Optimality of Myopic Policy for Multistate Channel Access

Kehao Wang, Lin Chen, Jihong Yu, and Duzhong Zhang

***Abstract*—We consider the multichannel access problem in which each of N channels is modeled as a multistate Markov chain. At each time instant, a transmitter accesses M channels and obtains some reward depending on the states of those chosen channels. The considered problem can be cast into a restless multiarmed bandit (RMAB) problem. It is well-known that solving the RMAB problem is PSPACE-hard. A natural alternative is to consider the myopic policy that maximizes the immediate reward but ignores the impact of the current strategy on the future reward. In this letter, we perform an analytical study on structure, optimality, and performance of the myopic policy for the considered RMAB problem. We show that the myopic policy has a simple robust structure that reduces channel selection to a round-robin procedure. The optimality of this simple policy is established for accessing $M = N - 1$ of N channels and conjectured for the general case of arbitrary M based on the structure of myopic policy.**

***Index Terms*—RMAB, myopic policy, PSPACE-Hard, optimality.**

I. Introduction

We consider a multi-channel communication system where a multi-antenna transmitter (Tx) chooses a set of channels to transmit data to multiple single-antenna receivers (Rxs). The fundamental object of our study is how the Tx does make its decision in selecting channels based on the feedback signals received over the channels in which it transmitted so as to maximize its utility (e.g., expected throughput).

In particular, we consider a set of N identical channels, each of which is characterized as an independent and identically distributed (i.i.d.) multi-state (i.e., X states, $X > 2$) discrete-time Markov chain, where state x_h, corresponds to the channel with higher signal to interference and noise ratio (SINR) than that of state x_l ($1 \leq x_l < x_h \leq X$). The objective of the SU is to seek a set of channels to access depending on the feedback signals so as to maximize the utility in the time horizon of interest. Obviously, the considered channel decision problem can be cast into the restless multi-armed bandit (RMAB) problem or partially observable Makov decision process (POMDP) [1], which is of fundamental importance in stochastic decision theory while proved to be PSPACE-Hard [2].

Thus far, very few results have been reported on the structure of the optimal policy of RMAB although the significant research efforts exist in the field. Hence, a natural alternative is to seek a simple myopic policy maximizing the immediate reward. For the case of *two-state*, Zhao *et al.* [3] established the structure of the myopic policy, and partly obtained the optimality for the case of i.i.d. channels. After that, Ahmad and Liu *et al.* [4] derived the optimality of the myopic sensing policy for the positively correlated i.i.d. channels for accessing one channel each time, and further extended the optimality to access multiple i.i.d. channels [5]. From another point, in [6], the authors extended i.i.d. channels [4] to non i.i.d. ones, and focused on a class of so-called *regular* functions, and derived closed-form sufficient conditions to guarantee the optimality of myopic sensing policy. For the complicated case of *multi-state*, in [7], the authors established the sufficient conditions for the optimality of myopic sensing policy in multi-state homogeneous channels with a strict constraint, i.e., the forth non-trivial assumption in [7].

In this letter, we establish the structure of the myopic policy for the proposed multi-state RMAB problem and obtain part of optimality without that non-trivial constraint. In particular, the contributions are two-folds:

- The structure of the myopic policy is shown to be a simple queue determined by the availability probability vector of channels provided that certain condition is satisfied for the transition matrix of multi-state channels.
- We establish a set of conditions under which the myopic policy is proved to be optimal for the case of accessing $N - 1$ of N channels and conjectured to be optimal in the general case. Furthermore, the optimality is verified by numerical simulation.

Notation: $e_i = [\underbrace{0, \cdots, 0}_{i-1}, 1, \underbrace{0, \cdots, 0}_{X-i}]$, $\mathbf{E} = [e_1, \cdots, e_X]'$.

II. Problem Formulation

We consider an N orthogonal channels communication system where each channel is characterized by a Markov chain of X states $\mathcal{X} = \{1, \ldots, X\}$, and the channel state transition probabilities $p_{i,j}, ij = 1, \ldots, X$.

$$\mathbf{P} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,X} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,X} \\ \vdots & \vdots & \ddots & \vdots \\ p_{X,1} & p_{X,2} & \cdots & p_{X,X} \end{pmatrix} = \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_X \end{pmatrix}.$$

Let $\mathbf{S}(t) \triangleq [S_1(t), \ldots, S_N(t)]$ denote the channel state vector where $S_i(t) \in \{1, \ldots, X\}$ is the state of channel i in slot t. A

Manuscript received July 24, 2015; accepted November 17, 2015. Date of publication November 25, 2015; date of current version February 12, 2016. This work was supported in part by National NSF of China under Grant 61303027, in part by the NSF of Hubei Province under Grant 2015CFB585, in part by the China Postdoctoral Science Foundation under Grant 2013M531753 and Grant 2014T70748, and the Fundamental Research Funds for the Central Universities (WUT:2015III012). The associate editor coordinating the review of this paper and approving it for publication was D. W. K. Ng.

K. Wang and D. Zhang are with the Key Laboratory of Fiber Optic Sensing Technology and Information Processing, Wuhan University of Technology, 430070 Hubei, China (e-mail: kehao.wang@whut.edu.cn; zhangduzhong@gmail.com).

L. Chen and J. Yu are with the Laboratoire de Recherche en Informatique (LRI), Department of Computer Science, University of Paris-Sud XI, 91405 Orsay, France (e-mail: lin.chen@lri.fr; jihong.yu@lri.fr).

Digital Object Identifier 10.1109/LCOMM.2015.2503770

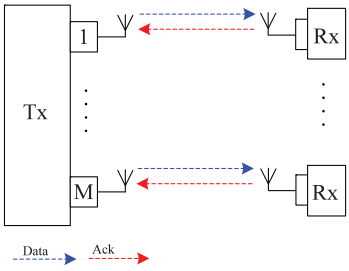

Fig. 1. System model with 1 Tx, M Rx, and N orthogonal channels.

Tx choose M channels to transmit data to M single-antenna Rxs, as shown in Fig. 1.

We assume that the multi-channel system operates in a synchronous slot fashion with the time slot indexed by t ($t = 1, 2, \ldots, T$), where T is the time horizon of interest. In a slot, for each of M antennas,i.e., m-th antenna, the Tx first choose a channel n for this antenna to transmit data packets to the m-th Rx, and then the Rx would respond Tx with feedback information over the same channel n (see Fig. 1).

We assume that the Tx can perfectly receive the feedback information sent by the Rx of the channel where it transmitted. Hence, at time slot t the Tx can collect the feedback observation $O_i(t)$ in the i channel, and equipped with M antennas, the Tx can receive the set $\mathcal{A}(t)$ ($\mathcal{A}(t) \subseteq \{1, 2, \ldots, N\}$, $|\mathcal{A}(t)| = M$) of N channels and furthermore, to obtain the observations $\mathcal{O}(t) = \{O_i(t) : i \in \mathcal{A}(t)\}$. For simplicity, we map the feedback $O_i(t)$ of channel i in state x to an integer x, which indicates the channel quality of channel i is in x level.

Considering the limited feedback information received by Tx (i.e., Tx only receiving feedback from M out of N channels), the channel state vector is only partially observable to the Tx for its decision. We define the information state vector $\Omega(t) \triangleq \{\mathbf{w}_i(t), i \in \mathcal{N}\}$ (referred to as belief vector), where $\mathbf{w}_i(t) = [w_{i1}(t) \cdots w_{iX}(t)]$ and $w_{ix}(t)$ is the estimated conditional probability that the channel $i \in \mathcal{N}$ is in x state (i.e., $S_i(t) = x$) given all the past observations and decisions by the Tx.

Given the information state $\Omega(t)$, the decision $\mathcal{A}(t)$ and the observations $\mathcal{O}(t)$, the belief vector for the Tx can be updated recursively using the following rule as shown in (1).

$$\mathbf{w}_i(t+1) = \begin{cases} P_x, & i \in \mathcal{A}(t),\ O_i(t) = x,\ x \in \mathcal{X} \\ \mathbf{w}_i(t)\mathbf{P}, & i \notin \mathcal{A}(t). \end{cases} \tag{1}$$

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

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

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

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{t=1}^{T} \beta^{t-1} R_\pi(\Omega(t)) \,\middle|\, \Omega(1)\right] \tag{3}$$

where $R_\pi(\Omega(t))$ is the reward collected in slot t under the accessing policy π with the initial belief vector $\Omega(1)$[1], and β ($0 \le \beta \le 1$) is the discount factor characterizing the feature that the future rewards are less valuable than the immediate reward. By treating the belief value of each channel as the state of each arm of a bandit, the optimization problem can be cast into a restless multi-armed bandit problem.

We assume that the reward obtained from accessing channel i at slot t depends on the state of the channel chosen at t, formally defined as follows:

$$R(S_i(t)) = r_x \text{ if } S_i(t) = x, \tag{4}$$

where, $r_X > \cdots > r_1$ indicates that the reward obtained in the high SINR channel state is larger than that in the low SINR, and $\mathbf{r} = [r_1 \cdots r_X]$.

In order to get more insight on the structure of the optimization problem formulated in (3) and the complexity to solve it, we derive the dynamic programming formulation of (3) as follows:

$$V_T(\Omega(T)) = \max_{\mathcal{A}(T)} \mathbb{E}\left[\sum_{i \in \mathcal{A}(t)} [\mathbf{w}_i \mathbf{r}']\right], \tag{5}$$

$$V_t(\Omega(t)) = \max_{\mathcal{A}(t)} \mathbb{E}\left[\sum_{i \in \mathcal{A}(t)} [\mathbf{w}_i \mathbf{r}'] + \beta \Gamma(\Omega(t))\right], \tag{6}$$

where

$$\Gamma(\Omega(t)) \triangleq \sum_{A_x \subseteq \mathcal{A}(t)}^{\mathcal{A}(t) = \cup_{x=1}^{X} A_x} \prod_{i \in A_1} \mathbf{w}_{i1} \cdots \prod_{j \in A_X} \mathbf{w}_{jX} \cdot V_{t+1}(\Omega(t+1)).$$

In the above equations, $V_t(\Omega(t))$ is the value function corresponding to the maximal expected reward from time slot t to T ($1 \le t \le T$) with $\Omega(t+1)$ following the evolution described in (1) given that the channels in the subset A_x ($x \in \mathcal{X}$) are observed in state x. In particular, the term $\Gamma(\Omega(t))$ corresponds to the expected accumulated discounted reward starting from slot $t+1$ to T, calculated over all possible realizations of the selected channels (i.e., channels in $\mathcal{A}(t)$).

Theoretically, the optimal policy can be obtained by solving the above dynamic programming. It is infeasible, however, due to the impact of the current action on the future reward, and in fact obtaining the optimal solution directly from the above recursive equations is computationally prohibitive. Hence, a natural alternative is to seek a simple myopic policy, formally defined as follows:

$$\hat{\mathcal{A}}(t) = \underset{\mathcal{A}(t)}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{i \in \mathcal{A}(t)} \mathbf{w}_i \mathbf{r}'\right]. \tag{7}$$

In the following sections we focus on structure and optimality of the myopic accessing policy in the multi-channel communication context.

[1]The initial belief $\mathbf{w}_i(1)$ can be set to $\mathbf{w}_0$ such that $\mathbf{w}_0(\mathbf{E} - \mathbf{P}) = 0$ if no information about the initial system state is available.

III. Structure of Myopic Policy

In this section, we show that the myopic policy has a simple and robust structure. Based on this structure, we prove that the myopic policy is optimal for $M = N - 1$ and give the conjecture for general M in the following section.

First, we let

$$\Pi(X) \triangleq \{(w_1, \ldots, w_X) : \sum_{i=1}^{X} w_i = 1, w_1, \ldots, w_X \geq 0\},$$

and then borrow some definitions about "ordering" for the following analysis.

Definition 1 (MLR ordering [8]): Let $\mathbf{w}_1, \mathbf{w}_2 \in \Pi(X)$ be any two belief vectors. Then $\mathbf{w}_1$ is greater than $\mathbf{w}_2$ with respect to the MLR ordering—denoted as $\mathbf{w}_1 \geq_r \mathbf{w}_2$, if

$$\mathbf{w}_{1i}\mathbf{w}_{2j} \leq \mathbf{w}_{2i}\mathbf{w}_{1j}, \quad i < j, i, j \in \{1, 2, \ldots, X\}.$$

Definition 2 (TP2 [8]): Matrix $\mathbf{P}$ is TP2 if all the second minors are nonnegative. Given $\mathbf{w}_1, \mathbf{w}_2 \in \Pi(X)$, then $\mathbf{w}_1\mathbf{P} \geq_r \mathbf{w}_2\mathbf{P}$ if $\mathbf{w}_1 \geq_r \mathbf{w}_2$ and $\mathbf{P}$ is TP2.

Assumption 1: There exists some L $(2 \leq L \leq X)$ such that $P_1\mathbf{P} \geq_r P_{L-1}$ and $P_X\mathbf{P} \leq_r P_L$.

Remark. This assumption guarantees that the probability vector is completely ordered in the probability space in the sense of MLR, which serves as the basis of tractable analysis of optimal policy.

Theorem 1 (Structure of Myopic Policy): Under Assumption 1, if $\mathbf{P}$ is TP2 and $P_X \geq_r \mathbf{w}_{\sigma_1}(1) \geq_r \cdots \geq_r \mathbf{w}_{\sigma_N}(1) \geq_r P_1$, the following stochastic order is kept at each slot.

1) The initial channel ordering $\mathbf{Q}(1)$ is determined by the initial belief vector:

$$\mathbf{w}_{\sigma_1}(1) \geq_r \cdots \geq_r \mathbf{w}_{\sigma_N}(1) \Rightarrow \mathbf{Q}(1) = (\sigma_1, \sigma_2, \ldots, \sigma_N)$$

2) The channels over which feedback x $(x \in \{L, \ldots, X\})$ are observed will stay at the head of the queue, and the channels over which feedback x $(x \in \{1, \ldots, L-1\})$are observed will be moved to the end of the queue while keeping their order unchanged;

Proof: Assume $\mathbf{Q}(t) = (\sigma_1, \ldots, \sigma_N,)$ at slot t. We thus have $P_X \geq_r \mathbf{w}_{\sigma_1}(t) \geq_r \cdots \geq_r \mathbf{w}_{\sigma_N}(t) \geq_r P_1$.

If channel σ_1 is observed to be state x $(L \leq x \leq X)$, then $\mathbf{w}_{\sigma_1}(t+1) = P_x \geq_r P_L \geq_r \mathbf{w}_{\sigma_2}(t)\mathbf{P} \geq_r \cdots \geq_r \mathbf{w}_{\sigma_N}(t)\mathbf{P}$ according to the assumption, and thus $\mathbf{Q}(t+1) = (\sigma_1, \ldots, \sigma_N)$ according to MLR of $\mathbf{w}$.

If channel σ_1 is observed in state x $(1 \leq x \leq L-1)$, then $\mathbf{w}_{\sigma_1}(t+1) = P_x \leq_r P_{L-1} \leq_r \mathbf{w}_{\sigma_N}(t)\mathbf{P} \leq_r \cdots \leq_r \mathbf{w}_{\sigma_2}(t)\mathbf{P}$, and further $\mathbf{Q}(t+1) = (\sigma_2, \ldots, \sigma_N, \sigma_1)$. ■

IV. Optimality of Myopic Policy

Let $V_t(\Omega; \mathcal{A}(t))$ the expected total discounted reward obtained by action $\mathcal{A}(t)$ in slot t followed by the myopic policy in future slots. We first establish some important auxiliary lemmas and then show the optimality of myopic policy.

Lemma 1: $V_t(\Omega; \mathcal{A}(t))$ is symmetrical about $\mathbf{w}_i, \mathbf{w}_j \in \mathcal{A}(t)$.

Proof: The proof is straightforward by noticing that both the immediate reward and the channel belief vector $\Omega(t+1)$ are unrelated with the order of $\mathbf{w}_i, \mathbf{w}_j$ since the myopic policy is adopted from slot $t+1$. ■

Lemma 2: It holds that $V_t(\Omega; \mathcal{A}(t))$ is an affine function, i.e.,

$$V_t(\cdots, \mathbf{w}_i, \cdots; \mathcal{A}(t)) = \sum_{j=1}^{X} \mathbf{w}_{ij} V_t(\cdots, e_j, \cdots; \mathcal{A}(t)), \forall i \in \mathcal{N}.$$

Proof: We prove the lemma by induction. It can be easily checked that the lemma holds for slot T. Assume that it holds for slot $T, \ldots, t+1$, we now prove it holds for slot t. We proceed by distinguishing the following two cases:

Case 1: $k \notin \mathcal{A}(t)$. In this case we have

$$V_t(\Omega; \mathcal{A}(t)) = \sum_{j \in \mathcal{A}(t)} [\mathbf{w}_j \mathbf{r}'] + \sum_{A_x \subseteq \mathcal{A}(t)}^{\mathcal{A}(t) = \cup_{x=1}^{X} A_x} \prod_{i \in A_1} \mathbf{w}_{i1} \cdots \prod_{j \in A_X} \mathbf{w}_{jX} \cdot V_{t+1}(\cdots, \mathbf{w}_k\mathbf{P}, \cdots).$$

By induction hypothesis, $V_{t+1}(\ldots, \mathbf{w}_k\mathbf{P}, \ldots)$ is an affine function of $\mathbf{w}_k\mathbf{P}$, and meanwhile, $\mathbf{w}_k\mathbf{P}$ is an affine transform of $\mathbf{w}_k$, thus $V_{t+1}(\cdots, \mathbf{w}_k\mathbf{P}, \cdots)$ is an affine function of $\mathbf{w}_k$. It follows that $V_t(\Omega(t); \mathcal{A}(t))$ is also an affine function of $\mathbf{w}_k$.

Case 2: $k \in \mathcal{A}(t)$. Let $m \notin \mathcal{A}(t)$ and $\mathcal{A}'(t) = \mathcal{A}(t) \setminus \{k\}$, we have

$$\begin{aligned} V_t(\Omega; \mathcal{A}(t)) &= \sum_{l \in \mathcal{A}(t)} [\mathbf{w}_l \mathbf{r}'] \\ &+ \sum_{A_x \subseteq \mathcal{A}(t)}^{\mathcal{A}(t) = \cup_{x=1}^{X} A_x} \prod_{i \in A_1} \mathbf{w}_{i1} \cdots \prod_{j \in A_X} \mathbf{w}_{jX} \cdot V_{t+1}(\cdots, \mathbf{w}_m\mathbf{P}, \cdots) \\ &= \sum_{l \in \mathcal{A}(t)} [\mathbf{w}_l \mathbf{r}'] + \sum_{A_x \subseteq \mathcal{A}(t) \setminus k}^{\mathcal{A}(t) \setminus k = \cup_{x=1}^{X} A_x} \prod_{i \in A_1} \mathbf{w}_{i1} \cdots \prod_{j \in A_X} \mathbf{w}_{jX} \\ &\times \left[\sum_{x=1}^{X} \mathbf{w}_k(x) V_{t+1}(\cdots, e_x, \ldots, \mathbf{w}_m\mathbf{P}, \cdots) \right] \end{aligned}$$

where, the third equality follows the induction hypothesis.

Obviously, $\sum_{l \in \mathcal{A}(t)} [\mathbf{w}_l \mathbf{r}']$ is an affine function of $\mathbf{w}_k$, the second term is also an affine function of $\mathbf{w}_k$. Therefore, $V_t(\mathbf{w}(t); \mathcal{A}(t))$ is an affine function of $\mathbf{w}_k$. ■

Lemma 2 can be applied one step further to prove the following corollary.

Corollary 1: For any $l, m \in \mathcal{N}$ it holds that

$$\begin{aligned} &V_t(\ldots, \mathbf{w}_l, \ldots, \mathbf{w}_m, \ldots; \mathcal{A}(t)) - V_t(\ldots, \mathbf{w}_m, \ldots, \mathbf{w}_l, \ldots; \mathcal{A}(t)) \\ &= \sum_{i=1}^{X} \sum_{j=i+1}^{X} (\mathbf{w}_{lj}\mathbf{w}_{mi} - \mathbf{w}_{li}\mathbf{w}_{mj}) \\ &\times \left[V_t(\ldots, e_j, \ldots, e_i, \ldots; \mathcal{A}(t)) - V_t(\ldots, e_i, \ldots, e_j, \ldots; \mathcal{A}(t)) \right]. \end{aligned}$$

Lemma 3: Let $\mathcal{A}(t) = \mathcal{N} \setminus \{j\}$ and $\mathcal{A}'(t) = \mathcal{N} \setminus \{i\}$ where $\mathbf{w}_i(t) \geq \mathbf{w}_j(t)$, it holds that $V_t(\Omega; \mathcal{A}(t)) \geq V_t(\Omega; \mathcal{A}'(t))$ if $\mathbf{P}$ is TP2 and $P_1 \leq_r \mathbf{w}_i(1) \leq_r P_X$.

Proof: The lemma can be easily checked that it holds for slot T. Assume that it holds for slot $T, \ldots, t+1$, we now prove it holds for slot t.

$$\begin{aligned} &V_t(\Omega; \mathcal{A}(t)) - V_t(\Omega; \mathcal{A}'(t)) \\ &\quad = V_t(\cdots, \mathbf{w}_i, \mathbf{w}_j; \mathcal{A}(t)) - V_t(\ldots, \mathbf{w}_j, \mathbf{w}_i; \mathcal{A}'(t)) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{X} \sum_{l=k+1}^{X} \left(\mathbf{w}_{il}\mathbf{w}_{jk} - \mathbf{w}_{ik}\mathbf{w}_{jl}\right) \\
&\quad \times \left[V_t\left(\cdots, e_l, e_k; \mathcal{A}(t)\right) - V_t\left(\cdots, e_k, e_l; \mathcal{A}'(t)\right)\right] \\
&= \sum_{k=1}^{X} \sum_{l=k+1}^{X} \left(\mathbf{w}_{il}\mathbf{w}_{jk} - \mathbf{w}_{ik}\mathbf{w}_{jl}\right) \Bigg[[r_l - r_k] \\
&\quad + \sum_{A_x(t) \subseteq \mathcal{A}(t)\setminus i}^{\mathcal{A}(t)\setminus i = \cup_{x=1}^{X} A_x(t)} \prod_{n \in A_1(t)} \mathbf{w}_{n1} \cdots \prod_{j \in A_X(t)} \mathbf{w}_{jX} \\
&\quad \times \left[V_{t+1}\left(\cdots, P_l, P_k\right) - V_{t+1}\left(\cdots, P_k, P_l\right)\right] \Bigg] \\
&\geq \sum_{k=1}^{X} \sum_{l=k+1}^{X} (\mathbf{w}_{il}\mathbf{w}_{jk} - \mathbf{w}_{ik}\mathbf{w}_{jl})(r_l - r_k) \\
&= \left(\mathbf{w}_i - \mathbf{w}_j\right)\mathbf{r}' \geq 0
\end{aligned}
$$

where, the first inequality follows induction hypothesis and $\mathbf{P}$ is TP2. ■

Theorem 2. The myopic policy is optimal if $M = N - 1$, $\mathbf{P}$ is TP2, and $\mathbf{w}_1(1) \leq_r \cdots \leq_r \mathbf{w}_N(1)$.

Proof: We prove the theorem by induction. In slot T, the optimality of the myopic policy is obvious. Assume that the myopic policy is also optimal for slot $T-1, \ldots, t+1$. We prove it holds for slot t.

To that end, we sort $\Omega(t)$ in the decreasing order such that $\mathbf{w}_1 \geq_r \cdots \geq_r \mathbf{w}_N$. To prove the optimality of myopic policy, we need to show that $V_t(\Omega; \mathcal{A}(t)) \geq V_t(\Omega; \mathcal{A}'(t))$ where $\mathcal{A}(t) = \{1, \ldots, N-1\} = \mathcal{N} \setminus \{N\}$ and $\mathcal{A}'(t)$ is any $N-1$ elements of $\mathcal{N}$. Without loss of generality, we assume $\mathcal{A}'(t) = \mathcal{N} \setminus \{l\}$. Noticing that $\mathbf{w}_l \geq_r \mathbf{w}_N$, it follows from Lemma 3 that $V_t(\Omega; \mathcal{A}(t)) \geq V_t(\Omega; \mathcal{A}'(t))$. ■

For the case of two-state, the myopic policy is conjectured to be optimal for arbitrary M and N [3] [10], and then the conjecture is proved in [9], [11]. For the case of multi-state, considering the same structure of myopic policy with [3], [9], we have the following similar conjecture:

Conjecture 1. The myopic policy is optimal for arbitrary M $(1 \leq M \leq N-1)$ if $\mathbf{P}$ is TP2 and $\mathbf{w}_1(1) \leq_r \cdots \leq_r \mathbf{w}_N(1)$.

V. Numerical Simulation

In this section, we study the average reward performance of *Myopic policy*, *Random policy*, and *Optimal policy* by two simplest scenarios ($N = 3$, $M = 1$, $\beta = 1$ and $N = 3$, $M = 2$, $\beta = 1$). In the two scenarios, if

$$
\mathbf{P} = \begin{pmatrix} 0.40 & 0.20 & 0.40 \\ 0.20 & 0.24 & 0.56 \\ 0.15 & 0.25 & 0.60 \end{pmatrix}, \mathbf{r}' = \begin{pmatrix} 0.0 \\ 0.8 \\ 1.0 \end{pmatrix},
$$

then Assumptions 1-3 of [7] hold except Assumption 4.

Fig. 2 shows that the average reward obtained by the myopic policy perfectly matches that of the optimal policy, which confirms our analytical results. We can also observe that the myopic policy outperforms the random policy to various extents. Given the exponential complexity of obtaining the optimal policy and the large number of trials in the random policy, the benefit of the myopic policy is well demonstrated.

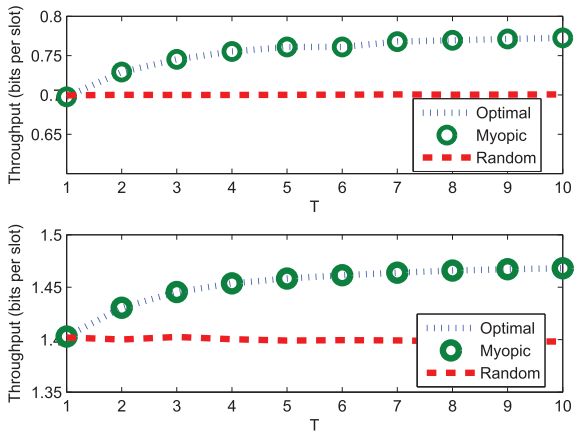

Fig. 2. Comparison ($N = 3$): upper plot: $M = 1$; lower plot: $M = 2$.

VI. Conclusion

In this letter, we have investigated the multi-channel access problem, which is formulated as a POMDP or RMAB. For the stochastically identical and independent channels, we have proved the optimality of the myopic policy for the case of $M = N - 1$, conjectured the optimality for the case of arbitrary $M < N$ based on its conservation queue structure of belief values, and then verified the optimality by numerical simulation. One of our future directions is to prove the conjecture about the optimality for generic M.

References

[1] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A pomdp framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Apr. 2010.

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

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

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

[5] S. Ahmad and M. Liu, "Multi-channel opportunistic access: A case of restless bandits with multiple plays," in *Proc. Allerton Conf.*, Monticello, IL, USA, Sep./Oct. 2009, pp. 1361–1368.

[6] K. Wang and L. Chen, "On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 300–309, Jan. 2012.

[7] Y. Ouyang and D. Teneketzis, "On the optimality of myopic sensing in multi-state channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 681–696, Jan. 2014.

[8] A. Muller and D. Stoyan, *Comparison Methods for Stochastic Models and Risk*. Hoboken, NJ, USA: Wiley, 2002.

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

[10] K. Liu, Q. Zhao, and B. Krishnamachari,"Dynamic multichannel access with imperfect channel state detection," *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2795–2807, May 2010.

[11] K. Wang, L. Chen, Q. Liu, and K. Al Agha, "On optimality of myopic sensing policy with imperfect sensing in multi-channel opportunistic access," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3854–3862, Sep. 2013.