Learning Efficient Correlated Equilibria

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

Abstract—The vast majority of the literature in distributed learning focuses on attaining convergence to Nash equilibria. Correlated equilibria, on the other hand, can often characterize collective behavior that is far more efficient than even the best Nash equilibrium. However, there are no distributed learning algorithms in the existing literature that coverage to specific correlated equilibria. In this paper, we provide one such algorithm. In particular, the proposed algorithm guarantees that the agents’ collective joint strategy will constitute an efficient correlated equilibrium with high probability. The key to attaining efficient correlated behavior as the result of a distributed learning process was incorporating a common random signal into the learning environment.

I. INTRODUCTION

Agents’ individual control laws are a crucial component of any multiagent system. These control laws, or learning algorithms, dictate how individual agents process locally available information to make a decision. A variety of factors determine the quality of a given learning algorithm, including informational dependencies, asymptotic guarantees, and convergence rates. Accordingly, significant research attention in both the engineering and social sciences 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], [11], [20], [22]. Recently, the research focus has shifted to ensuring convergence to alternative 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], [21], potential function maximizers [3], [15], welfare maximizing action profiles [1], [18], correlated equilibria [2], [6], [12], [16], among others.

In most of the 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. The importance of developing learning algorithms that converge to an efficient correlated equilibrium is driven by the fact that optimal system behavior can often be characterized by a correlated equilibrium. Unfortunately, the aforementioned learning algorithms, such a regret matching [12], merely provide convergence to the set of correlated equilibria. This means that the resulting behavior does not necessarily converge to – or even approximate – a specific correlated equilibrium at any instance of time.

The goal of this paper is to provide a simple distributed learning algorithm that provides convergence to the most efficient, i.e., welfare maximizing, correlated equilibrium. For concreteness, consider a mild variant of the Shapley game characterized by the following payoff matrix

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>M</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>1,ε</td>
<td>-ε,1</td>
<td>0,0</td>
</tr>
<tr>
<td>M</td>
<td>0,0</td>
<td>1,ε</td>
<td>-ε,1</td>
</tr>
<tr>
<td>B</td>
<td>-ε,1</td>
<td>1,0</td>
<td>1,ε</td>
</tr>
</tbody>
</table>

where ε > 0 is a small constant. In this game, there are two players (Row, Column), the row player has three actions (T,M,B), the column player has three actions (L,M,R), and the numbers highlight that payoffs to both players for each of the nine joint actions. The unique Nash equilibrium for this game is when each player uses a probabilistic strategy that selects each of the three actions with probability 1/3 and this yields an expected payoff of $\approx 1/3$ to each player. Alternatively, consider 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. Given this joint distribution, the expected payoff to each player is $\approx 1/2$ which exceeds the performance of the Nash equilibrium. Note that this distribution cannot be realized by independent strategies associated with the players. Rather, this joint distribution represents a specific correlated equilibrium.

As the above example demonstrates, deriving distributed learning algorithms that converge to efficient correlated equilibria is clearly desirable from a system-wide perspective. In line with this theme, a recent result in [16] proposed a distributed algorithm which guarantees that the empirical frequency of the agents’ collective behavior will converge to an efficient correlated equilibrium; however, the convergence in empirical frequencies is attained through deterministic cyclic behavior of the agents. For example, in the above Shapley game, the algorithm posed in [16] guarantees that the collective behavior of the agents will follow the cycle $(T,L) \rightarrow (T,M) \rightarrow (M,M) \rightarrow (M,R) \rightarrow (B,R) \rightarrow (B,L) \rightarrow (T,L)$ with high probability. Following this deterministic cycle results in an empirical frequency of play that equates to the efficient correlated equilibrium highlighted above; however, it is important to highlight that at any time instance the players are not playing a joint strategy in accordance with this efficient correlated equilibrium.

Predictable, cyclic behavior may be desirable from a system-wide perspective for many applications, e.g., data ferrying [5]. However, predictable, cyclic behavior could be exploited in many other situations, e.g., team versus team...
zero-sum games \[13\]. By viewing each team as a single player, classical results pertaining to two-player zero-sum games suggest that a team’s desired strategy is to play its security strategy, which constitutes a distribution over the team’s joint action space. These security strategies may often be impossible to realize through independent strategies of the team members. Accordingly, establishing distributed learning algorithms that can stabilize specific joint strategies, such as correlated equilibria, is necessary for providing strong performance guarantees in such settings.

The main contribution of this paper is the development of a distributed learning algorithm that ensures that the agents collectively play a joint distribution corresponding to the efficient correlated equilibrium. With regards to the Shapley game, our algorithm guarantees that the agents collectively play the highlighted joint distribution with high probability. Attaining such guarantees on the underlying joint strategy is non-trivial as we aim to ensure desired correlated behavior through the design of learning rules where individual agents make independent decisions in response to local information. The key element of our algorithm which makes this correlation possible is the introduction of a common random signal to the agents which is incorporated into their local decision-making rule. Another important feature of our algorithm is that it is completely uncoupled [8], i.e., agents make decisions based only on their received utility and their observation of the common random signal. In such settings, agents have no knowledge of the payoff or behavior of other agents, nor do they have any information regarding the structural form of their utility functions.

Lastly, there is a series of recent results focused on developing efficient centralized algorithm for computing specific correlated equilibria [14], [19]. 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 the 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 \(N = \{1, 2, \ldots, n\}\), and each agent \(i \in N\) is associated with a finite action set \(A_i\) and a utility function \(U_i : A \to [0, 1]\) where \(A = A_1 \times A_2 \times \cdots \times A_n\) denotes the joint action space. We will often represent such a game by the tuple \(G = \{N, \{U_i\}_{i \in N}, \{A_i\}_{i \in N}\}\).

In this paper we focus on a class of equilibria termed coarse correlated equilibria [2]. A coarse correlated equilibrium is characterized by a joint distribution \(q = \{q^a\}_{a \in A} \in \Delta(A)\), where \(\Delta(A)\) represents the simplex over the finite set \(A\), such that for any agent \(i \in N\) and action \(a'_i \in A_i\) we have

\[
\sum_{a \in A} U_i(a, a_{-i})q^a \geq \sum_{a \in A} U_i(a'_i, a_{-i})q^a,
\]

where \(a_{-i} = \{a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n\}\) denotes the collection of action of all players other than player \(i\).\(^{1}\) Informally, a coarse correlated equilibrium represents a joint distribution where each agent’s expected utility for going along with the joint distribution is at least as good as his utility for deviating to any fixed action. We say that a coarse correlated equilibrium \(q^*\) is efficient if it maximizes the sum of the expected payoffs of the agents, i.e.,

\[
q^* \in \arg \max_{q \in \text{CCE}} \sum_{i \in N} \sum_{a \in A} U_i(a)q^a.
\]

where CCE \(\subset \Delta(A)\) denotes the set of coarse correlated equilibria. It is well-known that CCE \(\neq \emptyset\).

The focus of this paper is on the derivation of a distributed learning algorithm that ensures that the collective behavior of the agents converges to an efficient coarse correlated equilibrium. Here, we adopt the framework of repeated one-shot games, where a static game \(G\) is repeated over time and agents are permitted to use observations from previous plays of the game to formulate a decision. More specifically, a repeated one-shot game yields a sequence of action profiles \(a(0), a(1), \ldots\), where at each time \(t \in \{0, 1, 2, \ldots\}\) the decision of each agent \(i\) is chosen independent according to the agent’s strategy at time \(t\), which we denote by \(p_i(t) = \{p_i^a(t)\}_{a \in A_i}, \in \Delta(A_i)\). In this paper, we focus on the case where the strategy of agent \(i\) at time \(t\) is selected according to a learning rule of the form

\[
p_i(t) = F_i(\{a_{i}(\tau), U_i(a(\tau))\}_{\tau=0,\ldots,t-1})
\]

which specifies how each agent processes available information to formate the agent’s strategy at the ensuing time-steps. Learning rules of the form (3) are termed completely uncoupled [8] and represent one of the most informationally restrictive classes of learning rules as 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 gauge the performance of a learning rule \(\{F_i\}_{i \in N}\) by the resulting asymptotic guarantees. With that goal in mind, let \(q(t) \in \Delta(A)\) represent the agents’ collective strategy at time \(t\), which is of the form

\[
q^{(a_1, \ldots, a_n)}(t) = p_1^{a_1}(t) \times \cdots \times p_n^{a_n}(t)
\]

where \(\{p_i(t)\}_{i \in N}\) are the individual agent strategies at time \(t\). The goal of this paper is to derive learning rules of the form (3) that guarantee that the agents’ collective strategy constitutes an efficient coarse correlated equilibrium the majority of the time, i.e., for all sufficient large times \(t\), we have

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

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 \(t \in \{0, 1, \ldots\}\) that is i.i.d. and drawn uniformly from the interval \([0, 1]\). Now, the considered distributed learning rule takes on the form

\[
p_i(t) = F_i(\{a_{i}(0), U_i(a(\tau), z(\tau))\}_{\tau=0,\ldots,t-1}). \tag{6}
\]

Here, this common signal can be exploited as a coordinating
entity to reach collective strategies that are not of the form (4).

III. A LEARNING ALGORITHM FOR ATTAINING EFFICIENT CORRELATED EQUILIBRIA

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

A. Preliminaries

Consider a situation where each agent $i \in N$ commits to a signal-based strategy of the form $s_i : [0, 1] \rightarrow A_i$ which associates with each signal $z \in [0, 1]$ an action $s_i(z) \in A_i$. With somewhat of an abuse of notation, we consider a associates with each signal $s$ equal length and associating each region with a specific

In essence, the considered strategies correspond to the agent ($S_i \in \cup_{k=1}^\omega \{A_i\}^k$ where $\omega \geq 1$ is a design parameter that identifies the granularization of the agent’s possible strategies. A strategy $s_i = (a_1^i, \ldots, a_m^i) \in S_i$, $m \leq \omega$, leads to a strategy of the form

$$s_i(z) = \begin{cases} 
    a_1^i & \text{if } z \in [0, 1/m) \\
    a_2^i & \text{if } z \in [1/m, 2/m) \\
    \vdots \\
    a_m^i & \text{if } z \in [(m-1)/m, 1]. 
\end{cases} \quad (7)$$

In essence, the considered strategies correspond to the agent breaking up the unit interval into at most $\omega$ regions of equal length and associating each region with a specific action. If the agents commit to a a strategy profile $s = (s_1, s_2, \ldots, s_N) \in S = \prod_{i \in N} S_i$, then the resulting joint strategy $q(s) = \{q^a(s)\}_{a \in A} \in \Delta(A)$ satisfies

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

where $I\{\cdot\}$ is the usual indicator function. We define the set of joint distributions that can be realized by the strategies $S_i$ as

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

B. Algorithm description

The forthcoming algorithm will be reminiscent of the trial and error learning algorithm introduced in [22]. Each agent will be associated with a baseline strategy, a baseline utility, and a mood that will influence the agent’s decision making as we highlight below.

We begin by defining a constant $c > n$ and an experimentation rate $\epsilon \in (0, 1)$. The algorithm proceeds as follows:

- At each instance of time $t \in \{0, 1, \ldots\}$, each agent will be playing a strategy $s_i(t) \in S_i$.
- If an agent selects a strategy $s_i(t)$ a time $t$, this agent commits to playing this strategy for $\bar{p} = [1 / \epsilon^{ac+2}]$ consecutive iterations. This gives the agent ample time to evaluate the strategy before revising his option.
- For mathematical convenience, we consider the case where the evaluation periods are aligned. Accordingly, agents will only revise their strategies at the specific time instances $\{\bar{p}, 2\bar{p}, 3\bar{p}, \ldots\}$. We will refer to the time periods $\{1, \ldots, \bar{p}\}$ as the first period, $\{\bar{p}+1, \ldots, 2\bar{p}\}$ as the second period, and so on.

The forthcoming algorithm identifies a rule by which each agent selects an ensuing strategy at the end of a given period. A core element associated with this rule is an internal state variable for each agent $i \in N$ of the form

$$x_i = (x_i^1, x_i^2, m_i)$$

where $x_i^1 \in S_i$ is the agent’s baseline strategy, $x_i^2 \in [0,1]$ is the agent’s baseline utility, and $m_i \in \{\text{Content, Discontent, Hopeful, Watchful}\}$ is the agent’s mood.

Here, the terminology of the moods is directly inherited from [22]. The algorithm will be divided into two distinct parts:

- Part #1: Agent Dynamics – The strategy selected by each agent $i \in N$ for period $k$, denoted by $s_i(k)$, depends purely on the state of agent $i$ at beginning of the $k$-th period, i.e.,

$$s_i(k) \sim \Pi_{\bar{a}}(x_i(k)). \quad (8)$$

For notational simplicity, let $x_i(k) = [x^1_i, x^2_i, m_i]$. The following describes how each agent selects the strategy $s_i(k)$.

- Content, $m_i = C$: When agent $i$ is content, the strategy $s_i(k)$ is chosen according to

$$\Pr [s_i(k) = s_i^1] = \begin{cases} 
1 - e^c & \text{if } s_i^1 = x_i^1 \\
\frac{e^c}{|A_i|} & \text{for any } s_i^1 = a_i \in A_i 
\end{cases} \quad (9)$$

Note that a strategy $s_i(k) = a_i$ implies that agent $i$ is committing to play action $a_i$ for the entire $k$-th period. A content player will primarily select its baseline strategy.

- Discontent, $m_i = D$: When agent $i$ is discontent, the strategy $s_i(k)$ is chosen randomly from the set of strategies $S_i$, i.e.,

$$\Pr [s_i(k) = s_i^1] = 1 / |S_i| \text{ for all } s^1_i \in S_i. \quad (10)$$

- Hopeful, $m_i = H$, or Watchful, $m_i = W$: When an agent is hopeful or watchful, the agent selects its baseline strategy, i.e., $s_i(k) = x_i^1$.

- Part #2: State Dynamics – At the end of the $k$-the period, each agent $i \in N$ updates its state according to a rule of the form

$$x_i(k+1) \sim \Pi_{\bar{a}}(s_i(k), u_i(k), x_i(k)). \quad (11)$$

where $u_i(k)$ denotes the average payoff over the $k$-th stage, i.e.,

$$u_i(k) = \frac{1}{p(k+1)} \sum_{\tau=pk+1}^{pk+2} U_i(s(z(\tau))), \quad s = s(k),$$

and $s(z) = (s_1(z), s_2(z), \ldots, s_N(z))$. For notational simplicity, let $x_i(k) = (x_i^1, x_i^2, m_i)$, $u_i = u_i(k)$, and $s_i = s_i(k)$. The following describes how each agent updates their local state variable at the end $k$-th period. Once again, observe that the agent’s mood plays a fundamental role in this update process.

- Content, $m_i = C$: When an agent is content, the ensuing state is highly dependent on whether or not $s_i = s_i^1$, in addition to how $s_i$ compares to $u_i^1$. The following table illustrates what the ensuing state $x_i(k+1)$ will be for all
possible conditions on \( s_i \) and \( u_i \):

<table>
<thead>
<tr>
<th>( s_i = s_i^h )</th>
<th>( s_i = s_i^w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u_i &gt; u_i^h + \epsilon )</td>
<td>( u_i &lt; u_i^h - \epsilon )</td>
</tr>
<tr>
<td>( s_i^h, u_i, C )</td>
<td>( s_i^h, u_i, W )</td>
</tr>
<tr>
<td>( s_i^w, u_i, C )</td>
<td>( s_i^w, u_i, C )</td>
</tr>
</tbody>
</table>

Note that a player only transitions to Hopeful or Watchful under the condition when \( s_i = s_i^h \).

- **Discontent**, \( m_i = D \): When an agent is discontent, the ensuing state is chosen according to

  \[
  x_i(k + 1) = \begin{cases} 
  [s_i^h, u_i, C] & \text{w.p. } \epsilon^{1-u_i}, \\
  [s_i^h, u_i, D] & \text{w.p. } 1 - \epsilon^{1-u_i}. 
  \end{cases}
  \]  

  (12) Hence, an agent is more likely to become content with strategies that yield higher average payoffs.

- **Hopeful**, \( m_i = H \): When an agent is hopeful, the new state of the agent is chosen according to

  \[
  x_i(k + 1) = \begin{cases} 
  [s_i^h, u_i, C] & \text{if } u_i \geq u_i^h - \epsilon, \\
  [s_i^h, u_i, W] & \text{if } u_i < u_i^h - \epsilon. 
  \end{cases}
  \]  

  (13) Hence, an agent becomes content if the realized payoff is high enough. Otherwise, the agent transitions to watchful.

- **Watchful**, \( m_i = W \): When an agent is watchful, the new state of the agent is chosen according to

  \[
  x_i(k + 1) = \begin{cases} 
  [s_i^h, u_i, H] & \text{if } u_i \geq u_i^h - \epsilon, \\
  [s_i^h, u_i, D] & \text{if } u_i < u_i^h - \epsilon. 
  \end{cases}
  \]  

  (14) Hence, an agent becomes discontent if the realized payoff is low enough. Otherwise, the agent transitions to hopeful.

**C. Main Result**

In order to provide the desired asymptotic guarantees, we need to ensure that the underlying game structure has a minor degree of coupling between the agents’ utility functions. The following definition of interdependence, taken from [22], captures the desired coupling.

**Definition 1**: A game \( G \) with agents \( N = \{1, 2, \ldots, n\} \) is said to be independent 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 \( U_i(a'_j, a_{-i}) \neq U_i(a_j, a_{-i}) \).

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

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

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

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

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

A few remarks are on order regarding Theorem 1. First, observe that the proposed algorithm is of the form (6). Second, the condition \( q(S) \cap \text{CCE} \neq \emptyset \) implies that the agents can realize specific joint distributions that are coarse correlated through the joint strategy set \( S \). When this is the case, the above algorithm ensures that the agents predominantly play a strategy \( s \in S \) where the resulting joint distribution \( q(s) \) corresponds to the efficient coarse correlated equilibrium. Alternatively, the condition \( q(S) \cap \text{CCE} = \emptyset \) implies that there are no agent strategies that can characterize a coarse correlated equilibrium. When that is the case, the above algorithm ensures that 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. Conclusion**

In this work we have extended the work of [16] to provide a distributed learning rule that ensures that the agents play strategies that constitute efficient coarse correlated equilibria. A mild variant of the proposed algorithm could also ensure that 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**


APPENDIX

A. Background: Resistance Trees

The proof of Theorem 1 involves identifying the support to the limiting distribution \( P^\epsilon \) as \( \epsilon \to 0 \). To accomplish this task, we will use the resistance tree theory developed in [23] and building from [9]. The following definitions and theorems are taken from [23]. We consider a family of Markov chains, \( M^\epsilon \) over finite state space \( \Omega \).

**Definition 2:** A Markov chain, \( M^\epsilon \) which is a perturbation of some nominal process \( M^0 \), parameterized by \( \epsilon \in (0, a) \) for some \( a > 0 \), is a regular perturbed process [23] if, for all \( \omega, \tilde{\omega} \in \Omega \):

1. \( M^\epsilon \) is aperiodic and irreducible for all \( \epsilon \in (0, a] \).
2. \( \lim_{\epsilon \to 0} M_{\omega \to 0}^\epsilon = M_{\omega \to \tilde{\omega}}^0 \).
3. \( M_{\omega \to 0}^\epsilon > 0 \) for some \( \epsilon \) implies there exists \( r(\omega \to \tilde{\omega}) \) such that
   \[
   0 < \lim_{\epsilon \to 0} \frac{M_{\omega \to \tilde{\omega}}^\epsilon}{\epsilon r(\omega \to \tilde{\omega})} < \infty
   \] (15)

The number \( r(\omega \to \tilde{\omega}) \) is referred to as the resistance of the transition \( \omega \to \tilde{\omega} \).

The resistance from one recurrent class, \( \Omega_i \), to another, \( \Omega_j \) is defined to be the total resistance of the lowest resistance path from \( \Omega_i \) to \( \Omega_j \). Then, define the weighted, directed graph \( G \) whose vertices are recurrent classes of the process \( P^0 \), and whose edge weights are given by the resistances between classes. Denote \( T_{\Omega_i} \) to be the set of all spanning trees of \( G \) rooted at class \( \Omega_i \).

**Definition 3:** A state \( \omega \in \Omega \) is said to be stochastically stable [7] if \( \lim_{\epsilon \to 0} \mu^\epsilon(\omega) > 0 \), where \( \mu^\epsilon \) is the stationary distribution corresponding to \( M^\epsilon \).

**Definition 4:** The stochastic potential of recurrent class \( \Omega_i \) is
\[
\gamma(\Omega_i) = \min_{t \in T_{\Omega_i}} \sum_{(x,y) \in t} r(x \to y)
\]

Using these definitions, we restate Theorem 4 from [23], which we will use in the proof of Theorem 1.

**Theorem 2 (23):** Let \( M^0 \) be a stationary Markov process on the finite state space \( \Omega \) with recurrent communication classes \( \Omega_1, \ldots, \Omega_j \). Let \( M^\epsilon \) be a regular perturbation of \( M^0 \), and let \( \mu^\epsilon \) be its unique stationary distribution for every small positive \( \epsilon \). Then:

1. As \( \epsilon \to 0 \), \( \mu^\epsilon \) converges to a stationary distribution \( \mu^0 \) of \( M^0 \).
2. \( \omega \) is stochastically stable if and only if \( \omega \) is contained in a recurrent class \( \Omega_i \) that minimizes \( \gamma(\Omega_i) \).

B. Alternate Statement of Theorem 1

Here, we provide an equivalent statement of Theorem 1 in order to use the resistance tree tools reviewed in Appendix A. Begin by noting that, if we allow the parameter \( \epsilon \) to vary, this results in an infinite number of joint agent states, \( x \in \prod_{i \in N} X_i \), since agent payoffs may take on many different values depending on the signals sent in each period. However, the specified agent decision making rules precisely specify a Markov chain over the finite state space we describe below. Analyzing emergent behavior over the corresponding finite state space simplifies analysis by allowing us to use the resistance tree tools of [23].

Define the system state corresponding to agent \( i \)'s internal state, \( x_i = [s_i^0, u_i^0, m_i] \), by

\[
y_i = \begin{cases} 
   s_i^b & \text{agent } i \text{'s baseline strategy} \\
   s_i^b & \text{joint trial strategy in } S \text{ played in the period when agent } i \text{ received payoff } u_i^b \\
   m_i & \text{mood (C, D, H, W)}
\end{cases}
\]

We denote the set of possible system states for agent \( i \) by \( Y_i \). The algorithm described above defines a Markov process, \( P^\epsilon \), parameterized by \( \epsilon > 0 \) over state space \( Y = \prod_{i \in N} Y_i \). For state \( y \in Y \) with \( y_i = [s_i^b, s_i^b, m_i] \) we will define the corresponding joint action by \( y^b := (s_i^b, s_i^b, \ldots, s_i^b) \).

Denote the set of states whose baseline joint action assignments induce coarse correlated equilibria by
\[
Y^{CCE} = \{ y \in Y : q(y^b) \in CCE \text{ and } m_i = C, \forall i \in N \}
\]

**Theorem 1 (alternate statement):** Let
\[
G = \{ N, \{U_i \}, \{A_i \} \}
\]
be a finite interdependent game with \( |N| \geq 2 \). If \( Y^{CCE} \) is nonempty, a state \( y \in Y \) is stochastically stable if and only if \( y \in Y^{CCE} \) and \( y \in \arg\max_{y \in Y^{CCE}} \sum_{i \in N} U_i(y^b) \). Furthermore, as \( \epsilon \to 0 \), the joint trial action assignment \( s \in S \) selected in any given period \( k \in N \) induces a coarse correlated equilibrium, \( q(s) \in CCE \), almost surely as \( \epsilon \to 0 \), and the probability of playing a joint action \( a \in A \) at any time during period \( k \) is given by \( q^a(s) \).

Otherwise, if \( Y^{CCE} = \emptyset \), a state \( y \in Y \) is stochastically stable if and only if \( y \in \arg\max_{y \in Y} \sum_{i \in N} U_i(y^b) \), where \( \tilde{Y} = \{ y \in Y : m_i = C, \forall i \} \).

Note that the process \( P^\epsilon \) is a regular perturbed process as defined in Appendix A. Moreover, by using Definition 3, it is straightforward to see that the two theorem statements are equivalent. Hence, we may proceed to prove Theorem 1 in the form above, using the resistance tree tools of [23]. In order to identify the stochastically stable states in \( Y \), we must first identify recurrent classes of the unperturbed process, and then characterize which of these recurrent classes are stochastically stable by applying Theorem 2. Accordingly, the proof of Theorem 1 has two parts: (A) characterize recurrent classes of the unperturbed process, \( P^0 \), and (B) characterize stochastically stable states of the perturbed process \( P^\epsilon \).

Before proceeding, we extend the utility function \( U_i \), for each \( i \in N \), to a function \( U_i : \Delta(A) \to \mathbb{R} \) on the simplex
over joint action space, $\Delta A$, as follows:

$$U_i(q) = \sum_{a \in A} U_i(a) q^a, \quad q = \{q^a\}_{a \in A} \in \Delta(A).$$

With an abuse of notation, we will write $U_i(s) := U_i(q(s))$.

C. Recurrent classes of the unperturbed process

We first define the unperturbed process $P^0$ over the state space, $Y$. We then show that $P^\epsilon$ converges to $P^0$ as $\epsilon \to 0$. Let $y_i(k) = [s_i^k, b_i, m_i]$ be agent $i$'s state at time $k \in \mathbb{N}$.

**Trial strategy in $P^0$:** Agent $i$'s trial strategy for the $k$th period, is chosen based on baseline strategy $s_i^0$, and mood $m_i$, according to the process in Section III-B, with $\epsilon = 0$.

**State updates in $P^0$:** System states in $Y$ evolve in process $P^0$ in parallel to the process over $X$ described in Section III-B, with $\epsilon = 0$. By noting that when $\epsilon = 0$, and joint strategy $s$ is played in period $k$, agent $i$ receives exactly payoff $U_i(s)$, it is straightforward to determine state transition probabilities in $P^0$.

**Convergence to the unperturbed process as $\epsilon \to 0$**

Note that, as $\epsilon \to 0$, $p = 1/\epsilon^{nc+2} \to \infty$. Then, since the signals $z(t)$ are sampled uniformly from $[0,1)$,

$$\lim_{p \to \infty} \frac{1}{p^{k+1}} \sum_{\tau = pk+1}^{p(k+1)} U_i(s(z(\tau))) = U_i(s)$$

where $s = (s_1, s_2, \ldots, s_n)$ is the joint action assignment for period $k$. Using this fact, it is straightforward to see that $P^\epsilon \to P^0$ as $\epsilon \to 0$.

**Lemma 1:** A state $y = (y_1, y_2, \ldots, y_n)$ belongs to a recurrent class of the unperturbed process if and only if it is of one of the following two forms:

**Form #1:** The state for every $i \in N$ takes the form: $y_i = [s_i^0, s_i, C]$, where $s = (s_1^0, s_2^0, \ldots, s_n^0)$. We represent the set of states of form #1 by $C^0$.

**Form #2:** The state for every $i \in N$ takes the form: $y_i = [s_i^0, s_i, D]$ with $s_i \in S_i$ and no restriction on $s_i$. We represent the set of states of form #2 by $D^0$.

We will prove Lemma 1 by showing the following:

I. $C^0$ and $D^0$ belong to recurrent classes of $P^0$.

II. For any state in a recurrent class of $P^0$, if $m_i = D$ for some $i \in N$, then $m_j = D$ for all $j \in N$.

III. For any state $y = (y_1, y_2, \ldots, y_n)$ in a recurrent class of $P^0$, with $y_i = [s_i^0, s_i, m_i]$ for each $i \in N$, either (i) $m_i = C$ and $s_i = (s_1^0, s_2^0, \ldots, s_n^0)$, OR (ii) $m_i = D$ for all $i \in N$.

**Proof:**

**Part I:** States of the form $C^0$ or $D^0$ belong to recurrent classes of $P^0$: In $P^0$, if $y \in C^0$, the system remains at state $y$ for all future time. Hence, states in $C^0$ are recurrent in the unperturbed process. Furthermore, for any states $y_i', y_i'' \in D^0$ there is positive probability in the unperturbed process of transitioning between the two states. Moreover, there is zero probability of exiting $D^0$ in process $P^0$. Hence $D^0$ is a recurrent class of $P^0$.

**Part II:** For any state in a recurrent class of $P^0$, if $m_i = D$ for some $i \in N$, then $m_j = D$ for all $j \in N$: Consider $y = (y_1, y_2, \ldots, y_n) \in Y$ with $y_i = [s_i^0, s_i, m_i], v_i \in N$, and $m_j = D$ for some $j \in N$. We will show that, from any state of this form, all agents eventually become discontent with probability $O(1)$ in $P^0$.

Let $J = \{j \in N : m_j = D\}$. If $J = N$ we are done, so assume $N \setminus J \neq \emptyset$. Let $s = (s_1, s_j, \ldots, s_n) = (s_1^0, s_j^0, s_n^0) \in S$ be the joint baseline strategy corresponding to state $y$, where $s_j \in \prod_{j \in J} S_j$ and $s_n \in \prod_{n \notin J} S_n$. Assume joint strategy $s$ was played in period $k - 1$, and that $y$ is the system state during period $k$. Recall that in process $P^0$, if $s$ is played during a given period, agent $i$ receives a utility of exactly $U_i(s)$, i.e., the agent state corresponding to $y_i$ is $x_i = [s_i^0, U_i(s_i^0), m_i]$.

We will show that some agent $i \notin J$ becomes discontent with probability $O(1)$.

By the interdependence condition, there exists an agent $i \notin J$ and joint strategy $(s_j, s_n)$ such that $U_i(s) \neq U_i(s_j, s_n)$. Assume that, as long as agent $i$ does not become discontent, no other agents in $N \setminus J$ become discontent; otherwise we are done. Then, in $P^0$ all agents in $N \setminus J$ play according to their baseline strategies $s_n$, and for any period $k' > k$, there is positive probability that agents in $J$ play any joint strategy $s_j \in \prod_{j \in J} S_j$. We will show, given any mood $m_i \in \{D, C, H, W\}$ there is a positive probability that agent $i$ becomes discontent within four periods. Since both joint strategies $s$ and $(s_j, s_n)$ are played with positive probability for all $t \geq 0$, this implies that agent $i$ eventually becomes discontent with probability $O(1)$ in $P^0$. Without loss of generality, assume $U_i(s) > U_i(s_j, s_n)$. For each possible mood of agent $i$, positive probability sequences of events leading it so become discontent are shown in Figure 1.

**Part III:** For any state $y = (y_1, y_2, \ldots, y_n)$ in a recurrent class of $P^0$, with $y_i = [s_i^0, s_i^b, m_i]$ for each $i \in N$, either (i) $m_i = C$ and $s_i = (s_1^0, s_2^0, \ldots, s_n^0)$, OR (ii) $m_i = D$ for all $i \in N$.

Let $y = (y_1, y_2, \ldots, y_n)$ be the state at time $k$ in $N \setminus J$, and let $s = (s_1^0, s_2^0, \ldots, s_n^0)$ be the corresponding baseline joint strategy. We will show that, starting from state $y$ and with probability $O(1)$, either all agents become discontent, or after at most two transitions, $s_i = s$ and $m_i = C$ for all $i \in N$.

From Part II, if any agent becomes discontent then all agents become discontent with probability $O(1)$. Hence, we may assume that no agent becomes discontent, implying that the joint strategy $s$ is played for all time periods $k' \geq k$. Suppose that $y_i = [s_i^0, s_i^b, m_i]$. It is straightforward to see that if $U_i(s) < U_i(s_i^b)$, agent $i$ will become discontent, which contradicts our assumption. Hence, $U_i(s) \geq U_i(s_i^b)$.

We show that, for any $m_i \in \{C, H, W\}$, agent $i$'s state transitions to $y_i' = [s_i, s, C]$ within two periods. For each mood, the following sequences of transitions occur with certainty in $P^0$, since the joint action assignment $s$ is played at each period, meaning agent $i$ receives utility $U_i(s') \geq U_i(s^b)$ in each period.

$$m_i = C : [s_i^b, s_i^b, C] \rightarrow [s_i^b, s_i^b, H] \rightarrow [s_i^b, s_i, C]$$
$$m_i = H : [s_i^b, s_i^b, H] \rightarrow [s_i^b, s_i, C]$$
$$m_i = W : [s_i^b, s_i^b, W] \rightarrow [s_i^b, s_i^b, H] \rightarrow [s_i^b, s_i, C].$$
Since all agents continue to play their baseline action assignments, the joint trial strategy is \( s \) for each subsequent period, and agent \( i \)'s state stays constant at \([s^b_i, s, C]\).

### D. Stochastically stable states

In this section we analyze behavior of the Markov chain \( P^\epsilon \) over \( Y \) as \( \epsilon \to 0 \). Recall, for any \( \epsilon > 0 \), the sequence of joint actions in the \( k \)th period is \( s(z(t_k^i)), s(z(t_k^j)), \ldots, s(z(t_k^n)) \), where \( s \) is the joint trial strategy in the \( k \)th period, and \( p = \left[ 1/ \epsilon^{p+2} \right] \). Agent \( i \) receives payoff \( u_i(k) = \frac{1}{p} \sum_{j=1}^{p} U_i(s(z(t_k^j))) \). Since the utility received depends on random signals \( z(t_k^1), \ldots, z(t_k^n) \), the payoff \( u_k \) does not necessarily equal \( U_i(s) \) as it did in the unperturbed chain. However, these payoff deviations vanish as \( \epsilon \to 0 \), and we may bound the probability that their magnitude exceeds \( \epsilon \) by using Chebyshev's inequality, as we show in the following lemma.

**Lemma 2:** Let \( s \in \mathcal{S} \) be the joint strategy played in period \( k \). Then

\[
\Pr \left[ u_i \not\in [U_i(s) - \epsilon, U_i(s) + \epsilon] \right] \leq \epsilon^{nc}.
\]

**Proof:** Since the signals \( z(t_k^1), z(t_k^2), \ldots, z(t_k^n) \) played in period \( k \) were chosen uniformly at random from \([0,1]\),

\[
\mathbb{E}[U_i(s(z(t_k^j))))] = \sum_{a \in \mathcal{A}} U_i(a)q^{a(s)} = U_i(s)
\]

for any \( \ell \in \{1,2,\ldots,n\} \). Moreover, because \( U_i(a) \in [0,1] \) for all \( a \in \mathcal{A} \), \( \Var[U_i(s(z(t_k^j))))] \leq 1 \). Then, using Chebyshev’s inequality and the fact that \( p = \left[ 1/ \epsilon^{n+2} \right] \),

\[
\Pr \left[ u_i - U_i(s) \geq \epsilon \right] \leq \frac{\Var[U_i(s(z(t))))]}{\epsilon^{2n}} \leq \epsilon^{nc}.
\]

### Recurrent classes

Recall \( D^0 \cup C^0 \) is precisely the set of recurrent classes of the unperturbed process. The set \( D^0 \) comprises a single recurrent class, since transitions between any two states in \( D^0 \) occur with positive probability in the unperturbed process. Each state \( y \in C^0 \) comprises a single recurrent class since, for any \( y \in C^0 \), the system remains in state \( y \) for all future time according to process \( P^0 \).

Define the set \( C^* \subseteq C^0 \) by:

\[
C^* = \{ y \in C^0 : U_i(s^0) \geq U_i(a_i, s^0_{-i}), \forall i \in N, a_i \in \mathcal{A}_i \}
\]

Here, \( (a_i, s^0_{-i}) = (s^b_i, \ldots, s^0_i, a_i, s^b_{i+1}, \ldots, s^b_n) \) is the joint strategy corresponding to a unilateral deviation by player \( i \in N \) to constant strategy \( a_i \). Thus, \( C^* \) is the set of states that induce CCE, i.e., \( y \in C^* \), then \( y \in Y^{CCE} \).

### Resistances between recurrent classes

Let \( s = (s^b_1, s^b_2, \ldots, s^b_n) \). For notational simplicity, define \( U^+_i := U_i(s) + \epsilon \) and \( U^-_i := U_i(s) - \epsilon \). In the following, we state a series of claims about the resistances between recurrence classes. We provide a proof for Claim 1 and omit proofs for the subsequent claims for brevity, since they follow in a similar manner.

**Claim 1:** For \( y \in C^0 \), with \( y_i = [s^b_i, s, C] \) for each \( i \in N \),

\[
r(D \to y) = \sum_{i \in N} (1 - U_i(s))
\]

**Proof:** Let \( u_i = \frac{1}{p} \sum_{j=1}^{p} U_i(z(t_k^j)) \) be the utility agent \( i \) received in the period immediately before the transition \( D \to y \). This transition occurs with probability

\[
\Pr[D \to y] = \sum_{i \in N} \int_0^1 \Pr(u_i = x)\epsilon^{1-x}dx.
\]

This is because, in order to transition from \( D \) to \( y \), each agent first receives utility \( u_i \), and then becomes content with probability \( \epsilon^{1-u_i} \). To show that the resistance of this transition is \( \sum_{i \in N} (1 - U_i(s)) \), we will show

\[
0 < \lim_{\epsilon \to 0} \frac{\Pr[D \to y]}{\epsilon} \leq \infty,
\]

satisfying (15). We begin by lower bounding \( \Pr[D \to y] \).

\[
\Pr[D \to y] = \prod_{i \in N} \int_0^1 \Pr(u_i = x)\epsilon^{1-x}dx
\]

\[
= \prod_{i \in N} \left( \int_0^{U^-_i} \Pr(u_i = x)\epsilon^{1-x}dx + \int_{U^-_i}^{U^+_i} \Pr(u_i = x)\epsilon^{1-x}dx \right)
\]

\[
\geq \prod_{i \in N} \int_{U^-_i}^{U^+_i} \Pr(u_i = x)\epsilon^{1-x}dx
\]

\[
\geq (a) \prod_{i \in N} \left( \int_{U^-_i}^{U^+_i} \Pr(u_i = x)dx \right)
\]
\[ \begin{align*}
&\geq \prod_{i \in N} e^{1-U^-_i} (1 - e^{nc}) \\
&= \prod_{i \in N} \left( e^{1-U^-_i} - e^{1-U^+_i + nc} \right) \\
&= e^{\sum_{i \in N} 1-U^-_i} + O(e^{nc})
\end{align*} \]

where (a) is from the fact that $e^{1-x}$ is continuous and increasing in $x$ for $\epsilon \in (0, 1)$, and (b) follows from Lemma 2. Then,

\[ \lim_{\epsilon \to 0} \frac{\Pr[D \to y]}{\epsilon^{\sum_{i \in N} 1-U^-_i}} \leq \lim_{\epsilon \to 0} \frac{e^{\sum_{i \in N} 1-U^+_i} + O(e^{nc})}{\epsilon^{\sum_{i \in N} 1-U^+_i}} = 1. \quad (17) \]

The desired result follows from (17) and (18).

**Claim 2:** For $y \in C^0 \setminus C^*$, $\tilde{y} \in C^0$, with $y_i = [\tilde{s}^b_i, s, C]$, and $\tilde{y}_i = [\tilde{s}^b_i, \tilde{s}, C]$ for each $i \in N$, where $s = (s^b_1, \ldots, s^b_n)$, $r(y \to \tilde{y}) \geq c + \sum_{i \in N: U_i(\tilde{y}) < U_i(y)} (1 - U_i(\tilde{y}))$.

**Claim 3:** For a sequence of transitions of the form $D \to y^0 \to y^1 \to \cdots \to y^m = y$, with $y^k \in C^0 \setminus C^*$ for each $k < m$, and $y \in C^0$, where $y_i = [\tilde{s}^b_i, s, C]$ for each $i \in N$, the resistance of this sequence of transitions satisfies $r(D \to y^0) + \sum_{k=1}^m r(y^k-1 \to y^k) \geq \sum_{i \in N} (1-U_i(s)) + mc$.

**Claim 4:** For a transition of the form $y \to \tilde{y}$, where $y \in C^*$ and $\tilde{y} \in C^*$, $r(y \to \tilde{y}) \geq 2c$.

**Claim 5:** For a transition of the form $y \to D$ where $y \in C^0 \setminus C^*$, $r(y \to D) \geq c$.

**Claim 6:** For a transition of the form $y \to D$ where $y \in C^0 \setminus C^*$, $r(y \to D) \geq 2c$.

**Claim 7:** For every $y \in C^0 \setminus C^*$, there exists a path $y = y^0 \to y^1 \to \cdots \to y^m \in C^* \cup D$ with resistance $r(y^j \to y^{j+1}) = c$, $\forall j \in \{0, 1, \ldots, m-1\}$.

**Stochastic potentials**

The following lemma specifies the stochastic potentials of each recurrent class. Using the resistances from Claims 1 to 7, these stochastic potentials follow from exactly the same arguments as in [16]; we omit the proof for brevity.

**Lemma 3:** The stochastic potentials for each recurrent class are:

\[ \begin{align*}
&\gamma(D) = c|C^0 \setminus C^*| + 2c|C^*| \\
&\gamma(y) = \left( |C^0 \setminus C^*| - 1 \right) c + 2c|C^*| + \sum_{i \in N} (1 - U_i(s)), \\
&\gamma(y^*) = |C^0 \setminus C^*| c + 2c(|C^*| - 1) + \sum_{i \in N} (1 - U_i(s)), \\
&y \in C^0 \setminus C^*
\end{align*} \]

where $s$ and $s^*$ are the baseline joint strategies corresponding to states $y$ and $y^*$ respectively.

We now use Lemma 3 to complete the proof of Theorem 1. For the first part, suppose $C^*$ is nonempty, and let $y \in \arg \max_y \sum U_i(s)$.

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

For $\tilde{y} \in C^0$, with $\tilde{y}_i = [\tilde{s}^b_i, \tilde{s}, C]$,

\[ \begin{align*}
\gamma(y) &= |C^0 \setminus C^*| c + 2c(|C^*| - 1) + \sum_{i \in N} (1 - U_i(s)) \\
&< |C^0 \setminus C^*| c + 2c(|C^*|) + \sum_{i \in N} (1 - U_i(\tilde{s})) \\
&= \gamma(\tilde{y}).
\end{align*} \]

For $\tilde{y} \in C^*$ with $\tilde{y}_i = [\tilde{s}^b_i, \tilde{s}, C]$ and $\tilde{y} \not\in \arg \max_y \sum U_i(s)$,

\[ \begin{align*}
\gamma(y) &= |C^0 \setminus C^*| c + 2c(|C^*| - 1) + \sum_{i \in N} (1 - U_i(s)) \\
&< |C^0 \setminus C^*| c + 2c(|C^*| - 1) + \sum_{i \in N} (1 - U_i(\tilde{s})) \\
&= \gamma(\tilde{y}).
\end{align*} \]

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

The second part of the theorem follows similarly by considering the case when $Y^{CCE} = \emptyset$. \(\blacksquare\)