# On a Restless Multi-Armed Bandit Problem with Non-Identical Arms

Naumaan Nayyar, Yi Gai, Bhaskar Krishnamachari
Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA
Email: {nnayyar, ygai, bkrishna}@usc.edu

***Abstract*—We consider the following learning problem motivated by opportunistic spectrum access in cognitive radio networks. There are $N$ independent Gilbert-Elliott channels with possibly non-identical transition matrices. It is desired to have an online policy to maximize the long-term expected discounted reward from accessing one channel at each time dynamically. While there is a stream of recent results on this problem when the channels are identical, much less is known for the harder case of non-identical channels. We provide the first characterization of the structure of the optimal policy for this problem when the channels can be non-identical, in the Bayesian case (when the transition matrices are known). We also provide the first provably efficient learning algorithm for a non-Bayesian version of this problem (when the transition matrices are unknown). Specifically, for the special case of two positively correlated channels, we use the structure we identify to develop a novel mapping to a different multi-armed bandit with countably-infinite arms, in which each arm corresponds to a threshold-based policy. Using this mapping, we propose a policy that achieves near-logarithmic regret for this problem with respect to an $\epsilon$-optimal solution.**

## I. Introduction

The problem of dynamic channel selection has recently been formulated and studied by many researchers [1], [2], [3], [4], [5], [6] under the framework of multi-armed bandits (MAB) [7]. In these papers, the channels are typically modelled as indpendent Gilbert-Elliott channels (i.e., described by two-state Markov chains, with a bad state "0" and a good state "1"). The objective is to develop a policy for the user to select a channel at each time based on prior observations so as to maximize some suitably defined notion of long-term reward. This can be viewed as a special case of restless multi-armed bandits (RMAB), a class of problems first introduced by Whittle [8]. Depending on whether the underlying state transition Matrix is known or unknown, the problem can be respectively classified as Bayesian (because in this case the belief that the channel will be good in the next time step can be updated exactly for all channels based on the prior observations) or non-Bayesian.

Many of the prior results apply to the Bayesian case, where it is assumed that the underlying Markov state transition matrices for all channels are known, so that the corresponding beliefs for each channel can be updated. For the special case when the channels evolve as *identical* chains, it has been shown that the Myopic policy is always optimal for $N = 2$, 3, and also optimal for any $N$ so long as the chains are positively correlated [1], [2]. In [3], the authors consider this problem in the general case when the channels can be *non-identical*. They show that a well-known heuristic, the Whittle's index exists for this problem, and can be computed in closed form. Moreover, it is shown in [3] that for the special case that channels are identical, the Whittle's index policy in fact coincides with the Myopic policy. However, as the Whittle's index is not in general the best possible policy, *a question that has remained open is identifying the optimal solution for general non-identical channels in the Bayesian case*.

In [1], it is also shown that when the channels are identical, the Myopic policy has a semi-universal structure, such that it requires only the determination of whether the transition matrix is positively or negatively correlated, not the actual parameter values. In [4], it has been shown that this structure can be exploited to obtain an efficient online learning algorithm for the non-Bayesian version of the problem (where the underlying transition matrix is completely unknown). In particular, Dai *et al.* [4] show that near logarithmic regret (defined as the difference between cumulative rewards obtained by a model-aware optimal-policy-implementing genie and that obtained by their policy) with respect to time[1] can be achieved by mapping two particular policies to arms in a different multi-armed bandit. For the more general case of non-identical arms, there have been some recent results that show near-logarithmic *weak* regret (measured with respect to the best possible single-channel-selection policy, which need not be optimal) [5], [6], [9]. Thus, *there are currently no strong results showing a provably efficient online learning algorithm for the case of non-identical channels in the non-Bayesian case*.

In this paper, we consider both the Bayesian and non-Bayesian versions of this two-state restless multi-armed bandit problem with non-identical channels. We make two main contributions:

- For the Bayesian version of the problem, when the underlying Markov transition matrices are known, we prove structural properties for the optimal policy. Specifically, we show that the decision region for a given

[1]Note that it is desirable to have sub-linear regret with respect to time for non-Bayesian multi-armed bandits, as this indicates that asymptotically the time-averaged reward approaches that obtained by the model-aware genie.

channel is contiguous with respect to the belief of that channel, keeping all other beliefs fixed.

- For the non-Bayesian version of the problem for the special case of $N = 2$ positively correlated possibly non-identical channels, we utilize the above-derived structure to propose a mapping to another multi-armed bandit problem, with a countably infinite number of arms, each corresponding to a possible threshold policy (one of which must be optimal). We present an online learning algorithm for this problem, and prove that it yields near-logarithmic regret with respect to any policy that achieves an expected discounted reward that is within $\epsilon$ of the optimal.

## II. Model

For our problem formulation, we will look at the following description of an RMAB. Consider the problem of probing $N$ independent Markovian channels. Each channel $i$ has two states - good (denoted by 1) and bad (denoted by 0), with transition probabilities $\{p_{01}^{(i)}, p_{11}^{(i)}\}$, for the transitions from 0 to 1 and 1 to 1 respectively. At each time $t$, the player chooses one channel $i$ to probe, denoted by the action $U(t)$, and receives a reward equal to the state, $S_i(t)$, of the channel (0 or 1). The objective is to design a policy that chooses the chain at each time to maximize a long-term reward, which will be mathematically formulated shortly.

It has been shown that a sufficient statistic to make an optimal decision is given by the conditional probabilty that each channel is in state 1 given all past observations and decisions [10]. We will refer to this as the belief vector, denoted by $\Omega(t) \triangleq \{\omega_1(t), ..., \omega_N(t)\}$, where $\omega_i(t)$ is the conditional probability that the $i$-th channel is in state 1. Given the sensing action $U(t)$ in observation slot $t$, the belief can be recursively updated as follows:

$$w_i(t+1) \quad = \quad \begin{cases} p_{11}^{(i)} & , \quad i \in U(t), S_i(t) = 1 \\ p_{01}^{(i)} & , \quad i \in U(t), S_i(t) = 0 \\ \tau(\omega_i(t)) & , \quad i \notin U(t) \end{cases} \quad (1)$$

where $\tau(\omega_i(t)) \triangleq \omega_i(t)p_{11}^{(i)} + (1 - \omega_i(t))p_{01}^{(i)}$ denotes the one-step belief update for unobserved channels.

In our study, we will focus on the discounted reward criterion. For a discount parameter $\beta$ ($0 \leq \beta < 1$), the reward is defined as,

$$\mathrm{E}_\pi \left( \sum_{t=0}^{\infty} \beta^t R_{\pi(\Omega(t))} | \Omega(0) = x_0 \right) \quad (2)$$

where $R_{\pi(\Omega(t))}$ is the reward obtained by playing strategy $\pi(\Omega(t))$, where $\pi : \Omega(t) \rightarrow U(t)$ is a policy, which is defined to be a function that maps the belief vector $\Omega(t)$ to the action in $U(t)$ in slot $t$.

The discounted reward criterion is used due to its inherently tunable nature that gives importance both to the immediate reward, unlike the average reward criterion, and the future, unlike the myopic criterion.

## III. Structure of the optimal policy

We will now derive the structure of the optimal policy for a 2-state Bayesian RMAB. For ease of exposition, we describe our result in the context of a problem with 2 channels. However, as we discuss, our key structural result in this section readily generalizes to any number of channels.

We first define the value function, which is the the expected reward for the player when the optimal policy is adopted, recursively. Next, we derive a key property of the optimal value function and use it to characterize the structure of the optimal policy.

### A. Optimal Value Function

In this part, we will derive the optimal value function, i.e., the expected discounted reward obtained by a player who uses the optimal policy. Let $V_\beta(\omega_1, \omega_2)$ denote the optimal value function with beliefs $\omega_1$ and $\omega_2$ of the two arms.

If the player chooses arm 1, and sees a good state (which occurs with probability $\omega_1$), he gets an immediate reward of 1 and a future reward of $\beta V_\beta(p_{11}^{(1)}, \omega_2 p_{11}^{(2)} + (1-\omega_2)p_{01}^{(2)})$. If he sees a bad state (which occurs with probability $1 - \omega_1$), he gets an immediate reward of 0 and a future reward of $\beta V_\beta(p_{01}^{(1)}, \omega_2 p_{11}^{(2)} + (1-\omega_2)p_{01}^{(2)})$.

Putting these together, the discounted payoff under the optimal policy given that the user chooses arm 1, is,

$$\begin{aligned} V_{\beta 1}(\omega_1, \omega_2) = \omega_1 (1 + \beta V_\beta(p_{11}, \omega_2 q_{11} + (1-\omega_2)q_{01})) \\ + (1-\omega_1)(\beta V_\beta(p_{01}, \omega_2 q_{11} + (1-\omega_2)q_{01})) \end{aligned} \quad (3)$$

Similarly, given that the player chooses arm 2, his discounted payoff under the optimal policy is,

$$\begin{aligned} V_{\beta 2}(\omega_1, \omega_2) = \omega_2 (1 + \beta V_\beta(\omega_1 p_{11} + (1-\omega_1)p_{01}, q_{11})) \\ + (1-\omega_2)(\beta V_\beta(\omega_1 p_{11} + (1-\omega_1)p_{01}, q_{01})) \end{aligned} \quad (4)$$

At each time instant, the optimal value function is the greater of the above two functions. Thus, the optimal value function is given by,

$$V_\beta(\omega_1, \omega_2) = \max\{V_{\beta 1}(\omega_1, \omega_2), V_{\beta 2}(\omega_1, \omega_2)\} \quad (5)$$

### B. Properties of optimal value function

***Lemma 1:*** $V_{\beta 1}(\omega_1, \omega_2)$ is linear in $\omega_1$ (keeping $\omega_2$ fixed) and $V_{\beta 2}(\omega_1, \omega_2)$ is linear in $\omega_2$ (keeping $\omega_1$ fixed).

*Proof:* These are trivial from the definition of $V_{\beta 1}$ and $V_{\beta 2}$. ■

***Lemma 2:*** $V_{\beta 1}(\omega_1, \omega_2)$ is convex in $\omega_2$, $V_{\beta 2}(\omega_1, \omega_2)$ is convex in $\omega_1$, and $V_\beta(\omega_1, \omega_2)$ is convex in $\omega_1$ and $\omega_2$.

*Proof:* This follows from convexity of optimal value functions proved in [11]. ■

### C. Characterization of Optimal Decision Region

We now present our main structural result.

***Theorem 1:*** The decision region where arm 1 is chosen is contiguous horizontally and the decision region for arm 2 is contiguous vertically. Mathematically, if $(\omega_1, \omega_2) \in \Phi_1$ and $(\omega_1', \omega_2) \in \Phi_1$ where $\omega_1 \leq \omega_1'$, $\Phi_1$ is the region where arm 1 is chosen in the decision region space, then $(\omega_1'', \omega_2) \in \Phi_1$, $\forall\ \omega_1'' \in [\omega_1, \omega_1']$ (and similarly for arm 2).

*Proof:* We will prove the theorem for the decision region for arm 2, and the result for arm 1 can be proved in an analogous manner.

Let $(\omega_1, \omega_2^{(1)}), (\omega_1, \omega_2^{(2)}) \in \Phi_2$, the decision region for arm 2. Then, for $\alpha \in [0, 1]$, we have,

$$\begin{aligned} & V_\beta(\omega_1, \alpha\omega_2^{(1)} + (1-\alpha)\omega_2^{(2)}) \\ & \leq \alpha V_\beta(\omega_1, \omega_2^{(1)}) + (1-\alpha)V_\beta(\omega_1, \omega_2^{(2)}) \\ & = \alpha V_{\beta 2}(\omega_1, \omega_2^{(1)}) + (1-\alpha)V_{\beta 2}(\omega_1, \omega_2^{(2)}) \\ & = V_{\beta 2}(\omega_1, \alpha\omega_2^{(1)} + (1-\alpha)\omega_2^{(2)}) \\ & \leq V_\beta(\omega_1, \alpha\omega_2^{(1)} + (1-\alpha)\omega_2^{(2)}) \end{aligned} \quad (6)$$

using the above lemmas.

Hence all the inequalities hold with equality sign, and

$$V_{\beta 2}(\omega_1, \alpha\omega_2^{(1)} + (1-\alpha)\omega_2^{(2)}) = V_\beta(\omega_1, \alpha\omega_2^{(1)} + (1-\alpha)\omega_2^{(2)})$$

Or in other words, $\alpha\omega_2^{(1)} + (1-\alpha)\omega_2^{(2)} \in \Phi_2$. ■

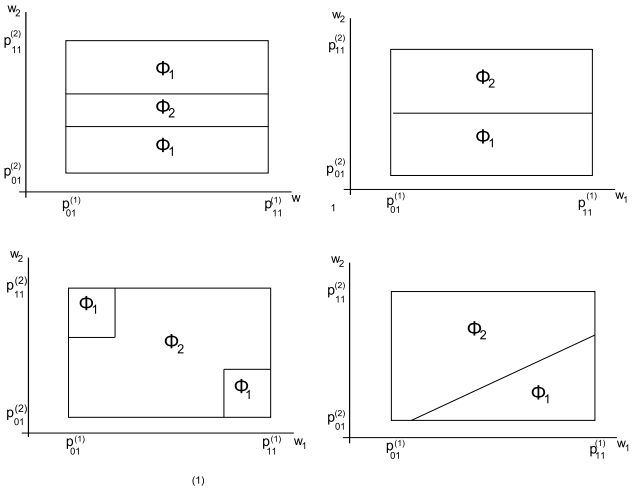


Fig. 1. Examples of possible decision regions satisfying the optimal structure.

***Lemma 3:*** It is sufficient to restrict the analysis of the decision region of any policy governed solely by the beliefs of the states of the arms to the boundary of the rectangle with vertices given by $\min\{p_{01}^{(1)}, p_{11}^{(1)}\}$, $\max\{p_{01}^{(1)}, p_{11}^{(1)}\}$, $\min\{p_{01}^{(2)}, p_{11}^{(2)}\}$, and $\max\{p_{01}^{(2)}, p_{11}^{(2)}\}$.

*Proof:* Recall that in our model, we had assumed the initial belief vector to be the vector of stationary probabilities of the Markov chains. It is easy to see that this lies between $\min\{p_{01}^{(i)}, p_{11}^{(i)}\}$ and $\max\{p_{01}^{(i)}, p_{11}^{(i)}\}$.

By definition of the belief update in (1), the belief of the state of an arm can only take values $p_{01}^{(i)}$, or $p_{11}^{(i)}$, or $\tau(\omega_i(t))$. From the recursive nature of $\tau$, the belief for an arm $i$ always lies between (and possibly including) $p_{01}^{(i)}$ and $p_{11}^{(i)}$ and tends to the stationary distribution of the Markov chain for the arm.

Hence, the two-dimensional velief vector lies in the rectangle formed by these extreme points. Further, since any policy always observes at least one of the states (and consequently giving rise to a belief of $p_{01}^{(i)}$ or $p_{11}^{(i)}$ for the observed arm $i$), the decision region is restricted to the boundary of the rectangle. ■

*Remark 1:* A noteworthy point is that we have used only the linearity of $V_{\beta i}(\omega_1, \omega_2)$ in the $i$-th argument and the convexity of the optimal value function in proving Theorem 1. Since these properties are both valid for number of arms, this immediately extends the theorem to those cases as well. Specifically, for an RMAB with $N$ arms, each having a good and a bad state, the decision region for arm $i$ is contiguous along the $i$-th dimension axis of the belief space. Similarly Lemma 3 can be generalized to any $N$ to indicate that it is sufficient to restrict attention to the hyper-plane boundaries of the corresponding $N$-dimensional hypercube.

We will now show that the optimal policy can only have 0, 2 or 4 intersection points with the decision boundary.

***Lemma 4:*** The optimal policy having the structure derived in Theorem 1 can have only 0, 2 or 4 intersection points with the decision boundary defined in Lemma 3.

*Proof:* We will outline the proof of this lemma in several steps. Recall from Lemma 3 that it is sufficient to restrict our attention to the boundary of the rectangle with vertices given by $\min\{p_{01}^{(1)}, p_{11}^{(1)}\}$, $\max\{p_{01}^{(1)}, p_{11}^{(1)}\}$, $\min\{p_{01}^{(2)}, p_{11}^{(2)}\}$, and $\max\{p_{01}^{(2)}, p_{11}^{(2)}\}$.

An immediate observation is that the number of intersection points of the optimal policy with the decision boundary is even (it cannot be odd because then the decision regions do not make sense). From Theorem 1, no edge of the decision boundary rectangle can have more than two points of intersection (if it does, then it violates one of the contiguity conditions). Also from Theorem 1, no two adjacent edges can both have two intersection points at the same time (if they do, one of them violates the contiguity condition).

The first and second observation imply that the optimal policy has 0, 2, 4, 6 or 8 intersection points with the boundary. The third implies that 8 points are not possible. It also implies that the only way to have 6 intersection points is to have two intersection points each on opposite edges. It is easy to see that this contradicts the contiguity condition for one of these edges. Hence, 6 is not possible. ■

*Remark 2:* An immediate consequence of Theorem 1, Lemma 3 and Lemma 4 is that the optimal policy has a threshold structure. A few example regions are illustrated in Fig 1.

## IV. Countable Policy Representation

We now focus our attention to the special case when $N = 2$ and both channels are positively correlated. We will show in this section, that in this case, the optimal policy can be represented in such a way that it must be one of a well-defined countably infinite set of policies.

***Lemma 5:*** For a positively-correlated 2-armed, 2-state RMAB, the belief update for an unobserved arm is monotonically increasing or decreasing towards the steady state probability.

*Proof:* For a positively-correlated Markov chain, $p_{11} \geq p_{01}$. From the defintion of belief update for an unobserverd

arm, $\tau$, in (1), the monotonic nature of belief update is established. ■

Because of this monotonicity property, the above-derived thresholds on the belief space can be translated into lower and upper thresholds on the time spent on a channel. Because of this, the optimal policy in this case can be represented as a mapping from $A$ to $B$, where $A$ is a vector of three binary indicators. $A[1]$ represents the current arm, $A[2]$ represents the current arm's state, and $A[3]$ represents the state of the other arm when it was last visited. $B[1]$, and $B[2]$ are lower and upper time thresholds on the time spent on the current channel that trigger a switch to the other channel. $B[1]$ and $B[2]$ can take on the countably infinite values of all natural numbers and also the symbol $\infty$ (to represent "never"). The corresponding policy maintains a counter $C(t)$ that is reset to 1 every time a new arm is played, and incremented by 1 each time the same arm has been played again. The meaning of this mapping is that whenever the condition in $A$ is satisfied, the policy should switch arms at the next time $t+1$ if the counter $C(t) > B[1]$ or if $C(t) < B[2]$.

A different way of putting this is that the optimal policy can be described by a 16-tuple, corresponding to all possible pairs $B[1]$, $B[2]$ in order, for each of the 8 different values that the three binary elements of $A$ can take on. This, in turn, implies that the optimal policy can be searched for among a countably infinite set (consisting of all such 16-tuples)[2].

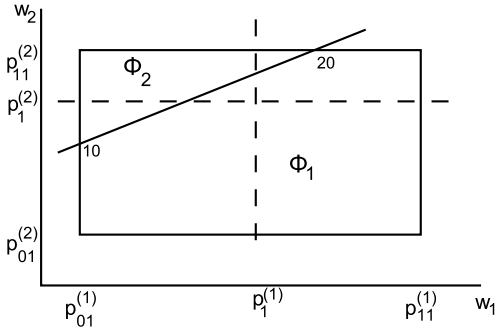


Fig. 2. Example of an optimal decision region to illustrate the countable policy representation

**Example**: We illustrate our countable policy representation with an example. Consider the case when the decision region looks like Fig 2, with the steady state probabilities of seeing a 1 in each channel represented by the vertical and horizontal dashed lines. For this example, the left-side threshold corresponds to the belief for arm 2 when it has not been played for 10 time steps assuming it was last observed in state 1 while arm 1 was observed just in the last step to be in state 0. Note that because this threshold is to the top of the steady state probability of arm 2, it could never reach this threshold if it was last observed to be in state 0. Similarly, the right-top threshold corresponds to the belief for arm 1 when it has not been played for 20 time steps assuming it was last observed in state 1 and arm 2 has just been observed in the last step to be in state 1. The optimal policy corresponding to this decision region can be depicted as in table I. For instance, when the last observed state on the previous arm was 0, and we observe the current arm, 1, to be in state 0, we will play it till the counter $C(t)$ crosses 10. The entries in the $B[1]$ and $B[2]$ give the conditions on $C(t)$ when the policy switches arms. Note that some entries may be irrelevant. For instance, we would never switch from arm 1 on a 1 because it lies entirely in decision region $\Phi_1$. Finally, the sixteen-tuple representation of this particular policy would then be simply $[10, \infty, \infty, \infty, 1, \infty, \infty, \infty, 1, \infty, \infty, \infty, 1, \infty, 1, 20]$

| current arm | state of current arm | last state of previous arm | $B[1]$ | $B[2]$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 10 | $\infty$ |
| 1 | 1 | 0 | $\infty$ | $\infty$ |
| 1 | 0 | 1 | 1 | $\infty$ |
| 1 | 1 | 1 | $\infty$ | $\infty$ |
| 2 | 0 | 0 | 1 | $\infty$ |
| 2 | 1 | 0 | $\infty$ | $\infty$ |
| 2 | 0 | 1 | 1 | $\infty$ |
| 2 | 1 | 1 | 1 | 20 |

TABLE I
THRESHOLD POLICY FOR FIG. 2

## V. THE NON-BAYESIAN CASE

We now turn to the non-Bayesian case for $N = 2$ positively correlated channels, to develop an online learning policy that takes advantage of the countable policy representation described in the previous section. In the non-Bayesian case, the underlying transition probability matrices are not known to the user, who must adaptively select arms over time based on observations.

### A. Mapping to infinite-armed MAB

The crux of our mapping is to consider each possible 16-tuple description of the threshold-based optimal policy as an arm in a new multi-armed bandit. As there are countably many of them, they can be renumbered as arm 1, 2, 3, ... Now, even through we do not know what the underlying transition probability matrices are, we know that the optimal policy corresponds to one of these arms. Each arm can be thought of as having a countably infinite number of states (these states correspond to the belief vector of the states of the arms). The states evolve in Markovian fashion depending on the policy being used.

Although the number of arms in the mapping has now increased dramatically (from two to infinity), this mapping simplifies a crucial aspect of the learning problem. Now we need only identify the single best arm, and no longer have to switch dynamically between them. The only other known strong result on non-Bayesian restless MAB [4] also performs a similar mapping; however in that case the mapping is only to two arms.

[2]We omit the details for brevity, but it is possible to map these tuples to a countable infinite set so that the optimal policy corresponds to some finite numbered element.

We first present a sublinear regret policy for a non-Bayesian MAB with countably infinite arms, each having an i.i.d reward. We then present a variant policy that also achieves a sublinear regret, with the key difference that it separates the exploration and exploitation phases. Finally, we give a policy for our 2-state, 2 positively correlated channels problem that builds on the variant policy for i.i.d. rewards, and show that it achieves a sublinear regret with respect to an $\epsilon$-optimality criterion.

### *B. MAB with countably infinite i.i.d. arms*

We consider a multi-armed bandit problem with countably infinite arms, where time is indexed by $n$. The reward got from an arm $i$ at $n$, denoted by $X_i(n)$, is an unknown random process which evolves i.i.d over time. At each time slot, an arm $i$ is selected under a strategy $\pi$, and the reward of the selected arm $X_i(n)$ is observed. The optimal arm is defined to be the arm with the highest expected reward, and we assume that the index of the optimal arm, i.e., $i^* = \arg\max_i \mathbb{E}[X_i]$, is finite. We assume that there exists a non-zero minimum difference between the expected rewards of the optimal arm and a suboptimal arm.

We evaluate a policy $\pi$ for this problem in terms of *regret*, which we define as follows:

$$\mathfrak{R}^{\pi}(n) = \mathbb{E}[\text{number of times suboptimal arms are played by policy } \pi \text{ in } n \text{ time slots}] \quad (7)$$

We present the following policy (Algorithm 1) for this problem, and prove an upper bound for the regret achieved by it. This policy, that we call UCB-CI, generalizes the well-known UCB1 policy of Auer *et al.* [12].

Let $n$ be the total number times the multiarmed bandit has been run, $n_i$ be the number of times arm $i$ has been selected. Let $M$ denote the set of arms added to the algorithm. $f(n)$ is a slowly growing function.

Define index of arm $i$ to be,

$$\textbf{index}_i(n) = \overline{X}_{i,n_i} + \sqrt{\frac{2\ln n}{n_i}} \quad (8)$$

where $\overline{X}_{i,n_i}$ is the sample mean of arm $i$ from $n_i$ selections of it.

**Algorithm 1** A policy for MAB with countably infinite arms yielding i.i.d rewards (UCB-CI)

**Initialize:** Add $f(0)$ arms to $M$. Select arms in $M$ once. Observe rewards and update $n$, $n_i$ and $\overline{X}_{i,n}$.
**for** slot $n = 1, 2, ...$ **do**
  Select the arm with the highest index. Observe reward and update $n$, $n_i$ and indices.
  Add additional arms to set $M$ s.t. the total number of arms is $f(n+1)$. Select each new arm once. Observe rewards and update $n$, $n_i$ and $\overline{X}_{i,n}$.
**end for**

***Theorem** 2:* For countably infinite i.i.d arms, when the number of arms in the set $M$ at time $n$ is upper bounded by $f(n)$, where $f(n)$ is a discrete, non-decreasing divergent function of time, using UCB-CI gives a regret that scales as [3] $O(f(n)\ln(n))$.

*Proof:* We will bound $T_i(n)$, the number of times arm $i$ ($\neq i^*$) is played in $n$ plays. Let $I_t$ be the arm played at time $t$ and $\{I_t = i\}$ be its indicator function.

Let $n \in \mathbb{N}$ be s.t. the set of arms in $M$ before play $n$, which we will, from now, denote by $M_n$, contains the optimal arm. Let $n_0$ be the smallest play number for which $M_{n_0}$ contains the optimal arm. For any arm $i \neq i^*$ in set $M_n$,

Define $\overline{X}_{i,T_i}$ as the mean of the observed rewards of arm $i$ for $T_i$ plays. Define $c_{t,s} = \sqrt{\frac{2\ln t}{s}}$ and $\Delta_i = \mathbb{E}[X_{i^*}] - \mathbb{E}[X_i]$.

Then,

$$\begin{aligned}
T_i(n) &= 1 + \sum_{t=f(0)+1}^{n} \{I_t = i\} \\
&= 1 + \sum_{t=f(0)+1}^{n_0} \{I_t = i\} + \sum_{t=n_0+1}^{n} \{I_t = i\} \\
&\leq C + l + \sum_{t=n_0+1}^{n} \{I_t = i, T_i(t-1) \geq l\} \\
&\leq C + l + \sum_{t=n_0+1}^{n} \{\overline{X}^*_{T^*(t-1)} + c_{t-1,T^*(t-1)} \\
&\qquad \leq \overline{X}_{i,T_i(t-1)} + c_{t-1,T_i(t-1)}, T_i(t-1) \geq l\} \\
&\leq C + l + \sum_{t=n_0+1}^{n} \{\min_{0<s<t} \{\overline{X}^*_s + c_{t-1,s}\} \\
&\qquad \leq \max_{l \leq s_i < t} \{\overline{X}_{i,s} + c_{t-1,s_i}\}\} \\
&\leq C + l + \\
&\quad \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=l}^{t-1} \{\overline{X}^*_s + c_{t,s} \leq \overline{X}_{i,s_i} + c_{t,s_i}\}
\end{aligned}$$

where $C = \sum_{t=f(0)+1}^{n_0} \{I_t = i\}$, which is a constant defined by $n_0$.

Following a similar analysis as in [12], $\overline{X}^*_s + c_{t,s} \leq \overline{X}_{i,s_i} + c_{t-1,s_i}$ implies at least one of the following

$$\overline{X}^*_s \leq \mathbb{E}[X_{i^*}] - c_{t,s} \quad (9)$$

$$\overline{X}_{i,s} \geq \mathbb{E}[X_i] + c_{t,s_i} \quad (10)$$

$$\mathbb{E}[X_{i^*}] < \mathbb{E}[X_i] + 2c_{t,s_i} \quad (11)$$

By Chernoff-Hoeffding bound,

$$\mathbb{P}\{\overline{X}^*_s \leq \mathbb{E}[X_{i^*}] - c_{t,s}\} \leq e^{-4\ln t} = t^{-4}$$
$$\mathbb{P}\{\overline{X}_{i,s} \geq \mathbb{E}[X_i] + c_{t,s_i}\} \leq e^{-4\ln t} = t^{-4}$$

For $l = \lceil \frac{8\ln n}{\Delta_i^2} \rceil$, the third event in (11) does not happen.

[3] We use asymptotic notation here for simplicity; the proof of Theorem 2 in fact shows that the upper bound of regret holds uniformly over time.

Thus, we get,

$$\begin{aligned}\mathbb{E}[T_i(n)] &\leq C + \lceil \frac{8\ln n}{\Delta_i^2} \rceil + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=l}^{t-1} 2t^{-4} \\ &\leq C + 1 + \frac{8\ln n}{\Delta_i^2} + \frac{\pi^2}{3} \\ &\leq C' + \frac{8\ln n}{\Delta_i^2}\end{aligned}$$

Since the number of arms at $n$-th play is upper bounded by $f(n)$, we have,

$$\begin{aligned}\mathfrak{R}^{UCB-CI}(n) &\leq C' f(n) + \frac{8\ln n f(n)}{\Delta_i^2} \\ &= O(f(n)\ln(n))\end{aligned}$$

■

### C. A variant policy for countably infinite i.i.d arms

We now present a variant policy, that we call DSEE-CI (for deterministic sequencing of exploration and exploitation for countably infinite arms), for the same model considered in the previous part, which is an extension of the DSEE policy in [13].

The main difference between DSEE-CI and UCB-CI is that the exploration and exploitation phases are separated out. Each arm is explored for the *same number of times* and then the arm having the highest sample mean is played for some time during the exploitation phase. It is important to note that the observations made in the exploitation phase are not used in updating the sample mean. This process is repeated with the relative duration of the exploitation phase increasing with time. The DSEE-CI policy is detailed in Algorithm 2. As before, $f(n)$ is a slowly growing function that upper bounds the number of arms in play at time $n$. $M$ is the actual number of arms in play at any given time and $M(n)$ denotes the set $M$ at time $n$. $|M|$ denotes the cardinality of set $M$. At any time, let $Z$ denote the cumulative number of slots used in the exploration phase up to that time. And let $c$ be a positive constant.

In the following theorem, we outline the proof of the upper bound of the regret of this algorithm. For simplicity, we adapt the proof of the UCB-CI policy for this; however, in principle it is possible to prove regret bounds under more general assumptions on the reward process (as shown in [13], for the finite arms case).

***Theorem** 3:* For countably infinite arms, when the number of arms in set $M$ at time $n$ is upper bounded by $f(n)$, where $f(n)$ is a discrete, increasing function of time, DSEE-CI yields a regret that is $O(g(n)\ln n)$, where $g(n)$ is a monotonic function of $f(n)$.

*Proof:* The proof follows along similar lines to Theorem 2. Note that the second term of the index defined in (8) is the same for each arm before the start of an exploitation phase. Thus, the first term, i.e, the sample mean, determines the index values for the arms. Also, by choosing $c$ large enough (larger than all $\frac{\mathbb{E}[T_i]}{\ln(n)}$ in Theorem 2, which depends on the minimum distance between the optimal arm and any suboptimal arm), it can be argued that a sub-optimal arm is played in the exploitation phase only $j(n)\ln n$ times, where $j(n)$ is a function of $f(n)$. Also, for the exploration phase, the cumulative length of the phase is always less than $Z = c|(M(n)|\ln n$. Thus, the total regret incurred by Algorithm 2 is $O(g(n)\ln n)$ where $g(n)$ is some function of $f(n)$, which can therefore be made to grow arbitrarily slowly. ■

**Algorithm 2** A variant policy for MAB with countably infinite arms and i.i.d. rewards (DSEE-CI)

**Initialize:** Add $f(0)$ arms to $M$. Select arms in $M$ once. Observe rewards and update $n_i$ and $\overline{X}_{i,n_i}$. Set $Z$ to be $f(0)$.
**while** 1 **do**
  **Exploitation:** Select the arm with the highest sample mean $\overline{X}_{i,n_i}$. Play the arm till slot $n'$ where $n'$ is determined by $Z = c|(M(t')|\ln n'$.
  **Exploration:** Add additional arms to set $M$ s.t. the total number of arms is $f(n'+1)$.
  For each old arm in $M$, play the arm once and observe its reward. Update $n_i$ and $\overline{X}_{i,m_i}$.
  For each newly adder arm in $M$, play the arm till it has been played for the same number of times as the old arms. Observe rewards and update $n_i$ and $\overline{X}_{i,n_i}$.
  Increment $Z$ by the number of slots used up in the exploration phase.
**end while**

### D. Online Learning Policy for non-Bayesian RMAB with 2 positively correlated channels

We are now in a position to tackle the problem of identifying the optimal policy for our problem (2-state RMAB with two positively correlated channels). Recall our mapping of the policies defined by the 16-tuples for the original problem to countably infinite arms in Section IV.

We define the arm that has the highest discounted reward as the optimal arm. It is given as:

$$i^* = \arg\max_i \mathbb{E}[\sum_{t=0}^{\infty} \beta^t R_i(t) | \Omega(0) = x_0] \tag{12}$$

where $R_i(t)$ is the reward obtained during the $t$-th play from arm $i$. Recall that we have set up the mapping in such a way that the index of this optimal arm is a finite number.

We now define an *$\epsilon$-optimal arm* to be an arm whose infinite horizon discounted reward is within $\epsilon$ of that of $i^*$.

Our regret definition with respect to an $\epsilon$-optimal arm is as follows.

$$\begin{aligned}\mathfrak{R}_\epsilon^\pi(n) = \mathbb{E}[&\text{number of times an arm whose discounted} \\ &\text{reward is lesser than that of the optimal arm } i^* \text{ by} \\ &\text{more than } \epsilon \text{ is played by strategy } \pi \text{ in } n \text{ time slots}]\end{aligned} \tag{13}$$

We now define a *playable state* and show that, using a predetermined strategy, we can reach a playable state in finite expected time from any initial state.

A *playable state* is any feasible belief vector that can be reached by a strategy that has selected each channel at least once. It is described by a four-tuple of its attributes: (*current channel, current channel's state, previous channel's last observed state and number of turns current channel has been played*) with respect to such a strategy. Note that since we define a playable state to be reached by some feasible strategy that selects a channel at each time, physically unrealizable states for a particular problem are eliminated.

***Lemma** 6:* There exists a predetermined strategy that takes any initial state to a playable state in finite expected time.

*Proof:* Our predefined strategy is as follows. We play the *previous channel* till we observe the *last observed state of that channel* for the playable state. We then switch to the other channel and play it for the *number of turns* in the playable state attribute. If we observe the *current channel's state* on the last turn, we stop. Otherwise, we switch to the *previous channel* and repeat the process.

Since the probability of reaching any channel's state is finite and non-zero with Markovian transitions (if it were zero, by definition of a playable state, a strategy that switches to the current channel on the *last observed state of the previous channel* would reach the playable state trivially), there is a finite expected time to reach the playable state because it is just a sequence of states with each having a positive probability of being reached. ■

From now onwards, $x_0$ will always denote a playable state. With our definition of regret, we can now equivalently restrict our horizon to any $T \geq T_0$ plays of an arm by the following lemma.

***Lemma** 7:* There exists a $T_0$ such that $\forall T \geq T_0$, an $\epsilon$-optimal arm for the finite horizon discounted reward criterion up to time $T$ is the same as that for the infinite horizon discounted criterion, if all rewards are finite and non-negative.

*Proof:* For the optimal arm $i^*$ defined by (12) and any $T > 0$, define $c(T)$ as

$$c(T) = \mathbb{E}[\sum_{t=T+1}^{\infty} \beta^t R_i^*(t) | \Omega_0 = x_0] \tag{14}$$

Since the rewards during a play are non-negative, this implies that $c(T)$ is a non-increasing function of $T$. Therefore, for any $0 < c_0 < \epsilon$, $\exists\, T_0$ s.t.

$$\mathbb{E}[\sum_{t=T_0+1}^{\infty} \beta^t R_i^*(t) | \Omega_0 = x_0] \leq c_0 < \epsilon. \tag{15}$$

Also note that for any sub-$\epsilon$-optimal arm $i'$,

$$\begin{aligned} \epsilon \leq\ & \mathbb{E}[\sum_{t=0}^{\infty} \beta^t R_{i^*}(t) | \Omega(0) = x_0] \\ & - \mathbb{E}[\sum_{t=0}^{\infty} \beta^t R_{i'}(t) | \Omega(0) = x_0] \end{aligned} \tag{16}$$

We then have,

$$\begin{aligned} & \mathbb{E}[\sum_{t=0}^{\infty} \beta^t R_{i^*}(t) | \Omega(0) = x_0] \\ & \quad - \mathbb{E}[\sum_{t=0}^{\infty} \beta^t R_{i'}(t) | \Omega(0) = x_0] \geq \epsilon \\ & \Rightarrow \mathbb{E}[\sum_{t=0}^{T_0} \beta^t R_{i^*}(t) | \Omega(0) = x_0] \\ & \quad - \mathbb{E}[\sum_{t=0}^{\infty} \beta^t R_{i'}(t) | \Omega(0) = x_0] \\ & \quad + \mathbb{E}[\sum_{t=T_0+1}^{\infty} \beta^t R_{i^*}(t) | \Omega(0) = x_0] \geq \epsilon \\ & \Rightarrow \mathbb{E}[\sum_{t=0}^{T_0} \beta^t R_{i^*}(t) | \Omega(0) = x_0] \\ & \quad - \mathbb{E}[\sum_{t=0}^{\infty} \beta^t R_{i'}(t) | \Omega(0) = x_0] > \epsilon - c_0 > 0 \end{aligned} \tag{17}$$

The strict inequality follows from the definition of $c_0$. Thus, by the non-increasing nature of the residual discounted reward as discussed above, $\forall T \geq T_0$, the $\epsilon$-optimal arms have a greater finite horizon discounted reward than even the infinite horizon discounted reward of any suboptimal channel, and consequently, its finite horizon discounted reward. ■

We present in Algorithm 3 a policy we refer to as R2PC for solving the 2-state RMAB problem with two positively correlated channels. The terms used in the description and the analysis of this policy are as follows.

A $T$-*play* of an arm is defined to be playing the arm for $T$ slots from the initial state $x_0$. An update of an index for an arm is calculated in the following way for the collection of the rewards obtained in the $T$-play of that arm. $m$ is the total number of $T$-plays and $m_i$ is the number of $T$-plays for arm $i$. Let $n(m)$ be the time at the beginning of $m$-th $T$-play. Let $Y_i(m)$ be the discounted reward got from the $T$-play of arm $i$, i.e., $Y_i(m) = \sum_{t=0}^{t=T} \beta^t R_i(n(m)+t)$. $\overline{Y}_{i,m_i}$ is the sample mean of the $m_i$ different observations of $Y_i$. As before, $f(n)$ is a slowly growing function of $n$ that upper bounds the total number of arms at the $n$-th play. $M$ is the set of arms at any time. $Z$ is the cumulative duration of the exploration phase up to a given time.

### *E. Analysis of R2PC*

We construct the proof of the regret bound for the R2PC policy in the following manner. First, we decompose the regret into that obtained by playing the predetermined strategy to reach $x_0$ and the part obtained during the $T$-play phases. We show that the reward obtained across $T$-plays of an arm is i.i.d and independent of the rewards from previous $T$-plays of other arms. Next, we prove the regret bound for the $T$-play. Finally, we show that the time taken to reach $x_0$ in each play of an arm is finite and hence, does not affect the order of the regret bound, thus completing our proof.

**Algorithm 3** An online learning policy for the two-state RMAB with two positively correlated channels (R2PC)

**Initialize:** Add $f(0)$ arms to $M$. Play the predetermined strategy to reach $x_0$. Select an arm in $M$. Play it for $T$ slots and observe the rewards. Repeat for each arm in $M$. Update $m$, $m_i$ and $\overline{Y}_{i,m_i}$. Set $Z$ to be $f(0)$.

**while** 1 **do**

**Exploitation:** Select the arm with the highest $\overline{Y}_{i,m_i}$. Play the arm till slot $n'$ where $n'$ is determined by $Z = c|(M(n')|\ln n'$.

**Exploration:** Add additional arms to set $M$ s.t. the total number of arms is $f(n'+1)$.

For each old arm in $M$, play the predetermined strategy to reach $x_0$. Then, play the arm for $T$ slots. Observe rewards and update $m$, $m_i$ and $\overline{Y}_{i,m_i}$

For each newly added arm in $M$, play the predetermined strategy to reach $x_0$. Then play the arm for $T$ slots. Repeat till the arm has the same $m_i$ as the old arms. Observe rewards and update $m$, $m_i$ and $\overline{Y}_{i,m_i}$.

Increment $Z$ by the number of $T$-plays in the exploration phase.

**end while**

***Lemma** 8:* The reward from each $T$-play of an arm is i.i.d. and independent of the rewards obtained in previous and future $T$-plays of any arm.

*Proof:* Since we start playing an arm and observing its rewards only from initial state $x_0$, the reward from each play of an arm is i.i.d. over each $T$-play because it is given by the state evolution of a Markov chain conditioned on the same initial state and played for the same duration. Also, by the same reasoning, the rewards obtained are independent of the past and future rewards for $T$-plays on any arm. ■

We now show our main theorem as follows.

***Theorem** 4:* The regret of the R2PC policy is bounded by[4] $O(h(n)\ln n)$.

*Proof:* Let us first consider the regret due to the $T$-plays. From Lemma 8, we know that the rewards obtained during each $T$-play of an arm in the exploration phase are i.i.d and independent of the previous and future plays of any other arm. Therefore, we can apply Theorem 2 and the regret achieved by the algorithm in the exploitation phase after $n$ slots is $O(g_1(n)\ln(n))$ where $g_1(n)$ is a monotonic function of $f(n)$.

We now show that the finite time taken by the predetermined strategy to reach $x_0$ does not affect the order of the regret bound. When we factor in the time taken to reach $x_0$, the expected length of time of the cummulative exploration phase at time $n$ is upper bounded by $c|M(n)|\ln(n)(T + \mathbb{E}[\text{time taken to reach } x_0]) \leq c' f(n)\ln(n)$ (from Lemma 6).

Thus, the total regret incurred by the R2PC policy, which is the sum of regrets during the exploration and the exploitation phases, is $O(h(n)\ln(n))$ where $h(n)$ is some monotonic function of $f(n)$. ■

[4]For Theorem 4 also, the regret bound holds uniformly over time.

## VI. Conclusion

In this paper, we have derived a structural result for the optimal policy for a two-state Bayesian restless multi-armed bandit with non-identical arms, under the infinite horizon discounted reward criterion. For the non-Bayesian version of this problem, in the special case of two positively correlated arms, we then developed a novel mapping to a different countably infinite-armed bandit problem. Using this mapping, we have proposed an online learning policy for this problem that yields near-logarithmic regret with respect to an $\epsilon$-optimal solution. Developing efficient learning policies for other cases remains an open problem. An alternative, possibly more efficient, approach to online learning in these kinds of problems might be to use the historical observations of each arm to estimate the P matrix, and use these estimates iteratively in making arm selection decisions at each time. It is, however, unclear at present how to prove regret bounds using such an iterative estimation approach.

## Acknowledgment

We would like to thank Yanting Liu at University of Southern California for her helpful discussions.

## References

[1] Q. Zhao, B. Krishnamachari, and K. Liu. On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance. *Wireless Communications, IEEE Transactions on*, 7(12):5431–5440, 2008.

[2] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari. Optimality of myopic sensing in multichannel opportunistic access. *Information Theory, IEEE Transactions on*, 55(9):4040–4050, 2009.

[3] K. Liu and Q. Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *Information Theory, IEEE Transactions on*, 56(11):5547–5567, 2010.

[4] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao. The non-bayesian restless multi-armed bandit: A case of near-logarithmic regret. In *proceedings of IEEE ICASSP*, pages 2940–2943, 2011.

[5] C. Tekin and M. Liu. Online learning in opportunistic spectrum access: A restless bandit approach. In *Proceedings of IEEE INFOCOM*, pages 2462–2470, 2011.

[6] H. Liu, K. Liu, and Q. Zhao. Logarithmic weak regret of non-bayesian restless multi-armed bandit. In *Proceedings of IEEE ICASSP*, pages 1968–1971, 2011.

[7] J. C. Gittins, K. D. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. Wiley, 2011.

[8] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.

[9] W. Dai, Y. Gai, and B. Krishnamachari. Efficient online learning for opportunistic spectrum access. *arXiv:1109.1552v1*, 2011.

[10] R. D. Smallwood and E. J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

[11] E. J. Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):282–304, 1978.

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

[13] K. Liu and Q. Zhao. Multi-armed bandit problems with heavy tail reward distributions. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, 2011.