HOW MUCH LEARNING IS SUFFICIENT IN INTERFERENCE GAMES?

Yi Su and Mihaela van der Schaar

Electrical Engineering Department, UCLA

ABSTRACT

This paper studies the learning behavior of self-interested users interacting in a two-user OR-channel interference game. We discuss how a strategic user should learn the behavior of its opponent, adapt its actions, and improve its own performance. Specifically, we investigate the trade-off that can be made by a user between learning duration and performance, if the opponent plays a mixed strategy. First, we assume a stationary opponent and we apply optimization theory and large deviations theory to analytically derive an upper bound of the minimum training required by the user given the tolerable performance loss and outage probability. Next, we extend the results to the cases, where an adaptive opponent plays a conditional strategy based on its bounded memory. By solving linear programs, we design optimized learning strategies that minimize an upper bound of the duration of learning against the adaptive opponent.

***Index Terms*—**Learning in games, OR interference channel

1. INTRODUCTION

Game theory has been extensively applied to study a broad class of problems in communications systems, where various entities interact in a self-interested, autonomous manner [1]. However, most research studying problems in non-cooperative settings focuses on deriving or proving the existence of equilibria in games. Most of the arguments are based on the hypothesis of exact common knowledge of payoffs and rationality, which can usually not be realized in the investigated informationally-decentralized communication scenarios. Alternatively, learning in games techniques have been studied to model the strategic behavior of interacting users acquiring information, building knowledge, and ultimately improving their performance, as well as designing and selecting the equilibrium at which they desire to operate [2]. Several learning models have been applied to solve multi-user interaction problems in both wireline and wireless network settings [3]-[5]. For instance, in [3], appropriate learning solutions are studied in distributed environments consisting of players with very limited information about their opponents, such as the Internet. A reinforcement learning algorithm is proposed to maximize the average throughput in sensor communications [4]. By modeling the interaction among non-cooperative nodes in wireless ad hoc networks as a repeated game, a reinforcement learning algorithm is proposed to design power control in wireless ad hoc networks [5], where it is shown that the learning dynamics can eventually converge to Nash equilibrium and achieve satisfactory performance. In the area of game theory and artificial intelligence, limited results are known about how to learn against stationary or even adaptive opponents [6][7].

As opposed to the previous works that focus on studying the long-run convergence behavior of certain learning algorithms, this paper aims to characterize and quantify the achievable performance of learning with limited observations. We study this problem using a simple setting: the OR interference channel shared by two competing users. We model the interaction between autonomous users as a game and analyze the learning behavior of a strategic user that has no prior knowledge about its opponent. Particularly, we consider the cases in which the opponent plays mixed strategies. We explicitly quantify the benefits that a user can derive in terms of its improved utility by having a longer learning duration. Starting from the case where an opponent plays a stationary policy, we use optimization and large deviations theory to derive an upper bound of the minimum observation duration given the required performance guarantee. Then, the results are extended to cases where an adaptive opponent plays conditional strategies based on its bounded memory. In this case, we formulate and solve linear programs such that a strategic user can manipulate its adaptive opponent and adapt itself to best estimate the competitor's strategy. While this paper focuses on studying the benefits of learning in a simple setting, our solutions can be generalized to more complicated applications that requires strategic learning solutions for communications systems.

2. SYSTEM MODEL

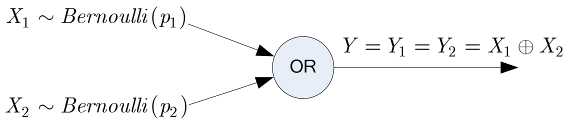
$X_1 \sim Bernoulli(p_1)$
$X_2 \sim Bernoulli(p_2)$
OR
$Y = Y_1 = Y_2 = X_1 \oplus X_2$

Fig. 1. Binary OR Channel

The considered system diagram is shown in Fig. 1 and the notations are given as follows. We denote by $\mathcal{I} = \{1,2\}$ the set of two participating *users*. For $i \in \mathcal{I}$, $-i$ denotes the complementary set $\mathcal{I} \setminus \{i\}$. Both users choose inputs from a binary alphabet $\mathcal{X} = \{0,1\}$. The input X_i of user i has a Bernoulli distribution with $\text{Prob}(x_i = 1) = p_i$ and $\text{Prob}(x_i = 0) = 1 - p_i$, i.e. $X_i \sim Bernoulli(p_i)$. The set of *actions* for user i is denoted as A_i, in which user i

determines its parameter $p_i \in [0,1]$. Let $\mathrm{A} = \mathrm{A}_1 \times \mathrm{A}_2$ be the set of action combinations and $a(n) = (a_1(n), a_2(n)) = (p_1^n, p_2^n) \in \mathrm{A}$ be the action combination at time n. Player *i*'s *mixed* actions set is the probability simplex over A_i, i.e. $\Delta(\mathrm{A}_i) = \left\{ g_i \in \mathcal{R}^{|\mathrm{A}_i|} : g_i \geq 0 \ componentwise \ and \ \mathbf{1}^T g_i = 1 \right\}$, where $\mathbf{1} = [1, \cdots, 1]^T \in \mathcal{R}^{|\mathrm{A}_i|}$. Both users observe the same output Y, which is the OR operation of the two input variables, i.e. $Y = Y_1 = Y_2 = X_1 \oplus X_2$ [8]. This channel originates from modeling the error characteristic of optical systems [9]. We assume that each user simply decodes its own signals by treating the other user's signal as noise. Under this decoding scheme, the multi-user OR channel can be transformed into two Z-channels shown in Fig. 2. For example, in user 1's asymmetric Z-channel model, the probability of 1 to 0 error is zero and that of 0 to 1 error is the probability p_2 that user 2 sends 1. Given the input distributions, user i's achievable rate is

$$\begin{aligned} R_i &= H(Y_i) - H(Y_i \mid X_i) \\ &= q(p_i + p_{-i}(1-p_i)) - (1-p_i) q(p_{-i}), \end{aligned} \tag{1}$$

where $q(x) = -x \log x - (1-x)\log(1-x)$. User's *utility functions* are determined based on their achievable rates, $u_i = R_i : \mathrm{A} \to \mathcal{R}$. It is easy to see that, if $p_i \neq 0$, user i's transmitted signal will cause interference to the other. Therefore, their utilities are coupled together by their actions. Summarizing, the tuple $\langle \mathcal{I}, (\mathrm{A}_i), (u_i) \rangle$ defines the model of interaction between the users [10]. Note that, in this interference game, users can observe their opponents' actions from its received signal. In this paper, a user's *observations* refer to the actions that its opponent took in the entire history.

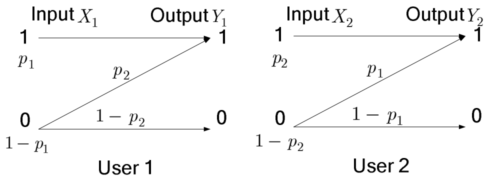

Fig. 2. Equivalent Z-channels for each user

3. LEARNING IN INTERFERENCE GAMES

This paper discusses how long a user should learn about its opponent's strategy, if the opponent behaves according to some initially unknown probability distributions over its action space A_{-i}. Specifically, using the learning scheme of fictitious play as an illustrative example, we derive an upper bound of the minimum required observation duration given the tolerable performance loss, and explicitly discuss the tradeoff of observation duration vs. performance. In this paper, the opponent is assumed to adopt two types of strategies. A strategy g_i of player i is a sequence of functions $(g_i^1, g_i^2, \cdots, g_i^n, \cdots)$, where the function g_i^n assigns a mixed action in $\Delta(\mathrm{A}_i)$ at each time n to each history $h^{n-1} = (a(1), a(2), \cdots, a(n-1))$. A strategy g_i of player i has finite memory if there exists a positive integer N_h such that only the history of the last N_h periods matters: for each $n > N_h$, the function g_i^n is of the form $g_i^n = g_i(a(n-N_h), a(n-N_h+1), \cdots, a(n-1))$, and we call this N_h-memory. The first type of strategy is named stationary strategy because the opponent always plays a stationary policy, i.e. g_i^n is fixed all the time. For the second type, the opponent plays a limited-memory adaptive strategy based on the actions that the strategic user took in the history $a_{-i}(n-N_h), a_{-i}(n-N_h+1), \cdots, a_{-i}(n-1)$. We limit the opponent's capabilities to these two types of strategies, because directly handling the opponents with entire history is intractable and these two types of behavior models capture the strategic nature of the opponent with limited-memory [6].

3.1. Fictitious Play

Over the past several decades, many learning algorithms, e.g. fictitious play and reinforcement learning, have been developed [2]. It is difficult to find the optimal learning scheme in general cases. Instead, we choose to fix the learning rule and explicitly quantify the achievable performance given the observation duration. This paper discusses the learning scheme of *fictitious play* [12], because the fact that the actions of the other user are observable in the interference game makes fictitious play the most efficient solution. The model of fictitious play is simply a count of the plays taken by the opponent in the past, and the observed frequencies are taken to represent the opponent's mixed strategy. Assume a stationary opponent chooses its action according to the probability mass function (pmf) $g_{-i}(p_{-i})$ at all times. We define an empirical frequency function

$$\gamma^n(p_{-i}) = \frac{k^n(p_{-i})}{\sum_{\tilde{p}_{-i} \in \mathrm{A}_{-i}} k^n(\tilde{p}_{-i})}, \tag{2}$$

where $k^n(a_{-i})$ is a counting function satisfying $k^0(a_{-i}) = 0, \ \forall a_{-i} \in \mathrm{A}_{-i}$ and

$$k^n(p_{-i}) = \begin{cases} k^{n-1}(p_{-i}) + 1, \ if \ p_{-i}^n = p_{-i} \\ k^{n-1}(p_{-i}) \quad , otherwise \end{cases}. \tag{3}$$

The strategic user approximates the actual pmf $g_{-i}(p_{-i})$ using the empirical frequency function $\gamma^n(p_{-i})$, and takes the best response that maximizes u_i by solving

$$\max_{p_i} \sum\nolimits_{p_{-i} \in \mathrm{A}_{-i}} \gamma^n(p_{-i}) R_i(p_i, p_{-i}). \tag{4}$$

We denote the performance in (4) with empirical frequency function $\gamma^n(p_{-i})$ as $U_i(\gamma^n)$, which differs from the performance with perfect information $g_{-i}(p_{-i})$, denoted as $U_i(g)$. Hence, an important question is how much a strategic user should learn when the opponent plays either stationary or adaptive policy, given its tolerable performance loss and outage probability.

3.2. Stationary Opponent

In this subsection, we investigate how much learning against a stationary opponent that fixes g_{-i}^{n} all the time is required. Intuitively, we know that the achievable rate will be improved when having more observations. Given the tolerable performance loss Δ_u and outage probability δ_u, the problem is formulated as

$$\begin{aligned} &\min n \\ &s.t.\ \text{Prob}\left(U_i(g) - U_i(\gamma^n) \geq \Delta_u\right) \leq \delta_u . \end{aligned} \tag{5}$$

Although there are several error bounds in statistical learning theory, e.g. Hoeffding's inequality [11], it is difficult to solve the problem in (5) because these bounds do not directly apply to this problem. However, we can find an upper bound for the optimum of (5). The key idea is to adopt tools from large deviations theory, which mainly concerns the asymptotic behavior of remote tails of sequences of probability distributions [13]. According to large deviations theory, the empirical frequency function $\gamma^n(p_{-i})$ of a random sample of size n drawn from $g_{-i}(p_{-i})$ satisfies

$$\text{Prob}\left(D(\gamma^n \,||\, g) \geq \delta\right) \leq \begin{pmatrix} n + |\mathrm{A}_{-i}| - 1 \\ |\mathrm{A}_{-i}| - 1 \end{pmatrix} 2^{-n\delta}, \forall \delta > 0, \tag{6}$$

where $D(p \,||\, q)$ is the Kullback-Leibler (KL) distance between two pmfs $p(x)$ and $q(x)$ [14]. Then, we need to convert the performance loss $U_i(g) - U_i(\gamma^n)$ into the KL distance $D(\gamma^n \,||\, g)$. Note these two metrics cannot always perfectly align with each other. Given Δ_u, we choose to find the minimum $\delta_{D_{\min}}$ such that $D(\gamma^n \,||\, g) \leq \delta_{D_{\min}}$ always leads to $U_i(g) - U_i(\gamma^n) \leq \Delta_u$. In other words, we are deriving an upper bound for problem (5). The key idea in determining the upper bound of observation duration is illustrated in Fig. 3. Using this figure, we divide the whole procedure into three steps and explain each step in details as follows.

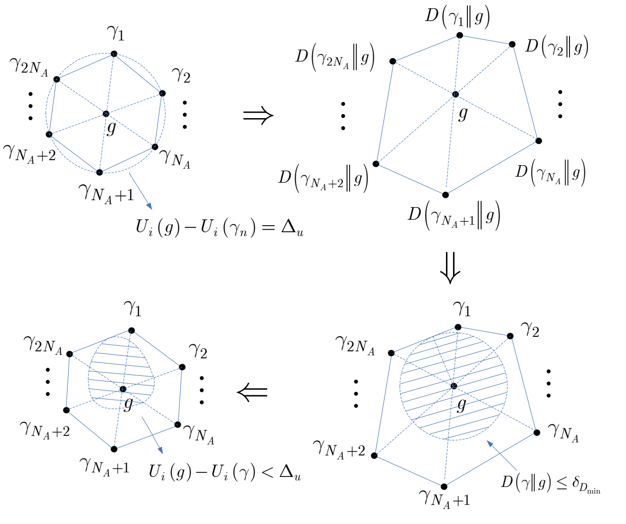

Fig. 3. Performance loss and KL distance

1) Extreme Points with Performance Loss Constraints

First, in the probability simplex $\Delta(\mathrm{A}_{-i})$, we construct a convex set that contains the true pmf. Let $\mathrm{A}^{comb} = \{\{k, j\} : k, j \in \{1, 2, \cdots, |\mathrm{A}_{-i}|\}\ and\ k < j\}$ and $(S)_n$ denote the n th element of S. We are interested in a total number of $2N_{\mathrm{A}}$ pmfs with $N_{\mathrm{A}} = \begin{pmatrix} |\mathrm{A}_{-i}| \\ 2 \end{pmatrix}$. These pmfs $\gamma_m \in \Delta(\mathrm{A}_{-i})$ are

$$\gamma_m\left(p = (\mathrm{A}_{-i})_n\right) = g_{-i}\left(p = (\mathrm{A}_{-i})_n\right), if\ n \notin \left(\mathrm{A}^{comb}\right)_m . \tag{7}$$

These pmfs can be rewritten as $\gamma_m(p, \delta)$ satisfying

$$\gamma_m\left(p = (\mathrm{A}_{-i})_n, \delta\right) = \begin{cases} g_{-i}\left(p = (\mathrm{A}_{-i})_n\right) - \delta, if\ n = \left(\left(\mathrm{A}^{comb}\right)_m\right)_1 \\ g_{-i}\left(p = (\mathrm{A}_{-i})_n\right) + \delta, if\ n = \left(\left(\mathrm{A}^{comb}\right)_m\right)_2 \\ g_{-i}\left(p = (\mathrm{A}_{-i})_n\right), \quad if\ n \notin \left(\mathrm{A}^{comb}\right)_m \end{cases}, \tag{8}$$

$m = 1, 2, \cdots, 2N_{\mathrm{A}}$. The $2N_{\mathrm{A}}$ pmfs are determined based on the tolerable performance loss Δ_u, and they are viewed as "extreme points". The extreme points are chosen to be $\gamma_m(p, \delta_m), m = 1, 2, \cdots, 2N_{\mathrm{A}}$. For $m = 1, 2, \cdots, N_{\mathrm{A}}$,

$$\delta_m = \begin{cases} g_{-i}\left(p = (\mathrm{A}_{-i})_l\right) with\ l = \left(\left(\mathrm{A}^{comb}\right)_m\right)_1, if\ S_\delta = \varnothing \\ \min \delta \in S_\delta \qquad , otherwise \end{cases}, \tag{9}$$

in which $S_\delta = \left\{\delta : U_i(g) - U_i(\gamma_m(\delta)) \geq \Delta_u\ and\ \delta \geq 0\right\}$, and

$$\delta_{m+N_{\mathrm{A}}} = \begin{cases} -g_{-i}\left(p = (\mathrm{A}_{-i})_l\right) with\ l = \left(\left(\mathrm{A}^{comb}\right)_m\right)_2, if\ S_{-\delta} = \varnothing \\ \min \delta \in S_{-\delta} \qquad , otherwise \end{cases}. \tag{10}$$

in which $S_{-\delta} = \left\{\delta : U_i(g) - U_i(\gamma_m(-\delta)) \geq \Delta_u\ and\ \delta \geq 0\right\}$, and the achievable rate of $\gamma_m(p, \delta_m)$ is denoted as $U_i(\gamma_m(\delta))$. Note when $n \in \left(\mathrm{A}^{comb}\right)_m$, $\gamma_m\left(p = (\mathrm{A}_{-i})_n, \delta_m\right)$ is set to zero if $S_\delta = \varnothing$ or $S_{-\delta} = \varnothing$ due to the non-negative property.

We focus on the convex set $\mathscr{B}$ formed by the convex hull of the extreme points, i.e. $\mathscr{B} = conv\left\{\gamma_m\left(p = (\mathrm{A}_{-i})_n, \delta_m\right), m = 1, \cdots, 2N_{\mathrm{A}}\right\}$ [16]. We can derive an upper bound of the minimum required learning duration in this convex set

Proposition 1: Any $\gamma \in \mathscr{B}$ satisfies $U_i(g) - U_i(\gamma) \leq \Delta_u$.

Proof: It is easy to verify that in (1), the achievable rate R_i is a concave function in p_i [1] and there always exist a unique zero $\frac{\partial U_i}{\partial p_i}$ for any $p_{-i} \in [0,1]$. Therefore, there exist an interval $\left[p_i^{\min}, p_i^{\max}\right]$ such that, for any γ with $U_i(g) - U_i(\gamma) \leq \Delta_u$, the maximizer of $U_i(\gamma)$ satisfies $p_i \in \left[p_i^{\min}, p_i^{\max}\right]$, and vice versa. In other words, for $m = 1, 2, \cdots, 2N_{\mathrm{A}}$, $\frac{\partial U_i(\gamma_m)}{\partial p_i} > 0, \forall p_i \in \left(0, p_i^{\min}\right)$ and $\frac{\partial U_i(\gamma_m)}{\partial p_i} < 0$, $\forall p_i \in \left(p_i^{\max}, 1\right)$. Note that $\forall \gamma \in \mathscr{B}$, $U_i(\gamma)$ are convex combination of $U_i(\gamma_m)$, $m = 1, 2, \cdots, 2N_{\mathrm{A}}$. It follows that $\arg\max_{p_i} U_i(\gamma) \in \left[p_i^{\min}, p_i^{\max}\right]$, because $\frac{\partial U_i(\gamma)}{\partial p_i} > 0$ for $p_i \in \left(0, p_i^{\min}\right)$ and $\frac{\partial U_i(\gamma_m)}{\partial p_i} < 0$ for $p_i \in \left(p_i^{\max}, 1\right)$. Hence, we can conclude that $U_i(g) - U_i(\gamma) \leq\ \Delta_u$ for any $\gamma \in \mathscr{B}$. ■

2) KL Distance Minimization in Extreme Points Set

In the first step, a convex set $\mathcal{B}$ is constructed based on the tolerable performance loss Δ_u. Now we apply large deviations theory to translate the performance loss Δ_u into another metric, KL distance δ_D. The basic idea is to solve an optimization problem to find the minimum KL distance $\delta_{D_{\min}}$ such that, for any γ that satisfies $D(\gamma \,||\, g) \leq \delta_{D_{\min}}$, we have $U_i(g) - U_i(\gamma) \leq \Delta_u$. The optimization problem is formulated as

$$\begin{aligned} &\min_{\gamma} D(\gamma \,||\, g) \\ &s.t.\ \gamma \in \mathcal{S}(\mathcal{B}), \end{aligned} \tag{11}$$

where $\mathcal{S}(\mathcal{B})$ represents the surface of the convex set $\mathcal{B}$, i.e. $\mathcal{S}(\mathcal{B}) = \mathcal{B} \setminus \operatorname{int}(\mathcal{B})$. Here we denote the interior of the set $\mathcal{B}$ as $\operatorname{int}(\mathcal{B})$ [15]. Note that the KL distance $D(\gamma \,||\, g)$ is convex in the pair (γ, g) and $\gamma \in \mathcal{S}(\mathcal{B})$ is a linear constraint [14]. Problem (11) essentially belongs to convex programming, and the optimal solution can be obtained efficiently [16]. Since the convex combinations of the extreme points in $\mathcal{B}$ cover the adjacent region of the true pmf g, we can see that the solution value $\delta_{D_{\min}}$ of the problem (11) that ensures $D(\gamma \,||\, g) \leq \delta_{D_{\min}}$ is sufficient to guarantee that $U_i(g) - U_i(\gamma) \leq \Delta_u$.

In the next step, we will apply large deviations theory to convert the bounded KL distance into an upper bound of minimum learning duration.

3) Minimum Observation Duration Calculation

In the previous step, we know that $D(\gamma \,||\, g) \leq \delta_{D_{\min}}$ always leads to $U_i(g) - U_i(\gamma) \leq \Delta_u$. Hence, an upper bound of problem (5) can be obtained by solving

$$\begin{aligned} &\min\ n \\ &s.t.\ \mathsf{Prob}\left(D(\gamma^n \,||\, g) \geq \delta_{D_{\min}}\right) \leq \delta_u. \end{aligned} \tag{12}$$

Applying formula (6), we have the following proposition:

Proposition 2: Suppose user i adopts fictitious play to update its empirical frequency function γ_{-i}^n and take the best-response action correspondingly. An upper bound N of the solution of problem (5) is

$$N = Q\left(\delta_{D_{\min}}, |\mathrm{A}_{-i}|, \delta_u\right), \tag{13}$$

in which $Q(x,y,z) = \min\left\{ n : \begin{pmatrix} n+y-1 \\ y-1 \end{pmatrix} \cdot 2^{-nx} \leq z \right\}$.

Proof: Combining (6) and (12), we know that any n that satisfies

$$\begin{pmatrix} n+|\mathrm{A}_{-i}|-1 \\ |\mathrm{A}_{-i}|-1 \end{pmatrix} 2^{-n\delta_{D_{\min}}} \leq \delta_u \tag{14}$$

is an upper bound of the solution for (5). Let $F(n) = \begin{pmatrix} n+|\mathrm{A}_{-i}|-1 \\ |\mathrm{A}_{-i}|-1 \end{pmatrix} 2^{-n\delta_{D_{\min}}}$. We have $\dfrac{F(n+1)}{F(n)} = \dfrac{n+|\mathrm{A}_{-i}|}{n+1} \cdot 2^{-\delta_{D_{\min}}}$ and $\lim\limits_{n\to\infty} \dfrac{F(n+1)}{F(n)} = 2^{-\delta_{D_{\min}}} < 1$. We can conclude that $\lim\limits_{n\to\infty} F(n) = 0$. As a result, by choosing $N = Q\left(\delta_{D_{\min}}, |\mathrm{A}_{-i}|, \delta_u\right)$ as the minimum integer satisfying the inequality (14), we obtain an upper bound of the optimum of (5). ■

Next, we provide some intuition to interpret the derived upper bound. Define $f : \mathcal{R} \to \mathcal{R}$ to be the non-increasing function that maps the tolerable performance loss into the KL distance. The upper bound of minimum learning duration can be rewritten as

$$N = Q\left(f(\Delta_u), |\mathrm{A}_{-i}|, \delta_u\right). \tag{15}$$

We can make several remarks about the upper bound:

Remark 1 **:** Decreasing the acceptable performance loss Δ_u will lead to a larger upper bound of the minimum observation duration.

Remark 2 **:** Decreasing the outage probabilty δ_u will increase the upper bound N.

Remark 3 **:** Adding more actions to enlarge the dimension of action space $|\mathrm{A}_{-i}|$ also increases the upper bound N.

3.3. Adaptive Opponent

Now we consider an N_h-memory adaptive opponent that updates its action strategies based on the observed history of $\left\{p_i^{n-N_h}, \cdots, p_i^{n-1}\right\}$ in its bounded memory [6]. The total number of possible states in its memory is $N_{\mathrm{A}_h} = |\mathrm{A}_i|^{N_h}$. Denote the pmf of the opponent's action strategy in the j th state as $g_j(p_i)$ and the index of the state to be $j_n = \pi\left(p_i^{n-N_h}, \cdots, p_i^{n-1}\right)$ when having $\left\{p_i^{n-N_h}, \cdots, p_i^{n-1}\right\}$ in its memory. We also denote $\pi^{-1}(j_n) = p_i^{n-N_h}, \cdots, p_i^{n-1}$ and $\left[\pi^{-1}(j_n)\right]_{m:n} = p_i^{n-N_h+m-1}, \cdots, p_i^{n-N_h+n-1}$.

A reasonable learning strategy against limited-memory adaptive opponent is to maintain separate empirical frequency function for each state, and update these functions based on the observed action of its opponent at each time. We name this strategy "conditional learning". Specifically, the empirical frequency functions $\gamma_1(p_{-i})$, $\gamma_2(p_{-i}), \cdots, \gamma_{N_{A_h}}(p_{-i})$ are updated according to

$$\gamma_j^n(p_{-i}) = \frac{k_j^n(p_{-i})}{\sum_{\tilde{p}_{-i} \in \mathrm{A}_{-i}} k_j^n(\tilde{p}_{-i})}, \tag{16}$$

where $k_j^n(p_{-i})$ takes the form of

$$k_j^n(p_{-i}) = \begin{cases} k_j^{n-1}(p_{-i}) + 1, \ if\ p_{-i}^n = p_{-i}\ and \\ \qquad\qquad j = \pi\left(p_i^{n-N_h}, \cdots, p_i^{n-2}, p_i^{n-1}\right) \\ k_j^{n-1}(p_{-i}) \qquad , otherwise \end{cases}. \tag{17}$$

If the opponent's bounded memory is in state j_k at time k and the strategic user takes action p_i^k, the conditional learning strategy will transits from state j_k to $j_{k+1} = \pi\left(\left[\pi^{-1}(j_k)\right]_{2:N_h}, p_i^k\right)$. At each time, the strategic user updates the empirical frequency function of the opponent's action using (16) and (17). Assuming that the initial state is

[1] The derived upper bound can be generalized to other cases as long as U_i is concave in the A_i.

$j_0 = \pi\left(p_i^{-N_h}, \cdots, p_i^{-1}\right)$, we want to design the strategic user's action sequence $p_i^0, p_i^1, \cdots, p_i^{n-1}$ with the minimal learning duration, while each state has the performance loss guarantee similar to (5). The problem is formulated as

$$\begin{aligned} &\min_{p_i^0, p_i^1, \cdots, p_i^{n-1}} \; n \\ &s.t.\ \mathsf{Prob}\left(U_i\left(g_j\right) - U_i\left(\gamma_j^n\right) \geq \Delta_u\right) \leq \delta_u,\ \forall j = 1, 2, \cdots, N_{\mathrm{A}_h}. \end{aligned} \tag{18}$$

An upper bound of the optimal solution in (18) can be derived by decomposing the problem into several separate sub-problems:

$$\begin{aligned} &\min\ n_k \\ &s.t.\ \mathsf{Prob}\left(U_i\left(g_k\right) - U_i\left(\gamma_k^n\right) \geq \Delta_u\right) \leq \delta_u \\ &\sum_{t=1}^{n} I\left(j_t = k\right) = n_k, \end{aligned} \tag{19}$$

in which $I(\bullet)$ is the indicator function and n_k represents the number of times that the opponent's memory is in state k. It is easy to see that problem (19) is exactly the same as (5). Based on the previous results, we can obtain an upper bound N_k for each sub-problem. Note that each state can transit into $|\mathrm{A}_i|$ different states. Therefore, an upper bound is achieved if the total number of transiting into any state k from time 0 to n is larger than N_k. Now we can design an optimized conditional learning strategy that minimizes the overall learning duration while satisfying the individual sub-problems (19) by solving a linear program. If the ending state at time n is j_e, the following formulation gives an upper bound of the solution in (18):

$$\begin{aligned} &\min \sum_{k=1}^{N_{A_h}} n_k \\ &s.t.\ n_k \geq N_k, && \forall k = 1, 2, \cdots, N_{\mathrm{A}_h} \\ &n_{kk'} \geq 0, && \forall k, k' = 1, 2, \cdots, N_{\mathrm{A}_h} \\ &\textstyle\sum_{k' \in F_k} n_{kk'} = n_k, && if\ k \neq j_e \\ &\textstyle\sum_{k' \in F_k} n_{kk'} = n_k - 1, && if\ k = j_e \\ &\textstyle\sum_{k' \in G_k} n_{k'k} = n_k, && if\ k \neq j_0 \\ &\textstyle\sum_{k' \in G_k} n_{k'k} = n_k - 1, && if\ k = j_0, \end{aligned} \tag{20}$$

where $F_k = \left\{k' : k' = \pi\left(\left[\pi^{-1}(k)\right]_{2:N_h}, p_i\right), \forall p_i \in \mathrm{A}_i\right\}$, $G_k = \left\{k' : k' = \pi\left(p_i, \left[\pi^{-1}(k)\right]_{1:N_h - 1}\right), \forall p_i \in \mathrm{A}_i\right\}$, and $n_{kk'}$ indicates the total number that the conditional learning algorithm transits from state k to state k' in the time interval of $[0, n]$. By solving this linear program, we can derive a conditional learning algorithm that starts from state j_0, ends in state j_e, and satisfies $n_k \geq N_k$. The upper bound of learning duration can be further minimized by enumerating all the possible ending states of $j_e \in \underbrace{\mathrm{A}_i \times \cdots \times \mathrm{A}_i}_{N_h}$.

4. SIMULATION RESULTS

Since the cases with an adaptive opponent are simply linear extensions of the stationary cases, we only provide the simulation results for stationary opponent in this paper. We assume a stationary opponent with $\mathrm{A}_{-i} = \{0.05, 0.5, 0.95\}$. The opponent can choose his actions based on the pmf $g_{-i}(p_{-i})$ with $g_{-i}(p_{-i} = 0.05) = 0.3$, $g_{-i}(p_{-i} = 0.5) = 0.4$, and $g_{-i}(p_{-i} = 0.95) = 0.3$. The contours of achievable rate $U_i(\gamma)$ and KL distance $D(\gamma \,||\, g)$ are shown in Fig. 4 and 5 separately. We can see from the figures that both $U_i(\gamma)$ and $D(\gamma \,||\, g)$ are convex in γ, which verifies proposition 1. We set the parameters in the problem (5) to be $\Delta_u = 10^{-4}$ and $\delta_u = 10^{-2}$. Fig. 6 and 7 illustrate the procedure of obtaining the upper bound in Section 3.2. Noting that $N_{\mathrm{A}} = |\mathrm{A}_{-i}| = 3$, six extreme points $\gamma_1, \cdots, \gamma_6$ are chosen in total, which are determined based on the utility-pmf curves in Fig. 6. These three plotted curves indicate the achievable rates for three pmfs, including $\gamma(p_{-i} = 0.5) = 0.4$, $\gamma(p_{-i} = 0.05) = 0.3$, or $\gamma(p_{-i} = 0.95) = 0.3$, i.e. $\gamma(p_{-i} = 0.5) + \gamma(p_{-i} = 0.05) = 0.7$. The convex hull of these extreme points $\gamma_1, \cdots, \gamma_6$ is the extreme point set $\mathscr{B}$. The red hexagon in Fig. 7 is the surface $\mathscr{S}(\mathscr{B})$ on which we minimize the KL distance. Solving the convex optimization problem (11) leads to $\delta_{D_{\min}} = 0.041233$. We can obtain $N = Q\left(0.041233, 3, 10^{-2}\right) = 582$ using (13). As shown in Fig. 7, if the observation duration is larger than N, the KL distance between the actual stationary action pmf $g_{-i}(p_{-i})$ and observed empirical frequency function $\gamma^N(p_{-i})$ will lie within the green circle with an outage probability less than 10^{-2}.

We also examine the tightness of the upper bound in different settings. The tolerable performance loss Δ_u is varied to be $10^{-3.5}, 10^{-4}, 10^{-4.5}$ and 10^{-5} while the outage probability δ_u is set as a constant 10^{-2}. In each scenario, we use Monte Carlo method and run 10^5 realizations to get the actual required learning duration N_a. The results are summarized in Table I. From the table, we can see that, under this setting, the bound is not very tight and the ratio of N_a / N is around 4 when Δ_u is varied. This can be explained by phenomena that the contours are similar with each other. Moreover, we can also see that wisely choosing the extreme points can enlarge $\mathscr{S}(\mathscr{B})$ and improve the tightness of the upper bound.

5. CONCLUSIONS

This paper studies the learning behavior in a two-user OR-channel interference game. In the existence of the mixed-strategy opponent, how much a strategic user should learn the response strategy of its opponent is discussed for both stationary and adaptive opponents. The derived results are useful for designing and evaluating future communications protocols with learning mechanisms.

6. REFERENCES

[1] E. Altman, T. Boulogne, R. El-Azouzi, T Jimenez, and L. Wynter, “A survey on networking games in telecommunications”, Computers and Operations Research, 2004.

[2] D. Fudenberg, and D. K. Levine, “The Theory of Learning in Games”, Cambridge, MIT Press, 1998.

[3] E. Friedman, and S. Shenker. "Learning and Implementation on the Internet." Manuscript. New Brunswick: Rutgers Univ., Department of Economics, 1997.

[4] C. Pandana and K.J.R. Liu, "Near Optimal Reinforcement Learning Framework for Energy-Aware Wireless Sensor Communications", IEEE JSAC special issue on Self-Organizing Distributed Collaborative Sensor Networks, Vol. 23, no 4, pp.788-797, April 2005.

[5] C. Long, Q. Zhang, B. Li, H. Yang, and X. Guan, "Non-Cooperative Power Control for Wireless Ad Hoc Networks with Repeated Games", IEEE JSAC special issue on Non-Cooperative Behavior in Networking, Vol. 25, pp. 1101-1112, Aug, 2007

[6] R. Powers and Y. Shoham, “Learning Against Opponents with Bounded Memory,” Proc. of IJCAI, pp. 817-822, 2005

[7] V. Conitzer and T. Sandholm, “Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents,” Proc. of the 20th Intern. Conf. on Machine Learning, pp. 83–90, 2003.

[8] A. Grant, B. Rimoldi, R. Urbanke, and P. Whiting, “Rate-Splitting Multiple Access for Discrete Memoryless Channels”, IEEE Trans. on Inform. Theory, 47(3):873–890, Mar. 2001.

[9] L.G. Tallini, S. Al-Bassam and B. Bose, “On the Capacity and Codes for the Z-Channel”, Proc. IEEE ISIT, p. 422, 2002.

[10] M. Osborne and A. Rubenstein, A Course in Game Theory. MIT Press, 1994

[11] O. Bousquet, S. Boucheron, and G. Lugosi, “Introduction to Statistical Learning Theory”, Advanced Lectures on Machine Learning Lecture Notes in Artificial Intelligence, Vol. 3176, pp. 169-207. Springer, 2004

[12] G. Brown, “Iterative solution of games by fictitious play,” Activity Analysis of Production and Allocation, John Wiley and Sons, New York, 1951

[13] I. Csiszár and P. C. Shields, “Information theory and statistics: a tutorial,” Communications and Information Theory, Vol.1, Issue.4, pp. 417-528, Dec. 2004

[14] T. M. Cover and J. A. Thomas, Elements of Information Theory. New York: Wiley, 2006.

[15] J. B. Conway, A Course in Functional Analysis, 2nd edition, Springer-Verlag, 1994.

[16] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

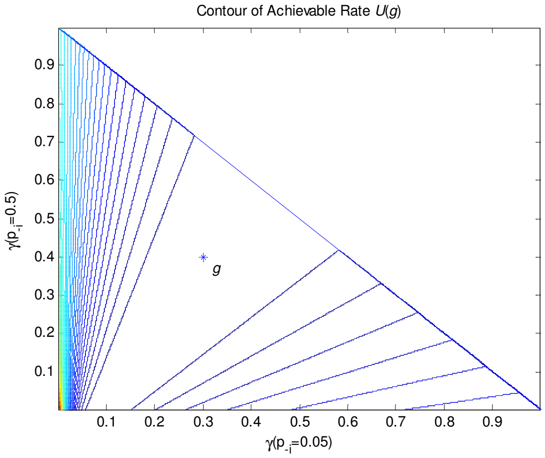

Fig. 4. Contour of achievable rate $U_i(\gamma)$

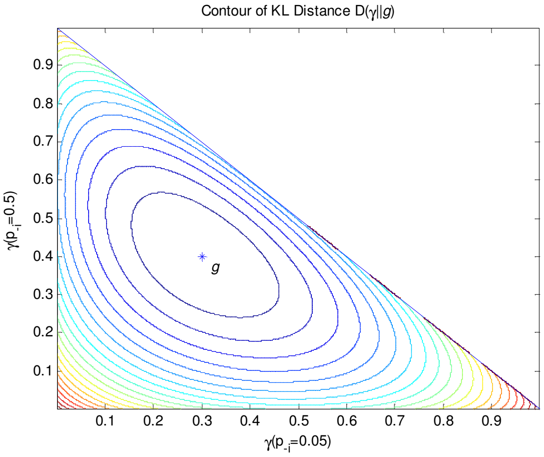

Fig. 5. Contour of KL distance $D(\gamma \,||\, g)$

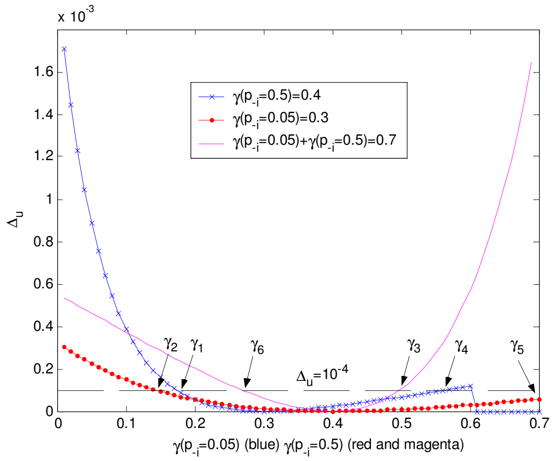

Fig. 6. Constructing the extreme points

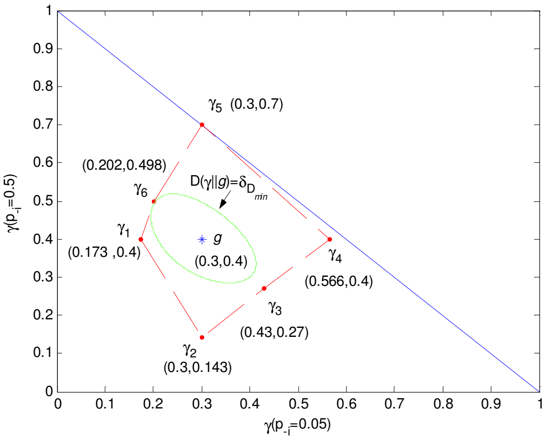

Fig. 7. KL distance minimization in $\mathcal{S}(\mathcal{B})$

Performance loss Δ_u	$10^{-3.5}$	10^{-4}	$10^{-4.5}$	10^{-5}
Actual value N_a	48	150	480	1570
Upper bound N	189	582	1865	6360

Table I. Learning duration for different performance loss requirements