# Learning Efficient Correlated Equilibria

Holly P. Borowski, Jason R. Marden, and Jeff S. Shamma

***Abstract*— The majority of distributed learning literature focuses on convergence to Nash equilibria. Correlated equilibria, on the other hand, can often characterize more efficient collective behavior than even the best Nash equilibrium. However, there are no existing distributed learning algorithms that converge to specific correlated equilibria. In this paper, we provide one such algorithm which guarantees that the agents' collective joint strategy will constitute an efficient correlated equilibrium with high probability. The key to attaining efficient correlated behavior through distributed learning involves incorporating a common random signal into the learning environment.**

## I. Introduction

Agents' control laws are a crucial component of any multiagent system. They dictate how individual agents process locally available information to make a decision. Factors that determine the quality of a learning algorithm include informational dependencies, asymptotic guarantees, and convergence rates. Hence, significant research has been directed at deriving distributed learning algorithms that perform well with regard to these performance metrics.

The majority of this research has focused on attaining convergence to (pure) Nash equilibria under stringent information conditions [4], [8]–[10], [21], [24]. Recently, the research focus has shifted to ensuring convergence to alternate types of equilibria that often yield more efficient behavior than Nash equilibria. In particular, results have emerged that guarantee convergence to Pareto efficient Nash equilibria [17], [22], potential function maximizers [3], [15], welfare maximizing action profiles [1], [18], and correlated equilibrium [2], [7], [11], [16], among others.

In most cases highlighted above, the derived algorithms guarantee (probabilistic) convergence to the specified equilibria. However, the class of correlated equilibria has posed significant challenges with regards to this goal. Learning algorithms that converge to an efficient correlated equilibrium are desirable because optimal system behavior can often be characterized by a correlated equilibrium. Unfortunately, the aforementioned learning algorithms, such as regret matching [11], merely converge to the *set* of correlated equilibria. This means that the long run behavior does not necessarily constitute – or even approximate – a specific correlated equilibrium at any instance of time.

This research was supported by AFOSR grant #FA9550-12-1-0359, ONR grant #N00014-09-1-0751, NSF grant #ECCS-1351866, and the NASA Aeronautics Scholarship Program.

H. P. Borowski is a graduate research assistant with the Department of Aerospace Engineering, University of Colorado, Boulder, holly.borowski@colorado.edu.

J. R. Marden is with the Department of Electrical, Computer, and Energy Engineering, University of Colorado, Boulder, jason.marden@colorado.edu.

J.S. Shamma is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, shamma@gatech.edu, and with King Abdullah University of Science and Technology (KAUST), jeff.shamma@kaust.edu.sa.

Here, we provide a simple distributed learning algorithm that converges to the most efficient, i.e., welfare maximizing, correlated equilibrium. For concreteness, consider a mild variant of the Shapley game with the following payoff matrix

| | L | M | R |
|---|---|---|---|
| T | 1,-$\varepsilon$ | -$\varepsilon$,1 | 0,0 |
| M | 0,0 | 1,-$\varepsilon$ | -$\varepsilon$,1 |
| B | -$\varepsilon$,1 | 0,0 | 1,-$\varepsilon$ |

where $\varepsilon > 0$ is a small constant. In this game, there are two players (Row, Column); the row player has three actions (T,M,B), and the column player has three actions (L,M,R). The numbers in the table above are the players' payoffs for each of the nine joint actions. The unique Nash equilibrium for this game occurs when each player uses a probabilistic strategy that selects each of the three actions with probability $1/3$. This yields an expected payoff of $\approx 1/3$ to each player. Alternatively, a joint distribution that places a mass of $1/6$ on each of the six joint actions that yield non-zero payoffs to the players yields an expected payoff of $\approx 1/2$ to each player. Note that this distribution cannot be realized by independent strategies associated with the two players, but instead represents a specific correlated equilibrium.

As the above example demonstrates, distributed learning algorithms that converge to efficient correlated equilibria can be desirable from a system-wide perspective. In line with this theme, a recent result in [16] proposed a distributed algorithm that guarantees that the empirical frequency of the agents' collective behavior will converge to an efficient correlated equilibrium; however, convergence in empirical frequencies is attained through deterministic cyclic behavior. Additional results presented in [14] rely upon looking for cyclic behavior against a bounded memory opponent.

Predictable, cyclic behavior may be desirable from a system-wide perspective for many applications, e.g., data ferrying [5]. However, such behavior could be exploited in many other situations, e.g., team versus team zero-sum games [12], [23]. By viewing each team as a single player, classical results for two-player zero-sum games suggest that a team's desired strategy is to play its security strategy, which can be characterized by a probability distribution over the team's joint action space. Distributed learning algorithms that can stabilize specific joint strategies, such as correlated equilibria, may be necessary for providing strong performance guarantees in such settings.

Here, we present a distributed learning algorithm that ensures agents play a joint strategy corresponding to the efficient correlated equilibrium. Attaining such guarantees on the underlying joint strategy is non-trivial as we aim to design learning rules where individual agents make independent decisions in response to local information. Incorporating a common random signal into agents' local decision making rules makes this correlation possible.

Recent research has also focused on efficient centralized algorithms for computing specific correlated equilibria [13], [19], [20]. Such algorithms often require a complete characterization of the game which is unavailable in many engineering multiagent systems. Hence, the applicability of such results to design and control of multiagent systems may be limited.

## II. Background

We consider the framework of finite strategic form games where there exists an agent set $N = \{1, 2, \dots, n\}$, and each agent $i \in N$ is associated with a finite action set $\mathcal{A}_i$ and a utility function $U_i : \mathcal{A} \to [0, 1]$ where $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$ denotes the joint action space. We represent such a game by the tuple $G = (N, \{U_i\}_{i\in N}, \{\mathcal{A}_i\}_{i\in N})$.

We focus on the class of coarse correlated equilibria [2]. A coarse correlated equilibrium is a joint distribution $q = \{q^a\}_{a\in\mathcal{A}} \in \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ represents the simplex over the finite set $\mathcal{A}$, such that for any $i \in N$ and $a'_i \in \mathcal{A}_i$,

$$\sum_{a\in\mathcal{A}} U_i(a_i, a_{-i})q^a \geq \sum_{a\in\mathcal{A}} U_i(a'_i, a_{-i})q^a, \tag{1}$$

where $a_{-i} = \{a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n\}$ denotes the actions of all players other than player $i$.[1] We say a coarse correlated equilibrium $q^*$ is *efficient* if it maximizes the sum of the expected payoffs of the agents, i.e.,

$$q^* \in \arg\max_{q\in \mathrm{CCE}} \sum_{i\in N}\sum_{a\in\mathcal{A}} U_i(a)q^a, \tag{2}$$

where $\mathrm{CCE} \subset \Delta(\mathcal{A})$ denotes the set of coarse correlated equilibria. It is well known that $\mathrm{CCE} \neq \emptyset$ for any game $G$.

We derive a distributed learning algorithm that ensures collective behavior of agents converges to an efficient coarse correlated equilibrium. We adopt the framework of repeated one-shot games, where a static game $G$ is repeated over time and agents make decisions based on observations of previous plays of the game. A repeated one-shot game yields a sequence of action profiles $a(0)$, $a(1)$, $\dots$, where at each time $t \in \{0, 1, 2, \dots\}$ the decision of each agent $i$ is chosen independently according to the agent's strategy at time $t$, denoted by $p_i(t) = \{p_i^{a_i}(t)\}_{a_i\in\mathcal{A}_i} \in \Delta(\mathcal{A}_i)$.

A learning rule dictates how each agent selects its strategy given available information from previous plays of the game. One type of learning rule, known as *completely uncoupled* or *payoff based* [8], takes on the form:

$$p_i(t) = F_i\left(\{a_i(\tau), U_i(a(\tau))\}_{\tau=0,\dots,t-1}\right) \tag{3}$$

Such learning rules are known as *completely uncoupled* [8] and represent one of the most informationally restrictive classes of learning rules since the only knowledge that each agent has about previous plays of the game is (i) the action the agent played and (ii) the utility the agent received.

We measure a learning rule's performance $\{F_i\}_{i\in N}$ by its asymptotic guarantees. Let $q(t) \in \Delta(\mathcal{A})$ represent the agents' collective strategy at time $t$, which is of the form

$$q^{(a_1,\dots,a_n)}(t) = p_1^{a_1}(t) \times \cdots \times p_n^{a_n}(t) \tag{4}$$

[1] We will express an action profile $a \in \mathcal{A}$ as $a = (a_i, a_{-i})$.

where $\{p_i(t)\}_{i\in N}$ are the individual agent strategies at time $t$. We derive learning rules that guarantee agents' collective strategy is an efficient coarse correlated equilibrium the majority of the time, i.e., for all sufficiently large times $t$,

$$\Pr\left[q(t) \in \arg\max_{q\in\mathrm{CCE}} \sum_{i\in N}\sum_{a\in\mathcal{A}} U_i(a)q^a\right] \approx 1. \tag{5}$$

Attaining this goal using learning rules of the form (3) is impossible as such rules do not allow for correlation between the players, i.e., the agents' collective strategies are restricted to being of form (4). Accordingly, we modify the learning rules in (3) by giving each agent access to a common random signal $z(t)$ at each period $t \in \{0, 1, \dots\}$ that is i.i.d. and drawn uniformly from the interval $[0, 1]$. Now, the considered distributed learning rule takes the form

$$p_i(t) = F_i\left(\{a_i(0), U_i(a(\tau), z(\tau))\}_{\tau=0,\dots,t-1}\right). \tag{6}$$

This common signal can be used as a coordinating entity to reach collective strategies beyond the form in (4).

## III. A learning algorithm for attaining efficient correlated equilibria

In this section, we present a learning rule of the form (6) that guarantees agents' collective strategy constitutes an efficient coarse correlated equilibrium the majority of the time. This algorithm achieves the desired convergence guarantees by using *signal-based strategies* to exploit the common random signal, $z(t)$.

### A. Preliminaries

Consider a situation where each agent $i \in N$ commits to a signal-based strategy of the form $s_i : [0, 1] \to \mathcal{A}_i$ which associates with each signal $z \in [0, 1]$ an action $s_i(z) \in \mathcal{A}_i$. With an abuse of notation, we consider a finite parameterization of such signal-based strategies, which we refer to as *strategies*, of the form $S_i = \cup_{\omega=1}^{\Omega}(\mathcal{A}_i)^\omega$ where $\Omega \geq 1$ is a design parameter identifying the granularization of the agent's possible strategies. A strategy $s_i = (a_i^1, \dots, a_i^\omega) \in S_i$, $\omega \leq \Omega$, defines a mapping of the form

$$s_i(z) = \begin{cases} a_i^1 & \text{if} \quad z \in [0, 1/\omega) \\ a_i^2 & \text{if} \quad z \in [1/\omega, 2/\omega) \\ \vdots & \quad\quad \vdots \\ a_i^\omega & \text{if} \quad z \in [(\omega-1)/\omega, 1]. \end{cases} \tag{7}$$

These strategies divide the unit interval into at most $\Omega$ regions of equal length and associate each region with a specific action in the agent's action set. If the agents commit to a strategy profile $s = (s_1, s_2, \dots, s_n) \in S = \prod_{i\in N} S_i$, the resulting joint strategy $q(s) = \{q^a(s)\}_{a\in\mathcal{A}} \in \Delta(\mathcal{A})$ satisfies

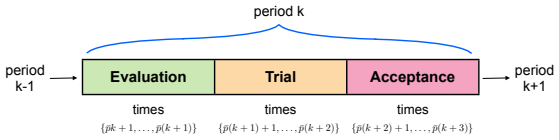
$$q^a(s) = \int_0^1 \prod_{i\in N} I\{s_i(z) = a_i\}dz$$

where $I\{\cdot\}$ is the indicator function. Lastly, the set of joint distributions that can be realized by the strategies $S$ is

$$q(S) = \{q \in \Delta(\mathcal{A}) : q(s) = q \text{ for some } s \in S\}.$$

### B. Algorithm description

The forthcoming algorithm is reminiscent of the trial and error learning algorithm introduced in [24] and can be viewed at a high level through the following diagram.




We begin by defining a constant $c > n$, an experimentation rate $\varepsilon \in (0,1)$, and the length of a phase to be $\bar{p} = \lceil 1/\varepsilon^{nc+1} \rceil$ time steps. A period consists of the evaluation, trial, and acceptance phases, and hence is $3\bar{p}$ time steps long. Let $x_i = x_i(k) = [s_i^b, m_i]$ represent that state of each agent $i \in N$ at the beginning of some period $k \in \{1, 2, \dots\}$.

**Agent Dynamics:** Here we describe how individual agents make decisions within a given period. Decisions of an agent $i \in N$ are influenced purely by its state at the beginning of the $k$-th period, $x_i(k)$, and by payoffs received during the $k$-th period. We specify agents' behavior during the $k$-th period for the three phases highlighted above.

**– *Evaluation Phase:*** The evaluation phase consists of the times $t \in \{\bar{p}3k+1, \dots, \bar{p}(3k+1)\}$. Throughout this phase, each agent commits to its baseline strategy $s_i^b$. At the end of the phase, each agent computes its average baseline utility,

$$u_i^b = \frac{1}{\bar{p}} \sum_{\tau=\bar{p}3k+1}^{\bar{p}(3k+1)} U_i\big(s_1^b(z(\tau), \dots, s_n^b(z(\tau))\big), \tag{8}$$

where $z(\tau)$ denotes the common random signal observed at time $\tau$. Here, $u_i^b$ is viewed as an assessment of the performance associated with the baseline strategy $s_i^b$.

**– *Trial Phase:*** After the evaluation phase comes the trial phase which consists of the times $t \in \{\bar{p}(3k+1)+1, \dots, \bar{p}(3k+2)\}$. During the trial phase each player $i \in N$ may try a strategy other than its baseline, and must commit to this trial strategy, $s_i^t \in S_i$, over the entire phase. Agents' trial strategies are selected according to the following rule:

- *Content,* $m_i = C$*:* When agent $i$ is content, its trial strategy, $s_i^t \in S_i$, is chosen according to the distribution

$$\Pr\left[s_i^t = s_i\right] = \begin{cases} 1-\varepsilon^c & \text{if } s_i = s_i^b \\ \varepsilon^c \,/\, |\mathcal{A}_i| & \text{for any } s_i = a_i \in \mathcal{A}_i \end{cases} \tag{9}$$

  Note that a content player predominantly selects its baseline strategy during the trial phase.
- *Discontent,* $m_i = D$*:* When agent $i$ is discontent, its trial strategy, $s_i^t$, is chosen randomly from the set $S_i$,

$$\Pr\left[s_i^t = s_i\right] = 1 \,/\, |S_i| \text{ for all } s_i \in S_i. \tag{10}$$

At the end of the trial phase, each agent computes its average utility:

$$u_i^t = \frac{1}{\bar{p}} \sum_{\tau=\bar{p}3(k+1)+1}^{\bar{p}(3k+2)} U_i\big(s_1^t(z(\tau), \dots, s_n^t(z(\tau))\big). \tag{11}$$

Here, $u_i^t$ is viewed as an assessment of the performance associated with the baseline strategy $s_i^t$.

**– *Acceptance Phase:*** The last phase is the acceptance phase which consists of times $t \in \{\bar{p}(3k+2)+1, \dots, \bar{p}(3k+3)\}$. The primary purpose of the acceptance phase is to further evaluate changes in the payoffs between $u_i^b$ and $u_i^t$. Each agent $i \in N$ commits to an acceptance strategy, denoted by $s_i^a \in S_i$, over the entire acceptance phase. Each agent's acceptance strategy is selected according to the following.

- *Content,* $m_i = C$*:* When agent $i$ is content, its acceptance strategy is chosen as follows:

$$s_i^a = \begin{cases} s_i^t & \text{if } u_i^t > u_i^b + \varepsilon, \\ s_i^b & \text{if } u_i^t \le u_i^b + \varepsilon. \end{cases} \tag{12}$$

  That is, players only repeat their trial strategy if their performance was high enough relative to the performance of the baseline strategy.
- *Discontent,* $m_i = D$*:* When agent $i$ is discontent, the acceptance strategy is set as $s_i^a = s_i^t$.

Following the acceptance phase, each agent computes its average utility:

$$u_i^a = \frac{1}{\bar{p}} \sum_{\tau=\bar{p}3(k+2)+1}^{\bar{p}(3k+3)} U_i\big(s_1^a(z(\tau), \dots, s_n^a(z(\tau))\big). \tag{13}$$

Here, $u_i^a$ is viewed as an assessment of the performance associated with the baseline strategy $s_i^a$.

**State Dynamics:** After the agent dynamics comes the state dynamics which specifies how the state of each agent evolves. The state of each agent $i \in N$ at the beginning of the $k+1$-st stage, i.e., $x_i(k+1)$, is influenced purely its state at the beginning of the $k$-th period, i.e., $x_i(k)$, strategies $s_i^b$, $s_i^t$ and $s_i^a$, and payoffs received during the $k$-th period. State dynamics are broken into the following cases:

*– Content and No Experimentation,* $m_i = C, s_i^t = s_i^b$*:* If agent $i$ was content at the start of the $k$-th period and did not experiment in the trial phase, its state at the beginning of the $(k+1)$-st period is chosen as follows:

$$x_i(k+1) = \begin{cases} \left[s_i^a = s_i^b,\, C\right] & \text{if } u_i^a \ge u_i^b - \varepsilon, \\ \left[s_i^a = s_i^b,\, D\right] & \text{if } u_i^a < u_i^b - \varepsilon. \end{cases} \tag{14}$$

Accordingly, if the agent's average payoff during the acceptance phase is low enough, then it will become discontent.

*– Content and Experimentation,* $m_i = C, s_i^t \neq s_i^b$*:* If agent $i$ was content at the start of the $k$-th period and experimented during the trial phase, its state at the beginning of the $(k+1)$-st period is chosen as

$$x_i(k+1) = [s_i^a, C]. \tag{15}$$

In this case the agent's average payoff during the acceptance phase does not impact its underlying state dynamics.

*– Discontent,* $m_i = D$*:* If agent $i$ was discontent at the start of the $k$-th period, its state at the beginning of the $(k+1)$-th period is chosen as follows

$$x_i(k+1) = \begin{cases} [s_i^a, C] & \text{w.p. } \varepsilon^{1-u_i^a}, \\ [s_i^a, D] & \text{w.p. } 1-\varepsilon^{1-u_i^a}. \end{cases} \tag{16}$$

Here, the agents are more likely to become content with strategies the yield higher average payoffs.

### C. Main Result

We focus on games where there is some degree of coupling between agents' utility functions. The following definition of interdependence, taken from [24], captures this notion.

*Definition 1:* A game $G$ with agents $N = \{1, 2, \dots, n\}$ is said to be *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 $U_i(a'_J, a_{-J}) \neq U_i(a_J, a_{-J})$.

The following theorem characterizes the limiting behavior associated with the proposed algorithm.

*Theorem 1:* Let $G = (N, \{U_i\}, \{\mathcal{A}_i\})$ be a finite interdependent game. First, suppose $q(S) \cap \mathrm{CCE} \neq \emptyset$. Given any probability $p < 1$, if the exploration rate $\varepsilon$ is sufficiently small, then for all sufficiently large times $t$,

$$\Pr\left[q(s(t)) \in \underset{q \in q(S) \cap \mathrm{CCE}}{\arg\max} \sum_{i \in N} \sum_{a \in \mathcal{A}} U_i(a) q^a\right] > p.$$

Alternatively, suppose $q(S) \cap \mathrm{CCE} = \emptyset$. Given any probability $p < 1$, if the exploration rate $\varepsilon$ is sufficiently small, then for all sufficiently large times $t$,

$$\Pr\left[q(s(t)) \in \underset{q \in q(S)}{\arg\max} \sum_{i \in N} \sum_{a \in \mathcal{A}} U_i(a) q^a\right] > p.$$

Observe that the proposed algorithm is of the form (6). Moreover, the condition $q(S) \cap \mathrm{CCE} \neq \emptyset$ implies the agents can realize specific joint distributions that are coarse correlated equilibria through the joint strategy set $S$. When this is the case, the above algorithm ensures the agents predominantly play a strategy $s \in S$ where the resulting joint distribution $q(s)$ corresponds to the efficient coarse correlated equilibrium. The condition $q(S) \cap \mathrm{CCE} = \emptyset$ implies there are no agent strategies that can characterize a coarse correlated equilibrium. When that is the case, the above algorithm ensures the agents predominantly play strategies that have full support on the action profiles $a \in \mathcal{A}$ that maximize the sum of the agents payoffs, i.e., $\arg\max_{a \in \mathcal{A}} \sum_{i \in N} U_i(a)$.

## IV. Proof of Theorem 1

The decision making process defined in Section III ensures the evolution of agents' states over periods $\{0, 1, 2, \dots\}$ can be represented as a finite ergodic Markov chain over state space $X = X_1 \times \cdots \times X_n$, where $X_i = S_i \times \{C, D\}$ denotes the set of possible states of agent $i$. Let $P^\varepsilon$ denote this Markov chain for some $\varepsilon > 0$. Proving Theorem 1 requires characterizing the stationary distribution of the family of Markov chains $\{P^\varepsilon\}_{\varepsilon > 0}$ for all sufficiently small $\varepsilon$. We use the theory of resistance trees for regular perturbed processes, introduced in [25], to accomplish this task. First we review this theory and then we prove Theorem 1.

### A. Background: Resistance Trees

Let $P^0$ be the transition matrix for some nominal Markov process, and let $P^\varepsilon$ be a perturbed version of this process where the size of the perturbation is $\varepsilon > 0$. We focus on the following class of Markov chains.

*Definition 2:* A family of Markov chains defined over a finite state space $X$, whose transition matrices are denoted by $\{P^\varepsilon\}_{\varepsilon > 0}$, is called a *regular perturbed process* of a nominal process $P^0$ if the following conditions are satisfied for all $x, x' \in X$:
(1) There exists a constant $c > 0$ such that $P^\varepsilon$ is aperiodic and irreducible for all $\varepsilon \in (0, c]$.
(2) $\lim_{\varepsilon \to 0} P^\varepsilon_{x \to x'} = P^0_{x \to x'}$.
(3) If $P^\varepsilon_{x \to x'} > 0$ for some $\varepsilon > 0$, then there exists a constant $r(x \to x') \geq 0$ such that

$$0 < \lim_{\varepsilon \to 0^+} \frac{P^\varepsilon_{x \to x'}}{\varepsilon^{r(x \to x')}} < \infty. \tag{17}$$

Constant $r(x \to x')$ is the *resistance* of transition $x \to x'$.

For any $\varepsilon > 0$, let $\mu^\varepsilon = \{\mu^\varepsilon_x\}_{x \in X} \in \Delta(X)$ denote the unique stationary distribution associated with $P^\varepsilon$. The theory of resistance trees presented in [25] provides efficient mechanisms for computing the support of the limiting stationary distribution, i.e., $\lim_{\varepsilon \to 0^+} \mu^\varepsilon$, commonly referred to as the stochastically stable states.

*Definition 3:* A state $x \in X$ is *stochastically stable* [6] if $\lim_{\varepsilon \to 0^+} \mu^\varepsilon_x > 0$, where $\mu^\varepsilon$ is the stationary distribution corresponding to $P^\varepsilon$.

We adopt the technique provided in [25] for identifying the stochastically stable states through a graph theoretic analysis over recurrent classes of the unperturbed process $P^0$. Let $Y_0, Y_1, \dots, Y_m$ denote the recurrent classes of $P^0$. Define $\mathcal{P}_{ij}$ to be the set of all paths connecting $Y_i$ to $Y_j$, i.e., a path $p \in \mathcal{P}_{ij}$ is of the form $p = \{(x_1, x_2), (x_2, x_3), \dots, (x_{k-1}, x_k)\}$ where $x_1 \in Y_i$ and $x_k \in Y_j$. The resistance associated with transitioning from $Y_i$ to $Y_j$ is defined as

$$r(Y_i, Y_j) = \min_{p \in \mathcal{P}_{ij}} \sum_{(x, x') \in p} r(x, x'). \tag{18}$$

Recurrent classes $Y_0, Y_1, \dots, Y_m$ satisfy: (i) there is a zero resistance path, i.e., a sequence of transitions each with zero resistance, from any state $x \in X$ to at least one state $y$ in one of the recurrent classes; (ii) for any recurrent class $Y_i$ and any states $y_i, y'_i \in Y_i$, there is a zero resistance path from $y_i$ to $y'_i$; and (iii) for any state $y_i \in Y_i$ and $y_j \in Y_j$, $Y_i \neq Y_j$, any path from $y_i$ to $y_j$ has strictly positive resistance.

The first step in identifying the stochastically stable states is to determine resistances between recurrent classes. The second step is to analyze spanning trees of the weighted, directed graph $\mathcal{G}$ whose vertices are recurrent classes of the process $P^0$, and whose edges are weighted by resistances between classes in (18). Denote $\mathcal{T}_i$ to be the set of all spanning trees of $\mathcal{G}$ rooted at recurrent class $Y_i$. We compute the stochastic potential of each recurrent class, defined as:

*Definition 4:* The *stochastic potential* of recurrent class $Y_i$ is

$$\gamma(Y_i) = \min_{\mathcal{T} \in \mathcal{T}_i} \sum_{(Y, Y') \in \mathcal{T}} r(Y, Y')$$

The following theorem characterizes the recurrent classes that are stochastically stable.

*Theorem 2 ( [25]):* Let $P^0$ be the transition matrix for a stationary Markov process over the finite state space $X$

with recurrent communication classes $Y_1, \ldots, Y_m$. For each $\varepsilon > 0$, let $P^\varepsilon$ be a regular perturbation of $P^0$ with a unique stationary distribution $\mu^\varepsilon$. Then:
(1) As $\varepsilon \to 0^+$, $\mu^\varepsilon$ converges to a stationary distribution $\mu^0$ of $P^0$.
(2) A state $x \in X$ is stochastically stable if and only if $x$ is contained in a recurrent class $Y_j$ that minimizes $\gamma(Y_j)$.

### B. Proof of Theorem 1

We begin by restating the contributions associated with Theorem 1 using the terminology of the previous section.

- If $q(S) \cap \text{CCE} \neq \emptyset$, then a state $x = \{x_i = [s_i, m_i]\}_{i \in N}$ is stochastically stable if and only if (i) $m_i = C$ for all $i \in N$ and (ii) the strategy profile $s = (s_1, \ldots, s_n)$ constitutes an efficient coarse correlated equilibrium, i.e.,

$$q(s) \in \underset{q \in q(S) \cap \text{CCE}}{\arg\max} \sum_{i \in N} \sum_{a \in \mathcal{A}} U_i(a) q^a. \tag{19}$$

- If $q(S) \cap \text{CCE} = \emptyset$, then a state $x = \{x_i = [s_i, m_i]\}_{i \in N}$ is stochastically stable if and only if (i) $m_i = C$ for all $i \in N$ and (ii) the strategy profile $s = (s_1, \ldots, s_n)$ constitutes an efficient action profile, i.e.,

$$q(s) \in \underset{q \in q(S)}{\arg\max} \sum_{i \in N} \sum_{a \in \mathcal{A}} U_i(a) q^a. \tag{20}$$

For convenience, and with an abuse of notation, define

$$U_i(s) := \sum_{a \in \mathcal{A}} U_i(a) q^a(s) \tag{21}$$

to be agent $i$'s expected utility with respect to distribution $q(s)$, where $s \in S$.

The proof of Theorem 1 will consist of the following steps:
(1) Define the unperturbed process, $P^0$.
(2) Determine the recurrent classes of process $P^0$.
(3) Establish transition probabilities of process $P^\varepsilon$.
(4) Determine the stochastically stable states of $P^\varepsilon$ using Theorem 2.

#### *Part 1: Defining the unperturbed process*

The unperturbed process $P^0$ is effectively the process defined in Section III where $\varepsilon = 0$. We highlight the attributes of the unperturbed process that may not be immediately clear.

- If agent $i$ is content, i.e., $x_i = [s_i^b, C]$, the trial action is $s_i^t = s_i^b$ with probability 1. Otherwise, if agent $i$ is discontent, the trial action is selected according to (10).
- The baseline utility $u_i^b$ in (8) associated with joint baseline strategy $s^b$ is now of the form

$$u_i^b = U_i(s^b). \tag{22}$$

  This results from invoking the law of large numbers since $\bar{p} = \lceil 1/\varepsilon^{nc+1} \rceil$. The trial utility $u_i^t$ and acceptance utility $u_i^a$ are also of the same form.
- A content player will only become discontent if $u_i^a < u_i^b$ where associated payoffs are computed according to (22).

#### *Part 2: Recurrent classes of the unperturbed process*

The second part of the proof analyzes the recurrent classes of the unperturbed process $P^0$ defined above. The following lemma identifies the recurrent classes of $P^0$.

*Lemma 1:* A state $x = (x_1, x_2, \ldots, x_n) \in X$ belongs to a recurrent class of the unperturbed process $P^0$ if and only if the state $x$ fits into one of following two forms:

- *Form #1:* The state for each agent $i \in N$ is of the form $x_i = [s_i^b, C]$ where $s_i^b \in S_i$. Each state of this form comprises a distinct recurrent classes. We represent the set of states of this form by $C^0$.
- *Form #2:* The state for each agent $i \in N$ is of the form $x_i = [s_i^b, D]$ where $s_i^b \in S_i$. All states of this form comprise a single recurrent class, represented by $D^0$.

We omit the proof of Lemma 1 for brevity.

#### *Part 3: Transition probabilities of process $P^\varepsilon$*

The transition probability $P^\varepsilon_{x \to x^+}$ for any $x, x^+ \in X$ is

$$P^\varepsilon_{x \to x^+} = \sum_{\tilde{s}^t \in S} \sum_{\tilde{s}^a \in S} \Big( \Pr[x^+ \mid s^t = \tilde{s}^t,\, s^a = \tilde{s}^a] \times \Pr[s^a = \tilde{s}^a \mid s^t = \tilde{s}^t] \Pr[s^t = \tilde{s}^t] \Big). \tag{23}$$

Strategy selections and state transitions also depend on state $x$; for brevity we do not explicitly write this dependence. Here, $s^t$ and $s^a$ represent joint trial and acceptance strategies during the period before the transition to $x^+$.

The detailed form of these transition probabilities may be determined using probabilities defined in Section III and by integrating over the possible payoffs each agent may receive, $u_i \in [0, 1]$. We omit this detailed discussion for brevity.

*Lemma 2:* The process $P^\varepsilon$ is a regular perturbation of $P^0$.

It is straightforward that $P^\varepsilon$ satisfies the first two conditions of Definition 2 with respect to $P^0$, and transition probabilities satisfy (17) due to the fact that the dominant terms in $P^\varepsilon_{x \to y}$ are polynomial in $\varepsilon$.

#### *Part 3: Determining the stochastically stable states*

We begin by defining

$$C^\star := \{x = \{[s_i, C]\}_{i \in N} \,:\, q(s) \in \text{CCE}\} \subseteq C^0$$

Here, we show that, if $C^\star$ is nonempty, then a state $x$ is stochastically stable if and only if $q(s)$ satisfies (19). The fact that $q(s)$ must satisfy (20) when $C^\star = \emptyset$ follows in a similar manner. To accomplish this task, we (1) establish resistances between recurrent classes, and (2) compute stochastic potentials of each recurrent class.

#### *Resistances between recurrent classes*

*Claim 1:* Resistances between recurrent classes satisfy:

- For $x \in C^0$ with corresponding joint strategy $s$, $r(D^0 \to x) = \sum_{i \in N} (1 - U_i(s))$.
- For a transition of the form $x \to y$, where $x \in C^\star$ and $y \in (C^0 \cup D^0) \setminus \{x\}$, $r(x \to y) \geq 2c$.
- For a transition of the form $x \to y$ where $x \in C^0 \setminus C^\star$ and $y \in (C^0 \cup D^0) \setminus \{x\}$, $r(x \to y) \geq c$.
- For every $x \in C^0 \setminus C^\star$, there exists a path $x = x^0 \to x^1 \to \cdots \to x^m \in C^\star \cup D^0$ with resistance $r(x^j \to x^{j+1}) = c$, $\forall j \in \{0, 1, \ldots, m-1\}$.

These resistances are computed in a similar manner to those in [16]; however, care must be taken due to the fact

that there is a small probability that average received utilities fall outside of $U_i(s) \pm \varepsilon$ during a phase in which joint strategy $s$ is played. The detailed proof is omitted for brevity.

*Stochastic potentials*

The following lemma specifies stochastic potentials of each recurrent class. Using resistances from Claim 1, the stochastic potentials follow from the same arguments as in [16]; we omit the proof for brevity.

*Lemma 3:* Let $x \in C^0 \setminus C^\star$ with corresponding joint strategy $s$, and let $x^\star \in C^\star$ with corresponding joint strategy $s^\star$. The stochastic potentials of each recurrent class are:

$$\gamma(D^0) = c|C^0 \setminus C^\star| + 2c|C^\star|,$$
$$\gamma(x) = \left(|C^0 \setminus C^\star| - 1\right) c + 2c|C^\star| + \sum_{i \in N}(1 - U_i(s)),$$
$$\gamma(x^\star) = |C^0 \setminus C^\star|c + 2c\left(|C^\star| - 1\right) + \sum_{i \in N}(1 - U_i(s^\star)),$$

We now use Lemma 3 to complete the proof of Theorem 1. For the first part, suppose $C^\star$ is nonempty, and let $x^\star \in \arg\max_{x \in C^\star} \sum U_i(s)$, where joint strategy $s$ corresponds to state $x$. Then,

$$\begin{aligned}\gamma(x^\star) &= |C^0 \setminus C^\star|c + 2c\left(|C^\star| - 1\right) + \sum_{i \in N}(1 - U_i(s^*)) \\ &< |C^0 \setminus C^\star|c + 2c|C^\star| \quad (\text{since } c \geq n) \\ &= \gamma(D).\end{aligned}$$

For $x \in C^0$,

$$\begin{aligned}\gamma(x^\star) &= |C^0 \setminus C^\star|c + 2c\left(|C^\star| - 1\right) + \sum_{i \in N}(1 - U_i(s^\star)) \\ &< |C^0 \setminus C^\star - 1|c + 2c\left(|C^\star|\right) + \sum_{i \in N}(1 - U_i(s)) \\ &= \gamma(x).\end{aligned}$$

For $x \in C^\star$ with $x \notin \arg\max_{x \in C^\star} \sum U_i(s)$,

$$\begin{aligned}\gamma(x^\star) &= |C^0 \setminus C^\star|c + 2c\left(|C^\star| - 1\right) + \sum_{i \in N}(1 - U_i(s^\star)) \\ &< |C^0 \setminus C^\star|c + 2c\left(|C^\star| - 1\right) + \sum_{i \in N}(1 - U_i(s) \\ &= \gamma(x).\end{aligned}$$

Applying Theorem 2, $x^\star$ is stochastically stable. Since all other states have strictly larger stochastic potential, *only* states $x^\star \in C^\star$ with $x^\star \in \arg\max_{x \in C^\star} \sum U_i(s)$ are stochastically stable. From state $x^\star$, if each agent plays according to its baseline strategy, then the probability that joint action $a \in \mathcal{A}$ is played at any given time is $\Pr(a = a') = q^{a'(s^\star)}$. This implies that a CCE which maximizes the sum of agents' payoffs is played with high probability as $\varepsilon \to 0$, after sufficient time has passed.

The second part of the theorem follows similarly by considering the case when $C^\star = \emptyset$. ∎

## V. Conclusion

The majority of distributed learning literature has focused on identifying learning rules that converge to Nash equilibria. However, alternate forms of behavior, such as correlated equilibrium, can often lead to significant improvements in system-wide behavior. This paper focuses on identifying learning rules that converge to joint distributions that do not necessarily constitute Nash equilibria. In particular, we have extended the work of [16] to provide a distributed learning rule that ensures agents play strategies that constitute efficient coarse correlated equilibria. A mild variant of the proposed algorithm could also ensure the agents play strategies that constitute correlated equilibria, as opposed to coarse correlated equilibria. Future work seeks to investigate the applicability of such algorithms in the context of team versus team zero-sum games.

## References


[1] I. Arieli and Y. Babichenko. Average Testing and the Efficient Boundary. *Journal of Economic Theory*, 147:2376–2398, 2012.

[2] R.J. Aumann. Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica*, 55(1):1–18, 1987.

[3] L. E. Blume. The statistical mechanics of strategic interaction. *Games and Economic Behavior*, 1993.

[4] O. Boussaton and J. Cohen. On the distributed learning of Nash equilibria with minimal information. *6th International Conference on Network Games, Control, and Optimization*, 2012.

[5] A.J. Carfang, E.W. Frew, and D.B. Kingston. A Cascaded Approach to Optimize Aircraft Trajectories for Persistent Data Ferrying. In *AIAA Gui*, 2013.

[6] D. Foster and H.P. Young. Stochastic evolutionary game dynamics. *Theoretical Population Biology*, 38:219–232, 1990.

[7] D.P. Foster and R.V. Vohra. Calibrated Learning and Correlated Equilibrium. *Games and Economic Behavior*, 21:40–55, October 1997.

[8] 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.

[9] P. Frihauf, M. Krstic, and T. Bas. Nash Equilibrium Seeking in Noncooperative Games. *IEEE Transactions on Automatic Control*, 57(5):1192–1207, 2012.

[10] B. Gharesifard and J. Cortes. Distributed convergence to Nash equilibria by adversarial networks with directed topologies. In *51st IEEE Conference on Decision and Control*, 2012.

[11] S. Hart and A. MasColell. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[12] Y.C. Ho and F.K. Sun. Value of Information in Two-Team Zero-Sum Problems. *Journal of Optimization Theory and Applications*, 14(5), 1974.

[13] A.X. Jiang and K. Leyton-Brown. Polynomial-time computation of exact correlated equilibrium in compact games. In *Proceedings of the Twelfth ACM Electronic Commerce Conference*, February 2011.

[14] Y. Lim. *Game Theoretic Distributed Coordination: Drifting Environments and Constrained Communication*. PhD thesis, Georgia Tech, 2014.

[15] J. R. Marden and J. S. Shamma. Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation. *Games and Economic Behavior*, 75(2):788–808, 2012.

[16] J.R. Marden. Selecting Efficient Correlated Equilibria Through Distributed Learning. *under submission*.

[17] J.R. Marden and H.P. Young. Payoff Based Dynamics for Multi-Player Weakly Acyclic Games. *SIAM Journal on Control and Optimization*, 48(1):373–396, 2009.

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

[19] LE Ortiz. Maximum entropy correlated equilibria. *AISTATS*, pages 347–354, 2007.

[20] C.H. Papadimitriou and T. Roughgarden. Computing correlated equilibria in multi-player games. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, volume 55, July 2005.

[21] J. Poveda and N. Quijano. Distributed Extremum Seeking for Real-Time Resource Allocation. In *American Control Conference*, 2013.

[22] B.S.R. Pradelski and H.P. Young. Learning efficient Nash equilibria in distributed systems. 2012.

[23] B. von Stengel and D. Koller. Team-Maxmin Equilibria. *Games and Economic Behavior*, 21(1-2):309 – 321, 1997.

[24] H. P. Young. Learning by trial and error. *Games and economic behavior*, 65:626–643, 2009.

[25] H.P. Young. The Evolution of Conventions. *Econometrica*, 61(1):57–84, 1993.