Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 791789, 12 pages
http://dx.doi.org/10.1155/2013/791789

*Research Article*

# Optimality of Multichannel Myopic Sensing in the Presence of Sensing Error for Opportunistic Spectrum Access

**Xiaofeng Jiang and Hongsheng Xi**

*Department of Automation, University of Science and Technology of China, Hefei 230021, China*

Correspondence should be addressed to Xiaofeng Jiang; jxf@mail.ustc.edu.cn

Received 19 January 2013; Revised 8 August 2013; Accepted 19 August 2013

Academic Editor: Alberto Cabada



The optimization problem for the performance of opportunistic spectrum access is considered in this study. A user, with the limited sensing capacity, has opportunistic access to a communication system with multiple channels. The user can only choose several channels to sense and decides whether to access these channels based on the sensing information in each time slot. Meanwhile, the presence of sensing error is considered. A reward is obtained when the user accesses a channel. The objective is to maximize the expected (discounted or average) reward accrued over an infinite horizon. This problem can be formulated as a partially observable Markov decision process. This study shows the optimality of the simple and robust myopic policy which focuses on maximizing the immediate reward. The results show that the myopic policy is optimal in the case of practical interest.

## 1. Introduction

There is a significant increase in the demand for radio spectrum with the emergence of new applications and the compelling need for mobile services in recent years. This is partly due to the increasing interest of consumers in convenient and ubiquitous wireless services, and the interest has been driving the evolution of wireless networks to high speed data networks. However, ever since the 1920s, in order to avoid the serious interference in wireless services, the wireless providers have been required to apply an exclusive license from the government. Today, it is becoming very difficult to find vacant bands to either deploy new services or to enhance the existing ones with most of the spectrum being already allocated [1]. On the other hand, not every channel in every band is in use all the time; a large number of vacant spectrum holes can be discovered in the spectrum [2]. A technique for opportunistic spectrum access can effectively utilize these spectrum holes.

The spectrum sensing for detecting spectrum holes is the precondition for opportunistic spectrum access. However, the existing spectrum sensing techniques has to face one main challenge: wideband sensing which is hard to be implemented for the main reason of hardware limitations [3]. The user usually uses a tunable narrowband bandpass filter at the radio frequency (RF) front-end to sense one channel at a time due to the costliness of a wideband RF front-end. Consequently, it is a lot of time delay for detecting all channels. Meng et al. [4] study this problem with the method of compressive sensing based on the sparse observations of sensing information. Equipped with frequency selective filters, the sparse sensing information vectors of multiple channels are linearly combined and compressed. Multiple channels thus can be sensed simultaneously. However, it is used in the case of less practical interest due to the requirements of sparse observations and frequency selective filters. The basic theory of compressive sensing is given in the work of [5–8]. Some other studies focus on the reliability of sensing information. Different SNR estimations and channel fading environments are considered in [9, 10] to improve the reliability of sensing information. Chen [11] studies the optimum number of collaborative users to get the tradeoff of the reliability and the complexity. The Byzantine attacks which come from malicious users and carry false sensing data are taken into account in [12].

The studies of [13–16] also exploit a method to solve the problem of wideband sensing by estimating the information of all channels with only a small amount of sensing results. The sensing procedure is modeled as a partially observed Markov decision process (POMDP). Zhao et al. [13] propose

this idea and a myopic sensing method. Wang et al. [14] exploit the impact of the rateless code. Lingcen et al. [15] modify the cost function of POMDP with the switching time.

We consider the communication system where a user has opportunistic access to multiple channels like the model of [13] but is limited to sensing and transmitting only on several channels at a given time due to its hardware limitation. Meanwhile, the presence of sensing error is also considered. We explore the problem to maximize the performance of opportunistic access when the past observations and the knowledge of the stochastic properties of these channels are given. This problem can be described as a partially observable Markov decision process since the user does not have full knowledge of the availabilities of all channels. We examine the optimality of the myopic policy for this problem in this study for the reason that the myopic policy is very simple and robust. Specially, we show that the myopic policy is optimal in the case of practical interest. Ahmad et al. [16] also study the optimality of the myopic sensing. However, we have discovered that the study of [16] is only suitable for the special case where the one-channel myopic sensing and the absence of sensing error are considered. If we consider the multichannel myopic sensing or the presence of sensing error, the study of [16] cannot hold its conclusion. The reason is that the mathematic method of the proof of [16] is very special for the case of one-channel myopic sensing, and Lemmas 2, 3, 4, and 5 of [16] cannot be improved to prove the optimality under the other conditions. We propose the proof of the optimality of multichannel myopic sensing in the presence of sensing error. Our mathematic method is rigorous and quite different from the method of [16]; we use two functions to give the proof, and the method is generally effective for such issues.

The rest of this paper is organized as follows. We formulate the problem in Section 2 and give the definition of the myopic policy in Section 3. We prove the optimality of the myopic policy in Section 4 and extend the results from the finite horizon to the infinite horizon. The numerical results of the performance comparison of the myopic policy and the optimal policy are given in Section 5, and the conclusion is drawn in Section 6.

## 2. System Model

Consider a spectrum consisting of $N$ independent and statistically identical wireless channels; each channel has two states {0(*occupied*), 1(*idle*)}, and the state transition is given by a two-state discrete time Markov chain shown in Figure 1. It is supposed that these channels evolve according to a synchronous time slot structure indexed by $t$, where $t = 1, 2, \ldots, T$. Specially, the states of all channels at time $t$ are denoted by $S(t) = [S_1(t), \ldots, S_N(t)]$, where $S_n(t) \in \{0, 1\}$.

We consider a user seeking the spectrum holes in these channels for opportunistic access. At the beginning of the time slot $t$, the user selects a set $A_1(t)$ of channels to sense and a set $A_2(t)$ of channels to access, where $A_2(t) \subseteq A_1(t)$. However, due to hardware limitation, $|A_1(t)|$ and $|A_2(t)|$ are usually much smaller than $N$. The user can only sense the channels in $A_1(t)$ and does not have the states of all channels $S(t)$, when a decision is made at time $t$. Consequently, the spectrum is not fully observable to the users. For clarity, we use $a(t) = \{A_1(t), A_2(t)\}$ to denote the action of the user at time $t$ and use $h_{A_1(t)}(t) = [h_{i_1}(t), \ldots, h_{i_{|A_1(t)|}}(t)]$ to denote the sensing information of channels, where $h_{A_1(t)}(t) \in \{0, 1\}^{|A_1(t)|}$ and $i_k$ is the index of the channel in $A_1(t)$. Specially, we cannot guarantee the absolute reliability of the sensing information in the presence of sensing error; the detection and false alarm probabilities should be taken into account. Here, the detection probability denoted by $p_d^{i_k}(t)$ and the false alarm probability denoted by $p_f^{i_k}(t)$ [17] are the conditional probabilities that the state of the channel $i_k$ is actually 0(*occupied*) when the observations are 0(*occupied*) and 1(*idle*), respectively.

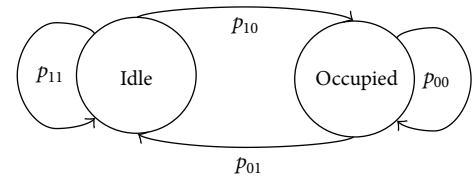


Figure 1: State transition of channels.

For making the optimal decision, a sufficient statistic of this system denoted by $\Lambda(t) = [\lambda_1(t), \ldots, \lambda_N(t)]$ is given, where $\lambda_n(t)$ is the conditional probability that the channel $n$ is idle at the beginning of the time slot $t$ given all past observations and actions, and this conditional probability is called the idle conditional probability. Due to the Markovian property of the channels, the future idle conditional probability is only a function of the current idle conditional probability and the action.

**Proposition 1.** *The relationship between* $\Lambda(t)$ *and* $\Lambda(t+1)$ *can be given by* (3).

*Proof.* We first consider case 1 that $n \in A_1(t)$ and $h_n(t) = 1$ to simplify the proof; the events $S_n(t+1) = 1$, $S_n(t) = 1$, $S_n(t) = 0$, and $h_n(t) = 1$ are denoted by $A$, $B_1$, $B_2$, and $C$, respectively. The value of $\lambda_n(t+1)$ in case 1 can be replaced by $P\{A \mid C\}$; then we can get

$$\begin{aligned} P\{A \mid C\} &= \sum_{i=\{1,2\}} P\{A, B_i \mid C\} \\ &= \sum_{i=\{1,2\}} \frac{P\{B_i, C\}}{P\{C\}} \frac{P\{A, B_i, C\}}{P\{B_i, C\}} \\ &= \sum_{i=\{1,2\}} P\{B_i \mid C\} P\{A \mid B_i, C\}. \end{aligned} \tag{1}$$

Since $A$ is independent of $C$ when $B_i$ has been determined, we can get

$$\begin{aligned} P\{A \mid C\} &= \sum_{i=\{1,2\}} P\{B_i \mid C\} P\{A \mid B_i\} \\ &= p_{11}\left(1 - p_f^n(t)\right) + p_{01} p_f^n(t) \end{aligned} \tag{2}$$

$$\lambda_n(t+1) = \begin{cases} P\{S_n(t+1)=1 \mid h_n(t)=1\} \\ \quad \text{if } n \in A_1(t), \quad h_n(t)=1 \\ P\{S_n(t+1)=1 \mid h_n(t)=0\} \\ \quad \text{if } n \in A_1(t), \quad h_n(t)=0 \\ P\{S_n(t+1)=1 \mid \lambda_n(t)\} \\ \quad \text{if } n \notin A_1(t) \end{cases}$$

$$= \begin{cases} P\{S_n(t+1)=1 \mid S_n(t)=1\} \\ \quad \times P\{S_n(t)=1 \mid h_n(t)=1\} \\ \quad + P\{S_n(t+1)=1 \mid S_n(t)=0\} \\ \quad \times P\{S_n(t)=0 \mid h_n(t)=1\} \\ \quad \text{if } n \in A_1(t), \quad h_n(t)=1 \\ P\{S_n(t+1)=1 \mid S_n(t)=1\} \\ \quad \times P\{S_n(t)=1 \mid h_n(t)=0\} \\ \quad + P\{S_n(t+1)=1 \mid S_n(t)=0\} \\ \quad \times P\{S_n(t)=0 \mid h_n(t)=0\} \\ \quad \text{if } n \in A_1(t), \quad h_n(t)=0 \\ P\{S_n(t+1)=1 \mid S_n(t)=1\} \\ \quad \times P\{S_n(t)=1 \mid \lambda_n(t)\} \\ \quad + P\{S_n(t+1)=1 \mid S_n(t)=0\} \\ \quad \times P\{S_n(t)=0 \mid \lambda_n(t)\} \\ \quad \text{if } n \notin A_1(t) \end{cases} \tag{3}$$

$$= \begin{cases} p_{11}\left(1-p_f^n(t)\right) + p_{01}p_f^n(t) \\ \quad \text{if } n \in A_1(t), \quad h_n(t)=1 \\ p_{11}\left(1-p_d^n(t)\right) + p_{01}p_d^n(t) \\ \quad \text{if } n \in A_1(t), \quad h_n(t)=0 \\ p_{11}\lambda_n(t) + p_{01}\left(1-\lambda_n(t)\right) \\ \quad \text{if } n \notin A_1(t). \end{cases}$$

The proofs of the other cases are similar. □

The objective of the user is to maximize its total (discounted or average) expected reward; let $J_\alpha^\pi(\Lambda(1))$ and $J^\pi(\Lambda(1))$ [18] denote the rewards, respectively. Here, $\pi$ denotes the policy of the user, $\alpha$ denotes the discounted factor, $\Lambda(1)$ is the initial probability distribution of all channels, and $E_{\Lambda(1)}^\pi$ represents the mathematical expectation which is determined by the initial probability distribution $\Lambda(1)$ and the policy $\pi$. Consequently, $J_\alpha^\pi(\Lambda(1))$ and $J^\pi(\Lambda(1))$ denote the discounted and average reward with the initial probability distribution $\Lambda(1)$ and the policy $\pi$, respectively.

An optimal policy should maximize the reward of the user, and this optimization problem can be formally defined as follows:

$$\max_\pi J_\alpha^\pi(\Lambda(1)) = \max_\pi E_{\Lambda(1)}^\pi\left[\lim_{T\to+\infty}\sum_{t=1}^{T}\alpha^{t-1}r(\Lambda(t),a(t))\right],$$
$$\max_\pi J^\pi(\Lambda(1)) = \max_\pi \limsup_{T\to+\infty}\frac{1}{T}E_{\Lambda(1)}^\pi\left[\sum_{t=1}^{T}r(\Lambda(t),a(t))\right], \tag{4}$$

where $r(\Lambda(t), a(t))$ denotes the immediate reward when the user implements the action $a(t)$, and we think that each channel to be accessed brings 1 unit of reward; thus $r(\Lambda(t), a(t)) = |A_2(t)|$. However, there is a selection problem of $A_2(t)$. Without loss of generality and for the greedy approach, all the channels whose states are sensed as 1 in $A_1(t)$ are selected into $A_2(t)$.

Then, we can give a recursive expression of the reward function:

$$R_T(\Lambda(T)) = \max_{a(T)} E[r(\Lambda(T), a(T))]$$

$$R_t(\Lambda(t)) = \max_{a(t)} E\left[r(\Lambda(t),a(t)) + \alpha R_{t+1}(\Lambda(t+1))\right]$$
$$= \max_{a(t)}\left\{\sum_{i\in A_1(t)}\lambda_i(t) + \alpha \times \sum_{A_2(t)\subseteq A_1(t)}\left[\left(\prod_{j\in A_2(t)}\lambda_j(t)\right)\times\left(\prod_{k\notin A_2(t),k\in A_1(t)}(1-\lambda_k(t))\right)\times R_{t+1}(\Lambda(t+1)\mid a(t))\right]\right\}, \tag{5}$$

where $E[r(\Lambda(t), a(t))]$ denotes the mathematical expectation of the immediate reward in the time slot $t$, $R_t(\Lambda(t))$ denotes the maximum expected reward that is accrued from time $t$ to $T$. Specially, $R_1(\Lambda(1)) = \max_\pi J_\alpha^\pi(\Lambda(1))$ when $T \to +\infty$. This proposition is proved in Section 4.3.

## 3. Myopic Policy

The myopic policy is essentially a greedy policy which maximizes the immediate expected reward in each time slot and ignores the future reward; this greedy policy has the minimal time complexity and computational complexity. The expression of the myopic policy can be given by

$$a(t) = \arg\max_{a(t)} E[r(\Lambda(t), a(t))] = \arg\max_{a(t)} \sum_{i\in A_1(t)}\lambda_i(t). \tag{6}$$

We can discover that the channels which have the $|A_1(t)|$ largest conditional probabilities in $\Lambda(t)$ are selected into $A_1(t)$ by the myopic policy. Consequently, the successive update of the idle conditional probability vector $\Lambda(t)$ can determine the action of the myopic policy at time $t$.

In particular, if $1 - p_f^n(t)$ is larger than all the idle conditional probabilities and $1 - p_d^n(t)$ is smaller than all the idle conditional probabilities for any $n$, the myopic policy requires only the initial condition $\Lambda(1)$ but not the precise values of $\{\Lambda(2), \Lambda(3), \ldots\}$. To give an explanation, we first simplify the expression of (3) by defining a function $f(x) = p_{11}x + p_{01}(1-x)$:

$$\lambda_n(t+1) = \begin{cases} f\left(1-p_f^n(t)\right) & \text{if } n \in A_1(t),\ h_n(t)=1 \\ f\left(1-p_d^n(t)\right) & \text{if } n \in A_1(t),\ h_n(t)=0 \\ f\left(\lambda_n(t)\right) & \text{if } n \notin A_1(t). \end{cases} \tag{7}$$

Due to the monotonicity of $f(x)$, $f(x)$ is a monotonically increasing function when $p_{11} \geq p_{01}$. The ordering of the idle conditional probabilities can be preserved when they are updated for the reason that $f(\lambda_i(t)) \geq f(\lambda_j(t))$ if $\lambda_i(t) \geq \lambda_j(t)$. If a channel is selected into $A_1(t)$, its idle conditional probability will become $1 - p_f^n(t)$ when it is observed as 1, or $1 - p_d^n(t)$ when the observation is 0. That is, the channel has the largest idle conditional probability if it is observed as 1, or the smallest idle conditional probability if it is observed as 0. The myopic policy can create a list which preserves the ordering of the idle conditional probabilities according to the initial condition $\Lambda(1)$. After each update in each time slot, the channels which are not observed do not change the list, the channels which are observed as 1 are selected into $A_2(t)$ and moved to the top of the list, and the channels which are observed as 0 are moved to the bottom of the list. Consequently, the myopic policy does not require the precise values of the updated idle conditional probabilities in this case.

We have an opposite situation when $p_{11} < p_{01}$, $f(x)$ is a monotonically decreasing function. The ordering of the idle conditional probabilities should be reversed when they are updated for the reason that $f(\lambda_i(t)) \geq f(\lambda_j(t))$ if $\lambda_i(t) \leq \lambda_j(t)$. The myopic policy also creates a list which preserve the ordering of the idle conditional probabilities. After each update in each time slot, the channels which are not observed reverse their locations, the channels which are observed as 1 are moved to the bottom of the list, and the channels which are observed as 0 are moved to the top of the list. Consequently, the myopic policy also does not require the precise values of the updated idle conditional probabilities in this case.

## 4. Optimality of Myopic Policy

In order to show the optimality of the myopic policy, we first define two functions which can denote the expected rewards obtained by the myopic policy and the arbitrary policy:

$$\begin{aligned} U_t(\Lambda) &= U_t(\{\lambda_1, \ldots, \lambda_N\}) \\ &= \left(\lambda_{N-|A_1^U|+1} + \cdots + \lambda_N\right) + \alpha \\ &\quad \times \sum_{A_2^U \subseteq A_1^U} \left[ \left( \prod_{j \in A_2^U} \lambda_j \right) \left( \prod_{k \notin A_2^U, k \in A_1^U} (1 - \lambda_k) \right) \right. \\ &\qquad \times U_{t+1}\Big( \left\{ f\left(1 - p_d^k\right) \mid k \notin A_2^U, k \in A_1^U \right\} \\ &\qquad\quad \bigcup \left\{ f(\lambda_i) \mid i \notin A_1^U \right\} \\ &\qquad\quad \left. \bigcup \left\{ f\left(1 - p_f^j\right) \mid j \in A_2^U \right\} \Big) \right] \\ U_T(\{\lambda_1, \ldots, \lambda_N\}) &= \lambda_{N-|A_1^U|+1} + \cdots + \lambda_N, \end{aligned} \tag{8}$$

where $U_t$ denotes the expected total reward obtained by the myopic policy from time $t$ on. $\Lambda$ denotes the sequence of the idle conditional probabilities of all channels; it is reordered to $\{\lambda_1, \ldots, \lambda_N\}$ by $U_t$ and $\lambda_1 \leq \cdots \leq \lambda_N$. $A_1^U$ denotes the set $\{N - |A_1^U| + 1, \ldots, N\}$ of channels which have been chosen to sense by the myopic policy. $A_2^U$ denotes the set of channels which have been chosen to access:

$$\begin{aligned} W_t(\Lambda) &= W_t\left(\left\{\lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}}\right\}\right) \\ &= \left(\lambda_{i_1} + \cdots + \lambda_{i_{|A_1^W|}}\right) + \alpha \\ &\quad \times \sum_{A_2^W \subseteq A_1^W} \left[ \left( \prod_{j \in A_2^W} \lambda_j \right) \left( \prod_{k \notin A_2^W, k \in A_1^W} (1 - \lambda_k) \right) \right. \\ &\qquad \times U_{t+1}\Big( \left\{ f\left(1 - p_d^k\right) \mid k \notin A_2^W, k \in A_1^W \right\} \\ &\qquad\quad \bigcup \left\{ f(\lambda_i) \mid i \notin A_1^W \right\} \\ &\qquad\quad \left. \bigcup \left\{ f\left(1 - p_f^j\right) \mid j \in A_2^W \right\} \Big) \right] \\ W_T\left(\left\{\lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}}\right\}\right) &= \lambda_{i_1} + \cdots + \lambda_{i_{|A_1^W|}}, \end{aligned} \tag{9}$$

where $W_t$ denotes the expected total reward obtained by the arbitrary policy at time $t$ and the myopic policy from time $t + 1$ on. $\Lambda$ denotes the sequence of the idle conditional probabilities of all channels. The channels corresponding to its last entries $\{\lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}}\}$ are selected into $A_1^W$. The arbitrary set $A_1^W$ is corresponding to the arbitrary policy. $A_1^W$ denotes the set $\{i_1, \ldots, i_{|A_1^W|}\}$ of channels which have been chosen to sense by the arbitrary policy. $A_2^W$ denotes the set of channels which have been chosen to access.

In particular,

$$W_t(\{\lambda_1, \ldots, \lambda_N\}) = U_t(\Lambda). \tag{10}$$

**Theorem 2.** *When $T$ is finite, the optimality of the myopic policy at times $1, 2, \ldots, T$ is equivalent to*

$$W_t(\Lambda) \leq U_t(\Lambda) \tag{11}$$

*for any $A_1^W$ and $t = 1, 2, \ldots, T$.*

*Proof.* We first prove the sufficiency inductively. The myopic policy is optimal at time $T$ for the reasons that $U_T$ is larger than any $W_T$ and $W_T$ obtained by the arbitrary policy.

Then, we suppose that the myopic policy is optimal at time $t + 1$. We have that $W_t$ is larger than the reward obtained by the same policy at time $t$ and the arbitrary policy from time $t + 1$ due to the induction hypothesis, and the immediate reward obtained by the myopic policy is the largest. Consequently, $U_t$ is larger than any $W_t$. $U_t$ is thus larger than the reward obtained by the arbitrary policy. The myopic policy is optimal at time $t$. The proof of the sufficiency is complete.

The proof of the necessity can also be obtained due to the optimality of the myopic policy. □

**Lemma 3.** *$U_t$ and $W_t$ are $N$ variable functions which are polynomial of order 1 for $t = 1, \ldots, T$.*

*Proof.* We prove this by induction over time $t$. $U_T$ and $W_T$ are polynomial of order 1 due to their definitions.

We suppose that $U_{t+1}$ and $W_{t+1}$ are polynomial of order 1. We have that $U_{t+1}(\{f(1-p_d^k) \mid k \notin A_2^U, k \in A_1^U\} \bigcup \{f(\lambda_i) \mid i \notin A_1^U\} \bigcup \{f(1-p_f^j) \mid j \in A_2^U\})$ and $U_{t+1}(\{f(1-p_d^k) \mid k \notin A_2^W, k \in A_1^W\} \bigcup \{f(\lambda_i) \mid i \notin A_1^W\} \bigcup \{f(1-p_f^j) \mid j \in A_2^W\})$ are polynomial of order 1 for the reason that $f$ is a linear function. Consequently, $U_t$ and $W_t$ are polynomial of order 1. The proof is complete. □

### 4.1. The Case of $p_{11} \geq p_{01}$

*Assumption 4.* The transition probabilities $p_{11}$ and $p_{01}$ are such that $p_{11} \geq p_{01}$.

The function $f$ is monotonically increasing under Assumption 4. For any $x, y \in [0,1]$ and $x \leq y$, we have $f(0) \leq f(x) \leq f(y) \leq f(1)$.

*Assumption 5.* We assume that for any discounted factor $\alpha \in [0,1]$ and all the idle conditional probabilities, the detection probability $p_d$ and the false alarm probability $p_f$ are such that

$$1 + \alpha \sum_{A_2 \subseteq A_1} \left[ \left( \prod_{j \in A_2} \lambda_j \right) \times \left( \prod_{k \notin A_2, k \in A_1} (1 - \lambda_k) \right) \left( p_d - p_f - 1 \right) \right] \geq 0. \tag{12}$$

For

$$\sum_{A_2 \subseteq A_1} \left[ \left( \prod_{j \in A_2} \lambda_j \right) \left( \prod_{k \notin A_2, k \in A_1} (1 - \lambda_k) \right) \right] = 1, \tag{13}$$

we can rewrite Assumption 5 as follows

$$\begin{aligned} & 1 + \alpha \left( p_d - p_f - 1 \right) \geq 0 \\ & \Longrightarrow p_d - p_f \geq 1 - \frac{1}{\alpha}. \end{aligned} \tag{14}$$

Assumption 5 is used to limit the reliability of the sensing information; we cannot make the optimal decision if the information is very unreliable. $p_d$ equals 1 and $p_f$ equals 0 in the absence of sensing error; Assumption 5 is always true.

**Theorem 6.** *The myopic policy is optimal under Assumptions 4 and 5 when $T$ is finite.*

*To prove this theorem, one should show that $U_t$ is larger than any $W_t$ for $t = 1, \ldots, T$ according to Theorem 2. One proves this inductively. Given that $U_i$ is larger than any $W_i$ for $i = t+1, \ldots, T$, one wants to show that $U_t$ is larger than any $W_t$. This relies on a number of lemmas introduced below.*

**Lemma 7.** *For all $\lambda_x \geq \lambda_y$, $x, y \in A_1^W$, one has*

$$\begin{aligned} & W_t\left( \left\{ \lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \ldots, \lambda_x, \lambda_y, \ldots \right\} \right) \\ & \quad = W_t\left( \left\{ \lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \ldots, \lambda_y, \lambda_x, \ldots \right\} \right). \end{aligned} \tag{15}$$

**Lemma 8.** *For any $\lambda_{j_1} \leq \cdots \leq \lambda_x \leq \lambda_y \leq \cdots \leq \lambda_{j_{N-|A_1^W|}}$ and $\lambda_{i_1} \leq \cdots \leq \lambda_{i_{|A_1^W|}}$, one has*

$$\begin{aligned} & W_t\left( \left\{ \ldots, \lambda_x, \lambda_y, \ldots, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}} \right\} \right) \\ & \quad = W_t\left( \left\{ \ldots, \lambda_y, \lambda_x, \ldots, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}} \right\} \right). \end{aligned} \tag{16}$$

*Proof.* Lemmas 7 and 8 are true according to the definition of $W_t$. □

**Lemma 9.** *One has*

$$\begin{aligned} & W_t\left( \left\{ \lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}} \right\} \right) \\ & \quad - W_t\left( \left\{ \lambda_{j_1}, \ldots, \lambda_{i_1}, \lambda_{j_{N-|A_1^W|}}, \ldots, \lambda_{i_{|A_1^W|}} \right\} \right) \\ & = \left( \lambda_{i_1} - \lambda_{j_{N-|A_1^W|}} \right) \\ & \quad \times \left[ W_t\left( \left\{ \lambda_{j_1}, \ldots, 0, 1, \ldots, \lambda_{i_{|A_1^W|}} \right\} \right) \right. \\ & \qquad \left. - W_t\left( \left\{ \lambda_{j_1}, \ldots, 1, 0, \ldots, \lambda_{i_{|A_1^W|}} \right\} \right) \right]. \end{aligned} \tag{17}$$

*Proof.* We use LHS and RHS to denote the left-hand side and the right-hand side of the equation, respectively. We can prove that $W_t(\{\lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}}\})$ is the first-order function of $\lambda_{j_{N-|A_1^W|}}$ and $\lambda_{i_1}$ according to Lemma 3. Consequently, we can suppose that

$$\begin{aligned} & W_t\left( \left\{ \lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}} \right\} \right) \\ & \quad = a\lambda_{j_{N-|A_1^W|}} + b\lambda_{i_1} + c\lambda_{j_{N-|A_1^W|}}\lambda_{i_1} + d, \end{aligned} \tag{18}$$

where $a$, $b$, $c$, and $d$ are irrelevant with $\lambda_{j_{N-|A_1^W|}}$ and $\lambda_{i_1}$. Consequently, we have that

$$\begin{aligned} \text{LHS} &= \left( a\lambda_{j_{N-|A_1^W|}} + b\lambda_{i_1} + c\lambda_{j_{N-|A_1^W|}}\lambda_{i_1} + d \right) \\ &\quad - \left( a\lambda_{i_1} + b\lambda_{j_{N-|A_1^W|}} + c\lambda_{j_{N-|A_1^W|}}\lambda_{i_1} + d \right) \\ &= (a-b)\left( \lambda_{j_{N-|A_1^W|}} - \lambda_{i_1} \right), \end{aligned} \tag{19}$$

$$\begin{aligned} \text{RHS} &= \left( \lambda_{i_1} - \lambda_{j_{N-|A_1^W|}} \right) \\ &\quad \times [(a0 + b1 + c0 \times 1 + d) \\ &\qquad - (a1 + b0 + c0 \times 1 + d)] \\ &= (a-b)\left( \lambda_{j_{N-|A_1^W|}} - \lambda_{i_1} \right) = \text{LHS}. \end{aligned} \tag{20}$$

The proof is complete. □

**Lemma 10.** *Consider Assumption 4. One has*

$$\begin{aligned} & U_t\left( \left\{ \lambda_z^1, \lambda_1, \ldots, \lambda_{N-1} \right\} \right) \\ & \quad - U_t\left( \left\{ \lambda_1, \ldots, \lambda_i, \lambda_z^2, \lambda_{i+1}, \ldots, \lambda_{N-1} \right\} \right) \\ & \quad \geq \lambda_z^1 - \lambda_z^2 \end{aligned} \tag{21}$$

*for any $\lambda_z^1 \leq \lambda_1 \leq \cdots \leq \lambda_z^2 \leq \cdots \leq \lambda_{N-1}$.*

*Proof.* We use LHS to denote the left-hand side of the inequality. We use $f^n(x)$ to denote $f(f(\cdots f(x)\cdots))$:

$$\begin{aligned} f^n(x) &= (p_{11}-p_{01})^n x + (p_{11}-p_{01})^{n-1} p_{01} + \cdots + p_{01} \\ f^n(x) - f^n(y) &= (p_{11}-p_{01})^n (x-y). \end{aligned} \tag{22}$$

Consequently, we have

$$f^n(x) - f^n(y) \geq \cdots \geq f(x) - f(y) \geq x - y \tag{23}$$

for any $x \leq y$. Therefore, we have

$$\begin{aligned} \text{LHS} &\geq \left(\lambda_i - \lambda_z^2\right) + \left(\lambda_{i-1} - \lambda_i\right) + \cdots + \left(\lambda_z^1 - \lambda_1\right) \\ &\geq \lambda_z^1 - \lambda_z^2. \end{aligned} \tag{24}$$

□

**Lemma 11.** *Consider Assumptions 4 and 5. For any* $\lambda_{j_1} \leq \cdots \leq \lambda_{j_{N-|A_1^W|}}$ *and* $\lambda_{i_1} \leq \cdots \leq \lambda_{i_{|A_1^W|}}$, *if* $\lambda_{j_{N-|A_1^W|}} \geq \lambda_{i_1}$, *one has*

$$\begin{aligned} &W_t\left(\left\{\lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}}\right\}\right) \\ &\quad \leq W_t\left(\left\{\lambda_{j_1}, \ldots, \lambda_{i_1}, \lambda_{j_{N-|A_1^W|}}, \ldots, \lambda_{i_{|A_1^W|}}\right\}\right). \end{aligned} \tag{25}$$

*Proof.* The inequality is true at time $T$. We have the following equation for any time $t < T$ according to Lemma 9:

$$\begin{aligned} &W_t\left(\left\{\lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}}\right\}\right) \\ &\quad - W_t\left(\left\{\lambda_{j_1}, \ldots, \lambda_{i_1}, \lambda_{j_{N-|A_1^W|}}, \ldots, \lambda_{i_{|A_1^W|}}\right\}\right) \\ &= \left(\lambda_{i_1} - \lambda_{j_{N-|A_1^W|}}\right) \\ &\quad \times \left[W_t\left(\left\{\lambda_{j_1}, \ldots, 0, 1, \ldots, \lambda_{i_{|A_1^W|}}\right\}\right)\right. \\ &\qquad \left. - W_t\left(\left\{\lambda_{j_1}, \ldots, 1, 0, \ldots, \lambda_{i_{|A_1^W|}}\right\}\right)\right]. \end{aligned} \tag{26}$$

Because $\lambda_{j_{N-|A_1^W|}} \geq \lambda_{i_1}, \lambda_{i_1} - \lambda_{j_{N-|A_1^W|}} \leq 0$, we use LHS to denote $W_t(\{\lambda_{j_1}, \ldots, 0, 1, \ldots, \lambda_{i_{|A_1^W|}}\}) - W_t(\{\lambda_{j_1}, \ldots, 1, 0, \ldots, \lambda_{i_{|A_1^W|}}\})$. According to the definition of $W_t$, we have

$$\begin{aligned} \text{LHS} = 1 + \alpha &\times \sum_{A_2^W \subseteq A_1^W} \left[\left(\prod_{j \in A_2^W - \{i_1\}} \lambda_j\right)\right. \\ &\times \left(\prod_{k \notin A_2^W, k \in A_1^W - \{i_1\}} (1 - \lambda_k)\right) \\ &\times \left(U_{t+1}\left(\left\{f\left(1 - p_f^{i_1}\right)\right\}\right.\right. \\ &\qquad \bigcup \left\{f\left(1 - p_f^j\right) \mid j \notin A_2^W - \{i_1\}\right\} \\ &\qquad \bigcup \left\{f(\lambda_i) \mid i \notin A_1^W\right\} \\ &\qquad \bigcup \left\{f\left(1 - p_d^k\right) \mid k \notin A_2^W, \right. \\ &\qquad\quad \left.\left. k \in A_1^W - \{i_1\}\right\} \right. \\ &\quad - U_{t+1}\left(\left\{f\left(1 - p_d^{j_{N-|A_1^W|}}\right)\right\}\right. \\ &\qquad \bigcup \left\{f\left(1 - p_d^k\right) \mid k \notin A_2^W, \right. \\ &\qquad\quad \left. k \in A_1^W - \{i_1\}\right\} \\ &\qquad \bigcup \left\{f(\lambda_i) \mid i \notin A_1^W\right\} \\ &\qquad \bigcup \left\{f\left(1 - p_f^j\right) \mid j \in A_2^W - \{i_1\}\right\} \\ &\qquad \left.\left.\left. \bigcup \{f(1)\}\right)\right)\right]. \end{aligned} \tag{27}$$

According to Lemma 10, we have

$$\begin{aligned} \text{LHS} \geq 1 &+ \alpha \sum_{A_2^W \subseteq A_1^W} \left[\left(\prod_{j \in A_2^W - \{i_1\}} \lambda_j\right)\right. \\ &\times \left(\prod_{k \notin A_2^W, k \in A_1^W - \{i_1\}} (1 - \lambda_k)\right) \\ &\left. \times \left(\left(0 - \left(1 - p_d^{j_{N-|A_1^W|}}\right)\right) + \left(1 - p_f^{i_1} - 1\right)\right)\right]. \end{aligned} \tag{28}$$

Consequently, LHS $\geq 0$ under Assumption 5:

$$\left(\lambda_{i_1} - \lambda_{j_{N-|A_1^W|}}\right) \text{LHS} \leq 0. \tag{29}$$

The proof is complete. □

Now we can give the proof of Theorem 6 with Lemmas 7, 8, and 11.

*Proof.* $W_T \leq U_T$ due to their definitions. For any time $t < T$, $\lambda_{j_1} \leq \cdots \leq \lambda_{j_{N-|A_1^W|}}$, $\lambda_{i_1} \leq \cdots \leq \lambda_{i_{|A_1^W|}}$, and $\lambda_1 \leq \cdots \leq \lambda_N$, we have

$$\begin{aligned} &W_t(\{\lambda_1, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_N, \lambda_i\}) \\ &\quad \leq W_t(\{\lambda_1, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_{N-1}, \lambda_i, \lambda_N\}) \\ &\quad \leq W_t(\{\lambda_1, \ldots, \lambda_{i-1}, \lambda_i, \lambda_{i+1}, \ldots, \lambda_N\}) \\ &\quad = U_t(\{\lambda_1, \ldots \lambda_N\}), \\ &W_t\left(\left\{\lambda_{j_1}, \ldots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \ldots, \lambda_{i_{|A_1^W|}}\right\}\right) \\ &\quad \leq W_t\left(\left\{\ldots, \lambda_{i_1 - 1}, \lambda_{i_1}, \lambda_{i_1 + 1}, \ldots, \right.\right. \\ &\qquad \left.\left. \lambda_{i_{|A_1^W|}-1}, \lambda_{i_{|A_1^W|}}, \lambda_{i_{|A_1^W|}+1}, \ldots\right\}\right) \end{aligned}$$

$$= W_t(\{\lambda_1, \dots \lambda_N\}) = U_t(\{\lambda_1, \dots \lambda_N\}). \tag{30}$$

The inequalities are true due to Lemmas 7, 8, and 11. Consequently, $W_t \le U_t$ at time $t = 1, \dots, T$. The myopic policy is optimal according to Theorem 2. The proof is complete. □

### 4.2. The Case of $p_{11} \le p_{01}$

*Assumption 12.* The transition probabilities $p_{11}$ and $p_{01}$ are such that $p_{11} \le p_{01}$.

The function $f$ is monotonically decreasing under Assumption 12. For any $x, y \in [0,1]$ and $x \le y$, we have $f(1) \le f(y) \le f(x) \le f(0)$.

*Assumption 13.* We assume that, for all the idle conditional probabilities, the discounted factor $\alpha$ is such that

$$1 - \alpha \sum_{A_2 \subseteq A_1} \left[ \left( \prod_{j \in A_2} \lambda_j \right) \times \left( \prod_{k \notin A_2, k \in A_1} (1 - \lambda_k) \right) \frac{|A_1|}{(1-\alpha)} \right] \ge 0. \tag{31}$$

Like Assumption 5, we can rewrite Assumption 13 as follows:

$$1 - \frac{\alpha |A_1|}{(1-\alpha)} \ge 0 \implies \alpha \le \frac{1}{(1 + |A_1|)}. \tag{32}$$

**Theorem 14.** *The myopic policy is optimal under Assumptions 12 and 13 when $T$ is finite.*

**Lemma 15.** *Consider Assumptions 12 and 13. For any $\lambda_{j_1} \le \cdots \le \lambda_{j_{N-|A_1^W|}}$ and $\lambda_{i_1} \le \cdots \le \lambda_{i_{|A_1^W|}}$, if $\lambda_{j_{N-|A_1^W|}} \ge \lambda_{i_1}$, one has*

$$W_t\left(\left\{\lambda_{j_1}, \dots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \dots, \lambda_{i_{|A_1^W|}}\right\}\right) \le W_t\left(\left\{\lambda_{j_1}, \dots, \lambda_{i_1}, \lambda_{j_{N-|A_1^W|}}, \dots, \lambda_{i_{|A_1^W|}}\right\}\right). \tag{33}$$

*Proof.* The inequality is true at time $T$. We have the following equation for any time $t < T$ according to Lemma 9:

$$\begin{aligned} &W_t\left(\left\{\lambda_{j_1}, \dots, \lambda_{j_{N-|A_1^W|}}, \lambda_{i_1}, \dots, \lambda_{i_{|A_1^W|}}\right\}\right) \\ &\quad - W_t\left(\left\{\lambda_{j_1}, \dots, \lambda_{i_1}, \lambda_{j_{N-|A_1^W|}}, \dots, \lambda_{i_{|A_1^W|}}\right\}\right) \\ &= \left(\lambda_{i_1} - \lambda_{j_{N-|A_1^W|}}\right) \\ &\quad \times \left[W_t\left(\left\{\lambda_{j_1}, \dots, 0, 1, \dots, \lambda_{i_{|A_1^W|}}\right\}\right) \right. \\ &\qquad \left. - W_t\left(\left\{\lambda_{j_1}, \dots, 1, 0, \dots, \lambda_{i_{|A_1^W|}}\right\}\right)\right]. \end{aligned} \tag{34}$$

Because $\lambda_{j_{N-|A_1^W|}} \ge \lambda_{i_1}$, $\lambda_{i_1} - \lambda_{j_{N-|A_1^W|}} \le 0$, we use LHS to denote $W_t(\{\lambda_{j_1}, \dots, 0, 1, \dots, \lambda_{i_{|A_1^W|}}\}) - W_t(\{\lambda_{j_1}, \dots, 1, 0, \dots, \lambda_{i_{|A_1^W|}}\})$. According to the definition of $W_t$, we have

$$\begin{aligned} \text{LHS} = 1 + \alpha &\times \sum_{A_2^W \subseteq A_1^W} \left[ \left( \prod_{j \in A_2^W - \{i_1\}} \lambda_j \right) \right. \\ &\times \left( \prod_{k \notin A_2^W, k \in A_1^W - \{i_1\}} (1 - \lambda_k) \right) \\ &\times \left( U_{t+1}\left( \left\{ f\left(1 - p_f^{i_1}\right) \right\} \right.\right. \\ &\qquad \bigcup \left\{ f\left(1 - p_f^{j}\right) \mid j \in A_2^W - \{i_1\} \right\} \\ &\qquad \bigcup \left\{ f(\lambda_i) \mid i \notin A_1^W \right\} \\ &\qquad \bigcup \left\{ f\left(1 - p_d^{k}\right) \mid k \notin A_2^W, \right. \\ &\qquad\quad \left. k \in A_1^W - \{i_1\} \right\} \\ &\qquad \left.\left. \bigcup \{f(0)\} \right) \right. \\ &- U_{t+1}\left( \{f(1)\} \right. \\ &\qquad \bigcup \left\{ f\left(1 - p_f^{j}\right) \mid j \in A_2^W - \{i_1\} \right\} \\ &\qquad \bigcup \left\{ f(\lambda_i) \mid i \notin A_1^W \right\} \\ &\qquad \bigcup \left\{ f\left(1 - p_d^{k}\right) \mid k \notin A_2^W, \right. \\ &\qquad\quad \left. k \in A_1^W - \{i_1\} \right\} \\ &\qquad \left. \left. \left. \bigcup \left\{ f\left(1 - p_d^{j_{N-|A_1^W|}}\right) \right\} \right) \right) \right]. \end{aligned} \tag{35}$$

We have the following inequalities due to the definition of $U_t$:

$$\begin{aligned} &U_{t+1}\left( \left\{ f\left(1 - p_f^{i_1}\right) \right\} \right. \\ &\qquad \bigcup \left\{ f\left(1 - p_f^{j}\right) \mid j \in A_2^W - \{i_1\} \right\} \\ &\qquad \bigcup \left\{ f(\lambda_i) \mid i \notin A_1^W \right\} \\ &\qquad \bigcup \left\{ f\left(1 - p_d^{k}\right) \mid k \notin A_2^W, k \in A_1^W - \{i_1\} \right\} \\ &\qquad \left. \bigcup \{f(0)\} \right) \\ &\ge 0 \end{aligned} \tag{36}$$

$$\begin{aligned} &U_{t+1}\left( \{f(1)\} \bigcup \left\{ f\left(1 - p_f^{j}\right) \mid j \in A_2^W - \{i_1\} \right\} \right. \\ &\qquad \bigcup \left\{ f(\lambda_i) \mid i \notin A_1^W \right\} \\ &\qquad \bigcup \left\{ f\left(1 - p_d^{k}\right) \mid k \notin A_2^W, k \in A_1^W - \{i_1\} \right\} \\ &\qquad \left. \bigcup \left\{ f\left(1 - p_d^{j_{N-|A_1^W|}}\right) \right\} \right) \end{aligned}$$

$$\begin{aligned}
&\le \left|A_1^W\right|\left(1+\alpha^1+\cdots+\alpha^{T-t}\right)\\
&\le \left|A_1^W\right|\left(1+\alpha^1+\cdots+\alpha^{+\infty}\right)\\
&= \frac{\left|A_1^W\right|}{(1-\alpha)}.
\end{aligned} \tag{37}$$

Then, we have

$$\begin{aligned}
\text{LHS} \ge 1 &+ \alpha \sum_{A_2^W \subseteq A_1^W} \left[\left(\prod_{j\in A_2^W-\{i_1\}} \lambda_j\right)\right.\\
&\times \left(\prod_{k\notin A_2^W, k\in A_1^W-\{i_1\}} (1-\lambda_k)\right)\\
&\left.\times\left(0-\frac{\left|A_1^W\right|}{(1-\alpha)}\right)\right].
\end{aligned} \tag{38}$$

Consequently, LHS $\ge 0$ under Assumption 13.

$$\left(\lambda_{i_1}-\lambda_{j_{N-|A_1^W|}}\right)\text{LHS} \le 0. \tag{39}$$

The proof is complete. □

Now we can give the proof of Theorem 14 with Lemmas 7, 8, and 15. The proof of Theorem 14 is similar with the proof of Theorem 6.

*4.3. The Case of $T \to +\infty$.* We discuss the optimality of the myopic policy in above subsections when $T$ is finite; now we consider the extensions of results when $T$ is infinite.

**Theorem 16.** *If the myopic policy is optimal when $T$ is finite, it is optimal when $T$ is infinite.*

*Proof.*

$$\begin{aligned}
&\max_\pi J_\alpha^\pi(\Lambda(1))\\
&= \max_\pi E_{\Lambda(1)}^\pi\left[\lim_{T\to+\infty}\sum_{t=1}^{T}\alpha^{t-1}r(\Lambda(t),a(t))\right]\\
&= \max_\pi \lim_{T\to+\infty} E_{\Lambda(1)}^\pi\left[\sum_{t=1}^{T}\alpha^{t-1}r(\Lambda(t),a(t))\right]\\
&= \max_\pi \lim_{T\to+\infty} J_T^\pi(\Lambda(1)),
\end{aligned} \tag{40}$$

where $J_T^\pi(\Lambda(1))$ denotes $E_{\Lambda(1)}^\pi[\sum_{t=1}^{T}\alpha^{t-1}r(\Lambda(t),a(t))]$ only in this proof.

For

$$\begin{aligned}
&\sum_{t=1}^{T}\alpha^{t-1}r(\Lambda(t),a(t))\\
&\quad\le |A_1|\left(1+\alpha+\cdots+\alpha^{T-1}\right)\\
&\quad\le |A_1|\left(1+\alpha+\cdots+\alpha^{+\infty}\right)\\
&\quad= \frac{|A_1|}{(1-\alpha)},
\end{aligned} \tag{41}$$

we can use the bounded convergence theorem to interchange $\lim_{T\to+\infty}$ and $E_{\Lambda(1)}^\pi$. Then, we consider the relationship of $\lim_{T\to+\infty}$ and $\max_\pi$, and two sequences are given as follows:

Sequence 1: $\{J_1^\pi(\Lambda(1)), J_2^\pi(\Lambda(1)), \ldots\}$

Sequence 2: $\{\max_\pi J_1^\pi(\Lambda(1)), \max_\pi J_2^\pi(\Lambda(1)), \ldots\}$.

We have the following inequalities:

$$\begin{aligned}
&E_{\Lambda(1)}^\pi\left[\sum_{t=1}^{T+1}\alpha^{t-1}r(\Lambda(t),a(t))\right]\\
&\quad\ge E_{\Lambda(1)}^\pi\left[\sum_{t=1}^{T}\alpha^{t-1}r(\Lambda(t),a(t))\right]\\
&\quad\Longrightarrow J_{T+1}^\pi(\Lambda(1)) \ge J_T^\pi(\Lambda(1))\\
&\max_\pi J_{T+1}^\pi(\Lambda(1)) \ge J_{T+1}^{\pi'}(\Lambda(1)) \ge J_T^{\pi'}(\Lambda(1))\\
&\quad\Longrightarrow \max_\pi J_{T+1}^\pi(\Lambda(1)) \ge \max_\pi J_T^\pi(\Lambda(1)),
\end{aligned} \tag{42}$$

where $\pi'$ denotes $\arg\max_\pi J_T^\pi(\Lambda(1))$. And for any $T$, we have

$$\begin{aligned}
J_T^\pi(\Lambda(1)) &\le \frac{|A_1|}{(1-\alpha)}\\
\max_\pi J_T^\pi(\Lambda(1)) &\le \frac{|A_1|}{(1-\alpha)}.
\end{aligned} \tag{43}$$

Consequently, Sequences 1 and 2 are monotonically increasing and bounded. We can conclude that Sequences 1 and 2 have finite limits due to the monotone convergence theorem. We thus have

$$\begin{aligned}
&\lim_{T\to+\infty}\max_\pi J_T^\pi(\Lambda(1)) - \lim_{T\to+\infty} J_T^\pi(\Lambda(1))\\
&\quad= \lim_{T\to+\infty}\left(\max_\pi J_T^\pi(\Lambda(1)) - J_T^\pi(\Lambda(1))\right) \ge 0
\end{aligned} \tag{44}$$

for any policy $\pi$ and $T$. Therefore, we have

$$\lim_{T\to+\infty}\max_\pi J_T^\pi(\Lambda(1)) \ge \max_\pi \lim_{T\to+\infty} J_T^\pi(\Lambda(1)). \tag{45}$$

For $\max_\pi J_T^\pi(\Lambda(1)) \in \{J_T^\pi(\Lambda(1)) \mid \pi\}$, we have

$$\lim_{T\to+\infty}\max_\pi J_T^\pi(\Lambda(1)) \le \max_\pi \lim_{T\to+\infty} J_T^\pi(\Lambda(1)). \tag{46}$$

We can conclude that

$$\lim_{T\to+\infty}\max_\pi J_T^\pi(\Lambda(1)) = \max_\pi \lim_{T\to+\infty} J_T^\pi(\Lambda(1)). \tag{47}$$

Consequently, we have

$$\begin{aligned}
&\max_\pi J_\alpha^\pi(\Lambda(1))\\
&\quad= \lim_{T\to+\infty}\max_\pi J_T^\pi(\Lambda(1))\\
&\quad= \lim_{T\to+\infty}\max_\pi E_{\Lambda(1)}^\pi\left[\sum_{t=1}^{T}\alpha^{t-1}r(\Lambda(t),a(t))\right]\\
&\quad= \lim_{T\to+\infty} R_1(\Lambda(1)).
\end{aligned} \tag{48}$$

The myopic policy is thus optimal when $T$ is infinite if the myopic policy is optimal when $T$ is finite. $R_1(\Lambda(1))$ has been defined in Section 2.

Then, we consider the uniqueness of the optimal policy. Let $\pi^*$ be the myopic policy, and $a^*(t) = \{A_1^*(t), A_2^*(t)\}$ is the action at time $t$; we have

$$\begin{aligned}
&\max_{\pi} J_{\alpha}^{\pi}(\Lambda(1)) \\
&\quad = \lim_{T\to+\infty} \max_{\pi} J_T^{\pi}(\Lambda(1)) \\
&\quad = \lim_{T\to+\infty} J_T^{\pi^*}(\Lambda(1)) \\
&\quad = \lim_{T\to+\infty} E_{\Lambda(1)}^{\pi^*}\left[\sum_{t=1}^{T}\alpha^{t-1} r\left(\Lambda(t), a^*(t)\right)\right] \\
&\quad = \sum_{i\in A_1^*(1)} \lambda_i + \alpha \\
&\qquad \times \sum_{A_2^*(1)\subseteq A_1^*(1)} \left[\left(\prod_{j\in A_2^*(1)} \lambda_j(1)\right)\right. \\
&\qquad\qquad \times \left(\prod_{k\notin A_2^*(1), k\in A_1^*(1)} (1-\lambda_k(1))\right) \\
&\qquad\qquad \times \lim_{T\to+\infty} E_{\Lambda(2)}^{\pi^*} \\
&\qquad\qquad \left.\times \left[\sum_{t=2}^{T}\alpha^{t-2} r\left(\Lambda(2), a^*(2) \mid a^*(1)\right)\right]\right].
\end{aligned} \tag{49}$$

For

$$\begin{aligned}
&\lim_{T\to+\infty} E_{\Lambda(2)}^{\pi^*}\left[\sum_{t=2}^{T}\alpha^{t-2} r\left(\Lambda(2), a^*(2) \mid a^*(1)\right)\right] \\
&\quad = \lim_{T\to+\infty} J_{T-1}^{\pi^*}\left(\Lambda(2) \mid a^*(1)\right) \\
&\quad = \lim_{T\to+\infty} J_T^{\pi^*}\left(\Lambda(2) \mid a^*(1)\right) \\
&\quad = J_{\alpha}^{\pi^*}\left(\Lambda(2) \mid a^*(1)\right),
\end{aligned} \tag{50}$$

we have

$$\begin{aligned}
&J_{\alpha}^{\pi^*}(\Lambda(1)) \\
&\quad = \sum_{i\in A_1^*(1)} \lambda_i + \alpha \\
&\qquad \times \sum_{A_2^*(1)\subseteq A_1^*(1)} \left[\left(\prod_{j\in A_2^*(1)} \lambda_j(1)\right)\right. \\
&\qquad\qquad \times \left(\prod_{k\notin A_2^*(1), k\in A_1^*(1)} (1-\lambda_k(1))\right) \\
&\qquad\qquad \left.\times J_{\alpha}^{\pi^*}\left(\Lambda(2) \mid a^*(1)\right)\right].
\end{aligned} \tag{51}$$

The above equation is the dynamic programming equation for the infinite horizon discounted reward problem. The uniqueness of the optimal policy can be proved due to the uniqueness of the dynamic programming solution.

The proof is complete. □

**Theorem 17.** *If the myopic policy is optimal for the discounted reward, it is optimal for the average reward.*

*Proof.* We first consider the Blackwell optimality [19, pp. 336–341] of the optimal policy for the discounted reward. The sequence of $\{\alpha_1, \alpha_2, \ldots\}$ is given, and $\alpha_1 \le \alpha_2 \le \cdots \le 1$, $\lim_{k\to+\infty}\alpha_k = 1$. For

$$\begin{aligned}
&\left(1-\alpha_k\right) J_{\alpha_k}^{\pi}(\Lambda(1)) \\
&\qquad \le \left(1-\alpha_k\right)\left|A_1\right|\left(1+\alpha_k+\alpha_k^2+\cdots\right) = \left|A_1\right| \\
&\left|J_{\alpha_k}^{\pi}(\Lambda) - J_{\alpha_k}^{\pi}(\Lambda(1))\right| \\
&\qquad \le \left|0-\left|A_1\right|\left(1+\alpha_k+\alpha_k^2+\cdots\right)\right| = \frac{\left|A_1\right|}{\left(1-\alpha_k\right)},
\end{aligned} \tag{52}$$

we can give the definition of $J(\Lambda(1))$ and $h(\Lambda)$ due to the boundedness of $(1-\alpha_k)J_{\alpha_k}^{\pi}(\Lambda(1))$ and $|J_{\alpha_k}^{\pi}(\Lambda) - J_{\alpha_k}^{\pi}(\Lambda(1))|$:

$$\begin{aligned}
J(\Lambda(1)) &= \max_{\pi} \lim_{k\to+\infty} \left(1-\alpha_k\right) J_{\alpha_k}^{\pi}(\Lambda(1)) \\
h(\Lambda) &= \max_{\pi} \lim_{k\to+\infty} \left(J_{\alpha_k}^{\pi}(\Lambda) - J_{\alpha_k}^{\pi}(\Lambda(1))\right)
\end{aligned} \tag{53}$$

for any $\Lambda$.

Then, we can give the average cost optimality equation (ACOE) [20]. Here, we calculate the reward:

$$\begin{aligned}
&J(\Lambda(1)) + h^{\pi}(\Lambda(1)) \\
&\quad = \max_{\pi} \lim_{k\to+\infty} \left[\left(1-\alpha_k\right) J_{\alpha_k}^{\pi}(\Lambda(1))\right. \\
&\qquad\qquad \left.+ \left(J_{\alpha_k}^{\pi}(\Lambda(1)) - J_{\alpha_k}^{\pi}(\Lambda(1))\right)\right] \\
&\quad = \max_{\pi} \lim_{k\to+\infty} \left[\left(1-\alpha_k\right) J_{\alpha_k}^{\pi}(\Lambda(1))\right].
\end{aligned} \tag{54}$$

For

$$\begin{aligned}
&J_{\alpha_k}^{\pi}(\Lambda(1)) \\
&= \sum_{i\in A_1(1)} \lambda_i + \alpha \\
&\quad \times \sum_{A_2(1)\subseteq A_1(1)} \left[\left(\prod_{j\in A_2(1)} \lambda_j(1)\right)\right. \\
&\qquad\qquad \times \left(\prod_{k\notin A_2(1), k\in A_1(1)} (1-\lambda_k(1))\right) \\
&\qquad\qquad \left.\times J_{\alpha}^{\pi}(\Lambda(2) \mid a(1))\right],
\end{aligned}$$

$$J_{\alpha_k}^{\pi}(\Lambda(1)) = \sum_{A_2(1)\subseteq A_1(1)} \left[ \left( \prod_{j\in A_2(1)} \lambda_j(1) \right) \times \left( \prod_{k\notin A_2(1), k\in A_1(1)} (1-\lambda_k(1)) \right) J_{\alpha_k}^{\pi}(\Lambda(1)) \right], \tag{55}$$

we have

$$\begin{aligned} J(\Lambda(1)) + h^{\pi}(\Lambda(1)) &= \max_{\pi} \lim_{k\to+\infty} \left[ J_{\alpha_k}^{\pi}(\Lambda(1)) - \alpha_k J_{\alpha_k}^{\pi}(\Lambda(1)) \right] \\ &= \max_{\pi} \lim_{k\to+\infty} \left[ \sum_{i\in A_1(1)} \lambda_i + \alpha_k \times \sum_{A_2(1)\subseteq A_1(1)} \left[ \left( \prod_{j\in A_2(1)} \lambda_j(1) \right) \times \left( \prod_{k\notin A_2(1), k\in A_1(1)} (1-\lambda_k(1)) \right) \times \left( J_{\alpha_k}^{\pi}(\Lambda(2) \mid a(1)) - J_{\alpha_k}^{\pi}(\Lambda(1)) \right) \right] \right]. \end{aligned} \tag{56}$$

Consequently, we have

$$\begin{aligned} J(\Lambda(1)) + h^{\pi}(\Lambda(1)) &= \max_{\pi} \left[ \sum_{i\in A_1(1)} \lambda_i + \sum_{A_2(1)\subseteq A_1(1)} \left[ \left( \prod_{j\in A_2(1)} \lambda_j(1) \right) \times \left( \prod_{k\notin A_2(1), k\in A_1(1)} (1-\lambda_k(1)) \right) \times h^{\pi}(\Lambda(2)) \right] \right]. \end{aligned} \tag{57}$$

We thus can conclude that the stationary deterministic policy realizing the pointwise maximum on the right-hand side of the ACOE is the average optimal policy due to the boundedness of $h^{\pi}(\Lambda(1))$, and $J(\Lambda(1))$ is the maximum average expected reward [20, Theorems 4.1–4.3].

For

$$J(\Lambda(1)) = \max_{\pi} \lim_{k\to+\infty} (1-\alpha_k) J_{\alpha_k}^{\pi}(\Lambda(1)), \tag{58}$$

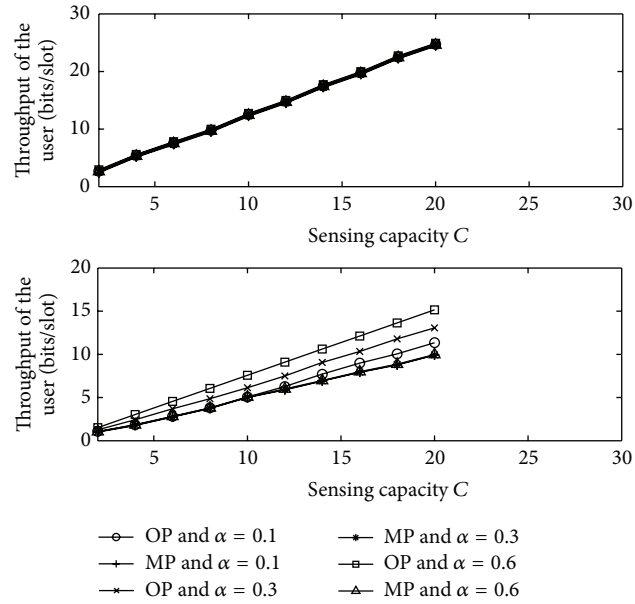


Figure 2: Performance comparison of OP and MP when $p_{11} \geq p_{01}$ and $p_{11} \leq p_{01}$.

the optimal policy for the discounted reward can maximize $J(\Lambda(1))$. Consequently, the myopic policy is optimal for the average reward when it is optimal for the discounted reward. The proof is complete. □

## 5. Numerical Results

We consider twenty independent channels with the same bandwidth $B = 1$ and transition probabilities $\{p_{11}, p_{01}\}$. The sensing capacity of the user who transmits the data on these channels is limited to $C$; that is, the user can sense $C$ channels in a sensing procedure. The transition probabilities are set as follows: $p_{11} = 0.8$ and $p_{01} = 0.5$ when $p_{11} \geq p_{01}$, $p_{11} = 0.3$ and $p_{01} = 0.5$ when $p_{11} \leq p_{01}$. The detection probability $p_d$ equals 0.9, and the false alarm probability $p_f$ equals 0.05. We present the numerical results to evaluate the performance of the optimal policy (OP) which is the dynamic programming solution and the myopic policy (MP).

We first use the throughput of the policies to evaluate the performance. The above subfigure of Figure 2 shows the performance comparison of OP and MP when $p_{11} \geq p_{01}$. The myopic policy is the optimal policy in this case for the reason that Assumptions 4 and 5 are met. We observe that the performance of OP is similar with MP's. The following subfigure shows the performance comparison of OP and MP when $p_{11} \leq p_{01}$. We observe that there is a large difference between OP and MP with the growth of the sensing capacity $C$. The reason is that Assumption 13 is not met. For example, the curves of MP and OP at $\alpha = 0.1$ separate when $C$ equals 11 for the reason that $\alpha \geq 1/(1+C)$.

Then, we use the collision probabilities which are the probabilities which the user accesses occupied channels to evaluate the performance. The collision probability is the key metric which measures the interference caused by the

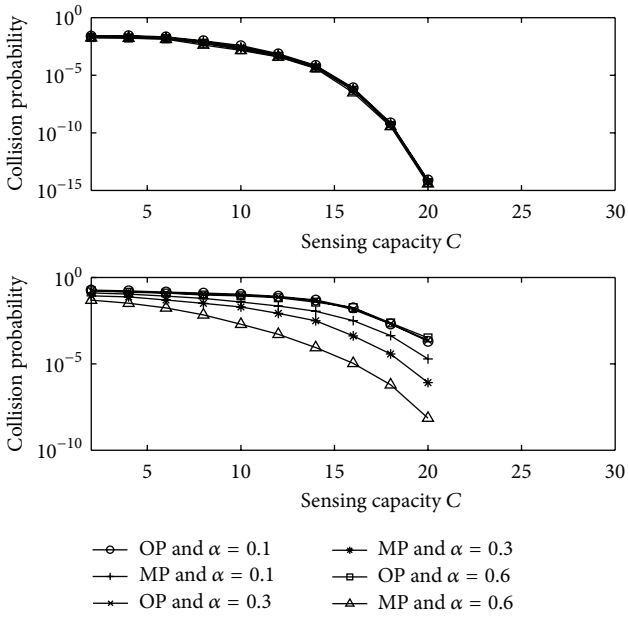


Figure 3: The comparison of collision probabilities of OP and MP when $p_{11} \geq p_{01}$ and $p_{11} \leq p_{01}$.

user. Figure 3 gives us similar results of Figure 2. The above subfigure shows the collision probabilities of OP and MP when $p_{11} \geq p_{01}$. We observe that they have the same collision probabilities for the reason that the myopic policy is the optimal policy. The following subfigure shows the collision probabilities of OP and MP when $p_{11} \leq p_{01}$. The myopic policies at $\alpha = 0.1, 0.3, 0.6$ have the same collision probabilities for the reason that they are the same policy. On the other hand, the optimal policies are different at different $\alpha$.

At last, we also give the comparison of the time complexity of OP and MP in Table 1. Since $p_{11}$, $p_{01}$, and $\alpha$ do not make impact on the time complexity of the policies, we mainly consider the variation of the sensing capacity. The first column of the table is the sensing capacity. The second and third columns show the time overhead of OP and MP, respectively, and the unit is second. The time overhead of MP is nearly 0 as MP does not need to calculate any parameter. In particular, the time overhead of OP is also almost 0 when the sensing capacity is 20 for the reason that OP can directly choose the channels which are observed as 1(idle). From Table 1, we can find that the time complexity of MP is much smaller than OP's.

## 6. Conclusion

We show the optimality of the simple and robust myopic policy for the infinite horizon discounted and average reward criteria in the case where the stochastic evolution of channels can be modeled as the independent and identically distributed two-state Markov chains. The myopic policy is optimal when the state transitions are positively correlated and the detection probability and the false alarm probability are limited. The myopic policy is also optimal when the state transitions are negatively correlated and the discounted factor is limited.

Table 1: Comparison of time complexity of OP and MP.

| $C$ | Optimal policy | Myopic policy |
|---|---|---|
| 1 | 1217.44 | 0.12 |
| 2 | 1012.68 | 0.03 |
| 4 | 697.82 | 0.07 |
| 6 | 473.34 | 0.06 |
| 8 | 334.71 | 0.01 |
| 10 | 198.34 | 0.05 |
| 12 | 127.3 | 0.11 |
| 14 | 79.57 | 0.06 |
| 16 | 39.12 | 0.07 |
| 18 | 22.37 | 0.02 |
| 20 | 0.07 | 0.08 |

## Acknowledgments

This work was partly supported by National Natural Science Foundations of China under Grant no. 61074033 and no. 61233003, Doctoral Fund of Ministry of Education of China under Grant no. 20093402110019.

## References

[1] G. Staple and K. Werbach, "The end of spectrum scarcity," *IEEE Spectrum*, vol. 41, no. 3, pp. 48–52, 2004.

[2] Shared Spectrum Company, "Spectrum Occupancy Report for New York City during the Republican Convention," September 2004, http://www.sharedspectrum.com/?section=measurements.

[3] Z. Quan, S. Cui, H. V. Poor, and A. H. Sayed, "Collaborative wideband sensing for cognitive radios: an overview of challenges and solutions," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 60–73, 2008.

[4] J. Meng, W. Yin, H. Li, E. Hossain, and Z. Han, "Collaborative spectrum sensing from sparse observations in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 2, pp. 327–337, 2011.

[5] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[6] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[7] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, 1967.

[8] F. F. Digham, M.-S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," in *Proceedings of the International Conference on Communications (ICC '03)*, pp. 3575–3579, May 2003.

[9] V. G. Chavali and C. R. C. M. da Silva, "Collaborative spectrum sensing based on a new SNR estimation and energy combining method," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 8, pp. 4024–4029, 2011.

[10] S. S. Jeong, W. S. Jeon, and D. G. Jeong, "Collaborative spectrum sensing for multiuser cognitive radio systems," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 5, pp. 2564–2569, 2009.

[11] Y. Chen, "Optimum number of secondary users in collaborative spectrum sensing considering resources usage efficiency," *IEEE Communications Letters*, vol. 12, no. 12, pp. 877–879, 2008.

[12] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of Byzantine attacks in cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 774–786, 2011.

[13] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: a POMDP framework," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589–600, 2007.

[14] X. Wang, W. Chen, and Z. Cao, "Partially observable Markov decision process-based MAC-layer sensing optimisation for cognitive radios exploiting rateless-coded spectrum aggregation," *IET Communications*, vol. 6, no. 8, pp. 828–835, 2012.

[15] W. Lingcen, W. Wei, and Z. Zhaoyang, "A POMDP-based optimal spectrum sensing and access scheme for cognitive radio networks with hardware limitation," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '12)*, pp. 1281–1286, 2012.

[16] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4040–4050, 2009.

[17] A. Ghasemi and E. S. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 131–136, November 2005.

[18] X.-R. Cao, *Stochastic Learning and Optimization*, Springer, New York, NY, USA, 2007.

[19] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice Hall, Englewood Cliffs, NJ, USA, 1987.

[20] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-time controlled Markov processes with average cost criterion: a survey," *SIAM Journal on Control and Optimization*, vol. 31, no. 2, pp. 282–344, 1993.