One Step Beyond Myopic Probing Policy: A Heuristic Lookahead Policy for Multi-Channel Opportunistic Access

Kehao Wang, Lin Chen, Quan Liu, Wei Wang, Fangmin Li

***Abstract*—In this paper, we consider the probing order and stopping problem arising from the identification of spectrum holes in multi-channel cognitive radio networks, in which a secondary user (SU) seeks to maximize the probability of finding an available channel while minimizing the related probing cost within a long time horizon. This problem can be casted into a restless multi-armed bandit (RMAB) problem, which is proved to be PSPACE-hard. The key point of this problem is the trade-off between *exploitation*, in which the SU stops probing once an available channel is identified, and *exploration*, in which the SU continues to probe new channels even after identifying an available channel in order to learn the system state to reduce probing cost in the future. To strike a desirable balance between the two conflicting objectives, we develop a heuristic channel probing policy, termed as ν-step lookahead policy, in which the SU makes its decision based on the prediction of system state within the future ν steps, with ν being a tunable parameter. We conduct an analytical study on the structure of the proposed ν-step lookahead policy, and demonstrate how the policy can be implemented with linear complexity with respect to the number of channels in the system via a detailed analysis on the 1-step lookahead policy. Numerical experiments between ν-step lookahead policy and myopic probing policy on two representative network scenarios demonstrate the effectiveness of the proposed ν-step lookahead policy.**

***Index Terms*—Opportunistic spectrum access (OSA), cognitive radio, restless multi-armed bandit (RMAB), myopic policy, heuristic policy, complexity**

I. Introduction

A. Background

With the rapid growth of wireless communications in recent years, so far almost all the exploitable spectra have been allocated for various wireless applications in different regions. Meanwhile, severe underutilization of the licensed spectrums at certain time or location has been observed by measurements of wireless spectrum usage [1], [2]. This observation has motivated the idea of opportunistic spectrum access (OSA), under which the unlicensed secondary users (SUs) can utilize the spectrum which is not occupied by the licensed primary users (PUs). As an enabling technique in implementing OSA paradigm, cognitive radio (CR) [3]–[6], allowing SUs to probe spectrum, analyze spectrum statistics, and adjust their transmissions according to the time-varying environment, has been claimed to be a hopeful solution to the conflicts between spectrum demand growth and spectrum underutilization [7]–[10].

To utilize channel opportunities of the PUs, the SU should probe the channels before transmission in order to determine whether the PUs are transmitting over them. In fact, the SU cannot probe all the channels each time due to resource constraint, i.e., hardware capability, energy consumption, and probing cost. Hence, the SU should decide which channels to probe in each time slot in order to utilize the spectrum opportunities as fully as possible. This decision process can be enhanced if relevant statistical information about these channels is taken into account. For example, with capability of probing multiple channels in a slot, the SU can probe the channels sequentially according to certain probing order (e.g., the descending order of availability probabilities of the channels) to gather information, and stops at a channel based on certain criterion (e.g., when successfully obtaining an available channel) to enter data transmission phase which takes the remainder of the slot duration. Therefore, the probing order and the stopping criterion should be jointly tuned to maximize the long-term objective of the SU (e.g., maximizing the average long-term throughput).

There are a body of related works in the literature addressing the probing order and optimal stopping problem where an SU continuously probes a set of selected channels until one channel is identified to be unoccupied. In this aspect, most of works focuses on the model of memoryless channel. In [11], the authors derived the optimal channel-sensing strategy for a single-user case with an assumption of recall and guess where the former allows the SU to access a previously sensed channel while the latter permits the SU to access a channel that has not yet been sensed. In [12], the authors showed that obtaining the optimal sensing strategy is computationally prohibitive, and then proposed polynomial complexity algorithms to ensure the

The work of K. Wang was supported by the National Natural Science Foundation of China under Grant 61303027, the China Postdoctoral Science Foundation 2013M531753, 2014T70748, and the Fundamental Research Funds for the Central Universities (WUT:2014-IV-067). The work of L. Chen was supported by the Agence Nationale de la Recherche (ANR) under grant Green-Dysan (ANR-12-IS03). The work of W. Wang was supported by the National Natural Science Foundation of China under Grant 61261130585. The work of F. Li was supported by the National Natural Science Foundation of China under Grant 61373024.

K. Wang is with the Key Laboratory of Fiber Optic Sensing Technology and Information Processing, Ministry of Education, Wuhan University of Technology, 430070 Wuhan, China (email: kehao.wang@whut.edu.cn).

Q. Liu and F. Li are with School of Information Engineering, Wuhan University of University, 430070 Wuhan, China (email: {quanliu,lifangmin}@whut.edu.cn).

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

W. Wang is with Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China (email: wangw@zju.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2014.2359917, IEEE Transactions on Wireless Communications

obtained reward at most ε less than that of the optimal strategy. In [13], the authors studied the optimal sensing order problem for a single-user case where neither recall nor guess is allowed with a simple sensing order for some special scenarios. In [14], the authors considered opportunistic channel sensing and access in cognitive radio networks when the sensing is imperfect and an SU can access up to a limited number of channels at a time, and derived asymptotically logarithmic regret performance for different scenarios. In [15], the optimal channel sensing order was derived for channels with homogeneous capacities, and the optimal sensing order problem for channels with heterogeneous capacities was shown to be NP-hard. In [16], a novel sequential sensing scheme was proposed based on suprathreshold stochastic resonance (SSR) to reduce the average sample number in a single sensing node. In [17], the authors focused on finding the appropriate sensing frequency during an SU's active data transmission on a licensed channel, and proposed detection schemes addressing channel state change and anomalous data to facilitate short-term sensing adaptation to the variations in sensed data. In [18], the authors studied sequential channel sensing and probing to resolve spectrum uncertainty and to search for good transmission opportunities for realtime traffic in CR networks by optimal stopping theory. In [19], a novel channel sensing and accessing strategy was proposed carefully to balance the channel statistics exploration and multichannel diversity exploitation such that the regret is in optimal logarithmic rate in time and polynomial in the number of channels.

B. Restless Multi-armed Bandit and Myopic Policy

The channel probing problem addressed in this paper is theoretically grounded in the restless multi-armed bandit (RMAB) problem, one of the most well-known generalizations of the classic multiarmed bandit (MAB) problem. The RMAB 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. Despite the significant research efforts in the field starting from almost half a century ago, the RMAB problem in its generic form still remains open and is notoriously reputed as a hard problem in the field of decision and control. 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.

However, a number of research works have recently appeared on a single-user case [21]–[26] under the restless multi-armed bandit (RMAB) framework, and have shown that the myopic policy, by sensing channels with the highest available probabilities, is optimal under some mild conditions when the SU senses k out of N channels each time. Ahmad and Liu *et al.* [22] derived the optimality of the myopic sensing policy for the positively correlated i.i.d. channels when the user is limited to access one channel (i.e., $k = 1$) each time, and further extended the optimality to the case of sensing multiple i.i.d. channels ($k > 1$) [23] for the scenario where the user gets one unit of reward for each channel sensed good. In [25], we extended i.i.d. channels [22] to non i.i.d. ones, and focused on a family of generic and important utility functions, termed as *regular* function, and derived closed-form conditions under which the myopic sensing policy is ensured to be optimal. For the imperfect sensing case, Liu and Zhao *et al.* [24] proved the optimality of the myopic policy for the case of two channels with a particular utility function and conjectured it for arbitrary N. Further, the optimality of the myopic policy was extended to access a fixed number k of channels in N homogeneous channels [26] and N heterogeneous channels [27], respectively.

In the above literature [21]–[27], a fixed number of channels are assumed to be sensed (or probed), and thus, the sensing (or probing) cost is constant at each slot and is omitted in those problems without losing performance. In practice, the number of channels an SU can sense varies according to its requirement and capability, and is upper-bounded by its resource constraint or probing limit. Under this context, the probing cost does not keep constant in each slot and cannot be omitted in the probing process. Thus, a natural question arises: how many channels should the SU probe each time taking into account the probing cost, in other words, when should the SU stop probing in each probing process? It is insightful to note that the key point to answer the question is the well-known trade-off between exploitation and exploration. Specifically, *exploitation* refers to starting immediate transmission over the channel which is probed to be available, without probing other channels, in order to leave more time for data transmission. On the other hand, *exploration* requires learning the system state by probing additional channels even after identifying an available channel, and obviously, *exploration* brings benefit in future slots because of the obtained channel information.

C. Our Contributions

In this paper, we develop a decision-making framework to analyze the probing channel problem with a variable number of channels when the system state can only be *partially* and *imperfectly* observable. Specifically, the SU can only probe k out of N ($k < N$) channels at each slot and, as a consequence, obtain partial system information. In practice, the probing error is unavoidable for the complicatedly varying conditions of channels from noise, fading, occupation etc, and thus, the SU can only obtain imperfect system information taking into account false alarm and missing detection.

Due to the hardness of the RMAB problem, each small step towards characterizing the optimal or near-optimal strategies is important and has its theoretic and engineering merits. Concerning the Markovian formulation of the RMAB problem addressed in the paper that captures the time dependence and evolution of channels, the state-of-the-art research strand is to seek simple myopic policies and study their optimality. We have also done some work [25]–[27] on establishing closed-form conditions under which the myopic policy is optimal. In this paper, we propose a heuristic channel probing policy, termed as ν-step lookahead policy, in which the SU makes its decision based on the prediction of the system states in the future ν slots, and then conduct an analytical study on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2014.2359917, IEEE Transactions on Wireless Communications

the structure of the proposed ν-step lookahead policy, and finally demonstrate how the policy can be implemented with linear complexity in terms of the number of channels in the system. Compared with the existing work on myopic probing policies, the novelty of the proposed heuristic algorithm and the correspondent analysis can be summarized as follows:

- In existing literature, the myopic policy requires that the number of probed channels to be fixed at each slot to establish the optimality result. In this work, we relax this constraint by investigating the case where the user can probe up to k channels each time. We note that such a small relaxation makes the problem much more difficult.
- Existing myopic policy maximizes only the immediate reward. In the heuristic algorithm we propose, the user has the choice anticipating ν slots by maximizing the aggregated ν-step rewards, where ν is a tunable parameter, by tuning which the user can achieve a desirable trade-off between optimality and complexity. We theoretically show how such ν-step lookahead policy can be implemented.
- We also incorporate the imperfect probing with non-identical probing error rate at each channel to make the analysis and results more generic.

The rest of this paper is organized as follows: Our problem is formulated Section II. In Section III, we propose an easily heuristic policy and derive the corresponding heuristic algorithm. Section IV presents the numerical experiment to support our claims. Finally, our conclusions are summarized in Section V.

II. Problem Formulation

In this section, we first introduce the system model, and then formulate the jointly optimal problem of channel probing order and stopping optimization into a RMAB one.

A. System Model

We consider a cognitive communication system in which an SU tries to access a set $\mathcal{N}$ of N primary channels, each channel $k \in \mathcal{N}$ given by a two state (0/occupied, 1/unoccupied) Markov chain with transition probabilities $\{p_{ij}^{(k)}\}_{i,j=0,1}$. The communication system is assumed to operate in a slotted fashion, and the time slots are indexed by t $(1 \leq t \leq T)$, where T is the time horizon of interest (until the SU gives up accessing the system). Specifically, we assume that channels go through state transition at the beginning of a slot. The length of each time slot is denoted as Δ, which is further divided into two parts: the probing phase and the transmission phase. Let $\delta = \alpha\Delta$ denote the time needed to probe one channel, the probing phase lasts $n\alpha\Delta$ if the user probes n channels, and the transmission phase consists of the rest of the time $(1-\alpha n)\Delta$.

The SU's objective is to maximize its throughput by choosing the appropriate set of channels and then probing them according to certain probing sequence. Let $\mathcal{A}(t)$ and $\mathcal{O}_A(t)$ denote the set of channels probed and the corresponding set of probed results, i.e., $\mathcal{O}_A(t) \triangleq \{O_i(t) \in \{1,0\}, i \in \mathcal{A}(t)\}$, by the SU at slot t. The SU is assumed to probe at most M $(1 \leq M < N)$ channels for the hardware limit and probing constraint, where $\alpha \leq \frac{1}{M}$ is required to ensure the existence of transmission phase. If at least one of the probed channel is probed to be unoccupied, the SU can successfully transmit one packet.[1]

Let $S_i(t)$ be the state of channel i at slot t. In our study, we take into consideration the imperfect probing which is characterized by the miss detection rate denoted as ζ_i and the false alarm rate denoted as ϵ_i, formally defined as follows:

$$\epsilon_i(t) \triangleq \Pr\{O_i(t)=0|S_i(t)=1\},$$
$$\zeta_i(t) \triangleq \Pr\{O_i(t)=1|S_i(t)=0\}.$$

Obviously, by imperfectly probing $|\mathcal{A}(t)|$ out of N channels at each slot t, the SU cannot observe the complete state information of the whole system. Hence, the SU has to infer the channel states from its past decision and observation history so as to make its future decision. Moreover, the current probing outcome further serves as statistics for 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 $0 \leq \omega_i(t) \leq 1$ is the conditional probability that channel i is not occupied.

Given the probing set $\mathcal{A}(t)$ and the detection outcomes $\mathcal{O}_A(t)$, the belief vector in $t+1$ slot can be updated recursively using Bayes Rule as shown in (1):

$$\omega_i(t+1) = \begin{cases} p_{11}^{(i)}, & i \in \mathcal{A}(t), O_i(t)=1 \\ \mathcal{T}_i(\varphi_i(\omega_i(t))), & i \in \mathcal{A}(t), O_i(t)=0 \\ \mathcal{T}_i(\omega_i(t)), & i \notin \mathcal{A}(t), \end{cases} \tag{1}$$

where,

$$\mathcal{T}_i(\omega_i(t)) \triangleq \omega_i(t)p_{11}^{(i)} + (1-\omega_i(t))p_{01}^{(i)}, \tag{2}$$
$$\varphi_i(\omega_i(t)) \triangleq \frac{\epsilon_i\omega_i(t)}{1-(1-\epsilon_i)\omega_i(t)}. \tag{3}$$

Note that the belief update under $O_i(t)=0$ results from the fact that the receiver cannot distinguish a failed transmission, i.e., collides with the primary traffic with probability $\zeta_i(1-\omega_i(t))$ from no transmission with probability $\epsilon_i\omega_i(t) + (1-\zeta_i)(1-\omega_i(t))$ [24].

B. Optimal Probing Order and Stopping Problem

We are interested in the SU's optimization problem to find a channel probing policy $\pi \triangleq [\pi_1, \cdots, \pi_T]$ that maximizes the expected accumulated discounted reward over a finite horizon, where, a probing policy π_t is defined as a mapping from the belief vector $\Omega(t)$ to $\mathcal{A}(t)$ at slot t:

$$\pi_t : \Omega(t) \to \mathcal{A}(t), 1 \leq |\mathcal{A}(t)| \leq M, t=1,2,\cdots,T.$$

Thus, the formal definition of the optimal probing problem $\mathbf{P}$ can be formulated as follows:

$$\mathbf{P}: \; \pi^* = \underset{\pi}{\operatorname{argmax}} \, \mathbb{E}_\pi \left[\sum_{t=1}^{T} \beta^{t-1} R\Big(\pi_t(\Omega(t)), \mathcal{O}_A(t)\Big) \middle| \Omega(1) \right], \tag{4}$$

[1] Our work can be extended to the case where the SU is equipped with more than one radio and can access multiple channels at a time.

where $R\Big(\pi_t(\Omega(t)), \mathcal{O}_A(t)\Big)$ is the SU's utility in slot t under the probing policy π_t with the observation set $\mathcal{O}_A(t)$ and the initial belief vector $\Omega(1)$[2], and $0 \le \beta \le 1$ is the discount factor characterizing the feature that the future rewards are less valuable than the immediate reward.

Generally, the RMAB problem with a fixed number of activated arms is proved to be PSPACE-hard [30] while the proposed problem $\mathbf{P}$ is a special RMAB problem with a variable number of arms to be activated. Hence, $\mathbf{P}$ is more complex than those RMAB problems with a fixed number of activated arms since the number of the probed channels at each slot is a random variable depending on the probing policy and the corresponding observation set. Meanwhile, the variable number of the probed channels, in return, has some impact on the channel probing.

C. When to Stop Probing?

Considering the exponential complexity of solving $\mathbf{P}$, a natural alternative to tackle $\mathbf{P}$ is to seek a myopic probing policy that maximizes the immediate reward [21]–[25], which corresponds to only focusing on exploitation while ignoring exploration from the probing order perspective. That is, the focus is how many channels to probe in each slot given the myopic probing order, i.e., probing channels according to the decreasing order of channel availability. In fact, the variable number of channels probed at each slot still reflects the intrinsic tradeoff between exploitation and exploration to some extent. The motivation of considering the myopic probing order is two-fold: 1) The myopic probing policy is ensured to be optimal under certain mild conditions in the literature [21]–[27]; 2) The myopic probing policy has a simple and robust structure which is easy to implement.

However, plenty of existing works on myopic policy of RMAB explicitly assume that the number of activated arms (corresponding to the number of channels to probe) is fixed. Actually, it is highly impossible for the SU to exactly probe a fixed number of channels in each slot due to the probing cost, i.e. energy or delay considered in this paper. Thus, the objective of the SU is to probe how many channels in each slot so as to maximize the expected aggregated reward.

For ease of presentation, assume that $\Omega(t)$ is sorted to $\omega_1(t) \ge \omega_2(t) \ge \cdots \ge \omega_N(t)$ in each slot t and then a channel list $l^0(t) \triangleq (1, 2, \cdots, N)$ which records channel index according to belief value[3]. Then the optimization problem on the number of channels to probe in each slot can be formulated as follows:

$$\mathbf{P_1}: \phi^* = \operatorname*{argmax}_{\phi} \mathbb{E}_\phi \left[\sum_{t=1}^{T} \beta^{t-1} R(n_t, \mathcal{O}_n(t)) \middle| \Omega(1) \right], \quad (5)$$

where, $\phi = [n_1, n_2, \cdots, n_T]$ and the first n_t channels are probed in slot t, i.e., $\mathcal{A}(t) = \{1, \cdots, n_t\}$.

It is insightful to note that $\mathbf{P_1}$ on the number of channels to probe hinges on the following tradeoff between exploitation and exploration: probing more channels can help SU learn and predict the future channel states, thus increasing the long-term reward, but at the price of sacrificing the reward at current slot since probing more channels reduces data transmission time, thus decreasing the throughput in the current slot.

Without loss of generality, we consider the following slot reward function $R(n_t, \mathcal{O}_n(t))$ in the normalized form:

$$R(n_t, \mathcal{O}_n(t)) = \begin{cases} 1 - C(n_t), & \text{if } \prod_{i=1}^{n_t} (1 - O_i(t)) = 0 \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

where $C(\cdot)$ is the monotonously increasing cost function, reflecting the time cost on channel probing and frequency switching. Specially, the first line of the RHS of (6) indicates that the SU can obtain a payoff $1 - C(n_t)$ as long as one of the first n_t channels are probed to be unoccupied, while the second line indicates the user obtains no payoff if none of the first n_t channels is probed to be unoccupied. By normalizing $\Delta = 1$, we have $C(n_t) = \alpha n_t$.

To better streamline our presentation, we introduce the pseudo cost function defined as follows:

$$\begin{aligned} q(n_t, \mathcal{O}_n(t)) &\triangleq 1 - R(n_t, \mathcal{O}_n(t)) \\ &= \begin{cases} C(n_t) = \alpha n_t, & \text{if } \prod_{i=1}^{n_t} (1 - O_i(t)) = 0 \\ C_0 = 1, & \text{otherwise.} \end{cases} \end{aligned} \quad (7)$$

Then the optimization problem $\mathbf{P_1}$ can be written as the following optimization problem $\mathbf{P_2}$:

$$\mathbf{P_2}: \phi^* = \operatorname*{argmin}_{\phi} \mathbb{E}_\phi \left[\sum_{t=1}^{T} \beta^{t-1} q(n_t, \mathcal{O}_n(t)) \middle| \Omega(1) \right]. \quad (8)$$

Remark. We would like to emphasize that the imperfect probing brings about the nonlinear propagation (i.e. (1)) of belief vector such that the LP relaxation in [28], [29] for perfect sensing cannot be adopted to provide guaranteed approximation ratio algorithms for the imperfect probing. Meanwhile, the proof concerning the PSPACE-Hardness of RMAB [30] shows that it is also PSPACE-Hard to justify whether the expected accumulated reward (or cost) of RMAB equals zero in the simplest scenario with two states. Hence, a practical and feasible alternative is to seek a simple and stable policy.

III. ν-Step Lookahead Policy

In the previous section, it was shown to be PSPACE-hard to obtain the optimal number of channels in the probing process for the proposed problem. As an alternative, we first analyze a feasible lower bound and upper one for the proposed problem, and then propose, based on the myopic policy, a heuristic policy (termed as ν-step lookahead policy) as well as the corresponding algorithm. Next, we take the case of $\nu = 1$ as an example to demonstrate how to calculate the relevant quantities in the proposed algorithm.

[2] If no information on the initial system state is available, each entry of $\Omega(1)$ can be set to the stationary distribution $\omega_0^{(i)} = \frac{p_{01}^{(i)}}{1+p_{01}^{(i)} - p_{11}^{(i)}}$, $1 \le i \le N$.

[3] The initial order of list is determined by the initial availability probability of each channel: $\omega_1(1) \ge \omega_2(1) \ge \cdots \ge \omega_N(1) \Rightarrow l^0(1) = (1, 2, \cdots, N)$.

A. Upper and Lower Bounds

Before giving the upper and lower bounds, we first state the following lemma which describes the structure of probing policy.

Lemma 1. *The SU should continue to probe new channel if all the probed channels are occupied.*

Proof: It is trivial to prove the lemma with noticing that by probing a new channel

- the cost for the current slot t will remain 1 according to (7) if the new channel is probed to be occupied, and will be smaller than 1 if the new channel is probed to be unoccupied;
- the SU can attain better reward in the future by exploring system state, i.e. probing a new channel.

■

Therefore, by Lemma 1, the SU would probe at least one channel at each slot. To show the upper bound, we construct a genie-aided system, where the SU, with the help of the genie, knows the actual state of all channels and then probes only one unoccupied channel to maximize its reward (or does not probe any channel if none of these channels is unoccupied). We denote the expected accumulated reward of the genie-aided system in the finite horizon time of T as U_g, and obviously, $U_g \leq (1-\alpha)T$. Hence, the expected accumulated reward of $\mathbf{P}$ is upper bounded by U_g.

For the problem $\mathbf{P}$, if the number of channels probed at each slot is constant, i.e. $|\mathcal{A}(t)| = \kappa$ ($\kappa \in [1, M]$), then the myopic policy is optimal under some mild conditions [26], [27], which are stated as follows:

Lemma 2. *Given that the SU probes a fixed number of channels each time, the myopic probing policy is optimal if*

1) $\epsilon_{max} \triangleq \max_{i\in\mathcal{N}}\{\epsilon_i\} \leq \frac{p_{01}(1-p_{11})}{p_{11}(1-p_{01})}$ *for the case of homogeneous channels [26].*
2) $\beta \max_{i\in\mathcal{N}}\{p_{11}^{(i)} - p_{01}^{(i)}\} \leq \frac{1}{2}$ *for the case of heterogeneous channels [27].*

Given the mild conditions according to Lemma 2, a feasible lower bound U_d of $\mathbf{P}$ can be set to the performance of the case in which the SU probes a fixed number of channels each time, denoted as $U_d = \max\big\{U_\kappa : \kappa \in [1, M]\big\}$, where $|\mathcal{A}(t)| = \kappa$ and

$$U_\kappa = \max_{\pi} \mathbb{E}_\pi \left[\sum_{t=1}^{T} \beta^{t-1} R\Big(\pi_t(\Omega(t)), \mathcal{O}_A(t)\Big)\middle|\, \Omega(1)\right]. \quad (9)$$

Thus, we say that a policy χ is called a good one if the performance U_χ achieved by the policy χ satisfies $U_d \leq U_\chi \leq U_g \leq (1-\alpha)T$. That is, the policy χ provides a guaranteed bound for the problem $\mathbf{P}$.

Remark. We would like to point out that the deterministic lower bound U_d is achieved under the myopic probing policy. Therefore, to obtain guaranteed performance, we should consider one of the variants of the myopic probing policy; otherwise, the proposed policy cannot provide guaranteed bound. Hence, in the next subsection, we will propose a heuristic policy which is based on the myopic probing policy.

B. Structure of ν-Step Lookahead Policy

Taking into account the exponential complexity of solving $\mathbf{P}_2$, we turn to the following heuristic strategy referred to as ν-step lookahead policy:

1) (Probing) at slot t, the SU probes the channels according to the decreasing order of the elements in $\Omega(t)$, and estimates the expected accumulated reward from the next ν slots (from $t+1$ to $t+\nu$, $t+\nu \leq T$), assuming that in the next ν slots, the SU stops probing new channels once an available one is found or the maximal number M of channels is reached.
2) (Stoping) at slot t, the SU stops probing new channels when the total reward in the current slot t plus that from slot $t+1$ to $t+\nu$ decreases.

Let $l^k(t)$ and $\Omega^k(t)$ ($k \leq M$) denote the channel index list and belief vector, respectively, according to the descending order of $\omega_i(t)$ ($1 \leq i \leq N$) after probing the first k best channels in slot t, and $l_j^k(t)$ denote the jth channel in $l^k(t)$. Given the initial belief vector $\Omega^0(t+1)$, then the correspondent channel list $l^0(t+1))$ is determined.

If the SU stops probing once a channel is probed to be unoccupied or M is reached, the expected accumulated pseudo cost $Q_{t+1}^{t+\nu}\Big(\Omega^0(t+1)\Big)$ accrued from the next ν slots can be written as follows

$$\begin{aligned} & Q_{t+1}^{t+\nu}\Big(\Omega^0(t+1)\Big) \\ & \triangleq \underbrace{\prod_{j=1}^{M}\big(1-\omega_{l_j^0(t+1)}(t+1)\big)\Big[C_0 + \beta\cdot Q_{t+2}^{t+\nu}\Big(\mathbb{T}\big(\Omega_0^M(t+1)\big)\Big)\Big]}_{term\ B} \\ & \quad + \sum_{i=1}^{M}\Big[\big[\omega_{l_i^0(t+1)}(t+1)\prod_{j=1}^{i-1}\big(1-\omega_{l_j^0(t+1)}(t+1)\big)\big]\cdot \\ & \qquad \underbrace{\big[C(i) + \beta\cdot Q_{t+2}^{t+\nu}\Big(\mathbb{T}\big(\Omega_1^i(t+1)\big)\Big)\big]}_{term\ A}\Big], \end{aligned}$$

where (a) term A denotes the pseudo cost when channel $l_i^0(t+1)$ is probed to be unoccupied while channels $l_1^0(t+1), \cdots, l_{i-1}^0(t+1)$ are probed to be occupied; (b) term B denotes the pseudo cost when the first M channels of $l^0(t+1)$ are probed to be occupied; (c) $\Omega_1^i(t+1)$ and $\Omega_0^i(t+1)$ denote the belief vectors where the channel $l_i^0(t+1)$ is probed to be occupied and unoccupied, respectively; (d) $\mathbb{T}$ denotes the mapping from $\Omega^k(t)$ to $\Omega^0(t+1)$ according to (1) at the beginning of slot $t+1$, i.e., $\mathbb{T}: \Omega^k(t) \rightarrow \Omega^0(t+1)$.

At each slot t, the ν-step lookahead policy can be implemented in a heuristic approach by transforming it into an optimal stopping problem, i.e., the user stops probing new channels when the total reward in the current slot plus that from slot $t+1$ to $t+\nu$ decreases. Mathematically, the number of channels to probe in the ν-step lookahead policy, denoted as $\overline{n}_t$, can be approximately written as follows:

$$\begin{aligned} \overline{n}_t = \inf\Big\{n_t : & C(n_t) + \beta Q_{t+1}^{t+\nu}\Big(\mathbb{T}\big(\Omega^{n_t}(t)\big)\Big) \\ & < C(n_t+1) + \beta \hat{Q}_{t+1}^{t+\nu}\Big(\Omega^{n_t}(t)\Big), 1 \leq n_t \leq M\Big\}, \end{aligned} \quad (10)$$

where (a) $\varsigma \triangleq l^0_{n_t+1}(t)$ denoting the (n_t+1)-th channel of $l^0(t)$; (b) $Q_{t+1}^{t+\nu}\Big(\mathbb{T}\big(\Omega^{n_t}(t)\big)\Big)$ is the expected accumulated pseudo cost from slot $t+1$ to $t+\nu$ when the first n_t channels of $l^0(t)$ are probed; (c) $\hat{Q}_{t+1}^{t+\nu}(\Omega^{n_t}(t))$ denotes the expected accumulated pseudo cost from slot $t+1$ to $t+\nu$ when channel ς is probed to be unoccupied with probability $(1-\epsilon_\varsigma)\omega_\varsigma(t)$ and occupied with probability $1-(1-\epsilon_\varsigma)\omega_\varsigma(t)$, i.e.,

$$\hat{Q}_{t+1}^{t+\nu}\Big(\Omega^{n_t}(t)\Big) \triangleq (1-\epsilon_\varsigma)\omega_\varsigma(t)Q_{t+1}^{t+\nu}\Big(\mathbb{T}\big(\Omega_1^{n_t+1}(t)\big)\Big) + (1-(1-\epsilon_\varsigma)\omega_\varsigma(t))Q_{t+1}^{t+\nu}\Big(\mathbb{T}\big(\Omega_0^{n_t+1}(t)\big)\Big). \tag{11}$$

C. Implementation of ν-Step Lookahead Policy

The following lemma further studies the structure of the ν-step lookahead policy by developing an optimal stopping algorithm, which decomposes the coupling of exploitation and exploration into two stages—exploitation and exploration, based on the structure of the cost function.

Algorithm 1 ν-step lookahead policy: executed for each slot t

```
Input: $\Omega^0(t)$, $l^0(t)$
Output: $n_t$
Initialization: $n_t = 0$
while $n_t < M$ do
    Probe the $(n_t+1)$th channel in $l^0(t)$
    Increase the number of the probed channels, i.e., $n_t = n_t + 1$
    if one of the first $n_t$ channels is probed to be unoccupied and the following inequality holds:
        $C(n_t)+\beta Q_{t+1}^{t+\nu}\Big(\mathbb{T}\big(\Omega^{n_t}(t)\big)\Big) < C(n_t+1)+\beta\hat{Q}_{t+1}^{t+\nu}\Big(\Omega^{n_t}(t)\Big)$   (12)
    then
        Terminate the algorithm by outputting $n_t$
    end if
end while
```

Lemma 3. *The ν-step lookahead policy can be implemented by Algorithm 1 with the computation complexity $O(M^{\nu+1})$.*

Proof: To solve $\overline{n}_t$ in (10), it suffices to show that the SU should (a) continue to probe new channels if all the probed channels are occupied or (b) stop probing new channels if at least one channel is probed to be unoccupied and the expected pseudo cost increases by probing a new channel.

The first action (a) is trivial to prove by Lemma 2. We now show the second action (b). If the SU stops at the current channel, the total cost can be written as $C(n_t) + \beta Q_{t+1}^{t+\nu}\Big(\mathbb{T}\big(\Omega^{n_t}(t)\big)\Big)$. Otherwise, the expected pseudo cost by assuming the SU probes a new channel $l^0_{n_t+1}(t)$ is $C(n_t+1) + \beta\hat{Q}_{t+1}^{t+\nu}\Big(\Omega^{n_t}(t)\Big)$. It can be noted that (10) is equivalent to the condition

$$C(n_t)+\beta Q_{t+1}^{t+\nu}\Big(\mathbb{T}\big(\Omega^{n_t}(t)\big)\Big) < C(n_t+1)+\beta\hat{Q}_{t+1}^{t+\nu}\Big(\Omega^{n_t}(t)\Big)$$

in Algorithm 1.

Noticing that the complexity of Algorithm 1 lies in the computation of (12), and thus it increases exponentially with ν, i.e. $O(M^{\nu+1})$. ■

Remark. It is insightful to note that the proposed ν-step lookahead policy can be decomposed into two steps: first *exploitation* and then *exploration*, which is different from the case where exploitation and exploration are tightly coupled.

- *Exploitation*: the SU exploits the current available information $\Omega(t)$ in a greedy way in order to find an unoccupied channel as soon as possible;
- *Exploration*: the SU continues to explore the system for long term gain once an unoccupied channel is probed. The *exploration* can be omitted if all the M best channels are probed to be occupied or if *exploration* does not increase gain in the long term (i.e., the condition in Algorithm 1 does not hold even once).

Before concluding this subsection, we reemphasize that the complexity of Algorithm 1 lies in the computation of (12) and thus increases exponentially with ν. On the other hand, a larger ν leads to better performance of the lookahead policy. Hence, the parameter ν can be tuned to achieve a desired tradeoff between efficiency and complexity.

D. Low-Complexity Implementation: One-Step Lookahead Policy

In the previous part, we have derived Algorithm 1 to implement the ν-step lookahead policy with exponential complexity. Thus, we focus on the system with i.i.d. channels and provide a mathematical analysis on the simplest case of $\nu = 1$, i.e., the one-step lookahead policy, to demonstrate how to calculate relevant quantities in (10) (11). The study on the one-step lookahead policy can provide structural insights on calculating expected pseudo cost, which is the foundation of the ν-step lookahead policy.

Before delving into the detailed analysis, the following lemma studies how the channel list should be updated when a new channel is probed.

Lemma 4. *For a system with positively correlated homogeneous i.i.d. channels, if $0 \le \epsilon \le \frac{p_{01}(1-p_{11})}{p_{11}(1-p_{01})}$, the channel probed to be unoccupied (occupied) should be moved to the head (tail) of the old channel list to form a new one.*

Proof: Assume the old channel list is $l^k(t) = (\sigma_1, \cdots, \sigma_N,)$ at slot t. We thus have $p_{11} \ge \omega_{\sigma_1}(t) \ge \cdots \ge \omega_{\sigma_N}(t) \ge p_{01}$. If channel σ_{k+1} is probed to be unoccupied, then $\omega_{\sigma_{k+1}}(t) = 1$, and further $l^{k+1}(t) = (\sigma_{k+1}, \sigma_1, \cdots, \sigma_k, \sigma_{k+2}, \cdots, \sigma_N)$ according to the descending order of ω. If channel σ_{k+1} is probed to be occupied, then $\omega_{\sigma_{k+1}}(t) = \varphi(\omega_{\sigma_{k+1}}(t)) \le p_{01}$, and further $l^{k+1}(t) = (\sigma_1, \cdots, \sigma_k, \sigma_{k+2}, \cdots, \sigma_N, \sigma_{k+1})$. ■

Given the system model presented in Subsection II-A, assume that the SU has probed k channels with at least one of them is unoccupied, the condition, by Algorithm 1, to decide whether to probe channel $k+1$ in the channel list can be written as:

$$\alpha > \beta\Big[Q_{t+1}^{t+1}\big(\mathbb{T}(\Omega^k(t))\big) - \hat{Q}_{t+1}^{t+1}\big(\Omega^k(t)\big)\Big]. \tag{13}$$

Without introducing ambiguity, we abuse $Q_{t+1}^{t+1}(\cdot)$ ($\hat{Q}_{t+1}^{t+1}(\cdot)$) and $Q(\cdot)$ ($\hat{Q}(\cdot)$), and show how to compute $Q\Big(\mathbb{T}\big(\Omega^k(t)\big)\Big)$ and $\hat{Q}\Big(\Omega^k(t)\Big)$ in an efficient way for homogeneous channels.

Assume that the channel list at the beginning of slot t is $l^0(t) = (1, 2, \cdots, N)$, sorted in the descending order of the belief values, and that m ($m \geq 1$) channels are probed to be unoccupied while $k - m$ are probed to be occupied among the k probed channels $\{1, \cdots, k\}$. It follows from Lemma 4 that m channels are moved to the head of the channel list and others to the tail, thus forming the new channel list $l^k(t)$.

The key point of the proposed policy is to decide whether to probe channel $k+1$. For tractable analysis, we introduce an auxiliary vector $\boldsymbol{X}\Big(\mathbb{T}\big(\Omega^k(t)\big), m\Big)$, defined as

$$\begin{aligned}
&\boldsymbol{X}\Big(\mathbb{T}(\Omega^k(t)), m\Big) \\
&\triangleq \begin{pmatrix} 1 \\ X_1\Big(\mathbb{T}(\Omega^k(t)), m\Big) \\ X_2\Big(\mathbb{T}(\Omega^k(t)), m\Big) \\ X_3\Big(\mathbb{T}(\Omega^k(t)), m+2\Big) \\ X_4\Big(\mathbb{T}(\Omega^k(t)), m+2\Big) \end{pmatrix} \\
&\triangleq \begin{pmatrix} 1 \\ \prod_{j=1}^{m}\Big(1 - \omega_{l_j^k(t)}(t+1)\Big) \\ 1 + \sum_{i=1}^{m}\prod_{j=1}^{i}\Big(1 - \omega_{l_j^k(t)}(t+1)\Big) \\ \prod_{j=m+2}^{M}\Big(1 - \omega_{l_j^k(t)}(t+1)\Big) \\ \sum_{i=m+2}^{M}\prod_{j=m+2}^{i}\Big(1 - \omega_{l_j^k(t)}(t+1)\Big) \end{pmatrix}.
\end{aligned}$$

The following lemma establishes an important structural property of $\boldsymbol{X}\Big(\mathbb{T}(\Omega^k(t)), m\Big)$ based on which $\boldsymbol{X}\Big(\mathbb{T}(\Omega^{k+1}(t)), m+1\Big)$ can be recursively derived no matter whether the channel $k+1$ is probed to be unoccupied or occupied.

Lemma 5. *The following recursive update on the auxiliary vector holds:*

- *If $k+1$ channel is probed to be unoccupied, $\boldsymbol{X}\Big(\mathbb{T}(\Omega_1^{k+1}(t)), m+1\Big) = \boldsymbol{H_1} \cdot \boldsymbol{X}\Big(\mathbb{T}(\Omega^k(t)), m\Big)$;*
- *If $k+1$ channel is probed to be occupied, $\boldsymbol{X}\Big(\mathbb{T}(\Omega_0^{k+1}(t)), m+1\Big) = \boldsymbol{H_2} \cdot \boldsymbol{X}\Big(\mathbb{T}(\Omega^k(t)), m\Big)$,*

where $\varsigma = l_{m+2}^k(t)$, $\varrho = l_{M+1}^k(t)$, $\eta = 1 - (1-\epsilon)p_{11}$,

$$\boldsymbol{H_1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \eta & 0 & 0 & 0 \\ 1 & 0 & \eta & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{1-\omega_\varsigma(t+1)} & 0 \\ -1 & 0 & 0 & 0 & \frac{1}{1-\omega_\varsigma(t+1)} \end{pmatrix},$$

$$\boldsymbol{H_2} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1-\omega_\varrho(t+1)}{1-\omega_\varsigma(t+1)} & 0 \\ -1 & 0 & 0 & \frac{1-\omega_\varrho(t+1)}{1-\omega_\varsigma(t+1)} & \frac{1}{1-\omega_\varsigma(t+1)} \end{pmatrix}.$$

Proof:

Case 1. When channel $l_{m+1}^k(t)$ is probed to be unoccupied, we have $\omega_{l_{m+1}^k(t)}(t+1) = (1-\epsilon)p_{11}$ according to (1) and false alarm rate. Recalling the definition of X_i ($i = 1, 2, 3, 4$), we have

$$\begin{cases} X_1(\mathbb{T}(\Omega_1^{k+1}(t)), m+1) = [1 - \omega_{l_{m+1}^k(t)}(t+1)]X_1(\mathbb{T}(\Omega^k(t)), m), \\ X_2(\mathbb{T}(\Omega_1^{k+1}(t)), m+1) = 1 + [1 - \omega_{l_{m+1}^k(t)}(t+1)]X_2(\mathbb{T}(\Omega^k(t)), m), \\ X_3(\mathbb{T}(\Omega_1^{k+1}(t)), m+3) = \frac{X_3(\mathbb{T}(\Omega^k(t)), m+2)}{1-\omega_{l_{m+2}^k(t)}(t+1)}, \\ X_4(\mathbb{T}(\Omega_1^{k+1}(t)), m+3) = \frac{X_4(\mathbb{T}(\Omega^k(t)), m+2)}{1-\omega_{l_{m+2}^k(t)}(t+1)} - 1. \end{cases}$$

It is straightforward to verify that $\boldsymbol{X}(\mathbb{T}(\Omega_1^{k+1}(t)), m+1) = \boldsymbol{H_1} \cdot \boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)$.

Case 2. When channel $l_{m+1}^k(t)$ is probed to occupied, we have $\omega_{l_{m+1}^k(t)}(t+1) = (1-\epsilon)\mathcal{T}(\varphi(\omega_{l_{m+1}^k(t)}(t)))$ according to (1) and false alarm rate. Note, if $M = N$, we have $\omega_{l_{M+1}^k(t)}(t+1) = \omega_{l_{m+1}^k(t)}(t+1)$ according to Lemma 4. Recalling the definition of X_i ($i = 1, 2, 3, 4$), we have

$$\begin{cases} X_1(\mathbb{T}(\Omega_0^{k+1}(t)), m) = X_1(\mathbb{T}(\Omega^k(t)), m), \\ X_2(\mathbb{T}(\Omega_0^{k+1}(t)), m) = X_2(\mathbb{T}(\Omega^k(t)), m), \\ X_3(\mathbb{T}(\Omega_0^{k+1}(t)), m+2) = X_3(\mathbb{T}(\Omega^k(t)), m+2)\frac{1-\omega_{l_{M+1}^k(t)}(t+1)}{1-\omega_{l_{m+2}^k(t)}(t+1)}, \\ X_4(\mathbb{T}(\Omega_0^{k+1}(t)), m+2) = \frac{X_4(\mathbb{T}(\Omega^k(t)), m+2)}{1-\omega_{l_{m+2}^k(t)}(t+1)} - 1 \\ \qquad\qquad + X_3(\mathbb{T}(\Omega^k(t)), m+2)\frac{1-\omega_{l_{M+1}^k(t)}(t+1)}{1-\omega_{l_{m+2}^k(t)}(t+1)}. \end{cases}$$

It is straightforward to verify that $\boldsymbol{X}(\mathbb{T}(\Omega_0^{k+1}(t)), m+1) = \boldsymbol{H_2} \cdot \boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)$.

■

After obtaining step-update of the auxiliary vector, three critical quantities $Q\Big(\mathbb{T}\big(\Omega^k(t)\big)\Big)$, $Q\Big(\mathbb{T}\big(\Omega_1^{k+1}(t)\big)\Big)$, and $Q\Big(\mathbb{T}\big(\Omega_0^{k+1}(t)\big)\Big)$ can be easily computed in an efficient fashion by using the auxiliary vector, as stated in the following lemma.

Lemma 6. $Q\Big(\mathbb{T}\big(\Omega^k(t)\big)\Big)$, $Q\Big(\mathbb{T}\big(\Omega_1^{k+1}(t)\big)\Big)$ *and* $Q\Big(\mathbb{T}\big(\Omega_0^{k+1}(t)\big)\Big)$ *are updated by*

$$\begin{aligned} Q\Big(\mathbb{T}\big(\Omega^k(t)\big)\Big) = \alpha\Big[&\boldsymbol{A_2}\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\boldsymbol{A_3}\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m) + \\ &\boldsymbol{A_1}\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\Big] \end{aligned} \tag{14}$$

$$\begin{aligned} Q\Big(\mathbb{T}\big(\Omega_1^{k+1}(t)\big)\Big) = \alpha\Big[&\boldsymbol{A_5}\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\boldsymbol{A_6}\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m) + \\ &\boldsymbol{A_4}\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\Big] \end{aligned} \tag{15}$$

$$\begin{aligned} Q\Big(\mathbb{T}\big(\Omega_0^{k+1}(t)\big)\Big) = \alpha\Big[&\boldsymbol{A_7}\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\boldsymbol{A_8}\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m) + \\ &\boldsymbol{A_1}\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\Big], \end{aligned} \tag{16}$$

where,

$$\begin{aligned} &\boldsymbol{A_1} = [0, 0, 1, 0, 0], \\ &\boldsymbol{A_2} = [0, 1 - \omega_{l_{m+1}^k(t)}(t+1), 0, 0, 0] \\ &\boldsymbol{A_3} = [1, 0, 0, \frac{1}{\alpha} - M - 1, 1], \end{aligned}$$

$$
\begin{aligned}
\boldsymbol{A_4} &= [1, 0, 1-(1-\epsilon)p_{11}, 0, 0]\\
\boldsymbol{A_5} &= [0, 1-(1-\epsilon)p_{11}, 0, 0, 0],\\
\boldsymbol{A_6} &= [0, 0, 0, \frac{1}{\alpha} - M - 1, 1]\\
\boldsymbol{A_7} &= [0, 1, 0, 0, 0],\\
\boldsymbol{A_8} &= [0, 0, 0, (\frac{1}{\alpha} - M)(1-(1-\epsilon)\mathcal{T}(\varphi(\omega_{l^k_{m+1}(t)}(t)))), 0].
\end{aligned}
$$

Proof: Assume that m of k channels are probed to be unoccupied while the remaining $k-m$ channels are probed to be occupied. Then $l^k(t)$ is obtained.

Case 1. If the SU does not probe channel $l^k_{m+1}(t)$, we have, by letting $f^{t+1}_n = 1-\omega_n(t+1)$ and separating channel $l^k_{m+1}(t)$ from others

$$
\begin{aligned}
&Q\Big(\mathbb{T}(\Omega^k(t))\Big)\\
=&\sum_{i=1}^{M} C(i)\omega_{l^k_i(t)}(t+1)\prod_{j=1}^{i-1} f^{t+1}_{l^k_j(t)} + \prod_{j=1}^{M} f^{t+1}_{l^k_j(t)}\\
=&\alpha\sum_{i=1}^{m} i\omega_{l^k_i(t)}(t+1)\prod_{j=1}^{i-1} f^{t+1}_{l^k_j(t)} + \prod_{j=1}^{M} f^{t+1}_{l^k_j(t)} +\\
&+\alpha f^{t+1}_{l^k_{m+1}(t)}\prod_{j=1}^{m} f^{t+1}_{l^k_j(t)}\sum_{i=m+2}^{M} i\omega_{l^k_i(t)}(t+1)\prod_{j=m+2}^{i-1} f^{t+1}_{l^k_j(t)}\\
&+\alpha(m+1)\omega_{l^k_{m+1}(t)}(t+1)\prod_{j=1}^{m} f^{t+1}_{l^k_j(t)}\\
=&\alpha\Big[1 + f^{t+1}_{l^k_1(t)} + \cdots + f^{t+1}_{l^k_1(t)}\cdots f^{t+1}_{l^k_{m-1}(t)} - m f^{t+1}_{l^k_1(t)}\cdots f^{t+1}_{l^k_m(t)}\Big]\\
&+\alpha(m+1)\omega_{l^k_{m+1}(t)}(t+1)\prod_{j=1}^{m} f^{t+1}_{l^k_j(t)} + \prod_{j=1}^{M} f^{t+1}_{l^k_j(t)}\\
&+\alpha f^{t+1}_{l^k_{m+1}(t)}\prod_{j=1}^{m} f^{t+1}_{l^k_j(t)} \times \Big[(m+2) + f^{t+1}_{l^k_{m+2}(t)} +\\
&f^{t+1}_{l^k_{m+2}(t)}\cdots f^{t+1}_{l^k_{M-1}(t)} - M f^{t+1}_{l^k_{m+2}(t)}\cdots f^{t+1}_{l^k_M(t)}\Big]\\
=&\alpha X_2(\mathbb{T}(\Omega^k(t)), m) + \alpha(m+1)(\omega_{l^k_{m+1}(t)}(t+1) - 1)\prod_{j=1}^{m} f^{t+1}_{l^k_j(t)}\\
&+\alpha f^{t+1}_{l^k_{m+1}(t)}\prod_{j=1}^{m} f^{t+1}_{l^k_j(t)} \times \Big[(m+2) + X_4(\mathbb{T}(\Omega^k(t)), m+2)\\
&+(\alpha - M - 1)X_3(\mathbb{T}(\Omega^k(t)), m+2)\Big]\\
=&\alpha X_2(\mathbb{T}(\Omega^k(t)), m) - \alpha(m+1) f^{t+1}_{l^k_{m+1}(t)} X_1(\mathbb{T}(\Omega^k(t)), m)\\
&+\alpha f^{t+1}_{l^k_{m+1}(t)} X_1(\mathbb{T}(\Omega^k(t)), m)\Big[(m+2) + X_4(\mathbb{T}(\Omega^k(t)), m+2)\\
&+(\alpha - M - 1)X_3(\mathbb{T}(\Omega^k(t)), m+2)\Big]\\
=&\alpha X_2(\mathbb{T}(\Omega^k(t)), m) + \alpha f^{t+1}_{l^k_{m+1}(t)} X_1(\mathbb{T}(\Omega^k(t)), m)\Big[1\\
&+X_4(\mathbb{T}(\Omega^k(t)), m+2) + (a - M - 1)X_3(\mathbb{T}(\Omega^k(t)), m+2)\Big]\\
=&\alpha\Big\{\boldsymbol{A_1}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\\
&+\boldsymbol{A_2}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\cdot\boldsymbol{A_3}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\Big\}.
\end{aligned}
$$

Case 2. If channel $l^k_{m+1}(t)$ (corresponding to channel $l^0_{k+1}(t)$) is probed to be unoccupied, then we have by separating channel $l^k_{m+1}(t)$ from others

$$
\begin{aligned}
Q\Big(\mathbb{T}(\Omega^{k+1}_1(t))\Big) &= \alpha\Big\{\boldsymbol{A_4}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\\
&+\boldsymbol{A_5}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\cdot\boldsymbol{A_6}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\Big\}.
\end{aligned}
$$

Case 3. If channel $l^k_{m+1}(t)$ is probed to be occupied,

$$
\begin{aligned}
Q\Big(\mathbb{T}(\Omega^{k+1}_0(t))\Big) &= \alpha\Big\{\boldsymbol{A_1}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\\
&+\boldsymbol{A_7}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\cdot\boldsymbol{A_8}\cdot\boldsymbol{X}(\mathbb{T}(\Omega^k(t)), m)\Big\}.
\end{aligned}
$$

■

Remark. Recall Algorithm 1 and (14)–(16), it can be verified that the one-step lookahead policy has a linear computational complexity $O(M^2)$.

IV. Numerical Experiments

In this section, we demonstrate the obtained results and gain further insight on the developed ν-step lookahead policy as well as the performance tradeoff hinging behind via a set of numerical experiments. Specifically, we present two typical scenarios, the homogeneous channels and the heterogeneous channels. In both scenarios, we are interested in the performance of average reward (throughput) of the following policies:

- *ν-step lookahead policy*;
- *Genie-aided policy*: it gives the upper bound of average reward;
- *Myopic policies* with $k = 1, 2, 3$: they form the lower bound of average reward;
- *Greedy policy*: the SU probes channels greedily and stops on the first available channel (or none of available channels is probed);
- *Random policy*: the SU probes channels randomly and stops on the first available channel (or none of available channels is probed).

Meanwhile, the results in this section provide a complementary quantitative study on the performance of the ν-step lookahead policy, which is not explicitly addressed in the analytical part.

A. Homogeneous Case

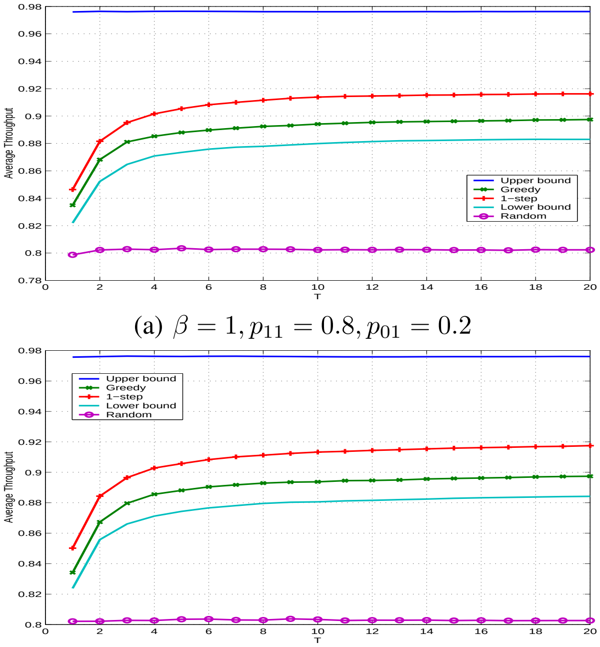
We first consider a homogeneous system with $N = 8$ i.i.d. channels and an SU is allowed to probe at most $M = 3$ channels each slot with $\epsilon = 0.02$ and $\alpha = 0.02$, respectively. The following two representative scenarios, corresponding to a strongly and weakly correlated channel model respectively, are studied:

- Case 1: $p_{11} = 0.8$, $p_{01} = 0.2$;
- Case 2: $p_{11} = 0.5$, $p_{01} = 0.4$.

Figure 1 compares average throughput performance between one-step lookahead policy, random policy, greedy policy and lower/upper bounds. It can be observed from the figure that after the stabilization, the one-step lookahead policy can further increase the throughput by approximately 5% with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2014.2359917, IEEE Transactions on Wireless Communications

respect to the lower bound (corresponding to the myopic policy with $k = 3$). Hence, one-step lookahead policy is a good one since it provides the guaranteed lower bound. Another observation is that the one-step lookahead policy is better that the greedy policy. As analyzed in the previous sections, this gain is due to the fact that the one-step lookahead policy can achieve a desired tradeoff between exploration and exploitation. Moreover, this benefit in throughput is especially attractive given the low complexity of one-step lookahead policy. As for the random policy, the performance is lower than the feasible lower bound, and thus it is not a good policy.



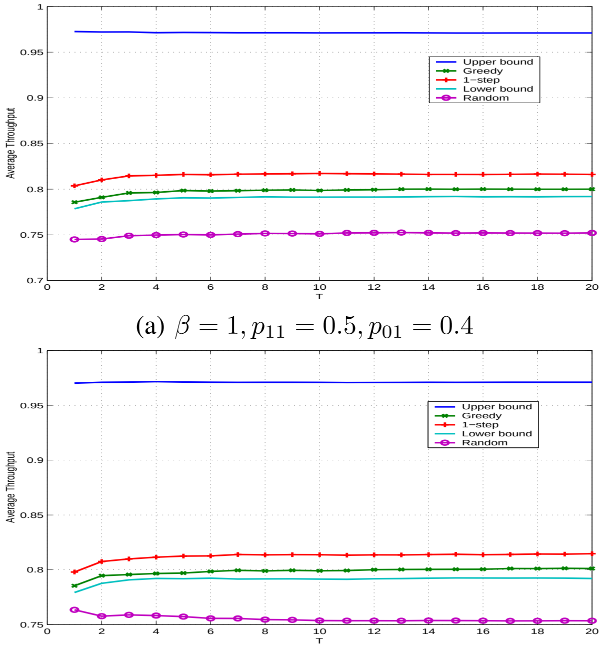
(a) $\beta = 1, p_{11} = 0.8, p_{01} = 0.2$

(b) $\beta = 0.95, p_{11} = 0.8, p_{01} = 0.2$

Fig. 1. Throughput comparison of 1-step lookahead policy, greedy policy, random policy, genie-aided policy, and myopic policy for homogeneous channels ($N = 8, M = 3, \alpha = 0.02, \epsilon = 0.02$)

Figure 2 illustrates the same comparison for Case 2. It can be noticed from the results that the performance gain in Case 2 is less significant compared to Case 1. This can be explained by the fact that the channel correlation in Case 2 is less significant in time than Case 1, and consequently, the effect of prediction is less important.

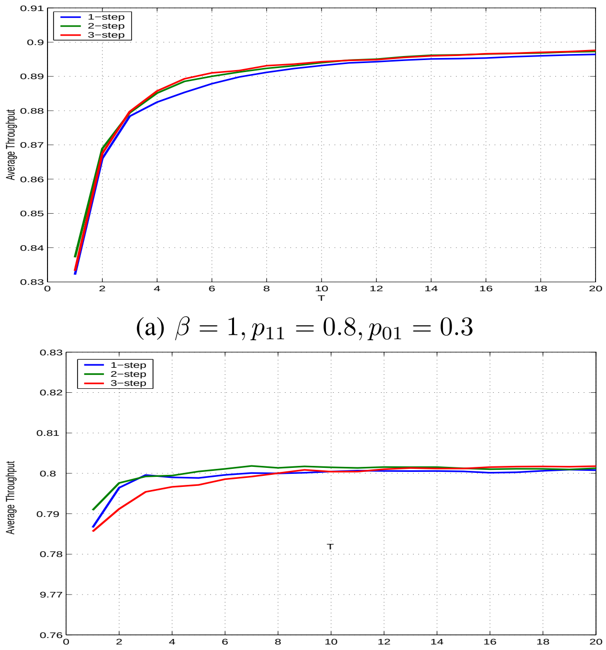
We then proceed to study the performance of the ν-step lookahead policy in the case of $\nu > 1$. Figure 3 shows the average throughput with $\nu = 1, 2, 3$ for Case 1 and Case 2, respectively. It can be observed that statistically, the increase of ν does not enhance the performance gain obviously, which justifies our focus on the one-step lookahead policy. More generically, by taking the complexity into account, we recommend to set $\nu = 1$ in a large variate of parameter settings.



(a) $\beta = 1, p_{11} = 0.5, p_{01} = 0.4$

(b) $\beta = 0.95, p_{11} = 0.5, p_{01} = 0.4$

Fig. 2. Throughput comparison of 1-step lookahead policy, greedy policy, random policy, genie-aided policy, and myopic policy for homogeneous channels ($N = 8, M = 3, \alpha = 0.02, \epsilon = 0.02$)



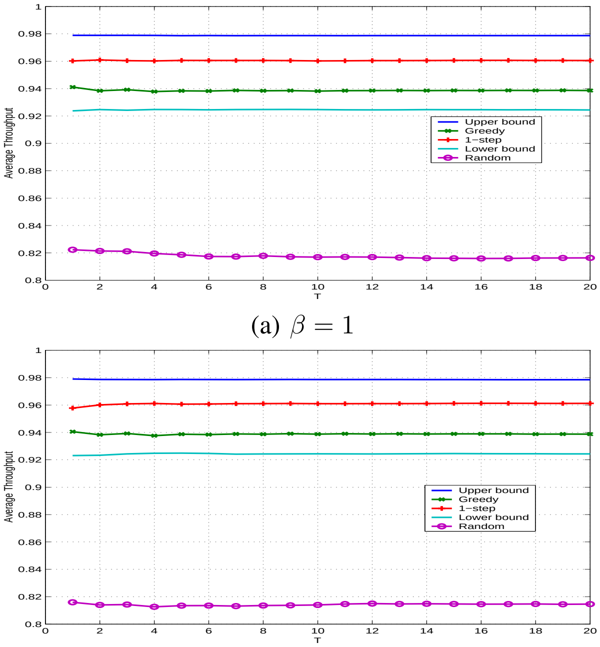
(a) $\beta = 1, p_{11} = 0.8, p_{01} = 0.3$

(b) $\beta = 1, p_{11} = 0.4, p_{01} = 0.3$

Fig. 3. Throughput comparison of 1-,2-,3-step lookahead policy for homogeneous channels ($N = 8, M = 3, \alpha = 0.02, \epsilon = 0.02$)

B. Heterogeneous Case with non i.i.d. Channels

We now proceed to evaluate the performance of the ν-step lookahead policy in heterogeneous systems with non i.i.d. channels. To this end, we randomly generate 100 heterogeneous systems with the following parameter setting: $N = 8$, $M = 3$, $\alpha = 0.02$, $\epsilon_i \in [0.01, 0.02]$, and $p_{11}^{(i)} > p_{01}^{(i)}$ $(1 \le i \le N)$. We plot the average throughput in Figure 4, and observe similar results as that of homogenous systems. That is, the ν-step lookahead policy statistically outperforms the greedy policy and the myopic policies with $k = 1, 2, 3$.



(a) $\beta = 1$

(b) $\beta = 0.95$

Fig. 4. Throughput comparison of 1-step lookahead policy, greedy policy, random policy, genie-aided policy, and myopic policyfor heterogeneous channels ($N = 8, M = 3, \beta = 1, \alpha = 0.02, \epsilon_i \in [0.01, 0.02]$)

V. Conclusion and Perspective

In CR networks, one of objectives is to minimize the probing time of each time slot and then to achieve more spectrum efficiency. In this paper, we investigate the decision-making optimization problem concerning the number of channels to probe, and demonstrate our research efforts to minimize the probing time in finding the first available channel in each time slot through probing other channels even after obtaining an available channel. Specially, the ν-step lookahead policy is proposed to shorten the probing time and improve the probing efficiency as well. In the proposed policy, the parameter ν allows to achieve a desired tradeoff between efficiency and computation complexity. Numerical experiments on several typical settings demonstrate the benefits of the proposed strategy.

References

[1] M. Lopez-Benitez, F. Casadevall, A. Umbert, J. Perez-Romero et al.. Spectral Occupation Measurements and Blind Standard Recognition Sensor for Cognitive Radio Networks. *Proc. of CROWNCOM 2009.* , pages 1–9, Jun. 2009.

[2] M. A. McHenry et al.. Spectrum occupancy measurements. *technical reports (Jan 2004 - Aug 2005)* . Available at: http://www.sharedspectrum.com/measurements.

[3] J. Mitola. Cognitive radio for flexible mobile multimedia communications. In *IEEE Int. Workshop on Mobile Multimedia Communications (MoMuC)*, Nov. 1999.

[4] A. Singh, M. R. Bhatnagar, and R. K. Mallik. Cooperative Spectrum Sensing in Multiple Antenna Based Cognitive Radio Network Using An Improved Energy Detector, In *IEEE Communications Letters*, 16(1), 64-67, Jan. 2012.

[5] A. Singh, M. R. Bhatnagar, and R. K. Mallik. Optimization of Cooperative Spectrum Sensing with an Improved Energy Detector over Imperfect Reporting Channels In *IEEE Vehicular Technology Conference (VTC-Fall)*, Sep. 2011.

[6] A. Singh, M. R. Bhatnagar, and R. K. Mallik. Performance Analysis of Multiple Sample Based Improved Energy Detector in Collabrative CR Networks In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2013

[7] Ian F. Akyildiz, Won-YeolLee, MehmetC. Vuran, ShantidevMohanty. A survey on Spectrum Management in Cognitive Radio Networks. *IEEE Communications Magazine*, 46(4):40–48, Apr. 2008.

[8] B.B. Wang, and K. J. Ray Liu. Advances in Cognitive Radio Networks: A Survey. *IEEE Journal of Selected Topics in Signal Processing*, 5(1):5–23, Feb. 2011.

[9] M.T. Masonta, M. Mzyece, N. Ntlatlapa. Spectrum Decision in Cognitive Radio Networks: A Survey. *IEEE Communications Surveys & Tutorials*,15(3), 1088–1107, Jul. 2013.

[10] E.Z. Tragos, S. Zeadally, A.G. Fragkiadakis, V.A. Siris. Spectrum Assignment in Cognitive Radio Networks: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials*,15(3), 1108–1135, Jul. 2013.

[11] N. B. Chang and M. Liu. Optimal channel probing and transmission scheduling for opportunistic spectrum access. *Proc. 13th ACM Annu. Int. Conf. MobiCom*, pages 27–38, Sep. 2007.

[12] K. Munagala S. Guha and S. Sarkar. Approximation schemes for information acquisition and exploitation in multichannel wireless networks. In *Proc. 44th Annu. Allerton Conf. Commun., Control Comput.*, Monticello, IL, Sep. 2006.

[13] H. Jiang, L. Lai, R. Fan, and H. V. Poor. Optimal selection of channel sensing order in cognitive radio. *IEEE Transactions on Wireless Communication*, 8(1):297–307, Jan. 2009.

[14] Z. Zhang, H. Jiang, P. Tan, and J. Slevinsky. Channel Exploration and Exploitation with Imperfect Spectrum Sensing in Cognitive Radio Networks. *IEEE Journal of Seleted Areas in Communications*, 31(3):429–441, Mar. 2013.

[15] H. Kim and K. G. Shin. Fast discovery of spectrum opportunities in cognitive radio networks. In *Proc. IEEE Symp. New Frontiers DySPAN*, Chicago, IL, Oct. 2008.

[16] Q. Li and Z. Li. A Novel Sequential Spectrum Sensing Method in Cognitive Radio Using Suprathreshold Stochastic Resonance. *IEEE Transactions on Vehicular Technology*, pp(99):1, Sep. 2013.

[17] Q. Liu, X. Wang and Yong Cui. Robust and Adaptive Scheduling of Sequential Periodic Sensing for Cognitive Radios. *IEEE Journal on Selected Areas in Communications*, 32(3):503-515, 2014.

[18] T. Shu, and H. Li. QoS-Compliant Sequential Channel Sensing for Cognitive Radios. *IEEE Journal on Selected Areas in Communications*, pp(99):1-13, 2014.

[19] B. Li, P. Yang, J. Wang et al.. Almost Optimal Dynamically-Ordered Channel Sensing and Accessing for Cognitive Networks. *IEEE Transactions on Mobile Computing*, pp(99):1, 2013.

[20] L.M. Lopez-Ramos, A.G. Marques, and J. Ramos Soft-decision sequential sensing for optimization of interweave Cognitive Radio networks. In *2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications* , 2013.

[21] Q. Zhao, and B. Krishnamachari, and K. Liu. On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance. *IEEE Transactions Wireless Communication*, 7(3):5413–5440, Dec. 2008.

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

[23] S. Ahmad and M. Liu. Multi-channel opportunistic access: a case of restless bandits with multiple plays. In *Allerton Conference*, Monticello, Il, Spet.-Oct. 2009.

[24] K. Liu, and Q. Zhao, and B. Krishnamachari. Dynamic multichannel access with imperfect channel state detection. *IEEE Transactions on Signal Processing*, 58(5):2795–2807, May 2010.

[25] K. Wang and L. Chen. On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach. *IEEE Transactions on Signal Processing*, 60(1):300–309, 2012.

[26] K. Wang, Q. Liu L. Chen, and Khaldoun Al Agha. On optimality of myopic sensing policy with imperfect sensing in multi-channel opportunistic access. *IEEE Transactions on Communications*, 61(9):3854–3862, 2013.

[27] K. Wang, L. Chen, and Q. Liu. On Optimality of Myopic Policy for Opportunistic Access with Non-identical Channels and Imperfect Sensing. *IEEE Transactions on Vehicular Technology*, 63(5):2478-2483, 2014.

[28] S. Guha and K. Munagala. Approximation algorithms for partial information based stochastic control with markovian rewards. In *Proc. IEEE*

Symposium on Foundations of Computer Science (FOCS), Providence, RI, Oct. 2007.

[29] S. Guha and K. Munagala. Approximation algorithms for restless bandit problems. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, New York, Jan. 2009.

[30] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queueing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

Kehao Wang received the B.S degree in Electrical Engineering, M.s degree in Communication and Information System from Wuhan University of Technology, Wuhan, China, in 2003 and 2006, respectively, and Ph.D in the Department of Computer Science, the University of Paris-Sud XI, Orsay, France, in 2012. In 2013, he joined the School of Information Engineering at the Wuhan University of Technology, where he is currently an associate professor. His research interests are cognitive radio networks, wireless network resource management, and embedded operating system.

Lin Chen received his B.E. degree in Radio Engineering from Southeast University, China in 2002 and the Engineer Diploma, Ph.D from Telecom ParisTech, Paris in 2005 and 2008, respectively. He also holds a M.S. degree of Networking from the University of Paris 6. He currently works as associate professor in the department of computer science of the University of Paris-Sud XI. His main research interests include modeling and control for wireless networks, security and cooperation enforcement in wireless networks and game theory.

Quan Liu is currently a Professor and the Dean of School of Information Engineering, Wuhan University of Technology, Wuhan, China. Her research interests include nonlinear system analysis and signal processing, etc. She has published more than 90 papers.

Wei Wang (StM'08-M'10) received the B.S. degree in Communication Engineering and the Ph.D. degree in Signal and Information Processing from Beijing University of Posts and Telecommunications, China in 2004 and 2009, respectively. Now, he is an associate professor with Department of Information Science and Electronic Engineering, Zhejiang University, China. From Sept. 2007 to Sept. 2008, he was a visiting student with University of Michigan, Ann Arbor, USA. Since Feb. 2013, he has also been a Hong Kong Scholar with Hong Kong University of Science and Technology, Hong Kong. His research interests mainly focus on cognitive radio networks, green communications, and radio resource allocation for wireless networks.

He is the editor of the book "*Cognitive Radio Systems*" (Intech, 2009) and serves as an editor for *Transactions on Emerging Telecommunications Technologies (ETT)*. He serves as TPC co-chair for CRNet 2010 and NRN 2011, symposium co-chair for WCSP 2013, tutorial co-chair for ISCIT 2011, and also serves as TPC member for major international conferences.

Fangmin Li received his B.S. degree from the Huazhong Univ. of Science and Technology (China) and his M.S. degree from the National Univ. of Defense Technology (China) in 1990 and 1997, respectively, both in Computer Science. In 2001, he received his Ph.D. degree at Zhejiang Univ. (China) in the field of Computer Science and Engineering. In 2002, he joined the School of Information Engineering at the Wuhan Univ. of Technology, where he is currently a full professor. His main research interests include ad hoc networks, new generation network architecture and embedded system design. His past research has been published in over 30 scientific journals and conference proceedings and 16 China patents. His research has been funded by the National Natural Science Foundation of China and other sources. He is a senior member of the China Computer Federation and executive member of the Sensor Network Technical Committee (China).