Channel Probing in Opportunistic Communication Systems

Channel Probing in Opportunistic Communication Systems

Matthew Johnston, Isaac Keslassy, Eytan Modiano

***Abstract*—We consider a multi-channel communication system in which a transmitter has access to M channels, but does not know the state of any of the channels. We model the channel state using an ON/OFF Markov process, and allow the transmitter to probe a single channel at predetermined probing intervals to decide over which channel to transmit. For models in which the transmitter must transmit over the probed channel, it has been shown that a myopic policy probing the channel most likely to be ON is optimal. In this work, we allow the transmitter to select a channel over which to transmit that is potentially different from the probed channel. For a system of two channels, we show that the choice of which channel to probe does not affect the throughput. For a system with many channels, we show that a probing policy that probes the channel that is second-most likely to be ON results in higher throughput. We extend the channel probing problem to dynamically choose when to probe based on probing history, and characterize the optimal probing policy for various scenarios.**

I. Introduction

Consider a communication system in which a transmitter has access to multiple channels over which to communicate. The state of each channel evolves independently from all other channels, and the transmitter does not know the channel states *a priori*. The transmitter is allowed to probe a single channel after a predefined time interval to learn the current state of that channel. Using the information obtained from the channel probes and the memory in the channel state process, the transmitter selects a channel in each time-slot over which to transmit, with the goal of maximizing throughput, or the number of successful transmissions.

This framework applies broadly to many opportunistic communication systems, in which there exists a tradeoff between overhead and performance. In many wireless communication systems, knowledge of the instantaneous channel state can improve the network throughput. For example, in an LTE network, transmitters can intelligently select subcarriers which have a high channel quality [**?**]. Additionally, in scenarios in which an adversarial jammer is attempting to block communication, channel probing may be used to find the frequency bands that yield the highest throughput. However, when there is a large number of channels over which to transmit, or a large number of users to transmit to, it may be impractical to learn the channel state information (CSI) of every channel before scheduling a transmission; consequently, it may be only practical for the transmitter to obtain partial channel state information, and use that partial CSI to make a decision. Therefore, the transmitter must decide *how much* information to obtain, and *which* information is needed in order to make efficient scheduling decisions.

In the context of channel probing, the decision of what information to obtain translates to the decision of which channel to probe. We refer to this decision as the *probing policy*. Similarly, the decision of how much information to acquire translates to deciding how often to probe channels for CSI. This decision is referred to throughout this work as the *probing interval*. We consider both the scenario in which the probing interval is constant between channel probes, and the scenario where the probing interval is a function of the channel probing history, and is allowed to vary from probe to probe.

Several works have studied channel probing policies in multichannel communication problems [3]–[9]. Of particular interest is the work in [10] and [11], in which the authors assume that after a channel is probed, the transmitter must transmit over that channel. They show that the optimal probing policy is a myopic policy, which probes the channel most likely to be ON. This model is also considered in [3], which characterizes the capacity region achievable and solves for the optimal policy as the limit of a sequence of linear programs in terms of state action frequencies with increasing state spaces. The work in [12] extends the common two-state channel model to a multi-state Markov model, and establishes the optimality of myopic policy in a system similar to that of [10] and [11], in which the transmitter must use a probed channel over which to transmit.

The works in [4]–[8] consider a model where the channel state is independent over time; thus, probing a channel in the current slot will yield no information about that channel

M. Johnston is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 02139 USA e-mail: (mrj@mit.edu).

I. Keslassy is with the Department of Electrical Engineering, Technion, Haifa, Israel, e-mail: (isaac@ee.technion.ac.il).

E. Modiano is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 02139 USA e-mail: (modiano@mit.edu).

This work was supported by NSF Grants CNS-0915988 and CNS-1217048, and ARO MURI Grant W911NF-08-1-0238.

This work was presented in part at the IEEE ISIT conference, July 2013, [1] and the IEEE GLOBECOM conference, December 2013 [2].

in the future. Furthermore, these works allow for multiple channel probes per time slot, and are concerned with finding the optimal subset of channels to probe. In [8], all the channel probes occur simultaneously, and the objective is to determine what subset of channels to probe. On the other hand, [4]–[6], [13], [14] considers a sequential channel probe problem. In this framework, transmitters are able to probe one channel at a time, and based on the result of that channel probe, decide whether to probe another channel, used one of the probed channels for transmission, or use an un-probed channel to save on the additional overhead of channel probing. These works are typically modeled as stopping-time stochastic optimization problems, where the optimization is concerned with a single time slot. The work in [13] shows the optimal stopping-time policy obeys a threshold structure, and can be described by an index policy. The work in [14] considers independent, Rayleigh faded channels and shows that a 1-step lookahead policy is optimal for this setting. The work in [**?**] also analyzes the sequential probing problem, but carefully considers the overhead associated with acquiring channel state information in an 802.11 implementation. Our paper differs from the above works as we restrict the transmitter to probing a single channel at each time slot due to the time and bandwidth associated with the CSI acquisition.

In [15], the authors consider allocating power to two channels, with channel states that vary over time according to a Markov Process. They formulate the rate allocation problem as a partially observable Markov decision process (POMDP) and show several properties of the optimal solution. Finally, the work in [16] assumes the controller has full CSI, but this information is delayed, in that it takes several time slots for the controller to learn the channel state of each channel.

In this work, we study the channel probing problem for wireless opportunistic communication, in which the transmitter is able to transmit over a channel other than that which was probed. This model aims to capture the benefit of opportunistically selecting channels based on a time-varying channel state. In a system with two channels, we show that the choice of which channel to probe does not affect the expected throughput. Additionally, we identify scenarios such that when the probability distribution of the channel state differs between the two channels, it is optimal to always probe one of the channels. For an system with an asymptotically large number of channels, we show that the myopic probing policy in [10], [11] is no longer optimal. Specifically, we prove using renewal theory that a simple policy, namely the policy which probes the channel that is second most likely to be ON, has a higher per-slot expected throughput. We characterize the per-slot throughput for these policies, and calculate the optimal fixed probing interval as a function of a fixed probing cost. Furthermore, we prove the optimality of this policy for a system of three channels, and conjecture that this policy is in fact optimal for systems with any number of channels. In the second half of the work, we extend our model to allow for a dynamic optimization of the probing intervals based on the results of past channel probes. We formulate the problem as a Markov decision process, and introduce a state action frequency approach to solve for the optimal probing intervals. For the case of an infinite system of channels, we explicitely characterize the optimal probing interval for various probing policies.

The remainder of this paper is organized as follows. We describe the model and problem formulation in detail in Section II. In Section III, we analyze the channel probing problem for a system with two channels. In section IV, we find the optimal probing policy for a system with three channels, and conjecture the optimal policy in a general system. We extend this to an infinite channel system in Section V, and apply renewal theory to show that the myopic policy is suboptimal by analytically computing the expected per-slot throughput of another policy, which is proven outperform the myopic policy of [10]. In Section VI, we solve for the optimal probing intervals when a fixed cost is associated with probing.

II. System Model

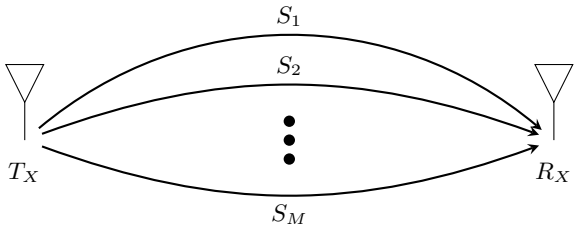


Figure 1: System model: transmitter and receiver connected through M independent channels

Consider a transmitter and a receiver that communicate using one of M independent channels, as shown in Figure 1. Assume time is slotted and at every time slot, each channel is either in an OFF state or an ON state. Channels are i.i.d. with respect to each other, and evolve across time according to a discrete time Markov process described by Figure 2.

At each time slot, the transmitter chooses a single channel over which to transmit. If that channel is in the ON state, then the transmission is successful; otherwise, the transmission fails. We assume the transmitter does not receive feedback regarding previous transmissions[1]. The objective is to maximize the expected sum-rate throughput, equal to the number of successful transmissions over time.

The transmitter obtains channel state information (CSI) by explicitly probing channels at predetermined intervals. In particular, the transmitter probes the receiver every k slots

[1]If such feedback exists in the form of higher layer acknowledgements, it arrives after a significant delay and is not useful for learning the channel state.

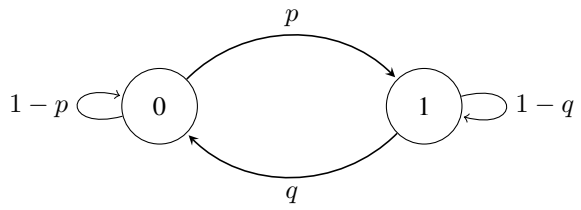


Figure 2: Markov Chain describing the channel state evolution of each independent channel. State 0 corresponds to an OFF channel, while state 1 corresponds to an ON channel.

for the state of one of the channels at the current time. Assume this information is delivered instantaneously, which was the same assumption as made in many previous works (e.g. [8], [10]). The transmitter uses the history of channel probes to make a scheduling decision. We emphasize that the transmitter may use a channel other than that which was probed for transmission. For example, if the transmitter probes a channel and it is found to be OFF, the transmitter can use a different channel for that transmission which is more likely to be ON.

A. Notation

Let $S_i(t)$ be the state of channel i at time t, where $S_i(t) = 1$ corresponds to a channel that is ON at time t, and $S_i(t) = 0$ corresponds to channel in the OFF state. The transmitter has an estimate of this state based on previous probes and the channel state distribution. Define the *belief* of a channel to be the probability that a channel is ON given the history of channel probes. For any channel i that was last probed k slots ago and was in state s_i, the belief x_i is given by

$$\begin{aligned} x_i(t) &= \mathbf{P}\big(\text{Channel } i \text{ is ON}|\text{probing history}\big) \\ &= \mathbf{P}\big(S_i(t) = 1|S_i(t-k) = s_i\big) \end{aligned} \tag{1}$$

where the second equality follows from the Markov property of the channel state process. The above probability is computed using the k-step transition probabilities of the Markov chain in Figure 2:

$$\begin{aligned} p_{00}^k &= \frac{q + p(1-p-q)^k}{p+q}, p_{01}^k = \frac{p - p(1-p-q)^k}{p+q} \\ p_{10}^k &= \frac{q - q(1-p-q)^k}{p+q}, p_{11}^k = \frac{p + q(1-p-q)^k}{p+q}. \end{aligned} \tag{2}$$

Throughout this work, we assume that $1 - p - q \geq 0$, corresponding to channels with "positive memory." The positive memory property ensures that a channel that was ON k slots ago is more likely to be ON at the current time, than a channel that was OFF k slots ago. This allows the transmitter to make efficient scheduling decisions without explicitly obtaining CSI at each time slot. Mathematically, this property is described by the set of inequalities:

$$p_{01}^i \leq p_{01}^j \leq p_{11}^k \leq p_{11}^l \qquad \forall i \leq j \quad \forall l \leq k. \tag{3}$$

As the CSI of a channel grows stale, the probability that the channel is in the ON state approaches π, the stationary distribution of the chain in Figure 2.

$$\lim_{k\to\infty} p_{01}^k = \lim_{k\to\infty} p_{11}^k = \pi = \frac{p}{p+q}. \tag{4}$$

Lastly, let $\tau^k(\cdot)$ be the function representing the change in belief of a channel over k time-slots when no new information regarding that channel is obtained.

$$\tau^k(x_i) = x_i p_{11}^k + (1 - x_i) p_{01}^k \tag{5}$$

This function will be used throughout this paper when analyzing the state transition properties of the system.

B. Optimal Scheduling

Since the objective is to maximize the expected sum-rate throughput, the optimal transmission decision at each time slot is given by the maximum likelihood (ML) rule, which is to transmit over the channel that is most likely to be ON, i.e. the channel with the highest belief. The expected throughput in a time slot is therefore given by

$$\max_i x_i(t). \tag{6}$$

where $x_i(t)$ is the belief of channel i at time t. Following the linearity of the state transition function $\tau^k(x_i)$ in (5), and the positive memory assumption, the optimal scheduling decision remains the same in between channel probes, as no additional CSI is obtained.

III. Two-Channel System

To begin, we consider a two-channel system, and formulate the problem of deciding which channel to probe using dynamic programming (DP), over a finite horizon of length N. Each index n corresponds to a time slot at which a probing decision is made. Assume there are k time slots between channel probes; thus, index n corresponds to time slot $t = kn$. The system state at each probing index n is equal to the vector $(x_1(n), x_2(n))$, the belief of channel 1 and channel 2 as defined in (1). Let $f^k(x_1, x_2)$ be the accumulated throughput over the k slots between channel probes, when channel 1 is probed. The function $f^k(x_1, x_2)$ is computed by conditioning on the result of the state of channel 1. If channel 1 is ON, which occurs with probability x_1, then the transmitter uses that channel for k slots, resulting in throughput $\sum_{i=0}^{k-1} p_{11}^i$. If the probed channel is OFF, then the other channel is used for transmission over those k slots, yielding throughput $\sum_{i=0}^{k-1} \tau^i(x_2)$. Consequently, the expected accumulated throughput is given by

$$f^k(x_1, x_2) = x_1 \sum_{i=0}^{k-1} p_{11}^i + (1 - x_1) \sum_{i=0}^{k-1} \tau^i(x_2) \tag{7}$$

Similarly, given the above definition, $f^k(x_2, x_1)$ is the accumulated throughput over the k slots between channel probes when channel 2 is probed.

We proceed by developing the DP value function for each probing decision. Let J_n^i be the expected reward after the nth probe if the choice is made to probe channel i at the current probing instance, and then follow the optimal probing policy for all subsequent probes. The expected reward after the last probe is given by:

$$J_N\big(x_1, x_2\big) = \max\Big(J_N^1\big(x_1, x_2\big), J_N^2\big(x_1, x_2\big)\Big) \quad (8)$$

$$J_N^1\big(x_1, x_2\big) = f^k(x_1, x_2) \quad (9)$$

$$J_N^2\big(x_1, x_2\big) = f^k(x_2, x_1) \quad (10)$$

Equations (9) and (10) follow since N is the final channel probe (in a time horizon of length N), and thus the only reward is the immediate reward, which is given by (7). At probing time $0 \leq n < N$, the expected reward function is defined recursively. If the decision at probe n is to probe channel 1, then an expected throughput of $f^k(x_1, x_2)$ is accumulated between probes, and at the next probe, the belief of channel 1 will be p_{11}^k (p_{01}^k) if the probed channel was ON (OFF), and the belief of channel two, which was not probed, will be $\tau^k(x_2)$. Thus, $J_n(x_1, x_2)$ is defined recursively as:

$$J_n\big(x_1, x_2\big) = \max\Big(J_n^1\big(x_1, x_2\big), J_n^2\big(x_1, x_2\big)\Big) \quad (11)$$

$$J_n^1\big(x_1, x_2\big) = f^k(x_1, x_2) + x_1 J_{n+1}(p_{11}^k, \tau^k(x_2)) + (1 - x_1) J_{n+1}(p_{01}^k, \tau^k(x_2)) \quad (12)$$

$$J_n^2\big(x_1, x_2\big) = f^k(x_2, x_1) + x_2 J_{n+1}(\tau^k(x_1), p_{11}^k) + (1 - x_2) J_{n+1}(\tau^k(x_1), p_{01}^k) \quad (13)$$

The dynamic program in (8)-(13) can be solved to compute the optimal probing policy for the two channel system. To begin with, we prove the following property of the immediate reward after probing, $f^k(x_1, x_2)$.

Lemma 1. $f^k(x_1, x_2) = f^k(x_2, x_1)$

The proof of Lemma 1 is given in the Appendix. Lemma 1 states that the immediate reward for probing channel 1 is the same as that for probing channel 2, for all probing intervals k. This is a consequence of the ability of the transmitter to choose over which channel to transmit *after* a channel probe, and accounts for the key difference between the model considered in this paper, and models considered in previous works [10], [11]. Using this result, we present the main result of this section.

Theorem 1. *For a two-user system with independent channels evolving over time according to an ON/OFF Markov chain with transition probabilities p and q, and probing epochs fixed at intervals of k slots, then for each channel probe, the total reward from probing channel 1 is equal to that of probing channel 2.*

Corollary 1. *The channel probing policy which always probes channel 1 (2) is optimal in a two-channel system.*

The proof of Theorem 1 is given in the Appendix, and follows using induction based on Lemma 1, and the affinity of the expected reward function in (8)-(13). Corollary 1 follows directly from Theorem 1. Intuitively, when a channel is probed, the transmitter receives information about the optimal channel to use until the next probe. For example, if the probed channel is ON, it is optimal to transmit over that channel until the next probe occurs. On the other hand, if the probed channel is OFF, it is optimal to transmit over the un-probed channel, because the belief of that channel will always be higher than that of the OFF channel, based on the inequalities in (3). Thus, the only information required from the channel probe is which channel to transmit over until the subsequent channel probe, and this information can be obtained through probing either channel.

This result is in contrast to the result in [11], which proves that the optimal decision is to probe the channel with the highest belief. However, their model assumed that a transmission must occur on the probed channel, whereas our model allows the transmitter to choose the channel over which to transmit based on the result of the probe. Consequently, the myopic policy of [11] is not a uniquely optimal policy in this setting.

Theorem 1 is used to determine the optimal fixed probing interval. Clearly, probing more frequently yields higher throughput, but requires more resources as well. To capture this, we associate a fixed cost c with each probe. The goal is to determine the probing interval k that maximizes the difference between throughput earned and cost accumulated.

Theorem 2. *Assume a fixed-interval probing scheme with probing cost c. The optimal probing interval is given by*

$$k^* = \arg\max_k \frac{\pi p_{10}^k - c(p+q)}{k(p+q)}. \quad (14)$$

Proof. From Corollary 1, the optimal probing policy is that which always probes channel 1. Under this policy, the belief of channel 2 equals the steady state probability of being in the ON state (π) given in (4). Channel 1 is probed every time, and will be on a fraction π of the time. When channel 1 is ON, a throughput of $\sum_{i=0}^{k-1} p_{11}^k$ is obtained, and when it is OFF, the throughput is simply πk, the expected throughput yielded by channel 2 over an interval of duration k. Consequently, the expected per-slot throughput accounting for the cost of probing is given by

$$\frac{1}{k}\left(-c + \pi \sum_{i=0}^{k-1} p_{11}^i + (1-\pi)\pi k\right) = \frac{-c}{k} + \pi + \frac{\pi p_{10}^k}{k(p+q)}. \quad (15)$$

The proof follows by maximizing the above expression with

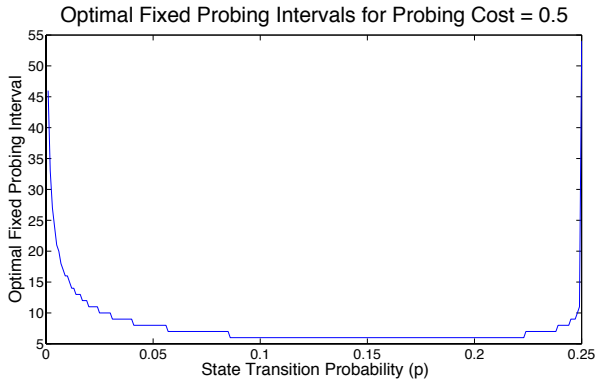


Figure 3: Optimal fixed probing interval for a two channel system as a function of state transition probability $p = q$. In this example, $c = 0.5$.

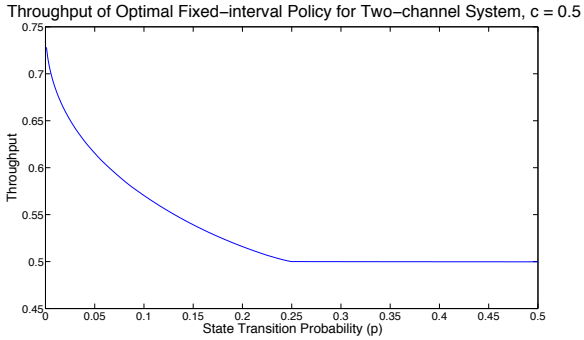


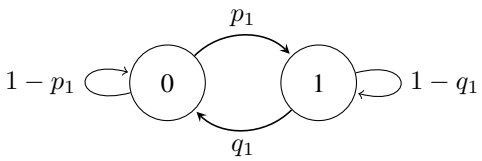
Figure 4: Throughput under the optimal fixed-interval probing policy for a two-channel system as a function of the state transition probability $p = q$. In this example, $p = q = 0.05$.

respect to k. □

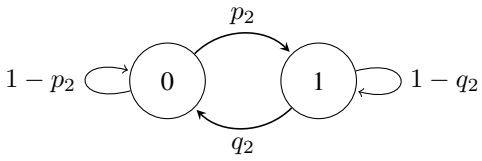
Figure 3 shows the optimal probing interval as a function of the state transition probability. As the state transition probability increases, each probe gives less information for the same cost. Thus, as the transition probability starts to decrease, the optimal probing interval decreases, since information needs to be obtained more frequently to account for the reduced information in each probe. As p continues to grow, the reward from probing becomes so small that the cost does not justify it, and eventually it becomes optimal to not probe.

Figure 4 shows the throughput under the optimal probing interval from Theorem 2 for various transition probabilities. At the state transition probability increases, throughput decreases. Note the optimal throughput does not drop below the steady state probability π, because at that point, it is optimal not to probe due to the high probing cost, and guess which channel to use.

Theorems 1 and 2 combine to characterize the optimal fixed-interval probing-policy for a two channel system. However, when the two channels are not identically distributed,



(a) Channel 1



(b) Channel 2

Figure 5: Two asymmetric Markov Chains, where $p_1 \leq 1/2$, $q_1 \leq 1/2$, $p_2 \leq 1/2$ and $q_2 \leq 1/2$

.

the optimal probing decision depends on the channel statistics, as shown in Section III-A. Furthermore, if the probing epochs are not fixed, the decision to probe depends on the results of the previous probe, yielding an advantage to probing one channel over the other, as shown in Section VI.

A. Heterogeneous Channels

In this section, we extend the results of the previous section to the case where the two channels differ statistically, i.e. channel 1 evolves in time according to the Markov chain in Figure 5a, and channel 2 evolves according to the chain in Figure 5b. Denote the k-step transition probability of channel 1 as $a^k_{i,j}$ and the k-step transition probability of channel 2 as $b^k_{i,j}$. Additionally, let π_1 and π_2 be the steady state ON probability of channel 1 and channel 2 respectively.

Intuitively, it is optimal to probe the channel with *more memory*, as that probe yields more information. For example, consider a channel that varies rapidly, with $p_1 = q_1 = \frac{1}{2} - \epsilon$, and a channel which rarely changes state, with $p_2 = q_2 = \epsilon$. Probing the low-memory channel provides accurate information for a few time slots, but that information quickly becomes stale, and the transmitter effectively guesses which channel is ON until the next probe. On the other hand, probing the high-memory channel yields information that remains accurate for many time slots after the probe. This intuition is confirmed in the following result.

Theorem 3. *For a two-user system with channel states evolving as described above, and probing instances fixed to intervals of k slots, if p_1, p_2, q_1, q_2 satisfy*

$$b^i_{11} \geq a^i_{11} \qquad \forall i, \tag{16}$$

then, the optimal probing policy is to probe channel 2 at all probing instances.

The proof of Theorem 3 is given in the Appendix, and follows by reverse induction over the channel probing

instances. To highlight its significance, we present the following corollaries.

Corollary 2. *Assume the two channels satisfy $\pi_1 = \pi_2$, and that $p_1+q_1 \geq p_2+q_2$. Then, the optimal policy is to always probe channel 2.*

Proof. We can rewrite the k-step transition probability of the second chain from (2) as follows.

$$\begin{aligned} b_{11}^i &= \frac{p_2 + q_2(1-p_2-q_2)^i}{p_2+q_2} \\ &= \pi_1 + (1-\pi_1)(1-p_2-q_2)^i \qquad (17) \\ &\geq \pi_1 + (1-\pi_1)(1-p_1-q_1)^i \qquad (18) \\ &= a_{11}^i \qquad (19) \end{aligned}$$

where (17) follows from the assumption that $\pi_1 = \pi_2$, and (18) follows from the assumption that $p_1 + q_1 \geq p_2 + q_2$. Therefore, $b_{11}^i \geq a_{11}^i$, and applying Theorem 3 concludes the proof. □

Corollary 3. *Assume the two channels satisfy $p_1 + q_1 = p_2 + q_2$, and that $\pi_1 \leq \pi_2$. Then, the optimal policy is to always probe channel 2.*

Proof. We can rewrite the k-step transition probability of the second chain from (2) as follows.

$$\begin{aligned} b_{10}^i &= \frac{q_2(1-(1-p_2-q_2)^i)}{p_2+q_2} \\ &= (1-\pi_2)(1-(1-p_1-q_1)^i) \qquad (20) \\ &\leq (1-\pi_1)(1-(1-p_1-q_1)^i) \qquad (21) \\ &= a_{10}^i \qquad (22) \end{aligned}$$

where (20) follows from the assumption that $p_1 + q_1 = p_2 + q_2$, and the inequality in follows from the assumption that $\pi_1 \leq \pi_2$. Since $b_{10}^i \leq a_{10}^i$, then $b_{11}^i \geq a_{11}^i$, and Theorem 3 can be applied to complete the proof. □

The above two corollaries describe scenarios where asymmetries in the channel statistics result in the optimal policy of always probing one of the two channels. This is in contrast to Theorem 1 where the channels are homogeneous, and probing either channel yields the same throughput. Corollary 2 states that if the channels are equally likely to be ON in steady state, the optimal decision is to probe the channel with the smaller $p_i + q_i$. In this context, $p_i + q_i$ is related to the rate at which the channel approaches the steady state. In particular, the Markov channel state approaches its stationary distribution exponentially at a rate equal to the second eigenvalue of the transition probability matrix, which for a two-state chain is $1-p-q$. The channel which approaches steady state more slowly is the channel with more memory, thus confirming our intuition that probing the channel with more memory is always optimal. Corollary 3 applies to a system in which the rate at which the steady state is reached is the same for both channels, but channel 2 is more likely to be ON in steady state than channel 1. In this case, it is optimal to probe the channel with the highest steady state probability of being ON at all probing instances.

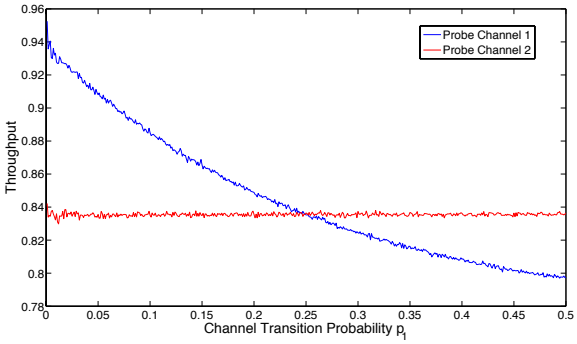


Figure 6: Throughput of 'Probe Channel 1' policy and 'Probe Channel 2' policy. In this example, p_1 is varied from 0 to $\frac{1}{2}$, and q_1 is chosen so $\pi = \frac{3}{4}$. The second channel satisfies $p_2 = \frac{1}{4}$ and $q_2 = \frac{1}{12}$, resulting in $\pi_2 = \pi_1$

Figure 6 plots the throughput obtained by the policy which always probes channel 1 versus the policy that always probes channel 2 for a sample set of parameters, measured through simulation. For the second channel, $p_2 = \frac{1}{4}$ and $q_2 = \frac{1}{12}$, so that $\pi_2 = \frac{3}{4}$. Then π_1 is fixed at $\frac{3}{4}$, but p_1 is varied from 0 to $\frac{1}{2}$. When channel 1 has less memory than channel 2, probing channel 1 yields much higher throughput than the alternative. In this example, when p_1 is very small, probing channel 1 results in a 15% throughput improvement over probing channel 2.

Theorem 1 and Theorem 3 describe scenarios in which probing one of the two channels at all probing instances is optimal. The simplicity of the optimal probing policy in these cases is an artifact of the transmitter only having two-channels from which to choose. Theorem 1 does not hold for systems with more than two channels. As the number of channels increases, a policy always probing one of the channels is suboptimal. Therefore, additional analysis is required for a system with more than two channels.

IV. Optimal Channel Probing over Finitely Many Channels

As mentioned above, for systems with more channels, i.e. $M > 2$, the policy of always probing one of the channels is suboptimal. In particular, the optimal probing policy is a function of the beliefs of the channels. In this section, we show that the policy which probes the channel with the second highest belief is optimal for a system of three channels, and conjecture an extension to a general system of finitely many channels.

A. Three Channel System

To begin, consider a system of three channels, with channel states identically distributed according to the Markov

chain in Figure 2. The following result characterizes the optimal channel probing policy as a function of the beliefs of the three channels.

Theorem 4. *In a system of three channels, where a single channel is probed every k slots, the optimal probing policy is to probe the channel with the second-highest belief.*

Denote by x_i the belief of the channel with the i^{th} largest belief. Thus, $x_1 \geq x_2 \geq x_3$. The probe second-best policy probes the channel with belief x_2. If that channel is ON, the transmitter uses that channel to transmit over for the next k slots. After these k slots, the best channel is the channel that was last probed, with belief $\tau^k(1)$, where τ^k is the information-decay function defined in (5). If on the other hand, the probed channel is OFF, the transmitter will use the channel with the highest belief among the remaining channels, x_1. After k slots, that channel will have belief $\tau^k(x_1)$, and the belief of the probed channel will be the smallest, at $\tau^k(0)$.

Define a function W_n as follows:

$$\begin{aligned} W_n(x_1,&x_2,x_3) \\ &\triangleq f^k(x_1,x_2) + x_2 W_{n+1}\big(\tau^k(1),\tau^k(x_1),\tau^k(x_3)\big) \\ &\quad + (1-x_2) W_{n+1}\big(\tau^k(x_1),\tau^k(x_3),\tau^k(0)\big) \end{aligned}$$

for all $0 \leq n \leq N$, where $f^k(\cdot)$ is the immediate reward function defined in (7). Let $W_{N+1}(x_1,x_2,x_3) = 0$ by convention. Note that $W_n(x_1,x_2,x_3)$ is the expected throughput of the probe second-best policy from time n onwards if and only if $x_1 \geq x_2 \geq x_3$. Additionally, if $x_2 \geq x_1 \geq x_3$, then $W_n(x_1,x_2,x_3)$ is the expected reward of the policy which probes the channel with the highest belief at index n, and then probes the channel with the second highest belief at all subsequent times. The following results hold for this definition of W_n, and is used to prove Theorem 4.

Lemma 2. *If $x_1 \geq x_2 \geq x_3$, then for all $0 \leq n \leq N$,*

$$W_n(x_1,x_2,x_3) \geq W_n(x_2,x_1,x_3) \tag{23}$$

Lemma 3. *If $x_1 \geq x_2 \geq x_3$, then for all $0 \leq n \leq N$,*

$$W_n(x_1,x_2,x_3) \geq W_n(x_1,x_3,x_2) \tag{24}$$

The proofs of Lemmas 2 and 3 are given in the Appendix.

Proof of Theorem 4. Without loss of generality, assume the beliefs of the three channels x_1,x_2,x_3 satisfy $x_1 \geq x_2 \geq x_3$. The proof follows using reverse induction on the probing index n. For $n = N$, probing the best channel yields throughput $W_N(x_2,x_1,x_3)$, while probing the second and third best channels yields throughput $W_N(x_1,x_2,x_3)$ and $W_N(x_1,x_3,x_2)$ respectively. By Lemma 2, $W_N(x_1,x_2,x_3) \geq W_N(x_2,x_1,x_3)$, and by Lemma 3, $W_N(x_1,x_2,x_3) \geq W_N(x_1,x_3,x_2)$; therefore, probing the second-best channel is optimal at $n = N$.

Now assume it is optimal to probe the second-best channel at probes $n+1,\ldots,N$. At probing instance n, the throughput of the three potential choices of channels are given by $W_n(x_2,x_1,x_3), W_n(x_1,x_2,x_3)$, and $W_n(x_1,x_3,x_2)$ for probing the best, second-best, and third best channels respectively. By Lemma 2, $W_n(x_1,x_2,x_3) \geq W_n(x_2,x_1,x_3)$, and by Lemma 3, $W_n(x_1,x_2,x_3) \geq W_n(x_1,x_3,x_2)$; therefore, probing the second-best channel is optimal at n as well. By induction, probing the second-best channel is optimal at all probing times. □

This result is exciting as it differs from the previous result in [10] which stated that the policy which probes the best channel is optimal for the model in which the transmitter must use the channel that was probed for transmission. In our model, the transmitter can collect CSI separately from the transmission decision, and therefore probing the second-best channel yields a higher throughput. Further intuition as to why the probe second-best policy is optimal is presented in Section V-B.

B. Arbitrary Number of Channels

Theorem 4 shows that the probe second-best policy is optimal for a system of three channels. In general, for $M > 3$, we conjecture that the probe second-best policy remains optimal.

Conjecture 1. *The probe second-best policy is optimal among all channel probing policies for fixed probing intervals k.*

The proof used for the $M = 3$ channel case does not extend to $M \geq 4$. In [10], the authors used a coupling argument to circumvent this issue and prove the optimality of the myopic policy for their setting for general networks. However, due to the additional complexity of the probe second-best policy, this coupling argument does not hold in our setting. Instead, we believe the general case can be proven by bounding the maximum difference in expected reward from being in a better state after probing the k^{th} best channel for $k \geq 2$, and proving that this extra reward must be less than the gain in the immediate expected reward that probing the second-best channel offers.

We have performed numerous simulations which support Conjecture 1. As an example, Table I presents the throughput obtained by different probing policies over varying numbers of channels. Observe that the probe second-best policy outperforms the other probing policies. However, the advantage of using the probe second-best policy over similar policies, such as probe best and probe third best, is relatively small.

In Figure 7, we compare the performance of the probe-best policy, the probe second-best policy, and probe third-best policy as a function of the number of channels in the system, for a fixed probing interval. We see that as the number of channels grows, the gap in performance between the

Simulation	3 Channels	5 Channels	7 Channels	10 Channels
Probe Channel 1	0.6955	0.6959	0.6957	0.6958
Probe Best Channel	0.7455	0.7640	0.7650	0.7659
Probe Second-Best	0.7553	0.7787	0.7799	0.7808
Probe Third Best	0.6849	0.7617	0.7691	0.7706
Probe Worst	0.6860	0.6804	0.6810	0.6806
Round Robin	0.7460	0.7649	0.7658	0.7661

Table I: Comparison of different probing policies for a fixed probing interval (6) and time horizon 2,000,000. State transition probability $p = q = 0.05$

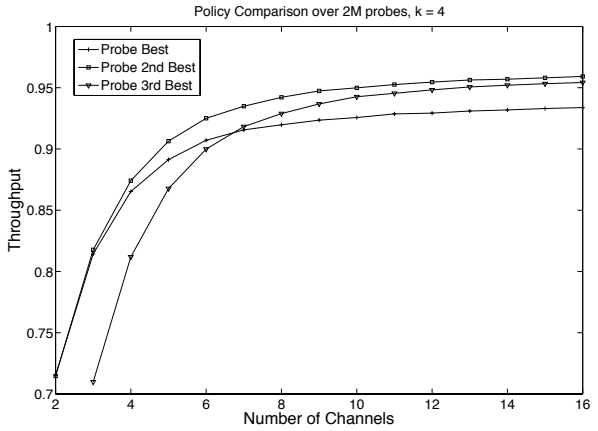


Figure 7: Comparison of the probe best policy, the probe second-best policy, and the probe third best policy as a function of the number of channels in the system. This simulation was run over 2 million probes, with each probe being at an interval of 4 time slots.

probe second-best policy and the probe second-best policy increases. Furthermore, the probe third best policy becomes more efficient as the number of channels increase, but does not reach the level of throughput of the probe second-best policy.

V. Infinite-Channel System

As the number of channels increases, the state space grows large and the probing formulation becomes more difficult to analyze. However, as the number of channels grows to infinity, we can introduce an assumption which affords various simplifications to the state space of the system. Whenever a probed channel is OFF, it is effectively removed from the system. This is because there always exists a channel which has not been probed in the previous N slots, for any finite N, and thus its belief is equal to the steady state ON probability π, and $p_{01}^k \leq \pi$ for all k. Therefore, since an OFF channel has belief $p_{01}^k \leq \pi$ for any finite k, it will never be optimal to transmit over that channel under the policies considered in this paper.

In this section, we use the infinite channel assumption to characterize the average throughput under several probing policies. We consider the myopic policy which is shown to be optimal for the model in [10], [11], as well as a round robing policy which probes channels sequentially. In addition, we characterize the throughput of the probe second-best policy, which is conjectured to be the optimal probing policy for a finite number of channels in Section IV, and prove that it outperforms the other two policies in this setting.

A. Probe-Best Policy

To begin, consider the *probe-best policy*, which probes the channel with the highest belief. This policy is commonly referred to as a myopic or greedy policy, as it maximizes the immediate reward without regard to future rewards. Intuitively, such a policy is advantageous as the channel with the highest belief is the most likely to be ON at the current time, yielding the highest expected throughput. Recall that this policy is shown to be optimal for the model in [10], [11]. For our model, we have the following results.

Theorem 5. *The state of the system is given by an infinite vector of beliefs for each channel. Without loss of generality, assume this vector is sorted as* $\mathbf{x} = \{x_1, x_2, \ldots\}$ *such that* $x_1 \geq x_2 \geq x_3 \ldots$. *The class of recurrent states under the probe-best policy satisfy* $x_1 \geq \pi$, *and* $x_i = \pi$ *for all other channels* $i \neq 1$.

Proof. The probe best policy probes the channel with belief x_1. If this channel is ON, its belief becomes p_{11}^1 in the next slot, and it remains the channel with the highest belief by the equality in 3. If that channel is OFF, it is removed from the system as per the infinite channel assumption. Therefore, the vector consisting of $x_i = \pi$ for all i is reachable from any state. This state corresponds to the transmitter having no information about the network. The only other state reachable from this state is reached when an ON channel is found, at which point, the state returns to a state satisfying $x_1 \geq \pi$, and $x_i = \pi \quad \forall i \neq 1$. □

Theorem 6. *Assume the transmitter makes probing decisions every* k *slots according to the probe best policy. The expected per-slot throughput is given by*

$$\mathbb{E}[Thpt] = \pi + \frac{\pi p_{10}^k}{k(p+q)(p_{10}^k t + \pi)} \tag{25}$$

Proof. We use renewal theory to compute the average throughput. Under the probe best policy, Theorem 5 states that only one channel can have belief greater than π. Define a renewal to occur immediately prior to probing a channel with belief π. Therefore, if a channel is probed and if it is OFF, it is removed from the system and a renewal occurs k slots later (before the next probe). If the channel is ON, that channel is probed at all future probing instances until it

is found to be OFF. The expected inter-renewal time $\bar{X}_B$ is given by

$$\bar{X}_B = (1-\pi)k + \pi(k\mathbb{E}(N)+k) \tag{26}$$
$$= k + k\pi\mathbb{E}(N) \tag{27}$$

where N is a random variable denoting the number of times an ON channel is probed before it is OFF, and is geometrically distributed with parameter p_{10}^k. Equation (27) reduces to

$$\bar{X}_B = k + \frac{\pi k}{p_{10}^k}. \tag{28}$$

The expected reward $\bar{R}_B$ incurred over a renewal interval is πk for the interval immediately after the OFF probe, and $\sum_{i=0}^{k-1} p_{11}^i$ for each subsequent ON probe. If the first probe is ON, then there will be N probes until the final OFF probe. Thus, the expected accumulated reward over a renewal interval is expressed as

$$\bar{R}_B = (1-\pi)\pi k + \pi(\pi k + \mathbb{E}[N]\sum_{i=1}^{k} p_{11}^i) \tag{29}$$
$$= \pi k + \pi\mathbb{E}\big[N\big]\sum_{i=0}^{k-1} p_{11}^i = \pi k + \frac{\pi\sum_{i=0}^{k-1} p_{11}^i}{p_{10}^k} \tag{30}$$

Using results from renewal-reward theory [17], the average per-slot reward is given by the ratio of the expected reward over the renewal interval divided by the expected length of that interval.

$$\frac{\bar{R}_B}{\bar{X}_B} = \frac{\pi k p_{10}^k + \pi\sum_{i=0}^{k-1} p_{11}^i}{k p_{10}^k + \pi k} = \pi + \frac{\pi p_{10}^k}{k(p+q)(p_{10}^k+\pi)} \tag{31}$$

□

Observe that the per-slot throughput is always larger than π, and decreases toward π as k increases. The probe best policy maximizes the immediate reward; however, the drawback of this policy is that when the probed channel is OFF, the transmitter has no knowledge of the state of the other channels as it searches for an ON channel, as described by Theorem 5. Consequently, transmitter probes channels with belief π until an ON channel is found, resulting in a low expected reward.

B. Probe Second-Best Policy

Now, consider a simple alternative policy, the *probe second-best* policy, which at each time slot probes the channel with the second-highest belief, and transmits on the channel with the highest belief after the channel probe. Consider channel state beliefs $x_1, x_2, x_3, \ldots$ where $x_1 \geq x_2 \ldots \geq x_i \ldots \geq \pi$. The probe-best policy of the Section V-A probes the channel with belief x_1. If it is ON, the transmitter uses that channel (resulting in throughput equal to 1 for the next slot) and if it is OFF, the transmitter uses the channel with the next highest belief x_2. Thus, the expected immediate reward of probing the best channel is given by

$$x_1 + (1-x_1)x_2 = x_1 + x_2 - x_1x_2, \tag{32}$$

The probe second-best policy instead probes the channel with belief equal to x_2. If this channel is ON, it transmits over that channel (resulting in throughput equal to 1) and otherwise transmits over the channel with highest belief, $x_!$. The expected immediate reward of probing the second-best channel is therefore equal to

$$x_2 + (1-x_2)x_1 = x_1 + x_2 - x_1x_2. \tag{33}$$

Hence, the probe second-best policy has the same immediate reward as the probe best policy. To understand how the probe second-best policy outperforms the probe-best policy, consider the following result, analogous to Theorem 5 for the probe best policy.

Theorem 7. *The state of the system is given by an infinite vector of beliefs for each channel. Without loss of generality, assume this vector is sorted as* $\mathbf{x} = \{x_1, x_2, \ldots\}$ *such that* $x_1 \geq x_2 \geq x_3 \ldots$. *The class of recurrent states under the probe second-best policy satisfy* $x_1 \geq x_2 \geq \pi$, *and* $x_i = \pi$ *for all other channels* $i \neq 1, 2$.

Proof. The probe second-best policy probes the channel with belief x_2. If this channel is ON, its belief becomes p_{11}^k iat the next probe, and it becomes the channel with the highest belief, while x_1 becomes the second highest belief. If the channel is OFF instead, it is removed from the system as per the infintie channel assumption. Therefore, the vector consisting of $x_1 \geq \pi$ and $x_i = \pi$ for all i is reachable from any state. This state corresponds to the transmitter having information of only one channel. From this state, by probing an ON channel, the system transitions into a state with two channels having belief greater than π; however, the system can never have more than two channels with $x_i > \pi$. □

By Theorem 7, since two channels can have belief greater than π under the probe second-best policy, when the probe second-best policy probes an OFF channel, the transmitter uses the channel with the next highest belief, while probing new channels to find another ON channel. This approach results in a higher expected throughput over that interval than under the probe best policy, which transmits on a channel with belief equal to the steady state probability π. It is this intuition that leads us to consider the probe second-best policy. The following theorem confirms our intuition, by showing that the probe second-best policy yields a higher throughput than the probe best policy.

Theorem 8. *The average reward of the probe second-best policy is greater than that of the probe best policy, for all fixed probing intervals* k.

Proof. Theorem 8 is proved using renewal theory to compute

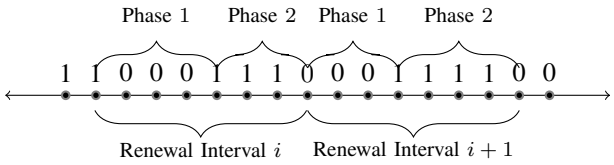


Figure 8: Illustration of renewal process. Points represent probing instances, and labels represent probing results. Each renewal interval consists of phase 1, and phase 2.

Time	0	k	$2k$	$3k$	$4k$	$5k$	$6k$
Best Channel Belief	p_{11}^{2k}	p_{11}^{3k}	p_{11}^{4k}	p_{11}^{k}	p_{11}^{k}	p_{11}^{k}	p_{11}^{2k}
Second-Best Belief	π	π	π	p_{11}^{5k}	p_{11}^{2k}	p_{11}^{2k}	π
Probe Result	0	0	1	1	1	0	-

Table II: Example renewal interval starting at time 0 and renewing at time $6k$. At each probing interval, the second-best channel is probed.

the average throughput of the probe second-best policy, and comparing it to that of the probe best policy. The key to the proof is in the definition of the renewal interval. We define a renewal to occur when the best channel has belief p_{11}^{2k}, and the second-best channel (and every other channel) has belief π. A renewal interval is divided into two phases: Phase 1 includes all the channel probes until a probe results in an ON channel, and phase 2 the subsequent probes until an OFF channel is probed. The division of renewal intervals into phases is illustrated in Figure 8. In Phase 1, the transmitter probes channels with belief π until an ON channel is probed, and in phase 2, the transmitter probes the second-best channel with belief greater than π until an OFF channel is probed. This definition ensures that the inter-renewal periods are i.i.d. The state evolution during an sample renewal interval is shown in Table II.

The expected inter-renewal time is given by $k\mathbb{E}(N_1 + N_2)$, where N_1 is the number of probes required to find an ON channel in phase 1, and is geometrically distributed with parameter π, and N_2 is the number of probes required until the next OFF probe in phase 2. The distribution of N_2 is dependent on N_1, and has the following distribution function.

$$N_2 = \begin{cases} 1 \text{ w.p. } p_{10}^{(N_1+2)k} \\ i \text{ w.p. } p_{11}^{(N_1+2)k} p_{10}^{2k} (p_{11}^{2k})^{i-2} \quad i \geq 2 \end{cases} \tag{34}$$

Therefore,

$$\bar{X}_{SB} = k\mathbb{E}(N_1 + N_2) = k\left(\frac{1}{\pi} + 1 + \frac{\mathbb{E}[p_{11}^{(2+N_1)k}]}{p_{10}^{2k}}\right) \tag{35}$$

During phase 1 of a renewal, the expected reward accumulated is given by

$$\bar{R}_{SB}^1 = \mathbb{E}\left[\sum_{i=0}^{(N_1-1)k-1} p_{11}^{i+2k} + \sum_{i=0}^{k-1} p_{11}^{i}\right]. \tag{36}$$

The first term is the throughput obtained from transmitting over the best channel while looking for an ON channel, which starts with belief p_{11}^{2k} and decays until an ON channel is found, as shown in Table II. In phase 2, the expected reward is given by

$$\bar{R}_{SB}^2 = \mathbb{E}\left[(N_2 - 1)\sum_{i=0}^{k-1} p_{11}^{i} + \sum_{i=0}^{k-1} p_{11}^{k+i}\right]. \tag{37}$$

For $N_2 - 1$ intervals of length k, the transmitter will transmit over a channel that was ON, yielding throughput $\sum_{i=0}^{k} p_{11}^{i}$. Then, for the last interval prior to the renewal, the best channel has belief p_{11}^{k}, and the expected accumulated throughput over that interval is $\sum_{i=0}^{k} p_{11}^{k+i}$. The average reward per time slot is given by

$$\frac{\bar{R}_{SB}^1 + \bar{R}_{SB}^2}{\bar{X}_{SB}} = \pi + \frac{\pi p_{10}^{k}(\pi + p_{10}^{2k})}{(p+q)k[\pi^2 + p_{10}^{2k}(1-(1-p-q)^k + \pi)]} \tag{38}$$

We can compute the difference between (38) and (25) from Theorem 6 as

$$\frac{\bar{R}_{SB}^1 + \bar{R}_{SB}^2}{\bar{X}_{SB}} - \frac{\bar{R}_B}{\bar{X}_B} = \frac{((1-p-q)^k \pi p_{10}^{k})^2}{k(p+q)(\pi + p_{10}^{k})(\pi^2 + p_{10}^{2k}(\pi + 1 - (1-p-q)^k))} \tag{39}$$

Since $p \leq \frac{1}{2}$ and $q \leq \frac{1}{2}$, we have $0 \leq (1-p-q)^k \leq 1$ for all k. Therefore, the expression in (39) is positive, completing the proof. □

Theorem 8 asserts that probing the channel with the second highest belief is a better policy than probing the channel with the highest belief under fixed-interval probing policies. A numerical comparison between these two policies is shown in Figure 9. This result is in sharp contrast to the result in [10] that shows that probing the channel with the highest belief is optimal. In our model, when a probed channel is OFF, the transmitter uses its knowledge of the system to transmit over another channel believed to be ON. In the model of [10], when an OFF channel is probed, the transmitter cannot schedule a packet in that slot. This difference in reward after probing leads to significantly different probing policies. This result also supports Conjecture 1, claiming that the probe second-best policy is optimal among all policies.

C. Round Robin Policy

It is of additional interest to consider a min-max policy, the *round robin policy*, which probes the channel for which the transmitter has the least knowledge. In a system with finitely many channels, the round robin policy probes all of the channels sequentially, always probing the channel which was probed longest ago. When the number of channels grows

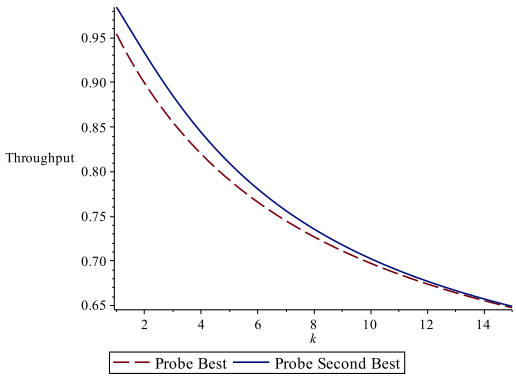


Figure 9: Comparison of the probe best policy and the probe 2nd best policy for varying probing intervals k. In this example, $p = q = 0.05$.

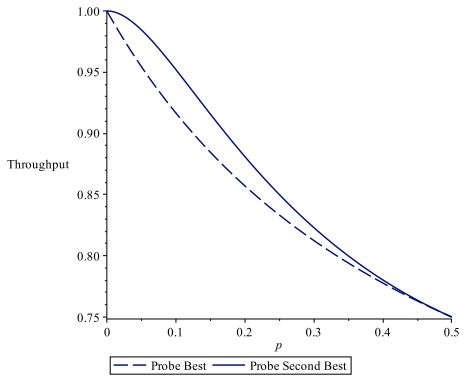


Figure 10: Comparison of the probe best policy and the probe 2nd best policy for varying state transition probabilities $p = q$.. In this example, $k = 1$.

to infinity, the transmitter always probes a channel that has previously never been probed. Consider channel state beliefs $x_1, x_2, x_3, \ldots$ where $x_1 \geq x_2 \ldots \geq x_i \ldots \geq \pi$. Under the round robin policy, a channel with belief π is probed; if that channel is ON it will be used by the transmitter (earning throughput 1) and otherwise the channel with the highest belief will be used (earning throughput x_1, the belief of the best channel). Thus, the immediate reward of round robin is given by:

$$\pi + (1 - \pi)x_1 = \pi + x_1 - \pi x_1. \tag{40}$$

By comparing (40) to (32), it is clear the immediate reward of the round robin policy is less than that of the probe best and the probe second-best policy. Interestingly, the following Theorem shows that the average per-slot throughput is the same for the round robin policy as the myopic probe best policy.

Theorem 9. *For all fixed k, the round robin policy has a per-slot average throughput of*

$$\mathbb{E}[Thpt] = \pi + \frac{\pi p_{10}^k}{k(p+q)(p_{10}^k + \pi)}, \tag{41}$$

the same as the probe best policy.

Proof. Let a renewal occur every time a new channel is probed and found to be ON. Since the result of each probe is an i.i.d. random variable with parameter π, the inter-renewal intervals are i.i.d.. The inter-renewal time $X_{RR} = k \cdot N$, where k is the time between probes, and N is a geometric random variable with parameter π, as defined in (4). Over that interval, the transmitter transmits over the last channel known to be ON, until a new ON channel is found. The expected reward earned over each renewal period is given by

$$\bar{R}_{RR} = \mathbb{E}\left[\sum_{i=0}^{N*k-1} p_{11}^i\right] \tag{42}$$

$$= \mathbb{E}\left[\pi N k + \frac{p_{10}^{Nk}}{p+q}\right] \tag{43}$$

$$= k + \frac{p_{10}^k}{p + q - q(1-p-q)^k}. \tag{44}$$

Thus, the time-average reward is given by

$$\frac{\bar{R}_{RR}}{\bar{X}_{RR}} = \pi + \frac{\pi p_{10}^k}{k(p+q)(\pi + p_{10}^k)}, \tag{45}$$

which is the same as the reward of the probe best policy in Theorem 6. □

Recall from Theorem 5, that under the probe best policy, at most one channel can have belief greater than π. In contrast, under the round robin policy many channels can have belief greater than π. Thus, Theorem 9 is surprising, since the round robin policy trades off immediate reward for increasing knowledge of the channel states, but yields the same average throughput as the probe best policy.

VI. Dynamic Optimization Of Probing Intervals

Until this point, we've assumed the transmitter chooses channels to probe at predetermined probing intervals. However, an alternate approach is to optimize the time until the next channel probe dynamically, as a function of the collected CSI. For example, after an ON probe, the transmitter has knowledge of a channel which yields high throughput, and therefore may not need to probe a new channel immediately. On the other hand, if that probed channel is OFF, the transmitter may benefit from probing a new channel in the near future to make up for lost throughput. In this example, the optimal probing policy sets the probing interval dynamically, based on the results of the previous probe.

In this section, the optimal dynamic probing policy is modeled as a stochastic control problem, where at each time

slot, a decision is made whether to probe a channel or not, and if so, which channel to probe.

A. Two-Channel System

To begin with, consider a system with only two channels. The optimal channel probing problem is formulated as a Markov Decision Process (MDP) or a Dynamic Programming problem (DP) over a finite horizon of length T. At each time slot, the system state is the vector consisting of the belief of each channel's state. After observing the system state at time t, the transmitter selects an action from a set of possible actions: probe channel 1, probe channel 2, probe neither channel. Thus, the expected reward function at time slot t is given by

$$J_t(x_1, x_2) = \max\{J_t^0(x_1, x_2), J_t^1(x_1, x_2), J_t^2(x_1, x_2)\}, \tag{46}$$

where J_t^0 is the expected reward given that neither channel is probed at the current slot, and J_t^1 and J_t^2 are the expected reward functions given that channel 1 or channel 2 is probed respectively. When the transmitter chooses to not probe either channel, the throughput obtained is given by the maximum of the channel beliefs, since the transmitter will transmit on the better of the two channels. Assume channel probes incur a cost of c. This channel cost represents the time that must be spent to execute the channel probing process, thus taking away from resources which could have been used for additional throughput. When a channel is probed and is ON, the transmitter uses that channel and a reward (throughput) of 1 is earned. On the other hand, if the probed channel is OFF, a unit throughput is earned only if the second channel is ON. Therefore, the terminal cost at time $t = T$ is given by

$$J_T^0(x_1, x_2) = \max(x_1, x_2), \tag{47}$$

$$J_T^1(x_1, x_2) = -c + x_1 + (1 - x_1)x_2, \tag{48}$$

$$J_T^2(x_1, x_2) = -c + x_2 + (1 - x_2)x_1 \tag{49}$$

For $t < T$, the reward function includes the expected future reward, based on the result of the channel probe. If the transmitter does not probe a channel, the state at the next slot is given by $(\tau(x_1), \tau(x_2))$, where $\tau(\cdot) = \tau^1(\cdot)$ is the information decay function in (5). If a channel is probed, then the belief of that channel in the following slot is either p or $1-q$ depending on whether the probe results in an OFF channel or an ON channel respectively. Thus, the recursive expected reward DP equations are given by

$$J_t^0(x_1, x_2) = \max(x_1, x_2) + J_{t+1}\big(\tau(x_1), \tau(x_2)\big) \tag{50}$$

$$J_t^1(x_1, x_2) = -c + x_1 + x_2 - x_1x_2 + x_1 J_{t+1}\big(1 - q, \tau(x_2)\big) + (1 - x_1)J_{t+1}\big(p, \tau(x_2)\big) \tag{51}$$

$$J_t^2(x_1, x_2) = -c + x_1 + x_2 - x_1x_2 + x_2 J_{t+1}\big(\tau(x_1), 1 - q\big) + (1 - x_2)J_{t+1}\big(\tau(x_1), p\big) \tag{52}$$

The maximizer of (46) is the optimal probing policy at time slot t as a function of the current state. Note that the state space is countably infinite, as each belief x_i has a one-to-one mapping to an (S, k) pair, where S is the state at the last channel probe, and k is the time since the last probe.

The following result states that this expected reward function is convex, which is used to characterize the region in which probing is optimal.

Theorem 10 (Convexity)**.** *For all t, $J_t(x_1, x_2)$ is convex in x_1 for fixed x_2, and is convex in x_2 for fixed x_1.*

The proof of Theorem 10 is given in the appendix.

Using the convexity of the expected reward function, we can find sufficient conditions for probing optimality for a given state.

Theorem 11. *If for any time slot t, the system state $(x_1(t), x_2(t))$ satisfies*

$$c \leq \min(x_1(t), x_2(t))\big(1 - \max(x_1(t), x_2(t))\big) \tag{53}$$

Then it is optimal to probe at slot t.

Proof.

$$J_t^0(x_1, x_2) = \max(x_1, x_2) + J_{t+1}\big(\tau(x_1), \tau(x_2)\big) \tag{54}$$

$$\leq \max(x_1, x_2) + x_1 J_{t+1}\big(1 - q, \tau(x_2)\big) + (1 - x_1)J_{t+1}\big(p, \tau(x_2)\big) \tag{55}$$

$$= \max(x_1, x_2) + J_t^1(x_1, x_2) + c - x_1 - x_2 + x_1x_2 \tag{56}$$

Where (55) follows from Theorem 10. Therefore, $J_t^0(x_1, x_2) - J_t^1(x_1, x_2) \leq 0$ if

$$c - x_1 - x_2 + x_1x_2 + \max(x_1, x_2) \leq 0 \tag{57}$$

$$c \leq \min(x_1, x_2)\big(1 - \max(x_1, x_2)\big) \tag{58}$$

□

Theorem 11 can be interpreted as when the belief of the two channels are sufficiently close together, it is optimal to probe (subject to probing cost). While the convexity bound yields sufficient conditions for probing optimality, necessary conditions do not follow directly from this analysis. Additionally, the convexity bound used in (55) is loose, and thus probing is often optimal even in states which do not satisfy the conditions of Theorem 11.

B. State Action Frequency Formulation

The channel probing MDP can also be modeled as an infinite horizon, average cost problem. In this case, it can be formulated as a linear program (LP) in terms of state action frequencies, which can be solved to determine the optimal policy. A state action frequency vector $\omega(\mathbf{s}; a)$ exists for each state and potential action, and corresponds to a stationary randomized policy such that $\omega(\mathbf{s}; a)$ equals the steady state probability that at a given time slot, the state is $\mathbf{s}$ and the action taken is a. For weakly communicating finite state and action MDP's, there exists a solution to the state action frequency LP that corresponds to a deterministic stationary policy [3]. The complete state action frequency LP is given in [18].

Figure 11 illustrates the structure of the solution to the state action frequency LP, showing the optimal decision as a function of the belief of channel 1 (x_1) and the belief of channel 2 (x_2). The system state can only reach a countable subset of the points on the x_1-x_2 plane. Under any policy, except for the policy where a channel is never probed, there is a single recurrent class of states, and only states in this class will have non-zero state action frequencies. From any recurrent state, if the optimal decision is not to probe, the system state will move to the next state $(\tau(x_1), \tau(x_2))$. The coordinates $(\tau^k(x_1), \tau^k(x_2))$ represent a line between (x_1, x_2) and (π, π) parameterized by k. Thus, while the transmitter refrains from probing, the system state follows a trajectory between the current state to (π, π). Based on this observation, and the results in Figure 11, we can characterize the structure of the optimal probing algorithm.

For a given set of parameters, there exists a probing-region, e.g. the dotted convex region in Figure 11, and a point (π, π), denoted by the dot in the center of Figure 11. At each time slot, if the current state lies outside of the probing region, the optimal decision is to not probe, and the state moves along the linear trajectory toward (π, π). When the state lies on or inside the probing region, the controller probes one of the channels. The state reached after the channel probe will correspond to a point on the edge of the unit square in Figure 11, since the belief of a probed channel is either 0 or 1. Then the process repeats, and the state will follow a new trajectory toward the point (π, π). Therefore, the region for which probing is optimal translates to a threshold policy, where probing becomes optimal after a certain time threshold, given by the distance between the point on the edge of the unit square, and the probing region.

If the point (π, π) lies outside of the probing region, then there exists a trajectory to (π, π) that does not intersect the probing region. If this is the case, the state after a channel probe will eventually be a point on the unit square such that the line between that point and (π, π) does not cross the probing optimality region. Thus, the optimal decision is to never probe, and the state monotonically approaches (π, π) along the linear trajectory. In this situation, all states

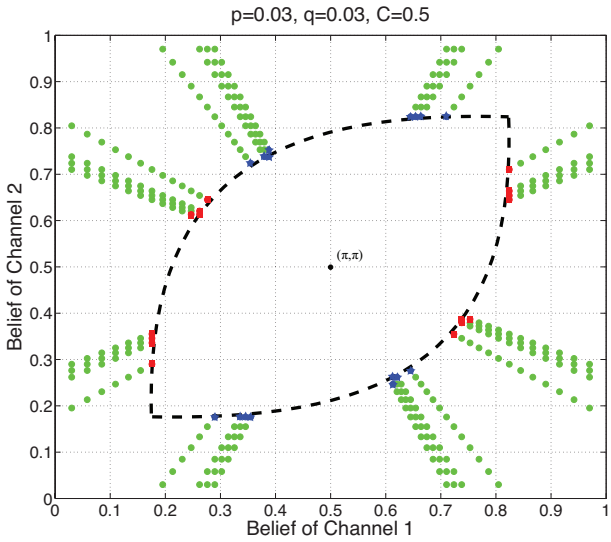


Figure 11: Optimal decisions based on SAFs. White space corresponds to transient states under the optimal policy, and green circles, red boxes, and blue stars correspond to recurrent states where the optimal action is to not probe, probe channel 1, and probe channel 2 respectively.

are transient under the optimal policy. This occurs when the probing cost is sufficiently large that obtaining any CSI is impractical. In summary, the optimal time between probes is given by the distance between the state immediately following a probe and the state on the boundary of the probing region, lying on the line between the current state and (π, π). To find the probing region, and the decisions to make at each point on the probing regions, the SAF LP must be solved.

C. Infinite-Channel System

For a system with more than two channels, the previous approaches can be used to formulate the problem of finding the optimal probing intervals. The drawback of these approaches is that the state space grows exponentially with the number of channels, and it becomes impractical to solve the MDP approach in Section VI-A and the state action frequency LP in section VI-B. However, in the asymptotic limit of the number of channels, the infinite channel assumption in Section V can be applied to greatly simplify the state space, and new approaches can be developed to characterize the optimal probing intervals. Clearly, these intervals are related to the underlying probing policy used to select the channels to probe. In this section, we consider two of the channel probing policies from Section V: the probe best policy and the round robing policy, and characterize the optimal intervals at which to probe.

To begin, assume the decision of which channel to probe is given by the probe-best policy. The optimal decision as to

whether to probe is a function of the state, and is described by the following Theorem.

Theorem 12. *For a system in which the transmitter only probes the channel with the highest belief, the optimal probing decision is to probe immediately after probing an OFF channel, and to probe k^* slots after probing an ON channel, where k^* is given by*

$$k^* = \arg\max_k \frac{1}{k\pi + p_{10}^k}\left(\frac{\pi p_{10}^k}{(p+q)} - c(\pi + p_{10}^k)\right) \quad (59)$$

Proof. As a result of Theorem 5, under the probe best policy, the belief of the best channel x_1 at every slot satisfies $x_1 \geq \pi$, and the belief of every other channel equals π. When a probed channel is OFF, it is removed from the system, and the belief of every channel is π, representing a state in which the transmitter has no knowledge of the system. The system remains in this state until an ON channel is found, as each OFF channel which is probed is removed from the system by the infinite channel assumption. If the optimal decision in this state is to not probe, then the transmitter never probes, since the state never changes. Thus, if it is optimal to probe in the state where the transmitter has no knowledge, then it is optimal to probe immediately after an OFF channel is probed. When a probed channel is ON, the highest belief is always $1-q$ in the next slot, and decays until that channel is probed again, as it will always remain the channel with the highest belief. Hence, there exists a threshold k^* after an ON probe such that after that time, it becomes optimal to probe.

Assume a probe occurs in the slot immediately after probing an OFF channel, and let k denote the number of slots after probing an ON channel until the best channel is probed again. Define a renewal to occur when the transmitter probes an OFF channel. It follows that the inter-renewal time is one slot if the next probed channel is OFF, and $1+kN$ if the probed channel is ON, where N is a random variable equal to the number of times the ON channel is probed until it turns OFF. Thus, the expected inter-renewal time is given by

$$\bar{X}_B = (1-\pi) + \pi(1 + k\mathbb{E}[N]) \quad (60)$$
$$= 1 + \pi k\mathbb{E}[N], \quad (61)$$

The random variable N is is geometrically distributed with parameter p_{10}^k. The reward accumulated over this interval is π if the probed channel is OFF, and N times $\sum_{i=0}^{k-1} p_{11}^i$ if the channel is ON, plus an additional π after the final OFF probe. A cost of c is incurred for each channel probe within this interval. The expected reward is given by

$$\bar{R}_B = (1-\pi)(\pi - c) + \pi\left(\mathbb{E}[N](\sum_{i=0}^{i} - c) + \pi - c\right) \quad (62)$$
$$= (\pi - c) + \pi\mathbb{E}[N](\sum_{i=0}^{k-1} p_{11}^i - c). \quad (63)$$

Therefore, the average per-time slot reward is given by the ratio of expected reward over a renewal interval to the expected length of the renewal interval:

$$\frac{\bar{R}_B}{\bar{X}_B} = \frac{p_{10}^k(\pi - c) + \pi(\sum_{i=0}^{k-1} p_{11}^i - c)}{p_{10}^k + k\pi} \quad (64)$$
$$= \pi - c\left(\frac{\pi + p_{10}^k}{k\pi + p_{10}^k}\right) + \frac{\pi p_{10}^k}{(p+q)(k\pi + p_{10}^k)} \quad (65)$$

The maximizing value of k in equation (65) is the optimal time k^* to wait after an ON probe. □

Theorem 12 characterizes the optimal probing interval under the probe best policy. If the probing policy changes, the optimal interval changes as well. However, the following result shows that under the round-robin policy, the optimal probing interval has a similar structure.

Theorem 13. *For a system in which the transmitter probes channels according to the round robin policy, the optimal decision is to probe a new channel immediately after probing an OFF channel, and to probe k' slots after probing an ON channel, where k' is given by*

$$k' = \arg\max_k \frac{-c(p+q) + p\mathbb{E}_N[\sum_{i=0}^{k+N-2} p_{11}^i]}{p(k-1) + p + q} \quad (66)$$

where N is a geometrically distributed random variable with parameter π.

Proof. In contrast to Theorem 12, there is no analog to Theorem 5 for round robin probing. Thus, we first prove the optimal form of the policy is a threshold policy, by proving the monotonicity of the expected reward function. Given the structure of the optimal policy, renewal theory is applied to characterize the optimal interval. To begin, we can write the expected reward as a function of k over a finite horizon.

$$J_T(k) = \max\left(p_{11}^k, -c + \pi + (1-\pi)p_{11}^k\right) \quad (67)$$
$$J_t(k) = \max\Big(p_{11}^k + J_{t+1}(k+1), -c + \pi\big(1 + J_{t+1}(1)\big) + (1-\pi)\big(p_{11}^k + J_{t+1}(k+1)\big)\Big) \quad (68)$$

where the left argument to the $\max(\cdot,\cdot)$ function represents the expected reward from not probing, and the right argument represents the expected reward from probing an unknown channel.

Under round robin probing, J_t is monotonically decreasing in k for all t. To see this, assume $t = T$, then assume k satisfies $\pi p_{10}^k \geq c$, then

$$J_T(k) = \max\left(p_{11}^k, -c + \pi + (1-\pi)p_{11}^k\right)$$
$$= p_{11}^k + \max(0, -c + \pi p_{10}^k) \quad (69)$$
$$= p_{11}^k - c + \pi(1 - p_{11}^k) = p_{11}^k(1-\pi) + \pi - c \quad (70)$$

which is monotonically decreasing in k, since p_{11}^k is a monotonically decreasing function of k. If on the other hand $\pi p_{10}^k \leq c$, then $J_T(k) = p_{11}^k$ which is monotonically decreasing in k.

Now assume $t \leq T$, and the hypothesis holds for $t+1, \ldots, T$, we will show using induction that it holds for t. Let $g(k) = p_{11}^k + J_{t+1}(k+1)$. By induction, $g(k)$ is monotonically decreasing in k, and using the analysis from the base case, the expression

$$J_n(k) = \max\left(g(k), -c + \pi\big(1 + J_{n+1}(1)\big) + (1-\pi)g(k)\right) \quad (71)$$

is also monotonically decreasing in k.

The remaining proof of Theorem 13 follows by reverse induction over the time horizon. Assume there is a k' such that it is optimal to probe at time T. Consider $k \geq k'$. It is optimal to probe if $c \leq \pi p_{10}^k$. However, $c \leq \pi p_{10}^{k'}$ since it is optimal to probe at k', and $p_{10}^k \geq p_{10}^{k'}$. Therefore, it is also optimal to probe at k.

Now consider $t \leq T$, and assume our induction hypothesis holds for $t+1$. The difference in the arguments to $\max(\cdot,\cdot)$ in (68) can be bounded as follows

$$-c + \pi(1 + J_{t+1}(1)) + (1-\pi)(p_{11}^k + J_{t+1}(k+1)) - p_{11}^k - J_{t+1}(k+1) \quad (72)$$
$$= -c + \pi(1 + J_{t+1}(1)) + -\pi(p_{11}^k + J_{t+1}(k+1)) \quad (73)$$
$$\geq -c + \pi(1 + J_{t+1}(1)) + -\pi(p_{11}^{k^*} + J_{t+1}(k^*+1)) \quad (74)$$
$$\geq 0. \quad (75)$$

where the first inequality holds from the monotonic property of the J function, and the second inequality holds from the assumption that it is optimal to probe for k'. Therefore, it is optimal to probe at t, and by induction, it is optimal to probe k' slots after an ON probe for some value of k'.

To characterize the optimal value of k', we introduce renewal theory using the renewals defined in Section V-C. Recall, a renewal occurs upon probing a channel which is ON. The expected time until the next renewal is the k' slots until the next probe, plus the number of slots it takes to find a new ON channel. Let N be the number of probes until an ON channel is found, which is geometrically distributed with parameter π. The expected inter-renewal time is given by

$$\bar{X}_R = \mathbb{E}_N[k + N - 1]. \quad (76)$$

Over this interval, a cost of c is incurred for each of the N channel probes, and at each time slot the transmitter uses the last known ON channel for transmission. Thus, the expected reward is given by

$$\bar{R}_R = \mathbb{E}_N\left[1 - Nc + \sum_{i=0}^{k+N-2} p_{11}^i\right]. \quad (77)$$

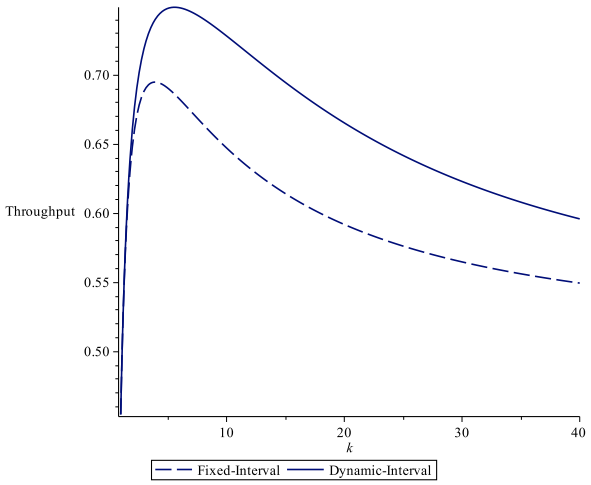
To determine the optimal k', we maximize the ratio of the expected reward to the expected length of the renewal interval, thus concluding the proof. □

Note that the optimal time to wait to probe after an ON probe under round robin (k') in (66) differs from the optimal k^* under the probe best policy in (59). Figure 13 plots the average reward of round robin and probe best for different values of k. Recall that under fixed probing intervals, Theorem 9 states that both policies have the same average reward. However, under dynamic probing intervals, the probe best policy outperforms the round robin policy. Figure 12 shows a comparison between expected throughput of the optimal fixed-interval probing policy and the optimal dynamic-interval policy under probe best and round robin. By looking at the maxima in these graphs, we observe that for the chosen parameters, introducing a dynamic probing-interval optimization yields an 8% gain in throughput under probe best, and a 5% gain in throughput under probe best.

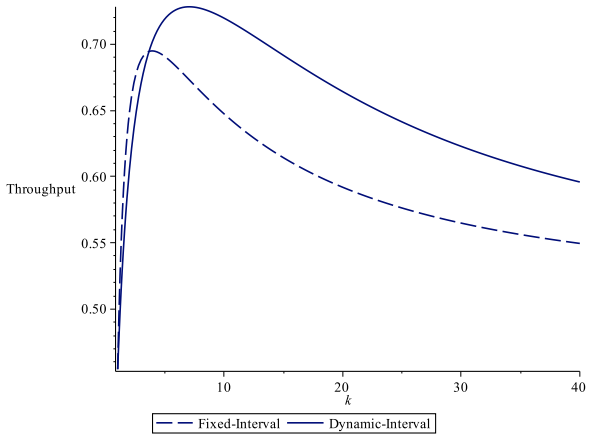
Based on the results of the fixed probing interval model, a natural extension to the above analysis is to consider the probe second-best policy, which was conjectured to be the optimal probing policy under fixed channel probing intervals. In contrast to probe best and round robin, the optimal time until the next probe under the probe second-best policy depends on the belief of the best channel after an ON channel is probed, and consequently, probe second-best does not have a single solution for the optimal probing interval after an ON channel has been probed. Thus, characterizing the optimal probing intervals is a more challenging problem in this context. It is an interesting and open problem to determine if the probe second-best policy is still optimal under dynamic probing intervals.

VII. Conclusion

This paper focuses on channel probing as a means of acquiring network state information, and optimizes the acquisition of this information in terms of which channels to probe and how often to probe these channels. In contrast to the work in [10], [11] that established the optimality of the myopic probe best policy, we showed that for a slightly modified model, these results no longer hold. Under



(a) Probe Best Probing Policy



(b) Round Robin Probing Policy

Figure 12: Comparison of the expected throughput of the probe best policy and the round robin policy under fixed intervals and under dynamic intervals. The x-axis plots k, the length of the interval. The maxima of each graph represents the optimal policy in each regime. In this example, $p = q = 0.05$ and $c = 0.5$.

a two channel system, we proved that probing either channel results in the same throughput, and under an infinite channel system, we proved that a simple alternative, the probe second-best policy, outperforms the probe best policy in terms of average throughput. We proved the optimality of the probe second-best policy in three channel systems, and conjecture that probing the second-best channel is the optimal decision in a general multi-channel system. Proving this conjecture is interesting, and remains an open problem.

Additionally, we showed that dynamically optimizing the probing intervals based on the results of the channel probe can additionally increase system throughput. We characterized the optimal probing intervals in a two channel system by formulating a markov decision problem, and using a state

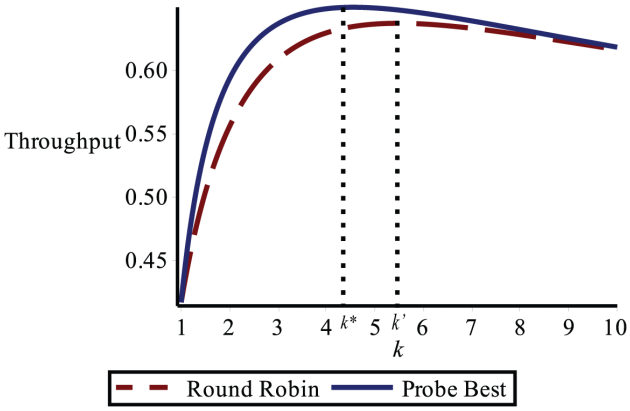


Figure 13: Comparison of the probe best policy and round robin for varying values of k, the minimum interval between probes. In this example, $p = q = 0.1$, and $c = 0.5$.

action frequency approach to solve the dynamic program. For the infinite channel case, we characterized the optimal probing intervals subject to a fixed probing policy, namely the probe best policy and the round robin probing policy. An extension to general probing polices, as well as a joint optimization over the probing decisions and the probing intervals is an interesting extension to this work.

VIII. Appendix

A. Proof of Lemma 1

Lemma 1: $f^k(x_1, x_2) = f^k(x_2, x_1)$

Proof of Lemma 1.

$$f^k(x_1, x_2) = x_2 \sum_{i=0}^{k-1} p_{11}^i + (1 - x_2) \sum_{i=0}^{k-1} \tau^i(x_1) \tag{78}$$

$$= \sum_{i=0}^{k-1} \left(x_2 p_{11}^i + (1 - x_2)\tau^i(x_1) \right) \tag{79}$$

$$= \sum_{i=0}^{k-1} \left(x_2 p_{11}^i + (1 - x_2)(\tau^k(x_1) = x_i p_{11}^i + (1 - x_1) p_{01}^i) \right) \tag{80}$$

$$= \sum_{i=0}^{k-1} \left(x_1 p_{11}^i + (1 - x_1)(\tau^k(x_2) = x_i p_{11}^i + (1 - x_2) p_{01}^i) \right) \tag{81}$$

$$= \sum_{i=0}^{k-1} \left(x_1 p_{11}^i + (1 - x_1)\tau^i(x_2) \right) = f^k(x_2, x_1) \tag{82}$$

□

B. Proof of Theorem 1

Proof of Theorem 1. This proof uses reverse induction on the probing index n. As a base case, consider $n = N - 1$.

$$J^1_{N-1}(x_1, x_2) = f^k(x_1, x_2) = f^k(x_2, x_1) = J^2_{N-1}(x_1, x_2) \tag{83}$$

Now assume $J^1_{n+1}(x_1, x_2) = J^2_{n+1}(x_1, x_2)$, and we prove this holds for index n.

First, we note that the function $f(x_1, x_2)$ is affine in both x_1 and x_2. To see this, consider $0 \leq \lambda \leq 1$.

$$\begin{aligned}
&\lambda f^k(a, x_2) + (1-\lambda) f^k(b, x_2) \\
&\quad = \sum_{i=0}^{k-1} \Big(\lambda a p^i_{11} + \lambda(1-a)\tau^j(x_2) \\
&\qquad + (1-\lambda) b p^i_{11} + (1-\lambda)(1-b)\tau^i(x_2) \Big) && (84) \\
&\quad = \sum_{i=0}^{k-1} \Big(p^i_{11}(\lambda a + (1-\lambda)b) + \tau^j(\tau^k(x_2)) \big(\lambda(1-a) \\
&\qquad + (1-\lambda)(1-b) \big) \Big) && (85) \\
&\quad = f^k(\lambda a + (1-\lambda)b, x_2) && (86)
\end{aligned}$$

As a consequence of Lemma 1, it also follows that

$$\lambda f^k(x_2, a) + (1-\lambda) f^k(x_1, b) = f^k(x_1, \lambda a + (1-\lambda)b) \tag{87}$$

Using the above fact, we can show that both J^1_{n+1} and J^2_{n+1} are affine as well.

$$\begin{aligned}
&\lambda J^1_{n+1}(a, x_2) + (1-\lambda) J^2_{n+1}(b, x_2) \\
&\quad = \lambda f^k(a, x_2) + \lambda \Big(a J_{n+2}(p^k_{11}, \tau^k(x_2)) \\
&\qquad + (1-a) J_{n+2}(p^k_{01}, \tau^k(x_2)) \Big) \\
&\qquad + (1-\lambda) f^k(b, x_2) + (1-\lambda) \Big(b J_{n+2}(p^k_{11}, \tau^k(x_2)) \\
&\qquad + (1-b) J_{n+2}(p^k_{01}, \tau^k(x_2)) \Big) && (88) \\
&\quad = f^k(\lambda a + (1-\lambda)b, x_2) \\
&\qquad + (\lambda a + (1-\lambda)b) J_{n+2}(p^k_{11}, \tau^k(x_2)) \\
&\qquad + (1 - \lambda a - (1-\lambda)b)(p^k_{01}, \tau^k(x_2)) && (89) \\
&\quad = J^1_{n+1}(\lambda a + (1-\lambda)b, x_2) && (90)
\end{aligned}$$

Similarly, since $J^1_n(x_1, x_2) = J^2_n(x_2, x_1)$, it is easy to show that J^2_{n+2} is affine in x_2 as well.

Using the results above, $J^1_n(x_1, x_2)$ is written as

$$\begin{aligned}
J^1_n(x_1, x_2) &= f^k(x_1, x_2) + x_1 J_{n+1}(p^k_{11}, \tau^k(x_2)) \\
&\quad + (1-x_1) J_{n+1}(p^k_{01}, \tau^k(x_2)) && (91) \\
&= f^k(x_1, x_2) + x_1 J^1_{n+1}(p^k_{11}, \tau^k(x_2)) \\
&\quad + (1-x_1) J^1_{n+1}(p^k_{01}, \tau^k(x_2)) && (92) \\
&= f^k(x_1, x_2) + J^1_{n+1}(\tau^k(x_1), \tau^k(x_2)) && (93) \\
&= f^k(x_1, x_2) + J^2_{n+1}(\tau^k(x_1), \tau^k(x_2)) && (94) \\
&= f^k(x_2, x_1) + x_2 J^2_{n+1}(\tau^k(x_1), p^k_{11}) \\
&\quad + (1-x_2) J^2_{n+1}(\tau^k(x_1), p^k_{01}) && (95) \\
&= f^k(x_2, x_1) + x_2 J_{n+1}(\tau^k(x_1), p^k_{11}) \\
&\quad + (1-x_2) J_{n+1}(\tau^k(x_1), p^k_{01}) && (96) \\
&= J^2_n(x_1, x_2) && (97)
\end{aligned}$$

where equations (92), (94), and (96) follow from the induction hypothesis, and equations (93) and (95) use the affinity of J^i_{n+1}, and Lemma 1.

□

C. Proof of Theorem 3

Theorem 3: For a two-user system with channel states evolving as described above, and probing instances fixed to intervals of k slots, if p_1, p_2, q_1, q_2 satisfy

$$b^i_{11} \geq a^i_{11} \qquad \forall i, \tag{98}$$

then, the optimal probing strategy is to probe channel 2 at all probing instances.

Proof of Theorem 3. This proof will use induction on the horizon length of the corresponding DP problem.

Define state transition functions

$$\tau^i_1(x) = a^i_{11} x + (1-x) a^i_{01} \tag{99}$$

$$\tau^i_2(x) = b^i_{11} x + (1-x) b^i_{01} \tag{100}$$

Base Case: Assume $n = N$. For this immediate-reward problem, the expected reward functions simplify to the following:

$$\begin{aligned}
&J^1_N(x_1, x_2) \\
&\quad = \sum_{i=0}^{k-1} \Big(x_1 \max(a^i_{11}, \tau^i_2(x_2)) + (1-x_1) \max(a^i_{01}, \tau^i_2(x_2)) \Big) && (101)
\end{aligned}$$

$$\begin{aligned}
&J^2_N(x_1, x_2) \\
&\quad = \sum_{i=0}^{k-1} \Big(x_2 \max(b^i_{11}, \tau^i_1(x_1)) + (1-x_2) \max(b^i_{01}, \tau^i_1(x_1)) \Big) && (102)
\end{aligned}$$

Since we have assumed that $b^i_{11} \geq a^i_{11}$, the following inequalities hold:

$$\begin{aligned}
b^i_{1,1} &\geq a^i_{1,1} \geq \tau^i_1(x_1) \\
b^i_{0,1} &\leq a^i_{0,1} \leq \tau^i_1(x_1)
\end{aligned} \tag{103}$$

Consequently, we can rewrite (102) as

$$J_n^2(s_1,k_1,s_2,k_2) = \sum_{i=0}^{k-1}\left(x_2 b_{11}^i + (1-x_2)\tau_1^i(x_1)\right)$$
$$= \sum_{i=0}^{k-1}\left(x_2 b_{11}^i + (1-x_2)x_1 a_{11}^i + (1-x_2)(1-x_1)a_{01}^i\right) \tag{104}$$

The proof of the base case differs slightly depending on the realization of s_2, so we will present two cases for each realization.

Case 1: $x_2 \geq \pi_2$. Equation (101) simplifies to

$$J_N^1(x_1,x_2)$$
$$= \sum_{i=0}^{k-1}\left(x_1\max(a_{11}^i,\tau_2^i(x_2)) + (1-x_1)\tau_2^i(x_2)\right)$$
$$= \sum_{i=0}^{k-1}\Big(x_1\max(a_{11}^i,\tau_2^i(x_2)) + (1-x_1)x_2 b_{1,1}^i + (1-x_1)x_2 b_{0,1}^i\Big) \tag{105}$$
$$= \sum_{i=0}^{k-1}\Big(x_1\max(a_{11}^i,\tau_2^i(x_2)) + x_2 b_{1,1}^i - x_1 x_2 b_{1,1}^i + (1-x_1)(1-x_2)b_{0,1}^i\Big) \tag{106}$$
$$= J_N^2(x_1,x_2) + \sum_{i=0}^{k-1}\Big(x_1\max(a_{11}^i,\tau_2^i(x_2)) - x_1 x_2 b_{1,1}^i + (1-x_1)(1-x_2)b_{0,1}^i - (1-x_2)x_1 a_{11}^i - (1-x_2)(1-x_1)a_{01}^i\Big) \tag{107}$$
$$\leq J_N^2(x_1,x_2) + \sum_{i=0}^{k-1}\Big(x_1\max(a_{11}^i,\tau_2^i(x_2)) - x_1 x_2 b_{1,1}^i - x_1(1-x_2)a_{11}^i\Big) \tag{108}$$
$$= J_N^2(x_1,x_2) + \sum_{i=0}^{k-1}\max\Big(x_1 a_{11}^i - (1-x_2)x_1 a_{11}^i, x_1\tau_2^i(x_2) - (1-x_2)x_1 a_{11}^i\Big) - x_1 x_2 b_{11}^i \tag{109}$$
$$= J_N^2(x_1,x_2) + \sum_{i=0}^{k-1}\max\Big(x_1 x_2(a_{11}^i - b_{1,1}^i), x_1(1-x_2)(b_{01}^i - a_{11}^i)\Big) \tag{110}$$
$$\leq J_N^2(x_1,x_2) \tag{111}$$

In the above, (108) and (111) follow from the fact that $b_{11}^i \geq a_{11}^i$.

Case 2: $x_2 \leq \pi_2$. Equation (101) simplifies to

$$J_N^1(x_1,x_2)$$
$$= \sum_{i=0}^{k-1}\left(x_1 a_{11}^i + (1-x_1)\max(a_{01}^i,\tau_2^i(x_2))\right)$$
$$= \sum_{i=0}^{k-1}\Big(x_1 a_{11}^i + (1-x_1)\max(a_{01}^i,\tau_2^i(x_2)) + (1-x_2)x_1 a_{11}^i - (1-x_2)x_1 a_{11}^i\Big) \tag{112}$$
$$= \sum_{i=0}^{k-1}\Big(x_1 x_2 a_{11}^i + (1-x_1)\max(a_{01}^i,\tau_2^i(x_2)) + (1-x_2)x_1 a_{11}^i\Big) \tag{113}$$
$$= J_N^2(x_1,x_2) + \sum_{i=0}^{k-1}\Big((1-x_1)\max(a_{01}^i,\tau_2^i(x_2)) + x_1 a_{11}^i x_2 - x_2 b_{11}^i - (1-x_2)(1-x_1)p_{01}^i\Big) \tag{114}$$
$$= J_N^2(x_1,x_2) + \sum_{i=0}^{k-1}\max\Big(x_2(x_1 a_{11}^i + (1-x_1)a_{01}^i) - x_2 b_{11}^i, x_1 x_2(a_{11}^i - b_{11}^i) + (1-x_1)(1-x_2)(b_{01}^i - a_{01}^i)\Big) \tag{115}$$
$$\leq J_N^2(x_1,x_2) \tag{116}$$

Where (114) results from applying (104), and (116) comes from the assumption that $a_{11}^i \leq b_{11}^i$.

Inductive Step: Assume that $J_l^1(x_1,x_2) \leq J_l^2(x_1,x_2)$, for all $n+1 \leq l \leq N$, we now prove that $J_n^1(x_1,x_2) \leq J_n^2(x_1,x_2)$. Therefore, the optimal cost to go $J_{n+1}(x_1,x_2) = J_{n+1}^2(x_1,x_2)$. By looking at expressions (101) and (102) from the analysis in the base case portion of the proof, we know that

$$\sum_{i=0}^{k-1}\left(x_1\max(a_{11}^i,\tau_2^i(x_2)) + (1-x_1)\max(a_{01}^i,\tau_2^i(x_2))\right) \leq \sum_{i=0}^{k-1}\left(x_2\max(b_{11}^i,\tau_1^i(x_1)) + (1-x_2)\max(b_{01}^i,\tau_1^i(x_1))\right) \tag{117}$$

To conclude the proof:

$$x_1 J_{n+1}^2(a_{11}^k,\tau_2^k(x_2)) + (1-x_1)J_{n+1}^2(a_{01}^k,\tau_2^k(x_2)) \tag{118}$$
$$= J_{n+1}^2(\tau_1^k(x_1),\tau_2^k(x_2)) \tag{119}$$
$$= x_2 J_{n+1}^2(\tau_1^k(x_1),a_{11}^k) + (1-x_2)J_{n+1}^2(\tau_1^k(x_1),a_{01}^k) \tag{120}$$

Where the above comes from the affinity of the function $J_n(x_1, x_2)$, shown in (88)-(88). □

D. Proof of Lemmas 2 and 3

Lemma 4. *Let $g(x, y)$ be any function satisfying $g(x, y) = ax + by + cxy + d$ for some constants a, b, c, d. Then,*

$$g(x, y) - g(y, x) = (x - y)(g(1, 0) - g(0, 1)) \tag{121}$$

Proof.

$$\begin{aligned} g(x, y) - g(y, x) &= ax + by + cxy + d - ay - bx - cyx - d \\ &= (x - y)(a - b) \\ &= (x - y)(g(1, 0) - g(0, 1)) \end{aligned}$$

□

Lemma 2: If $x_1 \geq x_2 \geq x_3$, then for all $0 \leq n \leq N$,

$$W_n(x_1, x_2, x_3) \geq W_n(x_2, x_1, x_3)$$

Proof of Lemma 2. The proof follows by reverse induction on the probing index n. For time $n = N$,

$$\begin{aligned} &W_N(x_1, x_2, x_3) - W_N(x_2, x_1, x_3) \\ &\qquad = f^k(x_1, x_2) - f^k(x_2, x_1) = 0 \end{aligned} \tag{122}$$

The last equality follows from Lemma 1. Assume the inductive hypothesis holds for $n + 1$:

$$\begin{aligned} &W_n(x_1, x_2, x_3) - W_n(x_2, x_1, x_3) \\ &\quad = (x_1 - x_2)(W_n(1, 0, x_3) - W_n(0, 1, x_3)) && (123) \\ &\quad = (x_1 - x_2)(f^k(1, 0) + W_{n+1}\big(\tau^k(1), \tau^k(x_3), \tau^k(0)\big) \\ &\qquad - f^k(0, 1) - W_{n+1}\big(\tau^k(1), \tau^k(0), \tau^k(x_3)\big) && (124) \\ &\quad = (x_1 - x_2)(W_{n+1}\big(\tau^k(1), \tau^k(x_3), \tau^k(0)\big) \\ &\qquad - W_{n+1}\big(\tau^k(1), \tau^k(0), \tau^k(x_3)\big) && (125) \\ &\quad \geq (x_1 - x_2)(W_{n+1}\big(\tau^k(1), \tau^k(0), \tau^k(x_3)\big) \\ &\qquad - W_{n+1}\big(\tau^k(1), \tau^k(0), \tau^k(x_3)\big) = 0 && (126) \end{aligned}$$

The inequality in (126) holds by the induction hypothesis of Lemma 3. □

Lemma 3: If $x_1 \geq x_2 \geq x_3$, then for all $0 \leq n \leq N$,

$$W_n(x_1, x_2, x_3) \geq W_n(x_1, x_3, x_2)$$

Proof of Lemma 3. The proof follows by reverse induction on the probing index n. For time $n = N$,

$$\begin{aligned} &W_N(x_1, x_2, x_3) \\ &\geq W_N(x_1, x_3, x_2) = f^k(x_1, x_2) - f^k(x_1, x_3) && (127) \\ &= (x_2 - x_3) \sum_{i=0}^{k-1} \big(p_{11}^i - \tau^i(x_1)\big) && (128) \\ &= (x_2 - x_3)(1 - x_1) \sum_{i=0}^{k-1} (p_{11}^i - p_{01}^i) \geq 0 && (129) \end{aligned}$$

where the inequality follows from the positive memory assumption on the channel. Now we assume the inductive hypothesis for Lemmas 2 and 3 hold for times at and after n.

$$\begin{aligned} &W_n(x_1, x_2, x_3) - W_n(x_1, x_3, x_2) \\ &\qquad = (x_2 - x_3)\big(W_n(x_1, 1, 0) - W_n(x_1, 0, 1)\big) && (130) \\ &= (x_2 - x_3)(f^j(x_1, 1) + W_{n+1}\big(\tau^k(1), \tau^k(x_1), \tau^k(0)\big) \\ &\qquad - r(x_1, 0) - W_{n+1}\big(\tau^k(x_1), \tau^k(1), \tau^k(0)\big) && (131) \\ &\geq (x_2 - x_3)\Big(W_{n+1}\big(\tau^k(1), \tau^k(x_1), \tau^k(0)\big) \\ &\qquad - W_{n+1}\big(\tau^k(x_1), \tau^k(1), \tau^k(0)\big)\Big) && (132) \\ &\geq (x_2 - x_3)\Big(W_{n+1}\big(\tau^k(1), \tau^k(x_1), \tau^k(0)\big) \\ &\qquad - W_{n+1}\big(\tau^k(1), \tau^k(x_1), \tau^k(0)\big)\Big) = 0 && (133) \end{aligned}$$

The inequality in (132) follows from (127) - (129). The inequality in (133) follows from the inductive hypothesis of Lemma 2. □

E. Proof of Theorem 10

Theorem 10: For all t, $J_t(x_1, x_2)$ is convex in x_1 for fixed x_2, and is convex in x_2 for fixed x_1.

Proof. In order to prove convexity, a number of supporting results are required.

Lemma 5 (Linearity)**.** *$J_t^1(x_1, x_2)$ is linear in x_1 for fixed x_2, and similarly, $J_t^2(x_1, x_2)$ is linear in x_2 for fixed x_1.*

Proof. We will prove the first half of this lemma here, and the other half follows using exactly the same steps but switching channel 1 and 2. Let $0 \leq \lambda \leq 1$.

$$\begin{aligned} &J_t^1(\lambda x_1 + (1 - \lambda)y_1, x_2) \\ &\quad = -c + \lambda x_1 + (1 - \lambda)y_1 + x_2 - (\lambda x_1 + (1 - \lambda)y_1)x_2 \\ &\qquad + (\lambda x_1 + (1 - \lambda)y_1)J_{t+1}\big(1 - q, \tau(x_2)\big) \\ &\qquad + (1 - (\lambda x_1 + (1 - \lambda)y_1))J_{t+1}\big(p, \tau(x_2)\big) && (134) \\ &\quad = \lambda(-c + x1 - x_1x_2) + (1 - \lambda)(-c + y_1 - y_1x_2) \\ &\quad + \lambda\Big(x_1J_{t+1}\big(1 - q, \tau(x_2)\big) + (1 - x_1)J_{t+1}\big(p, \tau(x_2)\big)\Big) \\ &\quad + (1 - \lambda)\Big(y_1J_{t+1}\big(1 - q, \tau(x_2)\big) + (1 - y_1)J_{t+1}\big(p, \tau(x_2)\big)\Big) && (135) \\ &\quad = \lambda J_t^1(x_1, x_2) + (1 - \lambda)J_t^1(y_1, x_2) && (136) \end{aligned}$$

□

Lemma 6 (Commutativity)**.**

$$J_t(x_1, x_2) = J_t(x_2, x_1) \tag{137}$$

Proof. This proof is by reverse induction on t. For T, we have

$$J_T(x_1,x_2) = \max\Big\{ \max(x_1,x_2), -c + x_1 + x_2 - x_1x_2, -c + x_2 + x_1 - x_2x_1 \Big\} \tag{138}$$

$$= \max\Big\{ \max(x_2,x_1), -c + x_2 + x_1 - x_2x_1, -c + x_1 + x_2 - x_1x_2 \Big\} \tag{139}$$

$$= J_T(x_2,x_1) \tag{140}$$

Now assume (137) holds for time $t+1$. Then we have

$$J_t^1(x_1,x_2) = -c + x_1 + x_2 - x_1x_2 + x_1 J_{t+1}\big(1-q, \tau(x_2)\big) + (1-x_1) J_{t+1}\big(p, \tau(x_2)\big) \tag{141}$$

$$= -c + x_2 + x_1 - x_2x_1 + x_1 J_{t+1}\big(\tau(x_2), 1-q\big) + (1-x_1) J_{t+1}\big(\tau(x_2), p\big) \tag{142}$$

$$= J_t^2(x_2,x_1) \tag{143}$$

Additionally, we have

$$J_t^0(x_1,x_2) = \max(x_1,x_2) + J_{t+1}\big(\tau(x_1), \tau(x_2)\big) \tag{144}$$

$$= \max(x_2,x_1) + J_{t+1}\big(\tau(x_2), \tau(x_1)\big) \tag{145}$$

$$= J_t^0(x_2,x_1) \tag{146}$$

Finally, we can use these two results to show that

$$J_t(x_1,x_2) = \max\big\{ J_t^0(x_1,x_2), J_t^1(x_1,x_2), J_t^2(x_1,x_2) \big\} \tag{147}$$

$$= \max\big\{ J_t^0(x_2,x_1), J_t^2(x_2,x_1), J_t^1(x_2,x_1) \big\} \tag{148}$$

$$= J_t(x_2,x_1) \tag{149}$$

The proof follows by induction. □

Let $\Phi_t(0)$, $\Phi_t(1)$, $\Phi_t(2)$ be the sets of (x_1,x_2) such that it is optimal to not probe, probe channel 1, and probe channel 2 respectively at time t.

Lemma 7 (Probe Symmetry)**.** *If* $(x_1,x_2) \in \Phi_t(1)$*, then* $(x_2,x_1) \in \Phi_t(2)$.

Proof. If $(x_1,x_2) \in \Phi(1)_t$, then $J_t^1(x_1,x_2) \geq J_t^2(x_1,x_2)$ and $J_t^1(x_1,x_2) \geq J_t^0(x_1,x_2)$. Using Lemma 6, we can then say that $J_t^2(x_2,x_1) \geq J_t^1(x_2,x_1)$ and $J_t^2(x_2,x_1) \geq J_t^0(x_2,x_1)$ which implies $(x_2,x_1) \in \Phi_t(2)$. □

Lemma 8 (No-Probe Symmetry)**.** *If* $(x_1,x_2) \in \Phi_t(0)$*, then* $(x_2,x_1) \in \Phi_t(0)$.

Proof. If $(x_1,x_2) \in \Phi_t(0)$, then $J_t^0(x_1,x_2) \geq J_t^1(x_1,x_2)$ and $J_t^0(x_1,x_2) \geq J_t^2(x_1,x_2)$. It follows from Lemma 6 that $J_t^0(x_1,x_2) = J_t^0(x_2,x_1)$ and $J_t^1(x_1,x_2) = J_t^1(x_2,x_1)$ which implies $J_t^0(x_2,x_1) \geq J_t^1(x_2,x_1)$. By a similar argument, we can show $J_t^0(x_2,x_1) \geq J_t^2(x_2,x_1)$, and therefore $(x_2,x_1) \in \Phi_t(0)$. □

Lemmas (5)-(8) combine to prove a convexity result on the expected reward function. The proof follows by reverse induction over t. For $t = T$,

$$J_T(x_1,x_2) = \max\Big\{ \max(x_1,x_2), -c + x_1 + x_2 - x_1x_2, -c + x_2 + x_1 - x_2x_1 \Big\} \tag{150}$$

is convex in each element since each argument to the maximum is convex (or affine) and the maximum of convex functions is also convex. Now consider $t < T$, and we assume that $J_{t+1}(x_1,x_2)$ is convex in x_1 for fixed x_2. To begin with, we note that

$$\tau(\lambda x_1 + (1-\lambda)y_1) = (1-q)(\lambda x_1 + (1-\lambda)y_1) + p(1 - \lambda x_1 - (1-\lambda)y_1) \tag{151}$$

$$= (1-q)\lambda x_1 + p\lambda(1-x_1) + (1-q)(1-\lambda)y_1 + p(1-\lambda)(1-y_1) \tag{152}$$

$$= \lambda\tau(x_1) + (1-\lambda)\tau(y_1) \tag{153}$$

First we consider the expected throughput after not probing.

$$J_t^0(\lambda x_1 + (1-\lambda)y_1, x_2) = \max(\lambda x_1 + (1-\lambda)y_1, x_2) + J_{t+1}\big(\tau(\lambda x_1 + (1-\lambda)y_1), \tau(x_2)\big) \tag{154}$$

$$\leq \lambda(\max(x_1,x_2)) + (1-\lambda)(\max(y_1,x_2)) + J_{t+1}\big(\lambda\tau(x_1) + (1-\lambda)\tau(y_1), \tau(x_2)\big) \tag{155}$$

$$\leq \lambda(\max(x_1,x_2)) + (1-\lambda)(\max(y_1,x_2)) + \lambda J_{t+1}\big(\tau(x_1), \tau(x_2)\big) + (1-\lambda) J_{t+1}\big(\tau(y_1), \tau(x_2)\big) \tag{156}$$

$$= \lambda(J_t^0(x_1,x_2)) + (1-\lambda)(J_t^0(y_1,x_2)) \tag{157}$$

where (155) holds by the convexity of $\max(x,\cdot)$ and linearity of $\tau(\cdot)$, and (156) holds from the induction hypothesis. Additionally, $J_t^1(x_1,x_2)$ is convex in x_1 by lemma 5. For $J_t^2(x_1,x_2)$, we have:

$$J_t^2(\lambda x_1 + (1-\lambda)y_1, x_2) = -c + \lambda x_1 + (1-\lambda)y_1 + x_2 - (\lambda x_1 + (1-\lambda)y_1)x_2 + x_2 J_{t+1}\big(\tau(\lambda x_1 + (1-\lambda)y_1), 1-q\big) + (1-x_2) J_{t+1}\big(\tau(\lambda x_1 + (1-\lambda)y_1), p\big) \tag{158}$$

$$= \lambda(-c + x_1 + x_2 - x_1x_2) + (1-\lambda)(-c + y_1 + x_2 - y_1x_2) + x_2 J_{t+1}\big(\lambda\tau(x_1) + (1-\lambda)\tau(y_1), 1-q\big) + (1-x_2) J_{t+1}\big(\lambda\tau(x_1) + (1-\lambda)\tau(y_1), p\big) \tag{159}$$

$$
\begin{aligned}
&\leq \lambda(-c + x_1 + x_2 - x_1 x_2) \\
&\quad + (1-\lambda)(-c + y_1 + x_2 - y_1 x_2) \\
&\quad + \lambda\Big(x_2 J_{t+1}\big(\tau(x_1), 1-q\big) + (1-x_2) J_{t+1}\big(\tau(x_1), p\big)\Big) \\
&\quad + (1-\lambda)\Big(x_2 J_{t+1}\big(\tau(y_1), 1-q\big) \\
&\quad + (1-x_2) J_{t+1}\big(\tau(y_1), p\big)\Big) \qquad (160) \\
&= \lambda(J_t^2(x_1, x_2)) + (1-\lambda)(J_t^2(y_1, x_2)) \qquad (161)
\end{aligned}
$$

Thus, each of $J_t^0(x_1, x_2)$, $J_t^1(x_1, x_2)$, and $J_t^2(x_1, x_2)$ is convex in x_1 for fixed x_2, and therefore $J_t(x_1, x_2)$ is convex in x_1 as well. The second half of the proof statement holds by symmetry. □

REFERENCES

[1] M. Johnston, E. Modiano, and I. Keslassy, "Channel probing in communication systems: Myopic policies are not always optimal," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1934–1938.

[2] M. Johnston and E. Modiano, "Optimal channel probing in communication systems: The two-channel case," in *Global Communications (GLOBECOM), 2013 IEEE International Symposium on*. IEEE, 2013.

[3] K. Jagannathan, S. Mannor, I. Menache, and E. Modiano, "A state action frequency approach to throughput maximization over uncertain wireless channels," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011.

[4] S. Guha, K. Munagala, and S. Sarkar, "Jointly optimal transmission and probing strategies for multichannel wireless systems," in *Information Sciences and Systems, 2006 40th Annual Conference on*. IEEE, 2006.

[5] ——, "Optimizing transmission rate in wireless channels using adaptive probes," in *ACM SIGMETRICS Performance Evaluation Review*. ACM, 2006.

[6] P. Chaporkar and A. Proutiere, "Optimal joint probing and transmission strategy for maximizing throughput in wireless systems," *Selected Areas in Communications, IEEE Journal on*, 2008.

[7] N. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," in *International Conference on Mobile Computing and Networking: Proceedings of the 13 th annual ACM international conference on Mobile computing and networking*, 2007.

[8] A. Gopalan, C. Caramanis, and S. Shakkottai, "On wireless scheduling with partial channel-state information," in *Proc. Ann. Allerton Conf. Communication, Control and Computing*, 2007.

[9] K. Kar, X. Luo, and S. Sarkar, "Throughput-optimal scheduling in multichannel access point networks under infrequent channel measurements," *Wireless Communications, IEEE Transactions on*, 2008.

[10] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *Information Theory, IEEE Transactions on*, 2009.

[11] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance," *Wireless Communications, IEEE Transactions on*, 2008.

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

[13] N. B. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," *Networking, IEEE/ACM Transactions on*, vol. 17, no. 6, pp. 1805–1818, 2009.

[14] Y. Xu, J. Wang, Q. Wu, Z. Zhang, A. Anpalagan, and L. Shen, "Optimal energy-efficient channel exploration for opportunistic spectrum usage," *Wireless Communications Letters, IEEE*, vol. 1, no. 2, pp. 77–80, 2012.

[15] J. Tang and B. Krishnamachari, "Power allocation over two identical gilbert-elliott channels," *arXiv preprint arXiv:1203.6630*, 2012.

[16] L. Ying and S. Shakkottai, "On throughput optimality with delayed network-state information," *Information Theory, IEEE Transactions on*, 2011.

[17] R. Gallager and R. Gallager, *Discrete stochastic processes*. Kluwer Academic Publishers, 1996.

[18] M. R. Johnston, "The role of control information in wireless link scheduling," Ph.D. dissertation, Massachusetts Institute of Technology, 2014.