Online Learning to Optimize Transmission over an Unknown Gilbert-Elliott Channel

Yanting Wu, Bhaskar Krishnamachari

Online Learning to Optimize Transmission over an Unknown Gilbert-Elliott Channel

Yanting Wu
Dept. of Electrical Engineering
University of Southern California
Email: yantingw@usc.edu

Bhaskar Krishnamachari
Dept. of Electrical Engineering
University of Southern California
Email: bkrishna@usc.edu

***Abstract*—This paper studies the optimal transmission policy for a Gilbert-Elliott Channel. The transmitter has two actions: sending aggressively or sending conservatively, with rewards depending on the action chosen and the underlying channel state. The aim is to compute the scheduling policy to determine which actions to choose at each time slot in order to maximize the expected total discounted reward. We first establish the threshold structure of the optimal policy when the underlying channel statistics are known. We then consider the more challenging case when the statistics are unknown. For this problem, we map different threshold policies to arms of a suitably defined multi-armed bandit problem. To tractably handle the complexity introduced by countably infinite arms and the infinite time horizon, we weaken our objective a little: finding a $(OPT-(\epsilon+\delta))$-approximate policy instead. We present the UCB-P algorithm, which can achieve this objective with logarithmic-time regret.**

I. Introduction

Communication over the wireless channels are affected by fading conditions, interference, path loss, etc. To gain a better utilization of wireless channels, a transmitter needs to adapt transmission parameters such as data rate and transmission power according to the communication channel states.

In this paper, we analyze mathematically a communication system operating over a 2-state Markov channel (known as the Gilbert-Elliott Channel) in a time-slotted fashion. The objective is for the transmitter to decide at each time, based on prior observations, whether to send data at an aggressive or conservative rate. The former incurs the risk of failure, but reveals the channels true state, while the latter is a safe but unrevealing choice. When the channel transition probabilities are known, this problem can be modelled as a Partially Observable Markov Decision Process (POMDP). This formulation is very closely related to a recent work pertaining to betting on Gilbert Elliott Channels [3], which considers three choices, and shows that a threshold-type policy consisting of one, two, or three thresholds depending on the parameters, is optimal. In our setting, we show that the optimal policy always has a single threshold that corresponds to a K-conservative policy, in which the transmitter adopts the conservative approach for K steps after each failure before reattempting the aggressive strategy. Unlike [3], however, our focus is on the case when the underlying state transition matrix is unknown. In this case, the problem of finding the optimal strategy is equivalent to finding the optimal choice of K. We map the problem to a Multi-armed bandit, where each possible K-conservative policy corresponds to an arm. To deal with the difficulties of optimizing the discounted cost over an infinite horizon, and the countably infinite arms that result from this mapping, we introduce approximation parameters δ, ϵ, and show that a modification of the well-known UCB1 policy guarantees that the number of times that the arms that are more than $(\epsilon+\delta))$ away from the optimal are played is bounded by a logarithmic function of time. In other words, we show that the time-averaged regret with respect to a $(OPT-(\epsilon+\delta))$ policy tends to zero.

We briefly review some other recent papers in the literature that have treated similar problems. Johnston and Krishnamurthy [5] consider the problem of minimizing the transmission energy and latency associated with transmitting a file across a Gilbert Elliott fading channel, formulate it as a POMDP, identify a threshold policy for it, and analyze it for various parameter settings. Karmokar *et al.* [6] consider optimizing multiple objectives (transmission power, delay, and drop probability) for packet scheduling across a more general finite-state Markov channel. Motivated by opportunistic spectrum sensing, several recent studies have explored optimizing sensing and access decisions over multiple independent but stochastically identical parallel Gilbert Elliott channels, in which the objective is to select one channel at each time, showing that a simple myopic policy is optimal [7], [8]. In [9], Dai *et al.* consider the non-Bayesian version of the sensing problem where the channel parameters are unknown, and show that when the problem has a finite-option structure, online learning algorithms can be designed to achieve near-logarithmic regret. In [10], Nayyar *et al.* consider the same sensing problem over two non-identical channels, derive the optimal threshold structure policy for it. For the non-Bayesian setting, show a mapping to countably infinite-armed multi-armed bandit, and also prove logarithmic regret with respect to a $(OPT-\delta)$ policy, similar to the approach adopted in this work for a different formulation.

The paper is organized as follows: section II introduces the model; section III gives the structure of the optimal policy when the underlying channel transition probabilities are known; section IV talks about K-conservative policies and we prove that the optimal policy corresponds to a K-conservative policy; section V discusses using multi-arm bandits to find out

978-3-901882-45-6/2012 - Copyright is with IFIP

optimal K; to address the two challenges: infinite number of arms, and infinite time horizon, we weaken our objective to find policies which are at most $(\epsilon+\delta)$ away from the optimal instead. We design the UCB-P algorithm to learn such policies. We present simulation results in section VI. Finally, section VII concludes the paper.

II. Model

For our problem setting, we consider the Gilbert-Elliott channel which is a Markov chain with two states: good (denoted by 1) or bad (denoted by 0). If the channel is good, it allows the transmitter to send data with a high rate successfully. However, if the channel is bad, it only allows transmitter to send data with a low rate successfully. The transition probabilities matrix is given as:

$$P = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} = \begin{bmatrix} 1-\lambda_0 & \lambda_0 \\ 1-\lambda_1 & \lambda_1 \end{bmatrix}. \tag{1}$$

Define $\alpha = \lambda_1 - \lambda_0$. We assume that the channel is positive correlated, which means $\alpha \geq 0$.

At the beginning of each time slot, the transmitter chooses one of the following two actions:

- Sending Conservatively (SC): the transmitter sends data with a low rate. No matter what the channel state is, it can successfully transmit a small number of bits. We assign a reward R_1 to this action. Since the transmission is always successful, the transmitter cannot learn the state if this action is chosen.
- Sending Aggressively (SA): the transmitter sends data with a high rate. If the channel is in good state, we consider the transmission successful and the transmitter can get a high reward $R_2(> R_1)$. If the channel is in bad state, sending with a high rate will cause high error rate and drop rate, we consider the transmission a failure and the transmitter gets a constant penalty C. We assume if the transmitter sends aggressively, it can learn the state of the channel. In other words, we assume that when the channel is in a bad state an aggressive transmission strategy will encounter and detect failure.

Because when sending conservatively, the state of the channel is not directly observable, the problem we consider in this paper turns out to be a Partially Observable Markov Decision Process (POMDP) problem. In [1], it has been shown that a sufficient statistic to make an optimal decision for this POMDP problem is the conditional probability that the channel is in state 1 given all past actions and observations. We call this conditional probability the belief, represented by $b_t = Pr\left[S_t = 1 | H_t\right]$ which H_t is the history of all actions and observations before t^{th} time slot. When sending aggressively, the transmitter learns the state of the channel; so the belief is λ_0 if the channel is bad or λ_1 if the channel is good.

The following is the expression for the expected rewards:

$$R(b_t, A_t) = \begin{cases} R_1 & \text{if } A_t = SC, \\ b_t R_2 - (1-b_t)C & \text{if } A_t = SA, \end{cases} \tag{2}$$

where b_t is the belief of the channel in good state and A_t is the action taken by the transmitter at time t.

In this paper, we consider the expected total-discounted reward to make decisions, which is defined as

$$E\left[\sum_{t=0}^{\infty} \beta^t R(b_t, A_t) | b_0 = p\right], \tag{3}$$

where $R(b_t, A_t)$ is the expected reward at t, $\beta(< 1)$ is a constant discounter factor, and b_0 is the initial belief.

III. Threshold Structure of the Optimal Policy

In this section, we discuss the optimal policy when the transition probabilities are known. We prove that the optimal policy is a one threshold policy and give a closed form formulation of the threshold.

Policy, denoted by π, is defined as a rule which maps belief probabilities to actions. We use $V^\pi(p)$ to represent the expected total-discounted reward the transmitter can get given the initial belief is p and policy is π:

$$V^\pi(p) = E_\pi\left[\sum_{t=0}^{\infty} \beta^t R(b_t, A_t) | b_0 = p\right]. \tag{4}$$

The aim is to find a policy having the greatest value of the expected total-discounted reward, denoted by $V(p)$:

$$V(p) = \max_\pi \left\{V^\pi(p)\right\}. \tag{5}$$

According to [2, Thm. 6.3], there exists a stationary policy which makes $V(p) = V^{\pi*}(p)$, thus $V(p)$ can be calculated by the following equation:

$$V(p) = \max_{A \in \{SA, SC\}} \left\{V_A(p)\right\}, \tag{6}$$

where $V_A(p)$ is the greatest value of the expected total-discounted reward by taking action A when the initial belief probability is p. $V_A(p)$ can be expressed as:

$$V_A(p) = R(p, A) + \beta E\left[V_A(p') | b_0 = p, A_0 = A\right], \tag{7}$$

where b_0 is the initial belief, A is the action taken by transmitter, and p' is the new belief after taking action A.

Sending conservatively: by taking this action, the belief changes from p to $T(p) = \lambda_0(1-p) + \lambda_1 p = \alpha p + \lambda_0$, hence,

$$V_{SC}(p) = R_1 + \beta V(T(p)). \tag{8}$$

Sending aggressively: by taking this action, the channel state is known, so

$$\begin{aligned} & V_{SA}(p) \\ = \ & (pR_2 - (1-p)C) + \beta[pV(\lambda_1) + (1-p)V(\lambda_0)]. \end{aligned} \tag{9}$$

This formulation turns out to be similar to the problem formulation in [3], except for two main differences as follows:
(1). We do not have a separate action for sensing.
(2). We introduce penalty when sending fails.

Theorem 3.1: The optimal policy has a single threshold structure.

$$\pi^*(p) = \begin{bmatrix} SC & \text{if} & 0 \leq p \leq \rho, \\ SA & \text{if} & \rho \leq p \leq 1, \end{bmatrix} \tag{10}$$

where p is the belief, and ρ is the threshold.

We omit the proof here because it follows in a straightforward manner from the results in [3]. Since the penalty does not change the linear property of function $V_{SA}(p)$, according to [3, Thm. 1, Thm.2], $V_\beta(p)$ is convex and nondecreasing, thus the optimal solution follows a threshold structure. However, unlike [3], which shows the existence of multiple thresholds, there is a single threshold for our problem setting.

A. Closed Form Expression of Threshold ρ

The threshold ρ is the solution of the following equation:

$$R_1 + \beta V(T(\rho)) = V_{SA}(\rho). \tag{11}$$

There are two possible scenarios for $T(\rho)$:

If $T(\rho) \leq \rho$, we have $V(T(\rho)) = V_{SC}(T(\rho))$, substituting in Eq. (11), we can get:

$$\rho = \frac{R_1 + C}{(R_2 + C) + \beta V(\lambda_1) - \beta \frac{R_1}{1-\beta}}. \tag{12}$$

Otherwise, we have $V(T(\rho)) = V_{SA}(T(\rho))$, substituting in Eq. (11), we can get Eq. (13) at the top of next page.

B. $V(\lambda_0)$ and $V(\lambda_1)$

To calculate $V(\lambda_1)$ and $V(\lambda_0)$, there are also two possible scenarios:

If $\lambda_1 \leq \rho$, then since $\lambda_0 \leq \lambda_1 \leq \rho$,

$$V(\lambda_1) = V(\lambda_0) = \frac{R_1}{1-\beta}. \tag{14}$$

Otherwise, $\lambda_1 > \rho$, $V(\lambda_1) = V_{SA}(\lambda_1)$, using Eq. (9), we get

$$V(\lambda_1) = \frac{\lambda_1 R_2 - (1-\lambda_1)C + \beta(1-\lambda_1)V(\lambda_0)}{1 - \beta\lambda_1}. \tag{15}$$

To get $V(\lambda_0)$, we adapt [3, Thm.4]:

$$\begin{aligned} V(\lambda_0) &= \max_{A \in \{SA, SC\}} \{V_A(\lambda_0)\} \\ &= \max\{R_1 + \beta V(T(\lambda_0)), V_{SA}(\lambda_0)\} \\ &= \max\{R_1 \frac{1-\beta^N}{1-\beta}, \\ &\quad \max_{0 \leq n \leq N-1} \{\frac{1-\beta^n}{1-\beta} R_1 + \beta^n V_{SA}(T^n(\lambda_0))\}\}. \end{aligned} \tag{16}$$

Since N is arbitrary and $0 \leq \beta < 1$, when $N \to \infty$, we have

$$V(p) = \max_{n \geq 0} \{R_1 \frac{1-\beta^n}{1-\beta} + \beta^n V_{SA}(T^n(\lambda_0))\}. \tag{17}$$

Using Eq. (9), Eq.(15), Eq. (17), we can get

$$\begin{aligned} & V(\lambda_0) \\ = & \max_{n \geq 0} \left\{ \frac{\frac{1-\beta^n}{1-\beta} R_1 + \beta^n \left(\kappa_n R_2 + \kappa_n C(1-\beta) - C\right)}{1 - \beta^{n+1}\left[1 - (1-\beta)\kappa_n\right]} \right\}, \end{aligned} \tag{18}$$

where $\kappa_n = \frac{T^n(\lambda_0)}{1-\beta\lambda_1} = \frac{(1-\alpha^n)\lambda_S}{1-\beta\lambda_1}$.

IV. K-Conservative Policies

In this section, we will discuss the K-conservative policy, where K is the number of time slots to send conservatively after a failure, before sending aggressively again. The Markov chain corresponding to K-conservative policy is as shown in Fig. 1.

The states are the number of time slots which the transmitter

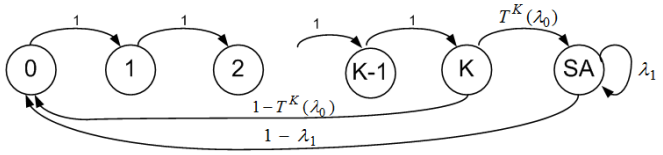

Fig. 1. K Conservative Policy Markov Chain.

has sent conservatively since last failure. There are $K+2$ states in this Markov chain. State 0 corresponds to the moment that sending aggressively fails, and it goes back to sending conservatively stage. State $K-1$ corresponds to that the transmitter has already sent conservatively for K time slots, and it will send aggressively next time slot. If the transmitter sends aggressively and succeeds, it goes to state SA and continues to send aggressively at the next time slot; otherwise it goes back to state 0. The probability that transmitter stays in state SA is λ_1. The transmitter has to wait K time slots before sending aggressively again, so the probabilities from state i to state $i+1$ is always 1 when $0 \leq i < K$.

There are $K+2$ states, each state corresponds to a belief and an action; belief and action determine the expected total-discounted reward. Thus given K, there are $K+2$ different expected total-discounted rewards.

Theorem 4.1: The threshold ρ policy structure is equivalent to a K_{opt}-Conservative policy structure, where K_{opt} is the number of time slots that a transmitter sends conservatively after a failure before sending aggressively again. If $\lambda_S > \rho$, $K_{opt} = \lceil \log_\alpha(1 - \frac{\rho(1-\alpha)}{\lambda_0}) \rceil - 1$, otherwise $K_{opt} = \infty$.

Proof: Whenever the transmitter sends aggressively but fails, it goes back to sending conservatively stage. The belief after a failure is λ_0, and it changes with time according to formula: $T^n(\lambda_0) = T(T^{n-1}(\lambda_0)) = \lambda_0(\frac{1-\alpha^{n+1}}{1-\alpha})$, where n is the number of time slots that the transmitter has sent conservatively since last time failure.

If $\lambda_S \leq \rho$, $T^n(\lambda_0)$ increases when n increases, when $n \to \infty$, $T^n(\lambda_0) \to \lambda_S$. The optimal policy is always sending conservatively, $K_{opt} = \infty$.

If $\lambda_S > \rho$, then there exists a finite integer K_{opt} which makes $T^{K_{opt}-1}(\lambda_0) < \rho$ and $T^{K_{opt}}(\lambda_0) \geq \rho$,

$$K_{opt} = \lceil \log_\alpha(1 - \frac{\rho(1-\alpha)}{\lambda_0}) \rceil - 1. \tag{19}$$

■

V. Online Learning for Unknown Channel

In this section, we will discuss how to find the optimal policy if the underlying channel's transition probabilities are

$$\rho = \frac{(1-\beta\lambda_1)R_1 + \lambda_0\beta R_2 + (1-\alpha\beta)(1-\beta)C + \beta(\beta-1)(1-\alpha\beta)V(\lambda_0)}{(1-\alpha\beta)(R_2 + (1-\beta)C + \beta(\beta-1)V(\lambda_0))}. \quad (13)$$

unknown. To find K_{opt}, we use the idea of mapping each K-conservative policy to a countable multi-armed bandits of countable time horizon. Now there are two challenges: (1). The number of arms can be infinite. (2). To get the true total discounted reward, each arm requires to be continuing played until time goes to infinity. To address these two challenges, we weaken our objective to find a suboptimal which is an $(OPT-(\epsilon+\delta))$ approximation of the optimal arm instead. Theorem 5.1 and theorem 5.2 address the two challenges respectively.

We define $(OPT-\epsilon)$ arm as the arm which gives $(OPT-\epsilon)$-approximation of the optimal arm no matter what the initial belief is.

Let arm SC correspond to the always sending conservatively policy, or $K_{opt} = \infty$.

Theorem 5.1: Given an ϵ and bound B on α, there exists a K_{max}, such that $\forall\ K \geq K_{max}$, the best arm in the arm set $C = \{0, 1, ...K, SC\}$ is an $(OPT-\epsilon)$ arm.

Proof: If $K > K_{opt}$ or $K_{opt} = \infty$, the optimal arm is already included in the arm sets.

If $K < K_{opt} < \infty$, suppose that transmitter has already sent conservatively for n time slots, let $k_{opt} = K_{opt} - n$, $k = K - n$, and $C' = R_2 + C + \frac{\beta}{1-\beta}(R_2 - R_1)$,

$$\begin{aligned}
& V^{\pi_{k_{opt}}}(p) - V^{\pi_k}(p) \quad (20)\\
= \ & [R_1\frac{1-\beta^{k_{opt}}}{1-\beta} + \beta^{k_{opt}}V_{SA}(T^{k_{opt}}(p))]\\
& -[R_1\frac{1-\beta^{k}}{1-\beta} + \beta^{k}V_{SA}(T^{k}(p))]\\
= \ & \beta^k[\frac{R_1}{1-\beta}(1-\beta^{k_{opt}-k})\\
& +\beta^{k_{opt}-k}V_{SA}(T^{k_{opt}}(p)) - V_{SA}(T^k(p))]\\
< \ & \beta^k[V_{SA}(T^{k_{opt}}(p)) - V_{SA}(T^k(p))]\\
= \ & \beta^k(T^{k_{opt}}(p) - T^k(p))(R_2 + C + \beta(V(\lambda_1) - V(\lambda_0)))\\
< \ & \beta^k(T^{K_{opt}}(\lambda_0) - T^K(\lambda_0))(R_2 + C + \beta(\frac{R_2}{1-\beta} - \frac{R_1}{1-\beta}))\\
= \ & \beta^k\alpha^{K+1}(1-\alpha^{K_{opt}-K})\lambda_S C'\\
< \ & \alpha^{K+1}C' < B^{K+1}C'.
\end{aligned}$$

Let $K_{max} = log_B\frac{\epsilon}{C'} - 1$,
when $K \geq K_{max}$, $V^{\pi_{k_{opt}}}(p) - V^{\pi_k}(p) < \epsilon$. ■

Theorem 5.2: Given an δ, there exists a T_{max} such that $\forall T \geq T_{max}$, an arm for the finite horizon total discounted reward up to time T is at most δ away from the infinite horizon total discounted reward.

Proof:

$$\begin{aligned}
& E[\sum_{t=0}^{\infty}\beta^t R_i(t)|b_0 = p] - E[\sum_{t=0}^{T_{max}}\beta^t R_i(t)|b_0 = p] \quad (21)\\
= \ & E[\sum_{t=T_{max}+1}^{\infty}\beta^t R_i(t)|b_0 = p]\\
< \ & \beta^{T_{max}+1}\frac{R_2}{1-\beta}.
\end{aligned}$$

When $T \geq \lceil \log_\beta \frac{\delta(1-\beta)}{R2} \rceil - 1$,
$E[\sum_{t=0}^{\infty}\beta^t R_i(t)|b_0 = p] - E[\sum_{t=0}^{T_{max}}\beta^t R_i(t)|b_0 = p] < \delta$. ■

We define period as time interval between arm switches, A-reward as the average $(OPT-\delta)$-finite horizon approximation total discounted reward at one period, regret as the number of time slots during which the transmitter uses policies which are more than $(\epsilon+\delta)$ away from the optimal policy. We design the UCB-Period (UCB-P) algorithm as shown in Algorithm 1. It is similar to UCB1 algorithm, but the time unit is period and the A-rewards are accumulated for each period.

Algorithm 1 Deterministic policy: UCB-P

Initialization: $L(\geq K_{max}+2)$ arms: arm $0, \cdots,$ arm $(L-2)$ and the SC arm. Play each arm for $T(\geq T_{max})$ time slots, then keep playing the arm until the arm hits arm state 0. Get initial A-reward of state 0 for each arm. Let $\bar{A} = A(i), (i = 0, 1, \cdots L-2, SC)$ as initial average A-reward for state 0 of each arm; $n_i = 1(i = 0, 1, \cdots L-2, SC)$ as initial number of periods arm i has been played, and $n = L$ as initial number of periods played so far.
for period $n = 1, 2, \cdots$ **do**
Select the arm with highest value of $\frac{(1-\beta)\bar{A}_i + C}{R_2 + C} + \sqrt{\frac{2ln(n)}{n_j}}$. Play the selected arm for a period. Update the average A-reward for state 0, n_i of the selected arm and n.
end for

Theorem 5.3: The regret of Algorithm UCB-P is bounded by $O(L(L+T)ln(t))$, where t is the number of time slots passed so far.

Proof: The procedure is similar to UCB1, [4, Thm.1] can be adapted.

A-reward is within the range of $[\frac{-C}{1-\beta}, \frac{R_2}{1-\beta}]$, where the left boundary corresponds to the transmitter sending aggressively but failing every time slot, and the right boundary corresponds to the transmitter sending aggressively and succeeding every time slot. We normalize the A-reward to be in the range of [0, 1], UCB1 algorithm shows that the number of time slots that selects non-optimal arm is bounded by $O((L-1)ln(n))$, where L is the number of arms and n the overall number of plays done so far.

The best arm is an $(OPT-(\epsilon+\delta))$ arm. If any other non-best arm $\tilde{K}$ hits SA states at the T^{th} time slots, the transmitter keeps playing that arm until sending fails, and the time playing the arm can be larger than T. Since the reward R_2 is the best reward the transmitter can get, these time slots do not count towards the regret. However, if such a $\tilde{K}^{th}$ arm hits state 0 just before T^{th} time slot, the transmitter needs to send conservatively for $\tilde{K}$ time slots. Since $\tilde{K} \leq L$, the arm can contribute regret for at most $(L+T)$ time slots. Thus, we will use $(L+T)$ to bound time slots generating regret in one period.

The number of plays selecting non-optimal arms in UCB1 is bounded by $O((L-1)ln(n))$. In our problem, the number of periods playing non-best arm is bounded by $O((L-1)ln(n))$, where n is the number of periods, $n < \frac{t}{T}$. In total, the number of time slots playing non-best arms is bounded by $O((L-1)(L+T)ln(\frac{t}{T})) = O(L(L+T)ln(t))$. Note that besides the best arm, some of the non-best arm in the arm sets may also give $(OPT-(\epsilon+\delta))$ approximation of the real expected total discounted reward, this only makes the regret smaller. $O(L(L+T)ln(n))$ is still an upper bound of the regret. More specifically, taking $L = K_{max}+2$ and $T = T_{max}$, the regret can be bounded by $O(K_{max}(K_{max}+T_{max})ln(n))$. ■

VI. Simulations

We start by analyzing the known underlying transition matrix case. We select 5 groups of transition probabilities, each corresponding to a different threshold, or equivalently, a different K_{opt}-conservative policy. The first corresponds to a scenario that the optimal policy is always sending conservatively. The other 4 correspond to different K_{opt}-conservative policies. In Fig. 2, the x axis represents different K-conservative policies and the y axis represents expected total discounted rewards. 5 curves correspond to different transition probabilities. The expected total discounted rewards get maximum when $K = K_{opt}$. Take the $K_{opt} = 4$ curve as an example,when $K < 4$, the total discounted reward increases when K increases and when $K > 4$, the total discounted reward decreases when K increase. The expected total discounted rewards get maximum when $K = 4$.

$R_1 = 1$; $R_2 = 2$; $C = 0.5$; $\beta = 0.75$

λ_0	λ_1	ρ	K_{opt}
0.36	0.91	0.5446	1
0.26	0.86	0.5060	2
0.16	0.96	0.4597	3
0.16	0.91	0.4553	4
0.01	0.61	0.5918	∞

TABLE I
THE OPTIMAL STRATEGY TABLE

Next, we consider the unknown transition probability matrix case. For $K \neq \infty$ scenarios, if α is bounded by 0.8, taking $\epsilon = 0.02$ and $\delta = 0.02$, we get $K_{max} = 26$ and $T_{max} = 20$. For the simulations, we take $L = 30$, $T = 100$. We run the simulations with different λ_0 and λ_1 and measure the

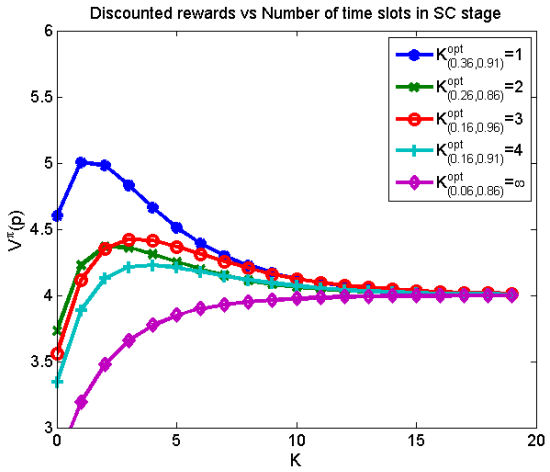

Fig. 2. Expected total discounted reward for K-conservative policies

percentage of time playing the $(OPT-(\epsilon+\delta))$ arm. Fig.3 is the simulation results running UCB-P algorithm. We can see when time goes to infinity, the percentage of time playing $(OPT-(\epsilon+\delta))$ arms approaches 100%.

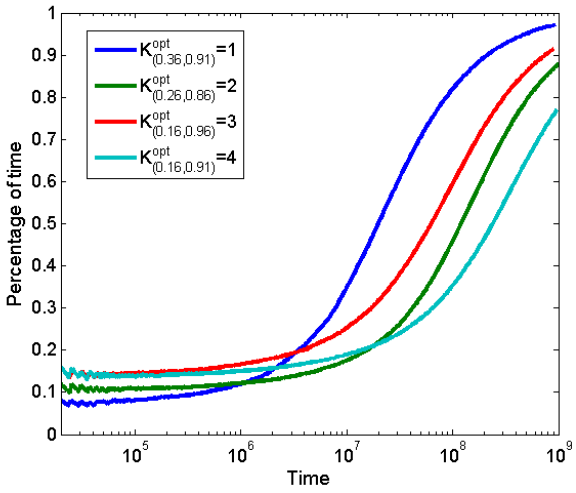

Fig. 3. Percentage of time that $OPT-(\epsilon+\delta)$ arms are selected by UCB-P algorithm

UCB-P algorithm, although mathematically proven to have logarithmic regret when $T \to \infty$, doesn't work that well in practice since it convergence slowly when the differences between arms are small. Thus, we use UCBP-TUNED algorithm, which the bound of UCB-P algorithm is tuned more finely. UCB1-TUNED [4] is adapted in our UCBP-TUNED algorithm. We use

$$V_k(s) \stackrel{def}{=} (\frac{1}{s}\Sigma_{\tau=1}^{s}(\frac{(1-\beta)A_{k,\tau}+C}{R_2+C})^2) - \frac{(1-\beta)\bar{A}_k+C}{R_2+C} + \sqrt{\frac{2\ln n}{s}}, \quad (22)$$

as an upper confidence bound for the variance of arm k, in which k is the arm index, and s is the number of periods that arm k is played during the first n periods, $A_{k,\tau}$ is the A-reward that arm k played the τth time.

We replace the upper confidence bound $\sqrt{2\ln(n)/n_j}$ of policy UCB-P with

$$\sqrt{\frac{\ln(n)}{n_j} min\{1/4, V_j(n_j)\}}, \tag{23}$$

and rerun the simulations. From Fig. 4, we can see that UCBP-TUNED algorithm converges much faster than UCB-P algorithm.

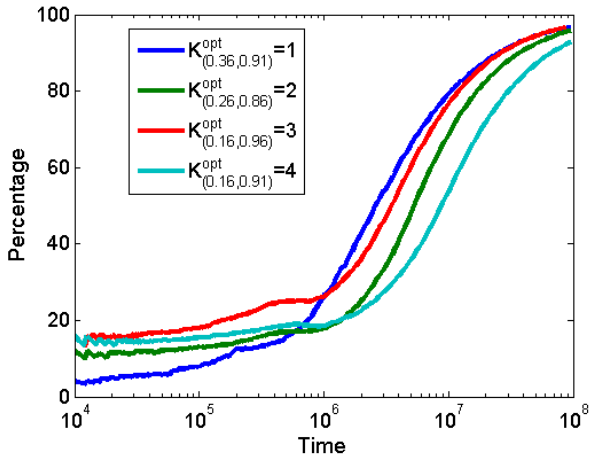

Fig. 4. Percentage of time that $OPT-(\epsilon+\delta)$ arms are selected by UCBP-TUNED algorithm

VII. Conclusion

This paper discusses the optimal policy of transmitting over a Gilbert Elliott Channel to maximize the expected total discounted rewards. If the underlying channel transition probabilities are known, the optimal policy has a single threshold. The threshold determines a K-conservative policy. If the underlying channel transition probabilities are unknown but a bound on α is known, we relax the requirement and show how to learn $(OPT-(\epsilon+\delta))$ policies. We designed UCB-P algorithm and the simulation results have shown that the percentage of selecting the $(OPT-(\epsilon+\delta))$ arms approaches 100% when $n \to \infty$. For future work, we plan to relax the assumption of known bound on α and consider ways to speed up the learning further.

Acknowledgment

This research was 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] R. Smallwood and E. Sondik, "The optimal control of partially observable Markov processes over a finite horizon", *Ops. Research*, pp.1071-1088, 1971

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

[3] A. Laourine and L. Tong, "Betting on Gilbert-Elliott Channels," *IEEE Transactions on Wireless Communications*, vol. 50, no.3, pp. 484-494, Feb. 2010

[4] P.Auer, N.Cesa-Bianchi, and P.Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, 47:235?56, 2002.

[5] L. A. Johnston and V. Krishnamurthy, "Opportunistic file transfer over a fading channel: A POMDP search theory formulation with optimal threshold policies," IEEE Transactions on Wireless Communications, vol. 5, pp. 394-405, 2006.

[6] A.K. Karmokar, D.V. Djonin, and V.K. Bhargava, "Optimal and suboptimal packet scheduling over correlated time varying flat fading channels", *IEEE Transactions On Wireless Communications*, Vol. 5, No. 2, February 2006.

[7] 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.

[8] 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.

[9] W. Dai, Y. Gai, B. Krishnamachari and Q. Zhao, "The Non-Bayesian Restless Multi-Armed Bandit: a Case of Near-Logarithmic Regret," *The 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May, 2011.

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