \rho_0 < \rho_1 < 1$.

For $i = 0, 1, 2,$ define $\gamma_i^{(0)}(\pi) = \pi$, $\gamma_i^{(k)} := \gamma_i\left(\gamma_i^{(k-1)}(\pi)\right)$, and $\gamma_{i,\infty} := \lim_{k\to\infty} \gamma_i^{(k)}(\pi)$. We can show that $0 < \mu_1 < \gamma_{1,\infty} < \gamma_{2,\infty} < \gamma_{0,\infty} < \mu_0 < 1$. See Fig. 3. The interval $(0, 1)$, the range of π is divided into five regions, denoted $A_1, \ldots, A_5$, as shown in Fig. 3.

1) For $\pi \in A_1$, $W(\pi) = \rho(\pi)$.
2) For $\pi \in A_2$, we will have the following cases
   a) If $\gamma_1(\pi) \geq \pi$, then $W(\pi) = \rho(\pi)$.
   b) If $\pi > \gamma_1(\pi)$, $\pi \leq \gamma_0(\pi)$, $\gamma_0(\gamma_1(\pi)) > \pi$, and $\gamma_1^{(2)}(\pi) \geq \pi$ then
   $$W(\pi) = \frac{(1-\beta)\left[\rho(\pi) + \beta\rho(\pi)\rho(\gamma_1(\pi))\right]}{(1-\beta\left[1-\rho(\pi)(1-\beta)\right])}.$$
   c) If $\pi > \gamma_1(\pi)$, $\pi \leq \gamma_0(\pi)$, $\gamma_0(\gamma_1(\pi)) > \pi$, and $\gamma_1^{(2)}(\pi) < \pi$, then
   $$W(\pi) = \frac{(1-\beta)C_1}{\left[1-(C_2+C_3+C_4)\right]},$$
   where
   $$\begin{aligned} C_1 &= \sum_{l=0}^{\tau_1-1} \beta^l \prod_{j=0}^{l} \rho\left(\gamma_1^{(j)}(\pi)\right) \\ C_2 &= \beta^{\tau_1} \prod_{j=0}^{\tau_1-1} \rho\left(\gamma_1^{(j)}(\pi)\right) \\ C_3 &= \sum_{l=1}^{\tau_1-1} \beta^{l+1} \prod_{j=1}^{l} \rho\left(\gamma_1^{(j-1)}(\pi)\right)\left(1-\rho\left(\gamma_1^{(l)}(\pi)\right)\right) \\ C_4 &= \beta(1-\rho(\pi)) \\ \tau_1 &:= \inf\left\{k \geq 1 : \gamma_1^{(k)}(\pi) \geq \pi\right\}. \end{aligned}$$
   d) If $\pi > \gamma_1(\pi)$, $\gamma_0(\pi) \geq \pi$, $\gamma_0(\gamma_1(\pi)) < \pi$ and $\gamma_1^{(2)}(\pi) < \pi$ then $W(\pi)$ is obtained numerically by performing the value iteration till convergence.
3) For $\pi \in A_3$ then the Whittle index is obtained via numerical computation as described above.
4) For $\pi \in A_4$, $W(\pi) = \rho(\pi) + \beta\gamma_2(\pi)(m-1)$.
5) For $\pi \in A_5$, then
   $$W(\pi) = m\pi\left(1-\beta(\lambda_0-\lambda_1)\right) + (1-\beta)c - \beta\lambda_1 m.$$
   where $m = \frac{\rho_0-\rho_1}{1-\beta(\mu_0-\mu_1)}$, $c = \frac{\rho_1 + \frac{\beta\mu_1(\rho_0-\rho_1)}{1-\beta(\mu_0-\mu_1)}}{1-\beta}$.

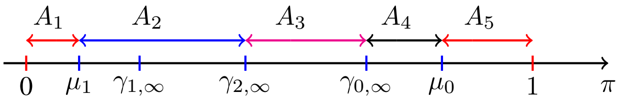


Fig. 3. The different cases to calculate $W(\pi)$ in Section V

We now provide a brief description of the key steps in obtaining the preceding expressions. The key idea is of course to solve $V_S(\pi, \eta_2) = V_{NS}(\pi, \eta_2)$ for η_2. This solution is $W(\pi)$. In general, $V_S(\pi, \eta_2)$ and $V_{NS}(\pi, \eta_2)$ do not have closed form expressions. The key step is to show that for fixed η_2, both $V_S(\pi, \eta_2)$ and $V_{NS}(\pi, \eta_2)$ have at most three connected components for fixed η_2. This fact, and the properties of the γs are then used to solve for η_2. For example, for $\pi \in A_1$, we have $0 \leq \pi \leq \mu_1$, $\gamma_0(\pi), \gamma_1(\pi) \geq \pi$ and $V_S(\pi, \eta_2) = \rho(\pi) + \beta\frac{\eta_2}{1-\beta}$ and $V_{NS}(\pi, \eta_2) = \frac{\eta_2}{1-\beta}$. Equating $V_S(\pi, \eta_2)$ and $V_{NS}(\pi, \eta_2)$ at $\pi = \pi_T$ and solving for η_2, we get $\eta_2 = \rho(\pi) = W(\pi)$. The other closed form expressions are similarly obtained. For the two cases for which we need to obtain $W(\pi)$ numerically, such a simplification is not possible.

VI. Concluding Remarks

Several interesting prospects for future work are open. We would of course like to know for sure if the single armed bandit indeed has a single threshold sampling policy. The complexity of the γ_is makes such a proof hard and the 'usual' techniques that have been used in the literature do not appear

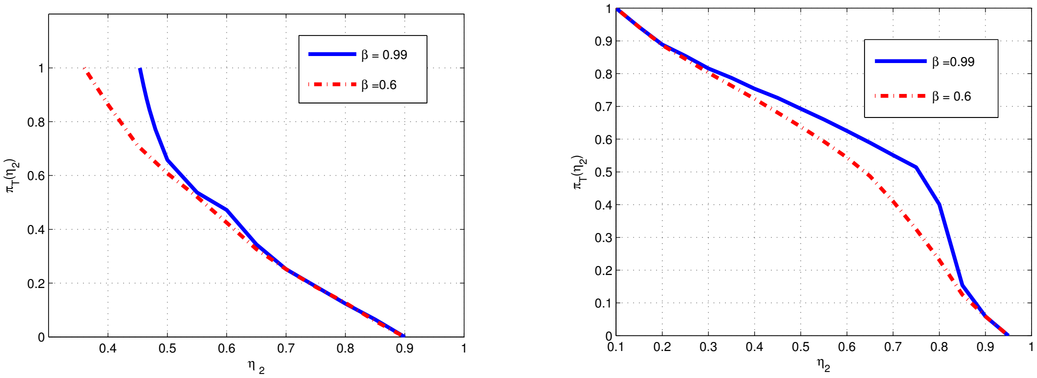


Fig. 2. $\pi_T(\eta_2)$ is plotted for $\beta = 0.6$ and $\beta = 0.99$. The left plot uses the parameter values $\rho_0 = \eta_0 = 0.1$, $\rho_1 = \eta_1 = 0.9$, $\mu_0 = 0.1$, $\mu_1 = 0.9$, $\lambda_0 = 0.9$, and $\lambda_1 = 0.1$. The right plot uses $\rho_0 = \eta_0 = 0.1$, $\rho_1 = \eta_1 = 0.95$, $\mu_0 = \lambda_0 = 0.9$, and $\mu_1 = \lambda_1 = 0.1$.

to be useful. The restriction on β in the main results are in the same spirit as that of [30]. The approximation is introduced here.

Since we do not have a closed-form expression for $V(\pi)$ and $W(\pi)$, provably good approximations may be sought. Also, since the Whittle index based policy is itself suboptimal, we could seek other suboptimal policies that can provide guarantees on the approximation to optimality.

References

[1] E. N. Gilbert, "Capacity of a Burst-Noise Channel," *Bell System Technical Journal*, vol. 39, no. 5, pp. 1253–1265, 1960.

[2] 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.

[3] J. Niño-Mora, "An index policy for dynamic fading-channel allocation to heterogeneous mobile users with partial observations," in *Proceedings of the Conference on Next Generation Internet Networks*, April 2008, pp. 231–238.

[4] J. Niño-Mora, "A restless bandit marginal productivity index for opportunistic spectrum access with sensing errors," in *Proceedings of Conference on Network Control Optimization (NET-COOP), LNCS 5894*, 2009, pp. 60–74.

[5] 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.

[6] C. Lott and D. Teneketzis, "On the optimality of the index rule in multi-channel allocation for single-hop mobile networks with multiple service classes," *Probability in the Engineering and Information Sciences*, vol. 14, pp. 259–297, 2010.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2053–2071, May 2008.

[13] 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.

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

[15] M. H. Veatch and L. M. Wein, "Scheduling a make-to-stock queue: Index policies and hedging points," *Operations Research*, vol. 44, no. 4, pp. 634–647, July-August 1996.

[16] 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, pp. 4220–4225.

[17] 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.

[18] K. Avrachenkov, U. Ayesta, J. Doncel, and P. Jacko, "Congestion control of TCP flows in Internet routers by means of index policy," *Computer Networks*, vol. 57, no. 17, pp. 3463–3478, 2013.

[19] K. Avrachenkov and V. S. Borkar, "Whittle index policy for crawling ephemeral content," *IEEE Transactions on Control of Network Systems*, 2016, DOI:10.1109/TCNS.2016.2619066.

[20] J. E. Niño-Mora, "Restless bandits, partial conservation laws and indexability," *Advances in Applied Probability*, vol. 33, pp. 76–98, 2001.

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

[22] 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.

[23] 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, October 1991.

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

[25] V. Krishnamurthy and R. J. Evans, "Hidden Markov model for multiarm bandits: a methodology for beam scheduling in multitarget tracking," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 2893–2908, December 2001.

[26] 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.

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

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

[29] Rahul Meshram, D. Manjunath, and Aditya Gopalan, "On the Whittle index for restless multi-armed hidden Markov bandits," arXiv:1603.04739, 2017.

[30] C. C. White III, "Optimal control-limit strategies for a partially observed replacement problem," *International Journal of System Science*, vol. 10, no. 3, pp. 321–331, 1979.