Optimality of Myopic Policy for a Class of Monotone Affine Restless Multi-Armed Bandits

Parisa Mansourifard[†] ,Tara Javidi[‡] ,Bhaskar Krishnamachari[†]
[†] Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA
[‡] Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA
emails: parisama@usc.edu, tjavidi@ucsd.edu, bkrishna@usc.edu

***Abstract*— We formulate a general class of restless multi-armed bandits with n independent and stochastically identical arms. Each arm is in a real-valued state $s \in [s_0, s_{max}]$. Selecting an arm with state s yields an immediate reward with expectation $R(s)$. The state of the arm that is selected stochastically jumps from its current value s to either s_{max} or s_0 with probability $p(s)$ or $1-p(s)$ respectively. The state of the arms that are not selected evolve according to a function $\tau(s)$. We prove that if $\tau(s)$, $p(s)$, and $R(s)$ are all monotonically increasing affine functions, and $\tau(s)$ is a contraction mapping, the simple myopic policy, which selects at each time the arm with the highest immediate reward, is optimal. This generalizes recent results in the literature pertaining to arms evolving as two-state Markov chains.**

I. Introduction

Multi-armed bandit (MAB) are a class of stochastic decision problems concerned with selecting from several alternatives at each time, in order to maximize the expected discounted or average reward obtained over some horizon, possibly infinite. They arise in a wide range of settings involving online learning and sequential control [1]. As the number of options that may be selected at each time are limited, and the reward process is typically stochastic, there is a fundamental trade-off in these problems between exploration and exploitation.

In the classical Bayesian *rested* multi-armed bandit problem, a single arm is selected to play at each time, and a state-dependent reward is obtained. The state of the played arm changes according to a known Markovian rule, while the remaining arms remain frozen. In [2], Gittins showed that the optimal policy has an index structure for classical MAB. Specifically an index can be assigned to the state of each arm, and the optimal policy is playing an arm with the largest index at each time. This index is referred to as the *Gittins index* [3].

Peter Whittle introduced the Bayesian restless multi-armed bandit (RMAB) problem [4], a generalization in which the state of all arms evolve in Markovian fashion at each time (even those that are not selected). Whittle showed that for RMAB, an index policy is not in general optimal. He proposed a Gittins-like index and showed it is optimal under a constraint on the average number of arms that can be played at each time. It has been shown that this class of problems is in fact PSPACE-hard [5]. Therefore the literature on these problems has emphasized the design of approximation algorithms ([6], [7], [8]) or the identification of special classes of RMAB for which particular heuristics are optimal ([9], [10], [11], [12]). Our contribution in this paper is of the latter type: we describe a general class of RMAB for which a simple index policy is optimal.

The simplest index policy is the *myopic* policy which ignores the impact of the current action on the future reward, and focuses on maximizing the current reward. For a specific opportunistic spectrum sensing RMAB problem arising in the domain of cognitive radio networks, recently several researchers ([10], [11], [12]) have shown that the myopic policy is optimal under certain conditions.

In this work, we formulate and consider a general class of restless multi-armed bandits with n independent and stochastically identical arms. Each arm is in a real-valued state $s \in [s_0, s_{max}]$. Selecting an arm with state s yields an immediate reward with expectation $R(s)$. The state of the arm that is selected stochastically jumps from its current value s to either s_{max} or s_0 with probability $p(s)$ or $1-p(s)$ respectively. The state of the arms that are not selected evolve according to a function $\tau(s)$. We show that if $\tau(s)$, $p(s)$, and $R(s)$ are all monotonically increasing affine functions, and $\tau(s)$ is a contraction mapping, the myopic policy is always optimal.

Our result is a significant generalization of the work in [10], [11] and [12], as the conditions under which a myopic policy is found to be optimal in those papers correspond to a specific setting of our formulation. Specifically, [10], [11] and [12] address a problem in which the arms can be in "good" or "bad" state and s is the conditional probability of jumping to the good state, $s_0 = p_{01}$, $s_{max} = p_{11}$. Then $p(s)$ and $R(s)$ are both simply equal to s, and $\tau(s)$ has a particular affine linear form obtained from a Bayesian belief update that satisfies the contraction criterion whenever the Markov chain for each arm is positively correlated.

In [13], the authors consider a class of reset processes that is related to our formulation. They show that the Whittle index can be computed for this class in closed form and is equivalent to the myopic policy, and present new results on its optimality in the asymptotic regime when the ratio of the number of arms selected at each time to the total number of arms tends to zero. They also extend prior results on optimality of the myopic policy in the finite regime for 2-state Markov chains, allowing for time-inhomogeneous chains and time-varying arm constraints. In contrast, our work focuses on the optimality of the myopic policy in the finite regime for a larger class of problems than two-state Markov chains.

The remainder of this paper is organized as follows. In Section II, we formulate our problem. In Section III, the optimal policy and the myopic policy are described. In Section III, we prove that in our case the myopic policy is optimal. Finally we conclude the paper in Section V.

II. Problem Formulation

We consider a restless multi-armed bandit problem in which only one arm can be played at each time. Assume there are n independent and statistically identical arms. Each arm is in a state changing over time, either played or not. After playing an arm, a reward can be achieved as a function of the state of played arm. The problem is to find an optimal policy of sequentially playing arms which maximizes the expected discounted reward achieved over finite horizon.

The finite horizon is denoted by T and time steps are indexed by $t = 1, 2, ..., T$. The state of arm j at time t, is given by $s_j(t) \in \mathbb{R}, j = 1, ..., n$, and the vector $\bar{s}(t) = [s_1(t), s_2(t), ..., s_n(t)]$ denotes the state of the system at time t. We have the following assumptions:

- The state of the arms can be any real number between s_0 and s_{max}, the lowest and the highest achievable state, respectively, *i.e.* $s_0 \leq s \leq s_{max}$.
- The expected reward collected by playing the arm a at time t, is a function of the state of that arm, denoted by $R(s_a(t))$.
- After playing the arm a, its state jumps to s_{max} with probability $p(s_a(t))$; otherwise it jumps to s_0. Then $p(s)$ is the probability of jumping from state s to s_{max} which is a function of the state of played arm.
- The state of not-played arms will be changed as a deterministic function of their states, *i.e.* $s_{j+1}(t) = \tau(s_j(t)), j \neq a$.

Briefly the state transition of the arms upon playing arm a is governed by the following:

$$\begin{aligned} & s_j(t+1) \\ & \quad = \begin{cases} s_{max} & \text{w.p. } p(s_j(t)) \text{, if } j = a \\ s_0 & \text{w.p. } 1 - p(s_j(t)) \text{, if } j = a \\ \tau(s_j(t)) & \text{w.p. } 1 \text{, if } j \neq a \end{cases} \\ & \forall j = 1, ..., n, \end{aligned} \tag{1}$$

For compactness we use an operator Γ for the state evolution of all arms described by (1) in a vector format. Γ is applied on the state vector $\bar{s}(t)$, upon playing arm a, as follows:

$$\begin{aligned} & \Gamma(\bar{s}(t), a) \\ & = \begin{cases} [\tau(\bar{s}_1^{a-1}(t)), s_{max}, \tau(\bar{s}_{a+1}^n(t))] & \text{w.p. } p(s_a(t)) \\ [\tau(\bar{s}_1^{a-1}(t)), s_0, \tau(\bar{s}_{a+1}^n(t))] & \text{w.p. } 1 - p(s_a(t)), \end{cases} \end{aligned} \tag{2}$$

where $\bar{s}_j^k(t)$ is the vector $[s_j(t), ..., s_k(t)]$, and $\tau(\bar{s}_j^k(t)) = [\tau(s_j(t)), ..., \tau(s_k(t))]$, for $1 \leq j \leq k \leq n$.

We assume the player uses a Markovian deterministic policy π which maps the current state vector, $\bar{s}(t)$, to the action of selecting a particular arm at time t. We denote this policy by the vector $\pi = [\pi(1), \pi(2), ..., \pi(T)]$ where $\pi(t) = a \in \{1, 2, ..., n\}$ means that the arm a is selected to play at time t. This is not restrictive because the current state vector is a sufficient statistic for the entire of observation history due to the Markovian dynamics of the underlying system.

The problem is maximizing the total discounted expected reward achieved in a finite horizon, over all admissible policies π. This maximization problem is written as follows:

$$\max_{\pi} J_T^{\pi}(\bar{s}) = \max_{\pi} E^{\pi}[\sum_{t=1}^{T} \beta^{t-1} R(s_{\pi(t)}(t)) | \bar{s}(1) = \bar{s}], \tag{3}$$

where $0 \leq \beta \leq 1$ is the discount factor and $\bar{s}$ is the initial state of the system. Note $\bar{s}$ is equal to $\bar{s}_1^n$ where we drop both the subscript and the superscript for notational simplicity. $R(s_{\pi(t)}(t))$ is the immediate expected reward collected by playing arm $\pi(t)$.

III. Optimal Policy and the Myopic Policy

The optimal policy is the policy π^* which maximizes $J_T^{\pi}(\bar{s})$ in (3). Note an optimal policy exists since the number of admissible policies are finite. This problem may be solved using dynamic programming (DP) and the following recursive equations:

$$
\begin{aligned}
&V_T(\bar{s}) = \max_{a=1,2,...,n} R(s_a), && (4)\\
&V_t(\bar{s}) = \max_{a=1,2,...,n} \{R(s_a) + \beta E[V_{t+1}(\Gamma(\bar{s}, a))]\}\\
&\;= \max_{a=1,2,...,n} \{R(s_a)\\
&\;+ \beta p(s_a) V_{t+1}(\tau(\bar{s}_1^{a-1}), s_{max}, \tau(\bar{s}_{a+1}^{n}))\\
&\;+ \beta(1 - p(s_a)) V_{t+1}(\tau(\bar{s}_1^{a-1}), s_0, \tau(\bar{s}_{a+1}^{n}))\},\\
&\quad t = 1, 2, ..., T-1, && (5)
\end{aligned}
$$

where $V_t(\bar{s})$, is the value function, or the maximum expected remaining reward accrued starting from time t when the current state is $\bar{s}$. $V_{a,t}(\bar{s}), a = 1, ..., n$, is the expected remaining reward accrued by playing arm a. it has two parts, the immediate expected reward obtained in time step t, and the maximum expected remaining reward starting from time $t + 1$ with the states updated according to the action a. Note, for all $t = 1, ..., T$, $V_t(\bar{s}) = \max_\pi J^\pi_{T-t+1}(\bar{s})$ with probability 1. In particular, $V_1(\bar{s}) = \max_\pi J^\pi_T(\bar{s})$.

A policy π^* is optimal if and only if for $t = 1, ..., T$, $a = \pi^*(t)$ achieves the maximum in (4). and (5) Because computing the optimal policy for a RMAB can be computationally intractable, there is a motivation to study the performance of simpler, possibly sub-optimal policies. One of the simplest policies is the myopic policy, which ignores the impact of the current action on the future reward. It focuses solely on maximizing the expected immediate reward. For problem (3), the myopic policy under state $\bar{s} = [s_1, s_2, ..., s_n]$ is given by

$$\pi^*(\bar{s}) = \arg \max_{a=1,2,...,n} R(s_a). \tag{6}$$

IV. OPTIMALITY OF MYOPIC POLICY

In this section, we will show that the myopic policy is optimal under the following conditions:

C1: $p(s)$, $R(s)$, and $\tau(s)$ are monotonically increasing functions of the state s. A function $X(s)$ is monotonically increasing if

$$X(s_1) \geq X(s_2) \;\; \forall s_1 \geq s_2, \tag{7}$$

C2: $p(s)$, $R(s)$, and $\tau(s)$ are affine functions of the state s, *i.e.*, they are in the form of $p_0 + \frac{p_{max} - p_0}{s_{max} - s_0} s$, $a_r + b_r s$ and $a_\tau + b_\tau s$, respectively.

C3: $\tau(s)$ is a contraction mapping, *i.e.*,

$$|\tau(s_1) - \tau(s_2)| \leq |s_1 - s_2| \;\; \forall s_1, s_2. \tag{8}$$

So $b_\tau \leq 1$. Intuitively, this property implies that the state of an arm, that is not played for a long time, converges to a steady state $s^* \in [s_0, s_{max}]$, (as per Banach fixed point theorem, [14]).

Using *C1*, the myopic policy of (6) is simplified to

$$\pi^*(\bar{s}) = \arg \max_{a=1,2,...,n} s_a. \tag{9}$$

The implementation of the myopic policy is as follows. We take the initial state $\bar{s}(1)$ and select an arm with the highest state. In subsequent steps, we will play the same arm if its state stays in s_{max}. Otherwise, it is moved to the lowest priority of playing and we will play other arm with the highest current state. Since *C1* is satisfied for $\tau(s)$, the ordering of the states of not-played arms is preserved.

Assume $W_t(s_1, ..., s_n), t = 1, ..., T$ indicate n-variable functions with a recursive representation as follows:

$$
\begin{aligned}
W_t(s_1, ..., s_n) &= R(s_1) + \beta p(s_1) W_{t+1}(s_{max}, \tau(\bar{s}_2^n))\\
&\quad + \beta(1 - p(s_1)) W_{t+1}(\tau(\bar{s}_1^{n-1}), s_0), \qquad (10)
\end{aligned}
$$

which is equal to the total expected reward earning by playing the arm with the lowest index at each time. If the state of played arm jumps to s_0, we will put it in nth index, unless keep it in the first index and repeat playing it for the next step. The state of other arms will be changed as (2).

Our main result is summarized in the following theorem:

Theorem: Under conditions *C1-3*, and $b_\tau \leq \frac{1}{\beta(1+\beta(p_{max}-p_0))}$ the myopic policy is optimal, *i.e.* if we sort the states such that $s_1 \geq s_2 \geq ... \geq s_n$, then we will have:

$$
\begin{aligned}
W_t(s_1, \bar{s}_2^{i-1}, s_i, \bar{s}_{i+1}^n) &\geq W_t(s_i, \bar{s}_2^{i-1}, s_1, \bar{s}_{i+1}^n),\\
&\forall t, 0 \leq t \leq T, \;\forall \bar{s} = [s_1, ..., s_n] \qquad (11)
\end{aligned}
$$

Proof: To prove the theorem, we will use backward induction on t. The optimality of the myopic policy at time T is straightforward from (4). Assuming satisfying (11) at times $t, t + 1, ..., T$, we prove some equalities and one inequality, given by lemmas 1-4. In lemma 1 and 2, we show the symmetry and affine linearity of $W_t(\bar{s})$. In lemma 3, we drive a simple expression for the difference between the W_t functions achieved by switching the playing order of different arms. In lemma 4, we prove that if the state of one arm at time t is changed, the difference between new W_t function and the previous one is less than an upper-bound. Using the lemmas, we will show that (11) holds at $t-1$, as well.

Lemma 1: $W_t(\bar{s})$ is an affine function of the states and the following equality holds:

$$\begin{aligned}\lambda W_t(\bar{s}_1^{j-1}, s, \bar{s}_{j+1}^n) &+ (1-\lambda)W_t(\bar{s}_1^{j-1}, s', \bar{s}_{j+1}^n)\\ &= W_t(\bar{s}_1^{j-1}, \lambda s + (1-\lambda)s', \bar{s}_{j+1}^n)\\ &\quad \forall j = 1, 2, ..., n, \forall \lambda. \end{aligned} \tag{12}$$

Proof: Affine linearity of $W_t(\bar{s})$ is obvious from (10), and *C2*. ■

Lemma 2: $W_t(\bar{s})$ has the symmetry property, *i.e.*,

$$\begin{aligned}W_t(s, \bar{s}_2^{j-1}, s', \bar{s}_{j+1}^n) &= W_t(s', \bar{s}_2^{j-1}, s, \bar{s}_{j+1}^n)\\ &\quad \forall\ 1 \le i \le j \le n. \end{aligned} \tag{13}$$

Proof: Since the arms are stochastically identical, exchanging the index and the state of the playing arm with another arm will not change W_t. ■

Lemma 3: For any $i = 1, ..., n$, we have:

$$\begin{aligned}W_t(s_1, \bar{s}_2^{i-1}, s_i, \bar{s}_{i+1}^n) &- W_t(s_i, \bar{s}_2^{i-1}, s_1, \bar{s}_{i+1}^n)\\ &= (\lambda_1 - \lambda_i) \times [W_t(U, \bar{s}_2^{i-1}, L, \bar{s}_{i+1}^n)\\ &\quad - W_t(L, \bar{s}_2^{i-1}, U, \bar{s}_{i+1}^n)], \end{aligned} \tag{14}$$

where,

$$U = \tau^{-1}(s_{max}), \quad L = \tau^{-1}(s_0), \tag{15a}$$

$$s_i = \lambda_i U + (1-\lambda_i)L, \tag{15b}$$

$$\frac{s_0 - L}{U - L} \le \lambda_i \le \frac{s_{max} - L}{U - L}. \tag{15c}$$

Note τ^{-1} is the inverse of the function τ.

Proof: From (12) and using (15b), we have:

$$\begin{aligned}W_t(s_1, &\bar{s}_2^{i-1}, s_i, \bar{s}_{i+1}^n)\\ &= \lambda_1 W_t(U, \bar{s}_2^{i-1}, s_i, \bar{s}_{i+1}^n)\\ &+ (1-\lambda_1)W_t(L, \bar{s}_2^{i-1}, s_i, \bar{s}_{i+1}^n)\\ &= \lambda_1[\lambda_i W_t(U, \bar{s}_2^{i-1}, U, \bar{s}_{i+1}^n)\\ &+ (1-\lambda_i)W_t(U, \bar{s}_2^{i-1}, L, \bar{s}_{i+1}^n)]\\ &+ (1-\lambda_1)[\lambda_i W_t(L, \bar{s}_2^{i-1}, U, \bar{s}_{i+1}^n)\\ &+ (1-\lambda_i)W_t(L, \bar{s}_2^{i-1}, L, \bar{s}_{i+1}^n)]. \end{aligned} \tag{16}$$

After computing $W_t(s_i, \bar{s}_2^{i-1}, s_1, \bar{s}_{i+1}^n)$ in the same way and subtracting it from $W_t(s_1, \bar{s}_2^{i-1}, s_i, \bar{s}_{i+1}^n)$, due to (13) the similar terms are cancelled, and the result follows. ■

lemma 4: If we change the state of ith arm at time t, the following upper bound hold for the difference between W_t functions:

$$\begin{aligned}W_t(\bar{s}_1^{i-1}, &s_i, \bar{s}_{i+1}^n) - W_t(\bar{s}_1^{i-1}, s'_i, \bar{s}_{i+1}^n)\\ &\le \beta^{i-1}\frac{R(\tau^{i-1}(s_i)) - R(\tau^{i-1}(s'_i))}{1 - \beta(p_{max} - p_0)}\\ &= \frac{(\beta b_\tau)^{i-1} b_r (s_i - s'_i)}{1 - \beta(p_{max} - p_0)}\\ &\forall t = 0, 1, ..., T, \end{aligned} \tag{17}$$

if $s_i \ge s'_i$.

Note p_{max} and p_0 represent the simplified notations for $p(s_{max})$ and $p(s_0)$, respectively.

Proof: The proof is inductive. For time T it is straightforward from (4) and the fact that $p_{max} - p_0 \le 1$. Assuming the inequality (17) holds for $t+1$, for time t we have:

$$\begin{aligned}W_t(\bar{s}_1^{i-1}, &s_i, \bar{s}_{i+1}^n) - W_t(\bar{s}_1^{i-1}, s'_i, \bar{s}_{i+1}^n)\\ &= \beta p(s_1)[W_{t+1}(s_{max}, \tau(\bar{s}_2^{i-1}), \tau(s_i), \tau(\bar{s}_{i+1}^n))\\ &- W_{t+1}(s_{max}, \tau(\bar{s}_2^{i-1}), \tau(s'_i), \tau(\bar{s}_{i+1}^n))]\\ &+ \beta(1 - p(s_1))[W_{t+1}(\tau(\bar{s}_2^{i-1}), \tau(s_i), \tau(\bar{s}_{i+1}^n), s_0)\\ &- W_{t+1}(\tau(\bar{s}_2^{i-1}), \tau(s'_i), \tau(\bar{s}_{i+1}^n), s_0)] \end{aligned} \tag{18a}$$

$$\begin{aligned}&\le \beta p(s_1)\frac{(\beta b_\tau)^{i-1} b_r(\tau(s_i) - \tau(s'_i))}{1 - \beta(p_{max} - p_0)}\\ &+ \beta(1 - p(s_1))\frac{(\beta b_\tau)^{i-2} b_r(\tau(s_i) - \tau(s'_i))}{1 - \beta(p_{max} - p_0)} \end{aligned} \tag{18b}$$

We get (18a) from (10). Then, from *C1* for $p(s)$ and (17) at time $t+1$, we get (18b). Then we have:

$$\begin{aligned}W_t(\bar{s}_1^{i-1}, &s_i, \bar{s}_{i+1}^n) - W_t(\bar{s}_1^{i-1}, s'_i, \bar{s}_{i+1}^n)\\ &\le p(s_1)\frac{(\beta b_\tau)^{i-1} b_r b_\tau (s_i - s'_i)}{1 - \beta(p_{max} - p_0)}\\ &+ (1 - p(s_1))\frac{(\beta b_\tau)^{i-1} b_r (s_i - s'_i)}{1 - \beta(p_{max} - p_0)} \end{aligned} \tag{19a}$$

$$\le \frac{(\beta b_\tau)^{i-1} b_r (s_i - s'_i)}{1 - \beta(p_{max} - p_0)} \tag{19b}$$

The equality (19b) is achieved from *C2*, *i.e.* $b_\tau \le 1$. ■

Two functions in the right side of (14) at time $t-1$ can be computed from (10), as following:

$$\begin{aligned}W_{t-1}(U, &\bar{s}_2^{i-1}, L, \bar{s}_{i+1}^n)\\ &= R(U) + \beta p(U) W_t(s_{max}, \tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n))\\ &+ \beta(1 - p(U))W_t(\tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n), s_0), \end{aligned} \tag{20a}$$

$$\begin{aligned}W_{t-1}(L, &\bar{s}_2^{i-1}, U, \bar{s}_{i+1}^n)\\ &= R(L) + \beta p(L) W_t(s_{max}, \tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n))\\ &+ \beta(1 - p(L))W_t(\tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n), s_0), \end{aligned} \tag{20b}$$

where we substitute $\tau(L)$ and $\tau(U)$ with s_0 and s_{max}, respectively, as defined in (15a). Then by substituting (20a) and (20b) in (14), we obtain:

$$\begin{aligned} & W_{t-1}(s_1, \bar{s}_2^{i-1}, s_i, \bar{s}_{i+1}^n) - W_{t-1}(s_i, \bar{s}_2^{i-1}, s_1, \bar{s}_{i+1}^n) \\ &\quad = (\lambda_1 - \lambda_i)[R(U) - R(L) \\ &\quad + \beta p(U) W_t(s_{max}, \tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n)) \\ &\quad + \beta(1 - p(U)) W_t(\tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n), s_0) \\ &\quad - \beta p(L) W_t(s_{max}, \tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n))] \\ &\quad - \beta(1 - p(L)) W_t(\tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n), s_0) \qquad (21a) \\ &\quad = (\lambda_1 - \lambda_i)[R(U) - R(L) \\ &\quad - \beta(1 - p(U))[W_t(\tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n), s_0) \\ &\quad - W_t(\tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n), s_0)] \\ &\quad - \beta p(L)[W_t(s_{max}, \tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n)) \\ &\quad - W_t(s_{max}, \tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n))]] \\ &\quad + \beta(p(U) - p(L))[W_t(s_{max}, \tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n)) \\ &\quad - W_t(\tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n), s_0)]. \qquad (21b) \end{aligned}$$

(21b) follows from straightforward manipulations. Applying (17), we obtain:

$$\begin{aligned} & W_{t-1}(s_1, \bar{s}_2^{i-1}, s_i, \bar{s}_{i+1}^n) - W_{t-1}(s_i, \bar{s}_2^{i-1}, s_1, \bar{s}_{i+1}^n) \\ &\quad \geq (\lambda_1 - \lambda_i)[R(U) - R(L) \\ &\quad - \beta(1 - p(U)) \frac{b_r(s_{max} - s_0)}{1 - \beta(p_{max} - p_0)} \\ &\quad - \beta p(L) \frac{b_r(s_{max} - s_0)}{1 - \beta(p_{max} - p_0)} + X] \qquad (22a) \\ &\quad = (\lambda_1 - \lambda_i)\times \\ &\quad [\frac{R(U) - R(L) - \beta b_r(s_{max} - s_0)}{1 - \beta(p_{max} - p_0)} + X] \qquad (22b) \\ &\quad = (\lambda_1 - \lambda_i)[\frac{b_r(U - L)[1 - \beta b_\tau]}{1 - \beta(p_{max} - p_0)} + X] \qquad (22c) \end{aligned}$$

where we use *C2* to conclude that $(R(U) - R(L))(p_{max} - p_0) - b_r(s_{max} - s_0)(p(U) - p(L)) = 0$ and (15a). X is as follows:

$$\begin{aligned} X = \beta(p(U) - p(L))[& W_t(s_{max}, \tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n)) \\ & - W_t(\tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n), s_0)], \end{aligned} \qquad (23)$$

We can have an upper bound for that:

$$\begin{aligned} X & \geq \beta(p(U) - p(L))[W_t(\tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n), s_{max}) \\ & - W_t(\tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n), s_0)] \qquad (24a) \\ & \geq \beta(p(U) - p(L))[W_t(\tau(\bar{s}_2^{i-1}), s_0, \tau(\bar{s}_{i+1}^n), s_0) \\ & - W_t(\tau(\bar{s}_2^{i-1}), s_{max}, \tau(\bar{s}_{i+1}^n), s_0)] \qquad (24b) \\ & \geq -\beta(p(U) - p(L)) \frac{b_r(\beta b_\tau)^{i-1}(s_{max} - s_0)}{1 - \beta(p_{max} - p_0)} \qquad (24c) \end{aligned}$$

where (24a) comes from using the result of theorem at time t and switching the position of s_{max} with all indexes $2, 3, ..., n$. For (24b) we use the fact that W_t is monotonically increasing in all states. Finally by Lemma 4 we get (24c).

So by using (24c) in (22c), we have:

$$\begin{aligned} & W_{t-1}(s_1, \bar{s}_2^{i-1}, s_i, \bar{s}_{i+1}^n) - W_{t-1}(s_i, \bar{s}_2^{i-1}, s_1, \bar{s}_{i+1}^n) \\ &\quad \geq \frac{\lambda_1 - \lambda_i}{1 - \beta(p_{max} - p_0)}[b_r(U - L)(1 - \beta b_\tau) \\ &\quad - \beta^i(p(U) - p(L)) b_r(\beta b_\tau)^{i-1}(s_{max} - s_0)] \\ &\quad = \frac{\lambda_1 - \lambda_i}{1 - \beta(p_{max} - p_0)} b_r(U - L) \\ &\quad [1 - \beta b_\tau - \beta^i b_\tau^{i-1}(p_{max} - p_0)] \\ &\quad \geq \frac{\lambda_1 - \lambda_i}{1 - \beta(p_{max} - p_0)} b_r(U - L)\times \\ &\quad [1 - \beta b_\tau(1 + \beta(p_{max} - p_0))] \qquad (25) \\ &\quad \geq 0 \end{aligned}$$

From $s_1 \geq s_i$ and *C1*, we have $\lambda_1 - \lambda_i \geq 0$ and $U - L \geq 0$. Using the fact that $\beta \leq 1$, and $\beta b_\tau \leq 1$ from *C3*, we got (25) we reach (25). Then with applying the assumption of theorem, *i.e.*, $b_\tau \leq \frac{1}{\beta(1+\beta(p_{max}-p_0))}$ the proof is complete. ■

V. CONCLUSION

Restless multi-armed bandit problems have long been known to be challenging to solve. Recent results in the literature [10], [11], [12] have identified special cases for which the simple myopic policy is optimal. Our results in this work have generalized these prior results beyond the specific setting of two-state Markov chains. We have shown that the myopic policy is optimal for reset processes with monotone affine state evolution and reward functions, where the evolution of the non-selected arms corresponds to a contraction mapping.

There are several avenues for future work. Using techniques from [15], we can easily extend the results in this paper to the selection of multiple arms at each time. In [16], the authors present three sufficient conditions under which the myopic policy is optimal for the RMAB problems involving 2-state Markov chains, including some cases involving non-identical arms. It would be of interest to consider whether such an approach could be applied to extend our results to show the optimality of myopic policy for non-identically evolving arms (under some additional conditions). We are interested in investigating the optimality of the myopic policy under further generalizations such as non-affine evolution and multi-dimensional states. We are also interested in identifying conditions for related problems where the myopic is not necessarily optimal but some other efficient, possibly index-based, policy is optimal.

REFERENCES

[1] J. Gittins, R. Weber, and K. Glazebrook, *Multi-armed bandit allocation indices*. Wiley Online Library, 1989.

[2] J. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177, 1979.

[3] P. Whittle, "Multi-armed bandits and the gittins index," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 143–149, 1980.

[4] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, pp. 287–298, 1988.

[5] C. Papadimitriou and J. Tsitsiklis, "The complexity of optimal queueing network control," in *Structure in Complexity Theory Conference, Proceedings of the Ninth Annual*, pp. 318–322, 1994.

[6] J. Nino-Mora, "Restless bandits, partial conservation laws and indexability," *Advances in Applied Probability*, vol. 33, no. 1, pp. 76–98, 2001.

[7] S. Guha and K. Munagala, "Approximation algorithms for partial-information based stochastic control with markovian rewards," in *Foundations of Computer Science, 48th Annual IEEE Symposium on*, pp. 483–493, 2007.

[8] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," *Journal of the ACM (JACM)*, vol. 58, no. 1, p. 3, 2010.

[9] C. Lott and D. Teneketzis, "On the optimality of an index rule in multichannel allocation for single-hop mobile networks with multiple service classes," *Probability in the Engineering and Informational Sciences*, vol. 14, no. 3, pp. 259–297, 2000.

[10] Q. Zhao and B. Krishnamachari, "Structure and optimality of myopic sensing for opportunistic spectrum access," in *Communications, IEEE International Conference on*, pp. 6476–6481, 2007.

[11] T. Javidi, B. Krishnamachari, Q. Zhao, and M. Liu, "Optimality of myopic sensing in multi-channel opportunistic access," in *Communications, IEEE International Conference on*, pp. 2107–2112, 2008.

[12] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *Information Theory, IEEE Transactions on*, vol. 55, no. 9, pp. 4040–4050, 2009.

[13] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5547–5567, 2010.

[14] V. Istratescu, *Fixed point theory an introduction*, vol. 7. Kluwer Academic Print on Demand, 2001.

[15] S. Ahmad and M. Liu, "Multi-channel opportunistic access: A case of restless bandits with multiple plays," in *Communication, Control, and Computing, 47th Annual Allerton Conference on*, pp. 1361–1368, 2009.

[16] K. Wang and L. Chen, "On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach," *Signal Processing, IEEE Transactions on*, no. 99, pp. 1–1, 2011.