On Optimality of Myopic Policy for Restless Multi-Armed Bandit Problem: An Axiomatic Approach

Kehao Wang and Lin Chen

***Abstract*—Due to its application in numerous engineering problems, the restless multi-armed bandit (RMAB) problem is of fundamental importance in stochastic decision theory. However, solving the RMAB problem is well known to be PSPACE-hard, with the optimal policy usually intractable due to the exponential computation complexity. A natural alternative approach is to seek simple myopic policies which are easy to implement. This paper presents a generic study on the optimality of the myopic policy for the RMAB problem. More specifically, we develop three axioms characterizing a family of generic and practically important functions termed as regular functions. By performing a mathematical analysis based on the developed axioms, we establish the closed-form conditions under which the myopic policy is guaranteed to be optimal. The axiomatic analysis also illuminates important engineering implications of the myopic policy including the intrinsic tradeoff between exploration and exploitation. A case study is then presented to illustrate the application of the derived results in analyzing a class of RMAB problems arising from multi-channel opportunistic access.**

***Index Terms*—Myopic policy, opportunistic spectrum access (OSA), restless multi-armed bandit (RMAB) problem.**

I. Introduction

THE restless multi-armed bandit (RMAB) problem, one of the most well-known generalizations of the classic multi-armed bandit (MAB) problem, is of fundamental importance in stochastic decision theory due to its generic nature and its application in numerous engineering problems such as wireless channel access, communication jamming and object tracking. The standard formulation of the RMAB problem can be briefly summarized as follows[1]: There is a bandit of N independent arms, each evolving as a two-state Markov process. At each time slot, a player chooses k $(1 \leq k \leq N)$ of the N arms to play and receives a certain amount of reward depending on the state of the played arms. Given the initial state of the system, the goal of the player is to find the optimal policy of playing the k arms at each slot so as to maximize the aggregated discounted long-term reward.

Despite the significant research efforts in the field, the RMAB problem in its generic form still remains open. Until today, few results are reported on the structure of the optimal policy. Obtaining the optimal policy for a general RMAB problem is often intractable due to the exponential computation complexity. Hence, a natural alternative is to seek simple myopic policies maximizing the short-term reward.[2] However, the optimality of such myopic policies is not always guaranteed. In such context, a natural while fundamentally important question arises: Under what conditions is the myopic policy guaranteed to be optimal?

In this paper, we answer the above posed question by performing an axiomatic study. More specifically, we develop three axioms characterizing a family of functions which we refer to as *regular* functions, which are generic and practically important. We then establish the optimality of the myopic policy when the reward function can be express as a regular function and when the discount factor is bounded by a closed-form threshold determined by the reward function. We also illustrate how the derived results, generic in nature, are applied to analyze a class of RMAB problems arising from multi-channel opportunistic access.

Compared with the existing literature addressing the optimality of the myopic policy of the RMAB problem such as [1], [2], the contribution of this paper is twofold.

1) When studying the optimality of the myopic policy, most existing works focus on the homogeneous case where each channel follows the identical Markov chain model, including our previous work [3] focusing on the optimality of the myopic policy. However, the analysis in [3] relies on some specific properties of the homogeneous channels to establish the optimality. These properties are no more applicable in the heterogeneous case where the Markov chains characterizing the channels are not identical, which requires an original study that cannot draw on existing results. To the best of our knowledge, very few results have been obtained for the heterogeneous case. Our work presented in this paper fills this void by establishing the conditions on the optimality of the myopic policy for the heterogeneous case.
2) In contrast to the research line followed by the related works in [1] and [2] aiming at showing the optimality/non-

Manuscript received April 18, 2011; revised August 16, 2011 and September 20, 2011; accepted September 21, 2011. Date of publication October 06, 2011; date of current version December 16, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Maja Bystrom.

K. Wang is with the School of Information Engineering, Wuhan University of Technology, 430070 Wuhan, China, and the Laboratoire de Recherche en Informatique (LRI), Department of Computer Science, the University of Paris-Sud XI, 91405 Orsay, France (e-mail: Kehao.Wang@whut.edu.cn, lri.fr).

L. Chen is with the Laboratoire de Recherche en Informatique (LRI), Department of Computer Science, University of Paris-Sud XI, 91405 Orsay, France (e-mail: Lin.Chen@lri.fr).

Digital Object Identifier 10.1109/TSP.2011.2170684

[1]Please refer to Section III for a detailed formulation of the RMAB problem studied in this paper.

[2]The myopic policy is also termed as greedy policy in the literature.

optimality of the myopic policy in given application scenarios, our work makes a more generic effort by focusing on the conditions ensuring the optimality without assuming any specific system setting.

From the methodological perspective, we adopt an axiomatic approach to streamline the analysis in the paper. On one hand, such axiomatic approach provides a hierarchical view of the addressed problem and leads to clearer and more synthetic analysis. On the other hand, the axiomatic approach also helps reduce the complexity of solving the RMAB problem and illustrates some important engineering implications behind the myopic policy.

The paper is organized as follows. Section II provides a brief summary on the related work on the RMAB problem in the literature. Section III formulates the RMAB problem and defines the myopic policy in the generic case. Section IV establishes the three axioms characterizing a family of generic functions and introduces the notion of regular functions. Section V further defines the pseudo value function and investigates the structural properties which are crucial to study the optimality of the myopic policy. Section VI establishes the conditions under which the myopic policy is optimal. Section VII provides a case study on the application of the major results. Finally, the paper is concluded in Section VIII.

II. Related Work

The root of the RMAB problem is the classic multi-armed bandit (MAB) problem in stochastic decision theory, originally proposed by Robbins [4]. In the standard MAB problem, a player activates one arm at each time slot and obtains a reward determined by the state of the activated arm. Only the activated arm changes its state as modeled by a Markov chain, with the states of the inactivated arms frozen. The objective is to maximize the long-term reward by choosing which arm to activate at each time slot. The breakthrough in characterizing the optimal policy is the seminal work of Gittins in [5] showing that there exists an index for each arm independent of the states of other arms and that playing the arm with the highest index results to be optimal. The index is later termed the Gittins index [6]. With the index structure of the myopic policy, the originally N-dimensional problem can be reduced to N independent one-dimensional problems.

However, when generalized to the RMAB problem, where the player is allowed to activate multiple arms and more importantly, the state of arms evolves even if the arm is not activated, the index-based policy is no more optimal. In fact, finding the optimal policy in the generic RMAB problem is shown to be PSPACE-hard by Papadimitriou *et al.* in [7]. Whittle proposed a heuristic index policy, called Whittle index policy [8] which are shown to be asymptotically optimal in certain limited regime under some specific constraints [9]. Unfortunately, not every RMAB problem has a well-defined Whittle index. Moreover, computing the Whittle index can be prohibitively complex. In this regard, Liu *et al.* studied in [10] the indexability of a class of RMAB problems relevant to dynamic multi-channel access applications. However, the optimality of the myopic policy based on Whittle index is not ensured in the general cases, especially when the arms follow non-identical Markov chains.

More recently, there are two major thrusts in the study of the myopic policy in the RMAB problem. Since the optimality of the myopic policy is not generally guaranteed, the first research thrust is to study how far it is to the optimal and design approximation algorithms and heuristic policies. The works of [11]–[13] follow this line of research. Specifically, a simple myopic policy, termed as greedy policy, is developed in [11] that yields a factor 2 approximation of the optimal policy for a subclass of scenarios referred to as *Monotone bandits*. The other thrust, more application-oriented, consists of establishing the optimality of the myopic policy in some specific application scenarios, particularly in the context of opportunistic spectrum access. The works in [1], [2], [14], and [15] belong to this category by focusing on specific forms of reward functions. More specifically, [1] studies the structure of the myopic sensing policy in the case where the user is allowed to sense one out of the N channels each slot and establishes the optimality of the myopic policy for $N = 2$. Reference [14] extends the work of [1] to the general case $N > 2$ by proving the optimality of the myopic sensing policy under certain conditions on the channel parameters and the discount factor in the utility function. [15] further relaxes the conditions and proves the optimality when the channels are positively correlated. Reference [2] studies the optimality of the myopic sensing policy when the user are allowed to sense multiple channels and transmit the packets on the idle channels. The myopic policy is showed to be optimal when channels are positively correlated under such reward model. Our previous work [16], however, shows that a slightly different structure of reward function can lead to totally contrary result. In a broader context, some researchers explore the non-Bayesian versions of the RMAB problem where the underlying Markov chains are unknown and have to be learned [17]–[19].

III. System Model and Problem Formulation

For the sake of concreteness, we present the system model and formulate the RMAB problem in the context of channel access in a multi-channel opportunistic communication system. Nevertheless, the model can be readily generalized to the generic RMAB problem and applied in a variety of applications. Therefore, the following description and the use of terms should be understood generically.

A. Multi-Channel Opportunistic Access Model

We consider a multi-channel opportunistic communication system, in which a user is able to access a set $\mathcal{N}$ of N independent channels, each characterized by a Markov chain of two states, *good* (1) and *bad* (0). The channel state transition matrix $\mathbf{P}_i$ for channel i ($i \in \mathcal{N}$) is given as follows:

$$\mathbf{P}^i = \begin{bmatrix} p_{11}^i & 1-p_{11}^i \\ p_{01}^i & 1-p_{01}^i \end{bmatrix}.$$

In our work, we focus on the *positively correlated* channel setting such that $p_{11}^i > p_{01}^i \quad \forall i \in \mathcal{N}$. Note that this channel setting corresponds to the realistic scenarios where the channel states are observed to evolve gradually over time. We assume that channels go through a state transition at the beginning of each slot t. The system operates in a synchronously time slotted fashion with the time slot indexed by $t (t = 1, 2, \ldots, T)$, where T is the time horizon of interest. This generic multi-channel opportunistic communication model can be naturally cast into the

opportunistic spectrum access (OSA) problem in cognitive radio systems where an unlicensed secondary user can opportunistically access the temporarily unused channels of the licensed primary users, with the availability of each channel evolving as an independent Markov chain.

Due to hardware constraints and energy cost, the user is allowed to sense only k ($1 \le k \le N$) of the N channels at each slot t. We denote the set of channels chosen by the user at slot t by $\mathcal{A}(t)$ where $\mathcal{A}(t) \subseteq \mathcal{N}$ and $|\mathcal{A}(t)| = k$. We assume that the user makes the channel selection decision at the beginning of each slot after the channel state transition. Based on the state of the sensed channels in slot t, denoted by $\mathbf{S}(t) \triangleq \{S_i(t), i \in \mathcal{A}(t)\}$ where $S_i(t) \in \{0,1\}$, the user obtains a certain amount of reward, characterized by the reward function $R(\mathbf{S}(t))$. A simple example of the reward function is $R(\mathbf{S}(t)) = \sum_{i\in\mathcal{A}(t)} S_i(t)$, meaning that the user gains one unit of reward for each channel sensed good (i.e., $S_i(t) = 1$), thus available for transmitting one packet on that channel. *The user's objective is to maximize the expected discounted long-term reward by designing a channel sensing policy that sequentially selects the channels to sense in each slot*. The detailed mathematical formulation of the optimization problem is given in next subsection.

Obviously, by sensing only k out of N channels, the user cannot observe the state information of the whole system. Hence, the user has to infer the channel states from its past decision and observation history so as to make its future decision. To this end, we define the *channel state belief vector* (hereinafter referred to as *belief vector* for briefness) $\Omega(t) \triangleq \{\omega_i(t), i \in \mathcal{N}\}$, where $0 \le \omega_i(t) \le 1$ is the conditional probability that channel i is in state good (i.e., $S_i(t) = 1$) at slot t given all past states, actions and observations.[3] Due to the Markovian nature of the channel model, the belief vector can be updated recursively using Bayes' rule as follows:

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

where

$$\tau_i(\omega_i(t)) \triangleq \omega_i(t)p_{11}^i + [1-\omega_i(t)]p_{01}^i \quad (2)$$

denotes the operator for the one-step belief update for non-sensed channels.

Lemma 1: If all channels are positively correlated, the following structural properties of $\tau_i(\omega_i(t))$ hold:

- $\tau_i(\omega_i(t))$ is monotonically increasing in $\omega_i(t)$;
- $p_{01}^i \le \tau_i(\omega_i(t)) \le p_{11}^i, \forall\, 0 \le \omega_i(t) \le 1$.

Proof: Noticing that $\tau_i(\omega_i(t))$ can be written as

$$\tau_i(\omega_i(t)) = (p_{11}^i - p_{01}^i)\omega_i(t) + p_{01}^i.$$

Lemma 1 holds straightforwardly. ∎

B. Optimal Sensing Problem and Myopic Sensing Policy

We are interested in the user's optimization problem to find the optimal sensing policy π^* that maximizes the expected total discounted reward over a finite horizon. Mathematically, a sensing policy π is defined as a mapping from the belief vector $\Omega(t)$ to the action (i.e., the set of channels to sense) $\mathcal{A}(t)$ in each slot t:

$$\pi:\ \Omega(t) \to \mathcal{A}(t), \quad |\mathcal{A}(t)| = k,\ t = 1,2,\ldots,T. \quad (3)$$

The following gives the formal definition of the optimal sensing problem:

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \mathbb{E}\left[\sum_{t=1}^{T} \beta^{t-1} R_\pi(\Omega(t)) \middle| \Omega(1)\right] \quad (4)$$

where $R_\pi(\Omega(t))$ is the reward collected in slot t under the sensing policy π with the initial belief vector $\Omega(1)$, $0 \le \beta \le 1$ is the discounting factor characterizing the feature that the future rewards are less valuable than the immediate reward.

To get more insight on the structure of the optimization problem and the complexity to solve it, we derive the dynamic programming formulation of (4) as follows:

$$V_T(\Omega(t)) = \max_\pi \mathbb{E}[R_\pi(\Omega(T))] = \max_{\substack{\mathcal{A}(T)\subseteq\mathcal{N} \\ |\mathcal{A}(T)|=k}} \mathbb{E}[R_\pi(\Omega(T))] \quad (5)$$

$$V_t(\Omega(t)) = \max_{\substack{\mathcal{A}(t)\subseteq\mathcal{N} \\ |\mathcal{A}(t)|=k}} \mathbb{E}\Bigg[R_\pi(\Omega(t)) + \beta \underbrace{\sum_{\mathcal{E}\subseteq\mathcal{A}(t)} \prod_{i\in\mathcal{E}} \omega_i(t) \prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}} (1-\omega_j(t)) V(\Omega(t+1))}_{\Gamma(\Omega(t))} \Bigg] \quad (6)$$

where $V_t(\Omega(t))$ is the value function corresponding to the maximal expected reward from time slot t to T($1 \le t \le T$) with the believe vector $\Omega(t+1)$ following the evolution described in (1) given that the channels in the subset $\mathcal{E}$ are sensed in state good and the channels in $\mathcal{A}(t)\backslash\mathcal{E}$ are sensed in state bad. Particularly, the term $\Gamma(\Omega(t))$ corresponds to the expected accumulated discounted reward starting from slot $t+1$ to T, calculated over all possible realizations of the selected channels (i.e., the channels in $\mathcal{A}(t)$).

Solving (4) using the above recursive iteration is computationally heavy due to the fact that the belief vector $\{\Omega(t), t = 1,2,\ldots,T\}$ is a Markov chain with uncountable state space, resulting the difficulty in tracing the optimal sensing policy π^*. Hence, a natural alternative is to seek simple myopic sensing policy which is easy to compute and implement that maximizes the immediate reward, formally defined as follows:

Definition 1 (Myopic Sensing Policy): Let the expected reward function $F(\Omega(t)) \triangleq \mathbb{E}[R_\pi(\Omega(t))]$ denote the expected immediate reward obtained in slot t under the sensing policy π. The myopic sensing policy, consists of sensing the k channels that maximizes $F(\Omega(t))$.

Despite its simple and robust structure, the optimality of the myopic sensing policy is not guaranteed. More specifically, when the channels are stochastically identical (i.e., all channels follow the same Markovian dynamics $\mathbf{P}^i = \mathbf{P}, \quad \forall i \in \mathcal{N}$) and

[3]The initial belief $\omega_i(1)$ can be set to $\frac{p_{01}^i}{p_{01}^i+1-p_{11}^i}$ if no information about the initial system state is available.

positively correlated, the myopic sensing policy is shown to be optimal when the user is limited to sense one channel each slot ($k = 1$) and obtains one unit of reward when the sensed channel is good [1]. The analysis of [15] and our work [16] further extend the study on the generic case where $k \geq 1$. However, the authors of [15] show that the myopic sensing policy is optimal if the user gets one unit of reward for each channel sensed to be good,[4] while our work [16] shows that the myopic sensing policy is not guaranteed to be optimal when the user's objective is to find at least one good channel.[5] Given that such nuance on the reward function leads to totally contrary results, a natural while fundamentally important question arises: how does the expected slot reward function $F(\Omega(t))$ impact the optimality of the myopic sensing policy? Or more specifically, under what conditions on $F(\Omega(t))$ is the myopic sensing policy guaranteed to be optimal?

In the sequel analysis in Sections IV–VI by performing an axiomatic study, we shall give affirmative answer to the above posed questions and study some important engineering implications behind the myopic sensing policy.

IV. Axioms

This section introduces a set of three axioms characterizing a family of generic and practically important functions, to which we refer as *regular* functions. The axioms developed in this section and the implied fundamental properties serve as a basis for the further analysis on the structure and the optimality of the myopic sensing policy in Sections V and VI.

Throughout this section, for the convenience of presentation, we sort the elements of the believe vector $\Omega(t) = [\omega_1(t), \ldots, \omega_N(t)]$ for each slot t such that $\mathcal{A} = \{1, \ldots, k\}$ (i.e., the user senses channel 1 to channel k) and let $\Omega_A \triangleq \{\omega_i, i \in \mathcal{A}\} = \{\omega_1, \ldots, \omega_k\}$.[6] The three axioms derived in the following characterize a generic function f defined on Ω_A.

Axiom (Symmetry): A function $f(\Omega_A) : [0,1]^k \to \mathbb{R}$ is symmetrical if $\forall i, j \in \mathcal{A}$ it holds that

$$f(\omega_1, \ldots, \omega_i, \ldots, \omega_j, \ldots, \omega_k) = f(\omega_1, \ldots, \omega_j, \ldots, \omega_i, \ldots, \omega_k).$$

Axiom (Monotonicity): A function $f(\Omega_A) : [0,1]^k \to \mathbb{R}$ is monotonically increasing if it is monotonically increasing in each variable ω_i, i.e., $\forall i \in \mathcal{A}$

$$\omega_i' > \omega_i \Longrightarrow f(\omega_1, \ldots, \omega_i', \ldots, \omega_k) > f(\omega_1, \ldots, \omega_i, \ldots, \omega_k).$$

Axiom (Decomposability): A function $f(\Omega_A) : [0,1]^k \to \mathbb{R}$ is decomposable if $\forall i \in \mathcal{A}$ it holds that

$$f(\omega_1, \ldots, \omega_i, \ldots, \omega_k) = \omega_i f(\omega_1, \ldots, 1, \ldots, \omega_k) + (1 - \omega_i) f(\omega_1, \ldots, 0, \ldots, \omega_k).$$

[4]Formally, in [15], the expected slot reward function is defined as $F(\Omega(t)) \triangleq \mathbb{E}[R_\pi(\Omega(t))] = \sum_{i \in \mathcal{A}(t)} w_i(t)$

[5]In our work [16], the expected slot reward function is defined as $F(\Omega(t)) = 1 - \prod_{i \in \mathcal{A}(t)} (1 - \omega_i(t))$

[6]For presentation simplicity, by slightly abusing the notations without introducing ambiguity, we drop the time slot index t.

Axioms 1 and 2 are intuitive. Axiom 3 on the decomposability states that $f(\Omega_A)$ can always be decomposed into two terms that replace ω_i by 0 and 1, respectively. The three axioms introduced in this section are consistent and non-redundant. Moreover, they can be used to characterize a family of generic functions, referred to as *regular* functions, defined as follows.

Definition 2 (Regular Function): A function is called regular if it satisfies all the three axioms.

The following definition studies the structure of the myopic sensing policy if the expected reward function is regular.

Definition 3 (Structure of Myopic Sensing Policy): Sort the elements of the belief vector in descending order such that $\omega_1 \geq \cdots \geq \omega_N$, if the expected reward function F is regular, then the myopic sensing policy, where the user is allowed to sense k channels, consists of sensing channel 1 to channel k.

Remark: In case of tie, we sort the channels in tie in the descending order of $\omega_i(t+1)$ calculated in (1). The argument is that larger $\omega_i(t+1)$ leads to larger expected payoff in next slot $t+1$. If the tie persists, the channels are sorted by indexes.

We would like to emphasize that the developed three axioms characterize a set of generic functions widely used in practical applications. To see this, we give two examples to get more insight: 1) The user gets one unit of reward for each channel that is sensed good. In this example, the expected reward function (for each slot), denoted as F, is the expected slot reward function is $F(\Omega_A) = \sum_{i=1}^k \omega_i$ and 2) the user gets one unit of reward if at least one channel is sensed good. In this example, the expected reward function is $F(\Omega_A) = 1 - \prod_{i=1}^k (1 - \omega_i)$. It can be verified that in both examples, F is regular by satisfying the three axioms.

V. Properties of Pseudo Value Function

Armed with the three axioms developed in the previous section, this section first defines the *pseudo value function* and then derives several fundamental properties of the pseudo value function, which are crucial in the study on the optimality of the myopic sensing policy.

To make the following presentation more convenient, we sort $\Omega(t)$ for each slot t in the descending order such that $\omega_1(t) \geq \omega_2(t) \geq \cdots \geq \omega_N(t)$ and let $\Omega_A \triangleq \{\omega_i, i \in \mathcal{A}\}$. We start by giving the formal definition of the pseudo value function in the recursive form.

Definition 4 (Pseudo Value Function): The pseudo value function, denoted as $W_t(\Omega)$, is recursively defined as in (7), shown at the bottom of the next page. $W_t^{\mathcal{A}(t)}(\Omega(t))$ is the expected total reward from slot t to T under the policy of sensing the channels in $\mathcal{A}(t)$ for slot t and then sensing the best k channels from slot $t+1$ to T. If $\mathcal{A}(t) = \{1, 2, \ldots, k\}$, then $W_t^{\mathcal{A}(t)}(\Omega(t))$ is the total reward generated by the myopic sensing policy.

It can be seen from backward induction that the myopic sensing policy is optimal if $W_t(\Omega(t))$ achieves its maximum with $\mathcal{A}(t) = \{1, 2, \ldots, k\}$. Before establishing the optimality of the myopic sensing policy in next section, this section investigates the basic structural properties of the pseudo value function, as stated in the following two lemmas.

Lemma 2 (Symmetry): If the expected reward function F is regular, the correspondent pseudo value function $W_t^{\mathcal{A}}(\Omega)$ is

symmetrical in any two channel $i, j \in \mathcal{A}$ or $i, j \in \mathcal{N} \setminus \mathcal{A}$ for all $t = 1, 2, \ldots, T$, i.e.,

$$W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_i, \ldots, \omega_j, \ldots, \omega_N) = W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_j, \ldots, \omega_i, \ldots, \omega_N).$$

Proof: The proof is given in the Appendix. ∎

Lemma 2 implies that a symmetrical pseudo value function is also robust against channel permutation given that all the permutated channels are sensed or none of them are sensed. Hence, it can be defined on two sets: the set of channels to be sensed and of those not to be sensed.

Lemma 3 (Decomposability): If the expected reward function F is regular, then the correspondent value function $W_t^{\mathcal{A}}(\Omega(t))$ is decomposable: i.e., $\forall i \in \mathcal{N}$ and $t = 1, \ldots, T$,

$$W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_i, \ldots, \omega_N) = \omega_i W_t^{\mathcal{A}}(\omega_1, \ldots, 1, \ldots, \omega_N) + (1 - \omega_i) W_t^{\mathcal{A}}(\omega_1, \ldots, 0, \ldots, \omega_N). \quad (8)$$

Proof: The lemma can be proven by backward induction noticing the structure of $W_t^{\mathcal{A}}(\Omega(t))$ in (7). ∎

Lemma 3 can be applied one step further to prove the following corollary.

Corollary 1: If the reward function $F(\Omega)$ is regular, then for any $l, m \in \mathcal{N}$ and $t = 1, \ldots, T$, it holds that

$$\begin{aligned} &W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_l, \ldots, \omega_m, \ldots, \omega_N) \\ &\quad - W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_m, \ldots, \omega_l, \ldots, \omega_N) \\ &\quad = (\omega_l - \omega_m) \left[W_t^{\mathcal{A}}(\omega_1, \ldots, 1, \ldots, 0, \ldots, \omega_N) \right. \\ &\qquad \left. - W_t^{\mathcal{A}}(\omega_1, \ldots, 0, \ldots, 1, \ldots, \omega_N) \right]. \end{aligned} \quad (9)$$

VI. Myopic Sensing Policy: Optimality Condition

Equipped with the results derived in Section V, we are ready to study the optimality of the myopic sensing policy in this section. We start by showing the following two important auxiliary lemmas (Lemma 4 and Lemma 5) and then establish the sufficient condition under which the optimality of the myopic sensing policy is ensured.

For the convenience of discussion, we firstly state some notations before developing the auxiliary lemmas. Let $\delta_p^{\max} \triangleq \max_{i \in \mathcal{N}} (p_{11}^i - p_{01}^i)$ and $\delta_p^{\min} \triangleq \min_{i \in \mathcal{N}} (p_{11}^i - p_{01}^i)$, let $\omega_{-i} \triangleq \{\omega_j, j \in \mathcal{A}, j \neq i\}$, and define

$$\begin{cases} \Delta_{\max} \triangleq \max\limits_{i \in \mathcal{N},\ \omega_{-i} \in [0,1]^{k-1}} \{F(1, \omega_{-i}) - F(0, \omega_{-i})\} \\ \Delta_{\min} \triangleq \min\limits_{i \in \mathcal{N},\ \omega_{-i} \in [0,1]^{k-1}} \{F(1, \omega_{-i}) - F(0, \omega_{-i})\}. \end{cases}$$

In Lemma 4, we consider two belief vectors $\Omega_l = [\omega_1, \ldots, \omega_l, \ldots, \omega_N]$ and $\Omega_l' = [\omega_1, \ldots, \omega_l', \ldots, \omega_N]$ that differ only in one element $\omega_l \leq \omega_l'$. Let $\mathcal{A}$ and $\mathcal{A}'$ denote the largest k elements in Ω_l and Ω_l', respectively,[7] Lemma 4 gives the lower bound and the upper bound on $W_t^{\mathcal{A}'}(\Omega_l') - W_t^{\mathcal{A}}(\Omega_l)$.

Lemma 4: If the expected reward function F is regular, $\forall l \in \mathcal{N}$, $\omega_l \leq \omega_l'$ and $1 \leq t \leq T$, it holds that

$$\begin{aligned} (\omega_l' - \omega_l)\Delta_{\min} &\leq W_t^{\mathcal{A}'}(\Omega_l') - W_t^{\mathcal{A}}(\Omega_l) \\ &\leq (\omega_l' - \omega_l)\Delta_{\max} \sum_{i=0}^{T-t} \beta^i \left(\delta_p^{\max}\right)^i, \\ &\qquad \text{if } l \in \mathcal{A} \text{ and } l \in \mathcal{A}'; \quad (10) \\ 0 &\leq W_t^{\mathcal{A}'}(\Omega_l') - W_t^{\mathcal{A}}(\Omega_l) \\ &\leq (\omega_l' - \omega_l)\delta_p^{\max}\Delta_{\max} \sum_{i=0}^{T-t-1} \beta^i \left(\delta_p^{\max}\right)^i, \\ &\qquad \text{if } l \notin \mathcal{A} \text{ and } l \notin \mathcal{A}'; \quad (11) \\ 0 &\leq W_t^{\mathcal{A}'}(\Omega_l') - W_t^{\mathcal{A}}(\Omega_l) \\ &\leq (\omega_l' - \omega_l)\Delta_{\max} \sum_{i=0}^{T-t} \beta^i \left(\delta_p^{\max}\right)^i, \\ &\qquad \text{if } l \notin \mathcal{A} \text{ but } l \in \mathcal{A}'. \quad (12) \end{aligned}$$

Proof: The proof is detailed in the Appendix. ∎

Remark: Lemma 4 bounds the difference between $W_t^{\mathcal{A}'}(\Omega_l')$ and $W_t^{\mathcal{A}}(\Omega_l)$ by distinguishing three cases. It is important to note that the case where $l \in \mathcal{A}$ but $l \notin \mathcal{A}'$ is impossible. Otherwise there exists $m \in \mathcal{A}'$ but $m \notin \mathcal{A}$. On one hand, it follows from $m \in \mathcal{A}'$ that $\omega_m > \omega_l'$ or in case of tie $\omega_m = \omega_l'$, channel m is chosen. On the other hand, it follows from $l \in \mathcal{A}$ that $\omega_l > \omega_m$ or in case of tie $\omega_m = \omega_l$, channel l is chosen. The two statements clearly contradict with each other noticing that $\omega_l' \geq \omega_l$.

We proceed one step further by considering $W_t^{\mathcal{A}'}(\Omega)$ and $W_t^{\mathcal{A}}(\Omega)$ with $\mathcal{A}'$ and $\mathcal{A}$ differing in one element in the sense that $l \in \mathcal{A}'$ and $m \in \mathcal{A}$ with $\omega_l > \omega_m$. Lemma 5 establishes the sufficient condition under which $W_t^{\mathcal{A}'}(\Omega) > W_t^{\mathcal{A}}(\Omega)$.

Lemma 5: $W_t^{\mathcal{A}'}(\Omega) > W_t^{\mathcal{A}}(\Omega)$ holds for $1 \leq t \leq T$ if the following two conditions are satisfied:

1) the expected slot reward function F is regular;
2) $\Delta_{\min} > \Delta_{\max} \sum_{i=1}^{T} \beta^i (\delta_p^{\max})^i$.

Proof: The case $t = T$ holds trivially as $\omega_l > \omega_m$. We now show that the lemma holds for $t < T$.

[7]The tie, if there exists, is resolved in the way as stated in the remark after Definition 3.

$$\begin{cases} W_T(\Omega(T)) = F(\omega_1(T), \ldots, \omega_k(T)); \\ W_r(\Omega(r)) = F(\omega_1(r), \ldots, \omega_k(r)) + \beta \sum\limits_{\mathcal{E} \subseteq \{1,2,\ldots,k\}} \prod\limits_{i \in \mathcal{E}} \omega_i(r) \prod\limits_{j \leq k, j \notin \mathcal{E}} (1 - \omega_j(r)) W_{r+1}(\Omega(r+1)),\ t < r < T; \\ W_t^{\mathcal{A}(t)}(\Omega(t)) = F(\Omega_{A(t)}) + \beta \underbrace{\sum\limits_{\mathcal{E} \subseteq \mathcal{A}(t)} \prod\limits_{i \in \mathcal{E}} \omega_i(t) \prod\limits_{j \in \mathcal{A}(t) \setminus \mathcal{E}} (1 - \omega_j(t)) W_{t+1}(\Omega(t+1))}_{\Gamma(\Omega(t))}. \end{cases} \quad (7)$$

By Corollary 1 and (7), we have

$$\begin{aligned} &W_t^{\mathcal{A}'}(\Omega) - W_t^{\mathcal{A}}(\Omega) \\ &\quad = (\omega_l - \omega_m)[W_t^{\mathcal{A}'}(\Omega_{10}(t)) - W_t^{\mathcal{A}}(\Omega_{01}(t))] \\ &\quad = (\omega_l - \omega_m)[F(1,\omega_{-i}) - F(0,\omega_{-i})] \\ &\qquad + (\omega_l - \omega_m)\beta \sum_{\mathcal{E}\subseteq\mathcal{A}'\setminus\{l\}} \prod_{i\in\mathcal{E}} \omega_i \prod_{j\in\mathcal{A}'\setminus\mathcal{E}\setminus\{l\}} (1-\omega_j) \\ &\qquad \times [W_{t+1}(\Omega_{10}(t+1)) - W_{t+1}(\Omega_{01}(t+1))] \end{aligned} \tag{13}$$

where $\Omega_{ab}(t+1)$ $(a,b \in \{0,1\})$ denotes the believe vector at slot $t+1$ with $\omega_l(t) = a$ and $\omega_m(t) = b$. It can be noticed that $\Omega_{01}(t+1)$ and $\Omega_{10}(t+1)$ differs only in two elements as illustrated by (14), shown at the bottom of the page.

We then develop

$$\begin{aligned} &W_{t+1}(\Omega_{10}(t+1)) - W_{t+1}(\Omega_{01}(t+1)) \\ &\quad = W_{t+1}(\Omega_{10}(t+1)) - W_{t+1}(\Omega_{00}(t+1)) \\ &\qquad - [W_{t+1}(\Omega_{01}(t+1)) - W_{t+1}(\Omega_{00}(t+1))]. \end{aligned}$$

Following Lemma 4, it holds that $W_{t+1}(\Omega_{10}(t+1)) \geq W_{t+1}(\Omega_{00}(t+1))$ and

$$\begin{aligned} &W_{t+1}(\Omega_{01}(t+1)) - W_{t+1}(\Omega_{00}(t+1)) \\ &\qquad \leq (p_{11}^m - p_{01}^m)\,\Delta_{\max} \sum_{i=0}^{T-t} \beta^i \left(\delta_p^{\max}\right)^i \end{aligned}$$

noticing that

$$\delta_p^{\max}\Delta_{\max} \sum_{i=0}^{T-t-1} \beta^i (\delta_p^{\max})^i < \Delta_{\max} \sum_{i=0}^{T-t} \beta^i (\delta_p^{\max})^i.$$

Noticing that $\sum_{i=1}^{T-t+1} \beta^i (\delta_p^{\max})^i$ is decreasing in t, if the two conditions in the lemma hold, it follows from (13) that

$$W_t^{\mathcal{A}'}(\Omega) - W_t^{\mathcal{A}}(\Omega) \geq \Delta_{\min} - \Delta_{\max} \sum_{i=1}^{T} \beta^i \left(\delta_p^{\max}\right)^i > 0$$

which completes our proof. ∎

Remark: It is insightful to note that the proof of Lemma 5 hinges on the fundamental trade-off between *exploitation*, by accessing the channel with the higher estimated good probability (channel l in the proof) based on currently available information (the belief vector) which greedily maximizes the immediate reward (i.e., F in the global utility function), and *exploration*, by sensing unexplored and probably less optimal channels (e.g., channel m in the proof) in order to learn and predict the future channel state, thus maximizing the long-term reward (i.e., the second term in the global utility function). If the user is sufficiently short-sighted (i.e., β is sufficiently small), exploitation naturally dominates exploration (i.e., the immediate reward overweighs the potential gain in future reward), resulting the better performance of sensing channel l w.r.t. m. The main result of Lemma 5 consists of quantifying this tradeoff between exploitation and exploration.

Armed with Lemma 5, we are now able to derive the central result of this section (Theorem 1) that can answer the questions posed at the end of Section III.

Theorem 1: The myopic sensing policy is optimal if the following two conditions hold: 1) the expected slot reward function F is regular and 2) $\Delta_{\min} > \Delta_{\max} \sum_{i=1}^{T} \beta^i (\delta_p^{\max})^i$.

Proof: We prove the theorem by backward induction. The theorem holds trivially for $t = T$. Assume that it holds for $T, T-1, \ldots, t+1$, i.e., the optimal sensing policy is to sense the best k channels from time slot $t+1$ to T. We now show that it holds for t.

To this end, assume, by contradiction, that given the belief vector $\Omega \triangleq \{\omega_{i_1}, \ldots, \omega_{i_N}\}$, the optimal sensing policy is to sense the best k channels from time slot $t+1$ to T and at slot t to sense channels $\{i_1, \ldots, i_k\} \neq \{1, \ldots, k\}$, given that the latter contains the best k channels in terms of belief values at slot t. There must exist i_m and i_l where $m \leq k < l$ such that $\omega_{i_m} < \omega_k \leq \omega_{i_l}$. It then follows from Lemma 5 that

$$W_t^{\{i_1,i_2,\ldots,i_k\}}(\Omega) < W_t^{\{\omega_{i_1},\omega_{i_2},\ldots,\omega_{i_{m-1}},\omega_{i_l},\omega_{i_{m+1}},\omega_{i_k}\}}(\Omega)$$

implying that sensing $\{i_1, \ldots, i_{m-1}, i_l, i_{m+1}, \ldots, i_k\}$ at slot t and then following the myopic sensing policy is better than sensing channels $\{i_1, \ldots, i_k\}$ at slot t and then following the myopic sensing policy, which contradicts with the assumption that the latter is the optimal sensing policy. This contradiction completes our proof. ∎

We conclude this section by studying the optimality of the myopic sensing policy for the case of infinite time horizon $T \to \infty$ in the following theorem. The proof follows straightforwardly from Theorem 1 by noticing that $\sum_{i=1}^{\infty} x^i = x/(1-x)$ for any $x \in [0,1]$.

Theorem 2: In the infinite horizon case $T \to \infty$, the myopic sensing policy is optimal if the following conditions hold: (1) F is regular; (2) $\Delta_{\min} > \dfrac{\beta\delta_p^{\max}\Delta_{\max}}{1-\beta\delta_p^{\max}}$.

VII. Application: Case Study

To illustrate the application of the results obtained in this paper, this section presents a comparative and synthetic analysis on the RMAB problem with different reward functions analyzed in [2] and [16]. Note that the different formulations of the RMAB problem in [2] and [16] are the motivating examples of our work, in which a nuance on the reward function leads to totally contrary results on the optimality of the myopic sensing policy, as summarized in Section III.

Consider a synchronously slotted cognitive radio communication system where an unlicensed secondary user can opportunistically access a N i.i.d. channels partially occupied by the

$$\begin{cases} \Omega_{10}(t+1):\ \omega_l(t+1) = \tau_l(\omega_l(t)) = p_{11}^l,\ \omega_m(t+1) = \tau_m(\omega_m(t)) = p_{01}^m \\ \Omega_{01}(t+1):\ \omega_l(t+1) = \tau_l(\omega_l(t)) = p_{01}^l,\ \omega_m(t+1) = \tau_m(\omega_m(t)) = p_{11}^m. \end{cases} \tag{14}$$

licensed primary users. The state of each channel follows the Markov chain presented in Section III with the good (bad, respectively) state representing that the channel is unoccupied (occupied) by the primary user. At the beginning of each time slot, the secondary user selects a subset $\mathcal{A}$ of k channels to sense and seeks to maximize its reward over T slots. The works in [2] and [16] focus on two specific reward functions and study the optimality of the myopic sensing policy in maximizing the aggregated reward.

In [2], the secondary user gets one unit of reward by accessing an unoccupied channel. Its objective is thus to find as many good channels as possible so as to maximize the throughput given that it can transmit on all the good channels. Formally, the expected slot reward function is $F_1(\Omega) \triangleq \sum_{i\in\mathcal{A}} \omega_i$, which is a regular and linear function. Noticing that in this case of N i.i.d. Markov channels, $\Delta_{\max} = \Delta_{\min}$, it holds that if $\delta_p^{\max} < 0.5$ the second condition in Theorem 1 holds for all $\beta \leq 1$. The myopic sensing policy is optimal in this case. This result is coherent with that obtained in [2] with a more stringent condition on the optimality. This is due to the fact that the analysis in [2] on the homogeneous channels is no longer applicable in the heterogeneous case. The generic analysis presented in this paper thus covers the homogeneous case at the price of more stringent conditions.

In [16], the secondary user can only transmit on one channel (e.g., due to hardware constraints). As a result, to maximize its throughput, it aims at maximizing the probability of finding at least one good channel. Formally, the expected slot reward function is $F_2(\Omega) \triangleq 1-\prod_{i\in\mathcal{A}}(1-\omega_i)$, which is regular. To study the optimality of the myopic sensing policy in this context, we apply Theorem 1. If the initial belief value $\omega_i(1) = p_{11}/(p_{11}+p_{01})$ for all $i \in \mathcal{N}$, by Lemma 1, we can show that

$$p_{01} \leq \omega_i(t) \leq p_{11},\ t = 1, 2, \ldots, T.$$

In this example, $\Delta_{\max} = (1-p_{01})^{k-1}$, $\Delta_{\min} = (1-p_{11})^{k-1}$. It then follows from Theorem 1 that the myopic sensing policy is optimal if

$$(1-p_{11})^{k-1} > (1-p_{01})^{k-1} \sum_{i=1}^{T} \beta^i (p_{11}-p_{01})^i.$$

This result confirms the result obtained in [16] that the myopic sensing policy is not always optimal, and further extends it by giving a sufficient condition under which the myopic sensing policy is ensured to be optimal.

Despite the focus of this section in the domain of opportunistic communication, the problem formulation is applicable in many other fields. One such example is the jamming problem where the jammer is constraint to jam only k of N channels with Markovian traffic and aims at maximizing its utility which can be modeled by functions such as F_1 and F_2 depending on the particular system setting. Another example is the opportunistic multiuser scheduling problem under imperfect channel state information which, studied in [20], has similar mathematical structure to the RMAB problem.

VIII. Conclusion

We have investigated the optimality of the myopic policy in the RMAB problem, which is of fundamental importance in many engineering applications. We have developed three axioms characterizing a family of generic and practically important functions which we refer to as regular functions. By performing a mathematical analysis based on the developed axioms, we have characterized the closed-form conditions under which the optimality of the myopic policy is ensured. The application of the derived results is demonstrated by analyzing a class of RMAB problems arising from multi-channel opportunistic access. As future work, a natural direction we are pursuing is to investigate the RMAB problem with multiple players with mutual conflicts and to study the structure and optimality of the myopic policy in that context.

Appendix A
Proof of Lemma 2

The lemma holds trivially for slot T noticing that $W_T(\Omega(T)) = F(\omega_1(T), \ldots, \omega_k(T))$, which is a regular function and is thus symmetrical.

We now show that $W_t^{\mathcal{A}}(\Omega(t))$ is symmetrical for $t < T$. Noticing the form of $W_t^{\mathcal{A}}(\Omega(t))$ given in (7), it suffices to show that $\Gamma(\Omega(t))$ is symmetrical in any $l, m \in \mathcal{A}(t)$ and any $l, m \in \mathcal{N} \setminus \mathcal{A}(t)$. We distinguish the following two cases:

- Case 1: $l, m \in \mathcal{A}(t)$;
- Case 2: $l, m \in \mathcal{N} \setminus \mathcal{A}(t)$.

For the first case, by rewriting $\Gamma(\Omega(t))$ in (7) and developing $\omega_l(t+1)$ and $\omega_m(t+1)$ in $\Omega(t+1)$, we have

$$\begin{aligned}
&\Gamma(\Omega_{l,m}(t)) \\
&\quad = \sum_{\mathcal{E}\subseteq\mathcal{A}(t)} \prod_{i\in\mathcal{E}} \omega_i(t) \\
&\qquad \times \prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}} (1-\omega_j(t)) W_{t+1}(\Omega(t+1)) \\
&\quad = \omega_l(t)\omega_m(t) \sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l,m\}} \prod_{i\in\mathcal{E}} \omega_i(t) \\
&\qquad \times \prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l,m\}} (1-\omega_j(t)) W_{t+1}(\Omega_{1,1}(t+1)) \\
&\qquad + (1-\omega_l(t))(1-\omega_m(t)) \sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l,m\}} \prod_{i\in\mathcal{E}} \omega_i(t) \\
&\qquad \times \prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l,m\}} (1-\omega_j(t)) W_{t+1}(\Omega_{0,0}(t+1)) \\
&\qquad + \omega_l(t)(1-\omega_m(t)) \sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l,m\}} \prod_{i\in\mathcal{E}} \omega_i(t) \\
&\qquad \times \prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l,m\}} (1-\omega_j(t)) W_{t+1}(\Omega_{1,0}(t+1)) \\
&\qquad + (1-\omega_l(t))\omega_m(t) \sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l,m\}} \prod_{i\in\mathcal{E}} \omega_i(t) \\
&\qquad \times \prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l,m\}} (1-\omega_j(t)) W_{t+1}(\Omega_{0,1}(t+1)),
\end{aligned}$$

where $\Omega_{l,m}(t) \triangleq [\omega_1(t), \ldots, \omega_l(t), \ldots, \omega_m(t), \ldots, \omega_N(t)]$, $\Omega_{a,b}(t+1)$ $(a, b \in \{0,1\})$ denotes the updated belief vector for slot $t+1$ under the belief vector $\Omega_{l,m}(t)$ with $\omega_l(t) = a$ and $\omega_m(t) = b$.

On the other hand, by exchanging l and m, following the similar notation and analysis, we have

$$\begin{aligned}
&\Gamma(\Omega_{m,l}(t))\\
&\quad=\sum_{\mathcal{E}\subseteq\mathcal{A}(t)}\prod_{i\in\mathcal{E}}\omega_i(t)\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}}(1-\omega_j(t))W_{t+1}(\Omega(t+1))\\
&\quad=\omega_m(t)\omega_l(t)\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l,m\}}\prod_{i\in\mathcal{E}}\omega_i(t)\\
&\qquad\times\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l,m\}}(1-\omega_j(t))W_{t+1}(\Omega_{1,1}(t+1))\\
&\qquad+(1-\omega_m(t))(1-\omega_l(t))\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l,m\}}\prod_{i\in\mathcal{E}}\omega_i(t)\\
&\qquad\times\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l,m\}}(1-\omega_j(t))W_{t+1}(\Omega_{0,0}(t+1))\\
&\qquad+\omega_m(t)(1-\omega_l(t))\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l,m\}}\prod_{i\in\mathcal{E}}\omega_i(t)\\
&\qquad\times\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l,m\}}(1-\omega_j(t))W_{t+1}(\Omega_{0,1}(t+1))\\
&\qquad+(1-\omega_m(t))\omega_l(t)\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l,m\}}\prod_{i\in\mathcal{E}}\omega_i(t)\\
&\qquad\times\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l,m\}}(1-\omega_j(t))W_{t+1}(\Omega_{1,0}(t+1))
\end{aligned}$$

It can be noticed that $\Gamma(\Omega_{l,m})=\Gamma(\Omega_{m,l})$ holds in this case.

For the second case, noticing that $l,m\in\mathcal{N}\setminus\mathcal{A}(t)$, we have

$$\Gamma(\Omega_{l,m}(t))=\sum_{\mathcal{E}\subseteq\mathcal{A}(t)}\prod_{i\in\mathcal{E}}\omega_i(t)\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}}(1-\omega_j(t))W_{t+1}(\Omega(t+1)).$$

Noticing that neither channel l nor channel m is sensed in slot t and that from slot $t+1$ to T, the user senses the k best channels, following the update (1), after sorting the elements in descending order, $\Omega_{m,l}(t)$ and $\Omega_{l,m}(t)$ generate the same belief vector $\Omega(t+1)$. It then follows that $\Gamma(\Omega_{m,l}(t))=\Gamma(\Omega_{l,m}(t))$.

Combining the results in both cases, it holds that $\Gamma(\Omega(t))$ is symmetrical. Hence, $W_t^{\mathcal{A}}(\Omega(t))$ is symmetrical, thus concluding the proof of Lemma 2.

Appendix B
Proof of Lemma 4

We prove the lemma by backward induction. For slot T, it is straightforward to check that (10) and (11) hold. We now prove (12). To this end, noticing that for $l\in\mathcal{A}'$, $l\notin\mathcal{A}$, $\mathcal{A}$ and $\mathcal{A}'$ differ in exactly one channel, let $m\in\mathcal{A}$ denote this channel. It follows from the definition of the myopic sensing policy that $\omega_l'\geq\omega_m\geq\omega_l$. We then have

$$\begin{aligned}
0&\leq(\omega_l'-\omega_m)\Delta_{\min}\leq W_T^{\mathcal{A}'}(\Omega_l')-W_T^{\mathcal{A}}(\Omega_l)\\
&\leq(\omega_l'-\omega_m)\Delta_{\max}\leq(\omega_l'-\omega_l)\Delta_{\max}.
\end{aligned}$$

Therefore, (12) holds for slot T.

Assume that Lemma 4 holds for $T,\ldots,t+1$, we now prove that it holds for slot t.

We first prove (10). By rewriting $\Gamma(\Omega)$ in (7) and developing $\omega_l(t+1)$ in $\Omega(t+1)$, we have

$$\begin{aligned}
&\Gamma(\Omega_l(t))\\
&\quad=\sum_{\mathcal{E}\subseteq\mathcal{A}(t)}\prod_{i\in\mathcal{E}}\omega_i(t)\\
&\qquad\times\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}}(1-\omega_j(t))W_{t+1}(\Omega(t+1))\\
&\quad=\omega_l(t)\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l\}}\prod_{i\in\mathcal{E}}\omega_i(t)\\
&\qquad\times\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l\}}(1-\omega_j(t))W_{t+1}(\Omega_1(t+1))\\
&\qquad+(1-\omega_l(t))\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l\}}\prod_{i\in\mathcal{E}}\omega_i(t)\\
&\qquad\times\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l\}}(1-\omega_j(t))W_{t+1}(\Omega_0(t+1))\\
&\quad=\left[\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l\}}\prod_{i\in\mathcal{E}}\omega_i(t)\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l\}}(1-\omega_j(t))\right]\\
&\qquad\times[\omega_l(t)W_{t+1}(\Omega_1(t+1))\\
&\qquad+(1-\omega_l(t))W_{t+1}(\Omega_0(t+1))]\\
&\quad=\left[\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l\}}\prod_{i\in\mathcal{E}}\omega_i(t)\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l\}}(1-\omega_j(t))\right]\\
&\qquad\times[W_{t+1}(\Omega_0(t+1))\\
&\qquad+\omega_l(t)(W_{t+1}(\Omega_1(t+1))-W_{t+1}(\Omega_0(t+1)))]
\end{aligned}$$

where $\Omega_a(t+1)$ $(a\in\{0,1\})$ denotes the updated belief vector for slot $t+1$ under the belief vector $\Omega_l(t)$ with $\omega_l(t)=a$.

By similar analysis on $\Gamma(\Omega_l')$, we have

$$\begin{aligned}
&\Gamma(\Omega_l'(t))\\
&\quad=\left[\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l\}}\prod_{i\in\mathcal{E}}\omega_i(t)\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l\}}(1-\omega_j(t))\right]\\
&\qquad\times[W_{t+1}(\Omega_0(t+1))\\
&\qquad+\omega_l'(t)(W_{t+1}(\Omega_1(t+1))-W_{t+1}(\Omega_0(t+1)))].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\Gamma(\Omega_l'(t))-\Gamma(\Omega_l(t))\\
&\quad=\left[\sum_{\mathcal{E}\subseteq\mathcal{A}(t)\setminus\{l\}}\prod_{i\in\mathcal{E}}\omega_i(t)\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}\setminus\{l\}}(1-\omega_j(t))\right]\\
&\qquad\cdot[\omega_l'(t)-\omega_l(t)][W_{t+1}(\Omega_1(t+1))-W_{t+1}(\Omega_0(t+1))].
\end{aligned}$$

Let $\mathcal{A}_0$ and $\mathcal{A}_1$ denote the set of channels sensed in slot $t+1$ based on the myopic policy (the set of k best channels) with the belief vector $\Omega_1(t+1)$ and $\Omega_0(t+1)$, it can be noted that $\Omega_1(t+1)$ and $\Omega_0(t+1)$ differ in one element ($\omega_l(t+1)=p_{11}^l$ in $\Omega_1(t+1)$ and p_{01}^l in $\Omega_0(t+1)$). Hence, $\mathcal{A}_0$ and $\mathcal{A}_1$ differ in at most one element. We distinguish two cases:

1) $\mathcal{A}_0 = \mathcal{A}_1$: for this case, it follows from the induction of (10) and (11) that

$$\begin{aligned}
&\left(p_{11}^l - p_{01}^l\right)\Delta_{\min} \\
&\quad \le W_{t+1}(\Omega_1(t+1)) - W_{t+1}(\Omega_0(t+1)) \\
&\quad \le \left(p_{11}^l - p_{01}^l\right)\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i,\ l \in \mathcal{A}_1, \\
&0 \le W_{t+1}(\Omega_1(t+1)) - W_{t+1}(\Omega_0(t+1)) \\
&\quad \le \delta_p^{\max}\left(p_{11}^l - p_{01}^l\right)\Delta_{\max}\sum_{i=0}^{T-t-2}\beta^i\left(\delta_p^{\max}\right)^i,\ l \notin \mathcal{A}_1.
\end{aligned}$$

2) $\mathcal{A}_1 \neq \mathcal{A}_0$: for this case, we further distinguish the following two subcases:
   a) $l \in \mathcal{A}_0$ but $l \notin \mathcal{A}_1$: for this subcase, there must exist $m \in \mathcal{N}$ such that $m \notin \mathcal{A}_0$ but $m \in \mathcal{A}_1$. Since the myopic sensing policy consists of choosing the k best channels, it holds that (1) $\omega_m \ge p_{11}^l$ as m is chosen in $\mathcal{A}_1$ but l is not and (2) $p_{01}^l \ge \omega_m$ as l is chosen in $\mathcal{A}_0$ but m is not. This contradicts with $p_{11}^l > p_{01}^l$ and implies that this subcase is impossible to happen.
   b) $l \in \mathcal{A}_1$ but $l \notin \mathcal{A}_0$: for this subcase, it follows from the induction of (12) that

$$\begin{aligned}
&0 \le W_{t+1}(\Omega_1(t+1)) - W_{t+1}(\Omega_0(t+1)) \\
&\quad \le (p_{11}^l - p_{01}^l)\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i.
\end{aligned}$$

Combing the analysis of Case 1 and Case 2, we have

$$\begin{aligned}
&0 \le \Gamma(\Omega_l'(t)) - \Gamma(\Omega_l(t)) \\
&\quad \le (\omega_l' - \omega_l)\left(p_{11}^l - p_{01}^l\right)\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i.
\end{aligned}$$

Noticing (7) that $W_t(\Omega_l'(t)) - W_t(\Omega_l(t)) = F(\Omega_l'(t)) - F(\Omega_l(t)) + \beta(\Gamma(\Omega_l'(t)) - \Gamma(\Omega_l(t)))$, we have

$$\begin{aligned}
(\omega_l' - \omega_l)\Delta_{\min} &\le W_t(\Omega_l'(t)) - W_t(\Omega_l(t)) \\
&\le (\omega_l' - \omega_l)\Delta_{\max} + \beta(\omega_l' - \omega_l) \\
&\quad \times \left(p_{11}^l - p_{01}^l\right)\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i \\
&\le (\omega_l' - \omega_l)\Delta_{\max} + \beta(\omega_l' - \omega_l) \\
&\quad \times \delta_p^{\max}\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i \\
&= (\omega_l' - \omega_l)\Delta_{\max}\sum_{i=0}^{T-t}\beta^i(\delta_p^{\max})^i.
\end{aligned}$$

We thus complete the proof of (10) for slot t.

We then prove (11). Noticing $l \notin \mathcal{A}'$ and $l \notin \mathcal{A}$, we have

$$\begin{cases}
\Gamma(\Omega_l(t)) = \displaystyle\sum_{\mathcal{E}\subseteq\mathcal{A}(t)}\prod_{i\in\mathcal{E}}\omega_i(t)\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}}(1-\omega_j(t))W_{t+1}(\Omega_l(t+1)) \\
\Gamma(\Omega_l'(t)) = \displaystyle\sum_{\mathcal{E}\subseteq\mathcal{A}(t)}\prod_{i\in\mathcal{E}}\omega_i(t)\prod_{j\in\mathcal{A}(t)\setminus\mathcal{E}}(1-\omega_j(t))W_{t+1}(\Omega_l'(t+1))
\end{cases}$$

where $\Omega_l(t+1)$ and $\Omega_l'(t+1)$ are the belief vector for slot $t+1$ generated by $\Omega_l(t)$ and $\Omega_l'(t)$ based on the belief update (1). We distinguish four cases.

1) $l \notin \mathcal{A}(r)$ and $l \notin \mathcal{A}'(r)$ for $t+1 \le r \le T$: i.e., l is not chosen from the slot $t+1$ to T in either scenario. For this case, it is straightforward to check that $\Gamma(\Omega_l'(t)) = \Gamma(\Omega_l(t))$, and furthermore $W_t(\Omega_l') = W_t(\Omega_l)$.
2) There exists $t_0 \ge t+1$ such that $l \notin \mathcal{A}(r)$ and $l \notin \mathcal{A}'(r)$ for $t+1 \le r < t_0$, $l \in \mathcal{A}(t_0)$ and $l \in \mathcal{A}'(t_0)$. For this case, it follows from the induction of (10) that

$$\begin{aligned}
&\left(p_{11}^l - p_{01}^l\right)^{t_0-t}(\omega_l'(t) - \omega_l(t))\Delta_{\min} \\
&\quad = (\omega_l'(t_0) - \omega_l(t_0))\Delta_{\min} \le W_{t_0}(\Omega_{A'(t_0)}) - W_{t_0}(\Omega_{A(t_0)}) \\
&\quad \le (\omega_l'(t_0) - \omega_l(t_0))\Delta_{\max}\sum_{i=0}^{T-t_0}\beta^i\left(\delta_p^{\max}\right)^i \\
&\quad = \left(p_{11}^l - p_{01}^l\right)^{t_0-t}(\omega_l'(t) - \omega_l(t))\Delta_{\max}\sum_{i=0}^{T-t_0}\beta^i\left(\delta_p^{\max}\right)^i.
\end{aligned}$$

Noticing that in this case, $\mathcal{A}(r) = \mathcal{A}'(r)$ for $t+1 \le r \le t_0$ and that $t+1 \le t_0$, it holds that

$$\begin{aligned}
&\left(p_{11}^l - p_{01}^l\right)(\omega_l'(t) - \omega_l(t))\Delta_{\min} \\
&\quad \le W_{t+1}(\Omega'(t+1)) - W_{t+1}(\Omega(t+1)) \\
&\quad \le \left(p_{11}^l - p_{01}^l\right)(\omega_l'(t) - \omega_l(t))\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i.
\end{aligned}$$

It then follows from (7) that

$$\begin{aligned}
&\beta\left(p_{11}^l - p_{01}^l\right)(\omega_l'(t) - \omega_l(t))\Delta_{\min} \\
&\quad \le W_t^{\mathcal{A}'}(\Omega_l') - W_t^{\mathcal{A}}(\Omega_l) \\
&\quad \le \beta\left(p_{11}^l - p_{01}^l\right)(\omega_l'(t) - \omega_l(t))\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i.
\end{aligned}$$

3) There exists $t_0 \ge t+1$ such that $l \notin \mathcal{A}(r)$ and $l \notin \mathcal{A}'(r)$ for $t+1 \le r < t_0$, $l \in \mathcal{A}'(t_0)$ and $l \notin \mathcal{A}(t_0)$. For this case, by the induction (12),

$$\begin{aligned}
&0 \le W_{t_0}(\Omega_{A'(t_0)}) - W_{t_0}\left(\Omega_{A(t_0)}\right) \\
&\quad \le (\omega_l'(t_0) - \omega_l(t_0))\Delta_{\max}\sum_{i=0}^{T-t_0}\beta^i\left(\delta_p^{\max}\right)^i.
\end{aligned}$$

It then follows from $t_0 \ge t+1$ and (1) that

$$\begin{aligned}
&0 \le W_{t+1}(\Omega'(t+1)) - W_{t+1}(\Omega(t+1)) \\
&\quad \le (\omega_l'(t+1) - \omega_l(t+1))\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i \\
&\quad = (\omega_l'(t) - \omega_l(t))\left(p_{11}^l - p_{01}^l\right)\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&0 \le W_t^{\mathcal{A}'}(\Omega_l') - W_t^{\mathcal{A}}(\Omega_l) \\
&\quad \le \beta(\omega_l'(t) - \omega_l(t))\left(p_{11}^l - p_{01}^l\right)\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i \\
&\quad \le (\omega_l'(t) - \omega_l(t))\delta_p^{\max}\Delta_{\max}\sum_{i=0}^{T-t-1}\beta^i\left(\delta_p^{\max}\right)^i.
\end{aligned}$$

4) There exists $t_0 \ge t+1$ such that $l \notin \mathcal{A}(r)$ and $l \notin \mathcal{A}'(r)$ for $t+1 \le r < t_0$, $l \notin \mathcal{A}'(t_0)$ and $l \in \mathcal{A}(t_0)$. For this case, it holds that $\mathcal{A}(r) = \mathcal{A}'(r)$ for $t \le r \le t_0 - 1$ and $\mathcal{A}(t_0)$ and $\mathcal{A}'(t_0)$ differ in one element, assume that $m \in \mathcal{A}'(t_0)$ and $m \notin \mathcal{A}(t_0)$. It follows from the definition

of the myopic sensing policy that $\omega_l(t_0) \geq \omega_m(t_0)$ and $\omega_l'(t_0) \leq \omega_m(t_0)$, which leads to contradiction since $\omega_l'(t+1) = p_{11}^l > \omega_l(t+1) = p_{01}^l$ leads to $\omega_l'(t_0) > \omega_l(t_0)$ following Lemma 1. This case is thus impossible.

Combing the analysis of the four cases, we complete the proof of (11) for slot t.

We now prove (12). For this case, there exists $m \in \mathcal{N}$ with $\omega_l' \geq \omega_m \geq \omega_l$ such that $\mathcal{A}$ and $\mathcal{A}'$ differ in one element: $\omega_l' \in \mathcal{A}'$ and $\omega_m \in \mathcal{A}$.[8] We have

$$\begin{aligned} &W_t^{\mathcal{A}'}(\Omega_l') - W_t^{\mathcal{A}}(\Omega_l) \\ &\quad = W_t^{\mathcal{A}'}(\omega_1, \ldots, \omega_l', \ldots, \omega_N) \\ &\qquad - W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_l, \ldots, \omega_N) \\ &\quad = W_t^{\mathcal{A}'}(\omega_1, \ldots, \omega_l', \ldots, \omega_N) \\ &\qquad - W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_l = \omega_m, \ldots, \omega_N) \\ &\qquad + W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_l = \omega_m, \ldots, \omega_N) \\ &\qquad - W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_l, \ldots, \omega_N). \end{aligned}$$

On one hand, we have shown that (10) holds for slot t. Hence, it holds that

$$\begin{aligned} 0 &\leq (\omega_l' - \omega_m)\Delta_{\min} \leq W_t^{\mathcal{A}'}(\omega_1, \ldots, \omega_l', \ldots, \omega_N) \\ &\quad - W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_l = \omega_m, \ldots, \omega_N) \\ &\leq (\omega_l' - \omega_m)\Delta_{\max} \sum_{i=0}^{T-t} \beta^i \left(\delta_p^{\max}\right)^i. \end{aligned}$$

On the other hand, we have shown that (11) holds for slot t. Hence, it holds that

$$\begin{aligned} 0 &\leq W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_l = \omega_m, \ldots, \omega_N) \\ &\quad - W_t^{\mathcal{A}}(\omega_1, \ldots, \omega_l, \ldots, \omega_N) \\ &\leq \delta_p^{\max}(\omega_m - \omega_l)\Delta_{\max} \sum_{i=0}^{T-t-1} \beta^i \left(\delta_p^{\max}\right)^i \\ &\leq (\omega_m - \omega_l)\Delta_{\max} \sum_{i=0}^{T-t} \beta^i \left(\delta_p^{\max}\right)^i. \end{aligned}$$

It then follows that

$$\begin{aligned} 0 &\leq W_t^{\mathcal{A}'}(\Omega_l') - W_t^{\mathcal{A}}(\Omega_l) \\ &\leq (\omega_l' - \omega_l)\Delta_{\max} \sum_{i=0}^{T-t} \beta^i \left(\delta_p^{\max}\right)^i. \end{aligned}$$

Thus, we complete the proof of (12).

Combining the above analysis, Lemma 4 is proven.

REFERENCES

[1] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance," *IEEE Trans. Wireless Commun.*, vol. 7, no. 3, pp. 5413–5440, 2008.

[2] S. Ahmad and M. Liu, "Multi-channel opportunistic access: A case of restless bandits with multiple plays," presented at the Allerton Conf., Monticello, IL, 2009.

[3] Q. Liu, K. Wang, and L. Chen, "On optimality of greedy policy for a class of standard reward function of restless multi-armed bandit problem," Computing Research Repository (CoRR), [Online]. Available: http://arxiv.org/abs/1104.5391 2011

[4] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, 1952.

[5] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Statist. Soc.*, ser. B, vol. 41, no. 2, pp. 148–177, 1979.

[6] P. Whittle, "Multi-armed bandits and the Gittins index," *J. Roy. Statist. Soc.*, ser. B, vol. 42, no. 2, pp. 143–149, 1980.

[7] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," *Math. Oper. Res.*, vol. 24, no. 2, pp. 293–305, 1999.

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

[9] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *J. Appl. Probab.*, vol. 27, no. 1, pp. 637–648, 1990.

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

[11] S. Guha and K. Munagala, "Approximation algorithms for partial-information based stochastic control with Markovian rewards," presented at the IEEE Symp. Found. Comput. Sci. (FOCS), Providence, RI, 2007.

[12] S. Guha and K. Munagala, "Approximation algorithms for restless bandit problems," presented at the ACM-SIAM Symp. Discrete Algorithms (SODA), New York, 2009.

[13] D. Bertsimas and J. E. Nino-Mora, "Restless bandits, linear programming relaxations, and a primal-dual heuristic," *Oper. Res.*, vol. 48, no. 1, pp. 80–90, 2000.

[14] T. Javidi, B. Krishnamachari, Q. Zhao, and M. Liu, "Optimality of myopic sensing in multi-channel opportunistic access," presented at the IEEE ICC, Beijing, China, May 2008.

[15] S. H. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, 2009.

[16] K. Wang and L. Chen, "On the optimality of myopic sensing in multi-channel opportunistic access: The case of sensing multiple channels," *IEEE Trans. Commun.*, 2011 [Online]. Available: http://arxiv.org/abs/1103.1784, submitted for publication

[17] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," presented at the INFOCOM, Shanghai, China, Apr. 2011.

[18] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao, "The non-Bayesian restless multi-armed bandit: A case of near-logarithmic regret," presented at the IEEE International Conf. Acoust., Speech, Signal Processing (ICASSP), Prague, Czech, May 2011.

[19] H. Liu, K. Liu, and Q. Zhao, "Logarithmic weak regret of non-Bayesian restless multi-armed bandit," presented at the IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Prague, Czech, May 2011.

[20] S. Murugesan, P. Schniter, and N. B. Shroff, "Opportunistic scheduling using ARQ feedback in multi-cell downlink," presented at the Asilomar Conf., Pacific Grove, CA, Nov. 2010.

[8] In case where $\omega_m = \omega_l$, it follows from the tie breaking rule of the myopic sensing policy that channel m has the priority over l.

Kehao Wang received the B.S. degree in electrical engineering and the M.S. degree in communication and information systems from Wuhan University of Technology, Wuhan, China, in 2003 and 2006, respectively.

He is currently working towards the Ph.D. degree in the Department of Computer Science, the University of Paris-Sud XI, Orsay, France, and in the School of Information Engineering, Wuhan University of Technology, Wuhan, China. His research interests are cognitive radio networks, wireless network resource management, and data hiding.

Lin Chen received the B.E. degree in radio engineering from Southeast University, China, in 2002, the Engineer Diploma from Telecom ParisTech, Paris, France, in 2005., and the M.S. degree of networking from the University of Paris 6, France.

He currently works as Assistant Professor in the Department of Computer Science of the University of Paris-Sud XI, France. His main research interests include modeling and control for wireless networks, security and cooperation enforcement in wireless networks, and game theory.