Dynamic Multichannel Access With Imperfect Channel State Detection

Keqin Liu, Qing Zhao, and Bhaskar Krishnamachari

***Abstract*—A restless multi-armed bandit problem that arises in multichannel opportunistic communications is considered, where channels are modeled as independent and identical Gilbert–Elliot channels and channel state detection is subject to errors. A simple structure of the myopic policy is established under a certain condition on the false alarm probability of the channel state detector. It is shown that myopic actions can be obtained by maintaining a simple channel ordering without knowing the underlying Markovian model. The optimality of the myopic policy is proved for the case of two channels and conjectured for general cases. Lower and upper bounds on the performance of the myopic policy are obtained in closed-form, which characterize the scaling behavior of the achievable throughput of the multichannel opportunistic system. The approximation factor of the myopic policy is also analyzed to bound its worst-case performance loss with respect to the optimal performance.**

***Index Terms*—Cognitive radio, dynamic multichannel access, myopic policy, restless multi-armed bandit.**

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 $\{p_{ij}\}_{i,j=0,1}$. This channel model has been commonly used to abstract physical channels with memory (see [2], [3] and references therein). An emerging application of this channel model is cognitive radio for opportunistic spectrum access where secondary users search in the spectrum for idle channels temporarily unused by primary users [4]. This two-state Markovian channel model has been shown to fit well with actual spectrum usage [5]–[7]. For this application, the "good" state represents an idle channel while the "bad" state an occupied channel.

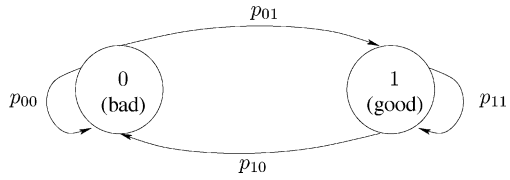

Fig. 1. The Gilbert–Elliot channel model.

Due to limited sensing capability, in each time slot, a user can only sense a subset of $M(M < N)$ channels and subsequently access those channels sensed to be in the good state. 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 to maximize the expected long-term reward.

B. Restless Multi-Armed Bandit and the Myopic Policy

The above dynamic multichannel access problem can be formulated as a partially observable Markov decision process (POMDP) for generally correlated channels [8], or a restless multi-armed bandit problem (RMAB) for independent channels as considered here. The maximum expected total reward of the multichannel opportunistic system is essentially the value function of an RMAB. Unfortunately, obtaining optimal solutions to a general restless bandit process is PSPACE-hard [9], and analytical characterizations of the performance of the optimal policy are often intractable.

In this paper, we show that for the special class of RMAB that arises in the dynamic multichannel access problem, simple structural policies exist that achieve a strong performance with low complexity. Specifically, we show that the myopic policy, which maximizes the expected immediate reward while ignoring the impact of the current action on the future, has a simple structure when the 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} (i.e., the sign of the correlation between the channel states in two consecutive

Manuscript received July 24, 2009; accepted December 30, 2009. Date of publication January 26, 2010; date of current version April 14, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ye (Geoffrey) Li. The work of K. Liu and Q. Zhao 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. The work of B. Krishnamachari was supported by the National Science Foundation under Grant CNS-0347621. This work was presented in part at the Second International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom), August 2007, and the Forty-Eighth IEEE Conference on Decision and Control, December 2009.

K. Liu and Q. Zhao are with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA (e-mail: kqliu@ucdavis.edu; qzhao@ucdavis.edu).

B. Krishnamachari is with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: bkrishna@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2010.2041600

1053-587X/$26.00 © 2010 IEEE

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

slots[1]). The myopic policy can thus be implemented without knowing the transition probabilities $\{p_{11}, p_{01}\}$ except their order, and it automatically tracks variations in the channel model provided that the order of p_{11} and p_{01} remains unchanged. Furthermore, we show that with such a simple and robust structure, the myopic policy achieves the optimal performance for $N = 2$. Numerical examples[2] suggest its optimality for the general case of $N > 2$.

To analytically characterize the performance of the myopic policy for $N > 2$, we develop closed-form lower and upper bounds on the steady-state throughput achieved by the myopic policy. The lower bound monotonically approaches to the upper bound as the number N of channels increases. This result thus defines the limiting performance of the myopic policy as N approaches to infinity. Furthermore, by analyzing a genie-aided system which provides an upper bound on the optimal performance, we characterize the approximation factor of the myopic policy to bound the worst-case performance loss of the myopic policy with respect to the optimal policy. Specifically, we show that the myopic policy achieves *at least* (M/N) of the optimal performance when channels are positively correlated, and at least $\max\{(1/2), (M/N)\}$ of the optimal performance when channels are negatively correlated.

This paper extends our earlier work in [10] that assumes perfect detection of the channel states. As shown in Sections II and III, communication constraints, namely, synchronization in channel selections between the transmitter and its receiver, require a different formulation of the problem when observations are imperfect, and the uncertainties in the sensed channel state complicate the analysis of the myopic policy. A detailed discussion on other related work is given in Section V.

II. Problem Formulation

In this section, we formulate the problem by considering the cognitive radio application. The general problem, however, finds applications in downlink scheduling in a fading environment, jamming and anti-jamming and target tracking in multi-agent systems.

A. System Model

Let $\mathcal{S}(t) \triangleq [S_1(t), \ldots, S_N(t)]$ 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 test[3]:

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

[1]It is easy to show that $p_{11} > p_{01}$ corresponds to the case where the channel states in two consecutive slots are positively correlated, i.e., for any distribution of $S(t)$, we have $\mathbb{E}[(S(t) - \mathbb{E}[S(t)])(S(t+1) - \mathbb{E}[S(t+1)])] > 0$, where $S(t)$ is the state of the Gilbert–Elliot channel in slot t. Similar, $p_{11} < p_{01}$ corresponds to the case where $S(t)$ and $S(t+1)$ are negatively correlated, and $p_{11} = p_{01}$ the case where $S(t)$ and $S(t+1)$ are independent.

[2]Actions given by the myopic policy and the optimal policy are compared numerically for randomly chosen p_{11} and p_{01} and $N = 3, 4$, and 5. Extensive numerical comparisons indicate the equivalence between the myopic policy and the optimal policy.

[3]We 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 performance of channel state detection is characterized by the receiver operating characteristic (ROC) which relates the probability of false alarm ϵ to the probability of miss detection δ:

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

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 of collision $\mathcal{P}_n(t)$ perceived by the primary network in any channel and in any slot should be 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 selections. This problem is formulated as a constrained POMDP in [8] for generally correlated channels. A separation principle has been established in [11] which decouples the design of the channel state detector and the access policy from that of the channel sensing policy. Specifically, the optimal joint design can be carried out in two steps: first to choose the channel state detector and the access policy to maximize the immediate reward under the collision constraint, and then to choose the sensing policy to maximize the total reward over a finite horizon of T slots. It has been shown in [11] that the first step can be solved in closed-form. Specifically, the optimal channel state detector is the Neyman–Pearson detector operating at the point that the probability δ of miss detection is equal to 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. Using the optimal design of the detector operating point and the access policy, we can then obtain the optimal sensing policy in the second step as an *unconstraint* POMDP problem, or an RMAB problem for independent channels.

The complexity issue in this problem, however, is only partially solved since the resulting unconstraint POMDP for designing the optimal sensing policy in the second step still suffers from the curse of dimensionality. The focus of this paper is on developing low-complexity sensing policies and studying their optimality and performance. Specifically in Section III, we will show that for independent and stochastically identical channels, the myopic sensing policy has a simple semi-universal structure and achieves the optimality for $N = 2$. For $N > 2$, we bound the steady-state throughput achieved by the myopic policy and establish its approximation factor with respect to the optimal policy.

Since failed transmissions may occur, acknowledgements (ACKs) are necessary to ensure guaranteed delivery.

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

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 in this slot. 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 $[S_1(t),\ldots,S_N(t)] \in \{0,1\}^N$ in slot t is not fully observable to the user. The user 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 [12]. Referred to as the belief vector, this sufficient statistic is denoted by $\Omega(t)\triangleq[\omega_1(t),\ldots,\omega_N(t)]$, 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 $\mathcal{K}(t) \in \{0\ (\text{NAK}), 1\ (\text{ACK})\}^M$ in each slot rather than the detection outcomes at the transmitter. Let $I(t)$ denote the sensing action that consists of the indexes of the M channels to be sensed 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:

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

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

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

Note that the belief update under $K_i(t)=0$ results from the fact that the receiver cannot distinguish a failed transmission (i.e., colliding with the primary users, which occurs with probability $\delta(1-\omega_i(t))$) from no transmission (which occurs with probability $\epsilon\omega_i(t)+(1-\delta)(1-\omega_i(t))$).

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

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

where $R(\pi_t(\Omega(t)))$ is the reward obtained when the belief is $\Omega(t)$ and channels $\pi_t(\Omega(t))$ are selected, $\Omega(1)$ is the initial belief vector, T is the horizon length, and $0\le\beta\le 1$ is the discount factor. This problem falls into the model of an RMAB by treating the belief value of each channel as the state of each arm of a restless 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 discounted reward that can be obtained starting from slot t up to T given the belief vector Ω. Given that the user takes action I and observes $\mathcal{K}=\{K_i\}_{i\in I}$, the expected discounted 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 discounted future reward $\beta V_{t+1}(\mathcal{T}(\Omega|I,\mathcal{K}))$, 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 equations:

$$V_t(\Omega(t))=\max_I\{\sum_{i\in I}\omega_i(t)(1-\epsilon) + \beta\mathbb{E}[V_{t+1}(\mathcal{T}(\Omega(t)|I,\mathcal{K}))]\} \quad (3)$$

$$V_T(\Omega(T))=\max_I\sum_{i\in I}\omega_i(T)(1-\epsilon). \quad (4)$$

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, Optimality, 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 prove that the myopic policy is optimal for $N=2$ and characterize the performance of the myopic policy for general M and N.

The myopic policy $\hat{\pi}$ 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 $\Omega=[\omega_1,\ldots,\omega_N]$ is simply given by

$$\hat{I}(\Omega)=\arg\max_I\sum_{i\in I}\omega_i. \quad (5)$$

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 $\omega_i(1)$ are bounded between p_{01} and p_{11}.

A2) The false alarm probability ϵ of the channel state detector is upper bounded by

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

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

Assumption A1) will only be used in Theorem 1 and part of Theorem 2. We note that if Assumption A1) does not hold, the initial belief values will be *transient* since the belief values starting from the second slot are always bounded between p_{01} and p_{11} [see (1)]. Consequently, the structure of the myopic policy given in Theorem 1 can be easily extended by treating the first slot separately from the future slots. We assume A1) in Theorem 1 for the ease of presentation. We also note that Assumption A1) will hold if the initial believe values are given by the stationary distribution ω_o [see (2)] of the underlying Markov chain, which is often the case in practical systems.

For Assumption A2), the allowed probability of miss detection δ (note that the optimal operating point is given by $\delta = \zeta$) 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

In this subsection, we show that the implementation of the myopic policy can be described with a simple queue structure. Specifically, all N channels are ordered in a queue, and in each slot, those M channels at the head of the queue are sensed. At the end of each slot, the positions of the channels in the queue are reordered based on the common observations of ACK/NAK. This structure of the myopic policy is detailed in Theorem 1 below.

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.

$$\omega_{n_1}(1) \geq \cdots \geq \omega_{n_N}(1) \Longrightarrow \mathbf{Q}(1) = (n_1, \ldots, n_N).$$

Under Assumptions A1) and A2), channels are reordered at the end of each slot according to the following simple rules. When $p_{11} \geq p_{01}$, the channels over which ACKs are observed will stay at the head of the queue, and the channels over which NAKs are observed will be moved to the end of the queue while keeping their order unchanged [see Fig. 2(a)]. When $p_{11} < p_{01}$, the channels over which NAKs are observed will stay at the head of the queue, and the channels over which ACKs are observed will be moved to the end of the queue. The order of the unobserved channels are reversed [see Fig. 2(b)].

Proof: See Appendix A. ■

The above simple structure suggests that the myopic sensing policy is particularly attractive in implementation: the myopic policy does not require any computation except maintaining a queueing order of channels. Contrast to the exponential complexity of solving the dynamic programming given by (3)–(4), the myopic policy has only a linear complexity with both N and T. Besides its simplicity, the myopic policy obviates the need for knowing the channel transition probabilities and automatically tracks variations in the channel model provided that the sign of the channel state correlation (i.e., the order of p_{11} and p_{01}) remains unchanged.

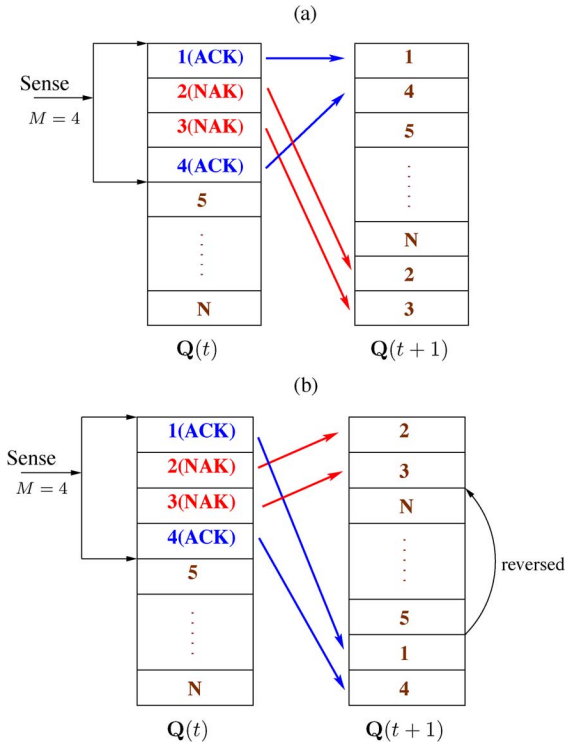

Fig. 2. The structure of the myopic policy ($M = 4$, the initial belief is assumed to satisfy $\omega_1(1) \geq \omega_2(1) \geq \cdots \geq \omega_N(1)$). (a) $p_{11} \geq p_{01}$, (b) $p_{11} < p_{01}$.

We point out that the structure of the myopic sensing policy in the presence of sensing errors is similar to that under the perfect sensing scenario given in [10]. The proof, however, is more involved since the observations here are acknowledgements and the state of the sensed channel cannot be inferred with certainty from a NAK.

Following the *belief-independence* property of this simple structure, we present the following corollary which allows us to work with a Markov reward process with a finite state space instead of one with an uncountable state space (i.e., the belief vectors) as we would encounter in a general POMDP.

Corollary 1: Let $\mathbf{Q}(t) = (n_1, n_2, \ldots, n_N)(n_i \in \{1, 2, \ldots, N\}\ \forall i)$ be the order of channels in slot t, where the myopic action $\hat{I}(t) = \{n_i\}_{i=1}^{M}$. Under Assumption A2), the ordered channel states $\vec{\mathbf{S}}(t) \triangleq [S_{n_1}(t), S_{n_2}(t), \ldots, S_{n_N}(t)]$ form a 2^N-state Markov chain, and the performance of the myopic policy is determined by the Markov reward process $(\vec{\mathbf{S}}(t), R(t))$ with $R(t) = \sum_{i=1}^{M} S_{n_i}(t)(1-\epsilon)$, where ϵ is the probability of false alarm.

Proof: See Appendix B. ■

Theorem 1 and Corollary 1 provide the foundation for analyzing the optimality and performance of the myopic policy in subsequent subsections.

C. Optimality

Theorem 2: The Optimality of the Myopic Policy: For $N = 2$, the myopic policy $\hat{\pi}$ is optimal in the following sense:

i) it maximizes the expected total discounted/undiscounted reward over a finite horizon under Assumptions A1) and A2), i.e.,

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

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

where $0 \leq \beta \leq 1$ is the discount factor;

ii) it maximizes the expected total discounted reward over an infinite horizon under Assumptions A1) and A2), i.e.,

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

where $0 \leq \beta < 1$;

iii) it maximizes the expected average reward over an infinite horizon under Assumption A2), i.e.,

$$\hat{\pi} = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} R(\pi_t(\Omega(t))) | \Omega(1) \right].$$

Proof: See Appendix C. ∎

Based on Theorem 2 and extensive numerical examples, we conjecture that the myopic policy is optimal for general M and N. Unfortunately, proving this conjecture or establishing sufficient conditions for the general optimality of the myopic policy appears to be challenging. Under the assumption of perfect channel state detection and single-channel sensing ($M = 1$), we have established in [10] the semi-universal structure of the myopic policy for general N and its optimality for $N = 2$. A recent follow-up work [13] has extended the optimality of the myopic policy to i) $N = 3$ and ii) $N \geq 3$ when channels are positively correlated. The proofs in [13] hinge on the simple structure of the myopic policy, which is preserved in the presence of sensing errors as shown in Theorem 1 (when applied to the case of $M = 1$). We thus hope that the optimality of the myopic policy in the general case also carries over to the imperfect sensing scenario. Extending the results in [13] to the imperfect sensing case is, however, highly nontrivial since sensing errors introduce further complications in the system dynamics (i.e., belief updates). Under the discounted reward criterion, our ongoing work shows that when the discount factor is less than $1/(M+1)$, the myopic policy is optimal for all M and N.

D. Performance

In this subsection, we analyze the performance of the myopic policy. Specifically, we establish lower and upper bounds on the system steady-state throughput achieved by the myopic policy. This result allows us to study the scaling behavior of the performance limit (based on the conjectured optimality of the myopic policy) of a multichannel opportunistic communications system as the number of channels increases. The lower bound also provides a closed-form worst-case performance of the myopic policy, which, combined with a closed-form upper bound on the system maximum throughput obtained based on a genie argument (see Section III-E), leads to characterizations of the approximation factor of the myopic policy to bound its worst case performance with respect to the optimal performance (see Section III-E).

1) Uniqueness of Steady-State Performance and Its Numerical Evaluation: We first establish the existence and uniqueness of the system steady-state performance under the myopic policy. The steady-state throughput under the myopic policy is defined as

$$U(\Omega(1)) \triangleq \lim_{T\to\infty} \frac{\hat{V}_{1:T}(\Omega(1))}{T} \tag{6}$$

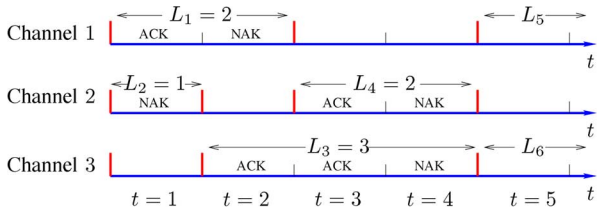

Fig. 3. The transmission period structure assuming $p_{11} \geq p_{01}, M = 2, N = 3$ (the realizations of channel order in the queue are as follows: $\mathbf{Q}(1) = [1,2,3], \mathbf{Q}(2) = [1,3,2], \mathbf{Q}(3) = \mathbf{Q}(4) = [3,2,1]$, and $\mathbf{Q}(5) = [1,3,2]$).

where $\hat{V}_{1:T}(\Omega(1))$ is the expected total reward obtained in T slots under the myopic policy when the initial belief is $\Omega(1)$. From Corollary 1, $U(\Omega(1))$ is determined by the Markov reward process $(\vec{\mathbf{S}}(t), R(t))$. It is easy to see that the 2^N-state Markov chain $\{\vec{\mathbf{S}}(t)\}$ is irreducible and aperiodic, thus has a limiting distribution. As a consequence, the limit in (6) exists, and the steady-state throughput U is independent of the initial belief value $\Omega(1)$.

Corollary 1 also provides a numerical approach to evaluating U by calculating the limiting (stationary) distribution of $\{\vec{\mathbf{S}}(t)\}$ whose transition probabilities can be directly obtained from the transition probabilities of the channel states. This numerical approach, however, does not provide an analytical characterization of the throughput U in terms of the number N of channels and the transition probabilities $\{p_{ij}\}$. In the next subsection, we obtain analytical bounds on U and its scaling behavior with respect to N based on a stochastic dominance argument.

2) Analytical Characterization of Throughput: From the structure of the myopic policy, the throughput is determined by how often each channel is moved to the end of the queue. When $p_{11} \geq p_{01}$, the event of moving a channel to the end of the queue is equivalent to a slot *without* reward on this channel. The opposite holds when $p_{11} < p_{01}$: moving a channel to the end of the queue corresponds to a slot *with* reward on this channel.

We thus introduce the concept of *transmission period (TP)* defined with respect to each channel; it is the time period when a channel is continuously sensed before being moved to the end of the queue (see Fig. 3 for an example). We index the transmission periods over all channels in the order of its starting point.[4] Let L_k denote the length of the kth TP. We then have a discrete-time random process $\{L_k\}_{k=1}^{\infty}$ taking values of positive integers.

Lemma 1: Under Assumption A2), we have

$$U = \begin{cases} M(1 - 1/\bar{L}), & p_{11} \geq p_{01} \\ M/\bar{L}, & p_{11} < p_{01} \end{cases} \tag{7}$$

where $\bar{L} = \lim_{K\to\infty} (\sum_{k=1}^{K} L_k)/(K)$ denotes the average length of a TP.

Proof: See Appendix D. ∎

Based on Lemma 1, throughput analysis is reduced to analyzing the average TP length $\bar{L}$, which is determined by the stationary distribution of the random process $\{L_k\}_{k=1}^{\infty}$. For $N = 2$ and channels are negatively correlated, $\{L_k\}_{k=1}^{\infty}$ is a first-order Markov chain where the stationary distribution can be solved in closed-form, leading to the closed-form average TP length $\bar{L}$. Unfortunately in general, $\{L_k\}_{k=1}^{\infty}$ is a random process with

[4]When several TPs have the same starting point, their indexes can be set arbitrarily.

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

high-order memory and it is difficult to 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 (7). Specifically, for $p_{11} \geq p_{01}$, a lower bound on U is obtained by constructing a first-order Markov chain whose stationary distribution is stochastically dominated by the stationary distribution of $\{L_k\}_{k=1}^{\infty}$. An upper bound on U is given by a first-order Markov chain whose stationary distribution stochastically dominates the stationary distribution of $\{L_k\}_{k=1}^{\infty}$. Similarly, bounds on U can be obtained for $p_{11} < p_{01}$.

Theorem 3: The Lower and Upper Bounds on the Throughput Achieved by the Myopic Policy: Define functions

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

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

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

$$\mathcal{G}[v(x,y,z,a,b)] \triangleq \frac{1}{\left(\frac{(2-y)z}{(1-y)^2} - x\right)v(x,y,z,a,b) + 1}. \quad (8)$$

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

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

$$\frac{f(c_1)(1-\epsilon)}{1-(p_{11}-f(c_1))(1-\epsilon)} \leq U \leq \frac{\omega_o(1-\epsilon)}{1-(p_{11}-\omega_o)(1-\epsilon)} \quad (9)$$

where ω_o is given by (2) and

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

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

$$\mathcal{G}[h(x_1,y_1,z_1,a_1,2N-4)] \leq U \leq \mathcal{G}\left[h\left(\frac{1}{x_1},1-z_1,1-y_1,a_1,3\right)\right] \quad (10)$$

where

$$x_1 = \frac{p_{01}}{p_{11}(p_{11}-p_{01})+p_{01}}$$
$$y_1 = 1-(1-\epsilon)(p_{11}(p_{11}-p_{01})+p_{01})$$
$$z_1 = (1-\epsilon)p_{01}$$
$$a_1 = (1-\epsilon)(\omega_o - p_{11})(p_{11}-p_{01}).$$

Proof: See Appendix E. ∎

For multi-channel sensing ($M > 1$), it is difficult to construct a first-order Markov process to stochastically dominate or be dominated by $\{L_k\}_{k=1}^{\infty}$. The main approach is to establish a uniform statistical bound on the distributions of all TPs based on the structure of the myopic policy. The bound is thus looser than those given in Theorem 3 when applied to the case of $M = 1$.

Theorem 4: Recall the definition of the operator $\mathcal{G}[\cdot]$ given in (8). Under Assumption A2), we have the following lower and upper bounds on throughput U when $M > 1$.

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

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

where

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

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

$$M\max\{\mathcal{G}[v_1(x_1,y_1,z_1,a_1,b_1)], \omega_o(1-\epsilon)\} \leq U \leq M\mathcal{G}\left[v_2\left(\frac{1}{x_1},1-z_1,1-y_1,a_1,b_1\right)\right] \quad (12)$$

where

$$v_1(\cdot) = 1-\left(\omega_o - (\omega_o - p_{11})(p_{11}-p_{01})^{2\lfloor \frac{N}{M} \rfloor - 2}\right)(1-\epsilon),$$
$$v_2(\cdot) = 1-(p_{11}(p_{11}-p_{01})+p_{01})(1-\epsilon),$$
$$x_1 = \frac{p_{01}}{p_{11}(p_{11}-p_{01})+p_{01}},$$
$$y_1 = 1-(p_{11}(p_{11}-p_{01})+p_{01})(1-\epsilon),$$
$$z_1 = (1-\epsilon)p_{01}.$$

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

Proof: See Appendix F. ∎

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

Proof: See Appendix G. ∎

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 $N \to \infty$ (under the conjecture on the 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.

E. Approximation Factor

As mentioned in Section III, the optimality of the myopic policy for general M and N is conjectured based on numerical results. In this section, we aim to establish analytic results on the

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

worst-case performance of the myopic policy 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. The main approach is to construct an upper bound on the maximum system throughput by considering a genie-aided system. Combining this upper bound with the lower bound on the performance achieved by the myopic policy as shown in Section III-D-2), we obtain a uniform lower bound on the approximation factor η, which is independent of channel parameters.

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 2 below.

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

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

$$\bar{U} \leq \min\left\{\left(Mp_{11} - \sum_{k=0}^{M}\binom{N}{k} d_k\right)(1-\epsilon), N\omega_o(1-\epsilon)\right\} \tag{13}$$

where

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

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

$$\bar{U} \leq \min\left\{\left(Mx - \sum_{k=0}^{M}\binom{N}{k} e_k\right)(1-\epsilon), N\omega_o(1-\epsilon)\right\} \tag{14}$$

where

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

Proof: See Appendix H. ∎

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 Section III-D-2), we bound the approximation factor η of the myopic policy as given in Theorem 5 below.

Theorem 5: Under Assumption A2), the approximation factor of the myopic policy is lower bounded by

$$\eta \geq \begin{cases} \frac{M}{N}, & \text{if } p_{11} > p_{01} \\ \max\left\{\frac{1}{2}, \frac{M}{N}\right\}, & \text{if } p_{11} < p_{01} \\ 1, & \text{if } p_{11} = p_{01} \text{ or } N = 2. \end{cases} \tag{15}$$

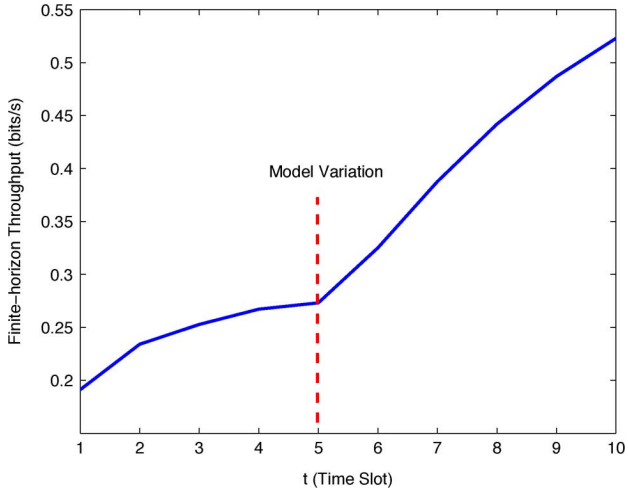

Fig. 4. Robustness of the myopic policy ($M = 1, N = 4, \epsilon = 0.0312$; for $t < 5, p_{11} = 0.6$ and $p_{01} = 0.1$; for $t \geq 5, p_{11} = 0.9$ and $p_{01} = 0.4$).

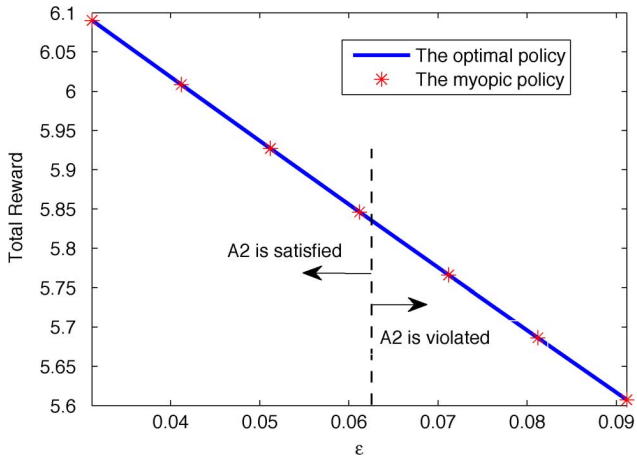

Fig. 5. Optimality of the myopic policy ($M = 1, N = 2, T = 10, p_{11} = 0.8, p_{01} = 0.2$).

Proof: See Appendix I. ∎

IV. Numerical Examples

Based on the simple semi-universal structure (see Theorem 1), the myopic policy can be implemented without knowing the channel transition probabilities except the order of p_{11} and p_{01}. As a result, the myopic policy is robust against model mismatch and automatically tracks variations in the channel model provided that the order of p_{11} and p_{01} remains unchanged. As show in Fig. 4, the transition probabilities change abruptly in the fifth slot, which corresponds to an improvement in the quality of the channels (for example, a drop in the traffic load of the primary system in the application of cognitive radio). From this figure, we can observe, from the change in the throughput increasing rate, that the myopic policy effectively tracks the model variations.

In Fig. 5, we plot both the optimal performance and the performance of the myopic policy as functions of the false alarm probability. We observe that the myopic policy can achieve the optimal performance even when Assumption A2) is violated, indicating that A2) is sufficient but not necessary for the optimality of the myopic policy.

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

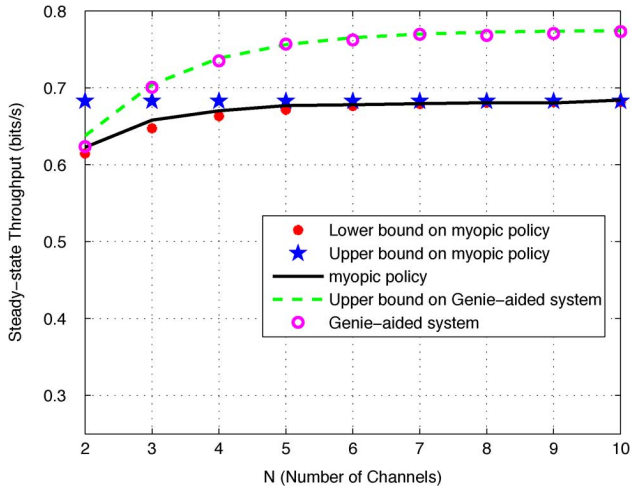

Fig. 6. Performance bounds of the myopic policy ($M = 1, p_{11} = 0.8, p_{01} = 0.2, \epsilon = 0.0312$).

To illustrate the tightness of the bounds on U and $\bar{U}$ given in Sections III-D-2) and III-E, we compare the performance of the myopic policy with the performance of the genie-aided system. From Fig. 6, we observe that the lower bound on the performance of the myopic policy quickly converges to the upper bound as N increases. We also observe from Fig. 6 that the bounds on the performance of the myopic policy are tight.

V. Discussions and Related Work

As discussed in Section II, in the context of cognitive radio for opportunistic spectrum access, the joint design of the channel state detector, the sensing policy for channel selection, and the access policy has been investigated within the framework of constrained POMDP in [8]. A separation principle has been established in [11] which reveals the optimality of the myopic approach in the design of the channel state detector and the access policy. Results obtained in this paper complement the separation principle by showing the structure and optimality of the myopic approach in designing the sensing policy. This paper also extends our earlier work [10] on myopic sensing with perfect channel state detection.

In [14], access strategies for a slotted secondary user searching for opportunities in an un-slotted primary network is considered under a continuous-time Markovian model of channel occupancy. In [15], the issue of power control is addressed under a discrete-time Markovian model of channel occupancy with perfect but delayed spectrum sensing. In [16], the authors consider the POMDP framework established in [8] under multichannel perfect sensing, where an approximation method is proposed with a closed-form bound on its performance loss with respect to the optimal policy. In [17], a heuristic sensing policy that selects the channel with the longest predicted time to be idle is proposed to reduce channel switching under general stochastic models for channel occupancy. In [18], the design of the optimal sensing and transmission strategies on a single channel is addressed under imperfect spectrum sensing, where the channel occupancy is modeled as a continuous-time semi-Markov renewal process. In [19], the authors extend the POMDP framework proposed in [8] 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 [19].

This paper also contributes to the general literature on restless multi-armed bandit problem. While an index policy was shown by Gittins in 1960s to be optimal for the classical bandit problems [20], the structure of the optimal policy for a general restless bandit process remains unknown, and the problem is shown to be PSPACE-hard [9]. The question whether an index policy is also optimal for RMAB has been pursued for decades. In 1988, Whittle proposed a Gittins-like heuristic index policy [21] for RMAB, which is asymptotically (in terms of the number of arms) optimal in certain limiting regime as shown by Weber and Weiss in 1990 [22]. Beyond this asymptotic result, relatively little is known about the structure of the optimal policies for a general restless bandit process. In fact, even the indexability of an RMAB is often difficult to establish [21], [23]. For a class of RMBP similar to that considered in this paper (the difference is that channels can be non-identical but channel state detection is perfect), the indexability of the RMAB is established and Whittle index obtained in closed-form in [24]–[26] under both discounted and average reward criteria. The structure and the optimality of Whittle index policy are also established in [25] and [26], for stochastically identical arms based on its equivalence to the myopic policy. The same class of RMBP is also considered in a parallel work in the context of multi-agent systems [27], where the indexability and the closed-form expression for Whittle index are obtained under the discounted reward criterion using a different approach. The structure and the optimality of Whittle index policy, however, were not considered in [27]. This paper differs from [24]–[27] by considering imperfect channel state detection. The indexability of the resulting RMBP and the connection between Whittle index policy and the myopic policy studied in this paper remain open in the case of imperfect sensing.

VI. Conclusion and Future Work

In this paper, we have analyzed the structure, optimality, and performance of the myopic sensing policy in multichannel opportunistic access under an independent and stochastically identical Gilbert–Elliot channel model with noisy state observations. We have established a simple and robust structure of the myopic sensing policy under certain conditions. The optimality of the myopic policy has been proved for two-channel systems and conjectured, based on numerical examples, for the general case. The performance of the myopic policy has been analyzed, which allows 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 resolving the optimality conjecture of the myopic policy and generalizing the results to stochastically non-identical channels by investigating Whittle index policy. The latter has been studied in [24]–[27] under the assumption of perfect channel state detection. It is also of interest to investigate the optimality of the myopic policy under system models beyond the Gilbert–Elliot channel. For example, dynamic multichannel access under a self-similar channel occupancy model has been formulated and studied in [28] where a universal structure of the myopic policy has been established. The optimality of the myopic policy and

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

the issue of imperfect channel state detection, however, remain open under the self-similar channel model.

APPENDIX A
PROOF OF THEOREM 1

Let $\mathbf{Q}(t) = (n_1, n_2, \ldots, n_N)(n_i \in \{1, 2, \ldots, N\}\ \forall i)$ be the queueing order of channels in slot t. We need to show that

$$\min_{1 \le i \le M} \omega_{n_i}(t) \ge \max_{M+1 \le j \le N} \omega_{n_j}(t). \tag{16}$$

We first establish the following properties of the operator $\Gamma(x)$ defined in (1).

- P1) $\Gamma(x)$ is an increasing function for $p_{11} \ge p_{01}$ and a decreasing function for $p_{11} < p_{01}$.
- P2) $\forall\ 0 \le x \le 1, p_{01} \le \Gamma(x) \le p_{11}$ for $p_{11} \ge p_{01}$ and $p_{11} \le \Gamma(x) \le p_{01}$ for $p_{11} < p_{01}$.
- P3) For $p_{11} \ge p_{01}$ and $\epsilon \le (p_{10}p_{01})/(p_{11}p_{00})$, we have $\Gamma((\epsilon\omega)/(\epsilon\omega + (1-\omega))) \le \Gamma(\omega')\ \forall p_{01} \le \omega, \omega' \le p_{11}$; for $p_{11} < p_{01}$ and $\epsilon \le (p_{00}p_{11})/(p_{01}p_{10})$, we have $\Gamma((\epsilon\omega)/(\epsilon\omega + (1-\omega))) \ge \Gamma(\omega')\ \forall p_{11} \le \omega, \omega' \le p_{01}$.

P1) and P2) follow directly from the definition of $\Gamma(x)$. To show P3) for $p_{11} \ge p_{01}$, it suffices to show $(\epsilon\omega)/(\epsilon\omega+(1-\omega)) \le p_{01}$ due to the monotonically increasing property of $\Gamma(x)$ and the bound on ω'. Noticing that $(\epsilon\omega)/(\epsilon\omega+(1-\omega))$ is an increasing function of both ω and ϵ, we arrive at P3) by using the upper bounds on ω and ϵ. Similarly, we can show P3) for $p_{11} < p_{01}$.

We now prove (16) by induction. For $t = 1$, (16) holds by the definition of $\mathbf{Q}(1)$. Assume that (16) is true for slot t. We show that it is also true for slot $t+1$.

Consider first $p_{11} \ge p_{01}$. For an $1 \le i \le M$ with $K_{n_i} = 1, \omega_{n_i}(t+1) = p_{11}$ which achieves the upper bound of the belief values (See P2). For an $1 \le j \le M$ with $K_{n_j} = 0, \omega_{n_j}(t+1)$ is upper bounded by those of unobserved channels due to P3). Among those channels over which NAKs are observed, the order of their believes remains unchanged in slot $t+1$ due to P1). Similarly, the order of the belief values of the unobserved channels also remains unchanged in slot $t+1$.

When $p_{11} < p_{01}$, for an $1 \le i \le M$ with $K_{n_i} = 1, \omega_{n_i}(t+1) = p_{11}$ which achieves the lower bound of the belief values (see P2)). For an $1 \le j \le M$ with $K_{n_j} = 0, \omega_{n_j}(t+1)$ is lower bounded by those of unobserved channels due to P3). Furthermore, the order of the belief values of the unobserved channels is reversed in slot $t+1$.

We thus proved (16) for all $t \ge 1$ under the structure of the myopic policy.

APPENDIX B
PROOF OF COROLLARY 1

$\vec{\mathbf{S}}(t)$ specifies the states of all channels as well as their order in slot t. Based on the structure of the myopic policy, $\vec{\mathbf{S}}(t)$ determines the probability distribution of $\vec{\mathbf{S}}(t+1)$, i.e., $\vec{\mathbf{S}}(t)$ is a Markov chain. Furthermore, the expected reward obtained on an idle channel is given by $1 - \epsilon$.

APPENDIX C
PROOF OF THEOREM 2

For $N = 2$, we only need to consider the nontrivial case of $M = 1$. We first consider an arbitrary finite horizon of T slots. Let $\hat{V}_t(\Omega)$ denote the expected total discounted reward obtained under the myopic policy starting from slot t, and $\hat{V}_t(\Omega; a)$ the expected total discounted reward obtained by sensing action a in slot t followed by the myopic policy in future slots. The proof is based on the following lemma which applies to a general POMDP and has been established in [10].

Lemma 3: For a T-horizon POMDP, the myopic policy is optimal if for $t = 1, \ldots, T$,

$$\hat{V}_t(\Omega) \ge \hat{V}_t(\Omega; a), \quad \forall a, \Omega. \tag{17}$$

We now prove Theorem 2. Considering all channel state realizations in slot t, we have

$$\begin{aligned}\hat{V}_t(\Omega; a) =& (1-\epsilon)\omega_a + \beta \sum_{s_1, s_2 \in \{0,1\}} \Pr[\mathbf{S}(t) = [s_1, s_2] | \Omega(t)] \\ & \cdot \hat{V}_{t+1}(\mathcal{T}(\Omega(t)|a, s_a) | \mathbf{S}(t) = [s_1, s_2])\end{aligned} \tag{18}$$

where $\hat{V}_{t+1}(\mathcal{T}(\Omega(t)|a, s_a)|\mathbf{S}(t) = [s_1, s_2])$ is the conditional expected total discounted reward obtained starting from slot $t+1$ given that the system state in slot t is $[s_1, s_2]$. Next, we establish two lemmas regarding the conditional value function of the myopic policy.

Lemma 4: Under Assumptions A1) and A2), the expected total discounted reward starting from slot t under the myopic policy is determined by the action $a(t-1)$ and the system state $\mathbf{S}(t-1)$ in slot $t-1$, hence independent of the belief vector $\Omega(t)$ at the beginning of slot t, i.e.,

$$\begin{aligned}&\hat{V}_t(\mathcal{T}(\Omega(t-1)|a, s_a)|\mathbf{S}(t-1) = [s_1, s_2]) \\ &\qquad = \hat{V}_t(\mathcal{T}(\Omega'(t-1)|a, s_a)|\mathbf{S}(t-1) = [s_1, s_2]).\end{aligned}$$

Adopting the simplified notation of $\hat{V}_t(a(t-1)|\mathbf{S}(t-1) = [s_1, s_2])$, We further have

$$\begin{aligned}&\hat{V}_t(a(t-1) = 1|\mathbf{S}(t-1) = [s_1, s_2]) \\ &\qquad = \hat{V}_t(a(t-1) = 2|\mathbf{S}(t-1) = [s_2, s_1]).\end{aligned} \tag{19}$$

Proof: Given $a(t-1)$ and $\mathbf{S}(t-1)$, the myopic actions in slots t to T, governed by the structure given in Theorem 1, are fixed for each sample path of system state and observation, independent of $\Omega(t)$. As a consequence, the total discounted reward obtained in slots t to T for each sample path is independent of $\Omega(t)$, so is the expected total discounted reward. (19) follows from the statistically identical assumption of channels. ■

Lemma 5: Under Assumptions A1) and A2), we have, $\forall t, a$,

$$\begin{aligned}&|\hat{V}_t(a(t-1) = a|\mathbf{S}(t-1) = [1,0]) - \hat{V}_t(a(t-1) \\ &\qquad = a|\mathbf{S}(t-1) = [0,1])| \le (1-\epsilon).\end{aligned} \tag{20}$$

Proof: Based on (19), it suffices to consider $a(t-1) = 1$. We prove for $p_{11} < p_{01}$ by reverse induction. The proof for $p_{11} > p_{01}$ is similar. The inequality in (20) holds for $t = T$ since $(1-\epsilon)$ is the maximum expected reward that can be obtained in one slot. Assume that the inequality holds for $t+1$. We show that it holds for t. Consider first $\hat{V}_t(a(t-1) = 1|\mathbf{S}(t-1) = [1,0])$. With probability $1-\epsilon$, the user successfully identifies that channel 1 is in the good state in slot $t-1$ and receives an acknowledgement at the end of slot $t-1$. According to the structure of the myopic policy, the user switches channel in slot t, i.e., $a(t) = 2$. The expected immediately reward in slot t is thus $p_{01}(1-\epsilon)$ since the state of channel 2 in slot $t-1$ is

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

0. We thus arrive at the first term of (21), where $\hat{V}_t(a(t-1)=1|\mathbf{S}(t-1)=[1,0])$ is given by the summation of $p_{01}(1-\epsilon)$ and the discounted future reward starting from slot $t+1$ conditioned on all four possible system states in slot t. With probability ϵ, a false alarm occurs in slot $t-1$, resulting in a NAK. The user thus stays in channel 1 in slot t: $a(t)=1$. We thus arrive at the second term of (21). Similarly, we obtain $\hat{V}_t(a(t-1)=1|\mathbf{S}(t-1)=[0,1])$ as given in (22), which follows from the fact that a NAK occurs in slot $t-1$ due to the given bad state of the chosen channel 1.

$$\begin{aligned}
&\hat{V}_t(1|[1,0]) \\
&\quad=(1-\epsilon)\{p_{01}(1-\epsilon)+\beta(p_{10}p_{00}\hat{V}_{t+1}(2|[0,0]) \\
&\qquad+p_{11}p_{01}\cdot\hat{V}_{t+1}(2|[1,1])+p_{11}p_{00}\hat{V}_{t+1}(2|[1,0]) \\
&\qquad+p_{10}p_{01}\hat{V}_{t+1}(2|[0,1]))\} \\
&\qquad+\epsilon\{p_{11}(1-\epsilon)+\beta(p_{10}p_{00}\hat{V}_{t+1}(1|[0,0]) \\
&\qquad+p_{11}p_{01}\cdot\hat{V}_{t+1}(1|[1,1]) \\
&\qquad+p_{11}p_{00}\hat{V}_{t+1}(1|[1,0])+p_{10}p_{01}\hat{V}_{t+1}(1|[0,1]))\} \quad (21) \\
&\hat{V}_t(1|[0,1]) \\
&\quad=p_{01}(1-\epsilon)+\beta(p_{00}p_{10}\hat{V}_{t+1}(1|[0,0]) \\
&\qquad+p_{01}p_{11}\hat{V}_{t+1}(1|[1,1]) \\
&\qquad+p_{11}p_{00}\hat{V}_{t+1}(1|[0,1])+p_{10}p_{01}\hat{V}_{t+1}(1|[1,0])). \quad (22)
\end{aligned}$$

Applying (19) and the upper bound on ϵ, we have

$$\begin{aligned}
&|\hat{V}_t(1|[0,1])-\hat{V}_t(1|[1,0])| \\
&\quad\le(1-\epsilon)p_{01}-(1-\epsilon)(\epsilon p_{11}+(1-\epsilon)p_{01}) \\
&\quad+\epsilon\beta|\hat{V}_{t+1}(1|[1,0])-\hat{V}_{t+1}(1|[0,1])|(p_{10}p_{01}-p_{11}p_{00}) \\
&\quad\le 2(1-\epsilon)\epsilon(p_{01}-p_{11}) \\
&\quad\le 2(1-\epsilon)\frac{p_{00}p_{11}}{p_{01}p_{10}}(p_{01}-p_{11}) \\
&\quad<(1-\epsilon),
\end{aligned}$$

where the last inequality follows from $(p_{01}-p_{11})(p_{11})/(p_{01})\le(1/4)$ and $p_{00}/p_{10}<1$. ■

We now show that (17) in Lemma 3 holds. Consider $\Omega(t)=[\omega_1(t),\omega_2(t)]$ with $\omega_1(t)>\omega_2(t)$, i.e., the myopic action in slot t is $a(t)=1$. Applying (19) and Lemma 5 to (18), we have

$$\begin{aligned}
&\hat{V}_t(\Omega;a=1)-\hat{V}_t(\Omega;a=2) \\
&\quad=(\omega_1-\omega_2)(1-\epsilon+\beta(\hat{V}_{t+1}(1|[1,0]) \\
&\qquad-\hat{V}_{t+1}(1|[0,1]))\ge 0.
\end{aligned}$$

We thus proved that the myopic policy is optimal for an arbitrary finite horizon of T slots under Assumptions A1) and A2).

By contradiction, it is easy to prove that the myopic policy maximizes both the expected total discounted reward and the expected average reward over an infinite horizon under Assumptions A1) and A2). Without Assumption A1), the structure of the myopic policy given in Theorem 1 still holds starting from the second slot. The myopic policy is thus optimal starting from the second slot even when Assumption A1) does not hold. Based on the ergodicity of the Markov reward process under the myopic policy (see Corollary 1), the expected average reward achieved by the myopic policy does not depend on the initial belief. The myopic policy thus achieves the maximum expected average reward starting from the second slot regardless of the belief values in the second slot. Combined with the fact that the expected immediate reward in the first slot does not contribute to the expected average reward over an infinite horizon, we conclude that the myopic policy maximize the expected average reward over an infinite horizon even when Assumption A1) dose not hold.

Appendix D
Proof of Lemma 1

Consider first $p_{11}\ge p_{01}$. Let k_T denote the number of events that a channel is moved to the end of the queue during a finite horizon of T slots.[5] Since such an event represents a loss of a unit of reward, the throughput U_T during the finite horizon is given below.

$$U_T=\frac{MT-k_T}{T}. \quad (23)$$

Let j_T denote the number of TPs during the finite horizon. We have $j_T=M+k_T$ since the event that a channel is moved to the end of the queue initializes a new TP. It is easy to see that $MT\le\sum_{i=1}^{M+k_T}L_k\le MT+\sum_{k+1}^{M+k_T}L_k$. Note that the length of a TP is finite almost surely. We thus have

$$\lim_{T\to\infty}\frac{MT}{k_T}=\lim_{T\to\infty}\frac{\sum_{i=1}^{M+k_T}L_k}{M+k_T}=\bar{L}. \quad \text{(a.s.)} \quad (24)$$

From (23) and (24), we have

$$U=\lim_{T\to\infty}U_T=M\lim_{T\to\infty}\left(1-\frac{k}{MT}\right)=M(1-1/\bar{L}). \quad \text{(a.s.)} \quad (25)$$

The case for $p_{11}<p_{01}$ can be similarly obtained by observing that the event that a channel is moved to the end of the queue represents a gain of one unit reward.

Appendix E
Proof of Theorem 3

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

Let ω_k denote the belief value of the chosen channel in the first slot of the kth TP. The length $L_k(\omega_k)$ of this TP has the following distribution:

$$\begin{aligned}
&\Pr[L_k(\omega_k)=l] \\
&\quad=\begin{cases}1-\omega_k(1-\epsilon), & l=1\\ \omega_k(1-\epsilon)^{k-1}p_{11}^{l-2}(1-p_{11}(1-\epsilon)), & l>1.\end{cases} \quad (26)
\end{aligned}$$

It is easy to see that if $\omega'\ge\omega$, then $L_k(\omega')$ stochastically dominates $L_k(\omega)$.

Note that the j-step belief update $\Gamma^j(\omega)$ when unobserved is given by

$$\Gamma^j(\omega)=\omega_o-(\omega_o-\omega)(p_{11}-p_{01})^j. \quad (27)$$

It is easy to obtain the following property on the convergence of $\Gamma^j(\omega)$: for $p_{11}\ge p_{01}$ and $\forall\omega\in[0,1],\Gamma^j(\omega)$ monotonically converges to ω_o as $j\to\infty$, where ω_o is the stationary distribution of the Gilbert–Elliot channel given in (2); for $p_{11}<p_{01}$ and $\forall\omega\in[0,1],\Gamma^{2j}(\omega)$ and $\Gamma^{2j+1}(\omega)$ converge, from opposite directions, to ω_o as $j\to\infty$.

[5]If multiple (say K) channels are simultaneously moved to the end of the queue, it is counted as K events.

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

Based on the structure of the myopic policy, we have $\omega_k = \Gamma^{(J_k+1)}(\epsilon x/(\epsilon x+1-x))$, where $J_k = \sum_{i=1}^{N-1} L_{k-i}$ denotes the number of consecutive slots in which the chosen channel has been unobserved since the last visit, and x denotes the belief value of the chosen channel in the slot of the last visit to this channel. From Assumption A2), $\Gamma(\epsilon x/(\epsilon x+1-x)) \leq \Gamma(p_{01}) \leq \omega_o$. Based on the convergence property of $\Gamma^j(\omega)$, we have $\omega_k \leq \omega_o \ \forall k$. $L_k(\omega_o)$ thus stochastically dominates $L_k(\omega_k)$, and the expectation of the former, $\overline{L_k(\omega_o)} = 1 + (\omega_o(1-\epsilon))/(1-p_{11}(1-\epsilon))$, leads to the upper bound of U given in (9).

Next, we prove the lower bound of U by constructing a hypothetical system where the initial belief value of the chosen channel in a TP is a lower bound of that in the real system. The average TP length in this hypothetical system is thus smaller than that in the real system, leading to a lower bound on U based on (7). Specifically, since $\omega_k = \Gamma^{(J_k+1)}(\epsilon x/(\epsilon x+1-x))$ and $J_k = \sum_{i=1}^{N-1} L_{k-i} \geq N + L_{k-1} - 2$, we have $\omega_k \geq \Gamma^{N+L_{k-1}-1}(\epsilon x/(\epsilon x+1-x)) \geq \Gamma^{N+L_{k-1}-1}(\epsilon p_{01}/(\epsilon p_{01}+1-p_{01}))$ based on the convergence property of $\Gamma^j(\omega)$. We thus construct a hypothetical system given by a first-order Markov chain $\{L'_k\}_{k=1}^{\infty}$ with the transition matrix $\mathbf{R} = \{r_{i,j}\}$, shown in the equation at the bottom of the page.

Lemma 6: The stationary distribution of the first order Markov chain $\{L'_k\}_{k=1}^{\infty}$ is stochastically dominated by the stationary distribution of $\{L_k\}_{k=1}^{\infty}$.

Proof: Let ω'_k denote the expected probability that the chosen channel is in state 1 in the first slot of the kth transmission period of $\{L'_k\}_{k=1}^{\infty}$. Assume in the kth transmission period, the distributions of L'_k and L_k both equal to the same distribution $\overrightarrow{\lambda}$, which may or may not be the stationary distribution of $\{L_k\}_{k=1}^{\infty}$. Next we show $\omega_{k+n} \geq \omega'_{k+n}$ for any $n \geq 1$ by induction.

When $n = 1$, we have

$$\begin{aligned}
\omega_{k+1} &= \sum_{l=1}^{\infty} \mathbb{E}_{L_{k-N+2},\cdots,L_{k-1}} \\
&\quad \times [\Gamma^{1+\sum_{i=k-N+2}^{k} L_i}\left(\frac{\epsilon x}{\epsilon x+1-x}\right) | L_k = l] \\
&\quad \times \Pr(L_k = l) \\
&\geq \sum_{l=1}^{\infty} \mathbb{E}_{L_{k-N+2},\cdots,L_{k-1}} \\
&\quad \times [\Gamma^{N-1+L_k}\left(\frac{\epsilon p_{01}}{\epsilon p_{01}+1-p_{01}}\right) | L_k = l] \\
&\quad \times \Pr(L_k = l) \\
&= \sum_{l=1}^{\infty} \Gamma^{N-1+l}\left(\frac{\epsilon p_{01}}{\epsilon p_{01}+1-p_{01}}\right) \lambda_l \\
&= \omega'_{k+1}. \qquad (28)
\end{aligned}$$

Assume $\omega_{k+n} \geq \omega'_{k+n}$, then

$$\begin{aligned}
&\omega_{k+n+1} \\
&= \sum_{l=1}^{\infty} \mathbb{E}_{L_{k+n-N+2},\cdots,L_{k+n-1}} \\
&\quad \times \left[\Gamma^{1+\sum_{i=k+n-N+2}^{k+n} L_i}\left(\frac{\epsilon x}{\epsilon x+1-x}\right) | L_{k+n} = l\right] \\
&\quad \times \Pr(L_{k+n} = l) \\
&\geq \sum_{l=1}^{\infty} \mathbb{E}_{L_{k+n-N+2},\cdots,L_{k+n-1}} \\
&\quad \times \left[\Gamma^{N-1+L_{k+n}}\left(\frac{\epsilon p_{01}}{\epsilon p_{01}+1-p_{01}}\right) | L_{k+n} = l\right] \\
&\quad \times \Pr(L_{k+n} = l) \\
&= \sum_{l=1}^{\infty} \Gamma^{N-1+l}\left(\frac{\epsilon p_{01}}{\epsilon p_{01}+1-p_{01}}\right) \\
&\quad \times \Pr(L_{k+n} = l). \qquad (29)
\end{aligned}$$

Since $\omega_{k+n} \geq \omega'_{k+n}$, by (26), we have

$$\begin{aligned}
&\Pr(L_{k+n} = l) \leq \Pr(L'_{k+n} = l), \quad \text{if } l = 1; \\
&\Pr(L_{k+n} = l) \geq \Pr(L'_{k+n} = l), \quad \text{if } l > 1. \qquad (30)
\end{aligned}$$

Since the smallest number in the series $\Gamma^{N-1+l}((\epsilon p_{01})/(\epsilon p_{01}+1-p_{01}))$ is the first one, by (30) and the fact that $\sum_{l=1}^{\infty} \Pr(L_{k+n} = l) = \sum_{l=1}^{\infty} \Pr(L'_{k+n} = l) = 1$, we have

$$\begin{aligned}
&\sum_{l=1}^{\infty} \Gamma^{N-1+l}\left(\frac{\epsilon p_{01}}{\epsilon p_{01}+1-p_{01}}\right) \Pr(L_{k+n} = l) \\
&\quad \geq \sum_{l=1}^{\infty} \Gamma^{N-1+l}\left(\frac{\epsilon p_{01}}{\epsilon p_{01}+1-p_{01}}\right) \Pr(L'_{k+n} = l) = \omega'_{k+n+1}. \qquad (31)
\end{aligned}$$

Combine (29) and (31), we have $\omega_{k+n+1} \geq \omega'_{k+n+1}$.

By the above induction, we have $\omega_{k+n} \geq \omega'_{k+n}$ for any $n \geq 1$. So the stationary distribution of the first order Markov chain $\{L'_k\}_{k=1}^{\infty}$ is dominated by the stationary distribution of $\{L_k\}_{k=1}^{\infty}$. ■

Let $\bar{L}'$ denote the average length of a transmission period of L'_k. Based on (7) and Lemma 6, $\bar{L}'$ leads to a lower bound on U. Last, we obtain closed-form $\bar{L}'$ by solving the stationary distribution of the first-order Markov chain $\{L'_k\}_{k=1}^{\infty}$.

Recall that $\mathbf{R} = \{r_{i,j}\}$ is the transition matrix of $\{L'_k\}_{k=1}^{\infty}$, where $r_{i,j}$ is given in the equation shown at the bottom of the page. Let $\mathbf{R}(:,k)$ denote the kth column of $\mathbf{R}$. We have

$$\begin{aligned}
\mathbf{1} - \mathbf{R}(:,1) &= \frac{\mathbf{R}(:,2)}{1-p_{11}(1-\epsilon)}, \mathbf{R}(:,k) \\
&= \mathbf{R}(:,2)(p_{11}(1-\epsilon))^{k-2}, (k \geq 2) \qquad (32)
\end{aligned}$$

$$r_{i,j} = \begin{cases} 1 - \Gamma^{N+i-1}\left(\frac{\epsilon p_{01}}{\epsilon p_{01}+1-p_{01}}\right), & i \geq 1, j = 1 \\ \Gamma^{N+i-1}\left(\frac{\epsilon p_{01}}{\epsilon p_{01}+1-p_{01}}\right)(1-\epsilon)^{j-1}(p_{11})^{j-2}(1-p_{11}(1-\epsilon)), & i \geq 1, j \geq 2. \end{cases}$$

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

where $\mathbf{1}$ is the unit column vector $[1,1,\ldots]^t$. By the definition of stationary distribution, we have, for $k=1,2,\ldots,$

$$[\lambda_1,\lambda_2,\ldots]\mathbf{R}(:,k)=\lambda_k \tag{33}$$

which, combined with (32), leads to

$$\begin{aligned}\lambda_1 &= 1-\frac{\lambda_2}{(1-p_{11}(1-\epsilon))},\\ \lambda_k &= \lambda_2(p_{11}(1-\epsilon))^{k-2}. \quad (k\geq 2).\end{aligned} \tag{34}$$

Substituting (34) into (33) for $k=2$ and solving for λ_2, we have $\lambda_2=f(c_1)(1-\epsilon)(1-p_{11}(1-\epsilon))$, where $f(c_1)$ is given in (9). From (34), we then have the stationary distribution as follows:

$$\lambda_k=\begin{cases}1-f(c_1)(1-\epsilon), & k=1\\ f(c_1)(1-\epsilon)(p_{11}(1-\epsilon))^{k-2}(1-p_{11}(1-\epsilon)), & k>1\end{cases} \tag{35}$$

which leads to $\overline{L'}=\sum_{k=1}^{\infty}k\lambda_k=1+(f(c_1)(1-\epsilon))/(1-p_{11}(1-\epsilon))$.

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

Let ω_k denote the belief value of the chosen channel in the first slot of the kth TP. Define the operator $c(\cdot)$ as $c(x)=(\epsilon x)/(\epsilon x+1-x)$. We have

$$\begin{aligned}&\Pr[L_k(\omega_k)=l]\\ &\quad=\begin{cases}\omega_k(1-\epsilon), & l=1\\ (1-\omega_k(1-\epsilon))\prod_{i=1}^{l-2}(1-(\Gamma\circ c)^i(\omega_k)(1-\epsilon))\\ \qquad\cdot(\Gamma\circ c)^{l-1}(\omega_k)(1-\epsilon), & l>1.\end{cases}\end{aligned}$$

Consider first the upper bound. We construct the following hypothetical system where the stationary distribution of a TP is stochastically dominated by the one in the real system. The average TP length in this hypothetical system is thus smaller that in the real system, leading to a upper bound on U based on (7). Specifically, the distribution of a TP in the hypothetical system has the following form:

$$\Pr[L'_k(\omega_k)=l]=\begin{cases}\frac{\Gamma(p_{11})}{p_{01}}\omega_k(1-\epsilon)+1-\frac{\Gamma(p_{11})}{p_{01}}, & l=1\\ (1-\omega_k(1-\epsilon))(1-p_{01}(1-\epsilon))^{k-2}\\ \qquad\cdot\Gamma(p_{11})(1-\epsilon), & l>1.\end{cases} \tag{36}$$

We first show that $L'_k(\omega_k)$ is stochastically dominated by $L_k(\omega_k)$. Note that $\Pr[L'_k(\omega_k)=l]\geq 0$ for all $l\in\mathcal{Z}^+$ and $\sum_{l=1}^{\infty}\Pr[L'_k(\omega_k)=l]=1$. The distribution of $L'_k(\omega_k)$ given in (36) is thus well-defined. Since $\Gamma(p_{11})\leq\Gamma\circ c(\omega)\leq p_{01}$ for any $p_{11}\leq\omega\leq p_{01}$, we have $\Pr[L'_k(\omega_k)=l]\leq\Pr[L_k(\omega_k)=l]$ for all $l\geq 2$. $L'_k(\omega_k)$ is thus stochastically dominated by $L_k(\omega_k)$.

It is easy to see that $L'_k(\omega')$ is stochastically dominated by $L'_k(\omega)$ if $\omega'\geq\omega$. $L'_k(\omega')$ is thus stochastically dominated by $L_k(\omega)$ if $\omega'\geq\omega$. Based on the structure of the myopic policy, it is clear that when L_{k-1} is odd, in the kth TP, the user will switch to the channel visited in the $(k-2)$th TP. As a consequence, the initial belief ω_k of the kth TP is given by $\omega_k=\Gamma^{(L_{k-1}+1)}(1)$. When L_{k-1} is even, we can show that $\omega_k\leq\Gamma^{(L_{k-1}+4)}(1)$. This is because that for L_{k-1} even, the user cannot switch to a channel visited $L_{k-1}+2$ slots ago, and $\Gamma^j(1)$ decreases with j for even j's and $\Gamma^j(1)\geq\Gamma^i(1)$ for any even j and odd i (based on the convergence property of $\Gamma^j(\omega)$). We thus construct a hypothetical system given by the first-order Markov chain $\{L'_k\}_{k=1}^{\infty}$ with the transition probabilities shown in the first equation at the bottom of the page. Similarly to Lemma 6, it can be shown that the stationary distribution of $\{L'_k\}_{k=1}^{\infty}$ is stochastically dominated by that of $\{L_k\}_{k=1}^{\infty}$. Furthermore the stationary distribution of $\{L'_k\}_{k=1}^{\infty}$ can be obtained in closed form by using an approach similar to that in Case 1, leading to the upper bound on U given in (10).

We now prove the lower bound. Consider the hypothetical system with the distribution of a TP as given below.

$$\Pr[L'_k(\omega_k)=l]=\begin{cases}\frac{p_{01}}{\Gamma(p_{11})}\omega_k(1-\epsilon)+1-\frac{p_{01}}{\Gamma(p_{11})}, & l=1\\ (1-\omega_k(1-\epsilon))(1-\Gamma(p_{11})(1-\epsilon))^{k-2}\\ \qquad\cdot p_{01}(1-\epsilon), & l>1.\end{cases} \tag{37}$$

Similarly, $L'_k(\omega_k)$ is well-defined and stochastically dominates $L_k(\omega_k)$. It is easy to see that $L'_k(\omega')$ stochastically dominates $L'_k(\omega)$ if $\omega'\leq\omega$. $L'_k(\omega')$ thus stochastically dominates $L_k(\omega)$ if $\omega'\leq\omega$.

Based on the structure of the myopic policy, $\omega_k=p_{11}^{(L_{k-1}+1)}$ when L_{k-1} is odd. When L_{k-1} is even, to find a lower bound on ω_k, we need to find the smallest odd j such that the last visit to the channel chosen in the kth TP is j slots ago. From the structure of the myopic policy, the smallest feasible odd j is $L_{k-1}+2N-3$, which corresponds to the scenario where all N channels are visited in turn from the $(k-N+1)$th TP to the kth TP with $L_{k-N+1}=L_{k-N+2}=\cdots=L_{k-2}=2$. We thus have $\omega_k\geq p_{11}^{(L_{k-1}+2N-3)}$. We then construct a hypothetical system given by the first-order Markov chain $\{L'_k\}_{k=1}^{\infty}$ with the transition probabilities shown in the second equation at the bottom of the page.

$$r_{i,j}=\begin{cases}\frac{\Gamma(p_{11})}{p_{01}}\Gamma^{i+1}(1)(1-\epsilon)+1-\frac{\Gamma(p_{11})}{p_{01}}, & \text{if } i \text{ is odd}, j=1\\ (1-\Gamma^{i+1}(1)(1-\epsilon))(1-p_{01}(1-\epsilon))^{j-2}\Gamma(p_{11})(1-\epsilon), & \text{if } i \text{ is odd}, j\geq 2\\ \frac{\Gamma(p_{11})}{p_{01}}\Gamma^{i+4}(1)(1-\epsilon)+1-\frac{\Gamma(p_{11})}{p_{01}}, & \text{if } i \text{ is even}, j=1\\ (1-\Gamma^{i+4}(1)(1-\epsilon))(1-p_{01}(1-\epsilon))^{j-2}\Gamma(p_{11})(1-\epsilon), & \text{if } i \text{ is even}, j\geq 2.\end{cases}$$

$$r_{i,j}=\begin{cases}\frac{p_{01}}{\Gamma(p_{11})}\Gamma^{i+1}(1)(1-\epsilon)+1-\frac{p_{01}}{\Gamma(p_{11})}, & \text{if } i \text{ is odd}, j=1\\ (1-\Gamma^{i+1}(1)(1-\epsilon))(1-p_{01}(1-\epsilon))^{j-2}\Gamma(p_{11})(1-\epsilon), & \text{if } i \text{ is odd}, j\geq 2\\ \frac{\Gamma(p_{11})}{p_{01}}\Gamma^{i+2N-3}(1)(1-\epsilon)+1-\frac{\Gamma(p_{11})}{p_{01}}, & \text{if } i \text{ is even}, j=1\\ (1-\Gamma^{i+2N-3}(1)(1-\epsilon))(1-p_{01}(1-\epsilon))^{j-2}\Gamma(p_{11})(1-\epsilon), & \text{if } i \text{ is even}, j\geq 2.\end{cases}$$

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

The stationary distribution of this hypothetical system leads to the lower bound on U given in (10).

APPENDIX F
PROOF OF THEOREM 4

We first prove that $U \geq M\omega_o(1-\epsilon)$. Note that $M\omega_o(1-\epsilon)$ is the steady-state throughput achieved by the random sensing policy that chooses M out of N channels with uniform probability (i.e., choose any set of M channels with probability $1/\binom{N}{M}$). Since the expected immediate reward under the random sensing policy in each slot is given by the expected sum of M randomly chosen belief values under any given policy (including the myopic policy) and the expected immediate reward under the myopic policy in each slot is given by the expected sum of the first M largest belief values, the throughput under the myopic policy is thus lower bounded by that under the random sensing policy. To complete the proof, we consider the following two cases.

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

Consider first the upper bound. Similarly to single-channel sensing, the belief value ω_k of the chosen channel in the first slot of the kth TP is upper bounded by ω_o. $L_k(\omega_o)$ thus stochastically dominates $L_k(\omega_k)$, and the expectation of the former leads to the upper bound on U given in (11).

We now consider the lower bound. Recall that $\omega_k = \Gamma^{(J_k+1)}(\epsilon x/(\epsilon x + 1 - x))$, where J_k denotes the number of consecutive slots in which the chosen channel has been unobserved since the last visit, and x denotes the belief value of the chosen channel in the slot of the last visit to this channel. Based on the structure of the myopic policy, the channel has the last priority when the user leaves it. It will take at least $\lfloor (N-M)/M \rfloor$ slots before the user returns to the same channel, i.e., $J_k \geq \lfloor N/M \rfloor - 1$. Based on the convergence property of $\Gamma^j(\omega)$, we have $\omega_k = \Gamma^{J_k+1}(\epsilon x/(\epsilon x + 1 - x)) \geq \Gamma^{\lfloor N/M \rfloor}(\epsilon x/(\epsilon x + 1 - x)) \geq \Gamma^{\lfloor (N/M) \rfloor}((\epsilon p_{01})/(\epsilon p_{01} + 1 - p_{01}))$. Thus, $L_k(\Gamma^{\lfloor (N/M) \rfloor}(\epsilon p_{01}/(\epsilon p_{01} + 1 - p_{01})))$ is stochastically dominated by $L_k(\omega_k)$, and the expectation of the former leads to the first argument in the maximization operator as given in the left-hand side of (11).

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

Consider first the upper bound. Let ω_k denote the belief value of the chosen channel in the first slot of the $k-$th TP. Based on the structure of the myopic policy, we have $\omega_k = \Gamma^{J_k+1}(1)$, where J_k denotes the number of consecutive slots in which the chosen channel has been unobserved since the last visit. From the convergence property of $\Gamma^j(\omega)$, we have $\omega_k = \Gamma^{J_k+1}(1) \leq \Gamma^2(1)$. Combined with the hypothetical system given in (36), $L'_k(\Gamma^2(1))$ is stochastically dominated by $L_k(\omega_k)$, and the expectation of the former leads to the upper bound on U given in (12).

We now consider the lower bound. Recall that $\omega_k = \Gamma^{J_k+1}(1)$. If J_k is odd, then $\Gamma^{J_k+1}(1) \geq \Gamma^{2\lfloor (N/M) \rfloor - 1}(1)$ since $2\lfloor (N/M) \rfloor - 1$ is an odd number (based on the convergence property of $\Gamma^j(\omega)$). If J_k is even, i.e., the user has stayed even slots before it returns this channel, then J_k is at least $2\lfloor (N-M)/M \rfloor$. We have $\omega_k = \Gamma^{J_k+1}(1) \geq \Gamma^{2\lfloor (N/M) \rfloor - 1}(1)$. Combined with the hypothetical system given in (37), $L'_k(\Gamma^{2\lfloor (N/M) \rfloor - 1}(1))$ stochastically dominates $L_k(\omega_k)$, and the expectation of the former leads to the first argument in the maximization operator as given in the left-hand side of (12).

APPENDIX G
PROOF OF COROLLARY 2

From the closed-form expressions of the lower bounds on U given in Theorem 3 and Theorem 4, it is easy to see that the lower bound is monotonically increasing with N. Let $x = |p_{11} - p_{01}|$. For $p_{11} > p_{01}$, after some simplifications, the lower bound (for large N) has the form $a + b/(x^{\lfloor (N/M) \rfloor} + c)$, where $a, b, c (c \neq 0)$ are constants. The upper bound is $a + b/c$. We have $(|a + b/(x^{\lfloor (N/M) \rfloor} + c) - a - b/c|)/(x^{(N/M)}) \to O(b/c^2)$ as $N \to \infty$. Thus, the lower bound converges to the upper bound with geometric rate $x^{(1/M)}$.

For $p_{11} < p_{01}$, the lower bound (for large N) has the form $d + e/(x^{2\lfloor (N/M) \rfloor - 1} + f)$, where $d, e, f (f \neq 0)$ are constants. It converges to $d + e/f$ as $N \to \infty$. We have $(|d + e/(x^{2\lfloor (N/M) \rfloor - 1} + f) - d - e/f|)/(x^{(2N)/M}) \to O(e/(xf^2))$ as $N \to \infty$. Thus, the lower bound converges with geometric rate $x^{(2/M)}$.

APPENDIX H
PROOF OF LEMMA 2

Similar to Corollary 1, the reward process under the myopic (optimal) policy of the genie-aided system is ergodic under Assumption A2). The maximum steady-state throughput $\bar{U}$ is thus well defined and independent of the initial belief vector. By noticing that $N\omega_o(1-\epsilon)$ is the throughput achieved when all N channels are sensed and subsequently accessed based on the sensing outcomes in each slot, we have that $\bar{U} \leq N\omega_o(1-\epsilon)$. To complete the proof, we consider the following two cases.

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

Based on the ergodicity of the reward process in the genie-aided system, the initial belief vector does not affect the optimal performance. Without loss of generality, assume the state of each channel starts from the stationary distribution ω_o. As a consequence, the number k of channels over which ACKs are observed falls into the binomial distribution $B(k, N, \omega_o(1-\epsilon))$ in every slot. Since the channels over which ACKs is observed will have the largest belief value p_{11} and other channels' belief values will be upper bounded by $\Gamma(\epsilon p_{11}/(\epsilon p_{11} + 1 - p_{11}))$ in the next slot, the expected reward obtained under the myopic policy will be upper bounded by the first argument in the minimization operator as given in the right-hand side of (13).

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

Similarly, we assume the state of each channel starts from the stationary distribution ω_o without loss of generality. The number k of channels over which ACKs are observed falls into the binomial distribution $B(k, N, \omega_o(1-\epsilon))$ in every slot. Since the channels over which ACKs are observed will have the smallest belief value p_{11} and other channels' belief values will be upper bounded by $\Gamma(\epsilon p_{11}/(\epsilon p_{11} + 1 - p_{11}))$ in the next slot, the expected reward obtained under the myopic policy will be upper bounded by the first argument in the minimization operator as given in the right-hand side of (14).

APPENDIX I
PROOF OF THEOREM 5

Consider first $p_{11} < p_{01}$. Based on Lemma 2 and Theorem 4, we have

$$\eta = \frac{U}{U^*} \geq \frac{U}{\bar{U}}$$

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.

$$\geq \frac{M\omega_o(1-\epsilon)}{\min\left\{M\Gamma\left(\frac{\epsilon p_{11}}{\epsilon p_{11}+1-p_{11}}\right)(1-\epsilon), N\omega_o(1-\epsilon)\right\}}$$
$$\geq \max\left\{\frac{M}{N}, \frac{1}{1+p_{01}-p_{11}}\right\} \geq \max\left\{\frac{M}{N}, \frac{1}{2}\right\}.$$

For $p_{11} > p_{01}$, based on Lemma 2 and Theorem 4, we have

$$\eta = \frac{U}{U^*} \geq \frac{U}{\bar{U}} \geq \frac{M\omega_o(1-\epsilon)}{N\omega_o(1-\epsilon)} = \frac{M}{N}.$$

For the trivial case of $p_{11} = p_{01}$, we note that the lower bound on U given in Theorem 4 agrees with the upper bound on $\bar{U}$ given in Lemma 2.

REFERENCES

[1] E. N. Gilbert, "Capacity of burst-noise channels," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–1265, Sep. 1960.

[2] M. Zorzi, R. Rao, and L. Milstein, "Error statistics in data transmission over fading channels," *IEEE Trans. Commun.*, vol. 46, pp. 1468–1477, Nov. 1998.

[3] L. A. Johnston and V. Krishnamurthy, "Opportunistic file transfer over a fading channel: A POMDP search theory formulation with optimal threshold policies," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, pp. 394–405, Feb. 2006.

[4] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, pp. 79–89, May 2007.

[5] S. D. Jones, N. Merheb, and I.-J. Wang, "An experiment for sensing based opportunistic spectrum access in CSMA/CA networks," in *Proc. 1st IEEE Int. Symp. New Frontiers Dynamic Spectrum Access Networks*, Nov. 2005, pp. 593–596.

[6] S. Geirhofer, L. Tong, and B. Sadler, "Dynamic spectrum access in the time domain: Modeling and exploiting white space," *IEEE Commun. Mag.*, vol. 45, no. 5, pp. 66–72, May 2007.

[7] D. Dalta, "Spectrum surveying for dynamic spectrum access networks," M.S. thesis, Univ. of Kansas, Lawrence, KS, Jan. 2007.

[8] 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. 2007.

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

[10] 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. 12, pp. 5431–5440, Dec. 2008.

[11] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2053–2071, May 2008.

[12] R. Smallwood and E. Sondik, "The optimal control of partially ovservable Markov processes over a finite horizon," *Oper. Res.*, pp. 1071–1088, 1971.

[13] S. H. Ahmad, M. Liu, T. Javadi, 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.

[14] Q. C. Zhao, S. Geirhofer, L. Tong, and B. M. Sadler, "Opportunistic spectrum access via periodic channel sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 785–796, Feb. 2008.

[15] L. Gao and S. Cui, "Power and rate control in cognitive radio networks via dynamic programming," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 4819–4827, Nov. 2009.

[16] S. Filippi, O. Cappé, F. Clérot, and E. Moulines, "A near optimal policy for channel allocation in cognitive radio," *Recent Adv. Reinforce. Learn.*, vol. 5323, pp. 69–81, Nov. 27, 2008.

[17] M. Hoyhtya, S. Pollin, and A. Mammela, "Performance improvement with predictive channel selection for cognitive radios," in *Proc. 1st Int. Workshop Cognitive Radio Advanced Spectrum Management*, Feb. 2008, pp. 1–5.

[18] S. Huang, X. Liu, and Z. Ding, "Optimal transmission strategies for dynamic spectrum access in cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 8, no. 12, pp. 1636–1648, Dec. 2009.

[19] J. Unnikrishnan and V. Veeravalli, "Algorithms for dynamic spectrum access with learning for cognitive radio," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 750–760, Feb. 2010.

[20] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Stat. Soc.*, ser. B, vol. 41, pp. 148–177, 1979.

[21] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25, pp. 287–298, 1988.

[22] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *J. Appl. Probab.*, vol. 27, pp. 637–648, 1990.

[23] J. E. Niño-Mora, "Restless bandits, partial conservation laws and indexability," *Adv. Appl. Probab.*, vol. 33, pp. 76–98, 2001.

[24] K. Liu and Q. Zhao, "A restless bandit formulation of opportunistic access: Indexability and index policy," in *Proc. 5th IEEE Conf. Sensor, Mesh Ad Hoc Communications Networks (SECON) Workshops*, Jun. 2008.

[25] K. Liu and Q. Zhao, "Channel probing for opportunistic access with multi-channel sensing," in *Proc. IEEE Asilomar Conf. Signals, Systems, Computers*, Oct. 2008.

[26] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle's index for dynamic multichannel access," *arXiv* 2008 [Online]. Available: http://arxiv.org/abs/0810.4658

[27] J. Le Ny, M. Dahleh, and E. Feron, "Multi-UAV dynamic routing with partial observations using restless bandit allocation indices," presented at the 2008 Amer. Control Conf. Seattle, WA, Jun. 2008.

[28] X. Xiao, K. Liu, and Q. Zhao, "Opportunistic spectrum access in self similar primary traffic," *EURASIP J. Adv. Signal Process. (Special Issue on Dynamic Spectrum Access for Wireless Networking)*.

Keqin Liu received the B.S. degree in automation from Southeast University, China, in 2005 and the M.S. degree in electrical engineering from the University of California at Davis, in 2008. He is currently working towards the Ph.D. degree at the University of California at Davis.

He is currently a Research Assistant in the Department of Electrical Engineering, University of California at Davis. His research interests are stochastic optimization in dynamic systems, distributed control in multiagent systems, and signal processing in wireless networks.

Qing Zhao received the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, in 2001.

In August 2004, she joined the Department of Electrical and Computer Engineering at University of California at Davis (UC Davis), where she is currently an Associate Professor. Her research interests are in the general area of dynamic systems and communication networks.

Dr. Zhao received the 2000 Young Author Best Paper Award from the IEEE Signal Processing Society and the 2008 Outstanding Junior Faculty Award from the UC Davis College of Engineering.

Bhaskar Krishnamachari received the B.E. degree in electrical engineering from The Cooper Union, New York, in 1998 and the M.S. and Ph.D. degrees from Cornell University, Ithaca, NY, in 1999 and 2002, respectively.

He is currently an Associate Professor and Ming Hsieh Faculty Fellow in the Department of Electrical Engineering at the University of Southern California, Los Angeles. His primary research interest is in the design and analysis of algorithms and protocols for wireless networks.

Authorized licensed use limited to: Nanjing University. Downloaded on November 10,2020 at 16:05:04 UTC from IEEE Xplore. Restrictions apply.