Opportunistic Spectrum Access Based on a Constrained Multi-Armed Bandit Formulation

Jing Ai and Alhussein A. Abouzeid

Abstract: **Tracking and exploiting instantaneous spectrum opportunities are fundamental challenges in opportunistic spectrum access (OSA) in presence of the bursty traffic of primary users and the limited spectrum sensing capability of secondary users. In order to take advantage of the history of spectrum sensing and access decisions, a sequential decision framework is widely used to design optimal policies. However, many existing schemes, based on a partially observed Markov decision process (POMDP) framework, reveal that optimal policies are non-stationary in nature which renders them difficult to calculate and implement. Therefore, this work pursues stationary OSA policies, which are thereby efficient yet low-complexity, while still incorporating many practical factors, such as spectrum sensing errors and a priori unknown statistical spectrum knowledge. First, with an approximation on channel evolution, OSA is formulated in a multi-armed bandit (MAB) framework. As a result, the optimal policy is specified by the well-known Gittins index rule, where the channel with the largest Gittins index is always selected. Then, closed-form formulas are derived for the Gittins indices with tunable approximation, and the design of a reinforcement learning algorithm is presented for calculating the Gittins indices, depending on whether the Markovian channel parameters are available a priori or not. Finally, the superiority of the scheme is presented via extensive experiments compared to other existing schemes in terms of the quality of policies and optimality.**

Index Terms: **Multi-armed bandit (MAB) problem, opportunistic spectrum access (OSA), partially observed Markov decision process (POMDP), reinforcement learning (RL).**

I. INTRODUCTION

This paper addresses opportunistic spectrum access (OSA), which is one particular promising class of dynamic spectrum access (DSA)[1] models [1] where primary users (licensees) do not have to alter their hardware or behavior, whereas secondary users (non-licensees) equipped with cognitive radios [2] search, identify and exploit instantaneous spectrum opportunities whenever and wherever primary users are not present. However, because of the time-varying nature of the channel/spectrum[2] statistics, as well as inaccuracies in detecting/sensing the spectrum, this problem poses fundamental challenges in finding optimal yet low complexity policies for channel access.

This problem can be cast in a *sequential* decision framework, since it allows to use statistical knowledge and past history to benefit future decisions. The main challenges are to (a) find an *accurate model* of the problem, (b) find an *optimal* solution for *the chosen model* that has reasonable complexity, and (c) evaluate the *quality* of the solution in realistic settings. The last point highlights the interplay between the first two challenges, and that finding an optimal solution to a model does not necessarily result in a good solution unless the model itself is reasonably accurate depiction of realistic conditions.

In this paper, this problem is formulated as a *constrained partially observable multi-arm bandit problem (C-POMAB)*. By exploiting its rich structure, we derive optimal low-complexity solution (in the form of closed form Gittins indices with tunable approximation). It is shown that (and when) the problem assumptions/approximations are accurate enough, as measured from simulations under realistic conditions. Notice that while the solution has low-complexity, C-POMAB still allows modeling Markovian (which is more realistic than i.i.d.) dynamics of primary network traffic, imperfect spectrum sensors and slowly varying statistics of primary network traffic. The key modeling assumption is regarding the "information state." The information state represents the conditional *probability* of a channel in a certain state based on the observation history so far. Our model assumes the information states of all channels that are not selected to be "frozen." This is a much less stringent assumption than assuming the actual channel states to be frozen [3] – it only assumes that the knowledge of the channel states (represented by the "information states") remains unchanged as long as the these channels are not sensed, which is intuitively appealing, and we test its validity by simulations. While we assume that the channel statistics are known, low complexity model-free algorithms are designed for calculating those Gittins indices regardless of whether the statistics of channel availability are known a priori or not.

For the purpose of appreciating the contribution of this work, we first very briefly (and loosely) introduce (for readability/completeness) the well known *Markov decision process* (MDP), *partially observable* MDP (POMDP) [4], *multi-arm bandit* (MAB) [5], [6], and *restless* MAB (RMAB) [7] problems and solutions. MDP refers to a sequential decision process in which the states are fully observable. If the states of the system are not observable, the problem is significantly more difficult to solve, and is called a POMDP. A *classic MAB problem* is a particular sequential decision processes which considers a decision maker and a set of n Markov reward processes. At each decision epoch, the decision maker selects a process (also called "arm"

Manuscript received October 1, 2008.

This material is based in part upon work supported by the National Science Foundation under grant number CNS-0546402.

J. Ai was with the Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180. He is now with Juniper Networks Inc., Sunnyvale, CA, 94089, email: jingai@juniper.net.

A. A. Abouzeid is with the Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, email: abouzeid@ecse.rpi.edu.

[1]The classic dedicated spectrum allocation strategies has resulted in *artificial* spectrum scarcity [8]. DSA offers a solution to this problem.

[2]Notice that we will use the "spectrum" and "channel" interchangeably in the rest of the paper.

1229-2370/09/$10.00 © 2009 KICS

in the literature) to use in the current period based on the states, transition probabilities and rewards of processes. The goal of the decision maker is to derive an optimal process selection policy to maximize some function of the expected total reward over the planning horizon. A classic MAB problem (a) admits no constraints on the optimization function and (b) does not allow the state evolution for the arms that have not been selected. The solution of the classic MAB relies on computing indices for all processes and selecting the one(s) with the highest index. This is called an *indexable* solution. In case the states are not directly observable, we call it a classic POMAB in this paper. It is well known that the solution of a classic POMAB problem lies in introducing a new state, called information state, that transforms the problem to a set of classic MAB problems, with indexable solutions. A restless MAB (RMAB) is similar to MAB except that the arms are allowed to evolve (i.e., removing the second constraint).

In this paper, the key new elements of C-POMAB problem/solution compared to the preceding well known problems are that (a) the optimization problem allows a constraint, and (b) it is shown that, by assuming the information state of non-selected arms to be "frozen," the solution is indexable, i.e., can be solved in a similar manner to classic POMAB. Allowing a constraint is extremely important, since we use this to *add a constraint on the total interference incurred on primary users*. That the optimal solution remains to be indexable is critical for finding low-complexity algorithm. In fact, we derive closed form expressions for these indices with tunable approximation.

The relation between the C-POMAB problem/solution and other related efforts in the literature that utilize a sequential decision framework can be summarized as follows. Optimal policies for a POMDP formulation of the problem is presented in [9] and [10] assuming the availability of each channel follows a Markov chain and their transition matrices are known to secondary users. However, the derivation of optimal policies suffers *exponential* computational complexity with respect to both the horizon length and the number of channels. Optimal polices for a MAB formulation is presented in [11] without assuming prior knowledge of primary traffic statistics. However, it assumes that the availability of each channel is independent from slot to slot, which is a strong assumption at odds with some practical measurements in [12]. Moreover, because of the independence assumption, it seems unlikely that it can be directly extended to the case where spectrum opportunities follow a Markov chain model[3]. A MAB formulation is also used in [1] but under a different channel access (open sharing) architecture. Notice that both of these MAB formulations [1], [11] do not take any constraints (unlike our C-POMAB), and thus are not able to consider or even explicitly quantify the amount of *interference* to primary networks induced by imperfect spectrum sensing[4]. Djonin *et al.* [13] propose a truncated MDP formulation with a *linear* complexity with respect to the horizon length, which is able to achieve a tunable trade-off between quality of policies and complexity depending on a chosen truncation parameter. However, due to the nature of dynamic programming formulation, it still incurs exponential computational cost with respect to the number of channels. Zhao *et al.* [15] propose a minimum time to availability (MTTA) heuristic as a myopic policy. Finally, most recently, Liu and Zhao [16] use an RMAB formulation and applied Whittle's index rule [7], which is shown to be *near-optimal* when channels are heterogenous and *optimal* when all channels are homogeneous. However, [16] does not take account of spectrum sensing errors, incurring interference to primary networks.

To summarize, the contribution of this paper is four-fold. First, with sound approximations, we formulate OSA as a C-POMAB problem, while explicitly considering interference constrains to primary networks induced by spectrum sensing errors. We not only show that its optimal policy is still indexable, but also extend the *separation principle* between spectrum sensing and access schemes in an infinite horizon setting, as compared to the main result in [9] in a finite-horizon setting. Second, assuming Markovian traffic dynamics with given transition probabilities, we are able to derive closed-form expressions of Gittins indices with tunable approximation, which indicates very low computational cost of computing the optimal policy. Particularly, when primary traffic statistics are identical, OSA policy would degenerate into a *myopic* policy in terms of a counting process (i.e., the number of spectrum access failures since the last success for a given channel), which is very *close* to the main result in [15] developed from a POMDP framework. Third, we design an online model-free learning (i.e., Q-learning) algorithm to calculate Gittins indices, which enables the scheme to adapt to slowly statistical varying (i.e., non-stationary) radio environments. Lastly, we demonstrate the superiority of the proposed OSA policy over other existing ones in terms of achieving a better trade-off between quality of policies and complexity via extensive simulations.

The rest of paper is organized as follows. Section II presents the system model. Section III formulates the C-POMAB problem. Section IV and V illustrate the Gittins indices calculation in a model-based and a model-free setting, respectively. Simulation results are provided in Section VI. Finally, Section VII concludes the paper and highlights a few future directions.

II. SYSTEM MODEL

We consider a pair of transmitter and receiver to sense and access a single channel at a time among n channels. Notice that the set of channels is denoted as $\mathcal{N} = \{1, 2, \cdots, n\}$ and the bandwidth of each channel i is denoted as w_i. Without loss of generality, we assume that the channel availability statistics among channels are mutually independent and dynamics of each channel can be modeled as a two-state discrete-time Markov chain with a transition matrix $P^i = [p^i_{jm}]_{j,m\in\mathcal{S}},\ i \in \mathcal{N}$ as shown in Fig. 1, where the state space is denoted as $\mathcal{S} = \{0\ (IDLE/GOOD), 1\ (BUSY/BAD)\}$.

Next, we assume that primary and secondary users access the spectrum aligning with time slots. The interval of a time slot is denoted as T_s and thus the time slot k represents the time interval $I_k = [kT_s, (k+1)T_s)$, where $k \in \{0, 1, \cdots\}$. In addition, we assume a saturated secondary transmitter that always has packets to send.

[3]Notice that a (semi-)Markov model more accurately characterizes the primary network traffic [9], [10], [14].

[4]See [9] for more information about imperfect spectrum sensing.

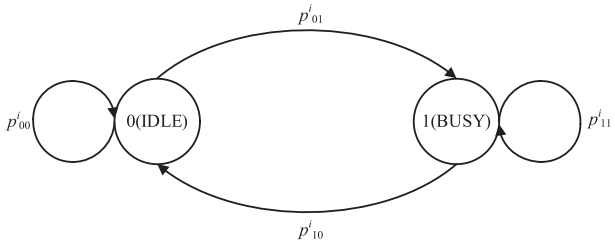

Fig. 1. The Markov model for channel i ($i \in \mathcal{N}$).

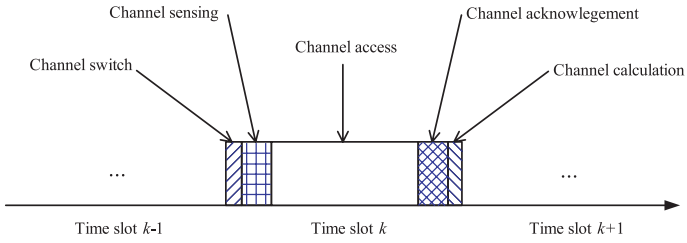

Fig. 2. Operations of OSA during a time slot.

As always, the secondary transmitter and receiver need to be firstly synchronized on a common channel to start communication and the handshake mechanism has been specified in [10]. In a typical time slot as illustrated in Fig. 2, the secondary transmitter starts to sense the selected channel at the beginning. Based on the sensing outcome, the secondary transmitter decides whether to send a packet or not. If selecting to send, it expects to receive an acknowledgement from the intended secondary receiver. Thus, by the end of a time slot, both secondary transmitter and receiver obtain the same knowledge on their communication. With the same set of decision rules (presented later in this paper), they compute the next common channel to sense and prepare to synchronize to it at the beginning of the following time slot.

Moreover, we do *not* neglect spectrum sensing errors. In OSA, it is known that secondary users need to sense the channel before access so as to achieve non-intrusive communication to primary users. However, with unreliable sensing outcomes induced by noise and fading, it is impossible to always observe the actual activities of primary users and make the *right* access decisions. Generally, there are two types of sensing errors: (a) *False alarms* occur when idle channels are detected as busy, quantified by $\epsilon \triangleq Pr\{o = 1|s = 0\}$, wasting the spectrum opportunity and, (b) *miss detections* occur when busy channels are detected as idle, quantified by $\delta \triangleq Pr\{o = 0|s = 1\}$, causing interference to primary users, where $o \in \mathcal{S}$ denotes the sensing outcome[5].

As a result, in presence of spectrum sensing uncertainty, a randomized spectrum access scheme, $\vec{q}$, is employed per slot without loss of generality. Specifically, the randomized spectrum access scheme, $\vec{q}$, can be specified by a 2-tuple (q^f, q^b) and q^f and q^b are defined as $q^f \triangleq Pr\{t = 1|o = 0\}$ and $q^b \triangleq Pr\{t = 1|o = 1\}$, where $t \in \{0\ (NO\ ACCESS), 1\ (ACCESS)\}$. Therefore, when the actual state of the sensed channel s is idle, the probability of receiving an acknowledgement, $e \in \{0\ (ACK), 1\ (NO\ ACK)\}$, under the randomized scheme $\vec{q}$ can be obtained as

$$\begin{aligned}\lambda &\triangleq Pr\{e = 0|s = 0, \vec{q}\}\\ &= Pr\{t = 1, o = 0|s = 0\} + Pr\{t = 1, o = 1|s = 0\}\\ &= Pr\{t = 1|o = 0, s = 0\}Pr\{o = 0|s = 0\}\\ &\quad + Pr\{t = 1|o = 1, s = 0\}Pr\{o = 1|s = 0\}\\ &= Pr\{t = 1|o = 0\}Pr\{o = 0|s = 0\}\\ &\quad + Pr\{t = 1|o = 1\}Pr\{o = 1|s = 0\}\\ &= q^f(1 - \epsilon) + q^b\epsilon\end{aligned} \tag{1}$$

which accounts for the only scenario that secondary users can gain throughput. Notice that we assume that, whenever accessing the channel, the secondary transmitter can get an ACK back without any error as long as the channel is truly idle in (1).

On the other hand, when the actual state of the sensed channel s is busy, it by no means that a packet can get through between secondary users, i.e., $Pr\{e = 1|s = 1, \vec{q}\} = 1$. Meanwhile, it may incur interference to primary users if the secondary transmitter decides to access the channel. The probability of collision to primary users can be obtained as

$$\begin{aligned}\xi &\triangleq Pr\{e = 1|s = 1, \vec{q}\}\\ &= Pr\{t = 1, o = 0|s = 1\} + Pr\{t = 1, o = 1|s = 1\}\\ &= Pr\{t = 1|o = 0\}Pr\{o = 0|s = 1\}\\ &\quad + Pr\{t = 1|o = 1\}Pr\{o = 1|s = 1\}\\ &= q^f\delta + q^b(1 - \delta).\end{aligned} \tag{2}$$

Combining (1) and (2), it is interesting to find out that false alarms by the spectrum sensor will only lower secondary users' throughput while miss detection by the spectrum sensor will only cause interference to primary users.

III. OSA IN A MULTI-ARMED BANDIT FRAMEWORK

In this section, we first formulate the problem in a MAB framework based on an assumption for approximation. Notice that, our OSA scheme becomes the solution of a constrained decision problem taking account of the trade-off between the expected discounted throughput of a pair of secondary users and the collision probability to primary users when considering sensing errors. Next, we study the solution structure of the problem and show that the optimal policy is still indexable, even with constraints, as in conventional MAB problems. As a result, spectrum sensing scheme and spectrum access scheme in our OSA can be derived separately.

A. MAB Formulation

We first define information states [3] of channels. They represent probability distributions on the "perceived" states based on past observations and actions. Given the system model presented in Section II, we specify that the action is composed by spectrum sensing action, $a \in \mathcal{N}$, as well as spectrum access policy, $\vec{q}$, performed at the beginning of the slot, and the observation, e, is the acknowledgement received at the end of the slot. Therefore, for any channel i, denoting the observation history at time slot $k-1$ as $\mathbf{e}_{k-1} = \{e_0, e_1, \cdots, e_{k-1}\}$

[5] In this paper, we regard ϵ and δ as model parameters and do not tune them as a part of joint PHY-MAC design as in [9].

and the action history at time slot k as $\mathbf{a}_{k-1} = \{a_0, \cdots, a_{k-1}\}$ and $\vec{\mathbf{q}}_{k-1} = \{\vec{q}_0, \cdots, \vec{q}_{k-1}\}$, its information state at time k is $\vec{x}_k^i = (x_k^i(0), x_k^i(1))$, and $x_k^i(s)$, $s \in \mathcal{S}$, is given by

$$\begin{aligned} x_k^i(s) &\triangleq Pr\{s_k^i = s | \mathbf{a}_{k-1}, \vec{\mathbf{q}}_{k-1}, \mathbf{e}_{k-1}\} \\ &= Pr\{s_k^i = s | a_{k-1}, \vec{q}_{k-1}, e_{k-1}\},\ s \in \mathcal{S},\ i \in \mathcal{N} \end{aligned} \tag{3}$$

where the equality follows by Markov property.

Next, for those channels ($\forall i \neq a_k$) that are not selected at time slot k, the evolution of information states is purely driven by the underlying channel dynamics, i.e., $\vec{x}_{k+1}^i = P^{i\prime}\vec{x}_k^i$[6]. However, it would lead to a restless bandit formulation [7] for our OSA problem, which is PSPACE-hard. Therefore, in order to derive tractable policies, we adopt a classic restless bandit approximation where we formulate our OSA problem in a multi-armed bandit framework. As a result, we assume that

$$\vec{x}_{k+1}^i = \vec{x}_k^i,\ \forall i \neq a_k \tag{4}$$

i.e., the *information* states of unselected bandit processes are frozen at the current time slot.

On the other hand, given channel i selected for sense and access at time slot k (i.e., $a_k = i$), the transition law for its information states can be derived as follows. We define a conditional probability, $u_k^i(j, m)$, that, at time slot k, the acknowledgement e_k is m, conditioned on the actual channel state s_k is j together with a randomized spectrum access action $\vec{q}_k$ as

$$u_k^i(j, m) \triangleq Pr\{e_k = m | s_k = j, \vec{q}_k\},\ j, m \in \mathcal{S}. \tag{5}$$

Similar to (1), we can obtain those probabilities. Accordingly, summarized in a matrix form, $U_k^i = [u_k^i(j, m)]_{j,m\in\mathcal{S}}$ can be written as

$$U_k^i = \begin{bmatrix} \lambda_k & 1 - \lambda_k \\ 0 & 1 \end{bmatrix}. \tag{6}$$

Consequently, given P^i and U_k^i, the information state of channel i can be updated by Bayes' rule as follows.

$$\vec{x}_{k+1}^i = \mathcal{F}(\vec{x}_k^i, \vec{q}_k, m) = \frac{U_k^i(m)P^{i\prime}\vec{x}_k^i}{\mathbf{1}_{|\mathcal{S}|}{}'U_k^i(m)\vec{x}_k^i},\ m \in \mathcal{S} \tag{7}$$

where $U_k^i(m) = diag[u_k(0, m), u_k(1, m)]$.

Since $x^i(0)$ and $x^i(1)$ are complementary, information state transition law by (7) regarding $x^i(0)$ can be simplified as

$$x_{k+1}^i(0) = \begin{cases} p_{00}^i, & \text{if } m = 0, \\ f(x^i(0)), & \text{if } m = 1, \end{cases} \quad j \in \mathcal{S} \tag{8}$$

where

$$f(x) = \frac{p_{00}(1 - \lambda)x + p_{10}(1 - x)}{1 - \lambda x} \tag{9}$$

and p_{00}, p_{10}, and λ are parameters. We can observe that information state $x_{k+1}^i(0) = p_{00}$, depending only on parameters of the i-th channel, when able to receive an ACK.

Next, referring to (1), we define the immediate expected reward achieved during the time slot k as

$$E\{R(\vec{\mathbf{x}}_k, a_k = i, \vec{q}_k)\} = \mathbf{1}_{|\mathcal{S}|}{}'U_k^i(0)\vec{x}_k^i w_i = \lambda_k x_k^i(0) w_i \tag{10}$$

[6]Notice that $'$ denotes transpose throughout the paper.

where $\vec{\mathbf{x}}_k = \{\vec{x}_k^1, \vec{x}_k^2, \cdots, \vec{x}_k^n\}$. At the same time, as a side effect of imperfect spectrum sensors, the corresponding interference probability to primary users can be quantified as by (2).

Therefore, combining all components above, we can formally state our goal as to derive an optimal policy, with respect to our assumption on "frozen" information states, $\pi : \vec{\mathbf{x}}_k \rightarrow (a_k, \vec{q}_k)$, satisfying

$$\max J_\pi = E\{\sum_{k=0}^{\infty} \gamma^k R(\vec{\mathbf{x}}_k, a_k, \vec{q}_k)\} \tag{11}$$

subject to

$$\xi \leq \xi_0,\ k \in \mathcal{T} \tag{12}$$

where γ is a discounted factor in $(0, 1)$ and ξ_0 is the maximum probability tolerated by primary users for any channel.

B. MAB Results

The following theorem states that the optimal policy of the above problem remains to be indexable, though it is different from conventional [5] MAB formulations.

Theorem 1: The optimal policy of the C-POMAB problem defined in (11) and (12) is still in a form of an index rule. Moreover, it leads to a special structure of optimal policy, which is composed by a *myopic* spectrum access scheme $\vec{q}$ per time slot and a *sequential* spectrum sensing scheme $\{a_k\}_{k\in\mathcal{T}}$ over an infinite horizon.

Proof: Since the maximum interference constraint ξ_o is enforced on all channels, we can assign any fixed $\vec{q} \in \{(q^f, q^b) | 0 \leq q^f \leq 1, 0 \leq q^b \leq 1, q^f\delta + q^b(1 - \delta) \leq \xi_0\}$ whenever selecting any channel without loss of generality. It is straightforward to see that our constrained MAB problem defined above is equivalent to an unconstrained MAB problem to maximize $J_\pi(\lambda)$ where $\pi : \vec{\mathbf{x}}_k \rightarrow a_k,\ k \in \mathcal{T}$ and thereby its optimal policy is in a form of an index rule [6].

Next, given that a series of conventional MAB problems parameterized on λ can be constructed for the same OSA problem, we proceed to show that $J_{\pi^*}(\lambda)$ monotonically increases with respect to λ, i.e., for any $\lambda_1 < \lambda_2$, $J_{\pi^*}(\lambda) < J_{\pi^*}(\lambda)$.

First, it is well known that a MAB problem can be decomposed into multiple independent optimal stopping problems (OSPs). As a result, regarding any channel i, its value functions with respect to the information state $\vec{x}^i$ with spectrum access policy $\vec{q}\,(\lambda)$ must satisfy the following Bellman equations

$$V(\vec{x}^i) = \max\left\{ \underbrace{\lambda w^i x^i(0)}_{\text{"stop"}}, \underbrace{\gamma \sum_{m\in\{0,1\}} V(\mathcal{F}(\vec{x}^i, \vec{q}, m))\mathbf{1}_{|\mathcal{S}|}{}'U^i(m)\vec{x}^i}_{\text{"continue"}} \right\}. \tag{13}$$

In (13), $\lambda w^i x^i(0)$ represents the immediate reward after successfully accessing the spectrum. Moreover, the secondary users do not pursue the same channel in the next time slot, which corresponds to the "stop" action in OSP. On the other hand, $\gamma \sum_{m\in\{0,1\}} V(\mathcal{F}(\vec{x}^i, \vec{q}, m))\mathbf{1}_{|\mathcal{S}|}{}'U^i(m)\vec{x}^i$ represents the future discounted reward if secondary users continue to pursue the same channel regardless of the result in the current slot, which corresponds to the "continue" action in OSP. Therefore,

the value function, $V(\vec{x}^i)$, represents the maximum expected reward when performing $\vec{q}$, starting from the information state $\vec{x}^i$. Consequently, to prove $J_{\pi^*}(\lambda)$'s monotonic property with respect to λ is attributed to demonstrate the monotonic property of $V(\vec{x}^i)$, $\forall \vec{x}^i$, with respect to λ. Therefore, for any $\vec{x}^i$, we compare $V(\vec{x}^i)$ with λ_1 and λ_2 such that $\lambda_1 < \lambda_2$.

As to the immediate reward, we have

$$\lambda_1 w^i x^i(0) - \lambda_2 w^i x^i(0) = (\lambda_1 - \lambda_2) w^i x^i(0) < 0. \tag{14}$$

It is easy to see that, when $\lambda_1 < \lambda_2$, we have $\lambda_1 w^i x^i(0) < \lambda_2 w^i x^i(0)$. Meanwhile, as to the future discounted reward, we have[7]

$$\begin{aligned} &\gamma \sum_{m\in\{0,1\}} V(\mathcal{F}(\vec{x}_k^i, \vec{q}^1, m)) \mathbf{1}_{|\mathcal{S}|}{}' U^i(m) \vec{x}^i \\ &- \gamma \sum_{m\in\{0,1\}} V(\mathcal{F}(\vec{x}_k^i, \vec{q}^2, m)) \mathbf{1}_{|\mathcal{S}|}{}' U^i(m) \vec{x}^i \\ &= \gamma [V(p_{00})\lambda_1 x^i(0) - V(p_{00})\lambda_2 x^i(0) \\ &\quad + V(f_{\lambda_1}(x^i(0)))(1 - \lambda_1 x^i(0)) \\ &\quad - V(f_{\lambda_2}(x^i(0)))(1 - \lambda_2 x^i(0))]. \end{aligned} \tag{15}$$

For further comparison, we need the following result that

$$\begin{aligned} V(\mathcal{F}(\vec{x}^i, \vec{q}^1, m=1)) \le \tau V(\mathcal{F}(\vec{x}^i, \vec{q}^1, m=0)) \\ + (1-\tau) V(\mathcal{F}(\vec{x}^i, \vec{q}^2, m=1)) \end{aligned} \tag{16}$$

where $\tau = (\lambda_2 - \lambda_1)x^i(0)/(1 - \lambda_1 x^i(0)) > 0$. This is because of the convexity of the value function $v(\cdot)$ due to $\max\{\cdot\}$ together with the equality below derived by some algebra work

$$\begin{aligned} \mathcal{F}(\vec{x}_k^i, \vec{q}^1, m=1) = \tau \mathcal{F}(\vec{x}_k^i, \vec{q}^1, m=0) \\ + (1-\tau)\mathcal{F}(\vec{x}_k^i, \vec{q}^2, m=1). \end{aligned} \tag{17}$$

By plugging (16) into (15), it becomes

$$\begin{aligned} &V(p_{00})\lambda_1 x^i(0) - V(p_{00})\lambda_2 x^i(0) + V(f_{\lambda_1}(x^i(0))) \\ &\quad \times (1 - \lambda_1 x^i(0)) - V(f_{\lambda_2}(x^i(0)))(1 - \lambda_2 x^i(0)) \\ &\le V(p_{00}) x^i(0)(\lambda_1 - \lambda_2 + \lambda_2 - \lambda_1) + V(f_{\lambda_2}(x^i(0))) \\ &\quad \times (1 - \lambda_1 x^i(0) - (\lambda_2 - \lambda_1)x^i(0) - 1 + \lambda_2 x^i(0)) \\ &= 0. \end{aligned} \tag{18}$$

Combining (14) and (18), the monotonic property of the value function, $V(x^i(0))$, with respect to the parameter λ holds. Therefore, in order to maximize $J_{\pi^*}(\lambda)$, we also need to maximize λ on top of the conventional MAB problem characterized by the index rule. □

Theorem 1 reveals that the optimal policies for spectrum sensing and spectrum access are orthogonal. In order to derive optimal policy as a whole, we need to derive optimal policy for each scheme, respectively, and then conduct them for each phase as shown in Fig. 2 for every time slot. Notice that, we obtain the *separation principle* from the above theorem under a MAB framework over an *infinite-horizon* as compared to the one in [9] independently developed from a POMDP framework over a finite horizon.

[7] $V(\vec{x}^i)$ and $V(x^i(0))$ are used interchangeable as below.

As a result, the corresponding optimal spectrum access scheme can be specified by a policy $\{\vec{q}_k\}$ that satisfies

$$\max \lambda \tag{19}$$

subject to

$$\xi \le \xi_0 \tag{20}$$

which can be derived by the following corollary.

Corollary 1: Given the characteristics of spectrum sensor (ϵ, δ) and the maximum interference constraint ξ_0 to primary networks, the optimal spectrum access scheme (q^{f^*}, q^{b^*}) is the solution of the linear program in (19) and (20), i.e.,

$$(q^{f^*}, q^{b^*}) = \begin{cases} (1, \frac{\xi_0 - \delta}{1-\delta}), & \delta \le \xi_0, \\ (\frac{\xi_0}{\delta}, 0), & \delta > \xi_0 \end{cases} \tag{21}$$

and thereby

$$\lambda^* = \begin{cases} 1 - \epsilon + \frac{\xi_0 - \delta}{1-\delta}\epsilon, & \delta \le \xi_0, \\ \frac{\xi_0}{\delta}(1-\epsilon), & \delta > \xi_0. \end{cases} \tag{22}$$

Proof: Firstly, from (20), the constraint of the linear program, we have

$$0 \le q^f \le \frac{\xi_0 - q^b(1-\delta)}{\delta}. \tag{23}$$

When $(\xi_0 - q^b(1-\delta))/\delta \ge 1$ (i.e., $\xi_0 \ge \delta$), the above constraint is relaxed for q^f and we have $q^{f^*} = 1$. Accordingly,

$$0 \le q^b \le \frac{\xi_0 - \delta}{1-\delta}. \tag{24}$$

Therefore, we have $q^{b^*} = (\xi_0 - \delta)/(1-\delta)$.

On the other hand, when $\xi < \delta$ and $0 \le \frac{\xi_0 - q^b(1-\delta)}{\delta} < 1$, we have

$$0 \le q^b \le \frac{\xi_0}{1-\delta}. \tag{25}$$

Then the objective function satisfies

$$\lambda \le \frac{\xi_0}{\delta}(1-\epsilon) + \left[\epsilon - (1-\epsilon)\frac{1-\delta}{\delta}\right] q^b. \tag{26}$$

Since $\epsilon/(1-\epsilon) < (1-\delta)/\delta$ usually holds for a reasonable spectrum sensor where ϵ and δ are small, λ is a decreasing function with respective to q^b in (25). Therefore, we have $q^{b^*} = 0$ and $q^{f^*} = \xi_0/\delta$ from (23).

Combining all the results above, (21) and (22) follow. □

From the above corollary, it is interesting to observe that, when $\delta = \xi_0$, the randomized spectrum access scheme will become a deterministic one where the secondary transmitter always trusts the spectrum sensing outcomes when making spectrum access decisions.

On the other hand, as to the optimal spectrum sensing scheme, given the optimal spectrum access is derived from Corollary 1, it is determined by the well-known Gittins index rule in a MAB framework as follows.

$$i_k^* = \arg\max_{i\in\mathcal{N}} \{v^i(\vec{x}_k^i)\} \tag{27}$$

where $v^i(\vec{x}_k^i)$ denotes the Gittins index of channel i given the information state $\vec{x}_k^i$ at time slot k. In the next two sections, we will illustrate how to calculate Gittins indices with and without knowing the parameters of channel availability, respectively.

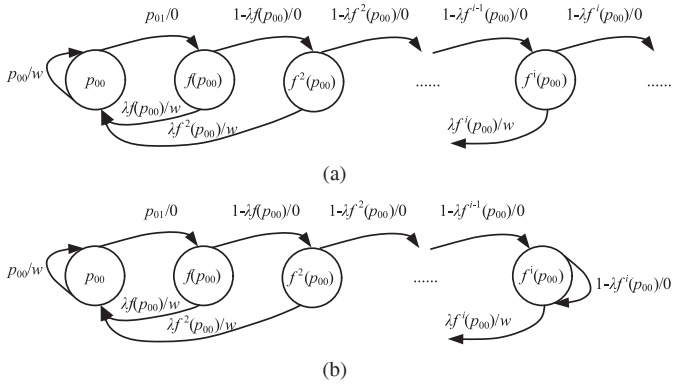

Fig. 3. Markov chain diagram of information states for a channel. Notice that numbers over lines denote transition probability/reward, respectively: (a) Original Markov chain with countably infinite states, and (b) truncated Markov chain with $I+1$ states.

IV. GITTINS INDICES CALCULATION FOR MARKOVIAN CHANNELS

In this section, we focus on Gittins indices calculation based on a Markov model for a single channel. First of all, from (8), we can observe that the information state space for any given channel can be represented as a countably infinite set $\{p_{00}, f(p_{00}), f^2(p_{00}), \cdots, f^i(p_{00}), \cdots\}$, where $f^i(\cdot) = \underbrace{f(f\cdots f(\cdot))}_{i}$[8] and thereby the information state $f^i(p_{00})$ can be achieved if there are i spectrum access failures (i.e., no ACK received) since the last success. Thereby, the underlying Markov chain can be drawn in Fig. 3(a). It shows that the function $f(\cdot)$ determines not only the evolution of information states but also the transition probabilities and rewards. Notice that when $p_{00} = p_{10}$, the channel dynamics degenerate into an i.i.d. process, which is the case that has been studied in [11] and not is the focus of our paper.

Before deriving Gittins indices, the following two theorems first illustrate the characteristics of information state space in terms of $f(\cdot)$.

Theorem 2: The series of information states $\{f^i(p_{00})\}_{i=0}^{\infty}$ converges as the iteration number i grows, i.e., there exists a fixed point $x^* \in (0,1)$ such that

$$\lim_{i\to\infty} f^i(x^*) = x^*. \tag{28}$$

When $p_{00} > p_{10}$, the series of information states $\{f^i(p_{00})\}$ decreases towards x^* as the iteration number i grows.

On the other hand, when $p_{00} < p_{10}$, the series of information states $\{f^i(p_{00})\}$ alternatingly increases and decreases towards (i.e., oscillates) x^* as the iteration number i grows.

Proof: See Appendix A. □

Notice that Theorem 2 reveals that a channel can be classified into two types depending on the order of transition probabilities (i.e., p_{00} vs. p_{10}). Specifically, (following the terminology from [17]), when $p_{00} > p_{10}$, the channel is called a *positively autocorrelated* channel, which states that the idle channel state is more likely followed by the idle channel state; while when $p_{00} < p_{10}$, the channel is called a *negatively autocorrelated* channel, which states that the idle channel state is more likely followed by the busy channel state. As we will see shortly after, the type of channel will greatly impact the Gittins indices calculation and thereby the pattern of its scheduling behavior.

Next, to ease Gittins indices calculation, we intend to use a finite number of information states as shown in Fig. 3(b) to represent the countably infinite information states series with any specified tolerance. From Theorem 2, we can expect that the actual Gittins indices of information states can be infinitely closely approximated as the increase of number of information states to be truncated. Specifically, given (ϵ, δ) of spectrum sensor together with ξ_0 as the maximum interference probability to primary users are reasonably small in typical scenarios such that λ to be employed becomes close to 1 according to (22), we can characterize the convergence rate of the information state series in the following theorem.

Theorem 3: When $\lambda \to 1$, given any ε, $0 < \varepsilon < 1$, the minimum number of information states $I+1$, such that $|f^i(x) - x^*| < \varepsilon$, $\forall i \geq I$, satisfies

Case 1: When $p_{00} > p_{10}$,

$$I = \left\lceil \frac{\log(\frac{\varepsilon}{p_{00}-p_{10}})}{\log(\frac{(p_{00}-p_{10})(1-\lambda)}{(1-\lambda p_{10})^2})} \right\rceil. \tag{29}$$

Case 2: When $p_{00} < p_{10}$,

$$I = \left\lceil \frac{\log(\frac{\varepsilon}{p_{10}-p_{00}})}{\log(\frac{(p_{10}-p_{00})(1-\lambda)}{1-\lambda p_{10}-\lambda p_{10}(p_{00}(1-\lambda)-p_{10}+1)})} \right\rceil. \tag{30}$$

Proof: See Appendix B. □

Now we can formally calculate Gittins index of a *truncated* Markov chain as follows. Notice that (31) provides a general rule of calculating the index, which can be interpreted as the maximum discounted reward rate per time slot.

$$v(x(i)) = \max_{\tau>1} E\{\frac{\sum_{k=0}^{\tau-1}\gamma^k R(x_k)|x_0 = x(i)}{\sum_{k=0}^{\tau-1}\gamma^k|x_0 = x(i)}\} \tag{31}$$

where $i \in \{0, 1, \cdots, I\}$[9]. Owing to the special state structure and Markovian process of our problem, in practical scenarios where λ is close to 1, it is possible to further simplify the index calculation by the following theorem.

Theorem 4: Given Markov chain is truncated with $I+1$ information states, Gittins index as a function of each state can be obtained by the following computation procedure.

Case 1: When $p_{00} > p_{10}$

$$v(x(i)) = \begin{cases} p_{00}w, & i = 0, \\ \frac{\lambda f^i(p_{00})\alpha_0 + \lambda f^i(p_{00})w}{\lambda f^i(p_{00})\beta_0 + 1}, & i = 1, \cdots, I \end{cases} \tag{32}$$

where α_0 and β_0 for each i can be computed as

$$\alpha_0 = \frac{\gamma p_{00}w + \sum_{j=1}^{i}\gamma^{j+1}p_{01}\lambda f^j(p_{00})w\prod_{k=1}^{j-1}(1-\lambda f^k(p_{00}))}{1-\gamma p_{00} - \sum_{j=1}^{i}\gamma^{j+1}p_{01}\lambda f^j(p_{00})\prod_{k=1}^{j-1}(1-\lambda f^k(p_{00}))} \tag{33a}$$

[8] Without loss of generality, we define that $f^0(x) = x$.

[9] From now on, we use the index of an information state i to replace the information state $x(i)$ for simplicity.

$$\beta_0=\frac{\gamma+\sum_{j=1}^{i}\gamma^{j+1}p_{01}\prod_{k=1}^{j-1}(1-\lambda f^k(p_{00}))}{1-\gamma p_{00}-\sum_{j=1}^{i}\gamma^{j+1}p_{01}\lambda f^j(p_{00})\prod_{k=1}^{j-1}(1-\lambda f^k(p_{00}))}. \tag{33b}$$

Case 2: When $p_{00} < p_{10}$ and λ is close to 1 such that $p_{00} < \lambda f(p_{00})$ and $f(p_{00}) \approx p_{10}$, given I is even

$$v(x(2i-1)) = \lambda f^{2i-1}(p_{00})w,\ i = 1,\cdots,\frac{I}{2}, \tag{34a}$$

$$v(x(2i)) = \begin{cases} \lambda f^{2i}(p_{00})w, & i = \frac{I}{2}, \\ \frac{(1-\lambda f^{2i}(p_{00}))\alpha_{2i+1}+\lambda f^{2i}(p_{00})w}{(1-\lambda f^{2i}(p_{00}))\beta_{2i+1}+1}, & i = \frac{I}{2}-1,\cdots,1, \\ \frac{p_{01}\alpha_1+p_{00}w}{p_{01}\beta_1+1}, & i = 0. \end{cases} \tag{34b}$$

Otherwise, given I is odd

$$v(x(2i-1)) = \lambda f^{2i-1}(p_{00})w,\ i = 1,\cdots,\frac{I+1}{2}, \tag{35a}$$

$$v(x(2i)) = \begin{cases} \frac{(1-\lambda f^{2i}(p_{00}))\alpha_{2i+1}+\lambda f^{2i}(p_{00})w}{(1-\lambda f^{2i}(p_{00}))\beta_{2i+1}+1}, & i = \frac{I-1}{2},\cdots,1, \\ \frac{p_{01}\alpha_1+p_{00}w}{p_{01}\beta_1+1}, & i = 0 \end{cases} \tag{35b}$$

where

$$\alpha_I = \frac{\gamma\lambda f^I(p_{00})w}{1-\gamma(1-\lambda f^I(p_{00}))} \tag{36a}$$

$$\beta_I = \frac{\gamma}{1-\gamma(1-\lambda f^I(p_{00}))} \tag{36b}$$

$$\alpha_i = \gamma(1-\lambda f^i(p_{00}))\alpha_{i+1} + \gamma\lambda f^i(p_{00})w \tag{37a}$$

$$\beta_i = \gamma(1-\lambda f^i(p_{00}))\beta_{i+1} + \gamma \tag{37b}$$

for $i = I-1,\cdots,1$, and

$$\alpha_0 = \frac{\gamma p_{01}\alpha_1 + \gamma p_{00}w}{1-\gamma p_{00}} \tag{38a}$$

$$\beta_0 = \frac{\gamma p_{01}\beta_1 + \gamma}{1-\gamma p_{00}}. \tag{38b}$$

Proof: See Appendix C. □

We gain the following insights from Theorem 4. When a channel is positively autocorrelated (i.e., $p_{00} > p_{10}$) and it is selected to sense and access, any failure will decrease its Gittins index and thus make the channel less "attractive" in the following scheduling. On the contrary, when a channel is negatively autocorrelated (i.e., $p_{00} < p_{10}$) and it is selected to sense and access, any failure will increase its Gittins index and thus make the channel more "attractive" in the following scheduling, since state 0 has the least Gittins index among all states. More importantly, our formulation from a MAB framework provides closed-forms (with tunable approximation) for deriving a stationary policy while many other works [9], [10], [13], based on a POMDP framework, require to solve a linear programming for non-stationary policies, implying a high computational and implementation cost.

It is interesting to find that, when the parameters and statistics of channel availability are identical, the spectrum sensing scheme would degenerate into a myopic policy. This is because when channels are homogeneous, comparison of Gittins indices in (27) is equivalent to comparison of indices of information states. Specifically, such a myopic policy can be stated by the following theorem.

Theorem 5: Let z^i denote the number of spectrum access failures since last success for channel i. For homogenous Markovian channels, our optimal spectrum sensing scheme can be specified as follows:

Case 1: When $p_{00} > p_{10}$, the optimal spectrum sensing scheme is to sense the channel i^* such that $i^* = \arg\min_{i\in\mathcal{N}}\{z^i\}$.

Case 2: When $p_{00} < p_{10}$, the optimal spectrum sensing scheme is to sense the channel i^* such that $i^* = \arg\max_{i\in\mathcal{N}}\{z^i\}$.

In case there exist multiple channels with the same z value, one channel will be randomly selected to break up the tie. Moreover, regardless of the above rule, in order to well exploit statistical knowledge, the myopic rule also needs to obey that: when channels are positively autocorrelated, a channel will always be selected until a spectrum access failure happens; while when channels are negatively autocorrelated, a channel will always be selected until a spectrum access success happens.

Proof: When $p_{00} > p_{10}$, Gittins index is monotonic decreasing with respect to z according to Theorem 4. Given channels are homogenous, comparison on Gittins indices of all channels is equivalent to comparison on z values. Therefore, by the optimal policy specified in (27), the result follows.

When $p_{00} < p_{10}$, state 0 has the least Gittins index among all states according to Theorem 4. Given channels are homogenous and all start by state 0, any spectrum access failure on a channel would make it has the largest Gittins index among all channels. Therefore, by the optimal policy specified in (27), the result follows. □

We notice that Zhao *et al.* [15] have also found a myopic rule when channels are homogenous, based on a POMDP framework. Their main results are that their myopic policy can be rigorously proved to be optimal when $n = 2$ but only conjectured to be optimal when $n > 2$ via extensive numerical and simulation results. Comparing those two myopic policies, we find that they are different in the capability of characterizing dynamics of unselected channels. Specifically, as a non-stationary policy, the counting process employed by the myopic policy [15] is able to predict the channel behavior even if it is not selected. As a result, it is able to finely differentiate channels according to the history of spectrum sensing and access decisions. However, on the other hand, owing to the limitation of our MAB formulation, information state of a channel is only updated whenever it is selected, though it indeed changes slowly as time goes by no matter it is selected or not. As one of the side effects, for example, when channel switching happens, there might be more than one channel having the same information states, which however do not indicate the same quality to be accessed. When this happens, our myopic policy is only able to randomly select one of them. Fortunately, as we see in Section VI, our MAB-based policy results in slight performance drop, which also implies that our scheme can also be an efficient policy in a more general case of heterogeneous channels.

V. MODEL-FREE GITTINS INDICES CALCULATION BY Q-LEARNING

In this section, we design a reinforcement learning algorithm to calculate Gittins indices without knowing the statistical parameters of channel availability. In order to realize the model-free calculation of Gittins indices, we adopt a reinforcement learning algorithm proposed in [18], which starts from a new interpretation of Gittins index as a *restart-in-state* problem [19]. Specifically, as to a restart-in-state-x problem for state x, we can imagine an MDP with only two actions, *continue* (C) or *return* (R) for state x and then continue from there. By using standard results of MDP theory, the value function with respect to state x satisfy

$$V(x) = \sup_{\tau>0} E\{\sum_{k=0}^{\tau-1} \gamma^k R(x(k)|x(0)=x) + \gamma^\tau V(x)\} \quad (39)$$

where τ is a random stopping time that the decision maker chooses to restart in state x in the restart-in-state-x problem. Combined with the results of Whittle [20], (39) implies that Gittins index of state x is

$$v(x) = (1-\gamma)V(x). \quad (40)$$

Therefore, Gittins index for a given state is characterized by the optimal value function of state x in the above "virtual" MDP. As a variant of a classic stopping problem, it is ready to be solved by Q-learning [21], a sample-based Monte-Carlo extension, via successive approximation.

Before presenting the Q-learning algorithm for Gittins index calculation, we first define several notations regarding state-action pairs as follows. We use z defined in Theorem 5 as the state and then specify that $z \in [0, M]$, where M is the maximum number of spectrum access failures, which can be tolerated for a channel. In addition, similar to [18], a Q-factor, representing each state-action pair in the decision problem, is defined by a 4-tuple as $Q(state\ z_j^i,\ action\ a,\ initial\ state\ z_m^i,\ channel\ i)$, where $a \in \{C, R\}$, $i \in \mathcal{N}$.

The basic operations of Q-learning is stated as follows. When accessing one channel at time slot k, the state of that channel changes and the associated reward is gained. In order to reduce the computational complexity when solving multiple MDPs simultaneously, such a transition is associated to not only all restart problems by taking the continue action starting at all states but also all restart problems by taking the restart action starting at that state. Therefore, there will be totally $2(M+1)$ Q-factor updates after each spectrum access. Moreover, in order to achieve an adequate sampling of states and actions, we use Boltzmann-distribution based channel selection. As a result, recall that Gittins index of channel i in state z_j^i is approximated by $(1-\gamma)Q(z_j^i, C, z_j^i, i)$ from (40), the randomized spectrum sensing scheme is specified by

$$Pr\{select\ channel\ i\} = \frac{e^{\frac{Q(z_j^i, C, z_j^i, i)}{B_k}}}{\sum_{i=1}^{N} e^{\frac{Q(z_j^i, C, z_j^i, i)}{B_k}}} \quad (41)$$

where B denotes Boltzmann temperature. Lastly, though Q-learning algorithm is highlighted to be a model-free scheme, the convergence rate on learning the "optimal" policy from history could be very slow. Therefore, in order to remedy the discrepancy that information states of unselected channels actually change all the time, we refresh $\{z_j^i\}$ for those channels which have not been scheduled for a long time.

Algorithm 1 Gittins Index Calculation by Q-Learning

Initialization

1: $k = 0$.
2: $Q(z_j^i, a, z_m^i, i) = 0,\ \forall i \in \mathcal{N},\ 0 \le z_j, z_m \le M,\ a \in \{C, R\}$.
3: $z_j^i = 0,\ \forall i \in \mathcal{N}$.
4: $N(i) = 0,\ \forall i \in \mathcal{N}$. (the number of time slots that channel i has not been scheduled)

At each time slot k

1: Update B_k and randomly select a channel i for spectrum sensing according to (41).
2: $N(i) = 0$ and $N(l) = N(l) + 1,\ \forall l \in \mathcal{N}\backslash\{i\}$.
3: Access the selected channel according to Theorem 1.
4: By observing its state transits from z_j^i to z_m^i and immediate reward r, Q-factors are updated as

$$Q(z_j^i, C, z_l^i, i) = (1-\alpha_k^i)Q(z_j^i, C, z_l^i, i) + \alpha_k^i(r + \gamma \max_{a\in\{C,R\}} Q(z_m^i, a, z_l^i, i)) \quad (42)$$

$$Q(z_l^i, R, z_j^i, i) = (1-\alpha_k^i)Q(z_l^i, R, z_j^i, i) + \alpha_k^i(r + \gamma \max_{a\in\{C,R\}} Q(z_m^i, a, z_j^i, i)) \quad (43)$$

where $0 \le z_l^i \le M$ and α_k^i is the learning rate.
5: $z_j^i = z_m^i$.
6: **if** $N(l) \ge 2 * n$ **then**
7: $z_j^l = 0$ and $N(l) = 0, l \in \mathcal{N}$.
8: **end if**
9: $k = k + 1$.

In summary, our Q-learning algorithm is formally presented in Algorithm 1. We can observe that, $2n(M+1)^2$ Q-factors need to be stored and only $2n$ Q-factors need to be updated after each decision, where the linear complexity in terms of both storage and computational cost with respect to the number of channels n indicates it as a low-complexity scheme.

VI. SIMULATION RESULTS

In this section, we present three sets of simulations: the first and second sets are for our MAB scheme with known channel statistical parameters with homogeneous and heterogeneous channels, respectively, and the third is for our MAB-based learning scheme without any known channel statistical parameters[10].

All simulations are conducted in Matlab [22] conforming to

[10]Notice that channel states in all simulations are normally evolved according to their transition probability matrix, as shown in Fig. 1, i.e., they are not (artificially) frozen at any point in time.

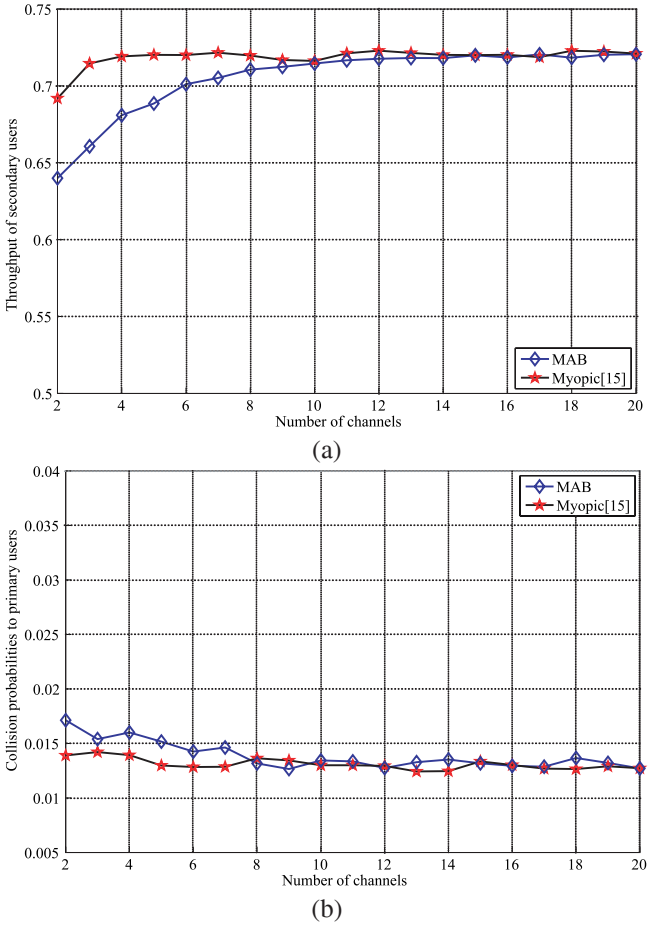

Fig. 4. Performance for positively autocorrelated homogeneous channels, where $p_{00} = 0.8$ and $p_{10} = 0.3$: (a) Throughput of secondary users, (b) collision probabilities to primary networks.

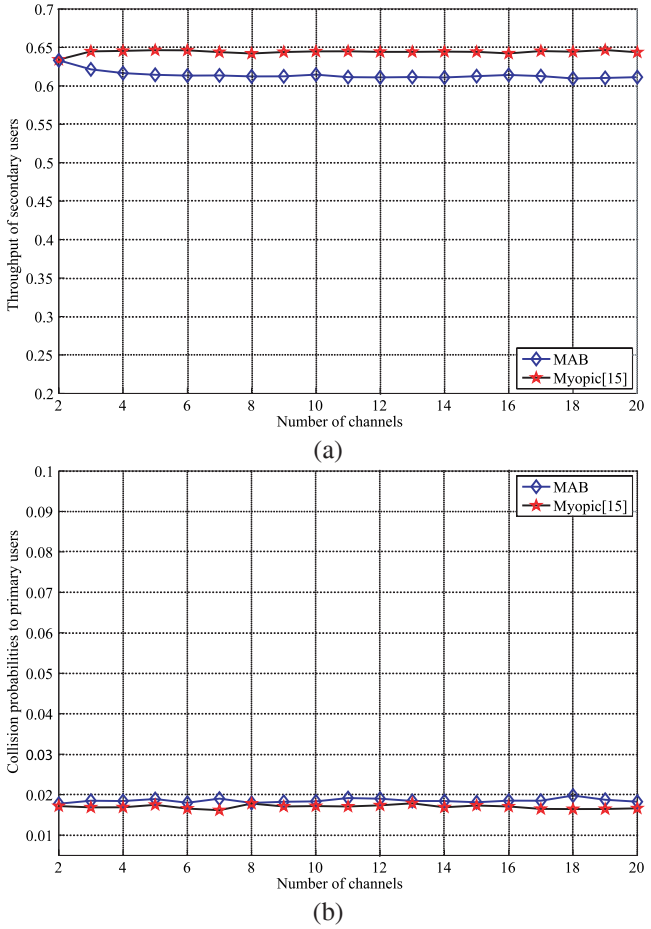

Fig. 5. Performance for negatively autocorrelated homogeneous channels, where $p_{00} = 0.3$ and $p_{10} = 0.8$: (a) Throughput of secondary users, and (b) collision probabilities to primary networks.

the system model described in Section II, and a set of parameters, $\epsilon = 0.0274$, $\delta = 0.05$, $\xi = 0.05$, and $w_i = 1$, $\forall\, i \in \mathcal{N}$ [15], are employed throughout simulations. Notice that every single data point to be shown in figures is taken by the mean of $1,000$ experiments with randomized channel realizations.

First, we consider scenarios with homogeneous channels, which are though unrealistic in practice. This is because that, the myopic policy derived in [15], has been conjectured to be optimal recently, though only the case for $n = 2$ is rigorously proved. On the other hand, our myopic policy derived by Theorem 5, is merely optimal in a MAB framework, which has been reached by making a "frozen." assumption on unobserved channels and thus efficient. In order to well understand the impact of such a critical assumption on sub-optimality, we compare the spectrum efficiency and collision probabilities to primary networks of the above schemes as Fig. 4 and 5.

Fig. 4 shows the performance in scenarios where homogeneous channels are positively autocorrelated. We can observe that the average throughput loss of our scheme in 100 slots to the myopic one is 7.5% when $n = 2$. As the number of channels increases, such a loss vanishes quickly and it is only 1.84% on average. Moreover, the average collision probabilities of our scheme (1.39%) is slightly higher than the one in [15] (1.31%).

Fig. 5 shows the performance in scenarios where homogeneous channels are negatively autocorrelated. We can observe that the average throughput loss of our scheme in 100 slots to the myopic one is almost zero when $n = 2$. As the increase of the number of channels, the the average throughput loss increases and quickly maintains at a level of 5.00%. Moreover, the average collision probabilities of our scheme (1.85%) is again slightly higher than the one in [15] (1.7%). Notice that, although the average throughput of our scheme can be regarded to be close to the optimal one in both cases, the patterns of average throughput loss are different. As stated in Theorem 5, when channels are *positively autocorrelated*, secondary users would stick on the same channel until the access fails; while when channels are *negatively autocorrelated*, secondary users would stick on the same channel until the access successes. In another word, channel switch happens when spectrum fails for *positively autocorrelated* channels while channel switch happens when spectrum successes for *negatively autocorrelated* channels. As a result, as the number of channels increases, when channels are *negatively autocorrelated*, the secondary user is more likely to switch to a favorable channel since it keeps changing channels if spectrum access fails, which corresponds to the diminished performance gap as shown in Fig. VI; on the other hand, when channels are *negatively autocorrelated*, it merely gets benefits by increasing the number of channels since it would get trapped

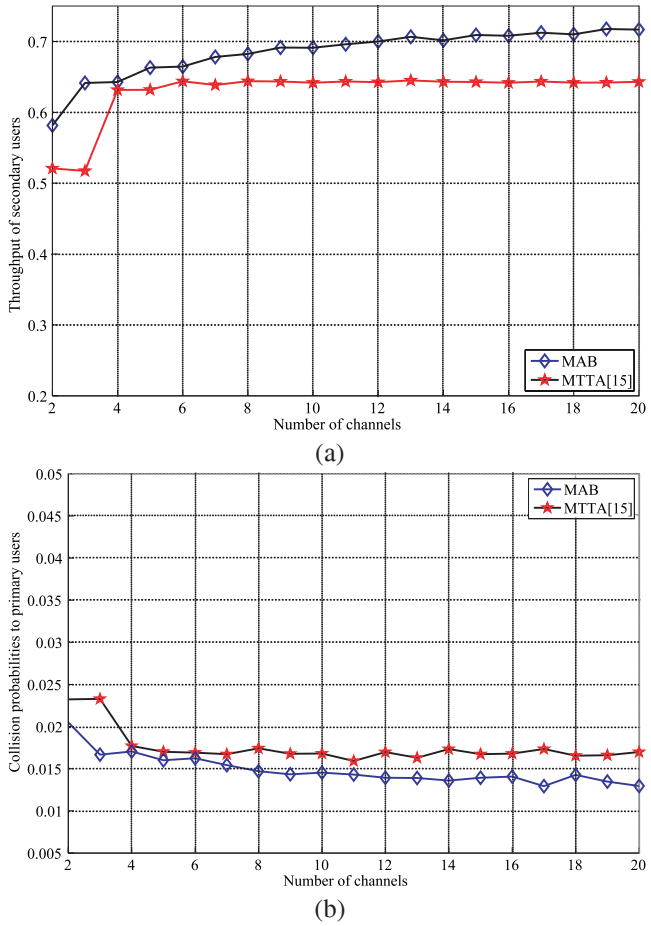

Fig. 6. Performance for heterogeneous channels: (a) Throughput of secondary users, and (b) collision probabilities to primary networks.

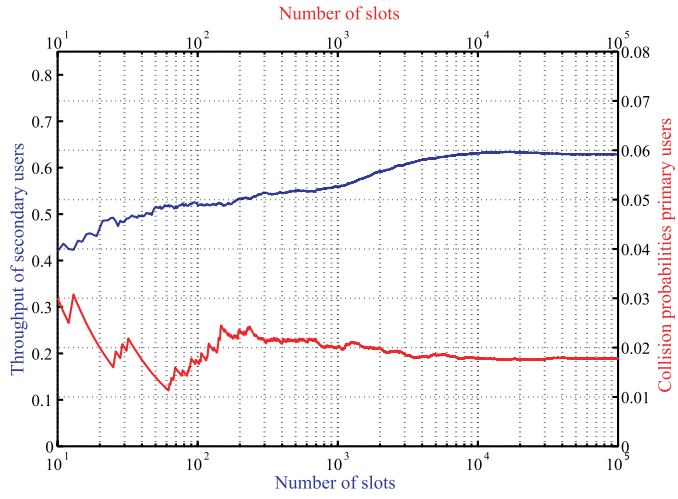

Fig. 7. Throughput of secondary users and collision probabilities vs. the number of time slots.

by some unfavorable channels for a while, which corresponds to insensitive performance gap as shown in Fig. VI.

Secondly, we consider scenarios with heterogeneous channels, which are more realistic in practice. Since it is very prohibitive to derive optimal policy as the length of horizon and the number of channels increase, we use MTTA, a near-optimal scheme proposed in [15], to compare with our scheme, instead. We find that our scheme given by (27) is generally better than MTTA. This is expected because our scheme is derived based on a solid theoretical foundation (i.e., bandit problem) while MTTA is devised based on heuristic. Notice that the performance gain depends on the setting of channel statistical parameters. For example, when we uniformly mix two types of channels used in the first set of simulations, Fig. 6 shows 8.23% gain in terms of average throughput in 100 slots. Correspondingly, the average collision probabilities of our scheme (1.49%) is lower than MTTA (1.75%).

Finally, we characterize the convergence rate of our Q-learning algorithm (i.e., Alg. 1). To the best of our knowledge, this is the first learning scheme for OSA in literature. We are interested in such a non-parameter approach since channel dynamics are usually non-stationary in reality, i.e., channel state transition probabilities would *slowly* change along the time. With the same parameters as the second set of simulations, Fig. 7 depicts the average throughput and average collision probabilities as the number of time slots increases for 8 channels. We can observe that both metrics continuously improve for the first $1,000$ time slots and then *stabilize* roughly after $4,000$ time slots. Notice that the slow convergence rate is inherent in the family Q-learning algorithms. Besides tuning Q-learning parameters, it is more sensible to limit the number of channels to be sensed and/or accessed to overcome such a drawback to some extent when applying in practice. In addition, we can also observe that both of the achieved metrics are slightly worse than the ones when channel statistical parameters are known as shown in Fig. 6, which can be regarded as the cost of lacking channel knowledge.

VII. CONCLUSIONS AND FUTURE WORK

We have considered the design of efficient low-complexity opportunistic spectrum access (OSA) policies, given the fact that the derivation of optimal policy is highly prohibitive with respect to either the horizon length or the number of channels. Specifically, by taking account of spectrum sensing errors, we are able to formulate OSA in a multi-armed bandit (MAB) framework with an approximation on information state evolution. We then derive closed-form expressions for Gittins indices of individual channels with tunable approximation, which are the essentials for the optimal policy, i.e., the well-known Gittins index rule, for MAB, no matter the channel parameters are available a priori or not. Finally, we show that our scheme achieves a better trade-off between quality of policies and complexity than other schemes in literature via extensive simulations.

We believe that this work opens doors for several avenues on developing low-complexity policies in future. For example, it could be more practical to extend our schemes to scenarios where primary users are not synchronized. Moreover, in future work we are interested in extending the analysis where a single pair of secondary users can access multiple channels at a time.

APPENDICES

I. PROOF OF THEOREM 2

Proof: First, we take the derivative of $f(x)$ as

$$f'(x) = \frac{(p_{00} - p_{10})(1-\lambda)}{(1-\lambda x)^2} \tag{44}$$

Regarding the monotonic property of $f(\cdot)$, it naturally lead to a discussion of the following two cases.

Case 1: When $p_{00} > p_{10}$, $f(x)$ is an increasing function with respect to x over $(0,1)$. We then prove that the series of information states $\{f^i(p_{00})\}_{i=0}^{\infty}$ is decreasing towards x^* by induction. For the $i=0$ case, we have

$$p_{00} - f(p_{00}) = \frac{(p_{00} - p_{10})(1-p_{00})}{1-\lambda p_{00}} > 0. \tag{45}$$

Therefore, the conclusion holds. Next, we assume that $f^l(p_{00}) > f^{l+1}(p_{00})$ holds for the $i=l$ case. For the $i=l+1$ case, recall that $f(\cdot)$ is an increasing function of x, after applying such a property on both side of the $i=l$ case, we can observe that the $i=l+1$ case also holds, i.e., $f^{l+1}(p_{00}) > f^{l+2}(p_{00})$, which completes the proof that the information state series is decreasing as i increases. As a result, $\{f^i(p_{00})\}_{i=0}^{\infty}$ decreases as i increases. On the other hand, we know that such a seises is bounded by $f(0) = p_{10}$. Combined the above results, we conclude that $\{f^i(p_{00})\}_{i=0}^{\infty}$ decreasingly converges towards a fixed point x^* as i increases.

Case 2: When $p_{00} < p_{10}$, $f(x)$ is a decreasing function with respect to x over $(0,1)$. We then prove that the series of information states $\{f^i(p_{00})\}_{i=0}^{\infty}$ is alternatively increasing and decreasing towards x^* by induction, i.e., $p_{00} < f^2(p_{00}) < \cdots$ for $i=2l$ and $l=0,1,\cdots$, and $f(p_{00}) > f^3(p_{00}) > \cdots$ for $i=2l+1$ and $l=0,1,\cdots$. Similarly to (45), we have $p_{00} < f(p_{00})$. Applying the monotonic property of the function $f(\cdot)$ on the both sides once and twice, we obtain $f(p_{00}) > f^2(p_{00})$ and $f^2(p_{00}) < f^3(p_{00})$, respectively. Moreover, we compare p_{00} and $f^2(p_{00})$ as

$$\begin{aligned} &p_{00} - f^2(p_{00}) \\ =&\frac{(p_{00}-p_{10})(1-p_{00})[p_{00}(1-\lambda)-p_{10}+1]}{1-\lambda p_{10}-\lambda p_{00}[p_{00}(1-\lambda)-p_{10}+1]} < 0. \end{aligned} \tag{46}$$

Therefore, given $p_{00} < f^2(p_{00})$, applying the monotonic property of the function $f(\cdot)$ on the both sides, we have $f(p_{00}) > f^3(p_{00})$. Combined all above results, we have $p_{00} < f^2(p_{00}) < f^3(p_{00}) < f(p_{00})$, which proves the $i=0$ case with $p_{00} = f^0(p_{00}) < f^2(p_{00})$ and $f(p_{00}) > f^3(p_{00})$ hold simultaneously. Next, we assume that $f^{2l}(p_{00}) < f^{2l+2}(p_{00}) < f^{2l+3}(p_{00}) < f^{2l+1}(p_{00})$ holds for the $i=l$ case, applying the monotonic property of the function $f(\cdot)$ on the both sides twice, then we have $f^{2l+2}(p_{00}) < f^{2l+4}(p_{00}) < f^{2l+5}(p_{00}) < f^{2l+3}(p_{00})$, which completes the proof of the $i=l+1$ case. Therefore, the series $\{p_{00}, f^2(p_{00}), \cdots\}$ is increasing while the series $\{f(p_{00}), f^3(p_{00}), \cdots\}$ is decreasing as i increases. On the other hand, both of series are bounded by $f(1) = p_{10}$ and thus converge. However, we still have to show that those two series converge to the same value x^*, i.e., there exists one and only one fixed point in the interval (p_{00}, p_{10}). Thus, we let $g(x) \triangleq (1-\lambda x)(f(x)-x)$ and we can show that

$$g(p_{00})g(p_{10}) = (p_{00}-1)(p_{00}-p_{10})(\lambda p_{10}-1)(p_{10}-p_{00}) < 0. \tag{47}$$

Given $g(x)$ is a quadric function, there would be only one $x^* \in (p_{00}, p_{10})$ such that $g(x^*) = 0$. Therefore, there exists only one solution for $f(x) = x$ and, we can conclude that both of two series converge to the same value, namely, x^*.

Combining results of the above two cases, we complete the proof. □

II. PROOF OF THEOREM 3

Proof: First, we characterize the fixed point x^* by solving $f(x) = x$ as

$$x^* = \frac{1+p_{10}-p_{00}(1-\lambda)-\sqrt{[1+p_{10}-p_{00}(1-\lambda)]^2-4\lambda p_{10}}}{2\lambda}. \tag{48}$$

Given that $\lambda \to 1$, we have

$$\lim_{\lambda \to 1} x^* = p_{10}. \tag{49}$$

Next, knowing the structural information on the convergence of the series of information states revealed by Theorem 2, we consider two cases accordingly as follows.

Case 1: When $p_{00} > p_{10}$, since $f''(x) = \frac{2\lambda(p_{00}-p_{10})(1-\lambda)}{(1-\lambda x)^3} > 0$, then $f(x)$ is a convex function and we have

$$f^l(p_{00}) - x^* \geq f'(x^*)(f^{l-1}(p_{00}) - x^*),\ l = 1, 2, \cdots. \tag{50}$$

Then iterating the inequalities from (50) on the right side until $l=1$, we have

$$\begin{aligned} f^l(p_{00}) - x^* &\geq (f'(x^*))^l(p_{00} - x^*) \\ &\approx (f'(p_{10}))^l(p_{00} - p_{10}) \end{aligned} \tag{51}$$

where the approximation follows by (49).

Therefore, we have

$$|(f'(p_{10}))^l(p_{00} - p_{10})| \leq |f^l(p_{00}) - x^*| \leq \varepsilon. \tag{52}$$

Plugging (44) and then rearranging, the result for case 1 follows.

Case 2: When $p_{00} < p_{10}$, we let $h(x) \triangleq f^2(x)$ and then explore the converge rate of the series $\{p_{00}, f^2(p_{00}), \cdots\}$ instead. Consequently, we have

$$h(x) = \frac{((p_{00}(1-\lambda)-p_{10})^2 - \lambda p_{10})x + p_{10}(p_{00}(1-\lambda)-p_{10}+1)}{1-\lambda p_{10} - \lambda(p_{00}(1-\lambda)-p_{10}+1)x} \tag{53}$$

and

$$h'(x) = \frac{(p_{00}-p_{10})^2(1-\lambda)^2}{(1-\lambda p_{10} - \lambda(p_{00}(1-\lambda)-p_{10}+1)x)^2} > 0. \tag{54}$$

Similarly, we find that $h(x)$ is a convex function by showing that $h''(x) > 0$. Following the same procedure, we have

$$|(h'(p_{10}))^{\frac{l}{2}}(p_{00}-p_{10})| \leq |h^{\frac{l}{2}}(p_{00}) - x^*| \leq \varepsilon. \tag{55}$$

Again, plugging (54) and then rearranging, the result for case 2 follows. □

III. PROOF OF THEOREM 4

Proof: We first let

$$\alpha_i = E\{\sum_{k=0}^{\tau^*-1} \gamma^k R(x_k)|x_0 = x(i)\} \tag{56a}$$

$$\beta_i = E\{\sum_{k=0}^{\tau^*-1} \gamma^k |x_0 = x(i)\} \tag{56b}$$

where τ^*, the optimal stopping time achieving the index value of state i, is a random variable conditioned on the initial state.

Next, by one-step memoryless property of a Markov chain and the special structure of our state transitions, we have

$$\alpha_0 = \gamma p_{01}\alpha_1 + \gamma p_{00}(w+\alpha_0) \tag{57a}$$

$$\beta_0 = \gamma p_{01}(1+\beta_1) + \gamma p_{00}(1+\beta_0) \tag{57b}$$

$$\alpha_i = \gamma(1-\lambda f^i(p_{00}))\alpha_{i+1} + \gamma\lambda f^i(p_{00})(w+\alpha_0) \tag{58a}$$

$$\beta_i = \gamma(1-\lambda f^i(p_{00}))(1+\beta_{i+1}) + \gamma\lambda f^i(p_{00})(1+\beta_0) \tag{58b}$$

for $i = 1, \cdots, I-1$, and

$$\alpha_I = \gamma(1-\lambda f^I(p_{00}))\alpha_I + \gamma\lambda f^I(p_{00})(w+\alpha_0) \tag{59a}$$

$$\beta_I = \gamma(1-\lambda f^I(p_{00}))(1+\beta_I) + \gamma\lambda f^I(p_{00})(1+\beta_0). \tag{59b}$$

By [23, Lemma 4.2], we start to calculate the largest Gittins index and its corresponding state, i.e., $i_1 = \arg\max_{0\leq i\leq I} \alpha_i/\beta_i$, where transition probabilities to other states are set to be zero. Thereby, we have

$$\frac{\alpha_i}{\beta_i} = \begin{cases} p_{00}w, & i = 0, \\ \lambda f^i(p_{00})w, & i = 1, \cdots, I. \end{cases} \tag{60}$$

In order to derive the maximum of $\{\alpha_i/\beta_i\}_{i=0}^I$, it naturally leads to two cases as revealed by Theorem 2.

Case 1: When $p_{00} > p_{10}$, state 0 has the largest Gittins index, i.e., $p_{00}w$, since the seises $\{f^i(p_{00})\}_{i=0}^\infty$ decreases as i increases. Then we proceed to calculate the second largest Gittins index among states $\{1, \cdots, I\}$. Again, by [23, Lemma 4.2], we only retain the transitions probabilities to themselves and state 0 and, calculate $i_2 = \arg\max_{1\leq i\leq I} \alpha_i/\beta_i$. Thereby, we have

$$\frac{\alpha_i}{\beta_i} = \frac{\lambda f^i(p_{00})\alpha_0 + \lambda f^i(p_{00})w}{\lambda f^i(p_{00})\beta_0 + 1}. \tag{61}$$

Given $f^i(p_{00})$ is a decreasing function as i increases, it is easy to see that α_i/β_i in (61) decreases as i increases as well. Therefore, state 1 has the second largest Gittins index. Similarly, we can further show that state i has the $(i+1)$th largest index, for $i = 2, \cdots, I$.

Specifically, for any $i \in \{1, \cdots, I\}$, α_0 and β_0 are determined by the following equation.

$$\alpha_0 = \gamma p_{00}w + \gamma p_{01}\alpha_1 + \gamma p_{00}\alpha_0 \tag{62a}$$

$$\beta_0 = \gamma + \gamma p_{01}\beta_1 + \gamma p_{00}\beta_0 \tag{62b}$$

$$\alpha_j = \gamma\lambda f^j(p_{00})w + \gamma(1-\lambda f^j(p_{00}))\alpha_{j+1} + \gamma\lambda f^j(p_{00})\alpha_0 \tag{62c}$$

$$\beta_j = \gamma + \gamma(1-\lambda f^j(p_{00}))\beta_{j+1} + \gamma\lambda f^j(p_{00})\beta_0 \tag{62d}$$

for $j = 1, \cdots, i-1$.

$$\alpha_i = \gamma\lambda f^i(p_{00})w + \gamma\lambda f^i(p_{00})\alpha_0 \tag{62e}$$

$$\beta_i = \gamma + \gamma\lambda f^i(p_{00})\beta_0 \tag{62f}$$

Next, we rearrange (62a), (62c) and (62e) in a matrix form as $\vec{\alpha} = A\vec{\alpha} + \vec{b}$, where $\vec{\alpha} = (\alpha_0, \cdots, \alpha_I)'$, $\vec{b} = (\gamma p_{00}w, \gamma\lambda f(p_{00})w, \cdots, \gamma\lambda f^i(p_{00})w)'$, and A is as (63).

Moreover, with $(I-A)\vec{\alpha} = \vec{b}$, we can derive α_0 by Cramer's rule as

$$\alpha_0 = \frac{\det[(I-A)_0]}{\det[(I-A)]} \tag{64}$$

where $(I-A)_0$ is the matrix formed by replacing the first column of by $\vec{b}$. After some algebra work, we can obtain α_0 as (33a). Following the same procedure, we can obtain β_0 as (33b). Therefore, the result follows for case 1.

Case 2: when $p_{00} < p_{10}$ and I is even, by (61) together with Theorem 2, we find that state 1 has the largest Gittins index, i.e., $\lambda f(p_{00})$. Then, we proceed to calculate the second largest Gittins index among states $\{0, 2, \cdots, I\}$. For any state i ($2 \leq i \leq I$), since it has zero transition probability to state 1, its α_i/β_i remains to be the same as the one in (60). However, for state 0, since it has a non-zero transition probability to state 1, α_0/β_0 becomes

$$\frac{\alpha_0}{\beta_0} = \frac{p_{01}\alpha_1 + p_{00}w}{p_{01}\beta_1 + 1} = \frac{\gamma p_{01}\lambda f(p_{00})w + p_{00}w}{\gamma p_{01} + 1} \tag{65}$$

from (57).

By Theorem 2, we know that $\alpha_2/\beta_2 = \min\{\lambda f^i(p_{00})w\}_{i=2}^I$ and $\alpha_3/\beta_3 = \max\{\lambda f^i(p_{00})ws\}_{i=2}^I$. Given λ is close to 1 such that $f(p_{00}) \approx p_{10}$, we compare α_0/β_0 and α_2/β_2 as

$$\begin{aligned} \frac{\alpha_0}{\beta_0} - \frac{\alpha_2}{\beta_2} &= \frac{\gamma p_{01}\lambda f(p_{00})w + p_{00}w}{\gamma p_{01} + 1} - \lambda f^2(p_{00})w \\ &\approx \frac{\gamma p_{01}\lambda p_{10}w + p_{00}w}{\gamma p_{01} + 1} - \lambda p_{10}w \\ &= \frac{p_{00} - \lambda p_{10}}{\gamma p_{01} + 1}w < 0. \end{aligned} \tag{66}$$

Therefore, $\alpha_0/\beta_0 = \min\{\alpha_i/\beta_i\}$, $i \in \{0, 2, \cdots, I\}$ and state 3 has the second largest Gittins index, i.e., $\lambda f^3(p_{00})$. Following the same procedure, we can find that state $2i-1$ has the ith largest Gittins index for $i = 1, \cdots, I/2$.

Next, we still have to calculate Gittins indices for remaining states, i.e., $\{0, 2, \cdots, I-2, I\}$. Except state I whose α_i/β_i is

$$A=\begin{bmatrix} \gamma p_{00} & \gamma p_{01} & \cdots & \cdots & \cdots & 0 \\ \gamma\lambda f(p_{00}) & 0 & \gamma(1-\lambda f(p_{00})) & \cdots & \cdots & 0 \\ \gamma\lambda f^2(p_{00}) & 0 & 0 & \gamma(1-\lambda f^2(p_{00})) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma\lambda f^{i-1}(p_{00}) & 0 & \cdots & \cdots & 0 & \gamma(1-\lambda f^{i-1}(p_{00})) \\ \gamma\lambda f^{i}(p_{00}) & 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} \tag{63}$$

still given by (60), α_{2i}/β_{2i} can be obtained from (57) and (58) by retaining transition probabilities to states $\{1,3,\cdots,I-1\}$ as

$$\frac{\alpha_{2i}}{\beta_{2i}}=\begin{cases}\frac{(1-\lambda f^{2i}(p_{00}))\gamma\lambda f^{2i+1}(p_{00})w+\lambda f^{2i}(p_{00})w}{(1-\lambda f^{2i}(p_{00}))\gamma+1} & i=\frac{I-1}{2},\cdots,1\\ \frac{p_{01}\alpha_1+p_{00}w}{p_{01}\beta_1+1} & i=0\end{cases} \tag{67}$$

On the other hand, as shown that $\alpha_{2i}/\beta_{2i}<\lambda f^{2i+2}(p_{00})w$ before (e.g., (66)), we conclude that state I has the $(\frac{I}{2}+1)$th largest Gittins index among states $\{0,2,\cdots,I-2,I\}$, i.e., $\lambda f^I(p_{00})w$.

Then, we continue to calculate the state with the $(\frac{I}{2}+2)$th largest Gittins index among states $\{0,2,\cdots,I-2\}$. By additionally allowing non-zero transition probabilities to state I, while all other values are unchanged from (67), α_{I-2}/β_{I-2} becomes

$$\frac{\alpha_{I-2}}{\beta_{I-2}}=\frac{(1-\lambda f^{I-2}(p_{00}))\alpha_{I-1}+\lambda f^{I-2}(p_{00})w}{(1-\lambda f^{I-2}(p_{00}))\beta_{I-1}+1} \tag{68}$$

where α_{I-1} and β_{I-1} above are determined by (36). In addition, given $f^{I-1}(p_{00})>f^I(p_{00})$ by Theorem 2 for this case, we can obtain that

$$\begin{aligned}\lambda f^{I-1}(p_{00})w &> \frac{\alpha_{I-1}}{\beta_{I-1}}\\ &= \frac{(1-\lambda f^{I-1}(p_{00}))\alpha_I+\lambda f^{I-1}(p_{00})w}{(1-\lambda f^{I-1}(p_{00}))\beta_I+1}\\ &> \lambda f^I(p_{00})w\end{aligned} \tag{69}$$

where $\alpha_I/\beta_I=\lambda f^I(p_{00})w$.

Combining (68) and (69) together with $f^{I-2}(p_{00})<f^I(p_{00})$, we have

$$\frac{\alpha_{I-2}}{\beta_{I-2}}>\lambda f^{I-2}(p_{00})w. \tag{70}$$

Recall that, from (67), $\alpha_{2i}/\beta_{2i}<\lambda f^{2i+2}(p_{00})w$, for $i\in\{0,2,\cdots,I-4\}$, with (70), we conclude that state $I-2$ has the $(I/2+2)$th largest Gittins index given by (68). Following the same procedure, we calculate Gittins indices for states $I-4, I-6,\cdots,0$ sequentially, which is similar to (68). Similarly, we can also prove the case where I is odd. Therefore, the result follows for case 2. □

ACKNOWLEDGMENTS

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal," *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, Aug. 1999.

[3] V. Krishnamurthy and R. J. Evans, "Hidden markov model multiarm bandits: A methodology for beam scheduling in multitarget tracking," *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 2893–2908, Dec. 2001.

[4] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable markov processes over a finite horizon," *Operations Research*, vol. 21, no. 5, pp. 1071–1088, 1973.

[5] D. P. Bertsekas, *Dynamic programming and optimal control.* 2nd ed., vol. 2, Athena Scientific, 2001.

[6] J. C. Gittins and D. M. Jones, "A dynamic allocation index for the sequential design of experiments," in *Progress in Statistics (European Meeting of Statisticians, Budapest, 1972)*, 1974, pp. 241–266.

[7] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Prob.*, vol. 25, pp. 287–298, 1988.

[8] R. Engelman *et al.* (2002, Nov.). Spectrum policy task force. Federal Commun. Commission. [Online]. Available: http://www.fcc.gov/sptf/files/SEWGFinalReport_1.pdf.

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

[10] 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, Apr. 2007.

[11] L. Lai *et al.*, "Cognitive medium acess: Exploration, exploitation and competition," submitted to *IEEE/ACM Trans. Netw.*, Oct. 2007.

[12] S. Geirhofer, L. Tong, and B. M. Sadler, "Dynamic spectrum access in the time domain: Modeling and exploiting white space," *IEEE Commun. Mag.*, vol. 45, no. 5, pp. 66–72, May 2007.

[13] D. V. Djonin, Q. Zhao, and V. Krishnamurthy, "Optimality and complexity of opportunistic spectrum access: A truncated markov decision process formulation," in *Proc. IEEE ICC*, June 2007, pp. 5787–5792.

[14] S. Geirhofer, L. Tong, and B. M. Sadler, "Dynamic spectrum access in wlan channels: Empirical model and its stochastic analysis," in *Proc. ACM TAPAS*, Boston, MA, USA, Aug. 2006.

[15] Q. Zhao, B. Krishnamachari, and K. Liu, "Low-complexity approaches to spectrum opportunity tracking," in *Proc. CrownCom*, Orlando, FL., Aug. 2007.

[16] K. Liu and Q. Zhao, "A restless bandit formulation of opportunistic access: Indexablity and index policy," in *Proc. IEEE Workshop on Netw. Technol. for Software Defined Radio (SDR) Networks*, June 2008.

[17] G. Koole, Z. Liu, and R. Righter, "Optimal transmission policies for noisy channels," *Operations Research*, vol. 49, no. 6, pp. 892–899, Nov. 2001.

[18] M. O. Duff, "Q-learning for bandit problems," in *Proc. IEEE ICML*, July 1995, pp. 209–217.

[19] M. N. Katehakis and J. Arthur F. Veinott, "The multi-armed bandits problem: Decomposition and computation," *Mathematics of Operations Research*, vol. 12, no. 2, pp. 262–268, May 1987.

[20] P. Whittle, "Multi-armed bandit and the gittins index," *J. Royal Statistical Society. Series B (Methodology)*, vol. 24, no. 2, pp. 143–149, 1980.

[21] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Mach. Learning*, vol. 16, no. 3, pp. 185–202, Sept. 1994.

[22] The MathWorks Inc., "Matlab.," [Online]. Available: http://www.mathworks.com/.

[23] P. P. Varaiya, J. C. Walrand, and C. Buyukkoc, "Extensions of the multiarmed bandit problem: The discounted case," *IEEE Trans. Autom. Control*, vol. 30, no. 5, pp. 426–439, May 1985.

Jing Ai received his Ph.D. degree in Computer Systems Engineering from Rensselaer Polytechnic Institute in August 2008. He received his B.E. and M.E. in the Electrical Engineering from Huazhong University of Science and Technology (HUST) in 2000 and 2002, respectively. He is now a Member of technical staff in Juniper Networks. His research interests include coverage and connectivity in wireless sensor networks, dynamic resource allocation, stochastic scheduling and cross-layer design in various types of wireless networks, e.g., wireless ad hoc networks, wireless mesh networks and cognitive radio networks.

Alhussein A. Abouzeid received the B.S. degree with honors from Cairo University, Cairo, Egypt in 1993, and the M.S. and Ph.D. degrees from University of Washington, Seattle, WA in 1999 and 2001, respectively, all in electrical engineering. He also received a post-graduate diploma in software engineering from the Information Technology Institute, The Cabinet of Egypt in 1994. He held summer appointments during his Ph.D. study with Allied Signal (now Honeywell), Redmond WA and Hughes Research Laboratories, Malibu, CA in 1999 and 2000, respectively.

Since December 2008, he has been a Program Director in the Computer and Network Systems Division, Directorate of Computer & Information Science & Engineering, National Science Foundation, Arlington, VA. Since 2001, he has been with the Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute (RPI), Troy, NY, where he is currently Associate Professor. Since 2007, he has been Deputy Director of the Center for Pervasive Computing & Networking at RPI. From 1994 to 1997 he was a Project Manager with Alcatel telecom, middle east regional office, Cairo, Egypt.

His research interests span various topics of computer networks with particular interest in wireless networks. He received the Faculty Early Career Development Award (CAREER) from the US National Science Foundation in 2006. He is a Member of IEEE and ACM and serves on the technical program and executive committees of various conferences. He is also a Member of the editorial board of Computer Networks (Elsevier).