extended abstract

Selecting Efficient Coarse Correlated Equilibria Through Distributed Learning*

Jason R. Marden
University of Colorado
jason.marden@colorado.edu

April 26, 2013

Abstract

A learning rule is completely uncoupled if each players behavior is conditioned only on his own realized payoffs, and does not need to know the actions or payoffs of anyone else. We demonstrate a simple, completely uncoupled learning rule such that, in any finite normal form game with generic payoffs, the players realized strategies implements a Pareto optimal coarse correlated (Hannan) equilibrium a very high proportion of the time. A variant of the rule implements correlated equilibrium a very high proportion of the time.

1 Introduction

This paper builds on a recent literature that seeks to identify learning rules that lead to equilibrium without the usual assumptions of perfect rationality and common knowledge. Of particular interest are learning rules that are simple to implement and require a minimum degree of information about what others in the population are doing. Such rules can be viewed as models of behavior in games with many dispersed agents and very limited observability. They also have practical application to the design of distributed control systems, where the agents can be designed to respond to their environment in ways that lead to desirable system-wide outcomes.

One can distinguish between various classes of learning rules depending on the amount of information they require. A rule is uncoupled if it does not require any knowledge of the payoffs of the other players [HMC03]. A rule is completely uncoupled if it does not require any knowledge of the actions or payoffs of the other players [FY06]. The latter paper identifies a family of completely uncoupled learning rules that come close to Nash equilibrium (pure or mixed) with high probability in two-person normal form games with generic payoffs. Subsequently, [GL07] showed that similar results hold for n-person normal form games with generic payoffs. Lastly, [MYAS09] exhibited a much simpler class of completely uncoupled rules that lead to Nash equilibrium in weakly acyclic games. These learning algorithms all have the feature that agents occasionally experiment with new strategies, which they adopt if they lead to higher realized payoffs.

In [You09], this approach was further developed by making an agent's search behavior dependent on his mood (an internal state variable). Changes in mood are triggered by changes in realized payoffs relative to the agents current aspiration level. Rules of this nature can be designed that select pure Nash equilibria in any normal form game with generic payoffs that has at least one pure Nash equilibrium. Moreover the rule can be designed so that it selects a Pareto optimal pure Nash equilibrium [PY10] or even a Pareto optimal action profile (irrespective of whether this action profile is a pure Nash equilibrium) [MYP11].

There is a quite different class of learning dynamics that leads to coarse correlated equilibrium (alternatively correlated equilibrium). These rules are based on the concept of no regret. They can be formulated so that they depend only on a players own realized payoffs, that is, they are completely uncoupled [FV97, HMC00, FL98]. However, while the resulting dynamics converge almost surely to the set of correlated equilibria, they do not necessarily converge to – or even approximate – correlated equilibrium behavior at a given point in time.

The contribution of this paper is to demonstrate a class of completely uncoupled learning rules that bridges these two approaches. In overall structure the rules are similar to the learning dynamics introduced in [You09, PY10, MYP11]. Like the no-regret rules our approach selects (coarse) correlated equilibria instead of Nash equilibria. Unlike no-regret learning, however, our rule 'leads to' equilibrium in the sense that players strategies actually constitute a coarse correlated equilibrium a high proportion of the time. In fact, as a bonus, they constitute a Pareto optimal coarse correlated equilibrium a high proportion of the time.

*This research was supported by AFOSR grant #FA9550-09-1-0538 and by ONR grant #N00014-09-1-0751.

2 Preliminaries

Let G be a finite strategic-form game with n agents. The set of agents is denoted by $N := \{1, ..., n\}$. Each agent $i \in N$ has a finite action set $\mathcal{A}_i$ and a utility function $U_i : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ denotes the joint action set. We shall henceforth refer to a finite strategic-form game simply as "a game." For any joint distribution $q = \{q_a\}_{a \in \mathcal{A}} \in \Delta(\mathcal{A})$ where $\Delta(\mathcal{A})$ denotes the simplex over the joint action set $\mathcal{A}$, we extend the definition of an agent's utility function in the usual fashion

$$U_i(q) = \sum_{a \in \mathcal{A}} U_i(a) q_a.$$

The set of coarse correlated equilibria is characterized by the set of joint distributions

$$\text{CCE} = \left\{ q \in \Delta(\mathcal{A}) : \sum_{a \in \mathcal{A}} U_i(a) q_a \geq \sum_{a \in \mathcal{A}} U_i(a'_i, a_{-i}) q_a, \forall i \in N, a'_i \in \mathcal{A}_i \right\}$$

which is by definition non-empty.

In this paper we focus on the derivation of learning rules that provide convergence to an efficient coarse correlated equilibria of the form

$$q^* \in \arg\max_{q \in CCE} \sum_{i \in N} U_i(q).$$

To that end, we consider the framework of repeated one-shot game where a given game G is repeated once each period $t \in \{0, 1, 2, \dots\}$. In period t, the agents simultaneously choose actions $a(t) = (a_1(t), ..., a_n(t))$ and receive payoffs $U_i(a(t))$. Agent $i \in N$ chooses the action $a_i(t)$ according to a probability distribution $p_i(t) \in \Delta(\mathcal{A}_i)$, which is the simplex of probability distributions over $\mathcal{A}_i$. We shall refer to $p_i(t)$ as the *strategy* of agent i at time t. We adopt the convention that $p_i^{a_i}(t)$ is the probability that agent i selects action a_i at time t according to the strategy $p_i(t)$. An agent's strategy at time t relies only on observations from times $\{0, 1, 2, ..., t-1\}$.

Different learning algorithms are specified by the agents' available information and the mechanism by which their strategies are updated as information is gathered. Here, we focus on one of the most informationally restrictive class of learning rules, termed *completely uncoupled* or *payoff-based*, where agents *only* have access to: (i) the action they played and (ii) the payoff they received. More formally, the strategy adjustment mechanism of agent i takes the form

$$p_i(t) = F_i \left(\{a_i(\tau), U_i(a(\tau))\}_{\tau = 0, \dots, t-1} \right). \tag{1}$$

Recent work has shown that for finite games with generic payoffs, there exist completely uncoupled learning rules that lead to Pareto optimal Nash equilibria [PY10] and also Pareto optimal action profile irrespective of whether or not they are a pure Nash equilibrium [MYP11]; see also [AB11, You09, FM86]. Here, we exhibit a different class of learning procedures that lead to efficient coarse correlated equilibria.

3 Algorithm Description

We will now introduce a payoff based learning algorithm which ensures that the empirical frequency of the joint actions converges in probability to the coarse correlated equilibrium which maximizes the sum of the players' average payoffs. To achieve this objective, each agent will select sequences of actions of at most length w as opposed to single actions. The setup is as follows:

- w: This represents the maximum length of a sequence of actions that any agent will play.
- $\vec{\mathcal{A}}_i = \cup_{k=1,\dots,w} \mathcal{A}_i^k$: Set of possible vectored actions (or sequence of actions) that player i can select.
- $\vec{a}_i \in \vec{\mathcal{A}}_i$: A chosen sequence of actions by player i. For example, if at time t agent i decides to play a sequence of actions $\vec{a}_i$ of length $l_i = |\vec{a}_i| \leq w$, then

$$\begin{aligned} a_i(t) &= \vec{a}_i(1) \\ a_i(t+1) &= \vec{a}_i(2) \\ \vdots &= \vdots \\ a_i(t+l_i-1) &= \vec{a}_i(l_i) \end{aligned}$$

The following algorithms follows the theme of [You09] where an agents search behavior is dependent on his mood (an internal state variable). Changes in mood are triggered by changes in realized payoffs relative to the agents current aspiration level. An informal description of the forthcoming algorithm is as follows:

- The main internal states of the players are Content (C) and Discontent (D).
- The feasible space for player $i \in N$ s the set $\vec{\mathcal{A}}_i$ of all action-sequences of length w or less, where w is a large positive integer, the same for all players.
- At any given time the (temporary) strategy of a player is to play a given sequence of actions $\vec{a}_i$ of some length $l_i \leq w$, and to repeat this sequence for L_i consecutive periods. This is called the players benchmark strategy. His benchmark utility is the utility he expects to get from playing a strategy. If his benchmark utility equals the utility he is getting from his current benchmark strategy we say the benchmarks are aligned.
- A content player occasionally experiments with a different strategy in which he plays a constant action for L_i consecutive periods. If the average payoff from this experiment leads to a higher payoff than his current benchmark (aspiration level), he switches to the new strategy, otherwise he reverts to his previous strategy. (The justification for a constant strategy is that he cannot be sure a new action is better without trying it out for a number of periods in succession.)
- A discontent player i chooses strategies at random from $\vec{\mathcal{A}}_i$. He switches spontaneously from D to C with a probability that is an increasing function of his current average payoff (in which case he takes his current strategy and its realized payoff as his new benchmarks).
- A player switches from C to D for sure if his payoff goes down for several periods in a row and he was not experimenting. He may also switch spontaneously from C to D with a very small probability.
- The relationship between these probabilities will be spelled out later. We also omit mention of certain states that are intermediate between C and D. Their role will become clear when we give the learning rule in detail.

3.1 Notation

At each point in time, the *action* of agent $i \in N$ can be represented by the tuple $[\vec{a}_i, a_i]$, where

- Agent i's ***sequence of actions*** is $\vec{a}_i \in \vec{\mathcal{A}}_i$.
- Agent i's ***current action*** is $a_i \in \vec{a}_i$.

At each point in time an agent's *state* can be represented by the tuple

$$x_i = \begin{cases} \vec{a}_i & : \text{ Trial sequence of actions} \\ u_i & : \text{ Payoff over trial sequence of actions} \\ k_i & : \text{ Element of trial sequence of actions currently on} \\ \vec{a}_i^b & : \text{ Baseline trial sequence of actions} \\ u_i^b & : \text{ Payoff over baseline trial sequence of actions} \\ m_i & : \text{ Mood (Content, Discontent, Hopeful, or Watchful)} \\ c_i^{H/W} & : \text{ Counter for number of times Hopeful/Watchful periods repeated} \\ L_i^{H/W} & : \text{ Number of times Hopeful/Watchful periods will be repeated} \end{cases}$$

- The first three components of the state $\{\vec{a}_i, u_i, k_i\}$ correspond to the sequence of actions that are currently being played by agent i. The sequence of actions is represented by $\vec{a}_i \in \vec{\mathcal{A}}_i$. The counter $k_i \in \{1, \ldots, |\vec{a}_i|\}$ keeps track of what component of $\vec{a}_i$ the agent should play next. Lastly, the payoff u_i represent the average payoff received over the first $(k_i - 1)$ iterations of the action sequence $\vec{a}_i$.
- The fourth and fifth components of the state $\{\vec{a}_i^b, u_i^b\}$ correspond to the baseline sequence of actions and baseline payoff. The benchmark sequence of actions is represented by $\vec{a}_i^b \in \vec{\mathcal{A}}_i$ and the benchmark payoff u_i^b captures the average payoff received for the baseline sequence of actions. The benchmark payoff is used as a gauge to determine whether experimentations with alternative sequence of actions is advantageous.
- The sixth component of the state is the mood m_i, which can take on four values: *content* (C), *discontent* (D), *hopeful* (H), and *watchful* (W). Each of the moods will lead to different types of behavior from the player as will be discussed in detail.

- The seventh and eighth components of the state $\{c_i^{H/W}, L_i^{H/W}\}$ represent counters on the number of times that either a Hopeful or Watchful mood has been repeated. The number $L_i^{H/W} \in \{0\} \cup \{w+1, \ldots, w^n + w\}$ prescribes the number of times that the intermediate state (hopeful or watchful) should be repeated. The number $c_i^{H/W} \in \{0, 1, 2, \ldots, w^n + w\}$ prescribes the number of times that the intermediate state (hopeful or watchful) has already been repeated. Accordingly, $c_i^{H/W} \leq L_i^{H/W}$. In the case when the mood is not hopeful or watchful, we adopt the convention that $c_i^{H/W} = L_i^{H/W} = 0$.

3.2 Formal Algorithm Description

We divide the dynamics into the following two parts: the agent dynamics and the state dynamics. Without loss of generality we shall focus on the case where agent utility functions are strictly bounded between 0 and 1, i.e., for any agent $i \in N$ and action profile $a \in \mathcal{A}$ we have $1 > U_i(a) \geq 0$. Lastly, we define a constant $c > n$ which will be utilized in the following algorithm.

Agent Dynamics: Fix an experimentation rate $\epsilon > 0$. The dynamics for agent i only rely on the state of agent i at that given time. Let $x_i(t) = [\vec{a}_i, u_i, k_i, \vec{a}_i^b, u_i^b, m_i, c_i^{H/W}, L_i^{H/W}]$ be the state of agent i at time t. For the following dynamics, each agent only has the opportunity to change strategies at the beginning of a planning window. Accordingly, if $k_i > 1$ then

$$\vec{a}_i(t) = \vec{a}_i \tag{2}$$
$$a_i(t) = \vec{a}_i(k_i) \tag{3}$$

where $\vec{a}_i(k_i)$ denotes the k_i-th component of the vector $\vec{a}_i$. If $k_i = 1$, then a player makes a decision based on the player's underlying mood:

- **Content** $(m_i = C)$: In this state, the agent chooses a sequence of actions $\vec{a}_i \in \vec{\mathcal{A}}_i$ according to the following probability distribution

$$\mathbf{Pr}\left[\vec{a}_i(t) = \vec{a}_i\right] = \begin{cases} 1 - \epsilon^c & \text{for } \vec{a}_i = \vec{a}_i^b \\ \frac{\epsilon^c}{|\mathcal{A}_i|} & \text{for any } \vec{a}_i = (a_i, \ldots, a_i) \in \mathcal{A}_i^{|\vec{a}_i^b|} \text{ where } a_i \in \mathcal{A}_i \end{cases} \tag{4}$$

  where $|\mathcal{A}_i|$ represents the cardinality of the set $\mathcal{A}_i$. The action is then chosen as

$$a_i(t) = \vec{a}_i(1; t)$$

  where $\vec{a}_i(1; t)$ denotes the first component of the vector $\vec{a}_i(t)$.[1]

- **Discontent** $(m_i = D)$: In this state, the agent chooses a sequence of actions $\vec{a}_i$ according to the following probability distribution:

$$\mathbf{Pr}\left[\vec{a}_i(t) = \vec{a}_i\right] = \frac{1}{|\vec{\mathcal{A}}_i|} \text{ for every } \vec{a}_i \in \vec{\mathcal{A}}_i \tag{5}$$

  Note that the benchmark action and utility play no role in the agent dynamics when the agent is discontent. The action is then chosen as

$$a_i(t) = \vec{a}_i(1; t).$$

- **Hopeful** $(m_i = H)$ or **Watchul** $(m_i = W)$: In either of these states, the agent selects his trial action, i.e.,

$$\vec{a}_i(t) = \vec{a}_i \tag{6}$$
$$a_i(t) = \vec{a}_i(1; t) \tag{7}$$

State Dynamics: First, the majority of the state components only change at the end of a sequence of actions. Let $x_i(t) = [\vec{a}_i, u_i, k_i, \vec{a}_i^b, u_i^b, m_i, c_i^{H/W}, L_i^{H/W}]$ be the state of agent i at time t, $a_i(t) = \vec{a}_i(k_i)$ be the action that agent i played at time t, and $U_i(a(t))$ be the utility player i received at time t. If $k_i < |\vec{a}_i|$, then

$$x_i(t) = \begin{Bmatrix} \vec{a}_i \\ u_i \\ k_i \\ \vec{a}_i^b \\ u_i^b \\ m_i \\ c_i^{H/W} \\ L_i^{H/W} \end{Bmatrix} \longrightarrow x_i(t+1) = \begin{Bmatrix} \vec{a}_i \\ \frac{k_i - 1}{k_i} u_i + \frac{1}{k_i} U_i(a(t)) \\ k_i + 1 \\ \vec{a}_i^b \\ u_i^b \\ m_i \\ c_i^{H/W} \\ L_i^{H/W} \end{Bmatrix}$$

[1] We could consider variations of deviations to stabilize alternative equilibria, e.g., correlated equilibria.

Otherwise, if $k_i = |\vec{a}_i|$ then the state is updated according to the underlying mood as follows: For shorthand notation, we define the running average of the payoff over the trial actions as

$$u_i(t) = \frac{k_i - 1}{k_i} u_i + \frac{1}{k_i} U_i(a(t)).$$

- **Content** ($m_i = C$): If $[\vec{a}_i, u_i(t)] = [\vec{a}_i^b, u_i^b]$, the state of agent i is updated as

$$x_i(t+1) \quad = \quad \begin{cases} \left[\vec{a}_i^b, u_i^b, 1, \vec{a}_i^b, u_i^b, C, 0, 0\right] & \text{with probability } 1 - \epsilon^{2c} \\ \left[\vec{a}_i^b, u_i^b, 1, \vec{a}_i^b, u_i^b, D, 0, 0\right] & \text{with probability } \epsilon^{2c} \end{cases} \tag{8}$$

  If $\vec{a}_i \neq \vec{a}_i^b$, the state of agent i is updated as

$$x_i(t+1) \quad = \quad \begin{cases} [\vec{a}_i, u_i(t), 1, \vec{a}_i, u_i(t), C, 0, 0] & \text{if } u_i(t) > u_i^b \\ \left[\vec{a}_i^b, u_i^b, 1, \vec{a}_i^b, u_i^b, C, 0, 0\right] & \text{if } u_i(t) \leq u_i^b \end{cases}$$

  If $\vec{a}_i(t) = \vec{a}_i^b$ but $u_i(t) \neq u_i^b$, the state of agent i is updated as

$$x_i(t+1) \quad = \quad \begin{cases} \left[\vec{a}_i^b, u_i^b, 1, \vec{a}_i^b, u_i^b, H, 1, L_i^H\right] & \text{if } u_i(t) > u_i^b \\ \left[\vec{a}_i^b, u_i^b, 1, \vec{a}_i^b, u_i^b, W, 1, L_i^W\right] & \text{if } u_i(t) < u_i^b \end{cases}$$

  where L_i^H (or L_i^W) is randomly selected from the set $\{w+1, \ldots, w^n + w\}$ with uniform probability.

- **Discontent** ($m_i = D$): The new state is determined by the transition

$$x_i(t+1) \quad = \quad \begin{cases} [\vec{a}_i, u_i(t), 1, \vec{a}_i, u_i(t), C, 0, 0] & \text{with probability } \epsilon^{1-u_i(t)} \\ [\vec{a}_i, u_i(t), 1, \vec{a}_i, u_i(t), D, 0, 0] & \text{with probability } 1 - \epsilon^{1-u_i(t)}. \end{cases}$$

- **Hopeful** ($m_i = H$): First, it is important to highlight that if the mood of any player $i \in N$ is Hopeful then $\vec{a}_i = \vec{a}_i^b$. The new state is determined as follows: If $c_i^H < L_i^H$, then

$$x_i(t+1) = \left[\vec{a}_i, u_i(t), 1, \vec{a}_i, u_i^b, H, c_i^H + 1, L_i^H\right].$$

  If $c_i^H = L_i^H$ and $u_i(t) \geq u_i^b$, then

$$x_i(t+1) \quad = \quad [\vec{a}_i, u_i(t), 1, \vec{a}_i, u_i(t), C, 0, 0]$$

  If $c_i^H = L_i^H$ and $u_i(t) < u_i^b$, then

$$x_i(t+1) = \left[\vec{a}_i, u_i(t), 1, \vec{a}_i, u_i^b, W, 1, L_i^W\right].$$

  where L_i^W is randomly selected from the set $\{w+1, \ldots, w^n + w\}$ with uniform probability.

- **Watchful** ($m_i = W$): First, it is important to highlight that if the mood of any player $i \in N$ is Watchful then $\vec{a}_i = \vec{a}_i^b$. The new state is determined as follows: If $c_i^W < L_i^W$, then

$$x_i(t+1) = \left[\vec{a}_i^b, u_i(t), 1, \vec{a}_i^b, u_i^b, W, c_i^W + 1, L_i^W\right].$$

  If $c_i^W = L_i^W$ and $u_i(t) < u_i^b$, then

$$x_i^s(t+1) \quad = \quad \left[\vec{a}_i^b, u_i(t), 1, \vec{a}_i, u_i(t), D, 0, 0\right]$$

  If $c_i^W = L_i^W$ and $u_i(t) \geq u_i^b$, then

$$x_i(t+1) = \left[\vec{a}_i, u_i(t), 1, \vec{a}_i^b, u_i^b, H, 1, n_i^H\right].$$

  where n_i^H is randomly selected from the set $\{w+1, \ldots, w^n + w\}$ with uniform probability.

4 Main Result

Before stating the main result we introduce a bit of notation. Let X denote the full set of states of the players. For a given state $x = (x_1, \dots, x_n)$ where $x_i = \left[\vec{a}_i, u_i, k_i, \vec{a}_i^b, u_i^b, m_i, c_i^{H/W}, L_i^{H/W}\right]$, define the ensuing sequence of baseline actions as follows: for every $k \in \{0, 1, 2, \dots\}$ and agent $i \in N$ we have

$$a_i(k|x_i) = \vec{a}_i^b(k + k_i)$$

where we write $\vec{a}_i^b(k + k_i)$ even in the case when $k + k_i > |\vec{a}_i^b|$ with the understanding that this implies the component $((k + k_i - 1) \mod |\vec{a}_i^b|) + 1$. We express the sequence of joint action profiles by $a(k|x) = (a_1(k|x_1), \dots, a_n(k|x_n))$. Define the average payoff over the forthcoming periods (provided that all players play according to their baseline action) for any player $i \in N$ and period $l \in \{1, 2, \dots\}$ as

$$u_i(0|x) = \frac{k_i - 1}{|\vec{a}_i|} u_i + \frac{|\vec{a}_i| - k_i + 1}{|\vec{a}_i|} \sum_{k=0}^{|\vec{a}_i| - k_i} U_i(a(k|x')), \tag{9}$$

$$u_i(l|x) = \frac{1}{|\vec{a}_i|} \sum_{k=l\cdot|\vec{a}_i| - k_i + 1}^{l\cdot|\vec{a}_i| - k_i} U_i(a(k|x')). \tag{10}$$

We will characterize the above dynamics by analyzing the empirical distribution of the joint distribution. To that end, define the empirical distribution of the joint actions associated with the baseline sequence of actions for a given state x by $q(x) = \{q_a(x)\}_{a \in \mathcal{A}} \in \Delta(\mathcal{A})$ where

$$q_a(x) = \lim_{t \to \infty} \frac{\sum_{\tau=0}^{t} I\{a = a(\tau|x)\}}{t + 1}, \tag{11}$$

$$= \frac{\sum_{\tau=1}^{\prod_{i \in N} |\vec{a}_i|} I\{a = a(\tau|x)\}}{\prod_{i \in N} |\vec{a}_i|}, \tag{12}$$

where $I\{\cdot\}$ represents the usual indicator function and the equality derives from the fact that players are repeating finite sequence of actions which ensures that for any $k \in \{0, 1, \dots\}$ we have

$$a(k|x) = a(k + \prod_{i \in N} |\vec{a}_i|).$$

Define the set of states which induce coarse correlated equilibria through repeated play of the baseline sequence of actions as

$$X^{\mathrm{CCE}} := \{x \in X : q(x) \in \mathrm{CCE}\}$$

Lastly, define the set of states X^* which induce coarse correlated equilibria and are consistent, i.e.,

$$X^* = \{x \in X : x \in X^{\mathrm{CCE}}, u_i(0|x) = u_i(k|x)\ \forall i \in N, k \in \{1, 2, \dots\}\}.$$

Note that in general the set X^* need not be empty. In fact, a sufficient condition for X^* to not be empty is

$$\left\{q \in \Delta(a) : q_a \in \cup_{k=1,\dots w} \left\{0, \frac{1}{k}, \dots, \frac{k-1}{k}, 1\right\} \text{ for all } a \in \mathcal{A}\right\} \cap X^{\mathrm{CCE}} \neq \emptyset.$$

The process described above can be characterized as a finite Markov chain parameterized by an exploration rate $\epsilon > 0$. The following theorem characterizing the *support* of the limiting stationary distribution, whose elements are referred to as the *stochastically stable states* [FY90]. More precisely, a state $x \in X$ is stochastically stable if and only if $\lim_{\epsilon \to 0^+} \mu(x, \epsilon) > 0$ where $\mu(x, \epsilon)$ is a stationary distribution of the process P^ϵ for a fixed $\epsilon > 0$. Our characterization requires a mild degree of genericity in the agents' payoff function which is summarized by the following notion of interdependence as introduced in [You09].

Definition 1 (Interdependence)**.** *An n-person game G on the finite action space $\mathcal{A}$ is interdependent if, for every $a \in \mathcal{A}$ and every proper subset of agents $J \subset N$, there exists an agent $i \notin J$ and a choice of actions $a'_J \in \prod_{j \in J} \mathcal{A}_j$ such that $U_i(a'_J, a_{-J}) \neq U_i(a_J, a_{-J})$.*

Theorem 1. *Let G be an finite interdependent game and suppose all players follow the above dynamics. If $X^* \neq \emptyset$, then a state $x \in X$ is stochastically stable if and only if $x \in X^*$ and*

$$\sum_{i \in N} u_i(0|x) = \max_{x' \in X^*} \sum_{i \in N} u_i(0|x').$$

If $X^* = \emptyset$, then a state $x \in X$ is stochastically stable if and only if

$$\sum_{i \in N} u_i(0|x) = \max_{a \in \mathcal{A}} \sum_{i \in N} U_i(a).$$

This theorem demonstrates that as the exploration rates $\epsilon \to 0^+$, the process will spend most of the time at the efficient coarse correlated equilibrium provide that the (discretized) set of coarse correlated equilibria in nonempty. If this set is empty, then the process will spend most of the time at the action profile which maximizes the sum of the agent's payoffs. We prove this theorem using the theory of resistance tree for regular perturbed processes developed in [You93]. We omit the details for brevity.

5 Conclusion

The results in this paper demonstrate that specific forms of correlated behavior can be attained through distributed learning rules with no explicit communication between the agents. While the players long run behavior is shown to be consistent with the most efficient coarse correlated equilibria, it is important to highlight that this results does not imply that the players' joint strategy at any given time constitutes an efficient coarse correlated equilibrium. Future work will focus on bridging this gap.

References

[AB11] I. Arieli and Y. Babichenko. Average testing and the efficient boundary. Discussion paper, Department of Economics, University of Oxford and Hebrew University, 2011.

[FL98] D. Fudenberg and D. Levine. *The Theory of Learning in Games*. MIT Press, Cambridge, MA, 1998.

[FM86] D. Fudenberg and E. Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, **54**:533–554, 1986.

[FV97] D.P. Foster and R. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, **21**:40–55, 1997.

[FY90] D.P. Foster and H.P. Young. Stochastic evolutionary games dynamics. *Journal of Theoretical Population Biology*, **38**:219–232, 1990.

[FY06] D.P. Foster and H.P. Young. Regret testing: Learning to play Nash equilibrium without knowing you have an opponent. *Theoretical Economics*, **1**:341–367, 2006.

[GL07] F. Germano and G. Lugosi. Global Nash convergence of Foster and Young's regret testing. *Games and Economic Behavior*, **60**:135–154, July 2007.

[HMC00] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[HMC03] S. Hart and A. Mas-Colell. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review*, 93(5):1830–1836, 2003.

[MYAS09] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma. Payoff based dynamics for multi-player weakly acyclic games. *SIAM Journal on Control and Optimization*, 48:373–396, February 2009.

[MYP11] J. R. Marden, H. P. Young, and L. Y. Pao. Achieving pareto optimality through distributed learning. under submission, 2011.

[PY10] B. R. Pradelski and H. P. Young. Learning efficient Nash equilibria in distributed systems. Discussion paper, Department of Economics, University of Oxford, 2010.

[You93] H. P. Young. The evolution of conventions. *Econometrica*, 61(1):57–84, January 1993.

[You09] H. P. Young. Learning by trial and error. *Games and Economic Behavior*, 65:626–643, 2009.