A simple learning rule with monitoring leading to Nash Equilibrium under delays

Siddharth Pal and Richard J. La*
Department of Electrical and Computer Engineering
and Institute for Systems Research,
University of Maryland, College Park, MD 20742.
Email: spal@umd.edu, hyongla@umd.edu

Abstract

We first propose a general game-theoretic framework for studying engineering systems consisting of *interacting* (sub)systems. Our framework enables us to capture the delays often present in engineering systems as well as asynchronous operations of systems. We model the interactions among the systems using a repeated game and provide a new simple learning rule for the players representing the systems. We show that if all players update their actions via the proposed learning rule, their action profile converges to a pure-strategy Nash equilibrium *with probability one*. Further, we demonstrate that the expected convergence time is finite by proving that the probability that the players have not converged to a pure-strategy Nash equilibrium decays *geometrically* with time.

1 Introduction

We are interested in studying learning in games – In particular, we examine a scenario where a set of agents or players interact with each other repeatedly and update their decisions based on the observed payoff history. This setting may be applicable to a wide range of applications in engineering fields, e.g., [1, 3, 9]. In many engineering systems, it is desirable or even necessary to ensure that the system reaches a desired operating point with *limited* communication and feedback without a centralized controller due to various system constraints.

A popular solution concept that is frequently adopted to approximate the desired operating point in engineering applications is a (pure-strategy) Nash equilibrium (PSNE). By now, there are several classes of learning rules that researchers proposed to help the agents learn through interactions and reach a PSNE under different settings. A set of closely related learning rules are summarized in Section 2 for the interested reader.

A key motivation for our study here is that, in many engineering systems, there are delays experienced by information exchanges or observations. Moreover, (the controllers of) various systems are not necessarily synchronized.

*This work was supported in part by National Institute of Standards and Technology and National Science Foundation.

Therefore, a common assumption in the literature that agents update their actions (nearly) simultaneously and observe consistent information, i.e., the payoffs generated by the same action profile for all agents, needs to be revisited. We take a first step towards addressing this issue in engineering systems. More specifically, we first introduce a general framework that enables us to capture the effects of both time-varying *delays* and *asynchrony* present in many engineering systems. Secondly, we propose a new simple adaptation strategy or learning rule with provable convergence properties. The proposed rule is quite intuitive in that agents may change their actions only when they have an incentive to do so or are exploring the game structure.

We prove that, if all agents employ the proposed rule, under a set of mild technical assumptions, the action profile selected by the agents converges to a PSNE almost surely (or with probability one). Furthermore, the probability that the agents have not reached a PSNE *decays geometrically* with time, thereby proving that the expected convergence time is finite. To the best of our knowledge, our work is the first one that (i) introduces a general framework for modeling more realistic engineering systems with delays and asynchronous operation of systems and (ii) proposes a simple learning rule with provable almost sure convergence to a PSNE in the presence of delays.

The rest of the paper is organized as follows: Section 2 provides a short summary of works that are closely related to our study. Section 3 describes the strategic-form repeated game setting we adopt. The system model, specifically the mathematical formulation of the delayed setting, is discussed in Section 4. The proposed rule is described in Section 5, followed by the main results in Section 6.

2 Related Literature

There is already a large volume of literature on learning in games; some of the earlier models and results can be found in, for instance, a manuscript by Fudenberg and Levine [5], and more recent works are well documented in a manuscript by Bianchi and Lugosi [2]. Learning rules in games in general aim to help agents learn equilibrium conditions in games, oftentimes with special structure, e.g., identical interest games, potential games (PGs), weakly acyclic games (WAGs), and congestion games.

While a PSNE is not guaranteed to exist in an arbitrary game, it is shown to exist for PGs [20]. This led to further research in this field with problems being formulated as PGs. For instance, Arslan et al. [1] model the autonomous vehicle-target assignment problem as a PG and also discuss procedures for vehicle utility design. They present two learning rules and show their convergence to PSNEs. In [13], Marden et al. study large-scale games with many players with large strategy spaces. They generalize the notion of PGs and propose a learning rule that ensures convergence to a PSNE in the class of games they consider with applications to congestion games.

The WAGs were first studied in a systemic manner in [24]. Since then, there has been considerable interest in WAGs. For instance, Marden et al. [10] establish the relations between cooperative control problems (e.g., consensus problem) and game theoretic models. They propose the better-reply-with-inertia dynamics and apply it to a class of games, which they call *sometimes weakly acyclic games*, to address time-varying objective functions and action sets. We note that there are other payoff-based learning rules with provable convergence to Nash equilibria (with

high probability) in WAGs, e.g., [12, 14]. In [17], Pal and La consider a generalization of WAGs, called *generalized weakly acyclic games* (GWAGs), which contain WAGs as special cases. They also propose a simple learning rule that guarantees almost sure convergence of agents' action profile to a PSNE in GWAGs.

Another well-known class of learning rules is based on *regrets*. Regrets capture the additional payoff an agent could have received by playing a different action. There are several regret-based learning rules, e.g., [6, 7, 8], which guarantee convergence to either Nash equilibria or correlated equilibria in an appropriate sense. In [11], Marden et al. propose regret-based dynamics that achieve almost sure convergence to a strict PSNE in WAGs.

The study of (distributed) systems with delays has been of significant interest to the control theory community; please see [21] for a model of asynchronous distributed computation that motivated our work to some extent. Fang and Antsaklis [4] study the consensus problem of discrete-time multi-agent systems in an asynchronous framework. Xiao and Wang [22] extend the above work by investigating consensus problems in the presence of time-varying communication network topologies and information delays. In addition to the studies in control theoretic settings, researchers also studied the effects of delays in evolutionary games, e.g., [16, 23]. In this paper we consider the repeated game setting and study the convergence of action profile to a PSNE under an asynchronous framework motivated by the earlier studies.

3 Learning in games

In this section, we first define the strategic-form repeated game we adopt for our study.

A. Finite stage game: Let $\mathcal{P} = \{1, 2, \ldots, n\}$ be the set of agents or players of a *finite* game $\mathbf{G}$. The pure action space of agent $i \in \mathcal{P}$ and the joint action space of all agents are denoted by $\mathcal{A}_i = \{1, 2, \ldots, A_i\}$ and $\mathcal{A} := \prod_{i \in \mathcal{P}} \mathcal{A}_i$, respectively.[1] The payoff function of agent i is given by $U_i : \mathcal{A} \to \mathbb{R} := (-\infty, \infty)$.

An agent $i \in \mathcal{P}$ chooses its action according to a probability distribution $\mathbf{p}_i \in \Delta(\mathcal{A}_i)$, where $\Delta(\mathcal{A}_i)$ denotes the probability simplex over $\mathcal{A}_i$. If the probability distribution $\mathbf{p}_i$ is degenerate, i.e., it puts probability of one on a single action, it is called a *pure* strategy. Otherwise, agent i is said to play a *mixed* strategy.

Given an action profile $\mathbf{a} = (a_1, a_2, \ldots, a_n) \in \mathcal{A}$, $\mathbf{a}_{-i}$ denotes the action profile of all the agents other than agent i, i.e., $\mathbf{a}_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$. Similarly, given $J \subset \mathcal{P}$, $\mathbf{a}_J = (a_j,\ j \in J)$ denotes the set of actions picked by agents in J. An action profile $\mathbf{a}^\star \in \mathcal{A}$ is a PSNE of $\mathbf{G}$ if, for every agent $i \in \mathcal{P}$,

$$U_i(a_i^\star, \mathbf{a}_{-i}^\star) \geq U_i(a_i, \mathbf{a}_{-i}^\star) \text{ for all } a_i \in \mathcal{A}_i \setminus \{a_i^\star\}. \tag{1}$$

The PSNE is said to be *strict* if the inequality in (1) is strict for all agents $i \in \mathcal{P}$.

B. Infinitely repeated game: In an (infinitely) repeated game, the above finite stage game $\mathbf{G}$ is repeated at each time $t \in \mathbb{N} := \{1, 2, \ldots\}$. At time t, agent $i \in \mathcal{P}$ chooses its action $a_i(t)$ according to a strategy $\mathbf{p}_i(t) \in \Delta(\mathcal{A}_i)$.

We denote the action profile chosen by the agents at time t by $\mathbf{a}(t) = (a_i(t),\ i \in \mathcal{P})$. Based on the selected action

[1] It is reasonable to assume that the action spaces are finite because, in many cases, both the observations and actions are *quantized* to facilitate the information exchange in real systems.

profile $\mathbf{a}(t)$, each agent $i \in \mathcal{P}$ receives a payoff $U_i(\mathbf{a}(t))$ at time t. The agents are allowed to adapt their strategies based on the history of payoff information.

4 System model

As mentioned in Section 1, we are interested in scenarios in which the actions adopted by agents may not take effect immediately and/or the payoff information generated by the system(s) may not be observable to the agents right away. Obviously, there are many different ways in which one can capture and model the effects of such delays. In this section, we describe the system model we assume for our study.

First, we assume that there are multiple (sub)systems: For each agent $i \in \mathcal{P}$, there is a separate (sub)system that is responsible for generating the payoff of the agent. We call it system i. This model describes, for instance, a large system consisting of multiple *interacting* systems, where the agents can be viewed as (controllers of) interacting systems that comprise the overall system. An example with three systems is shown in Figure 1. In such scenarios, depending on the structure of the system, the new actions taken by various systems may experience varying delays before affecting other systems.

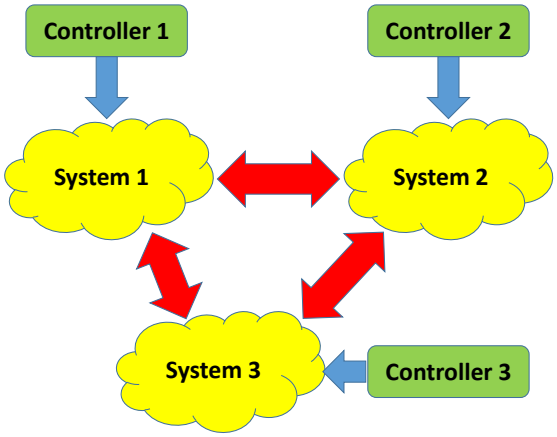

Figure 1: System model.

In the system model, there are two types of delays – (i) forward delays and (ii) feedback delays. The forward delays refer to the amount of time that elapses before agents' actions go into effect after they are adopted. On the other hand, the feedback delays are the amount of time it takes for payoff information to become available for the agents to observe (after it is generated). These are illustrated in Figure 2 with two systems. In the figure, after agent i, $i = 1, 2$, updates its action, the forward delay experienced before system i sees the new action is shown as a (red or green) solid arrow (with label 'A'), whereas the time that elapsed before the other system sees it appears as a dotted arrow. The feedback experienced by payoff information generated by a system before the corresponding agent sees it is shown as a (blue or purple) solid arrow (with label 'F').

We shall model these delays using sequences of random variables (rvs) as follows. First, we assume that the agents

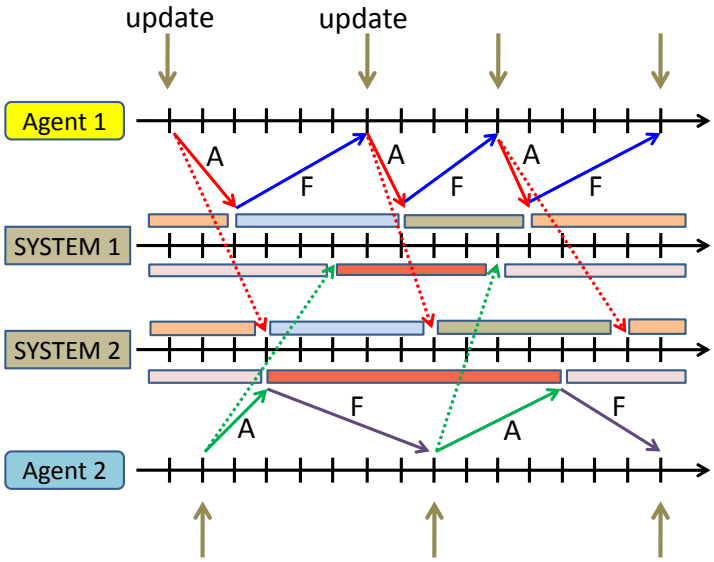

Figure 2: Forward and feedback delays in the system.

choose their initial actions $\mathbf{a}(1) = (a_i(1); i \in \mathcal{P})$ according to some (joint) distribution $\mathcal{G}$. The times at which each agent updates its action are given by some discrete-time stochastic process: For each $i \in \mathcal{P}$, let $\mathcal{T}^i := \{T_k^i; k \in \mathbb{N}\}$ be a *strictly increasing* sequence of times at which agent i updates its action with $T_1^i = 1$. In other words, T_k^i is the time at which agent i chooses its action for the kth time.

The inter-update times are denoted by $U_k^i := T_{k+1}^i - T_k^i$, $k \in \mathbb{N}$. Each inter-update time U_k^i is given by a sum of two rvs – X_k^i and Y_k^i. The rv X_k^i models the forward delay experienced by the kth action of agent i chosen at time T_k^i, i.e., $a_i(T_k^i)$; we assume that the kth action $a(T_k^i)$ goes into effect at system i after experiencing a forward delay of X_k^i at time $T_k^i + X_k^i =: R_k^i$, at which point system i generates payoff feedback information. This payoff feedback experiences a delay of Y_k^i and is received by agent i at time $T_{k+1}^i (= R_k^i + Y_k^i)$, which agent i uses to select the $(k+1)$th action at time T_{k+1}^i. For notational simplicity, we denote the pair (X_k^i, Y_k^i) by Z_k^i and the sequences $\{Z_k^i, k \in \mathbb{N}\}$ and $\{R_k^{i,j},\ k \in \mathbb{N}\}$ by $\mathcal{Z}^i$ and $\mathcal{R}^{i,j}$, respectively.

Similarly, the actions chosen by agent i at time T_k^i may not be seen by system j, $j \neq i$, immediately; instead, it takes effect at system j at time $T_k^i + V_k^{i,j} =: R_k^{i,j}$, where $V_k^{i,j}$ is the forward delay that action $a_i(T_k^i)$ experiences before it begins to affect system j. For every $i, j \in \mathcal{P}, i \neq j$, let $\mathcal{V}^{i,j} = \{V_k^{i,j}, k \in \mathbb{N}\}$.

Because the initial actions chosen by the agents at time $t = 1$ may not take effect at the systems right away, in order to complete the system model, we impose the following initial conditions: For each system i, we assume that there is some action profile $\bar{\mathbf{a}}^i \in \mathcal{A}$ that is in effect at the system. In other words, for each $i \in \mathcal{P}$, system i sees action $\bar{a}_i^i$ from agent i till $R_1^i = 1 + X_1^i$ and $\bar{a}_j^i$ from agent j, $j \neq i$, till $R_1^{j,i} = 1 + V_1^{j,i}$.

In order to describe how the payoff feedback is generated by the systems for each agent $i \in \mathcal{P}$, we introduce the

following variables. For all $t \in \mathbb{N}$ and $i, j \in \mathcal{P}$,

$$\tau_j^i(t) = \begin{cases} \max\{T_k^i \in \mathcal{T}^i \mid T_k^i \leq t\} & \text{if } i = j, \\ \max\{R_k^{j,i} \in \mathcal{R}^{j,i} \mid R_k^{j,i} \leq t\} & \text{if } i \neq j. \end{cases}$$

In other words, $\tau_j^i(t)$, $j \neq i$, denotes the last time a new action of agent j went into effect at system i prior to time t. Otherwise, $\tau_i^i(t)$ is the last time agent i updated its action before time t. For notational ease, we define $\mathbf{a}^i(t)$, $t \in \mathbb{N}$, to be the action profile of other agents, which is in effect at system i at time t, i.e., $\mathbf{a}^i(t) = (a_j(\tau_j^i(t)),\ j \in \mathcal{P} \setminus \{i\})$. Therefore, the payoff feedback generated for agent i at time t is based on $\mathbf{a}^i(t)$.

In order to make progress, we assume that the following assumptions hold.

A1 The rvs $\{\mathcal{Z}^i, \mathcal{V}^{i,j}, j \neq i\}$ for different agents $i \in \mathcal{P}$ are mutually independent.

A2. For every $i \in \mathcal{P}$,

$$\mathbf{P}\left[U_k^i < \infty, V_k^{i,j} < \infty \text{ for all } j \neq i \text{ and } k \in \mathbb{N}\right] = 1.$$

A3. For all $i, j \in \mathcal{P}$, $j \neq i$,

$$\mathbf{P}\left[R_k^{i,j} < R_{k+1}^{i,j} \text{ for all } k \in \mathbb{N}\right] = 1.$$

A4. Let $\mathcal{F}_k^i := \big((Z_\ell^i, V_\ell^{i,j}, j \neq i), 1 \leq \ell \leq k\big)$. There exist $\eta > 0$ and $\Delta_\eta < \infty$ such that, for all $i \in \mathcal{P}$, $k \in \mathbb{N}$ and $\tau \in \mathbf{Z}_+ := \{0, 1, 2, \ldots\}$, with probability one (w.p.1)

$$\mathbf{P}\left[U_{k+1}^i \leq \Delta_\eta + \tau, V_{k+1}^{i,j} \leq \Delta_\eta + \tau \text{ for all } j \neq i \mid U_{k+1}^i \geq \tau, \mathcal{F}_k^i\right] \geq \eta.$$

Assumption A1 states that all forward and backward delays experienced by actions picked by different agents are mutually independent. Assumption A2 ensures that they are proper random variables. Assumption A3 guarantees that every system sees the actions from an agent in the same order they were adopted. Assumption A4 essentially implies that the distributions of rvs $\{U_k^i, V_k^{i,j}, j \neq i\}$ do not have a heavy tail.

We feel that, while these assumptions are technical, they are not very restrictive and likely to hold in many cases of practical interest. For example, constant delays and geometric delays satisfy the aforementioned assumptions.

5 Proposed Algorithm

In this section, we present our proposed algorithm, called the Reduced Simple Experimentation with Monitoring (RSEM), which can be viewed as a special case of the SEM algorithm outlined in [18]. Under the RSEM algorithm, at each time $t \in \mathbb{N}$, every agent is at one of two possible states – Explore (E) or Converged (C). We denote the state of agent $i \in \mathcal{P}$ at time t by $s_i(t)$. Let $\mathcal{S} := \{E, C\}^n$, and the state vector $\mathbf{s}(t) := (s_i(t),\ i \in \mathcal{P}) \in \mathcal{S}$. The rule governing the update of an agent's state will be explained shortly.

5.1 Action updates

In this subsection, we first describe how the agents choose their actions according to the payoff feedback they receive at the times of updates. Fix $\delta > 0$. For each $i \in \mathcal{P}$, let $\Delta_\delta(\mathcal{A}_i)$ denote the subset of the probability simplex over $\mathcal{A}_i$ such that, for all $\mathbf{q} = (q(a_i), a_i \in \mathcal{A}_i) \in \Delta_\delta(\mathcal{A}_i)$, we have $q(a_i) \geq \delta$ for all $a_i \in \mathcal{A}_i$. Initially, at time $T_1^i = 1$, the agents choose their actions according to some (joint) distribution as stated before. For $k \geq 2$, agent $i \in \mathcal{P}$ updates its action at time T_k^i according to the rule provided below. We assume that agent i continues to play action $a_i(T_k^i)$ between T_k^i and $T_{k+1}^i - 1$, i.e., $a_i(t) = a_i(T_k^i)$ for all $t \in \{T_k^i, \ldots, T_{k+1}^i - 1\}$.

Action Selection Rule: Fix $\delta \in (0, 1/(\max_{i\in\mathcal{P}} |\mathcal{A}_i|))$.

---

1. if $s_i(T_k^i) = E$
   - pick $\mathbf{q} \in \Delta_\delta(\mathcal{A}_i)$;
   - choose each action $a_i \in \mathcal{A}_i$ with probability $q(a_i)$ and set $a_i(T_k^i) = a_i$;
2. else (i.e., $s_i(T_k^i) = C$)
   - set $a_i(T_k^i) = a_i(T_k^i - 1)$;

It is clear that, under RSEM, an agent may choose a new action only if it is at state E. Otherwise, it continues to play the same action employed at the previous time.

5.2 State dynamics

As explained in the previous subsection, under the RSEM rule, the state of an agent plays a key role in its action selection. Hence, the dynamics of $\mathbf{s}(t)$, $t \in \mathbb{N}$, play a major role in the algorithm. In this subsection, we explain how the agents update their states based on the received payoff feedback.

At time $t = 1$, we assume that all agents are at state E, i.e., $\mathbf{s}(1) = (E, E, \ldots, E)$. Because agent i receives new payoff information only at T_k^i, $k \in \mathbb{N}$, we allow the agents to update their states only at T_k^i, $k \in \mathbb{N}$, just *before* choosing the action. We assume that the payoff information available to agent i at time T_k^i, $k \geq 2$, is of the form $\mathbf{I}_i(T_k^i) = (U_i(a_i, \mathbf{a}^i(R_{k-1}^i));\ a_i \in \mathcal{A}_i)$. In other words, agent i knows the payoff it would receive for each available action given the action profile of the other agents in effect (at the agent i's system) at the time the payoff feedback is generated, i.e., R_{k-1}^i.

The state of agent i at time T_k^i depends on (i) the payoff information vector $\mathbf{I}_i(T_k^i)$ if $s_i(T_{k-1}^i) = E$ and (ii) the payoff information vectors available at time T_{k-1}^i and T_k^i if $s_i(T_{k-1}^i) = C$. We assume $\mathbf{I}_i(T_1^i) = (0, \ldots, 0)$ for all $i \in \mathcal{P}$. First, for each agent $i \in \mathcal{P}$, define a mapping $BR_i : \mathbb{R}^{|\mathcal{A}_i|} \times \mathcal{A}_i \to 2^{\mathcal{A}_i}$, where

$$BR_i(\mathbf{I}, a_i) = \left\{ a_i^* \in \mathcal{A}_i \;\middle|\; I(a_i^*) > I(a_i) \right\},\ \mathbf{I} = (I(a_i'),\ a_i' \in \mathcal{A}_i) \in \mathbb{R}^{|\mathcal{A}_i|} \text{ and } a_i \in \mathcal{A}_i.$$

It is clear from the definition that $BR_i(\mathbf{I}, a_i)$ is the set of indices of the values in $\mathbf{I}$ which are larger than the value assigned to a_i.

State Update Rule: Fix $p \in (0, 1)$.

S1. if $s_i(T^i_{k-1}) = E$

- if $BR_i(\mathbf{I}_i(T^i_k), a_i(T^i_k - 1)) \neq \emptyset$, then $s_i(T^i_k) = E$
- else $\big($i.e., $BR_i(\mathbf{I}_i(T^i_k), a_i(T^i_k - 1)) = \emptyset\big)$

$$s_i(T^i_k) = \begin{cases} E & \text{with probability } p \\ C & \text{with probability } 1 - p \end{cases}$$

S2. else (i.e., $s_i(T^i_{k-1}) = C$)

- if $BR_i(\mathbf{I}_i(T^i_k), a_i(T^i_k - 1)) \neq \emptyset$, then $s_i(T^i_k) = E$
- else $\big($ i.e., $BR_i(\mathbf{I}_i(T^i_k), a_i(T^i_k - 1)) = \emptyset\big)$
  - if $\mathbf{I}_i(T^i_k) \neq \mathbf{I}_i(T^i_{k-1})$, then $s_i(T^i_k) = E$
  - else $\big($ i.e., $\mathbf{I}_i(T^i_k) = \mathbf{I}_i(T^i_{k-1})\big)$, then $s_i(T^i_k) = C$

In a nutshell, agent i transitions to or remains at state E if either $BR_i(\mathbf{I}_i(T^i_k), a_i(T^i_k - 1)) \neq \emptyset$ or $\mathbf{I}_i(T^i_k) \neq \mathbf{I}_i(T^i_{k-1})$. The second condition means that the payoffs that agent i would receive using available actions have changed from the last time of update. Hence, the agent prefers to remain in the Explore state.

6 Main results

Assumption 1 *(Interdependence Assumption) For every $\mathbf{a} = (a_i,\ i \in \mathcal{P}) \in \mathcal{A}$ and $J \subsetneq \mathcal{P}$, there exist an agent $i \notin J$ and $\mathbf{a}^*_J \in \prod_{j \in J} \mathcal{A}_j$ such that $U_i(\mathbf{a}) \neq U_i(\mathbf{a}^*_J, \mathbf{a}_{-J})$.*

Assumption 1 simply states that, given any action profile and a strict subset J of the agents, we can find another agent $i \notin J$ whose payoff would change if the agents in J changed their actions to $\mathbf{a}^*_J$. Put differently, it implies that it is not possible to partition the set of agents into two subsets that do not interact with each other. The interdependence assumption has been used in the literature, for instance, to prove the convergence of action profile to efficient equilibria or Pareto optimal point [15, 19, 25]. In addition, Pal and La [18] show under the interdependence assumption that long run equilibria under their proposed Simple Experimentation with Monitoring (SEM) rule are PSNEs with a certain level of *resilience*, which can be chosen using a tunable parameter of the SEM rule.

Under this interdependence assumption, we can show that, if all agents update their actions according to the RSEM rule, the action profile converges almost surely to a PSNE.

Theorem 6.1 *Suppose that the game* $\mathbf{G}$ *satisfies the interdependence assumption and has a nonempty set of PSNE(s) denoted by* $\mathcal{A}_{NE}$ *and that Assumptions A1-A4 hold. Let* $\mathbf{a}(t), t \in \mathbb{N}$*, be the sequence of action profiles generated by the RSEM rule. Then, the action profile* $\mathbf{a}(t)$ *converges to a PSNE w.p.1. In other words, w.p.1, there exist* $T^* < \infty$ *and* $\mathbf{a}^\star \in \mathcal{A}_{NE}$ *such that* $\mathbf{a}(t) = \mathbf{a}^\star$ *for all* $t \geq T^*$.

Proof. A proof of the theorem is provided in Appendix A. ■

In addition to the almost sure convergence of the action profile, we can establish that the probability that the action profile has not converged to a PSNE decays geometrically with time t.

Theorem 6.2 *Under the same settings assumed in Theorem 6.1, there exist* $C < \infty$ *and* $\eta \in (0, 1)$ *such that*

$$1 - \mathbf{P}\left[\mathbf{a}(t) \in \mathcal{A}_{NE} \text{ for all } t \geq k\right] \leq C \cdot \eta^k. \tag{2}$$

Theorem 6.2 follows directly from Corollary A.2 in the proof of Theorem 6.1 in Appendix A.

7 Conclusion

We studied the repeated interactions among agents in a wide range of engineering systems with delays in a game-theoretic framework and proposed a new learning rule for allowing the agents or players to converge to a PSNE almost surely. The interactions are modelled as a discrete-time system with a fixed time unit. We are currently working to extend our framework to a continuous-time case where the system delays can be state-dependent as well.

A Proof of Theorem 6.1

The theorem will be proved with the help of several lemmata we introduce. Their proofs are provided in Appendices B through D. Throughout the appendices, $\mathbf{s}^\star = (C, C, \ldots, C)$ and, for any $\mathbf{s} \in \mathcal{S}$,

$$C(\mathbf{s}) = \{i \in \mathcal{P} \mid s_i = C\} \text{ and } E(\mathbf{s}) = \{i \in \mathcal{P} \mid s_i = E\}.$$

The first lemma states that, if the action profile at time t is not a PSNE, then even if all agents are at state C at time t, there is positive probability that at least one agent will transition to state E after $3\Delta_\eta$ periods (where Δ_η is the constant introduced in Assumption A4 in Section 4).

Lemma A.1 *For every* $\mathbf{a} \notin \mathcal{A}_{NE}$ *and* $t \in \mathbb{N}$*,*

$$\mathbf{P}\left[E(\mathbf{s}(t + 3\Delta_\eta)) \neq \emptyset \mid \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a})\right] \geq \zeta_0 > 0. \tag{3}$$

The second lemma shows that, if there is at least one agent at state E at time $t \in \mathbb{N}$, there is positive probability that the number of agents at state E will increase after a finite number of periods.

Lemma A.2 *For every $r \in \{1, 2, \ldots, n-1\}$, there exists $0 < D_1 \leq 4\Delta_\eta$ such that, for every $t \in \mathbb{N}$ and $\mathbf{z} = (\mathbf{s}, \mathbf{a}) \in \mathcal{Z}$ with $|E(\mathbf{s})| = r$, we have*

$$\mathbf{P}\left[|E(\mathbf{s}(t + D_1))| \geq r + 1 \mid \mathbf{z}(t) = \mathbf{z}\right] \geq \varsigma_r > 0. \tag{4}$$

The following corollary now follows from Lemmas A.1 and A.2, by repeatedly applying Lemma A.2 until all agents switch to state E.

Corollary A.1 *There exists $0 < D \leq 4n\Delta_\eta$ such that, for all $\mathbf{z} \in \mathcal{Z} \setminus \mathcal{Z}_{NE}$ and $t \in \mathbb{N}$,*

$$\mathbf{P}\left[|E(\mathbf{s}(t + D)| = n \;\middle|\; \mathbf{z}(t) = \mathbf{z}\right] \geq \mu > 0. \tag{5}$$

The final lemma has two parts; first, it states that, if all agents are at state E at some time t, then for any $\mathbf{z}^\star \in \mathcal{Z}_{NE}$, there is positive probability that they will reach any $\mathbf{z}^\star$ after $3\Delta_\eta$ periods. Second, if the agents are at some $\mathbf{z}^\star \in \mathcal{Z}_{NE}$ at time t, with positive probability they will remain at $\mathbf{z}^\star$ for good.

Lemma A.3 *(i) Suppose that $\mathbf{z}(t) = \mathbf{z} = (\mathbf{s}, \mathbf{a})$ where $|E(\mathbf{s})| = n$, i.e., all agents are at state E. Then, for all $\mathbf{z}^\star = (\mathbf{s}^\star, \mathbf{a}^\star) \in \mathcal{Z}_{NE}$, we have*

$$\mathbf{P}\left[\mathbf{z}(t + 4\Delta_\eta) = \mathbf{z}^\star \mid \mathbf{z}(t) = \mathbf{z}\right] \geq \rho_1 > 0.$$

(ii) For every $\mathbf{z}^\star \in \mathcal{Z}_{NE}$,

$$\mathbf{P}\left[\mathbf{z}(t') = \mathbf{z}^\star \textit{ for all } t' \geq t \mid \mathbf{z}(t) = \mathbf{z}^\star\right] \geq \rho_2 > 0.$$

The following corollary is a consequence of the above lemmas.

Corollary A.2 *There exist $0 < \tilde{D} \leq 4(n + 1)\Delta_\eta$ such that, for all $\mathbf{z} \in \mathcal{Z} \setminus \mathcal{Z}_{NE}$, $\mathbf{z}^\star \in \mathcal{Z}_{NE}$ and $t \in \mathbb{N}$, we have*

$$\mathbf{P}\left[\mathbf{z}(t') = \mathbf{z}^\star \textit{ for all } t' \geq t + \tilde{D} \;\middle|\; \mathbf{z}(t) = \mathbf{z}\right] \geq \tilde{\mu} > 0. \tag{6}$$

We now proceed with the proof of Theorem 6.1. Lemma A.1 shows that, if the action profile is not a PSNE at time t, then after a finite number of periods, at least one agent will be at state E. Lemma A.2 then claims that, whenever there is at least one agent at state E, after finitely many periods, all agents will be state E (Corollary A.1). Once all agents are at state E, Lemma A.3 asserts that they can reach any $\mathbf{z}^\star \in \mathcal{Z}_{NE}$ with positive probability after a finite number of periods and stay there forever. Since this argument can be made starting with any $\mathbf{z} \notin \mathcal{Z}_{NE}$, it is clear that the number of periods spent at $\mathbf{z} \in \mathcal{Z} \setminus \mathcal{Z}_{NE}$ will be finite with probability 1, and the theorem follows.

B Proof of Lemma A.1

Because $\mathbf{a} \notin \mathcal{A}_{NE}$, when the agents adopt $\mathbf{a}$, there is at least one agent, say agent i^*, with an incentive to deviate from a_{i^*}. We will prove that the state of agent i will transition to E with positive probability after $3\Delta_\eta$ periods.

Let $T_\ell(t) := \{t + \ell \cdot \Delta_\eta + 1, \ldots, t + (\ell+1)\Delta_\eta\}$, $\ell \in \mathbf{Z}_+$. For each $i \in \mathcal{P}$, let $k_i^\star : \mathbb{N} \to \mathbf{Z}_+$, where

$$k_i^\star(t) = \max\{k \in \mathbb{N} \mid T_k^i \leq t\}.$$

We define the following four events:

$$\begin{aligned}
\mathcal{E}_1 &= \{\tau_i^i(t) + V_{k_i^\star(t)}^{i,j} \leq t + \Delta_\eta \text{ for all } i, j \in \mathcal{P}, i \neq j\} \\
\mathcal{E}_2 &= \{\mathcal{T}_{i^*} \cap T_\ell(t) \neq \emptyset,\ \ell = 1, 2\} \\
\mathcal{E}_3 &= \{a_i(t') = a_i \text{ for all } i \in E(\mathbf{s}(t')),\ t' = t+1, \ldots, t + 3\Delta_\eta\} \\
\mathcal{E}_4 &= \{i^* \in E(\mathbf{s}(t'+1)) \text{ if } i^* \in E(\mathbf{s}(t')) \text{ for all } t' = t, \ldots, t + 3\Delta_\eta - 1\}.
\end{aligned}$$

The event $\mathcal{E}_1$ implies that the action profile at time t is seen by all systems by time $t + \Delta_\eta$. The second event $\mathcal{E}_2$ simply states that agent i^* updates its action at least once in each of the intervals T_1 and T_2. Event $\mathcal{E}_3$ requires any agent at state E between $t+1$ and $t + 3\Delta_\eta$ to choose the same action it did at time t (which will happen with strictly positive probability). Finally, the fourth event $\mathcal{E}_4$ demands that the agent i^*, once it switches to state E (which will happen by time $t + 3\Delta_\eta$ if the events $\mathcal{E}_1$ through $\mathcal{E}_3$ take place because $\mathbf{a} \in \mathcal{A}_{NE}$), remain at state E till time $t + 3\Delta_\eta$.

We use the following lower bound to complete the proof.

$$\begin{aligned}
&\mathbf{P}\left[E(\mathbf{a}(t + 3\Delta_\eta)) \neq \emptyset \mid \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a})\right] \\
&\geq \mathbf{P}\left[E(\mathbf{a}(t + 3\Delta_\eta)) \neq \emptyset, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4 \mid \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a})\right] \\
&= \mathbf{P}\left[E(\mathbf{a}(t + 3\Delta_\eta)) \neq \emptyset \mid \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4, \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a})\right] \mathbf{P}\left[\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4 \mid \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a})\right] \qquad (7)
\end{aligned}$$

From the explanations of the events $\mathcal{E}_1$ through $\mathcal{E}_4$ above, if these four events take place, it is clear that at least one agent, namely agent i^*, will be at state E at time $t + 3\Delta_\eta$. Hence, the first conditional probability in (7) is one. Hence, if we can show that the second conditional probability is also positive, the theorem is proved.

First, we rewrite $\mathbf{P}\left[\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4 \mid \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a})\right]$ as follows.

$$\begin{aligned}
&\mathbf{P}\left[\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4 \mid \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a})\right] \\
&= \mathbf{P}\left[\mathcal{E}_1, \mathcal{E}_2 \mid \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a})\right] \cdot \mathbf{P}\left[\mathcal{E}_3, \mathcal{E}_4 \mid \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a}), \mathcal{E}_1, \mathcal{E}_2\right]. \qquad (8)
\end{aligned}$$

The first term in (8) is lower bounded by η^{n+2} from Assumption A4. In addition, from the description of the algorithm, the second term is lower bounded by $\delta^{3n\Delta_\eta} \cdot p^{3\Delta_\eta}$. Putting together, we have the following lower bound.

$$\mathbf{P}\left[E(\mathbf{a}(t + 3\Delta_\eta)) \neq \emptyset \mid \mathbf{z}(t) = (\mathbf{s}^\star, \mathbf{a})\right] \geq (\delta^n \cdot p)^{3\Delta_\eta} \cdot \eta^{n+2} =: \zeta_0 > 0$$

C Proof of Lemma A.2

We consider two cases.

c1. There exists at least one agent $i^+ \in C(\mathbf{s}(t))$ such that $\mathbf{I}_{i^+}(\tau_{i^+}^{i^+}(t)) \neq (U_{i^+}(a_{i^+}, \mathbf{a}_{-i^+}(t));\ a_{i^+} \in \mathcal{A}_{i^+})$. In other words, its payoff information vector received at the last time of update is different from what it would receive if its system generated the payoff information vector in response to the current action profile at time t.

c2. There is no such agent, i.e., for all $i \in C(\mathbf{s}(t))$, we have $\mathbf{I}_i(\tau_i^i(t)) = (U_i(a_i, \mathbf{a}_{-i}(t));\ a_i \in \mathcal{A}_i)$.

Case c1: First, we define the following events.

$$\mathcal{E}_1' = \{\tau_i^i(t) + V_{k_i^*(t)}^{i,j} \leq t + \Delta_\eta \text{ for all } i, j \in \mathcal{P}, i \neq j\}$$
$$\mathcal{E}_2' = \{\mathcal{T}_{i^+} \cap T_\ell(t) \neq \emptyset,\ \ell = 1, 2\} \text{ (where } T_\ell, \ell = 1, 2, \text{ are as defined in Appendix B)}$$
$$\mathcal{E}_3' = \{a_i(t') = a_i \text{ for all } i \in E(\mathbf{s}(t')),\ t' = t+1, \ldots, t+3\Delta_\eta\}$$
$$\mathcal{E}_4' = \{E(\mathbf{s}(t')) \subseteq E(\mathbf{s}(t'+1)) \text{ for all } t' = t, \ldots, t+3\Delta_\eta - 1\}$$

Using a similar argument used in Appendix B, we obtain

$$\begin{aligned}
&\mathbf{P}\left[|E(\mathbf{s}(k+3\Delta_\eta))| \geq r+1 \mid \mathbf{z}(k) = \mathbf{z}\right] \\
&\geq \mathbf{P}\left[|E(\mathbf{s}(k+3\Delta_\eta))| \geq r+1, \mathcal{E}_1', \mathcal{E}_2', \mathcal{E}_3', \mathcal{E}_4' \mid \mathbf{z}(k) = \mathbf{z}\right] \\
&= \mathbf{P}\left[|E(\mathbf{s}(k+3\Delta_\eta))| \geq r+1 \mid \mathcal{E}_1', \mathcal{E}_2', \mathcal{E}_3', \mathcal{E}_4', \mathbf{z}(k) = \mathbf{z}\right] \cdot \mathbf{P}\left[\mathcal{E}_1', \mathcal{E}_2', \mathcal{E}_3', \mathcal{E}_4' \mid \mathbf{z}(k) = \mathbf{z}\right]. \qquad (9)
\end{aligned}$$

From the assumption on agent i^+, if events $\mathcal{E}_1'$ through $\mathcal{E}_4'$ take place, by time $t + 3\Delta_\eta$, it will switch its state to E. Thus, the first conditional probability in (9) is equal to one. Also, following an analogous argument in Appendix B, the second conditional probability is lower bounded by $(\delta \cdot p)^{3n\Delta_\eta} \cdot \eta^{n+2} > 0$.

Case c2: Recall that, in this case, for every agent $i \in C(\mathbf{s}(t))$, we have $I_i(\tau_i^i(t)) = (U_i(a_i, \mathbf{a}_{-i}(t));\ a_i \in \mathcal{A}_i)$. Assumption 1 implies that there exists $i' \notin E(\mathbf{s}(t)) =: J_1$ and $\mathbf{a}_{J_1}^*$ such that if the agents in J_1 adopt the actions in $\mathbf{a}_{J_1}^*$ while the other agents choose the same action stipulated by $\mathbf{a}(t)$, then agent i''s payoff information vector changes. As we will prove, this implies that there is positive probability that agent i' will switch its state to E after $4\Delta_\eta$ periods.

To this end, we define the following events

$$\mathcal{E}_1^+ = \{\tau_i^i(t+\Delta_\eta) + V_{k_i^*(t+\Delta_\eta)}^{i,j} \leq t + 2\Delta_\eta \text{ for all } i, j \in \mathcal{P}, i \neq j\}$$
$$\mathcal{E}_2^+ = \{\mathcal{T}_{i'} \cap T_\ell(t) \neq \emptyset,\ \ell = 2, 3\}$$
$$\mathcal{E}_3^+ = \{E(\mathbf{s}(t')) \subseteq E(\mathbf{s}(t'+1)) \text{ for all } t' = t, \ldots, t+4\Delta_\eta - 1\}$$
$$\mathcal{E}_4^+ = \{\mathcal{T}_i \cap T_0(t) \neq \emptyset \text{ for all } i \in J_1\}$$
$$\mathcal{E}_5^+ = \{a_i(t') = \tilde{a}_i \text{ for all } i \in \mathcal{P} \text{ and } t' = T_{k_i^*(t)+1}^i, \ldots, t+4\Delta_\eta\}$$

where

$$\tilde{a}_i = \begin{cases} a_i(t) & \text{if } i \notin J_1, \\ a_i^* & \text{if } i \in J_1. \end{cases}$$

The rest of the proof follows from a similar argument.

$$\begin{aligned}
&\mathbf{P}\left[|E(\mathbf{s}(k+3\Delta_\eta))| \geq r+1 \mid \mathbf{z}(k) = \mathbf{z}\right] \\
&\geq \mathbf{P}\left[|E(\mathbf{s}(k+3\Delta_\eta))| \geq r+1, \mathcal{E}_1^+, \mathcal{E}_2^+, \mathcal{E}_3^+, \mathcal{E}_4^+, \mathcal{E}_5^+ \mid \mathbf{z}(k) = \mathbf{z}\right] \\
&= \mathbf{P}\left[|E(\mathbf{s}(k+3\Delta_\eta))| \geq r+1 \mid \mathcal{E}_1^+, \mathcal{E}_2^+, \mathcal{E}_3^+, \mathcal{E}_4^+, \mathcal{E}_5^+, \mathbf{z}(k) = \mathbf{z}\right] \cdot \mathbf{P}\left[\mathcal{E}_1^+, \mathcal{E}_2^+, \mathcal{E}_3^+, \mathcal{E}_4^+, \mathcal{E}_5^+ \mid \mathbf{z}(k) = \mathbf{z}\right].
\end{aligned}$$

From the definitions of the above events, the first conditional probability is one because agent i' will have switched its state to E (from C) by time $t + 4\Delta_\eta$ if the events $\mathcal{E}_1^+$ through $\mathcal{E}_5^+$ take place. In addition, the second conditional probability is lower bounded by $\eta^{n+2+r} (\delta \cdot p)^{4n\Delta_\eta} = \varsigma_r (< \zeta_0)$.

D Proof of Lemma A.3

We first prove Lemma A.3(i). First, define the following events.

$$\begin{aligned}
\mathcal{E}_1^\# &= \{\mathcal{T}_i \cap T_0(t) \neq \emptyset \text{ for all } i \in \mathcal{P}\} \\
\mathcal{E}_2^\# &= \{\tau_i^i(t + \Delta_\eta) + V^{i,j}_{k_i^*(t+\Delta_\eta)} \leq t + 2\Delta_\eta \text{ for all } i, j \in \mathcal{P}, i \neq j\} \\
\mathcal{E}_3^\# &= \{a_i(t') = a_i^\star \text{ for all } i \in \mathcal{P} \text{ and } t' = T^i_{k_i^*(t)+1}, \ldots, t + 4\Delta_\eta\} \\
\mathcal{E}_4^\# &= \{\mathcal{T}_i \cap T_\ell(t) \neq \emptyset \text{ for all } i \in \mathcal{P} \text{ and } \ell = 2, 3\} \\
\mathcal{E}_5^\# &= \{s_i(T^i_{k_i^*(t+3\Delta_\eta)+1}) = C \text{ for all } i \in E(\mathbf{s}(t + 3\Delta_\eta))\}
\end{aligned}$$

Events $\mathcal{E}_1^\#, \mathcal{E}_2^\#, \mathcal{E}_3^\#$ together imply that all agents update their actions to $a_i^\star$ during $T_0(t)$ and their actions go into effect by $t + 2\Delta_\eta$. Events $\mathcal{E}_4^\#$ and $\mathcal{E}_5^\#$ state that all agents update at least once during the intervals $T_2(t)$ and $T_3(t)$, and switch their state to C. Hence, together these events mean that all agents are at state C and adopting the PSNE $\mathbf{a}^\star$ at time $t + 4\Delta_\eta$. Therefore,

$$\begin{aligned}
&\mathbf{P}\left[\mathbf{z}(t + 4\Delta_\eta) = \mathbf{z}^\star \mid \mathbf{z}(t) = \mathbf{z}\right] \\
&\geq \mathbf{P}\left[\mathbf{z}(t + 4\Delta_\eta) = \mathbf{z}^\star \mid \mathcal{E}_1^\#, \mathcal{E}_2^\#, \mathcal{E}_3^\#, \mathcal{E}_4^\#, \mathcal{E}_5^\#, \mathbf{z}(t) = \mathbf{z}\right] \cdot \mathbf{P}\left[\mathcal{E}_1^\#, \mathcal{E}_2^\#, \mathcal{E}_3^\#, \mathcal{E}_4^\#, \mathcal{E}_5^\# \mid \mathbf{z}(t) = \mathbf{z}\right] \qquad (10)
\end{aligned}$$

As argued above, the first conditional probability in (10) is one. The second conditional probability can be lower bounded by $\eta^{4n} \cdot \delta^{4n\Delta_\eta} \cdot (1 - p)^n$. This completes the proof of Lemma A.3(i).

Before we prove Lemma A.3(ii), note that even though the agents are at $\mathbf{z}^\star \in \mathcal{Z}_{NE}$ at time t, it is still possible for some agents to transition to state E due to the delays in the system. Hence, the conditional probability in the lemma is strictly positive as we will show, but is in general less than one.

First, we define the following events.

$$\begin{aligned}
\mathcal{E}_1^- &= \{\mathcal{T}_i \cap T_\ell(t) \neq \emptyset \text{ for all } i \in \mathcal{P} \text{ and } \ell = 0, 1, 2\} \\
\mathcal{E}_2^- &= \{a_i(t') = a_i^\star \text{ for all } i \in \mathcal{P} \text{ and } t' = T^i_{k_i^*(t)+1}, \ldots, t + 3\Delta_\eta\} \\
\mathcal{E}_3^- &= \{\tau_i^i(t) + V^{i,j}_{k_i^*(t)} \leq t + \Delta_\eta \text{ for all } i, j \in \mathcal{P}, i \neq j\} \\
\mathcal{E}_4^- &= \{s_i(T^i_{k_i^*(t+2\Delta_\eta)+1}) = C \text{ for all } i \in E(\mathbf{s}(t + \Delta_\eta))\}
\end{aligned}$$

Note that events $\mathcal{E}_1^-$ through $\mathcal{E}_4^-$ mean that all agents continue to play the action profile $\mathbf{a}^\star$ between time t and $t + 3\Delta_\eta$ (and afterwards); all systems see $\mathbf{a}^\star$ after time $t + \Delta_\eta$ (event $\mathcal{E}_3^-$) and all agents update during the interval $T_1(t)$ (event $\mathcal{E}_1^-$). Thus, when the agents update during the interval $T_2(t)$, conditional on event $\mathcal{E}_2^-$, all agents will see the payoff in response to $\mathbf{a}^\star$. Finally, the agents' states will have changed to C by time $t + 3\Delta_\eta$ (event $\mathcal{E}_4^-$) and, as a result, they will keep playing $\mathbf{a}^\star$ after time $t + 3\Delta_\eta$.

Following essentially the same argument, we have

$$\begin{aligned}&\mathbf{P}\left[\mathbf{z}(t')=\mathbf{z}^\star \text{ for all } t'\geq t \;\middle|\; \mathbf{z}(t)=\mathbf{z}\right]\\&\geq \mathbf{P}\left[\mathbf{z}(t')=\mathbf{z}^\star \text{ for all } t'\geq t \;\middle|\; \mathbf{z}(t)=\mathbf{z}, \mathcal{E}_1^-, \mathcal{E}_2^-, \mathcal{E}_3^-, \mathcal{E}_4^-\right]\cdot \mathbf{P}\left[\mathcal{E}_1^-, \mathcal{E}_2^-, \mathcal{E}_3^-, \mathcal{E}_4^-, \mid \mathbf{z}(t)=\mathbf{z}\right], \qquad (11)\end{aligned}$$

where the first conditional probability in (11) is one. The second conditional probability can be lower bounded by $\eta^{3n}\cdot\delta^{3n\Delta_\eta}(1-p)^n$.

References

[1] G. Arslan, J.R. Marden, and J. S. Shamma, "Autonomous vehicle-target assignment: A game-theoretical formulation," *Journal of Dynamic Systems, Measurement, and Control*, 129(5):584-596, Apr. 2007.

[2] N.C. Bianchi, and G. Lugosi, "Prediction, learning, and games," Cambridge University Press, 2006.

[3] S.H. Chun and R.J. La, "Secondary spectrum trading - auction-based framework for spectrum allocation and profit sharing," *IEEE/ACM Trans. of Networking*, 21(1):176-189, Feb. 2013.

[4] L. Fang, and P.J. Antsaklis, "Information consensus of asynchronous discrete-time multi-agent systems," Proceedings of American Control Conference, 2005.

[5] D. Fudenberg and D.K. Levine, *The Theory of Learning in Games*, The MIT Press, 1998.

[6] F. Germano and G. Lugosi, "Global Nash convergence of Foster and Young's regret testing," *Games and Economic Behavior*, 60(1):135-154, Jul. 2007.

[7] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica* 68(5):1127-1150, Sep. 2000.

[8] S. Hart and A. Mas-Colell, "A reinforcement procedure leading to correlated equilibrium," *Economics Essays*, ed. by G. Debreu, W. Neuefeind and W. Trockel, pp.181-200, Springer Berlin Heidelberg, 2001.

[9] R.J. La and V. Anantharam, "Optimal routing control: repeated game approach," *IEEE Trans. on Automatic Control*, 47(3):437-450, Mar. 2002.

[10] J.R. Marden, G. Arslan and J.S. Shamma, "Connections between cooperative control and potential games illustrated on the consensus problem," Proceedings of European Control Conference (ECC), 2007.

[11] J.R. Marden, G. Arslan and J.S. Shamma, "Regret based dynamics: convergence in weakly acyclic games," Proceedings of the 6th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2007.

[12] J.R. Marden, H.P. Young, G. Arslan and J.S. Shamma, "Payoff-based dynamics for multiplayer weakly acyclic games," SIAM Journal on Control and Optimization, 48(1):373-396, 2009.

[13] J.R. Marden, G. Arslan and J.S. Shamma, "Joint strategy fictitious play with inertia for potential games," *IEEE Trans. on Automatic Control*, 54(2):208-220, Feb. 2009.

[14] J.R. Marden and J.S. Shamma, "Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation," Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing, Monticello (IL), Oct. 2010.

[15] J.R. Marden, H.P. Young and L.Y. Pao, "Achieving pareto optimality through distributed learning," Proceedings of the 51st IEEE Conference on Decision and Control, Maui (HI), Dec. 2014.

[16] J. Miękisz, M. Matuszak, and J. Poleszczuk, "Stochastic stability in three-player games with time delays," *Dynamic Games and Applications* 4.4 (2014): 489-498.

[17] S. Pal and R.J. La, "A simple learning rule in games and its convergence to pure-strategy Nash equilibria," Proceedings of American Control Conference, Chicago (IL), Jul. 2015.

[18] S. Pal and R.J. La "A simple learning rule for resilient Nash equilibria," preprint available at http://www.ece.umd.edu/∼hyongla.

[19] B.S.R. Pradelski and H.P. Young, "Learning efficient Nash equilibria in distributed systems," *Games and Economic Behavior*, 75:882-897, 2012.

[20] R.W. Rosenthal, "A class of games possessing pure-strategy Nash equilibria," *International Journal of Game Theory*, 2(1):65-67, 1973.

[21] J.N. Tsitsiklis, D.P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE transactions on automatic control* 31.9 (1986): 803-812.

[22] F. Xiao, and L. Wang, "State consensus for multi-agent systems with switching topologies and time-varying delays," *International Journal of Control* 79.10 (2006): 1277-1284.

[23] T. Yi, and W. Zuwang, "Effect of time delay and evolutionarily stable strategy," *Journal of theoretical biology* 187.1 (1997): 111-116.

[24] H.P. Young, "The evolution of conventions," *Econometrica: Journal of the Econometric Society*, 61(1):57-84, Jan. 1993.

[25] H.P. Young, "Learning by trial and error," *Games and economic behavior* 65.2 (2009): 626-643.