www.ietdl.org

Published in IET Signal Processing
Received on 13th May 2011
Revised on 17th January 2012
doi: 10.1049/iet-spr.2011.0185

ISSN 1751-9675

Optimality of greedy policy for a class of standard reward function of restless multi-armed bandit problem

K. Wang[1,2] *Q. Liu*[1] *L. Chen*[2]

[1]*School of Information, Wuhan University of Technology, Hubei 430070, People's Republic of China*
[2]*LRI, Department of Computer Science, University of Paris-Sud XI, Orsay 91405, France*
E-mail: kehao.wang@whut.edu.cn

Abstract: In this study, the authors consider the restless multi-armed bandit problem, which is one of the most well-studied generalisations of the celebrated stochastic multi-armed bandit problem in decision theory. However, it is known to be PSPACE-Hard to approximate to any non-trivial factor. Thus, the optimality is very difficult to obtain because of its high complexity. A natural method is to obtain the greedy policy considering its stability and simplicity. However, the greedy policy will result in the optimality loss for its intrinsic myopic behaviour generally. In this study, by analysing one class of so-called standard reward function, the authors establish the closed-form condition about the discounted factor β such that the optimality of the greedy policy is guaranteed under the discounted expected reward criterion, especially, the condition $\beta = 1$ indicating the optimality of the greedy policy under the average accumulative reward criterion. Thus, this kind of standard reward function can easily be used to judge the optimality of the greedy policy without any complicated calculation. Some examples in cognitive radio networks are presented to verify the effectiveness of the mathematical result in judging the optimality of the greedy policy.

1 Introduction

We consider the system consisting of n uncontrolled Markov chains evolving independently in the discrete time. Each of those chains is an independent identically distributed (iid) two-state Markov process. The two states are denoted as 'good' state (state 1) and 'bad' state (state 0). The transition probabilities are p_{ij}, $i, j = 0,1$. In each time instance of the system, a secondary user (SU) is allowed to select k out of the n processes according to its strategy, and to observe their states (assuming the precise observation), while those processes not selected by the user will evolve according to their rules. The user would obtain some reward determined by the combination of those observed states of the k selected processes, that is collecting no reward if those states of k processes are observed 'bad'. The above selecting, observing and collecting process repeats until the user does not access the system. Obviously, this is a multi-armed bandit (MAB) problem [1] as well as partially observed Markov decision process (POMDP) problem which has been used and studied in the [2, 3]. Unfortunately, obtaining optimal solutions to a general restless bandit process is PSPACE-Hard [4], and analytical characterisations of the performance of the optimal policy are often intractable. Hence the greedy policy governing the channel selection is the suitable choice because it only focuses on the maximisation of the immediate reward ignoring its affect on the future reward. However, the greedy policy is generally not optimal.

Recently, arise two main research directions addressing the greedy policy of this kind of MAB problem. The first one is to seek the constant-factor approximation algorithm, such as [5] where a 68-approximation developed via the linear programming relaxation under the condition of $p_{11} > 0.5 > p_{01}$ for each arm, and a 2-approximation policy for a class of monotone restless bandit problem [6]. The relevant application in dynamic multichannel access is given in [7], where the authors established the indexability and obtained Whittle index in closed form for both discounted and average reward criteria. In [8], the authors developed efficient sampling policies – link sampling and node sampling – based on the Whittle's indices for tracking the topology of dynamic networks under sampling constraints, and proved its indexability under certain conditions. In [9], an analysis for simultaneous sensing of multiple primary user activity in cognitive radios was presented from a signal-processing perspective. In [10], downlink spatial multiplexing techniques were proposed to enable multiple SUs to share spectrum simultaneously without harmfully interfering the primary users.

Another research direction is to explore the optimal condition of greedy policy corresponding to a concrete application or scenario. Our work follows this line. Although many literatures have studied this problem, the immediate reward function in those works focuses only on the linear combination of those observed states, that is in [11], the optimality of the greedy policy was proved in choosing $k = 1$ of n channels in the case of positively correlated channels, and then extended to choose $k > 1$ channels in [12]. In our previous work [13, 14], we have

extended the work in [11] on another line to the scenario where the immediate reward function is the simplest non-linear combination of those observed states, and proved that the greedy policy is generally not optimal, which is contrary to the result of [12] where the immediate reward function is the linear combination of the observed states. The contrary conclusion makes it necessary to study the impact of immediate reward function on the optimality of greedy policy, which is one of the major incentives for this paper.

From the technical perspective, the optimality of greedy policy needs user prefer to exploit rather than to explore. One simplest approach to implement this mechanism is to adjust the balance between exploitation and exploration by the discounted factor β. In contrast, we notice the different optimality resulting from the nuance of immediate reward functions [12, 13], and then we focus only on a generic and basic class of immediate reward function formulated by the combination of variables of order 1, referred to as standard reward function. Therefore our objective is to derive the sufficient condition of the discounted factor such that the greedy policy is guaranteed to be optimal for the so-called standard reward function under the discounted accumulative reward criterion. If the discounted factor $\beta = 1$, the optimality of greedy policy for the discounted accumulative reward can be promoted to the optimality for the average expected reward on the time horizon of interest. Therefore we can judge the optimality of the greedy policy for the discounted accumulative and average expected reward according to the closed-form condition of β. To the best of our knowledge, very few results had been reported from this perspective.

Compared with other existing work on the optimality of greedy policy in MAB problem, our contribution is 3-fold:

- We analyse one special class of MAB problem where the immediate reward function is so-called standard one, and derive that the discounted accumulative reward function is also a standard reward function. Furthermore, we establish the optimality of greedy policy under the discounted accumulative reward criterion when $p_{11} > p_{01}$. The theoretical results demonstrate that the greedy policy choosing the best 1 or $n-1$ out of n channels is optimal when $0 < \beta \leq 1$. For the case of choosing k $(1 < k < n-1)$ channels, the greedy policy is optimal only when the discounted factor satisfies a simple closed-form condition.
- The major technique developed in this paper is largely based on the analytic properties of standard reward function, completely different from [11, 12] relying on the coupling argument. Besides the significant and practical application in cognitive radio networks, this technique serves as the key criterion to judge the optimality of greedy policy when the immediate reward function is the combination of the standard functions in other scenarios.
- We analyse two practical models in cognitive radio networks. The first model in cognitive radio networks involves the sensing order problem where the SU selects k $(1 < k < n)$ of n channels in order to maximise the probability of finding an idle channel. It is obvious that the immediate reward function is the order 1 non-linear combination of the availability probability of selected channels. The result demonstrates that the greedy policy is not optimal generally under the average expected reward, which is coherent with [13]. The second model is that a user chooses $k(1 \leq k < n)$ channels to access and receive a reward on the channel in good state. The immediate reward function is the linear combination of the availability probability of those selected channels. Our derived result is consistent with those in [11, 12] where the myopic policy choosing any number of channels is optimal.

The rest of the paper is organised as follows. Our model is formulated in Section 2. Section 3 analyses standard reward function. Section 4 gives the optimality theorem of the myopic policy. Three applications are given in Section 5. Finally, our conclusions are summarised in Section 6.

2 Problem formulation

As outlined in the introduction, we consider a user trying to access the system consisting of n independent and statistically identical channels, each given by a two state Markov chain. The set of n channels is denoted by $\mathcal{N}$, each indexed by $i = 1, 2, \ldots, n$, and the state of channel i denoted by $S_i(t) =$ (good), 0(bad). The system operates in discrete time steps indexed by t $(t = 1, 2, \ldots, T)$, where T is the time horizon of interest. Specifically, we assume that channels go through state transition at the beginning of slot t and then at time t the user makes the channel selection decision. Limited by hardware or sensing cost, at time t the user is allowed to choose k $(1 \leq k < n)$ of the n channels to sense, the chosen channel set denoted by $a^k(t) \subset \mathcal{N}$, $|a^k(t)| = k$.

Obviously, the user cannot observe the whole states $\boldsymbol{S}(t) = [0, 1]^n$ of the underlying system (i.e. the states of n channels). We know that a sufficient statistic of such a system for optimal decision making, or the information state of the system, is given by the conditional probability that each channel is in state 1 given all past actions and observations [2]. We denote this information state (also called belief vector) by $\boldsymbol{\Omega}(t) = [\omega_1(t), \ldots, \omega_n(t)] \in [0, 1]^n$, where $\omega_i(t)$ is the conditional probability that channel i is in state 1 at time t. Owing to the Markovian nature of the channel model, the future information state is only a function of the current information state and the current action, that is, it is independent of past history given the current information state and action. Given that the information state at time t is $\boldsymbol{\Omega}(t) \triangleq \{\omega_i(t), i \in \mathcal{N}\}$ and the sensing policy $a^k(t) \subset \mathcal{N}$ is taken, the belief vector at time $t+1$ can be updated using Bayes rule as shown in (1)

$$\omega_i(t+1) = \begin{cases} p_{11}, & i \in a^k(t), S_i(t) = 1 \\ p_{01}, & i \in a^k(t), S_i(t) = 0 \\ \tau(\omega_i(t)), & i \notin a^k(t) \end{cases} \tag{1}$$

where, $\tau(\omega_i(t)) = \omega_i(t)p_{11} + [1 - \omega_i(t)]p_{01}$, and $p_{11} > p_{01}$ is assumed in the rest of the paper.

The objective is to maximise the discounted accumulative reward over a finite horizon given in the following problem

$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \beta^{t-1} R_{\pi_t}(\boldsymbol{\Omega}(t)) | \boldsymbol{\Omega}(1)\right] \tag{2}$$

where $R_{\pi_t}(\boldsymbol{\Omega}(t))$ is the reward collected with the initial belief vector $\boldsymbol{\Omega}(1)$ [If no information on the initial system state is available, each entry of $\boldsymbol{\Omega}(1)$ can be set to the stationary distribution $\omega_0 = ((p_{01})/(1 + p_{01} - p_{11}))$.] when channels in the set $a^k(t) = \pi_t(\Omega(t))$ are selected, π_t specifies a mapping from the current information state $\boldsymbol{\Omega}(t)$ to a channel selection action $a^k(t) = \pi_t(\ \Omega(t)) \subset \mathcal{N}$.

Let $V_t(\Omega)$ be the value function, which represent the maximum expected discounted accumulative reward obtained from t to T given the initial belief vector $\boldsymbol{\Omega}(1)$. Let

$p_{01}[x]$ and $p_{11}[x]$ denote the vector $[\boldsymbol{p}_{01}, \ldots, \boldsymbol{p}_{01}]$ and $[\boldsymbol{p}_{11}, \ldots, \boldsymbol{p}_{11}]$ of length x. Thus, we arrive at the following optimality equation

$$\begin{aligned} V_T(\Omega(t)) &= \max_{a^k(t)\subset\mathcal{N}} E[R(\Omega(t))] = \max_{a^k(t)\subset\mathcal{N}} F(\Omega(t)) \\ V_t(\Omega(t)) &= \max_{a^k(t)\subset\mathcal{N}} [F(\Omega(t)) + \beta K_t(\Omega(t))] \\ K_t(\Omega(t)) &= \sum_{e\in\mathcal{P}(a^k(t))} \prod_{i\in e} \omega_i \prod_{j\in a^k(t)\setminus e} (1-\omega_j) V_{t+1} \\ &(p_{11}[|e|], \tau(\omega_{k+1}(t)), \ldots, \tau(\omega_n(t)), p_{01}[k-|e|]) \end{aligned} \tag{3}$$

where, $\mathcal{P}(a^k(t))$ represents the power set generated by the set $a^k(t)$, the expected immediate reward $F(\Omega(t))$ is $F{:}\Omega(t) \to R$, and $|e|$ is the cardinality of set e. On right side of (3), the reward that can be collected from slot t consists of two parts: the expected immediate reward $F(\Omega(t))$ and the future discounted accumulative reward $\beta K_t(\Omega(t))$ calculated by summing over all possible realisations of the k selected channels. In $K_t(\Omega(t))$, the channel state probability vector consists of three parts: a sequence of p_{11}s indicating those channels sensed to be in state 1 at time t; a sequence of values $\tau(\omega_j(t))$ for all $j \notin a^k(t)$; and a sequence of p_{01}s indicating those channels sensed to be in state 0 at time t.

Considering the computational complexity of the recursive structure (3), we should seek other policies but the optimal policy. One of the simplest policy is the greedy one in which the objective is to maximise the expected immediate reward $F(\Omega(t))$ at each time step. Thus, the greedy policy is given as follows

$$\hat{a}^k(t) = \arg\max_{a^k(t)\subset\mathcal{N}} F(\Omega(t)) \tag{4}$$

In the following sections, we will derive the sufficient condition of β to guarantee the optimality of the greedy policy. The key mathematical symbols used in this paper is tabulated in Table 1.

3 Standard reward function

In this section, we will define a class of standard reward function based on three basic and generic assumptions, and then prove the value function $V_t(\Omega)$ is also standard under the greedy policy.

Table 1 Index of mathematical symbols

β	discounted factor
n	total channel number
$\mathcal{N}$	set {1, 2, …, n}
k	channel number chosen at each slot
$S_i(t)$	the state of channel i at slot t
$\boldsymbol{S}(t)$	state vector at slot t
$\omega_i(t)$	the probability of channel i in state 1 at slot t
$\boldsymbol{\Omega}(t)$	belief vector at slot t
$a^k(t)$	sensing policy at slot t
$\hat{a}^k(t)$	greedy policy at slot t
$\mathcal{P}(a^k(t))$	power set generated by the core $a^k(t)$
$F(\Omega(t))$	expected immediate reward at slot t
$K_t(\Omega(t))$	expected accumulative reward
$V_t(\Omega(t))$	value function

3.1 Definition of standard reward function

For simplicity, we assume that $a^k(t) = \omega_1(t), \ldots, \omega_k(t)$, and then use $a^k(t) = 1, \ldots, k$ and $a^k(t) = \omega_1(t), \ldots, \omega_k(t)$ alternatively. Especially, we drop the time slot index of $\omega_i(t)$, and abuse $\omega_i(t)$ and ω_i alternatively without introducing ambiguity.

Three fundamental while natural assumptions about the immediate reward function are listed as follows:

Assumption 1 (symmetry): The immediate reward function $F(\Omega(t))$ is symmetric about $\omega_i(t)$, $\omega_j(t) \in a^k(t)$, that is

$$\begin{aligned} &F(\omega_1(t), \ldots, \omega_i(t), \ldots, \omega_j(t), \ldots, \omega_n(t)) \\ &\quad = F(\omega_1(t), \ldots, \omega_j(t), \ldots, \omega_i(t), \ldots, \omega_n(t)) \end{aligned} \tag{5}$$

Assumption 2 (affine): The immediate reward function $F(\Omega(t))$ is order 1 polynomial [$F(\Omega(t))$ is affine in each variable if all other variables hold constant] *of* $\omega_i(t)$, $1 \le i \le n$, that is

$$\begin{aligned} &F(\omega_1(t), \ldots, \omega_{i-1}(t), \omega_i(t), \omega_{i+1}(t), \ldots, \omega_n(t)) \\ &= \omega_i(t) F(\omega_1(t), \ldots, \omega_{i-1}(t), 1, \omega_{i+1}(t), \ldots, \omega_n(t)) \\ &\quad + (1-\omega_i(t)) F(\omega_1(t), \ldots, \omega_{i-1}(t), 0, \omega_{i+1}(t), \ldots, \omega_n(t)) \end{aligned} \tag{6}$$

Assumption 3 (monotonicity): The immediate reward function $F(\Omega(t))$ increases monotonically with $\omega_i(t)$, $1 \le i \le n$, that is

$$\begin{aligned} \omega_i'(t) \ge \omega_i(t) &\Rightarrow F(\omega_1(t), \ldots, \omega_i'(t), \ldots, \omega_n(t)) \\ &\ge F(\omega_1(t), \ldots, \omega_i(t), \ldots, \omega_n(t)) \end{aligned} \tag{7}$$

Note these assumptions are necessary and non-redundant. Moreover, these three assumptions are used to define a class of general functions, referred to as 'standard' immediate reward functions.

Definition 1: A reward function is standard one if it satisfies the aforementioned three assumptions.

On the basis of the three assumptions, we can obtain the following structure of the greedy policy for this class of standard reward function.

Definition 2: We assume that $\omega_1(t) \ge \omega_2(t) \ge \ldots \ge \omega_n(t)$ at slot t, then the greedy policy is to choose the first k best channels, that is to say, $\hat{a}^k(t) = \{1, 2, \ldots, k\}$.

In order to see the intrinsic structure of the standard immediate reward function, we give three basic examples.

Example 1: Considering the scenario in [12] where the user gets one unit of reward for each channel sensed good. In this example, the expected immediate reward function is $F(\Omega) = \sum_{i=1}^{k} \omega_i$. It can be easily verified that $F(\Omega)$ satisfies the above three assumptions and thus is 'standard'.

Example 2: Considering the scenario where the user gets one unit of reward only if all the channels are sensed to be good. Thus, the immediate reward function is formulated by $F(\Omega) = \prod_{i=1}^{k} \omega_i$, which is 'standard' one.

Example 3: Consider the scenario in [13] where the user gets one unit of reward if at least one channel is sensed good. In

doi: 10.1049/iet-spr.2011.0185

this case, the expected immediate reward function is $F(\Omega) = 1 - \prod_{i=1}^{k}(1-\omega_i)$, which is standard according to the three assumptions.

3.2 Feature of accumulative reward function

In this part, the accumulative reward function $V_t(\Omega(t))$ (also called value function) is proved to be standard reward function under the greedy policy, which consists of the main part of the proof for the optimality of greedy policy in the next section.

Lemma 1 (symmetry): Under the greedy policy from slot $t+1$, $V_t(\Omega(t))$ is symmetric about $\omega_i(t)$, $\omega_j(t)$, $1 \le i, j \le k$, that is

$$V_t(\omega_1(t), \ldots, \omega_i(t), \ldots, \omega_j(t), \ldots, \omega_n(t)) = V_t(\omega_1(t), \ldots, \omega_j(t), \ldots, \omega_i(t), \ldots, \omega_n(t))$$

Proof: The proof is given in Appendix 2. □

Lemma 2 (affine): Under the greedy policy from slot $t+1$, $V_t(\Omega(t))$ is an affine function of $\omega_i(t)$, $1 \le i \le n$ when all other $\omega_j(t)$, $j \ne i$, $1 \le j \le n$ hold constant.

Proof: The proof is given in Appendix 3. □

Lemma 3 (monotonicity): Under the greedy policy from slot $t+1$, $V_t(\Omega(t))$ increases monotonically with $\omega_i(t)$, $1 \le i \le n$, that is

$$\omega_i'(t) \ge \omega_i(t) \Rightarrow V_t(\omega_1(t), \ldots, \omega_i'(t), \ldots, \omega_n(t)) \ge V_t(\omega_1(t), \ldots, \omega_i(t), \ldots, \omega_n(t))$$

Proof: The proof is given in Appendix 4. □

Lemma 4: Under the greedy policy from slot $t+1$, $V_t(\Omega(t))$ is a standard reward function.

Proof: It is obvious that $V_t(\Omega(t))$ is a standard reward function according to Lemmas 1–3. □

In this section, we analyse the feature of the class of standard reward function $V_t(\Omega(t))$ under the greedy policy, and the optimality of the greedy policy for this class of function will be explored in the next section.

4 Optimality of greedy policy for standard reward function

In this section, we first give the main theorem of optimality for the class of standard reward function, which states the sufficient condition of discounted factor for the optimality of greedy policy. After introducing some useful lemmas, we will give the complete proof of the theorem of optimality.

Let ω_{-i} denote the vector except the ith element ω_i, and define

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

It is easy to verify that $F'_{\max} \ge F'_{\min} \ge 0$ based on the three basic assumptions.

The main theorem of optimality is firstly stated as follows:

Theorem 1: The greedy policy is optimal for $V_t(\Omega(t))$ when $p_{01} \le \omega_i(1) \le p_{11}$, $1 \le i \le n$ if $F(\Omega(t))$ is a standard reward function and β satisfies the following condition

$$0 \le \beta \le \frac{F'_{\min}}{F'_{\max}(1-(1-p_{11})^{n-k-1})} \tag{8}$$

In order to prove the Theorem 1, we introduce some useful lemmas firstly. Note Lemmas 5–7 hold under condition (8) in the rest of the paper.

Lemma 5: For $p_{11} \ge \omega_i \ge \omega_{i+1} \ge p_{01}$ $(k+1 \le i \le n-1)$, under the greedy policy from slot $t+1$, we have the following inequality for $t = 1, 2, \ldots, T$ if (8) holds

$$V_t(\omega_1, \ldots, \omega_i, \omega_{i+1}, \ldots, \omega_n) - V_t(\omega_1, \ldots, \omega_{i+1}, \omega_i, \ldots, \omega_n) \ge 0 \tag{9}$$

Lemma 6: For $1 > \omega_1(t) \ge \omega_2(t) \ge \cdots \ge \omega_n(t) > 0$, under the greedy policy from slot $t+1$, we have the following inequality for $t = 1, 2, \ldots, T$ if (8) holds

$$V_t(\omega_1, \ldots, \omega_{n-1}, \omega_n) - V_t(\omega_n, \omega_1, \ldots, \omega_{n-1}) \le F'_{\max} \tag{10}$$

Lemma 7: For $p_{11} \ge x \ge y \ge p_{01}$, under the greedy policy from slot $t+1$, we have the following inequality for $t = 1, 2, \ldots, T$ if (8) holds

$$V_t(\omega_1, \ldots, \omega_{k-1}, x, y, \ldots, \omega_n) - V_t(\omega_1, \ldots, \omega_{k-1}, y, x, \ldots, \omega_n) \ge 0 \tag{11}$$

Remark: We would like to point out the complicated dependence in the proving process where Lemma 5 depends on Lemmas 2, 6 and 7, Lemma 6 depends on Lemmas 6 and 7 and Lemma 7 depends on Lemmas 7 and 6. Therefore we give the proof of Lemmas 5, 6 and 7 together by backward induction over time horizon in Appendix 5.

After obtaining the Lemmas 5–7, we are ready to prove the Theorem 1.

Proof: The basic approach is by induction on t. It is obvious that the myopic policy is optimal at T. Now, assuming the optimality of the myopic policy for $t+1, \ldots, T-1$, we shall show the myopic policy is also optimal for t. Denote $i_1, \ldots, i_n$ as any one of permutations of $\mathcal{N}$. To prove the optimality of greedy policy in slot t, we need to prove

$$V_t(\omega_1, \ldots, \omega_k, \ldots, \omega_n) \ge V_t(\omega_{i_1}, \ldots, \omega_{i_k}, \ldots, \omega_{i_n}) \tag{12}$$

The proving process is same as the bubble sort algorithm, comparing each pair of adjacent items and swapping them if they are in the wrong order according to Lemmas 1, 5 and 7 until no swaps are needed, which indicates that the list is sorted to $V_t(\omega_1, \ldots, \omega_k, \ldots, \omega_n)$. The optimality of greedy policy at slot t is guaranteed. Therefore the Theorem 1 is concluded. □

Corollary 1: The greedy policy is optimal if choosing 1 out of n channels for $0 < \beta \le 1$ if $p_{11} > p_{01}$.

Proof: When $k = 1$, according to Lemmas 1–3, we have $F(\Omega(t)) = a\omega_i(t)$, $a > 0$, hence

$$\frac{F'_{\min}}{F'_{\max}(1-(1-p_{11})^{n-k-1})} = \frac{1}{1-(1-p_{11})^{n-2}} > 1 \quad (13)$$

According to Theorem 1, we have the conclusion. □

Corollary 2: The greedy policy is optimal if choosing $n-1$ out of n channels for $0 < \beta \leq 1$.

Proof: In case of $k = n-1$, we have

$$\left[\frac{F'_{\min}}{F'_{\max}(1-(1-p_{11})^{n-k-1})}\right]_{k=n-1} \to \infty \quad (14)$$

Hence, the greedy policy is optimal according to Theorem 1. □

5 Applications in cognitive radio network

To illustrate the application of the mathematical results derived in the previous section, three typical scenarios [12, 13] described by standard reward function are presented here, which demonstrate that the different optimality is completely from different forms of immediate reward function.

5.1 Application 1

An application is in a synchronously slotted cognitive radio network where a SU can opportunistically access a set of n i.i.d. channels partially occupied by primary users. The state of each channel i in time slot t, denoted by $S_i(t)$, is modelled by a discrete time two-state Markov chain. At the beginning of each slot t, the SU selects a subset $\mathcal{A}(t)$ of channels to sense. If at least one of the sensed channels is in the idle state (i.e. unoccupied by any primary user), the SU transmits its packet and collects one unit of reward. Otherwise, the SU cannot transmit, thus obtaining no reward. The decision procedure is repeated for each slot. The objective is to maximise the average reward over T slots, that is to say, the discounted factor $\beta = 1$. Obviously, we have the immediate reward function $F(\Omega(t)) = 1 - \prod_{i\in\mathcal{A}(t)}(1-\omega_i(t))$. Therefore the greedy policy is to choose the best k channels by (4). According to Theorem 1, we have $F'_{\max} = (1-p_{01})^{k-1}$, $F'_{\min} = (1-p_{11})^{k-1}$ if $p_{01} \leq \omega_i(1) \leq p_{11}$, $1 \leq i \leq n$. Therefore the greedy policy, choosing the best k out of n channels, is optimal if the discounted factor β satisfies the following condition

$$0 \leq \beta \leq \frac{(1-p_{11})^{k-1}}{(1-p_{01})^{k-1}(1-(1-p_{11})^{n-k-1})}$$

Obviously, the upper bound cannot achieve 1 generally. Thus, the greedy policy, in general, is not optimal for the average reward over time horizon, which was proved in our previous work [13]. In particular, the greedy policy, choosing the best $k = 1$ or $n-1$ out of n channels is optimal for $\beta = 1$ according to the Corollaries 1 and 2.

5.2 Application 2

Consider the problem of probing n independent Markov chains. Each one has two states – good (1) and bad (0) – with transition probabilities p_{11}, p_{01} across chain. Assuming $p_{11} > p_{01}$. A player selects k chains to probe according to its preference (policy) and obtain a reward for each probed chain in the good state. We assume that the reward is affine function of the probability of the selected channel in the good state, that is, $u_i(t) = a\omega_i(t)$, $a > 0$, then we have the immediate reward function: $F(\Omega(t)) = a\sum_{i=1}^{k}\omega_i(t)$. As $F'_{\max} = F'_{\min} = a$, thus, $0 \leq \beta \leq 1 < (1/(1-(1-p_{11})^{n-k-1}))$. We have the following lemma by Theorem 1:

Lemma 8: The greedy policy of choosing the first k best channels is optimal for $0 < \beta \leq 1$.

Obviously, this result is consistent with [11, 12].

5.3 Application 3

Consider the scenario where a player detects n independent Markov chains. Each one has two states – good (1) and bad (0) – with transition probabilities p_{11}, p_{01} ($p_{11} > p_{01}$) across chain. The player selects k chains to detect according to its policy and obtain one unit of reward if all detected channels are good; otherwise, no reward. We assume that the probability of i channel in good state at time t is $\omega_i(t)$, then we have the immediate reward function: $F(\Omega(t)) = \Pi_{i=1}^{k}\omega_i(t)$. Therefore the greedy policy is to detect the first k best channels, and $F'_{\max} = p_{11}^{k-1}$, $F'_{\min} = p_{01}^{k-1}$. We have the following conclusion by Theorem 1

$$0 \leq \beta \leq \frac{p_{01}^{k-1}}{p_{11}^{k-1}(1-(1-p_{11})^{n-k-1})}$$

So in case of $1 < k < n-1$ the greedy policy is not optimal generally for $\beta = 1$, while choosing the best $k = 1$ or $k = n-1$ out of n channels is optimal for $0 < \beta \leq 1$.

6 Conclusion

In this study, we have considered a class of POMDP problem arisen in the fields of cognitive radio network, server scheduling and downlink scheduling in cellular systems, characterised by the so-called standard reward function. For this class of POMDP, we establish the optimal condition of the greedy policy focusing only on the maximisation of the immediate reward. The technical approach analysing this problem is purely mathematical, and thus is general for other models involving the recursive backward induction on the time horizon. The future direction is to investigate non-i.i.d Markov chain model through the proposed method, and another more challenging work is to extend standard reward function by dropping at least one of the three basic assumptions.

7 Acknowledgments

The authors thank the editor and the anonymous referee for their valuable comments and suggestions that improved the clarity and quality of this manuscript. This work was supported in part by China Scholarship Council (No. 2010695006), the Chenguang Youth Science and Technology Development Program of Wuhan (No. 201271031371) and the National Natural Science Foundation of China (No. 51175389).

8 References

1 Whittle, P.: 'Multi-armed bandits and the Gittins index', *J. Royal Stat. Soc. B.*, 1980, **42**, (2), pp. 143–149

doi: 10.1049/iet-spr.2011.0185

2 Smallwood, R., Sondik, E.: 'The optimal control of partitially observable markov processes over a finite horizon', *Oper. Res.*, 1971, **21**, (5), pp. 1071–1088
3 Zhao, Q., Tong, L., Swami, A., Chen, Y.: 'Decentralized cognitive mac for opportunistic spectrum access in ad hoc networks: a POMDP framework', *IEEE J. Sel. Areas Commun.*, 2007, **25**, (3), pp. 589–600
4 Papadimitriou, C.H., Tsitsiklis, J.N.: 'The complexity of optimal queueing network control', *Math. Oper. Res.*, 1999, **24**, (2), pp. 293–305
5 Guha, S., Munagala, K.: 'Approximation algorithms for partial-information based stochastic control with markovian rewards'. Proc. IEEE Symp. on Foundations of Computer Science (FOCS), Providence, RI, October 2007, pp. 483–493
6 Guha, S., Munagala, K.: 'Approximation algorithms for restless bandit problems'. Proc. ACM-SIAM Symp. on Discrete Algorithms (SODA), New York, USA, December 2009, pp. 28–37
7 Liu, K., Zhao, Q.: 'Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access', *IEEE Trans. Inf. Theory*, 2010, **56**, (11), pp. 5547–5567
8 He, T., Anandkumar, A., Agrawal, D.: 'Index-based sampling policies for tracking dynamic networks under sampling constraints'. INFOCOM 2011, Shanghai, China, April 2011, pp. 1233–1241
9 Sheikh, F., Masud, S., Bing, B.: 'Harmonic power detection in wideband cognitive radios', *IET Signal Process.*, 2009, **3**, (1), pp. 40–50
10 Cumanan, K., Krishna, R., Xiong, Z., Lambotharan, S.: 'Multiuser spatial multiplexing techniques with constraints on interference temperature for cognitive radio networks', *IET Signal Process.*, 2010, **4**, (6), pp. 666–672
11 Ahmand, S., Liu, M., Javidi, T., Zhao, Q., Krishnamachari, B.: 'Optimality of myopic sensing in multichannel opportunistic access', *IEEE Trans. Inf. Theory*, 2009, **55**, (9), pp. 4040–4050
12 Ahmad, S., Liu, M.: 'Multi-channel opportunistic access: A case of restless bandits with multiple players'. Proc. Allerton Conf. Commun. Control Comput., Monticello, IL, October 2009, pp. 1361–1368
13 Wang K., Chen L., Al Agha, K., Liu, Q.: 'On optimality of myopic policy in opportunistic spectrum access: the case of sensing multiple channels and accessing one channel'. IEEE Wireless Communications Letters, DOI: 10.1109/WCL.2012.12.120326
14 Wang, K., Chen, L.: 'On optimality of myopic policy for restless multi-armed bandit problem: an axiomatic approach', *IEEE Trans. Signal Process.*, 2012, **60**, (1), pp. 300–309

9 Appendices

9.1 Appendix 1. Proof of Lemma 9

Lemma 9: Assume $a^k(t) = \omega_1(t), \ldots \omega_k(t)$, $K_t(\Omega(t))$ is symmetric about $\omega_i(t)$, $\omega_j(t)$ for all $1 \le i, j \le k$, that is

$$K_t(\omega_1(t), \ldots, \omega_i(t), \ldots, \omega_j(t), \ldots, \omega_n(t)) = K_t(\omega_1(t), \ldots, \omega_j(t), \ldots, \omega_i(t), \ldots, \omega_n(t))$$

Proof: For the conciseness of presentation, we introduce a variable $K_t^m(\Omega(t))$ as follows

$$K_t^m(\Omega(t)) = \sum_{\substack{e \in \mathcal{P}(a^k(t)) \\ |e|=m}} \prod_{i \in e} \omega_i \prod_{j \in a^k(t) \setminus e} (1-\omega_j) V_{t+1}(p_{11}[|e|], \tau(\omega_{k+1}), \ldots, \tau(\omega_n), p_{01}[k-|e|]) \quad (15)$$

As e is the subset of the power set $\mathcal{P}(a^k(t))$ generated by the core $a^k(t)$, thus $0 \le |e| \le k$, and furthermore, $K_t(\Omega(t)) = \sum_{m=0}^{k} K_t^m(\Omega(t))$. Obviously, $V_{t+1}(p_{11}[|e|], \tau(\omega_{k+1}), \ldots, \tau(\omega_n), p_{01}[k-|e|])$ is unrelated with $a^k(t)$, so we only need to prove the $k+1$ coefficients are symmetric about $\omega_i(t)$, $\omega_j(t)$ for all $1 \le i, j \le k$, that is

$$\mathcal{C}_t^m = \sum_{\substack{e \in \mathcal{P}(a^k(t)) \\ |e|=m}} \prod_{i \in e} \omega_i \prod_{j \in a^k(t) \setminus e} (1-\omega_j), \quad 0 \le m \le k$$

is symmetric about $\omega_i(t)$, $\omega_j(t)$. On the basis of the feature of power set $\mathcal{P}(a^k(t))$, it is simple to obtain that $\mathcal{C}_t^m$ $(0 \le m \le k)$ is symmetric about any two $\omega_i(t)$, $\omega_j(t) \in a^k(t)$. Therefore $K_t(\Omega(t))$ is symmetric about $\omega_i(t)$, $\omega_j(t) \in a^k(t)$. □

9.2 Appendix 2. Proof of Lemma 1

1. According to assumption 1, for $1 \le i \ne j \le k$ in time slot T, as $V_{\mathcal{T}}(\Omega(T)) = F(\Omega(T))$, then it is easy to verify that $V_{\mathcal{T}}(\Omega(T))$ is symmetric.
2. Assume $V_{T-1}(\Omega(t)), \ldots, V_{t+2}(\Omega(t))$, $V_{t+1}(\Omega(t))$ are symmetric, then at slot t we have $V_t(\Omega(t)) = F(\Omega(t)) + \beta K_t(\Omega(t))$. On the basis of Assumption 1, $F(\Omega(t))$ is symmetric. According to Lemma 9 (Appendix 1), $K_t(\Omega(t))$ is also symmetric. Hence, $V_t(\Omega(t))$ is symmetric.

9.3 Appendix 3. Proof of Lemma 2

1. According to Assumption 2, in time slot T, $F(\Omega(T))$ is affine function of $\omega_i(T)$, $1 \le i \le n$. Hence, $V_{\mathcal{T}}(\Omega(T)) = F(\Omega(T))$ is also affine function of $\omega_i(T)$.
2. Assume $V_{T-1}(\Omega(T-1)), \ldots, V_{t+2}(\Omega(t+2))$, $V_{t+1}(\Omega(t+1))$ are affine functions, we prove it also holds for slot t. Two cases should be considered as follows:

Case 1: Channel $\omega_i \notin a^k(\mathrm{t}) = \{\omega_1, \ldots, \omega_k\}$, we have

$$V_t(\Omega(t)) = F(\Omega(t)) + \beta \sum_{e \in \mathcal{P}(a^k(t))} \prod_{p \in e} \omega_p \prod_{q \in a^k(t) \setminus e} (1-\omega_q) \times V_{t+1}(p_{11}[|e|], \ldots, \tau(\omega_i), \ldots, \tau(\omega_n), p_{01}[k-|e|])$$

By the induction hypothesis, $V_{t+1}(\Omega(t+1))$ is the affine function of $\tau(\omega_i)$, and meanwhile, $\tau(\omega_i)$ is an affine transform of ω_i, thus $V_{t+1}(\Omega(t+1))$ is the affine function of ω_i. Considering $F(\Omega(t))$ is unrelated with ω_i, we have $V_t(\Omega(t))$ is the affine function of ω_i.

Case 2: Channel $\omega_i \in a^k(t)$, let $a^{k-1}(t) = a^k(t) - \{\omega_i\}$, we have (see equation at the bottom of the page)

By Assumption 2, $F(\omega_1, \ldots, \omega_i, \ldots, \omega_k)$ is the affine function of ω_i. Obviously, the second term of the right hand of the above formulation is also the affine function of ω_i. Therefore $V_t(\Omega(t))$ is the affine function of ω_i.

$$V_t(\Omega(t)) = F(\Omega(t)) + \beta \sum_{e \in \mathcal{P}(a^k(t))} \prod_{p \in e} \omega_p \prod_{q \in a^k(t) \setminus e} (1-\omega_q) V_{t+1}(p_{11}[|e|], \tau(\omega_{k+1}), \ldots, \tau(\omega_n), p_{01}[k-|e|])$$

$$= F(\omega_1, \ldots, \omega_i, \ldots, \omega_k) + \beta \sum_{m=0}^{k-1} \sum_{|e|=m e \in \mathcal{P}(a^{k-1}(t))} \prod_{p \in e} \omega_p \prod_{q \in a^{k-1}(t) \setminus e} (1-\omega_q) \{\omega_i V_{t+1}(p_{11}[|e|], p_{11}, \tau(\omega_{k+1}), \ldots, \tau(\omega_n), p_{01}[k-|e|])$$

$$+ (1-\omega_i) V_{t+1}(p_{11}[|e|], \tau(\omega_{k+1}), \ldots, \tau(\omega_n), p_{01}, p_{01}[k-|e|])\}$$

Combining the two cases, we have $V_t(\Omega(t))$ is the affine function of ω_i. Lemma 2 is concluded.

9.4 Appendix 4. Proof of Lemma 3

1. The lemma holds trivially for slot T considering $V_T(\Omega(T)) = F(\Omega(T))$, which is the increasing function with ω_i.
2. Assume $V_{T-1}(\Omega(T-1)), \ldots, V_{t+2}(\Omega(t+2)), V_{t+1}(\Omega(t+1))$ increase monotonically, we prove it is true for slot t by two different cases.

Case 1: Channel $\omega_i \notin a^k(t)$, we have

$$V_t(\Omega(t)) = F(\Omega(t)) + \beta \sum_{e\in\mathcal{P}(a^k(t))} \prod_{p\in e} \omega_p \prod_{q\in a^k(t)\setminus e} (1-\omega_q) \times V_{t+1}(p_{11}[|e|],\ \ldots,\ \tau(\omega_i),\ \ldots,\ \tau(\omega_n),\ p_{01}[k-|e|])$$

Obviously, $\tau(\omega_i)$ increases with ω_i when $p_{11} > p_{01}$, and $V_{t+1}(\Omega(t+1))$ increases with $\tau(\omega_i)$ according to the induction hypothesis, thus $V_{t+1}(\Omega(t+1))$ also increases with ω_i. As $F(\Omega(t))$ is unrelated with ω_i, we have $V_t(\Omega(t))$ is the increasing function of ω_i.
Case 2: Channel $\omega_i \in a^k(t)$, let $a^{k-1}(t) = a^k(t) - \{\omega_i\}$, we have (see equation at the bottom of the page)

where, the first term, $F(\omega_1, \ldots, \omega_i, \ldots, \omega_k)$, of the right hand of the above formulation increases monotonically with ω_i according to Assumption 3, and the second term is also the increasing function of ω_i because (see equation at the bottom of the page)

where, $p_{11} \geq \tau(\omega_{k+1}) \geq \ldots \geq \tau(\omega_n) \geq p_{01}$ according to the increasing monotonicity of $\tau(\omega_i)$ in ω_i when $p_{11} > p_{01}$, and thus, each term in bracket is larger than or equals zero according to the induction hypothesis.

Therefore we have $V_t(\Omega(t))$ increases monotonically with ω_i through the two cases and complete the proof.

9.5 Appendix 5. Proof of Lemmas 5–7

Proof: The proving process is based on backward induction in three steps as follows:

Step 1: In slot T, these lemmas hold trivially noticing $V_T(\Omega(T) = F(\Omega(T)))$.

For Lemma 5

$$V_T(\omega_1, \ldots, \omega_i, \omega_{i+1}, \ldots, \omega_n) - V_T(\omega_1, \ldots, \omega_{i+1}, \omega_i, \ldots, \omega_n) = F(\omega_1, \ldots, \omega_k) - F(\omega_1, \ldots, \omega_k) = 0$$

For Lemma 6

$$\begin{aligned} &V_T(\omega_1, \ldots, \omega_{n-1}, \omega_n) - V_T(\omega_n, \omega_1, \ldots, \omega_{n-1}) \\ &= F(\omega_1, \ldots, \omega_{k-1}, \omega_k) - F(\omega_n, \omega_1, \ldots, \omega_{k-1}) \\ &= (\omega_k - \omega_n)(F(\omega_1, \ldots, \omega_{k-1}, 1) - F(\omega_1, \ldots, \omega_{k-1}, 0)) \\ &\leq F'_{\max} \end{aligned}$$

where, the second equality is due to Lemmas 1 and 2.

$$\begin{aligned} V_t(\Omega(t)) &= F(\Omega(t)) + \beta \sum_{e\in\mathcal{P}(a^k(t))} \prod_{p\in e} \omega_p \prod_{q\in a^k(t)\setminus e} (1-\omega_q) V_{t+1}(p_{11}[|e|],\ \tau(\omega_{k+1}),\ \ldots,\ \tau(\omega_n),\ p_{01}[k-|e|]) \\ &= F(\omega_1,\ \ldots,\ \omega_i,\ \ldots,\ \omega_k) + \beta \sum_{m=0}^{k-1} \sum_{|e|=m e\in\mathcal{P}(a^{k-1}(t))} \prod_{p\in e} \omega_p \prod_{q\in a^{k-1}(t)\setminus e} (1-\omega_q) \\ &\quad \times \{\omega_i V_{t+1}(p_{11}[m],\ p_{11},\ \tau(\omega_{k+1}),\ \ldots,\ \tau(\omega_n),\ p_{01}[k-m]) \\ &\quad + (1-\omega_i) V_{t+1}(p_{11}[m],\ \tau(\omega_{k+1}),\ \ldots,\ \tau(\omega_n),\ p_{01},\ p_{01}[k-m])\} \\ &= F(\omega_1,\ \ldots,\ \omega_i,\ \ldots,\ \omega_k) + \sum_{m=0}^{k-1} \sum_{|e|=m e\in\mathcal{P}(a^{k-1}(t))} \prod_{p\in e} \omega_p \prod_{q\in a^{k-1}(t)\setminus e} (1-\omega_q) \\ &\quad \times \{\omega_i [V_{t+1}(p_{11}[m],\ p_{11},\ \tau(\omega_{k+1}),\ \ldots,\ \tau(\omega_n),\ p_{01}[k-m]) \\ &\quad - V_{t+1}(p_{11}[m],\ \tau(\omega_{k+1}),\ \ldots,\ \tau(\omega_n),\ p_{01},\ p_{01}[k-m])] \\ &\quad + V_{t+1}(p_{11}[m],\ \tau(\omega_{k+1}),\ \ldots,\ \tau(\omega_n),\ p_{01},\ p_{01}[k-m])\} \\ &\quad \geq 0 \end{aligned}$$

$$\begin{aligned} &V_{t+1}(p_{11}[m],\ p_{11},\ \tau(\omega_{k+1}),\ \tau(\omega_{k+2}),\ \ldots,\ \tau(\omega_{n-1}),\ \tau(\omega_n),\ p_{01}[k-m]) \\ &\quad - V_{t+1}(p_{11}[m],\ \tau(\omega_{k+1}),\ \tau(\omega_{k+2}),\ \ldots,\ \tau(\omega_{n-1}),\ \tau(\omega_n),\ p_{01},\ p_{01}[k-m]) \\ &\quad = [V_{t+1}(p_{11}[m],\ p_{11},\ \tau(\omega_{k+1}),\ \tau(\omega_{k+2}),\ \ldots,\ \tau(\omega_{n-1}),\ \tau(\omega_n),\ p_{01}[k-m]) \\ &\quad\quad - V_{t+1}(p_{11}[m],\ \tau(\omega_{k+1}),\ \tau(\omega_{k+1}),\ \tau(\omega_{k+2}),\ \ldots,\ \tau(\omega_{n-1}),\ \tau(\omega_n),\ p_{01}[k-m])] \\ &\quad\quad + [V_{t+1}(p_{11}[m],\ \tau(\omega_{k+1}),\ \tau(\omega_{k+1}),\ \tau(\omega_{k+2}),\ \ldots,\ \tau(\omega_{n-1}),\ \tau(\omega_n),\ p_{01}[k-m]) \\ &\quad\quad - V_{t+1}(p_{11}[m],\ \tau(\omega_{k+1}),\ \tau(\omega_{k+2}),\ \tau(\omega_{k+2}),\ \ldots,\ \tau(\omega_{n-1}),\ \tau(\omega_n),\ p_{01}[k-m])] \\ &\quad\quad + \cdots + [V_{t+1}(p_{11}[m],\ \tau(\omega_{k+1}),\ \tau(\omega_{k+2}),\ \ldots,\ \tau(\omega_{n-1}),\ \tau(\omega_n),\ \tau(\omega_n),\ p_{01}[k-m]) \\ &\quad\quad - V_{t+1}(p_{11}[m],\ \tau(\omega_{k+1}),\ \tau(\omega_{k+2}),\ \ldots,\ \tau(\omega_{n-1}),\ \tau(\omega_n),\ p_{01},\ p_{01}[k-m])] \\ &\quad\quad \geq 0 \end{aligned}$$

doi: 10.1049/iet-spr.2011.0185

For Lemma 7

$$V_T(\omega_1, \ldots, \omega_{k-1}, x, y, \ldots, \omega_n) - V_T(\omega_1, \ldots, \omega_{k-1}, y, x, \ldots, \omega_n)$$
$$= F(\omega_1, \ldots, \omega_{k-1}, x) - F(\omega_1, \ldots, \omega_{k-1}, y)$$
$$= (x-y)(F(\omega_1, \ldots, \omega_{k-1}, 1) - F(\omega_1, \ldots, \omega_{k-1}, 0))$$
$$\geq (x-y)F'_{\min} \geq 0$$

Step 2: Assume at $T-1, \ldots, t+1$, Lemmas 5 (Induction Hypothesis 1, IH1), 6 (Induction Hypothesis 2, IH2), and 7 (Induction Hypothesis 3, IH3) hold, we thus prove these lemmas also hold at slot t.

Step 3: At slot t,

For Lemma 5 (see equation at the bottom of the page)

where, $a^k(t) = \omega_1, \ldots, \omega_k$, the first equality is due to Lemma 2, the inequality is due to IH1 if $|e| + i - k - 1 \geq k$, IH3 if $|e| + i - k - 1 = k - 1$, and Lemma 1 if $|e| + i - k - 1 < k - 1$.

For Lemma 6, we have the following decomposition according to Lemma 2 (see (16))

Therefore we analyse the above formulation through four cases as follows:

Case 1: For the first term of the right hand of (16), we denote $a^{k-1}(t) = \{\omega_1, \omega_2, ..., \omega_{k-1}\}$, and thus have where, the first inequality is due to Lemma 3. (see equation at the bottom of the page)

Case 2: For the second term of the right hand of (16), we denote $a^{k-1}(t) = \{\omega_1, \omega_2, \ldots, \omega_{k-1}\}$, and have (see equation at the bottom of the page)

Case 3: For the third term of the right hand of (16), we denote $a^{k-1}(t) = \{\omega_1, \omega_2, \ldots, \omega_{k-1}\}$, and have (see equation at the bottom of the page)

$$V_t(\omega_1, \ldots, \omega_k, \ldots, \omega_i, \omega_{i+1}, \ldots, \omega_n) - V_t(\omega_1, \ldots, \omega_k, \ldots, \omega_{i+1}, \omega_i, \ldots, \omega_n)$$
$$= (\omega_i - \omega_{i+1})(V_t(\omega_1, \ldots, \omega_{i-1}, 1, 0, \omega_{i+2}, \ldots, \omega_n) - V_t(\omega_1, \ldots, \omega_{i-1}, 0, 1, \omega_{i+2}, \ldots, \omega_n))$$
$$= (\omega_i - \omega_{i+1})\beta \sum_{e \in \mathcal{P}(a^k(t))} \prod_{i \in e} \omega_i \prod_{j \in a^k(t) \setminus e} (1 - \omega_j)$$
$$\times \{V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{i-1}), p_{11}, p_{01}, \pi(\omega_{i+2}), \ldots, \pi(\omega_n), p_{01}[k - |e|])$$
$$- V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{i-1}), p_{01}, p_{11}, \pi(\omega_{i+2}), \ldots, \pi(\omega_n), p_{01}[k - |e|])\}$$
$$\geq 0$$

$$V_t(\omega_1, \ldots, \omega_{k-1}, \omega_k, \ldots, \omega_{n-1}, \omega_n) - V_t(\omega_n, \omega_1, \ldots, \omega_{k-1}, \omega_k, \ldots, \omega_{n-1})$$
$$= \omega_k \omega_n [V_t((\omega_1, \ldots, \omega_{k-1}, 1, \omega_{k+1}, \ldots, \omega_{n-1}, 1) - V_t(1, \omega_1, \ldots, \omega_{k-1}, 1, \omega_{k+1}, \ldots, \omega_{n-1})]$$
$$+ \omega_k(1 - \omega_n)[V_t((\omega_1, \ldots, \omega_{k-1}, 1, \omega_{k+1}, \ldots, \omega_{n-1}, 0) - V_t(0, \omega_1, \ldots, \omega_{k-1}, 1, \omega_{k+1}, \ldots, \omega_{n-1})]$$
$$+ (1 - \omega_k)\omega_n[V_t((\omega_1, \ldots, \omega_{k-1}, 0, \omega_{k+1}, \ldots, \omega_{n-1}, 1) - V_t(1, \omega_1, \ldots, \omega_{k-1}, 0, \omega_{k+1}, \ldots, \omega_{n-1})]$$
$$+ (1 - \omega_k)(1 - \omega_n)[V_t((\omega_1, \ldots, \omega_{k-1}, 0, \omega_{k+1}, \ldots, \omega_{n-1}, 0) - V_t(0, \omega_1, \ldots, \omega_{k-1}, 0, \omega_{k+1}, \ldots, \omega_{n-1})]) \quad (16)$$

$$V_t(\omega_1, \omega_2, \ldots, \omega_{k-1}, 1, \omega_{k+1}, \ldots, \omega_{n-1}, 1) - V_t(1, \omega_1, \omega_2, \ldots, \omega_{k-1}, 1, \omega_{k+1}, \ldots, \omega_{n-1})$$
$$= F(\omega_1, \omega_2, \ldots, \omega_{k-1}, 1) - F(1, \omega_1, \omega_2, \ldots, \omega_{k-1}) + \beta \sum_{e \in \mathcal{P}(a^{k-1}(t))} \prod_{i \in e} \omega_i \prod_{j \in a^{k-1}(t) \setminus e} (1 - \omega_j)$$
$$\times \{V_{t+1}(p_{11}[|e|], p_{11}, \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), \pi(\omega_n), p_{01}[k - 1 - |e|])$$
$$- V_{t+1}(p_{11}[|e|], p_{11}, \pi(\omega_k), \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}[k - 1 - |e|])\}$$
$$= \beta \sum_{e \in \mathcal{P}(a^{k-1}(t))} \prod_{i \in e} \omega_i \prod_{j \in a^{k-1}(t) \setminus e} (1 - \omega_j)$$
$$\times \{V_{t+1}(p_{11}[|e|], p_{11}, \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{11}, p_{01}[k - 1 - |e|])$$
$$- V_{t+1}(p_{11}[|e|], p_{11}, p_{11}, \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}[k - 1 - |e|])\}$$
$$\leq 0 \leq F'_{\max}$$

$$V_t(\omega_1, \omega_2, \ldots, \omega_{k-1}, 1, \omega_{k+1}, \ldots, \omega_{n-1}, 0) - V_t(0, \omega_1, \omega_2, \ldots, \omega_{k-1}, 1, \omega_{k+1}, \cdots, \omega_{n-1})$$
$$= F(\omega_1, \omega_2, \ldots, \omega_{k-1}, 1) - F(0, \omega_1, \omega_2, \ldots, \omega_{k-1}) + \beta \sum_{e \in \mathcal{P}(a^{k-1}(t))} \prod_{i \in e} \omega_i \prod_{j \in a^{k-1}(t) \setminus e} (1 - \omega_j)$$
$$\times \{V_{t+1}(p_{11}[|e|], p_{11}, \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}, p_{01}[k - 1 - |e|])$$
$$- V_{t+1}(p_{11}[|e|], p_{11}, \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}, p_{01}[k - 1 - |e|])\}$$
$$= F(\omega_1, \omega_2, \ldots, \omega_{k-1}, 1) - F(0, \omega_1, \omega_2, \ldots, \omega_{k-1}) \leq F'_{\max}$$

where, the first inequality is due to IH3 when $|e|+1=k$, the second one due to IH2, and the second equality due to Lemma 1 when $|e|+1<k$, noticing $0\le|e|\le k-1$.

Case 4: For the forth term of the right hand of (16), we denote $a^{k-1}(t)=\{\omega_1, \omega_2, \ldots, \omega_{k-1}\}$, and have (see equation at the bottom of the page)

where, the first inequality is due to IH2 and the third equality is due to Lemma 1.

Combing the results of four cases and (16), we have

$$\begin{aligned}&V_t(\omega_1, \omega_2, \ldots, \omega_{k-1}, \omega_k, \ldots, \omega_{n-1}, \omega_n)\\&\quad - V_t(\omega_n, \omega_1, \omega_2, \ldots, \omega_{k-1}, \omega_k, \ldots, \omega_{n-1})\\&\le \omega_k\omega_n 0 + \omega_k(1-\omega_n)F'_{\max} + (1-\omega_k)\omega_n\beta F'_{\max}\\&\quad + (1-\omega_k)(1-\omega_n)\beta F'_{\max} \le F'_{\max}\end{aligned}$$

To this end, we complete the proof of Lemma 6.

$$\begin{aligned}&V_t(\omega_1, \omega_2, \ldots, \omega_{k-1}, 0, \omega_{k+1}, \ldots, \omega_{n-1}, 1) - V_t(1, \omega_1, \omega_2, \ldots, \omega_{k-1}, 0, \omega_{k+1}, \ldots, \omega_{n-1})\\&= F(\omega_1, \omega_2, \ldots, \omega_{k-1}, 0) - F(1, \omega_1, \omega_2, \ldots, \omega_{k-1}) + \beta \sum_{e\in\mathcal{P}(a^{k-1}(t))}\prod_{i\in e}\omega_i \prod_{j\in a^{k-1}(t)\setminus e}(1-\omega_j)\end{aligned}$$

$$\begin{aligned}&\times \{V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{11}, p_{01}, , p_{01}[k-1-|e|])\\&-V_{t+1}(p_{11}[|e|], p_{11}, p_{01}, \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}[k-1-|e|])\}\\&\le -F'_{\max} + \beta \sum_{e\in\mathcal{P}(a^{k-1}(t))}\prod_{i\in e}\omega_i \prod_{j\in a^{k-1}(t)\setminus e}(1-\omega_j)\times\end{aligned}$$

$$\begin{aligned}&\{V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{11}, p_{01}, , p_{01}[k-1-|e|])\\&-V_{t+1}(p_{11}[|e|], p_{01}, p_{11}, \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}[k-1-|e|])\}\\&= -F'_{\max} + \beta \sum_{e\in\mathcal{P}(a^{k-1}(t))}\prod_{i\in e}\omega_i \prod_{j\in a^{k-1}(t)\setminus e}(1-\omega_j)\end{aligned}$$

$$\begin{aligned}&\{V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{11}, p_{01}, , p_{01}[k-1-|e|])\\&-V_{t+1}(p_{01}, p_{11}[|e|], p_{11}, \pi(\omega_{k+1}), \cdots, \pi(\omega_{n-1}), p_{01}[k-1-|e|])\}\\&\le -F'_{\max} + \beta \sum_{e\in\mathcal{P}(a^{k-1}(t))}\prod_{i\in e}\omega_i \prod_{j\in a^{k-1}(t)\setminus e}(1-\omega_j)\end{aligned}$$

$$\begin{aligned}&\{V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \cdots, \pi(\omega_{n-1}), p_{11}, p_{01}, , p_{01}[k-1-|e|])\\&+F'_{\max} - V_{t+1}(p_{11}[|e|], p_{11}, \pi(\omega_{k+1}), \cdots, \pi(\omega_{n-1}), p_{01}, p_{01}[k-1-|e|])\}\\&\le -F'_{\max} + \beta F'_{\max} \le (\beta-1)F'_{\max} \le F'_{\max}\end{aligned}$$

$$\begin{aligned}&V_t(\omega_1, \omega_2, \ldots, \omega_{k-1}, 0, \omega_{k+1}, \ldots, \omega_{n-1}, 0) - V_t(0, \omega_1, \omega_2, \ldots, \omega_{k-1}, 0, \omega_{k+1}, \ldots, \omega_{n-1})\\&= F(\omega_1, \omega_2, \ldots, \omega_{k-1}, 0) - F(0, \omega_1, \omega_2, \ldots, \omega_{k-1}) + \beta \sum_{e\in\mathcal{P}(a^{k-1}(t))}\prod_{i\in e}\omega_i \prod_{j\in a^{k-1}(t)\setminus e}(1-\omega_j)\end{aligned}$$

$$\begin{aligned}&\times \{V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}, p_{01}, p_{01}[k-1-|e|])\\&- V_{t+1}(p_{11}[|e|], p_{01}, \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}, p_{01}[k-1-|e|])\}\\&= \beta \sum_{e\in\mathcal{P}(a^{k-1}(t))}\prod_{i\in e}\omega_i \prod_{j\in a^{k-1}(t)\setminus e}(1-\omega_j)\end{aligned}$$

$$\begin{aligned}&\times \{V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}, p_{01}, p_{01}[k-1-|e|])\\&\times V_{t+1}(p_{11}[|e|], p_{01}, \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}, p_{01}[k-1-|e|])\}\\&= \beta \sum_{e\in\mathcal{P}(a^{k-1}(t))}\prod_{i\in e}\omega_i \prod_{j\in a^{k-1}(t)\setminus e}(1-\omega_j)\end{aligned}$$

$$\begin{aligned}&\times \{V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-2}), \pi(\omega_{n-1}), p_{01}, p_{01}, p_{01}[k-1-|e|])\\&-V_{t+1}(p_{01}, p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}, p_{01}[k-1-|e|])\}\\&\le \beta \sum_{e\in\mathcal{P}(a^{k-1}(t))}\prod_{i\in e}\omega_i \prod_{j\in a^{k-1}(t)\setminus e}(1-\omega_j)\end{aligned}$$

$$\begin{aligned}&\times \{V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \pi(\omega_{k+2}), \ldots, \pi(\omega_{n-2}), \pi(\omega_{n-1}), p_{01}, p_{01}, p_{01}[k-1-|e|])\\&+F'_{\max} - V_{t+1}(p_{11}[|e|], \pi(\omega_{k+1}), \ldots, \pi(\omega_{n-1}), p_{01}, p_{01}, p_{01}[k-1-|e|])\}\\&\le \beta F'_{\max}\end{aligned}$$

doi: 10.1049/iet-spr.2011.0185

For Lemma 7 (see equation at the bottom of the page)

where, the third inequality is from condition (8) and the first inequality is due to the following inequality,

$$\begin{aligned}\Delta V &= V_{t+1}(p_{11}[|e|],p_{11},p_{01},\pi(\omega_{k+2}),\ldots,\pi(\omega_n),p_{01}[k-1-|e|]) \\ &\quad - V_{t+1}(p_{11}[|e|],p_{11},\pi(\omega_{k+2}),\ldots,\pi(\omega_n),p_{01},p_{01}[k-1-|e|]) \\ &\geq -\Big(1-\prod_{j=k+2}^{n}(1-\omega_j)\Big)F'_{\max} \qquad (17)\end{aligned}$$

Note, if $\pi(\omega_{k+2}(t)) = \cdots = \pi(\omega_n(t)) = p_{01}$, then $\Delta V = 0$, which corresponds to the event that $\omega_{k+2}(t) = \cdots = \omega_n(t) = 0$ at slot t. Obviously, this event happens with the probability $\prod_{j=k+2}^{n}(1-\omega_j)$. Thus with the probability $1-\prod_{j=k+2}^{n}(1-\omega_j)$, there exists at least i $(k+2 \leq i \leq n)$ such that $\pi(\omega_i) > p_{01}$ and furthermore, $\Delta V \neq 0$. According to IH2 and IH4, we have $\Delta V \geq -F'_{\max}$ with probability $1-\prod_{j=k+2}^{n}(1-\omega_j)$, which is (17).

Therefore we finish the whole proving process of Lemmas 5–7.

$$\begin{aligned}&V_t(\omega_1,\ \ldots,\ \omega_{k-1},x,y,\ \ldots,\ \omega_n) - V_t(\omega_1,\ \ldots,\ \omega_{k-1},y,x,\ \ldots,\ \omega_n) \\ &= (x-y)(V_t(\omega_1,\ \ldots,\ \omega_{k-1},1,0,\ \ldots,\ \omega_n) - V_t(\omega_1,\ \ldots,\ \omega_{k-1},0,1,\ \ldots,\ \omega_n)) \\ &= (x-y)\{(F(\omega_1,\ \ldots,\ \omega_{k-1},1) - (\omega_1,\ \ldots,\ \omega_{k-1},0)) + \beta \sum_{e\in\mathcal{P}(a^{k-1}(t))}\prod_{i\in e}\omega_i \prod_{j\in a^{k-1}(t)\setminus e}(1-\omega_j) \\ &\quad \times [V_{t+1}(p_{11}[|e|],p_{11},p_{01},\pi(\omega_{k+2}),\ \ldots,\ \pi(\omega_n),p_{01}[k-1-|e|]) \\ &\quad - V_{t+1}(p_{11}[|e|],p_{11},\pi(\omega_{k+2}),\ \ldots,\ \pi(\omega_n),p_{01},p_{01}[k-1-|e|])]\} \\ &\geq (x-y)\left\{(F(\omega_1,\ \ldots,\ \omega_{k-1},1) - F(\omega_1,\ \ldots,\ \omega_{k-1},0)) - \beta\left(1-\prod_{j=k+2}^{n}(1-\omega_j)\right)F'_{\max}\right\} \\ &\geq (x-y)\left\{F'_{\min} - \beta\left(1-\prod_{j=k+2}^{n}(1-\omega_j)\right)F'_{\max}\right\} \\ &= (x-y)\left(1-\prod_{j=k+2}^{n}(1-\omega_j)\right)F'_{\max}\left[\frac{F'_{\min}}{F'_{\max}\left(1-\prod_{j=k+2}^{n}(1-\omega_j)\right)} - \beta\right] \\ &\geq 0\end{aligned}$$