Power Allocation over Two Identical Gilbert-Elliott Channels

Junhua Tang
School of Electronic Information
and Electrical Engineering
Shanghai Jiao Tong University, China
Email: junhuatang@sjtu.edu.cn

Parisa Mansourifard
Ming Hsieh Department
of Electrical Engineering
Viterbi School of Engineering
University of Southern California
Email: parisama@usc.edu

Bhaskar Krishnamachari
Ming Hsieh Department
of Electrical Engineering
Viterbi School of Engineering
University of Southern California
Email: bkrishna@usc.edu

Abstract—**We study the problem of power allocation over two identical Gilbert-Elliot communication channels. Our goal is to maximize the expected discounted number of bits transmitted over an infinite time horizon. This is achieved by choosing among three possible strategies: (1) betting on channel 1 by allocating all the power to this channel, which results in high data rate if channel 1 happens to be in good state, and zero bits transmitted if channel 1 is in bad state (even if channel 2 is in good state) (2) betting on channel 2 by allocating all the power to the second channel, and (3) a balanced strategy whereby each channel is allocated half the total power, with the effect that each channel can transmit a low data rate if it is in good state. We assume that each channel's state is only revealed upon transmission of data on that channel. We model this problem as a partially observable Markov decision processes (MDP), and derive key threshold properties of the optimal policy. Further, we show that by formulating and solving a relevant linear program the thresholds can be determined numerically when system parameters are known.**

I. Introduction

Adaptive power control is an important technique to select the transmission power of a wireless system according to channel condition to achieve better network performance in terms of higher data rate or spectrum efficiency. While there has been some recent work on power allocation over stochastic channels [1], [2], the problem of optimal adaptive power allocation across multiple stochastic channels with memory is challenging and poorly understood. In this paper, we analyze a simple but fundamental problem. We consider a wireless system operating on two stochastically identical independent parallel transmission channels, each modeled as a slotted Gilber-Elliott channel (i.e. described by two-state Markov chains, with a bad state "0" and a good state "1"). Our objective is to allocate the limited power budget to the two channels dynamically so as to maximize the expected discounted number of bits transmitted over time. Since the channel state is unknown when power allocation decision is made, this problem is more challenging than it looks like.

Recently, several works have explored different sequential decision-making problems involving Gilbert-Elliott channels [3], [4], [5], [6], [7]. In [3],[4], the authors consider selecting one channel to sense/access at each time among several identical channels, formulate it as a restless multi-armed problem, and show that a simple myopic policy is optimal whenever the channels are positively correlated over time. In [5], the authors study the problem of dynamically choosing one of three transmitting schemes for a single Gilbert-Elliott channel in an attempt to maximize the expected discounted number of bits transmitted. And in [6], the authors study the problem of choosing a transmitting strategy from two choices emphasizing the case when the channel transition probabilities are unknown. While similar in spirit to these two studies, our work addresses a more challenging setting involving two independent channels. A more related two-channel problem is studied in [7], which characterizes the optimal policy to opportunistically access two non-identical Gilber-Elliott channels (generalizing the prior work on sensing policies for identical channels [3],[4]). While we address only identical channels in this work, the strategy space explored here is richer because in our formulation of power allocation, it is possible to use both channels simultaneously whilst in [3], [7] only one channel is accessed in each time slot.

In this paper, we formulate our power allocation problem as a partially observable Markov decision process (POMDP). We then treat the POMDP as a continuous state MDP and develop the structure of the optimal policy (decision). Our main contributions are the following: (1) we formulate the problem of dynamic power allocation over parallel Markovian channels, (2) using the MDP theory, we theoretically prove key threshold properties of the optimal policy for this particular problem, (3) through simulation based on linear programming, we demonstrate the existence of the 0-threshold and 2-threshold structures of the optimal policy, and (4) we demonstrate how to numerically compute the thresholds and construct the optimal policy when system parameters are known.

II. Problem Formulation

A. Channel model and assumptions

We consider a wireless communication system operating on two parallel channels. Each channel is described by a slotted Gilbert-Elliott model which is a one dimensional Markov chain $G_{i,t}(i \in \{1,2\}, t \in \{1,2,...,\infty\})$ with two states: a good state denoted by 1 and a bad state denoted by 0 (i is the channel number and t is the time slot). The channel transition probabilities are given by $Pr[G_{i,t} = 1|G_{i,t-1} = 1] = \lambda_1, i \in \{1, 2\}$ and $Pr[G_{i,t} = 1|G_{i,t-1} = 0] = \lambda_0, i \in \{1, 2\}$. We assume the two channels are identical and independent of each other, and channel transitions occur at the beginning of each time slot. We also assume that $\lambda_0 \leq \lambda_1$, which is the positive correlation assumption commonly used in the literature.

The system has a total transmission power of P. At the beginning of time slot t, the system allocates transmission power $P_1(t)$ to channel 1 and $P_2(t)$ to channel 2, where $P_1(t) + P_2(t) = P$. We assume the channel state is not directly observable at the beginning of each time slot. That is, the system needs to allocate the transmission power to the two parallel channels without knowing the channel states. If channel $i(i \in \{1, 2\})$ is used at time slot t by allocating transmission power $P_i(t)$ on it, the channel state of the elapsed slot is revealed at the end of the time slot through channel feedback. But if a channel is not used, that is, if transmission power is 0 on that channel, the channel state of the elapsed slot remains unknown at the end of that slot.

B. Power allocation strategies

To simplify the problem, we assume the system may allocate one of the following three power levels to a channel: $0, P/2$, or P. That is, based on the belief in the channel state of channel i for the current time slot t, the system may decide to give up the channel ($P_i(t) = 0$), use it moderately ($P_i(t) = P/2$) or use it fully($P_i(t) = P$). Since the channel state is not directly observable when the power allocation is done, the following circumstances may occur. If a channel is in bad state, no data is transmitted at all no matter what the allocated power is. If a channel is in good state, and power $P/2$ is allocated to it, it can transmit R_l bits of data successfully during that slot. If a channel is in good condition and power P is allocated to it, it can transmit R_h bits of data successfully during that slot. We assume $R_l < R_h < 2R_l$.

We define three power allocation strategies(actions): balanced, betting on channel 1, and betting on channel 2. Each strategy is explained in detail as follows.

Balanced: For this action (denoted by B_b), the system allocates the transmission power evenly on both channels, that is, $P_1(t) = P_2(t) = P/2$, for time slot t. This corresponds to the situation when the system cannot determine which of the channels is more likely to be in good state, so it decides to "play safe" by using both of the channels.

Betting on channel 1: For this action (denoted by B_1), the system decides to "gamble" and allocate all the transmission power to channel 1. That is, $P_1(t) = P, P_2(t) = 0$ for time slot t. This corresponds to the situation when the system believes that channel 1 is in a good state and channel 2 is in a bad state.

Betting on channel 2: For this action (denoted by B_2), the system put all the transmission power in channel 2, that is, $P_2(t) = P, P_1(t) = 0$ for time slot t.

Note that for strategies B_1 and B_2, if a channel is not used, the system (transmitter) will not acquire any knowledge about the state of that channel during the elapsed slot.

C. POMDP formulation

At the beginning of a time slot, the system is confronted with a choice among three actions. It must judiciously select actions so as to maximize the total expected discounted number of bits transmitted over an infinite time span. Because the state of the channels is not directly observable, the problem in hand is a Partially Observable Markov Decision Process (POMDP). In [8], it is shown that a sufficient statistic for determining the optimal policy is the conditional probability that the channel is in the good state at the beginning of the current slot given the past history (henceforth called belief) [5]. Denote the belief of the system by a two dimension vector $\mathbf{x}_t$=$(x_{1,t}, x_{2,t})$, where $x_{1,t} = \Pr[G_{1,t} = 1|\hbar_t]$, $x_{2,t} = \Pr[G_{2,t} = 1|\hbar_t]$, where $\hbar_t$ is all the history of actions and observations at the current slot t. By using this belief as the decision variable, the POMDP problem is converted into an MDP with the uncountable state space $([0, 1], [0, 1])$ [5].

Define a policy π as a rule that dictates the action to choose, *i.e.*, a map from the belief at a particular time to an action in the action space. Let $V^\pi(\mathbf{p})$ be the expected discounted reward with initial belief $\mathbf{p} = (p_1, p_2)$, that is, $x_{1,0} = \Pr[G_{1,0} = 1|\hbar_0] = p_1$, $x_{2,0} = \Pr[G_{2,0} = 1|\hbar_0] = p_2$, where the superscript π denotes the policy being followed. Define $\beta(\in [0, 1))$ as the discount factor, the expected discounted reward has the following expression

$$V^\pi(\mathbf{p}) = E_\pi[\sum_{t=0}^{\infty} \beta^t g_{a_t}(\mathbf{x}_t)|\mathbf{x}_0 = \mathbf{p}], \tag{1}$$

where E_π represents the expectation given that the policy π is employed, t is the time slot index, a_t is the action chosen at time t, $a_t \in \{B_b, B_1, B_2\}$. The term $g_{a_t}(\mathbf{x}_t)$ denotes the expected reward acquired when the belief is $\mathbf{x_t}$ and the action a_t is chosen:

$$g_{a_t}(\mathbf{x}_t) = \begin{cases} x_{1,t}R_l + x_{2,t}R_l, & \text{if} \quad a_t = B_b \\ x_{1,t}R_h, & \text{if} \quad a_t = B_1 \\ x_{2,t}R_h, & \text{if} \quad a_t = B_2 \end{cases}. \tag{2}$$

Now we define the value function $V(\mathbf{p})$ as

$$V(\mathbf{p}) = \max_\pi V^\pi(\mathbf{p}), \quad \text{for all} \quad \mathbf{p} \in ([0, 1], [0, 1]). \tag{3}$$

A policy is said to be stationary if it is a function mapping the state space $([0, 1], [0, 1])$ into the action space $\{B_b, B_1, B_2\}$. Ross proved in [9] (Th.6.3) that there exists a stationary policy π^* such that $V(\mathbf{p}) = V^{\pi^*}(\mathbf{p})$. The value function $V(\mathbf{p})$ satisfies the Bellman equation

$$V(\mathbf{p}) = \max_{a \in \{B_b, B_1, B_2\}} \{V_a(\mathbf{p})\}, \tag{4}$$

where $V_a(\mathbf{p})$ is the value acquired by taking action a when the initial belief is $\mathbf{p}$. $V_a(\mathbf{p})$ is given by

$$V_a(\mathbf{p}) = g_a(\mathbf{p}) + \beta E^{\mathbf{y}}[V(\mathbf{y})|\mathbf{x}_0 = \mathbf{p}, a_0 = a], \tag{5}$$

where $\mathbf{y}$ denotes the next belief when the action a is chosen and the initial belief is $\mathbf{p}$. The term $V_a(\mathbf{p})$ is explained next for the three possible actions.

a) Balanced (action B_b): If this action is taken, and the current belief is $\mathbf{p} = (p_1, p_2)$, the immediate reward is $p_1 R_l + p_2 R_l$. Since both channels are used, the channel quality of both channels during the current slot is then revealed to the transmitter. With probability p_1 the first channel will be in good state and hence the belief of channel 1 at the beginning of the next slot will be λ_1. Likewise, with probability $1 - p_1$ channel 1 will turn out to be in bad state and hence the updated belief of channel 1 for the next slot is λ_0. Since channel 2 and channel 1 are identical, channel 2 has similar belief update. Consequently if action B_b is taken, the value function evolves as

$$\begin{aligned}
& V_{B_b}(p_1, p_2) \\
= \;& p_1 R_l + p_2 R_l + \beta[(1-p_1)(1-p_2)V(\lambda_0, \lambda_0) \\
+ \;& p_1(1-p_2)V(\lambda_1, \lambda_0) + (1-p_1)p_2 V(\lambda_0, \lambda_1) \\
+ \;& p_1 p_2 V(\lambda_1, \lambda_1)]. \qquad (6)
\end{aligned}$$

b) Betting on channel 1(action B_1): If this action is taken, and the current belief is $\mathbf{p} = (p_1, p_2)$, the immediate reward is $p_1 R_h$. But since channel 2 is not used, its channel state remains unknown. Hence if the belief of channel 2 during the elapsed time slot is p_2, its belief at the beginning of the next time slot is given by

$$T(p_2) = p_2\lambda_1 + (1-p_2)\lambda_0 = \alpha p_2 + \lambda_0, \qquad (7)$$

where $\alpha = \lambda_1 - \lambda_0$. Consequently, if this action is taken, the value function evolves as

$$\begin{aligned}
& V_{B_1}(p_1, p_2) = p_1 R_h + \\
& \beta[(1-p_1)V(\lambda_0, T(p_2)) + p_1 V(\lambda_1, T(p_2))]. \qquad (8)
\end{aligned}$$

c) Betting on channel 2(action B_2): Similar to action B_1, if action B_2 is taken, the value function evolves as

$$\begin{aligned}
& V_{B_2}(p_1, p_2) = p_2 R_h + \\
& \beta[(1-p_2)V(T(p_1), \lambda_0) + p_2 V(T(p_1), \lambda_1)], \qquad (9)
\end{aligned}$$

where

$$T(p_1) = p_1\lambda_1 + (1-p_1)\lambda_0 = \alpha p_1 + \lambda_0. \qquad (10)$$

Finally the Bellman equation for our power allocation problem reads as follows

$$V(\mathbf{p}) = \max\{V_{B_b}(\mathbf{p}), V_{B_1}(\mathbf{p}), V_{B_2}(\mathbf{p})\}. \qquad (11)$$

III. Structure of the Optimal Policy

From the above discussion we understand that an optimal policy exists for our power allocation problem. In this section, we try to derive the optimal policy by first looking at the features of its structure.

A: Properties of value function

Lemma 1. *$V_{B_i}(p_1, p_2), i \in \{1, 2, b\}$ is affine with respect to p_1 and p_2 and the following equalities hold:*

$$\begin{aligned}
& V_{B_i}(cp + (1-c)p', p_2) = cV_{B_i}(p, p_2) + (1-c)V_{B_i}(p', p_2), \\
& V_{B_i}(p_1, cp + (1-c)p') = cV_{B_i}(p_1, p) + (1-c)V_{B_i}(p_1, p'), \; (12)
\end{aligned}$$

where $0 \le c \le 1$ is a constant; and we say $f(x)$ is affine with respect to x if $f(x) = a + cx$, with constant a and c.

Refer to [10] for proof.

Lemma 2. *$V_{B_i}(p_1, p_2), i \in \{1, 2, b\}$ is convex in p_1 and p_2.*

Proof: The convexity of $V_{B_i}, i \in \{1, 2, b\}$ in p_1 and p_2 follows from its affine linearity in Lemma 1. ■

Lemma 3. *$V(p_1, p_2) = V(p_2, p_1)$, that is, $V(p_1, p_2)$ is symmetric with respect to the line $p_1 = p_2$ in the belief space.*

Refer to [10] for proof.

B: Properties of the decision regions of policy $\pi*$

We use Φ_a to denote the set of beliefs for which it is optimal to take the action a. That is,

$$\Phi_a = \{(p_1, p_2) \in ([0,1],[0,1]), V(p_1, p_2) = V_a(p_1, p_2)\}, \\ a \in \{B_b, B_1, B_2\}. \; (13)$$

Definition 1. *Φ_a is said to be contiguous along p_1 dimension if we have $(x_1, p_2) \in \Phi_a$ and $(x_2, p_2) \in \Phi_a$, then $\forall x \in [x_1, x_2]$, we have $(x, p_2) \in \Phi_a$. Similarly, we say Φ_a is contiguous along p_2 dimension if we have $(p_1, y_1) \in \Phi_a$ and $(p_1, y_2) \in \Phi_a$, then $\forall y \in [y_1, y_2]$, we have $(p_1, y) \in \Phi_a$.*

Theorem 1. *Φ_{B_b} is contiguous in both p_1 and p_2 dimensions. Φ_{B_1} is contiguous in p_1 dimension, and Φ_{B_2} is contiguous in p_2 dimension.*

Refer to [10] for proof.

Theorem 2. *If belief (p_1, p_2) is in Φ_{B_1}, then belief (p_2, p_1) is in Φ_{B_2}. In other words, the decision regions of B_1 and B_2 are mirrors with respect to the line $p_1 = p_2$ in the belief space.*

Refer to [10] for proof.

Theorem 3. *If belief (p_1, p_2) is in Φ_{B_b}, then belief (p_2, p_1) is in Φ_{B_b}. That is, the decision region of B_b is symmetric with respect to the line $p_1 = p_2$ in the belief region.*

Refer to [10] for proof.

Lemma 4. *After each channel is used once, the belief state is the four sides of a rectangle determined by four vertices at $(\lambda_0, \lambda_0), (\lambda_0, \lambda_1), (\lambda_1, \lambda_0), (\lambda_1, \lambda_1)$ (Figure 1 (a)).*

Refer to [10] for proof.

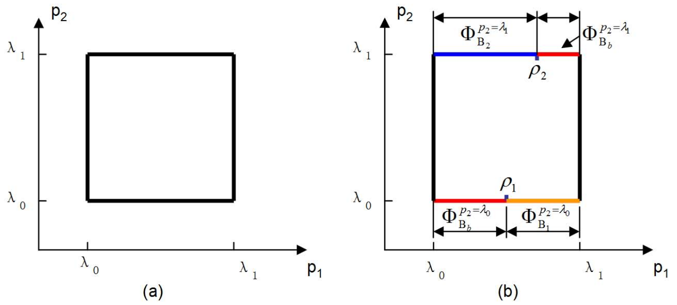
Theorem 4. *Let $p_1 \in [\lambda_0, \lambda_1]$, $p_2 = \lambda_0$, there exists a threshold $\rho_1 (\lambda_0 \le \rho_1 \le \lambda_1)$ such that $\forall p_1 \in [\lambda_0, \rho_1], (p_1, \lambda_0) \in \Phi_{B_b}$. (Figure 1(b))*

Refer to [10] for proof.

Theorem 5. *Let $p_1 \in [\lambda_0, \lambda_1]$, $p_2 = \lambda_1$, there exists a threshold $\rho_2 (\lambda_0 \le \rho_2 \le \lambda_1)$ such that $\forall p_1 \in [\rho_2, \lambda_1], (p_1, \lambda_1) \in \Phi_{B_b}$. (Figure 1(b))*

Refer to [10] for proof.

Lemma 5. *In case of $p_2 = \lambda_0$, it is not optimal to take action B_2. In case of $p_2 = \lambda_1$, it is not optimal to take action B_1.*



p2
λ 1
λ 0
λ 0
λ 1
p1
(a)
p2
$\Phi_{B_2}^{p_2=\lambda_1}$
$\Phi_{B_b}^{p_2=\lambda_1}$
λ 1
ρ_2
ρ_1
λ 0
$\Phi_{B_b}^{p_2=\lambda_0}$
$\Phi_{B_1}^{p_2=\lambda_0}$
λ 0
λ 1
p1
(b)

Fig. 1. (a) The feasible belief space. (b) The threshold on p_1 ($p_2 = \lambda_0(\lambda_1)$).

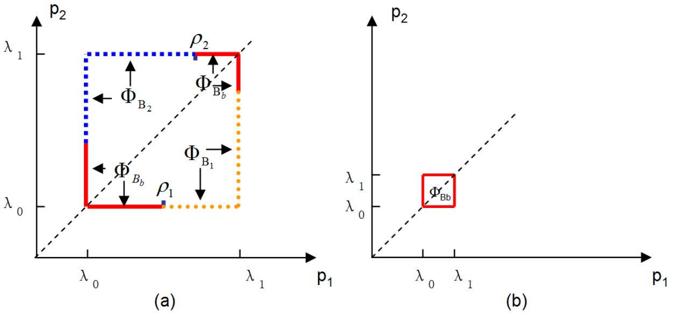


Fig. 2. Structure of optimal policy.

Refer to [10] for proof.

C: The structure of the optimal policy

Theorem 6. *The optimal policy has a simple threshold structure and can be described as follows (Figure 2):*

$$\pi^*(p_1,\lambda_0) = \begin{cases} B_b, & \textit{if} \quad \lambda_0 \le p_1 \le \rho_1 \\ B_1, & \textit{if} \quad \rho_1 < p_1 \le \lambda_1 \end{cases}, \qquad (a)$$

$$\pi^*(p_1,\lambda_1) = \begin{cases} B_b, & \textit{if} \quad \rho_2 \le p_1 \le \lambda_1 \\ B_2, & \textit{if} \quad \lambda_0 \le p_1 < \rho_2 \end{cases}, \qquad (b)$$

$$\pi^*(\lambda_0,p_2) = \begin{cases} B_b, & \textit{if} \quad \lambda_0 \le p_2 \le \rho_1 \\ B_2, & \textit{if} \quad \rho_1 < p_2 \le \lambda_1 \end{cases}, \qquad (c)$$

$$\pi^*(\lambda_1,p_2) = \begin{cases} B_b, & \textit{if} \quad \rho_2 \le p_2 \le \lambda_1 \\ B_1, & \textit{if} \quad \lambda_0 \ge p_2 < \rho_1 \end{cases}. \qquad (d)$$

$$(14)$$

Refer to [10] for proof.

From the above analysis we understand that the optimal policy has a simple threshold structure. And it is critical to find the two thresholds ρ_1 and ρ_2.

Theorem 7. *Let* $\delta_{i,j}(k_1,k_2) = V_{B_i}(k_1,k_2) - V_{B_j}(k_1,k_2), (i \in \{1,2,b\}, j \in \{1,2,b\})$, ρ_1 *can be calculated as follows*

1) if $T(\lambda_0) < \rho_2$, $T(\lambda_0) \le \rho_1$

$$\rho_1 = \frac{\lambda_0 R_l + \beta\lambda_0\delta_{2,b}(\lambda_0,\lambda_1)}{R_h - R_l + \beta\lambda_0(\delta_{1,b}(\lambda_1,\lambda_1) + \delta_{2,b}(\lambda_0,\lambda_1))}, \qquad (15)$$

2) if $T(\lambda_0) < \rho_2$, $T(\lambda_0) > \rho_1$

$$\rho_1 = \frac{\lambda_0 R_l + \beta(1-\lambda_0)\delta_{b,2}(\lambda_0,\lambda_0)}{R_h - R_l + \beta\lambda_0\delta_{1,b}(\lambda_1,\lambda_1) + \beta(1-\lambda_0)\delta_{b,2}(\lambda_0,\lambda_0)}, \qquad (16)$$

3) if $T(\lambda_0) \ge \rho_2$, $T(\lambda_0) \le \rho_1$

$$\rho_1 = \frac{\lambda_0 R_l + \beta\lambda_0\delta_{2,b}(\lambda_0,\lambda_1)}{R_h - R_l + \beta\lambda_0\delta_{2,b}(\lambda_0,\lambda_1) + \beta(1-\lambda_0)\delta_{b,1}(\lambda_1,\lambda_0)}, \qquad (17)$$

4) if $T(\lambda_0) \ge \rho_2$, $T(\lambda_0) > \rho_1$, ρ_1 *is calculated in (18).*

Refer to [10] for proof.

Theorem 8. *Let* $\delta_{i,j}(k_1,k_2) = V_{B_i}(k_1,k_2) - V_{B_j}(k_1,k_2), (i \in \{1,2,b\}, j \in \{1,2,b\})$, *the threshold* ρ_2 *is calculated as follows*

1) if $T(\rho_2) \ge \rho_2$ *and* $T(\rho_2) > \rho_1$

$$\rho_2 = \frac{\lambda_1(R_h - R_l) - \beta\lambda_1\delta_{2,b}(\lambda_0,\lambda_1) - \beta(1-\lambda_1)\delta_{b,1}(\lambda_0,\lambda_0)}{R_l - \beta\lambda_1\delta_{2,b}(\lambda_0,\lambda_1) - \beta(1-\lambda_1)\delta_{b,1}(\lambda_0,\lambda_0)}, \qquad (19)$$

2) if $T(\rho_2) \ge \rho_2$ *and* $T(\rho_2) \le \rho_1$

$$\rho_2 = \frac{\lambda_1(R_h - R_l) - \beta\lambda_1\delta_{2,b}(\lambda_0,\lambda_1))}{R_l - \beta\lambda_1\delta_{2,b}(\lambda_0,\lambda_1) - \beta(1-\lambda_1)\delta_{b,1}(\lambda_1,\lambda_0)}, \qquad (20)$$

3) if $T(\rho_2) < \rho_2$, $T(\rho_2) > \rho_1$

$$\rho_2 = \frac{\lambda_1(R_h - R_l) - \beta(1-\lambda_1)\delta_{b,1}(\lambda_0,\lambda_0))}{R_l - \beta\lambda_1\delta_{2,b}(\lambda_1,\lambda_1) - \beta(1-\lambda_1)\delta_{b,1}(\lambda_0,\lambda_0)}, \qquad (21)$$

4) if $T(\rho_2) < \rho_2$, $T(\rho_2) \le \rho_1$

$$\rho_2 = \frac{\lambda_1(R_h - R_l)}{R_l - \beta\lambda_1\delta_{2,b}(\lambda_1,\lambda_1) - \beta(1-\lambda_1)\delta_{b,1}(\lambda_1,\lambda_0)}. \qquad (22)$$

The proof of this theorem is similar to that of theorem 7 and is omitted here.

IV. Simulation based on linear programming

Linear programming is one of the approaches to solve the Bellman's equation in (4). Based on [11], we model our problem as the following linear program:

$$\min \sum_{\mathbf{p}\in\mathbb{X}} V(\mathbf{p}),$$
$$\text{s.t.} \quad g_a(\mathbf{p}) \; + \; \beta\sum_{\mathbf{y}\in\mathbb{X}} f_a(\mathbf{p},\mathbf{y})V(\mathbf{y}) \le V(\mathbf{p}),$$
$$\forall\mathbf{p}\in\mathbb{X}, \forall a \in \mathbb{A}_{\mathbf{p}}, \qquad (23)$$

where $\mathbb{X}$ is the space of belief state, $\mathbb{A}_{\mathbf{p}}$ is the set of available actions for state $\mathbf{p}$. The state transition probabilities $f_a(\mathbf{p},\mathbf{y})$ is the probability that the next state will be $\mathbf{y}$ given that the current state is $\mathbf{p}$ and the current action is $a \in \mathbb{A}_{\mathbf{p}}$. The optimal policy can be generated according to

$$\pi(\mathbf{p}) = \arg\max_{a\in\mathbb{A}_{\mathbf{p}}}(g_a(\mathbf{p}) + \beta\sum_{\mathbf{y}\in\mathbb{X}} f_a(\mathbf{p},\mathbf{y})V(\mathbf{y})). \qquad (24)$$

We used the LOQO solver on NEOS Server [12] with AMPL input [13] to obtain the solution of equation (23). Then we used MATLAB to construct the policy according to equation (24).

Figure 3 shows the AMPL solution of value function for the following set of parameters: $\lambda_0 = 0.1, \lambda_1 = 0.9, \beta = 0.9, R_l = 2, R_h = 3$. The corresponding optimal policy is

$$\rho_1 = \frac{\lambda_0 R_l + \beta\lambda_0\delta_{2,1}(\lambda_0,\lambda_1) + \beta(1-\lambda_0)\delta_{b,1}(\lambda_0,\lambda_0)}{R_h - R_l + \beta\lambda_0\delta_{2,1}(\lambda_0,\lambda_1) + \beta(1-\lambda_0)(\delta_{b,1}(\lambda_1,\lambda_0) + \delta_{b,1}(\lambda_0,\lambda_0))}. \tag{18}$$

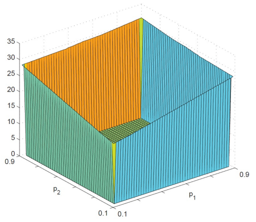


Fig. 3. Value function.

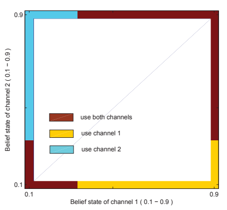


Fig. 4. Optimal policy.

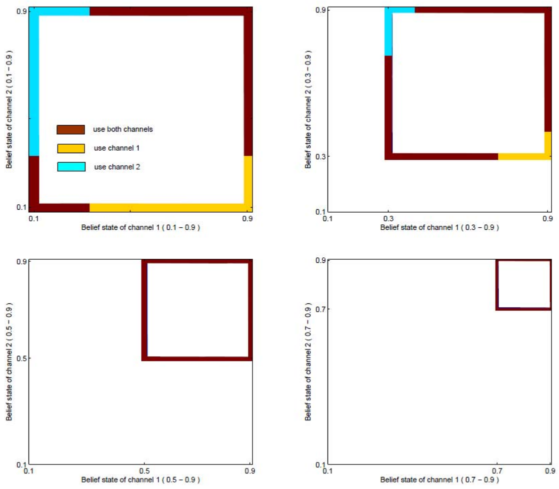


Fig. 5. Optimal policy with increasing λ_0 ($R_l = 2$, $R_h = 3$) .

shown in Figure 4. The structure of the policy in Figure 4 clearly shows the properties we gave in Theorems 1 to 5.

In order to observe the effect of parameters λ_0, λ_1, R_l and R_h on the structure of optimal policy, we have conducted simulation experiments for varying parameters. Figure 5 shows the policy structure when λ_0 varies from 0.1 to 0.7, while the rest of the parameters remain the same as in the above experiment. We can observe in Figure 5 that when λ_0 increases from 0.1 to 0.3, the decision region of action B_b occupies a bigger part of the belief space. Whilst when λ_0 is 0.5 or greater, the whole belief space falls in the decision region of action B_b, meaning that it is optimal to always use both channels in the set of this experiment when $\lambda_0 > 0.5$. Due to space limit, we cannot present more simulation results here, but the structure types in Theorem 6 is clearly observed in Figure 5.

V. Conclusion

In this paper we have shown the structure of the optimal policy by theoretical analysis and simulation. Knowing that this problem has a 0 or 2 threshold structure reduces the problem of identifying optimal performance to finding the (only up to 2) threshold parameters. In settings where the underlying state transition matrices are unknown, this could be exploited by using a multiarmed bandit (MAB) formulation to find the best possible thresholds (similar to the ideas in the papers [6] and [7]). Also, we would like to investigate the case of non-identical channels, and derive useful results for more than 2 channels, possibly in the form of computing the Whittle index [14], if computing the optimal policy in general turns out to be intractable.

Acknowledgment

The authors would like to thank Yi Gai and Yanting Wu at USC for their helpful discussions. This work is partially supported by National Key Basic Research Program of China under grant 2013CB329603, Natural Science Foundation of China under grant 61071081 and 60932003. This research was also sponsored in part by the U.S. Army Research Laboratory under the Network Science Collaborative Technology Alliance, Agreement Number W911NF-09-2-0053, and by the Okawa Foundation, under an Award to support research on “Network Protocols that Learn”.

References

[1] X. Wang, D. Wang, H. Zhuang, and S. D. Morgera, “Fair energy-efficient resource allocation in wireless sensor networks over fading TDMA channels,” *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 28, no. 7, pp. 1063–1072, 2010.

[2] Y. Gai and B. Krishnamachari, “Online learning algorithms for stochastic water-filling,” in *Information Theory and Applications Workshop (ITA 2012)*, 2012.

[3] Q. Zhao, B. Krishnamachari, and K. Liu, “On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431–5440, 2008.

[4] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, “Optimality of myopic sensing in multi-channel opportunistic access,” *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4040–4050, 2009.

[5] A. Laourine and L. Tong, “Betting on gilbert-elliot channels,” *IEEE Transactions on Wireless communications*, vol. 9, pp. 723–733, February 2010.

[6] Y. Wu and B. Krishnamachari, “Online learning to optimize transmission over unknown gilbert-elliot channel,” in *WiOpt*, 2012.

[7] N. Nayyar, Y. Gai, and B. Krishnamachari, “On a restless multi-armed bandit problem with non-identical arms,” in *Allerton*, 2011.

[8] R. D. Smallwood and E. J. Sondik, “The optimal control of partially observable markov processes over a finite horizon,” *Operations Research*, vol. 21, pp. 1071–1088, September-October 1973.

[9] S. M. Ross, *Applied Probability Models with Optimization Applications*. San Francisco: Holden-Day, 1970.

[10] J. Tang, P. Mansourifard, and B. Krishnamachari, “Power allocation over two identical gilbert-elliott channels,” in *http://arxiv.org/abs/1203.6630*, 2012.

[11] D. P. D. Farias and B. V. Roy, “The linear programming approach to approximate dynamic programming,” *Operations Research*, vol. 51, pp. 850 – 865, November-December 2002.

[12] “NEOS server for optimization.” http://neos.mcs.anl.gov/neos/.

[13] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Modeling Language for Mathematical Programming*. Brooks/Cole Publishing Company, 2002.

[14] P. Whittle, “Multiarmed bandits and the gittins index,” *Journal of the Royal Statistical Society*, vol. 42, no. 2, pp. 143–149, 1980.