

On the Myopic Policy for a Class of Restless Bandit Problems with Applications in Dynamic Multichannel Access

Keqin Liu and Qing Zhao

Abstract— We consider a class of restless multi-armed bandit problems that arises in multi-channel opportunistic communications, where channels are modeled as independent and stochastically identical Gilbert-Elliot channels and channel state observations are subject to errors. We show that the myopic channel selection policy has a semi-universal structure that obviates the need to know the Markovian transition probabilities of the channel states. Based on this structure, we establish closed-form lower and upper bounds on the steady-state throughput achieved by the myopic policy. Furthermore, we characterize the approximation factor of the myopic policy to bound its worst-case performance loss with respect to the optimal performance.

Index Terms— Dynamic multi-channel access, restless multi-armed bandit, myopic policy

I. INTRODUCTION

A. Dynamic Multichannel Access

We consider the following stochastic optimization problem that arises in multichannel opportunistic communications. Assume that there are N independent and stochastically identical Gilbert-Elliot channels [1]. As illustrated in Fig. 1, the state of a channel — “good” or “bad” — indicates the desirability of accessing this channel and determines the resulting reward. The transitions between these two states follow a discrete-time Markov chain with transition probabilities $\{p_{ij}\}_{i,j \in \{0,1\}}$. This channel model has been commonly used to abstract physical channels with memory. Consider, for example, the emerging application of cognitive radios for opportunistic spectrum access where secondary users search in the spectrum for idle channels temporarily unused by primary users [2]. For this application, the good state represents an idle channel while the bad state an occupied channel.

In each time slot, a user chooses M out of the N channels to sense and subsequently access channels sensed to be in the good states. Sensing is subject to errors: a good channel may be sensed as bad and *vice versa*. Accessing a good channel results in a unit reward, and no access or accessing a bad channel leads to zero reward. The design objective is the optimal sensing policy for dynamic channel selection in order to maximize the expected long-term reward.

This work was supported by the Army Research Office under Grant W911NF-08-1-0467 and by the National Science Foundation under Grants ECS-0622200 and CCF-0830685.

Keqin Liu and Qing Zhao are with the Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616, USA {kqliu, qzhao}@ucdavis.edu.

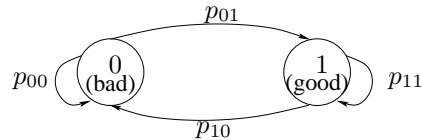


Fig. 1. The Gilbert-Elliot channel model.

B. Restless Multi-armed Bandit and Myopic Policy

This problem can be formulated as a partially observable Markov decision process (POMDP) for generally correlated channels [3], or a restless multi-armed bandit process (RMBP) for independent channels considered here. The maximum expected total reward of the multi-channel opportunistic system is essentially the value function of an RMBP. Unfortunately, obtaining optimal solutions to a general restless bandit process is PSPACE-hard [4], and analytical characterization of the performance of the optimal policy is often intractable.

We thus focus on the low-complexity myopic policy which has been shown to be optimal for this class of restless bandit problems under certain conditions (see Sec. I-C). Specifically, we establish a simple structure of the myopic policy when false alarm probability of the channel state detector is below a certain value. This structure is semi-universal: it is independent of the Markovian transition probabilities except the order of p_{11} and p_{01} . Based on this structure, we develop closed-form lower and upper bounds on the steady-state throughput under the myopic policy that monotonically tighten as the number N of channels increases. When each channel is positively correlated ($p_{11} \geq p_{01}$), we further obtain the limiting performance of the myopic policy as N approaches to infinity. Furthermore, by considering a genie-aided system, we develop an upper bound on the optimal performance, which provides a performance benchmark for the myopic policy. This result, coupled with the lower bound on the performance of the myopic policy, leads to an analytical characterization of the approximation factor of the myopic policy.

C. Related Work

The design of multi-channel opportunistic access was formulated as a constrained POMDP in [3] for general correlated channels. A separation principle has been established in [5] which decouples the design of channel state detector and access policies from that of channel sensing policy. The channel sensing policy is then reduced to an unconstrained POMDP problem. Under single-channel sensing ($M = 1$), the structure of the myopic sensing policy has been established in [6] when the false alarm probability is below

a certain value. Based on this structure, the optimality of the myopic policy has been proved for $N = 2$ and conjectured for $N > 2$ based on numerical examples [6]. In this paper, we extend the structure of the myopic policy to multi-channel sensing scenarios and characterize its performance and approximation factor.

This paper also extends our earlier work in [7] that assumes perfect detection of the channel states, where the structure of the myopic policy has been established for all N and its optimality proved for $N = 2$ and conjectured for $N > 2$. A recent follow-up work [8] has extended the optimality of the myopic policy to all N under the condition of $p_{11} \geq p_{01}$. For the same model in [7] except assuming non-identical channels, Whittle's index policy under both discounted and average reward criteria has been established in [9]. The structure and the optimality of Whittle's index policy are also established in [9] for stochastically identical channels based on its equivalence to the myopic policy. The same problem is also considered in a parallel work in the context of multi-agent systems [10], where Whittle's index is established under the discounted reward criterion using a different approach. The structure and the optimality of Whittle's index policy, however, were not considered in [10].

Other related work on multi-channel opportunistic access can be found in [11]–[13]. In [11], the authors consider the POMDP framework established in [3] and an approximation method is proposed. In [12], a heuristic sensing policy is proposed to reduce channel switching under general stochastic models for channel occupancy. In [13], the authors extend the POMDP framework established in [3] by considering the use of analog channel measurements (instead of acknowledgement) in the belief update. In this case, a dedicated control channel is needed to ensure that the secondary transmitter and its receiver select the same channel for communication. The issue of unknown parameters in the distribution of the primary signal is also addressed in [13].

II. PROBLEM FORMULATION

In this section, we formulate the problem by considering the cognitive radio application.

A. System Model

Let $\mathcal{S}(t) \triangleq [S_1(t), \dots, S_N(t)]$ denote the channel states, where $S_n(t) \in \{0 \text{ (bad/busy)}, 1 \text{ (good/idle)}\}$ is the state of channel n in slot t . At the beginning of each slot, the user first decides which M channels to sense for potential access. Once a channel (say channel n) is chosen, the user detects the channel state, which can be considered as a binary hypothesis test¹:

$$\mathcal{H}_0 : S_n(t) = 1 \text{ (good/idle) vs. } \mathcal{H}_1 : S_n(t) = 0 \text{ (bad/busy).}$$

The performance of channel state detection is characterized by the receiver operating characteristic (ROC) which relates

¹We consider here the nontrivial cases with p_{01} and p_{11} in the open interval of (0,1). When they take the special value of 0 or 1, channel state detection can be simplified. Extensions to such special cases are straightforward.

the probability of false alarm ϵ and the probability of miss detection δ :

$$\epsilon \triangleq \Pr\{\text{decide } \mathcal{H}_1 | \mathcal{H}_0 \text{ is true}\}, \quad \delta \triangleq \Pr\{\text{decide } \mathcal{H}_0 | \mathcal{H}_1 \text{ is true}\}.$$

Based on the imperfect detection outcome in slot t , the user chooses an access action $\Phi_n(t) \in \{0 \text{ no access}, 1 \text{ access}\}$ that determines whether to access channel n for transmission. We note that the design should be subject to a constraint on the probability of accessing a busy channel, which causes interference to primary users. Specifically, the probability $\mathcal{P}_n(t)$ of collision perceived by the primary network in any channel and in any slot is capped below a predetermined threshold ζ , *i.e.*,

$$\mathcal{P}_n(t) \triangleq \Pr(\Phi_n(t) = 1 | S_n(t) = 0) \leq \zeta, \quad \forall n, t.$$

This constrained stochastic optimization problem requires the joint design of the channel state detector (*i.e.*, how to choose the detection thresholds to trade off false alarms with miss detections), the access policy that decides the transmission probabilities based on imperfect detection outcomes, and the sensing policy for channel selection. This problem is formulated as a constrained POMDP in [3] for generally correlated channels. A separation principle has been established in [5] showing that the optimal detector is the Neyman-Pearson detector with the probability δ of miss detection given by the maximum allowable probability ζ of collision, and the optimal access policy is to simply trust the detection outcomes: transmit over a channel if and only if it is detected as idle. Thus, the user can obtain a unit reward on a chosen channel if and only if it is idle and detected correctly (*i.e.*, no false alarm). The optimal sensing policy can then be designed using the optimal detector and the optimal access policy without the constraint on accessing a busy channel, which becomes an unconstrained POMDP addressed here. The objective is to maximize the expected total reward over a horizon of T slots by choosing judiciously a sensing policy that governs channel selection in each slot.

Since failed transmissions may occur, acknowledgements (ACKs) are necessary to ensure guaranteed delivery. Specifically, when the receiver successfully receives a packet from a channel, it sends an acknowledgement to the transmitter over the same channel at the end of the slot. Otherwise, the receiver does nothing, *i.e.*, a NAK is defined as the absence of an ACK, which occurs when the transmitter did not transmit over this channel or transmitted but the channel is busy. We assume that acknowledgements are received without error since acknowledgements are always transmitted over idle channels.

B. Restless Multi-Armed Bandit Formulation

Due to limited and imperfect sensing, the system state $[S_1(t), \dots, S_N(t)] \in \{0, 1\}^N$ in slot t is not fully observable to the user. It can, however, infer the state from its decision and observation history. It has been shown that a sufficient statistic of the system for optimal decision making is given by the conditional probability that each channel is in state 1 given all past decisions and observations [14]. Referred

to as the belief vector, this sufficient statistic is denoted by $\Omega(t) \triangleq [\omega_1(t), \dots, \omega_N(t)]$, where $\omega_i(t)$ is the conditional probability that $S_i(t) = 1$. In order to ensure that the user and its intended receiver tune to the same channels in each slot, channel selections should be based on common observations: the acknowledgements $\mathcal{K}(t) \in \{0 (NAK), 1 (ACK)\}^M$ in each slot rather than the detection outcomes at the transmitter. Let $I(t)$ denote the sensing action that consists of M channels to sense in slot t . Given the sensing action $I(t)$ and the observations $\{K_i(t) \in \{0, 1\} : i \in I(t)\}$ in slot t , the belief vector for slot $t + 1$ can be obtained via the Bayes rule.

$$\omega_i(t+1) = \begin{cases} p_{11}, & i \in I(t), K_i(t) = 1 \\ \Gamma\left(\frac{\epsilon\omega_i(t)}{\epsilon\omega_i(t)+1-\omega_i(t)}\right), & i \in I(t), K_i(t) = 0 \\ \Gamma(\omega_i(t)), & i \notin I(t) \end{cases} \quad (1)$$

where the operator $\Gamma(\cdot)$ is defined as

$$\Gamma(x) \triangleq xp_{11} + (1-x)p_{01}.$$

A sensing policy π specifies a sequence of functions $\pi = [\pi_1, \pi_2, \dots, \pi_T]$ where π_t maps a belief vector $\Omega(t)$ to a sensing action $I(t)$ for slot t . Multi-channel opportunistic access can thus be formulated as the following stochastic optimization problem.

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=1}^T R(\pi_t(\Omega(t))) | \Omega(1) \right],$$

where $R(\pi_t(\Omega(t)))$ is the reward obtained when the belief is $\Omega(t)$ and channels $\pi_t(\Omega(t))$ are selected, and $\Omega(1)$ is the initial belief vector. This problem falls into the model of an RMBP by treating the belief value of each channel as the state of each arm of a bandit. If no information on the initial system state is available, each entry of $\Omega(1)$ can be set to the stationary distribution ω_o of the underlying Markov chain:

$$\omega_o = \frac{p_{01}}{p_{01} + p_{10}}. \quad (2)$$

Let $V_t(\Omega)$ be the value function, which represents the maximum expected total reward that can be obtained starting from slot t given the current belief vector Ω . Given that the user takes action I and observes $\mathcal{K} = \{K_i\}_{i \in I}$, the expected reward that can be accumulated starting from slot t consists of two parts: the expected immediate reward $\sum_{i \in I} \omega_i(1 - \epsilon)$ and the maximum expected future reward $V_{t+1}(\mathcal{T}(\Omega|I, \mathcal{K}))$, where $\mathcal{T}(\Omega|I, \mathcal{K})$ denotes the updated belief vector for slot $t + 1$ after incorporating action I and observations \mathcal{K} as given in (1). Averaging over all possible observations \mathcal{K} and maximizing over all actions I , we arrive at the following optimality equation.

$$\begin{aligned} V_t(\Omega(t)) &= \max_I (\sum_{i \in I} \omega_i(t)(1 - \epsilon) + \mathbb{E}[V_{t+1}(\mathcal{T}(\Omega(t)|I, \mathcal{K}))]), \\ V_T(\Omega(T)) &= \max_I \sum_{i \in I} \omega_i(T)(1 - \epsilon). \end{aligned}$$

In theory, the optimal policy π^* and its performance $V_1(\Omega(1))$ can be obtained by solving the above dynamic programming. Unfortunately, due to the impact of the current action on the future reward and the uncountable space of the belief vector, obtaining the optimal solution using directly the above recursive equations is computationally prohibitive.

Even when approximate numerical solutions can be obtained, they do not provide insight into system design or analytical characterizations of the optimal performance $V_1(\Omega(1))$.

III. STRUCTURE AND PERFORMANCE OF THE MYOPIC POLICY

In this section, we show that the myopic sensing policy has a simple and robust structure. Based on this structure, we characterize the performance and approximation factor of the myopic policy. Due to the space limit, we omit all proofs which can be found in [15].

A myopic policy ignores the impact of the current action on the future reward, focusing solely on maximizing the expected immediate reward $\mathbb{E}[R(I(t))]$. Myopic policies are thus stationary. The myopic action \hat{I} under belief state $\Omega = [\omega_1, \dots, \omega_N]$ is simply given by

$$\hat{I}(\Omega) = \arg \max_I \sum_{i \in I} \omega_i. \quad (3)$$

In general, obtaining the myopic action in each slot requires the recursive update of the belief vector Ω as given in (1), which requires the knowledge of the transition probabilities $\{p_{ij}\}$. However, for the problem at hand, we show that the myopic policy has a simple and robust structure that does not need the precise knowledge of the transition probabilities.

A. Assumptions

The following two assumptions are adopted in this paper.

A1: The initial belief values are bounded between p_{01} and

p_{11} .

A2:

$$\epsilon \leq \frac{\min\{p_{01}, p_{11}\}(1 - \max\{p_{01}, p_{11}\})}{\max\{p_{01}, p_{11}\}(1 - \min\{p_{01}, p_{11}\})}.$$

Assumption A1 will only be used in Theorem 1 which describes the structure of the myopic policy. We note that the structure can be directly extended if assumption A1 does not hold. We assume A1 in Theorem 1 for the easy of presentation.

For Assumption A2, the allowed probability of miss detection δ plays a major role since ϵ can be reduced to an arbitrarily small value at the price of increased δ . However, both ϵ and δ can be improved by increasing the sensing/detection time (*i.e.*, taking more measurements). The caveat is the reduced transmission time for a given slot length. This interesting tradeoff between the complexity of the detector at the physical layer and the transmission strategy at the Medium Access Control (MAC) layer of a communication network can be complex and is beyond the scope of this paper.

B. Structure

The implementation of the myopic policy can be described with a queue structure: all N channels are ordered in a queue, and in each slot, those M channels at the head of the queue are sensed.

Theorem 1: The Semi-Universal structure of the myopic policy

The initial channel ordering $\mathbf{Q}(1)$ is determined by the initial belief vector as given below.

solve its stationary distribution. Under single-channel sensing ($M = 1$), the approach is to construct first-order Markov chains that stochastically dominate or are dominated by $\{L_k\}_{k=1}^{\infty}$. The stationary distributions of these first-order Markov chains, which can be obtained in closed-form, lead to lower and upper bounds on U according to (5).

Theorem 2: Define functions

$$f(x) \triangleq \frac{\omega_o - x}{1 - x(1 - \epsilon) \left(1 - \frac{(p_{11} - p_{01})(1 - p_{11}(1 - \epsilon))}{1 - (p_{11} - p_{01})p_{11}(1 - \epsilon)}\right)},$$

$$h(x, y, z, a, b) \triangleq \frac{1 - \omega_o(1 - \epsilon) + a}{1 - a \left(\frac{(y(p_{11} - p_{01})^2 + (p_{11} - p_{01})^{b+1})z}{1 - (p_{11} - p_{01})y} - x \right)},$$

and for any function $v(\cdot)$ of vector $[x, y, z, a, b]$, define the following operator

$$\mathcal{G}[v(x, y, z, a, b)] \triangleq \frac{1}{\left(\frac{(2-y)z}{(1-y)^2} - x\right)v(x, y, z, a, b) + 1}. \quad (6)$$

Under assumption A2, we have the following lower and upper bounds on the throughput U when $M = 1$.

- *Case 1:* $p_{11} \geq p_{01}$

$$\frac{f(c_1)(1 - \epsilon)}{1 - (p_{11} - f(c_1))(1 - \epsilon)} \leq U \leq \frac{\omega_o(1 - \epsilon)}{1 - (p_{11} - \omega_o)(1 - \epsilon)},$$

where ω_o is given by (2) and

$$\begin{aligned} c_1 &= (\omega_o - c_2)(p_{11} - p_{01})^{N-1}, \\ c_2 &= \frac{p_{01}(1 - p_{01} + \epsilon p_{11})}{1 - p_{01} + \epsilon p_{01}}. \end{aligned}$$

- *Case 2:* $p_{11} < p_{01}$

$$\mathcal{G}[h(x_1, y_1, z_1, a_1, 2N-4)] \leq U \leq \mathcal{G}[h(\frac{1}{x_1}, 1 - z_1, 1 - y_1, a_1, 3)],$$

where $x_1 = \frac{p_{01}}{p_{11}(p_{11} - p_{01}) + p_{01}}$,

$$y_1 = 1 - (1 - \epsilon)(p_{11}(p_{11} - p_{01}) + p_{01}),$$

$$z_1 = (1 - \epsilon)p_{01},$$

$$a_1 = (1 - \epsilon)(\omega_o - p_{11})(p_{11} - p_{01}).$$

For multi-channel sensing ($M > 1$), it is difficult to construct first-order Markov process to stochastically dominate or be dominated by $\{L_k\}_{k=1}^{\infty}$. Instead, we establish a uniform statistical bound on the distributions of all TPs based on the structure of the myopic policy. The bounds on the throughput when applied to $M = 1$ are thus looser than those under single-channel sensing scenarios as given in Theorem 2.

Theorem 3: Recall the definition of the operator $\mathcal{G}[\cdot]$ given in (6). Under assumption A2, we have the following lower and upper bounds on throughput U when $M > 1$.

- *Case 1:* $p_{11} \geq p_{01}$

$$M(1 - \epsilon) \max\left\{\frac{c_3}{1 - (p_{11} - c_3)(1 - \epsilon)}, \omega_o\right\} \leq U \leq \frac{M\omega_o(1 - \epsilon)}{1 - (p_{11} - \omega_o)(1 - \epsilon)},$$

where $c_3 = \omega_o - (\omega_o - \frac{\epsilon p_{01}}{\epsilon p_{01} + 1 - p_{01}})(p_{11} - p_{01})^{\lfloor \frac{N}{M} \rfloor}$.

- *Case 2:* $p_{11} < p_{01}$

$$\begin{aligned} M \max\{\mathcal{G}[v_1(x_1, y_1, z_1, a_1, b_1)], \omega_o(1 - \epsilon)\} &\leq U \\ &\leq M\mathcal{G}[v_2(\frac{1}{x_1}, 1 - z_1, 1 - y_1, a_1, b_1)], \end{aligned}$$

where

$$v_1(\cdot) = 1 - (\omega_o - (\omega_o - p_{11})(p_{11} - p_{01})^{2\lfloor \frac{N}{M} \rfloor - 2})(1 - \epsilon),$$

$$v_2(\cdot) = 1 - (p_{11}(p_{11} - p_{01}) + p_{01})(1 - \epsilon),$$

$$x_1 = \frac{p_{01}}{p_{11}(p_{11} - p_{01}) + p_{01}},$$

$$y_1 = 1 - (p_{11}(p_{11} - p_{01}) + p_{01})(1 - \epsilon),$$

$$z_1 = (1 - \epsilon)p_{01}.$$

Note that a_1 and b_1 can be arbitrary since they are arguments of the constant functions v_1 and v_2 .

Corollary 2: For $p_{11} > p_{01}$, the lower bound on throughput U increasingly converges to the constant upper bound at geometrical rate $(p_{11} - p_{01})^{\frac{1}{M}}$ as N increases; for $p_{11} < p_{01}$, the lower bound on U increasingly converges to a constant at geometrical rate $(p_{01} - p_{11})^{\frac{2}{M}}$.

The monotonicity of the difference between the upper and lower bounds with respect to N illustrates that the performance of the multichannel opportunistic system improves with the number N of channels, as suggested by intuition. For $p_{11} \geq p_{01}$, the upper bound gives the limiting performance of the system when $N \rightarrow \infty$ (under the conjectured optimality of the myopic policy). However, for a fixed sensing capacity M , the throughput in the multichannel opportunistic system saturates quickly as the number of channels goes to infinity (see Corollary 2). Since the saturating rate is decreasing with M , for a system consisting of a large number of channels, it is crucial to enhance the sensing capacity M to the level under which the saturation can be avoided in order to fully exploit the opportunities offered by a large number of channels.

D. Approximation Factor of the Myopic Policy

Although the optimality of the myopic policy is proved for $N = 2$ and conjectured for general scenarios based on numerical results, proving this conjecture or establishing simple sufficient conditions for the general optimality of the myopic policy appears to be challenging. Under discounted reward criterion, we have shown that so long as the discount factor is less than $1/(M + 1)$, the myopic policy is optimal for all N . In this section, we take a further step toward the optimality of the myopic policy by characterizing its approximation factor with respect to the optimal policy. Specifically, we will bound from below the approximation factor η of the myopic policy, which is defined as the ratio of the throughput achieved by the myopic policy to that achieved by the optimal policy.

In the genie-aided system, the secondary user still senses, accesses, and accrues rewards among M channels in each slot. However, at the end of each slot, the genie will inform the secondary user the observations (ACK/NAK) that would have been obtained from all *unobserved* channels if they had also been sensed and subsequently accessed based on the sensing outcomes. As a consequence, the secondary user will obtain ACK/NAK from all N channels at the end of each slot. Clearly, the optimal policy in the genie-aided system is given by the myopic policy since the current sensing action

will not affect the belief transitions as well as the future reward. The optimal performance of the genie-aided system can thus be upper bounded as given in Lemma 1 below.

Lemma 1: Define $x \triangleq \frac{\epsilon p_{11}^2 + p_{01} - p_{01} p_{11}}{\epsilon p_{11} + 1 - p_{11}}$. Under assumption A2, the maximum steady-state throughput \bar{U} in the genie-aided system is upper bounded as given below.

- *Case 1:* $p_{11} \geq p_{01}$

$$\bar{U} \leq \min\{Mp_{11} - \sum_{k=0}^M \binom{N}{k} d_k, N\omega_o\}(1 - \epsilon),$$

where $d_k = (M-k)(p_{11}-x)(\omega_o(1-\epsilon))^k(1-\omega_o(1-\epsilon))^{N-k}$.

- *Case 2:* $p_{11} < p_{01}$

$$\bar{U} \leq \min\{Mx - \sum_{k=0}^M \binom{N}{k} e_k, N\omega_o\}(1 - \epsilon),$$

where $e_k = (M-k)(x-p_{11})(\omega_o(1-\epsilon))^{N-k}(1-\omega_o(1-\epsilon))^k$.

The throughput of the genie-aided system provides an upper bound on the optimal performance and a performance benchmark of all sensing policies, including the myopic policy. Combining the lower bound on the throughput achieved by the myopic policy as given in Sec. III-C.2, we bound the approximation factor η of the myopic policy as given in Theorem 4 below.

Theorem 4: Under assumption A2, the approximation factor of the myopic policy is lower bounded by

$$\eta \geq \begin{cases} \frac{M}{N}, & \text{if } p_{11} > p_{01} \\ \max\{\frac{1}{2}, \frac{M}{N}\}, & \text{if } p_{11} < p_{01} \\ 1, & \text{if } p_{11} = p_{01} \text{ or } N = 2 \end{cases}. \quad (7)$$

IV. NUMERICAL EXAMPLES

In this section, we demonstrate the tightness of the bounds on U given in Sec. III-C.2 and the upper bound on the optimal performance given by the genie-aided system given in Sec. III-D. We obtain the performance of the myopic policy and the optimal performance in the genie-aided system by Monte Carlo simulations. From Fig. 3, we observe that the lower bound on the performance of the myopic policy quickly converges to the upper bound as $N \rightarrow \infty$ when channels are positively correlated. We also observe from Fig. 3 that the upper bound on the optimal performance in the genie-aided system is tight.

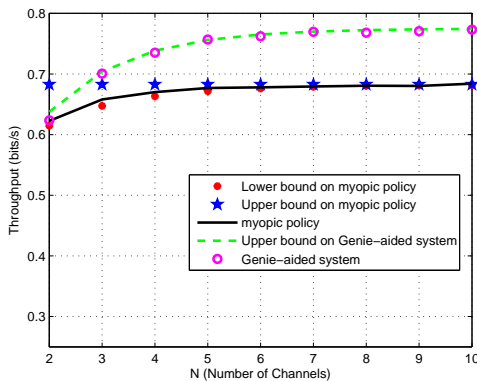


Fig. 3. Performance bounds of the myopic policy ($M = 1$, $p_{11} = 0.8$, $p_{01} = 0.2$, $\epsilon = 0.0312$).

V. CONCLUSION AND FUTURE WORK

In this paper, we have analyzed the performance of the myopic sensing policy in multi-channel opportunistic access under an independent and stochastically identical Gilbert-Elliott channel model with noisy state observations. The obtained analytical results allow us to bound the worst case performance of the myopic policy and to systematically examine the impact of the number of channels and channel dynamics (transition probabilities) on the system performance. Future work includes proving the optimality conjecture of the myopic policy, and generalization to independent and stochastically non-identical channel model by investigating Whittle's index policy.

REFERENCES

- [1] E.N. Gilbert, "Capacity of burst-noise channels," *Bell Syst. Tech. J.*, vol. 39, pp. 1253-1265, Sept. 1960. WA), pp. 331-335, June, 1995.
- [2] Q. Zhao and B. Sadler, "A Survey of Dynamic Spectrum Access," *IEEE Signal Processing magazine*, vol. 24, pp. 79-89, May, 2007.
- [3] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework," in *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589-600, April, 2007.
- [4] C. H. Papadimitriou and J. N. Tsitsiklis, "The Complexity of Optimal Queueing Network Control," in *Mathematics of Operations Research*, Vol. 24, No. 2, pp. 293-305, May, 1999.
- [5] Y. Chen, Q. Zhao, and A. Swami, "Joint Design and Separation Principle for Opportunistic Spectrum Access in the Presence of Sensing Errors," in *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2053-2071, May, 2008.
- [6] Q. Zhao, B. Krishnamachari, and K. Liu, "Low-Complexity Approaches to Spectrum Opportunity Tracking," in *Proc. of the 2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications*, August, 2007.
- [7] Q. Zhao, B. Krishnamachari, and K. Liu, "On Myopic Sensing for Multi-Channel Opportunistic Access: Structure, Optimality, and Performance," in *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431-5440, Dec., 2008.
- [8] S. H. Ahmad, M. Liu, T. Javadi, Q. Zhao and B. Krishnamachari, "Optimality of Myopic Sensing in Multi-Channel Opportunistic Access," *IEEE Transactions on Information Theory*, Sept., 2009.
- [9] K. Liu and Q. Zhao, "Indexability of Restless Bandit Problems and Optimality of Whittle's Index for Dynamic Multichannel Access," submitted to *IEEE Transactions on Information Theory*, November, 2008. Available at <http://arxiv.org/abs/0810.4658> (conference versions appeared in *Proc. of the 5th IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON) Workshops*, June, 2008 and *Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, October, 2008).
- [10] J. Le Ny, M. Dahleh, E. Feron, "Multi-UAV Dynamic Routing with Partial Observations using Restless Bandit Allocation Indices," in *Proceedings of the 2008 American Control Conference*, Seattle, WA, June, 2008.
- [11] S. Filippi, O. Cappé, F. Clérot, and E. Moulines, "A Near Optimal Policy for Channel Allocation in Cognitive Radio," *Recent Advances in Reinforcement Learning*, Vol. 5323, pp. 69-81, Nov. 27, 2008.
- [12] M. Hoyhtya, S. Pollin, A. Mammela, "Performance improvement with predictive channel selection for cognitive radios," in *Proc. of the First International Workshop on Cognitive Radio and Advanced Spectrum Management*, Feb., 2008.
- [13] J. Unnikrishnan and V. Veeravalli, "Algorithms for Dynamic Spectrum Access with Learning for Cognitive Radio," to appear in *IEEE Transactions on Signal Processing*. Available at <http://arxiv.org/abs/0807.2677>.
- [14] R. Smallwood and E. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Operations Research*, pp. 1071-1088, 1971.
- [15] K. Liu, Q. Zhao, and B. Krishnamachari, "Dynamic Multichannel Access with Imperfect Channel State Detection," submitted to *IEEE Transactions on Signal Processing*, July, 2009.