Dynamic spectrum access under partial observations: A restless bandit approach

Nima Akbarzadeh
Electrical and Computer Engineering
McGill University
Montreal, Canada
nima.akbarzadeh@mail.mcgill.ca

Aditya Mahajan
Electrical and Computer Engineering
McGill University
Montreal, Canada
aditya.mahajan@mcgill.ca

***Abstract*—We consider a communication system where multiple unknown channels are available for transmission. Each channel is a channel with state which evolves in a Markov manner. The transmitter has to select L channels to use and also decide the resources (e.g., power, rate, etc.) to use for each of the selected channels. It observes the state of the channels it uses and receives no feedback on the state of the other channels. We model this problem as a partially observable Markov decision process and obtain a simplified belief state. We show that the optimal resource allocation policy can be identified in closed form. Once the optimal resource allocation policy is fixed, choosing the channel scheduling policy may be viewed as a restless bandit. We present an efficient algorithm to check indexability and compute the Whittle index for each channel. When the model is indexable, the Whittle index policy, which transmits over the L channels with the smallest Whittle indices, is an attractive heuristic policy.**

***Index Terms*—Channel scheduling, resource allocation, restless bandits.**

I. Introduction

Dynamic spectrum access is a key component of various applications including cognitive radio networks, resource constraint jamming and opportunistic over fading channels [1]. In such models, a transmitter has to transmit data over a set of time-varying channels. The transmitter does not know the current state of the channels based on past observations and the channel statistics.

When the transmitter knows the state of the channels, the system can be modeled as a Markov decision process (MDP). However, in many applications, the state of the channel is not known and is observed only sporadically. Such system can be modeled as a partially obvservable Markov decision process (POMDP). However, using MDP or POMDP models to obtain optimal scheduling and resource allocation policies suffers from the curse of dimensionality: the number of joint states for the channels is exponential in the number of channels; in case of POMDPs, the state space is also exponential in the number of states of the channel. Therefore, most of the results in the literature have analyzed such problems under simplifying modeling assumptions.

The problem of transmitting over a two-state Gilbert-Elliot channel under rate or power constraints has been investigated in [2], [3]. For a two-state channel, the posterior belief on the state of the channel is characterized by a real number. Under different modelling assumptions [2], [3] show that the optimal strategy is characterized by a threshold on the posterior belief.

Opportunistic scheduling over multiple two-state Gilbert-Elliot channels has been investigated in [4], [5]. When the channels are identical, the myopic policy is opitmal [5]. Similar results were obtained for multi-state channels in [6].

More general models of dynamic spectrum access have been investigated in [7]–[15]. Many of these results rely on modelling the scheduling problem as a restless bandit [16] and using the Whittle index heuristic. However, these papers either consider a fully-observable channel state [9]–[12] or restrict to two-state channels in case of partially observable channel state [13]–[15].

In this paper we present a restless bandit approach for dynamic spectrum access under partial observations. Unlike the previous papers in the literature we do not restrict to two-state channels. Therefore, the belief state in our model lies in a multi-dimensional simplex. Our main idea is to exploit the structure of the reachable set of belief states to identify an alternative information state which lies in a countable set. We then use results from countable state MDPs to approximate the countable information state by a finite information state. Using this approximation, we develop efficient algorithms to check for indexability and to compute Whittle index. When the problem is indexable, the set of channels to transmit may be chosen according to the Whittle index policy.

II. Model and Problem Formulation

A. The communication channels

Consider a communication system consisting of n independent channels indexed by the set $\mathcal{N} \coloneqq \{1, \dots, n\}$. Each channel $i \in \mathcal{N}$ is a channel with state; the state process $\{S_t^i\}_{t\geq 0}$, $S_t^i \in \mathcal{S}^i$, is a time-homogeneous Markov chain with initial distribution π_0^i and transition probability matrix P^i. The state processes $\{S_t^i\}_{t\geq 0}$, $i \in \mathcal{N}$, are independent across channels. We assume that $\mathcal{S}^i$ is an ordered finite set. We use $\boldsymbol{\mathcal{S}}$ to denote $\prod_{i\in\mathcal{N}} \mathcal{S}^i$ and $\boldsymbol{S}_t = (S_t^1, \dots S_t^n) \in \boldsymbol{\mathcal{S}}$ to denote the state of all channels.

Let $\mathcal{R} = \{\emptyset, r_1, \dots, r_k\}$ denote a set of allocable resources (e.g., rate, power, bandwidth, etc.), where $\emptyset$ indicates that no resources are allocated. If we transmit over channel $i \in \mathcal{N}$

978-1-7281-0954-1/19/$31.00 ©2019 IEEE

using resource $r \in \mathcal{R}$ when its state is $s \in \mathcal{S}^i$, then we receive a reward $\rho^i(s, r)$, where $\rho^i(s, \emptyset) = 0$ for all $s \in \mathcal{S}^i$.

Example 1 (Model of [2]) Suppose channel $i \in \mathcal{N}$ is a two-state Gilbert-Elliot channel with a good and bad state, denoted by $\mathcal{S}^i = \{s_{\text{BAD}}, s_{\text{GOOD}}\}$. The transmitter has the option of transmitting conservatively at a low rate (denoted by r_{LOW}) or transmitting aggressively at a high rate (denoted by r_{HIGH}). When transmitting conservatively, the transmission is always successful and r_{LOW} bits are communicated. When transmitting aggressively, the transmission is successful if the channel is in a good state, in which case r_{HIGH} bits are communicated; if the channel is a bad state, then the communication is unsuccessful and no data is communicated. The reward function may be written as

$$\rho^i(s, r) = \begin{cases} r_{\text{LOW}}, & \text{if } r = r_{\text{LOW}} \\ r_{\text{HIGH}}, & \text{if } r = r_{\text{HIGH}} \text{ and } s = s_{\text{GOOD}} \\ 0, & \text{otherwise.} \end{cases}$$

□

Example 2 (Model of [3]) Analogous to Example 1, suppose channel $i \in \mathcal{N}$ is a two-state Gilbert-Elliot channel with a good and bad state, denoted by $\mathcal{S}^i = \{s_{\text{BAD}}, s_{\text{GOOD}}\}$. The transmitter has the option of either transmitting at high power (denoted by p_{HIGH}), or transmitting at low power (denoted by p_{LOW}) or not transmitting (denoted by $\emptyset$). When the channel is in a bad state, no transmission is successful. When the channel is in a good state, transmitting at high power achieves a bit-rate of r_{HIGH} while transmitting at low power achieves a bit-rate of r_{LOW}. Thus, the reward function may be written as

$$\rho^i(s, r) = \begin{cases} r_{\text{LOW}}, & \text{if } r = p_{\text{LOW}} \text{ and } s = s_{\text{GOOD}} \\ r_{\text{HIGH}}, & \text{if } r = P_{\text{HIGH}} \text{ and } s = s_{\text{GOOD}} \\ 0, & \text{otherwise.} \end{cases}$$

□

Example 3 (Model of [5]) Analogous to the previous examples, suppose channel $i \in \mathcal{N}$ is a two-state Gilbert-Elliot channel with a good and bad state, denoted by $\mathcal{S}^i = \{s_{\text{BAD}}, s_{\text{GOOD}}\}$. The transmitter has the option of either transmitting (denoted by $r = 1$) or not (denoted by $r = \emptyset$). When the transmitter transmits and the channel is in s_{GOOD}, the transmission is successful, otherwise the transmission is not successful. Thus, the reward can be written as

$$\rho^i(s, r) = \begin{cases} 1, & \text{if } r = 1 \text{ and } s = s_{\text{GOOD}} \\ 0, & \text{otherwise.} \end{cases}$$

□

Example 4 (Model of [6]) In contrast to the previous examples, suppose channel $i \in \mathcal{N}$ is a multi-state channel with an ordered state space $\mathcal{S}^i$. The transmitter has the option of either transmitting (denoted by $r = 1$) or not (denoted by $r = \emptyset$). The probability of success depends on the state of the channel and it is denoted by p_s. Thus, the reward function is

$$\rho^i(s, r) = \begin{cases} p_s, & \text{if } r = 1 \\ 0, & \text{otherwise.} \end{cases}$$

□

B. The transmitter

A transmitter wants to communicate over the communication system described above. At time t, it makes two decisions: it selects L channels indexed by $\mathcal{L}_t \subset \mathcal{N}$ and chooses resources $\{R_t^i\}_{i \in \mathcal{L}_t}$, $R_t^i \in \mathcal{R}$ for those channels. For ease of notation, we denote these decisions by $\boldsymbol{A}_t = (A_t^1, \dots, A_t^n)$ and $\boldsymbol{R}_t = (R_t^1, \dots, R_t^n)$, where $A_t^i = 1$ if $i \in \mathcal{L}_t$ and $(A_t^i, R_t^i) = (0, \emptyset)$ for $i \notin \mathcal{L}_t$.

When the system is in state $\boldsymbol{S}_t$ and the transmitter chooses to transmit over $\boldsymbol{A}_t$ channels using $\boldsymbol{R}_t$ resources, we receive a reward

$$\rho(\boldsymbol{S}_t, \boldsymbol{R}_t, \boldsymbol{A}_t) = \sum_{i \in \mathcal{N}} \rho^i(S_t^i, R_t^i) A_t^i. \tag{1}$$

After transmitting, the transmitter completely observes the states S_t^ℓ for all $\ell \in \mathcal{L}_t$; it receives no new information about other channels. We denote this observation by $\boldsymbol{Y}_t := (Y_t^1, \dots, Y_t^n)$, where

$$Y_t^i = \begin{cases} S_t^i, & \text{if } A_t^i = 1 \\ \mathfrak{E}, & \text{if } A_t^i = 0, \end{cases} \tag{2}$$

where $\mathfrak{E}$ denotes the event "no observation".

The decisions $\boldsymbol{A}_t$ and $\boldsymbol{R}_t$ are chosen based on the history of observations and decisions up to time t by the following *decision policies*:

$$\boldsymbol{A}_t = f_t(\boldsymbol{Y}_{0:t-1}, \boldsymbol{R}_{0:t-1}, \boldsymbol{A}_{0:t-1}), \tag{3}$$

$$\boldsymbol{R}_t = g_t(\boldsymbol{Y}_{0:t-1}, \boldsymbol{R}_{0:t-1}, \boldsymbol{A}_{0:t}), \tag{4}$$

where $\boldsymbol{Y}_{0:t-1}$ is a short-hand notation for $(\boldsymbol{Y}_0, \dots, \boldsymbol{Y}_{t-1})$ and similar interpretations hold for $\boldsymbol{A}_{0:t-1}$ and $\boldsymbol{R}_{0:t-1}$. The collection of decision rules $\boldsymbol{f} = (f_0, f_1, \dots)$ and $\boldsymbol{g} = (g_0, g_1, \dots)$ are called the *channel-selection strategy* and the *resource allocation strategy*, respectively. We refer to $(\boldsymbol{f}, \boldsymbol{g})$ as the *communication strategy*.

We consider the infinite horizon discounted reward performance metric. Given a discount factor $\beta \in (0, 1)$, the performance of a communication strategy $(\boldsymbol{f}, \boldsymbol{g})$ is given by

$$J(\boldsymbol{f}, \boldsymbol{g}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \rho(\boldsymbol{S}_t, \boldsymbol{R}_t, \boldsymbol{A}_t)\right]. \tag{5}$$

C. The optimization problem

We are interested in the following optimization problem.

Problem 1 *Given a discount factor $\beta \in (0, 1)$, a set of resources $\mathcal{R}$, and the state space, transition probability, and reward function $(\mathcal{S}^i, P^i, \rho^i)_{i \in \mathcal{N}}$ for all channels, choose a communication strategy $(\boldsymbol{f}, \boldsymbol{g})$ to maximize $J(\boldsymbol{f}, \boldsymbol{g})$ given by* (5)*, where the maximum is taken over all history dependent strategies of the form* (3) *and* (4)*.*

The problem formulated above is a multi-stage optimization problem where the decision maker has only partial observations of the state of the channels. Therefore, the system is a partially observable Markov decision process (POMDP) [17].

978-1-7281-0954-1/19/$31.00 ©2019 IEEE

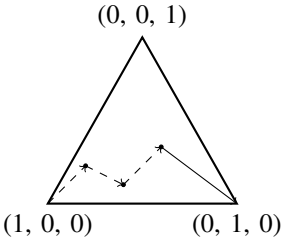

Fig. 1. Belief state dynamics for a 3-state channel i in the simplex $\Delta(\{1,2,3\})$. Dashed arrows show a sample realizations of the belief state evolution under $A_t^i = 0$ for three time steps and the solid arrow shows a sample realization of the belief state evolution under $A_t^i = 1$.

III. Simplification of the optimization problem

A. A simplified belief state for Problem 1

As argued above, Problem 1 is a POMDP. Following standard results in POMDPs, we define the belief state $\Theta_t \in \Delta(\boldsymbol{\mathcal{S}})$ of the system as follows: for any $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$,

$$\Theta_t(\boldsymbol{s}) = \mathbb{P}(\boldsymbol{S}_t = \boldsymbol{s} \,|\, \boldsymbol{Y}_{0:t-1}, \boldsymbol{R}_{0:t-1}, \boldsymbol{A}_{0:t-1}).$$

Note that Θ_t is a random variable that takes values in $\Delta(\boldsymbol{\mathcal{S}})$. Using standard results in POMDPs [17], we have the following.

Proposition 1 *In Problem 1, Θ_t is a sufficient statistic for $(\boldsymbol{Y}_{0:t-1}, \boldsymbol{R}_{0:t-1}, \boldsymbol{A}_{0:t-1})$. Therefore, the is no loss of optimality in restricting attention to communication strategies of the form $\boldsymbol{A}_t = \tilde{f}_t(\Theta_t)$ and $\boldsymbol{R}_t = \tilde{g}_t(\Theta_t, \boldsymbol{A}_t)$.* □

Note that it is also possible to write a dynamic program using Θ_t, but for reasons that will become apparent, we are not presenting the dynamic program in detail.

We first present a simplified information state. For that matter, for every $i \in \mathcal{N}$, define the belief $\Pi_t^i \in \Delta(\mathcal{S}^i)$ as follows: for any $s^i \in \mathcal{S}^i$,

$$\Pi_t^i(s^i) = \mathbb{P}(S_t^i = s^i \,|\, Y_{0:t-1}^i, R_{0:t-1}^i, A_{0:t-1}^i).$$

Similar to Θ_t, Π_t^i, $i \in \mathcal{N}$, are also distribution-valued random variables. The time evolution of these individual beliefs can be written as follows. For any t,

$$\Pi_{t+1}^i = \begin{cases} \Pi_t^i \cdot P^i, & \text{if } A_t^i = 0, \\ \delta_{S_t^i}, & \text{if } A_t^i = 1. \end{cases} \tag{6}$$

An example of such dynamics is depicted in Fig. 1.

Our first simplification for the structure of optimal communication policies is the following.

Proposition 2 *For any $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$, we have*

$$\Theta_t(\boldsymbol{s}) = \prod_{i \in \mathcal{N}} \Pi_t^i(s^i), \quad \textit{a.s.}. \tag{7}$$

Let $\boldsymbol{\Pi}_t$ denote $(\Pi_t^1, \dots, \Pi_t^n)$. Then, there is no loss of optimality in restricting attention to communication strategies of the form $\boldsymbol{A}_t = \hat{f}_t(\boldsymbol{\Pi}_t)$ and $\boldsymbol{R}_t = \hat{g}_t(\boldsymbol{\Pi}_t, \boldsymbol{A}_t)$. □

PROOF Eq. (7) follows from the conditional independence of channels, and the nature of the observation function. The structure of the optimal policies then follow immediately from Proposition 1. ■

B. The optimal resource allocation strategy

In this section, we show that the optimal rate allocation strategy can be computed offline. In particular, we have the following.

Proposition 3 *Define $\hat{g}^{i,*} \colon \Delta(\mathcal{S}^i) \times \{0,1\} \to \mathcal{R}$ as follows*

$$\hat{g}^{i,*}(\pi^i, 0) = \emptyset,$$
$$\hat{g}^{i,*}(\pi^i, 1) = \arg\max_{r^i \in \mathcal{R}} \sum_{s^i \in \mathcal{S}^i} \pi^i(s^i) \rho^i(s^i, r^i).$$

Let $\hat{g}^*(\boldsymbol{\pi}) = (\hat{g}^{1,*}(\pi^1, 1), \dots, \hat{g}^{n,*}(\pi^n, 1))$. Then, for any choice of the channel-selection strategy $\boldsymbol{f}$, the time-homogeneous strategy $(\hat{g}^*, \hat{g}^*, \dots)$ is an optimal resource allocation strategy. □

PROOF We define the expected instantaneous reward as

$$\begin{aligned} \rho(\boldsymbol{\Pi}_t, \boldsymbol{R}_t, \boldsymbol{A}_t) &= \mathbb{E}[\rho(\boldsymbol{S}_t, \boldsymbol{R}_t, \boldsymbol{A}_t) | \boldsymbol{\Pi}_t] \\ &= \sum_{i \in \mathcal{N}} \sum_{s^i \in \mathcal{S}^i} \Pi_t^i(s^i) \rho^i(s^i, R_t^i) A_t^i. \end{aligned} \tag{8}$$

By (5) and (8), we have

$$J(\boldsymbol{f}, \boldsymbol{g}) = \mathbb{E}^{\boldsymbol{f}, \boldsymbol{g}} \Bigg[\sum_{t=0}^{\infty} \beta^t \sum_{i \in \mathcal{N}} f(\Pi_t^i) \sum_{s^i \in \mathcal{S}^i} \Pi^i(s^i) \rho^i \left(s^i, g\left(\Pi_t^i, f(\Pi_t^i)\right)\right) \Bigg].$$

Let $J^* := \max_{\boldsymbol{f}} \max_{\boldsymbol{g}} J(\boldsymbol{f}, \boldsymbol{g})$ be the optimal performance measure. The key idea is that the evolution of the belief state $\boldsymbol{\Pi}_t$ depends only on $\boldsymbol{A}_t$ and not on $\boldsymbol{R}_t$ as shown in (6). Then, we have

$$\begin{aligned} J^* &= \max_{\boldsymbol{f}} \max_{\boldsymbol{g}} \mathbb{E}^{\boldsymbol{f}, \boldsymbol{g}} \Bigg[\sum_{t=0}^{\infty} \beta^t \sum_{i \in \mathcal{N}} f(\Pi_t^i) \sum_{s^i \in \mathcal{S}^i} \Pi_t^i(s^i) \rho^i(s^i, g(\Pi_t^i, f(\Pi_t^i))) \Bigg] \\ &= \max_{\boldsymbol{f}} \max_{\boldsymbol{g}} \mathbb{E}^{\boldsymbol{f}} \Bigg[\mathbb{E}^{\boldsymbol{g}} \Bigg[\sum_{t=0}^{\infty} \beta^t \sum_{i \in \mathcal{N}} f(\Pi_t^i) \sum_{s^i \in \mathcal{S}^i} \Pi_t^i(s^i) \rho^i(s^i, g(\Pi_t^i, f(\Pi_t^i))) \Bigg| \Pi_t^i \Bigg] \Bigg] \\ &= \max_{\boldsymbol{f}} \mathbb{E}^{\boldsymbol{f}} \Bigg[\sum_{t=0}^{\infty} \beta^t \sum_{i \in \mathcal{N}} f(\Pi_t^i) \max_{\boldsymbol{g}} \mathbb{E}^{\boldsymbol{g}} \Bigg[\sum_{s^i \in \mathcal{S}^i} \Pi_t^i(s^i) \rho^i(s^i, g(\Pi_t^i, f(\Pi_t^i))) \Bigg| \Pi_t^i \Bigg] \Bigg] \\ &= \max_{\boldsymbol{f}} \mathbb{E}^{\boldsymbol{f}} \Bigg[\sum_{t=0}^{\infty} \beta^t \sum_{i \in \mathcal{N}} f(\Pi_t^i) \max_{\boldsymbol{g}} \sum_{s^i \in \mathcal{S}^i} \Pi_t^i(s^i) \rho^i(s^i, g(\Pi_t^i, f(\Pi_t^i))) \Bigg]. \end{aligned}$$

978-1-7281-0954-1/19/$31.00 ©2019 IEEE

Therefore, there is no loss of optimality to define $\hat{g}^{i,*}$ as given the proposition statement. ∎

Thus, $\boldsymbol{R}_t$ can be optimized in an *open-loop* manner. The strategy given in Proposition 3, denoted by $\hat{g}^*$, maximizes the per-step expected reward.

The timeline of the sequence of events for channel i is shown below.

$$\ldots \to \underbrace{\Pi_t^i \to A_t^i \to R_t^i \to Y_t^i \to \rho_t^i}_{\text{time } t} \to \Pi_{t+1}^i \to \ldots$$

At time t, first, the transmitter forms a belief over the true state of the channel. Then, based on the belief state, it selects L channels. Then, based on the belief state and the channel-selection strategy, it specifies the resource allocated to the channels. Next, it transmits using these resources, and observes the state of the selected channels and receives a pay-off accordingly. The process then repeats at time $t+1$.

IV. Restless bandits, indexability, and the computation of Whittle index

Fix the resource allocation strategy as specified in Proposition 3. Then, the problem of choosing the optimal channel-selection strategy may be viewed as a restless bandit. In particular, we may think of $\{\Pi_t^i\}_{t\geq 0}$, $i \in \mathcal{N}$, as bandit processes. The transmitter can *activate* L of these processes. If processes i is activated in state π, its next state is one of the "corner" states $\{\delta_s : s \in \mathcal{S}^i\}$, where the probability that the next state is δ_s is equal to $\pi(s)$. The activated process also yields an expected reward

$$\bar{\rho}^i(\pi) = \max_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}^i} \pi(s)\rho^i(s,r).$$

If process i is not activated, it remains *passive* in state π, then it does not yield any reward and its next state is $\pi \cdot P^i$. Thus, the dynamics of the model are the same as (6) and the per-step reward function is given by $\bar{\rho}^i(\Pi_t^i)A_t^i$ for arm i at time t. Such a setup is the standard restless bandit model [16].

When a restless bandit problem satisfies a technical condition known as *indexability*, then a low-complexity *index strategy* known as the Whittle index can be proposed that is optimal in certain cases [18] and performs close to optimal for many applications. In the rest of this section, we provide the definition of indexability, Whittle index and propose an algorithm for computing the Whittle index.

A. Restless bandit formulation

The main idea of restless bandit formulation is to decompose the coupled n-channel optimization problem to n independent one-channel optimization problems. In particular, the optimization problem for channel $i \in \mathcal{N}$ considers the process $\{\Pi_t^i\}_{t\geq 0}$ with dynamics (6) but a modified per-step reward of $(\bar{\rho}^i(\pi) - \lambda)a^i$. One may view λ as the cost for transmitting over channel i.

The performance of any time-homogeneous policy $\tilde{f}^i : \Delta(\mathcal{S}^i) \to \{0,1\}$ is given by

$$J_\lambda^i(\tilde{f}^i) := \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \left(\bar{\rho}^i(\Pi_t^i) - \lambda\right) A_t^i\right]. \tag{9}$$

Then, we consider the following optimization problem.

Problem 2 *Given channel $i \in \mathcal{N}$, the discount factor $\beta \in (0,1)$, the cost $\lambda \in \mathbb{R}$, and the belief state space, transition probability, reward function tuple $(\Delta(\mathcal{S}^i), P^i, \rho^i)$, choose a policy $\tilde{f}^i$ to maximize $J_\lambda^i(\tilde{f}^i)$ given by* (9)*.*

Problem 2 is a Markov decision process which can be solved by dynamic programming as explained below.

Theorem 1 *Let $V_\lambda^i : \Delta(\mathcal{S}^i) \to \mathbb{R}$ be the unique fixed point of equation*

$$V_\lambda^i(\pi) = \max_{a\in\{0,1\}} Q_\lambda^i(\pi, a) \tag{10}$$

where

$$Q_\lambda^i(\pi,0) = \beta V_\lambda^i(\pi \cdot P^i)$$
$$Q_\lambda^i(\pi,1) = \bar{\rho}_\lambda^i(\pi) - \lambda + \beta \sum_{s\in\mathcal{S}^i} \pi(s) V_\lambda^i(\delta_s).$$

Let $\bar{f}_\lambda^i(\pi)$ denote the $\arg\max$ *of the right hand side of* (10) *where we set $\bar{f}_\lambda^i(\pi) = 1$ if $Q_\lambda^i(\pi,0) = Q_\lambda^i(\pi,1)$. Then, the time-homogeneous policy $\bar{f}_\lambda^i$ is optimal for Problem 2.* □

PROOF The proof follows immediately from Markov decision theory. ∎

B. Indexability and Whittle index

To define indexability of a process, we require the notion of *passive set*. The passive set $\mathcal{P}_\lambda^i$ is the set of states where being passive is optimal for process i, i.e.,

$$\mathcal{P}_\lambda^i = \left\{\pi \in \Delta(\mathcal{S}^i) : \bar{f}_\lambda^i(\pi) = 0\right\}.$$

Definition 1 (Indexability) For any $\lambda_1, \lambda_2 \in \mathbb{R}$ if $\mathcal{P}_\lambda^i$ is weakly increasing in λ, i.e.,

$$\lambda_1 \leq \lambda_2 \implies \mathcal{P}_{\lambda_1}^i \subseteq \mathcal{P}_{\lambda_2}^i,$$

then process i is indexable. □

Definition 2 (Whittle index) The Whittle index of belief state π of arm i is defined as the smallest value of λ for which π is not part of the passive set $\mathcal{P}_\lambda^i$, i.e.,

$$w^i(\pi) = \inf\left\{\lambda \in \mathbb{R} : \pi \notin \mathcal{P}_\lambda^i\right\}. \qquad \square$$

Equivalently, the Whittle index $w^i(\pi)$ is the smallest value of λ for which the optimal policy is indifferent between activating channel i or set it as passive when the belief state of the channel is π. The restless bandit problem is indexable if all channels are indexable.

978-1-7281-0954-1/19/$31.00 ©2019 IEEE

C. Information states

The dynamic programming of Problem 2 has a continuous state space, which makes it difficult to solve. In this section, we introduce a new information state which is countable and at the same time, equivalent to the belief state. We then use standard results from countable state MDPs to develop an approximate dynamic program for Problem 2.

Let $O_t^i \in \mathcal{S}^i$ denote the last observed state of channel i and $K_t^i \in \mathbb{N}$ denote the time since the last observation. Then, we have

$$(O_{t+1}^i, K_{t+1}^i) = \begin{cases} (S_t^i, 0) & \text{if } A_t^i = 1 \\ (O_t^i, K_t^i + 1) & \text{if } A_t^i = 0. \end{cases} \tag{11}$$

Suppose the initial state π_0^i is of the form $\delta_{o_0} \cdot (P^i)^{k_0}$ for some $O_0 \in \mathcal{S}^i$ and $k_0 \in \mathbb{Z}_{\geq 0}$. Then, we have the following.

Lemma 1 *At any time t, $\Pi_t^i = \delta_{O_t^i} \cdot (P^i)^{K_t^i}$ almost surely.* □

PROOF This result can be proved by induction.

- *Basis of induction*: The initial state π_0^i is of the form $\delta_{o_0} \cdot (P^i)^{k_0}$ for some $o_0 \in \mathcal{S}^i$ and $k_0 \in \mathbb{Z}_{\geq 0}$.
- *Induction step*: Now assume that the realization π_{t-1}^i is of the form $\delta_{o_{t-1}^i} \cdot (P^i)^{k_{t-1}^i}$. Comparing dynamics (6) and (11), we get that π_t^i is $\delta_{o_t^i} \cdot (P^i)^{k_t^i}$. ■

By a slight abuse of notation, define $\bar{\rho}^i(o,k) = \bar{\rho}^i(\delta_o \cdot (P^i)^k)$. Then, Theorem 1 is equivalent to the following theorem.

Theorem 2 *Let $W_\lambda^i : \mathcal{S}^i \times \mathbb{Z}_{\geq 0} \to \mathbb{R}$ be the unique fixed point of the following equation*

$$W_\lambda^i(o,k) = \max\{\beta W_\lambda^i(o,k+1), \\ \bar{\rho}^i(o,k) - \lambda + \beta \sum_{s \in \mathcal{S}^i} (P^i)_{os}^k W_\lambda^i(s,1)\}. \tag{12}$$

Let $\hat{f}_\lambda^i(o,k)$ denote the $\arg\max$ *of the right hand side of* (12) *where we set $\hat{f}_\lambda^i(o,k) = 1$ if the two argument inside $\max\{\cdot,\cdot\}$ are equal. Then, the time-homogeneous policy $\hat{f}_\lambda^i$ is optimal for Problem 2.* □

PROOF The proof follows immediately from Theorem 1 and Lemma 1. ■

D. Finite State Approximation

The dynamic program method described in Theorem 2 has a countable state space. We now provide a finite state approximation of it.

Theorem 3 *Given $m \in \mathbb{N}$, let $\mathbb{N}_m := \{0, \dots, m\}$ and $W_{\lambda,m}^i : \mathcal{S}^i \times \mathbb{N}_m \to \mathbb{R}$ denote the unique fixed point of the following policy equation*

$$W_{\lambda,m}^i(o,k) = \max\{\beta W_{\lambda,m}^i(o, k+1 \wedge m), \\ \bar{\rho}^i(o,k) - \lambda + \beta \sum_{s \in \mathcal{S}^i} (P^i)_{os}^k W_{\lambda,m}^i(s,0)\}. \tag{13}$$

Let $\hat{f}_{\lambda,m}^i(o,k)$ denote the $\arg\max$ *of the right hand side of* (13) *where we set $\hat{f}_{\lambda,m}^i(o,k) = 1$ if the two argument inside $\max\{\cdot,\cdot\}$ are equal. Then, we have the following:*
(i) Let $\rho_{\max}^{i,\lambda} = \max_{o,k} \rho^i(o,k) - \lambda$, then

$$\|W_\lambda^i(o,0) - W_{\lambda,m}^i(o,0)\|_\infty \leq \frac{\beta^{m+1} \rho_{\max}^{i,\lambda}}{1-\beta}, \forall o \in \mathcal{S}^i.$$

(ii) $\lim_{m\to\infty} W_{\lambda,m}^i(o,k) = W_\lambda^i(o,k)$, $\forall (o,k) \in \mathcal{S}^i \times \mathbb{Z}_{\geq 0}$.
(iii) Let $\hat{f}_\lambda^{i,*}(\cdot,\cdot)$ be any fixed point of $\{\hat{f}_{\lambda,m}^i(\cdot,\cdot)\}_{m\geq 1}$. Then, the policy $\hat{f}_\lambda^{i,*}(\cdot,\cdot)$ is optimal for Problem 2. □

PROOF (i): The payoff obtained by the approximate optimal policy starting from $(o,0)$, $\forall o \in \mathcal{S}^i$ would be the same as the optimal for times $\{0,\dots,m\}$ and after that the per-step pay-off would differ at most $\rho_{\max}^{i,\lambda}$. Thus, the bound holds.
(ii) & (iii): The sequence of finite-state models described above is an *augmentation type approximation sequence* (see [19, Definition 2.5.3]). As a result, a limit point of $\hat{f}_\lambda^{i,*}$ exists and the final result holds [19, Proposition B.5, Theorem 4.6.3].

E. Whittle index calculations

In this section, we demonstrate a set of algorithms by which the Whittle index policy is constructed. The first one is a binary search algorithm by which Whittle index for each information state of a channel is obtained. By an abuse of notation, we use $w^i(o,k)$ to denote Whittle index of information state (o,k) for channel i.

Let us assume that there are activation costs LB and UB such that $\hat{f}_{\texttt{LB}}^i(o,k) = 1$ and $\hat{f}_{\texttt{UB}}^i(o,k) = 0$ for all $(o,k) \in \mathcal{S}^i \times \mathbb{N}_m$. Given any $\lambda \in \mathbb{R}$, we define the next critical cost as

$$\Lambda_c^i(\lambda) = \inf\{w^i(o,k) : \lambda < w^i(o,k), o \in \mathcal{S}^i, k \in \mathbb{N}_m\}.$$

If no such $w^i(o,k)$ exists then we set $\Lambda_c^i(\lambda) = \infty$. An algorithm to compute the critical cost is given in Alg. 1 where the policy $\hat{f}_\lambda^i$ for a given λ is computed using the approximate dynamic programming of Theorem 3.

Algorithm 1 Critical Cost Finder: $\Lambda_c^i(\lambda)$

Input: λ, UB, ϵ
$\lambda_l \leftarrow \lambda$, $\lambda_u \leftarrow$ UB
if $\hat{f}_{\lambda_u}^i = \hat{f}_\lambda^i$ **then**
  return ∞
end if
while $|\lambda_u - \lambda_l| \geq \epsilon$ **do**
  $\lambda_c \leftarrow (\lambda_l + \lambda_u)/2$
  if $\hat{f}_{\lambda_c}^i = \hat{f}_{\lambda_l}^i$ **then**
    $\lambda_l \leftarrow \lambda_c$
  else
    $\lambda_u \leftarrow \lambda_c$
  end if
end while
return λ_u

The calculation of the Whittle index is carried out in Alg. 2. This algorithm is constructing $\{w^i(o,k) : o \in \mathcal{S}^i, k \in \mathbb{N}_m\}$ in an ascending manner while ensuring that the passive sets

978-1-7281-0954-1/19/$31.00 ©2019 IEEE

Algorithm 2 Whittle Index Calculation

Input: i, LB, UB, ϵ
$\lambda_l \leftarrow$ LB
while $\mathcal{P}_{\lambda_l} \neq \mathcal{S}^i \times \mathbb{N}_m$ **do**
  Compute $\lambda_c = \Lambda_c^i(\lambda_l)$ (using by Alg. 1)
  $\mathcal{M}_0 = \{(o,k) : \hat{f}^i_{\lambda_c}(o,k) = 1 \text{ and } \hat{f}^i_{\lambda_l}(o,k) = 0\}$
  $\mathcal{M}_1 = \{(o,k) : \hat{f}^i_{\lambda_c}(o,k) = 0 \text{ and } \hat{f}^i_{\lambda_l}(o,k) = 1\}$
  if $\mathcal{M}_1 \neq \emptyset$ **then**
    return "The problem is not indexable."
  else
    $w^i(o,k) \leftarrow \lambda_c$, for all $(o,k) \in \mathcal{M}_0$.
    $\lambda_l = \lambda_c$
  end if
end while
return $w^i(o,k)$ for all $(o,k) \in \mathcal{S}^i \times \mathbb{N}_m$.

are increasing. Fig. 2 presents an illustrative example of it. Note that if we find two values λ_1 and λ_2 in the above set such that $\lambda_1 \leq \lambda_2$ but $\mathcal{M}_1 \neq \emptyset$, that means that $\mathcal{P}_{\lambda_1} \not\subseteq \mathcal{P}_{\lambda_2}$ and therefore, the process is not indexable. If $\mathcal{M}_1 = \emptyset$, the by definition, the process is indexable and we identify the Whittle index.

As the final step, the Whittle index heuristic prescribes that *at each time, obtain the Whittle index corresponding to current information state of all channels and transmit over the L channels with the smallest Whittle index*. The algorithm is shown in Alg. 3.

Algorithm 3 Whittle Index Heuristic

Compute $w^i(o^i,k)$, $\forall k \in \mathbb{N}_m$, $\forall o^i \in \mathcal{S}^i$, and $\forall i \in \mathcal{N}$ by Alg. 2.
$t = 0$.
while $t \geq 0$ **do**
  Observe (O_t^i, K_t^i), $\forall i \in \mathcal{N}$.
  Transmit over the channels with the L smallest $w^i(O_t^i, K_t^i)$ using the resources $g^{i,*}(\delta_{O_t^i} \cdot (P^i)^{K_t^i}, 1)$.
  $t = t + 1$.
end while

V. Conclusion

We consider dynamic spectrum access for transmitting over multiple channels with partially observed channel state. We model this problem as a POMDP and identified a simplified information state. To circumvent the curse of dimensionality, we convert the POMDP to a restless bandit and use the Whittle index heuristic. There are two challenges in using the Whittle index heuristic for belief-valued processes: (i) showing that the model is indexable and (ii) computing the Whittle index. We exploit the structure of the reachable set of beliefs to convert the belief-valued process into a countable-state process and then approximate the countable-state process by a finite-state process. This approximation allows us to develop low-complexity algorithms to check whether each channel is indexable and if so, compute the Whittle index for each information state.

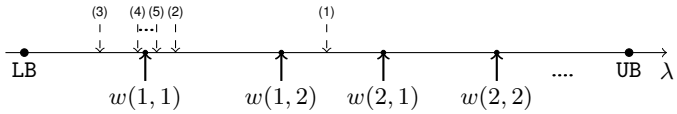

Fig. 2. The figure depicts the binary search algorithm to find the Whittle indices sequentially from left to right. The dashed arrows show the points selected by Alg. 1 sequentially labeled to find $w(1,1)$ using LB and UB. To find $w(1,2)$, Alg. 1 starts using $w(1,1)$ and UB as the bounds.

References

[1] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.
[2] A. Laourine and L. Tong, "Betting on gilbert-elliot channels," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 723–733, 2010.
[3] J. Tang, P. Mansourifard, and B. Krishnamachari, "Power allocation over two identical gilbert-elliott channels," in *International Conference on Communications (ICC)*, June 2013, pp. 5888–5892.
[4] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, 2007.
[5] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, 2009.
[6] Y. Ouyang and D. Teneketzis, "On the optimality of myopic sensing in multi-state channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 681–696, 2014.
[7] S. Guha, K. Munagala, and S. Sarkar, "Jointly optimal transmission and probing strategies for multichannel wireless systems," in *Annual Conf. Information Sciences and Systems*, 2006, pp. 955–960.
[8] A. Gopalan, C. Caramanis, and S. Shakkottai, "On wireless scheduling with partial channel-state information," in *Proc. Ann. Allerton Conf. Communication, Control and Computing*, 2007.
[9] F. Cecchi and P. Jacko, "Scheduling of users with Markovian time-varying transmission rates," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 1, 2013, pp. 129–140.
[10] U. Ayesta, M. Erausquin, and P. Jacko, "A modeling framework for optimizing the flow-level scheduling with time-varying channels," *Performance Evaluation*, vol. 67, no. 11, pp. 1014–1029, 2010.
[11] I. Taboada, F. Liberal, and P. Jacko, "An opportunistic and non-anticipating size-aware scheduling proposal for mean holding cost minimization in time-varying channels," *Performance Evaluation*, vol. 79, pp. 90–103, 2014.
[12] S. Aalto, P. Lassila, and P. Osti, "Whittle index approach to size-aware scheduling for time-varying channels with multiple states," *Queueing Systems*, vol. 83, no. 3-4, pp. 195–225, 2016.
[13] I. Taboada, P. Jacko, U. Ayestaa, and F. Liberal, "Opportunistic scheduling of flows with general size distribution in wireless time-varying channels," in *IEEE International Teletraffic Congress*, 2014, pp. 1–9.
[14] R. Meshram, D. Manjunath, and A. Gopalan, "On the Whittle index for restless multiarmed hidden Markov bandits," *IEEE Trans. Autom. Control*, vol. 63, no. 9, pp. 3046–3053, 2018.
[15] S. H. A. Ahmad and M. Liu, "Multi-channel opportunistic access: A case of restless bandits with multiple plays," in *Annual Allerton Conf. Comm., Control, and Computing*. IEEE, 2009, pp. 1361–1368.
[16] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.
[17] K. J. Astrom, "Optimal control of Markov processes with incomplete state information," *Journal of mathematical analysis and applications*, vol. 10, no. 1, pp. 174–205, 1965.
[18] A. Mahajan and D. Teneketzis, "Multi-armed bandits," in *Foundations and Applications of Sensor Management*. Springer-Verlag, 2008, pp. 121–151.
[19] L. I. Sennott, *Stochastic dynamic programming and the control of queueing systems*. John Wiley & Sons, 2009, vol. 504.

978-1-7281-0954-1/19/$31.00 ©2019 IEEE