Dyn Games Appl (2019) 9:24–46
https://doi.org/10.1007/s13235-018-0244-z

Learning to Play Efficient Coarse Correlated Equilibria

Holly P. Borowski[1] · **Jason R. Marden**[2] · **Jeff S. Shamma**[3]

Published online: 10 March 2018

Abstract The majority of the distributed learning literature focuses on convergence to Nash equilibria. Coarse 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 coarse correlated equilibria. In this paper, we provide one such algorithm, which guarantees that the agents' collective joint strategy will constitute an efficient coarse 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.

Keywords Game theory · Networked control · Multiagent systems · Distributed control

This research was supported by ONR grant #N00014-17-1-2060, NSF grant #ECCS-1638214, the NASA Aeronautics scholarship program, the Philanthropic Educational Organization, and the Zonta International Amelia Earhart fellowship program, and funding from King Abdullah University of Science and Technology (KAUST).

✉ Jason R. Marden
jrmarden@ece.ucsb.edu

Holly P. Borowski
hollyboro@gmail.com

Jeff S. Shamma
jeff.shamma@kaust.edu.sa

1 Numerica Corporation, 5042 Technology Parkway #100, Fort Collins, CO 80528, USA

2 Department of Electrical and Computer Engineering, University of California, Harold Frank Hall, Rm 5161, Santa Barbara, CA 93106, USA

3 Computer, Electrical and Mathematical Science and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955–6900, Saudi Arabia

1 Introduction

The theory of learning in games focuses on identifying how different types of equilibria can emerge through processes where individual agents independently adjust their behavior in response to locally available information. While the majority of this research focuses on (pure) Nash equilibria [8,11–14,31,35], there is also extensive research focusing on other solution concepts which often are more efficient than Nash equilibria. In particular, existing results focus on convergence to Pareto efficient Nash equilibria [26,32], potential function maximizers [1,7,24], welfare maximizing action profiles [4,27], and the set of (coarse) correlated equilibria [5,9,16,17,23], among others.

Most of the algorithms highlighted above guarantee (probabilistic) convergence to the specified equilibria. However, the convergence results associated with distributed learning for coarse correlated equilibria are significantly weaker than those for other equilibrium classes, merely guaranteeing convergence to the *set* of coarse correlated equilibria [16,17]. In other words, these convergence guarantees only ensure that the empirical frequency of play will approach the set of coarse correlated equilibria. This means that the long run behavior does not necessarily constitute—or even approximate—a specific coarse correlated equilibrium at any instance of time.

In this paper, we seek to identify whether specific forms of coarse correlated equilibria can emerge through distributed learning. Throughout, we focus our attention on the coarse correlated equilibrium that optimizes social welfare, i.e., maximizes the sum of the agents' expected payoffs. For concreteness, consider a mild variant of the Shapley game with the following payoff matrix

	L	M	R
T	1,-ε	-ε,1	0,0
M	0,0	1,-ε	-ε,1
B	-ε,1	0,0	1,-ε

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 approximately $1/3$ to each player. Alternatively, a joint distribution that places a mass of $1/6$ on each of the six joint actions that yield nonzero payoffs to the players yields an expected payoff of approximately $1/2$ to each player. Note that this distribution cannot be realized by independent strategies associated with the two players, but instead represents a specific coarse correlated equilibrium.

The central question that we address in this manuscript is whether there are distributed algorithms that lead the collective behavior to such an efficient coarse correlated equilibrium. In line with this theme, a recent result in [23] proposes a distributed algorithm that guarantees the empirical frequency of play will converge to an efficient coarse correlated equilibrium; however, convergence in empirical frequencies is attained through deterministic cyclic behavior. For example, in the above-modified Shapley game, the algorithm posed in [23] guarantees that the collective behavior of the agents could follow a deterministic cycle of the form $(T, L) \to (T, M) \to (M, M) \to (M, R) \to (B, R) \to (B, L) \to (T, L)$ with high probability. Following this deterministic cycle results in an empirical frequency of play that equates to the efficient coarse correlated equilibrium highlighted above; however, at any

time instance, the players are *not* playing a joint strategy in accordance with this efficient coarse correlated equilibrium.

In this paper, we present a novel distributed learning algorithm that ensures the agents collectively play a joint strategy corresponding to the efficient coarse correlated equilibrium with high probability. That is, the agents' collective choice at any time step will be drawn according to a joint distribution that is characterized by the efficient coarse correlated equilibrium. With regard to the Shapley game, our algorithm guarantees that the agents collectively play the highlighted joint distribution with high probability. The key element of our proposed algorithm that makes this correlation possible is the introduction of a common random signal to the agents, which is incorporated into their local decision-making rule. The distributed learning rule effectively leverages this common random signal to attain the desired level of correlation in the agents' strategies.

This result complements a recent thread in distributed learning focused on attaining desirable performance guarantees while minimizing the informational dependence of the learning rule. More specifically, recent research has focused on a class of learning rules termed completely uncoupled or payoff based [9,11,16–18], where agents make decisions based only on information pertaining to their received payoffs. In such settings, agents do not have knowledge of the payoff or behavior of other agents, nor do they have any information regarding the structural form of their utility functions. There are several notable results from this research thread including the derivation of completely uncoupled learning rules that converge to pure Nash equilibria in weakly acyclic games [26], Nash equilibria [11,13], Pareto efficient Nash equilibria [32], potential function maximizers [24], efficient action profiles [4,27], efficient (coarse) correlated equilibria in terms of empirical frequencies [23], among others. The contributions of this manuscript add to this literature, demonstrating that we can also attain efficient (coarse) correlated equilibria in terms of the agents' period-by-period joint strategies.

We conclude with two side notes. First, game theory has emerged as a valuable framework for the design of agents' local control laws in a distributed engineering system. In this setting, the agents are programmable components that are required to make local independent decisions in response to locally available information, c.f., [2,15,21,22,25,28]. When coupled with agent objective functions that are either designed or inherited, a game theoretic learning rule prescribes these control laws by dictating how each agent will revise its behavior based on its available information. Accordingly, significant research has focused on deriving distributed learning rules that possess desirable asymptotic performance guarantees while enabling the agents to make decisions based on limited information. Furthermore, recent research has also identified the importance of deriving distributed algorithms for attaining efficient coarse correlated equilibria, e.g., team versus team zero-sum games [19], peer-to-peer file sharing systems [33], and access control for wireless communications [3]. In particular, [3] demonstrates that the optimal system behavior associated with the problem of access control for wireless communications is in fact the efficient coarse correlated equilibrium studied in this paper.

Secondly, we want to highlight the difference between the approach presented in this manuscript and recent results on centralized algorithms for computing specific coarse correlated equilibria [20,29,30]. The prevailing question from this literature is on the feasibility of deriving centralized algorithms that can compute a given coarse correlated equilibrium for any game in a reasonable period of time. The applicability of such results to distributed engineering systems is limited due to the reliance on a complete characterization of the game which is often unavailable in the distributed systems of interest. However, a recent result in [6] demonstrates that computing any non-trivial coarse correlated equilibrium with welfare strictly better than the worst possible welfare is NP-hard. This result demonstrates that for

arbitrary game structures, the convergence rates associated with our proposed algorithm will be quite poor in a worst-case sense. However, it remains an open question as to whether specific game structures can be exploited to provide improved guarantees with regard to convergence rates.

2 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 $\mathcal{A}_i$ 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 $G = (N, \{U_i\}_{i\in N}, \{\mathcal{A}_i\}_{i\in N})$.

In this paper, we focus on the class of coarse correlated equilibria [5]. A coarse correlated equilibrium is characterized by 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 agent $i \in N$ and action $a_i' \in \mathcal{A}_i$,

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

where $a_{-i} = \{a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n\}$ 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 expected 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.,

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

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

This paper focuses on deriving a distributed learning algorithm that ensures the collective behavior of the 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 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 independently accordingly to the agent's strategy at time t, which we denote 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* [11], takes on the form:

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

Completely uncoupled learning rules 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 gauge the performance of a learning rule $\{F_i\}_{i\in N}$ by the resulting asymptotic guarantees. With that goal in mind, let $q(t) \in \Delta(\mathcal{A})$ represent the agents' collective strategy at time t, which is of the form

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

[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. The goal of this paper is to derive learning rules that guarantee the agents' collective strategy constitutes 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 because 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

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

As we show in the following section, this common signal can be used as a coordinating entity to reach collective strategies beyond the form given in (4).

3 A Learning Algorithm for Attaining Efficient Correlated Equilibria

In this section, we present a specific learning rule of the form (6) that guarantees 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*.

3.1 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

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

These strategies divide the unit interval into at most Ω 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 $s = (s_1, s_2, \ldots, s_n) \in S = \prod_{i \in N} S_i$, the resulting joint strategy $q(s) = \{q^a(s)\}_{a \in \mathcal{A}} \in \Delta(\mathcal{A})$ satisfies

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

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

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

3.2 Informal Algorithm Description

The forthcoming algorithm is reminiscent of the trial and error learning algorithm introduced in [35] and can be viewed at a high level through the following diagram (Fig. 1).

The times $\{1, 2, \ldots\}$ will be broken up into periods of length $3\bar{p}$ where $\bar{p} > 1$ is an interval whose length will be defined formally below. At the beginning of each period k, each agent $i \in N$ has a local state variable of the form $x_i(k) = [s_i^b, m_i]$ where $s_i^b \in S_i$ is the agent's baseline strategy and m_i is the agent's mood. The agent's baseline strategy corresponds to the strategy the agent is accustomed to playing. The agent's mood m_i, which can either be CONTENT or DISCONTENT, dictates how likely each agent is to select its baseline strategy during a given period. Roughly speaking, a content agent is more likely to select its baseline strategy, while a discontent agent is more likely to try an alternate strategy.

Each period $k > 0$, which consists of the time steps $\{3\bar{p}k + 1, \ldots, 3\bar{p}(k+1)\}$, will be broken up into three distinct phases called *evaluation*, *trial*, and *acceptance*. The behavior of the agents in each of these phases is highlighted below:

- *Evaluation Phase*: The first phase is the *evaluation phase*. In this phase, each agent establishes a baseline utility, u_i^b, associated with its current baseline strategy, s_i^b. All agents commit to their baseline strategies during this entire phase.
- *Trial Phase*: The second phase is the *trial phase*. During this phase, each agent has the opportunity to experiment with an alternate trial strategy, s_i^t, in order to determine whether changing its baseline strategy could be advantageous. An agent's mood determines how likely it is to experiment. In particular, a content agent will use its baseline strategy s_i^b during the trial phase with high probability. On the other hand, a discontent player is likely to experiment with a trial strategy $s_i^t \neq s_i^b$. The exact probabilities associated with this selection process will be described in detail in the forthcoming section.
- *Acceptance Phase*: The third phase is the *acceptance phase*. Here, an agent who experimented during the trial phase decides whether to accept its trial strategy or revert to its baseline strategy. Agents who did not experiment during the trial phase commit to their

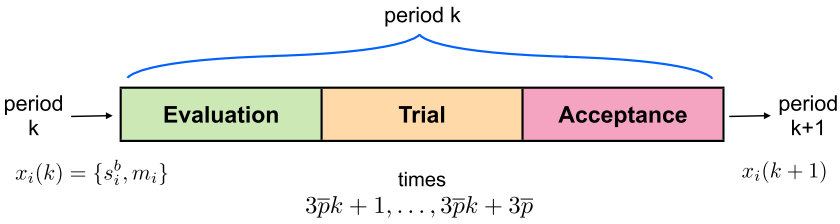

Fig. 1 This figure highlights the phases of the learning algorithm within each time period

baseline strategies and observe payoff changes which occur due to others' changes in strategy.

3.3 Formal Algorithm Description

We begin by defining a constant $c > n$, an experimentation rate $\varepsilon \in (0, 1)$, and the length of a phase to be $\bar{p} = \lceil 1/\delta^{nc+1} \rceil$ time steps, for some small $\delta \in (0, 1)$. 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 $k \in \{1, 2, \dots\}$. We will formally present the algorithm using the same general structure given in previous section.

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 $t \in \{3\bar{p}k+1, \dots, 3\bar{p}k+\bar{p}\}$. 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,

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

where $z(\tau)$ denotes the common random signal observed at time τ. 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 $t \in \{(3\bar{p}k + \bar{p}) + 1, \dots, 3\bar{p}k + 2\bar{p}\}$. 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

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

A strategy $s_i^t = a_i$ means that agent i commits to playing action a_i for the entire trial phase of the k-th period, i.e., the strategy does not depend on the common random signal. Observe 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,

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

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

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

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 $t \in \{(3\bar{p}k + 2\bar{p}) + 1, \ldots, 3\bar{p}k + 3\bar{p}\}$. 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:
$$s_i^a = \begin{cases} s_i^t & \text{if } u_i^t > u_i^b + \delta, \\ s_i^b & \text{if } u_i^t \leq u_i^b + \delta. \end{cases} \tag{12}$$
    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:
$$u_i^a = \frac{1}{\bar{p}} \sum_{\tau=(3\bar{p}k+2\bar{p})+1}^{3\bar{p}k+3\bar{p}} U_i\big(s_1^a(z(\tau), \ldots, s_n^a(z(\tau))\big). \tag{13}$$
    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., $x_i(k)$, the strategies s_i^b, s_i^t and s_i^a, and the payoffs received during the k-th period. The 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:

  - If $u_i^a \geq u_i^b - \delta$,
$$x_i(k+1) = \begin{cases} \left[s_i^a = s_i^b,\ C\right] & \text{w.p. } 1 - \varepsilon^{2c}, \\ \left[s_i^a = s_i^b,\ D\right] & \text{w.p. } \varepsilon^{2c}. \end{cases} \tag{14}$$
  - If $u_i^a < u_i^b - \delta$,
$$x_i(k+1) = \left[s_i^a = s_i^b,\ D\right] \tag{15}$$
    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
$$x_i(k+1) = \left[s_i^a, C\right]. \tag{16}$$
  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

$$x_i(k+1) = \begin{cases} [s_i^a, C] & \text{w.p. } \varepsilon^{1-u_i^a}, \\ [s_i^a, D] & \text{w.p. } 1-\varepsilon^{1-u_i^a}. \end{cases} \tag{17}$$

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

3.4 Main Result

Throughout this paper, we focus on games where there is some degree of coupling between the utility functions of the agents. The following definition of interdependence, taken from [35], captures this notion of coupling.

Definition 1 A game G with agents $N = \{1, 2, \ldots, n\}$ 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 $U_i(a'_J, a_{-J}) \neq U_i(a_J, a_{-J})$.

Roughly speaking, the definition of interdependence states that it is not possibly to partition the group of agents into two sets whose actions do not impact one another's payoffs.

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 $q(S) \cap \text{CCE} \neq \emptyset$. Given any probability $p < 1$, if the exploration rate ε is sufficiently small, and if $\delta = \varepsilon$, then for all sufficiently large times t,*[2]

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

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

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

We prove Theorem 1 in "Appendix B."

A few remarks are in 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 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. Alternately, the condition $q(S) \cap \text{CCE} = \emptyset$ 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., $\arg\max_{a \in \mathcal{A}} \sum_{i \in N} U_i(a)$.

[2] For the proof of Theorem 1, we require $\delta = \varepsilon$. However, in practice, fixing $\delta > \varepsilon$ in order to shorten the period length, $\bar{p}$, often yields similar results, as we demonstrate in Example 1.

3.5 Illustrative Example

Here, we present an example where agents update their strategies according to the algorithm above, and their actions converge to an efficient coarse correlated equilibrium.

Example 1 Consider a game with two players, (Row, Column), and the following payoff matrix:

	L	M	R
T	0, 0	0, 1	0.85, 0.75
M	1, 0	0, 0	0, 0
B	0.75, 0.85	0, 0	0, 0

The efficient coarse correlated equilibrium in this game places probability 0.5 on joint action (T, R), and probability 0.5 on joint action (B, L), i.e.,

$$q^{(T,R)} = q^{(B,L)} = 0.5, \tag{18}$$

and $q^a = 0$ for $a \notin \{(T, R), (B, L)\}$. The expected utility associated with this coarse correlated equilibrium is $U_i(q) = 0.8$.

For the values of ε shown below, we simulated our algorithm for 50 times over 10^6 iterations with parameters $\Omega = 2$ and $\delta = 0.14$. The table below shows the percentage of the last 1×10^4 iterations spent in the efficient coarse correlated equilibrium. Note that as ε decreases, more time is spent in the efficient coarse correlated equilibrium, as predicted by Theorem 1.

ϵ	% time in efficient CCE
0.2	17%
0.1	19%
0.08	24%
0.06	27%
0.04	44%
0.02	67%
0.01	94%

4 An Extension for Correlated Equilibrium

The above algorithm can also be adapted to allow for convergence to efficient correlated equilibria. A correlated equilibrium is characterized by a joint distribution $q = \{q^a\}_{a \in \mathcal{A}} \in \Delta(\mathcal{A})$, where for any agent $i \in N$ and actions $a_i', a_i'' \in \mathcal{A}_i$,

$$\sum_{a_{-i}} U_i(a_i', a_{-i}) q^{(a_i', a_{-i})} \geq \sum_{a_{-i}} U_i(a_i'', a_{-i}) q^{(a_i', a_{-i})}. \tag{19}$$

Informally, a 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 expected utility for deviating to any fixed action, *conditioned on the underlying signal*. It is straightforward to show that any correlated equilibrium is a coarse correlated equilibrium; however,

the converse is not true. We will call a correlated equilibrium q^* *efficient* if it maximizes the sum of the expected payoffs of the agents, i.e.,

$$q^* \in \arg\max_{q \in \mathrm{CE}} \sum_{i \in N} \sum_{a \in \mathcal{A}} U_i(a) q^a, \tag{20}$$

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

We will now state the revised learning rule that provides convergence to the efficient correlated equilibrium. The new algorithm is identical to the aforementioned algorithm for coarse correlated equilibria where the trial phase takes on the following form:

– *Trial Phase*: Given a baseline strategy s_i^b and two actions $a_i, a_i' \in \mathcal{A}_i$, define a new strategy $s_i^b(a_i \to a_i')$ where for any $z \in [0, 1]$

$$s_i^b(z|a_i \to a_i') = \begin{cases} s_i^b(z) & \text{if } s_i^b(z) \neq a_i, \\ a_i' & \text{if } s_i^b(z) = a_i. \end{cases}$$

Informally, the strategy $s_i^b(z|a_i \to a_i')$ takes the signals associated with a_i in strategy s_i^b and reassigns them to a_i', keeping everything else the same. Denote the set of admissible trial strategies as $S_i(s_i^b) = \{s_i^b(a_i \to a_i') : a_i, a_i' \in \mathcal{A}_i\}$. Note that for any baseline strategy s_i^b, $|S_i(s_i^b)| = |\mathcal{A}_i|^2$.

The agents' trial strategies are now 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

$$\Pr\left[s_i^t = s_i\right] = \begin{cases} 1 - \varepsilon^c & \text{if } s_i = s_i^b \\ \varepsilon^c \,/\, |\mathcal{A}_i|^2 & \text{for any } s_i = S_i(s_i^b) \end{cases} \tag{21}$$

- *Discontent*, $m_i = D$: When agent i is discontent, its trial strategy, s_i^t, is chosen randomly from the set S_i,

$$\Pr\left[s_i^t = s_i\right] = 1 \,/\, |S_i| \text{ for all } s_i \in S_i. \tag{22}$$

All remaining steps of the algorithm are identical.

The following theorem characterizes the limiting behavior associated with this new algorithm for correlated equilibrium.

Theorem 2 *Let $G = (N, \{U_i\}, \{\mathcal{A}_i\})$ be a finite interdependent game. First, suppose $q(S) \cap \mathrm{CE} \neq \emptyset$. Given any probability $p < 1$, if the exploration rate ε is sufficiently small, and if $\delta = \varepsilon$, then for all sufficiently large times t,*

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

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

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

While this new algorithm is slightly more complex than the previous algorithm for coarse correlated equilibrium, the general premise of the algorithm and the accompanying proof are virtually identical. Hence, we do not include the proof in this manuscript to avoid redundancy.

5 Conclusion

The majority of distributed learning literature focuses on identifying learning rules that converge to Nash equilibria. However, alternate forms of behavior, such as (coarse) 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 provide a distributed learning rule, similar in spirit to the learning rule in [23], that ensures agents collectively play a joint strategy that constitute an efficient coarse correlated equilibrium. A mild variant of the proposed algorithm also ensures the agents collectively play a joint strategy that constitute an efficient correlated equilibria. Future work seeks to investigate the applicability of such algorithms in the context of team versus team zero-sum games.

Appendix

The formulation of the decision-making process defined in Sect. 3 ensures that the evolution of the agents' states over the periods $\{0, 1, 2, \dots\}$ can be represented as a finite ergodic Markov chain over the state space

$$X = X_1 \times \cdots \times X_n \tag{23}$$

where $X_i = S_i \times \{C, D\}$ denotes the set of possible states of agent i. Let P^ε denote this Markov chain for some $\varepsilon > 0$, and $\delta = \varepsilon$. Proving Theorem 1 requires characterizing the stationary distribution of the family of Markov chains $\{P^\varepsilon\}_{\varepsilon>0}$ for all sufficiently small ε. We employ the theory of resistance trees for regular perturbed processes, introduced in [34], to accomplish this task. We begin by reviewing this theory and then proceed with the proof of Theorem 1.

A Background: Resistance Trees

Define P^0 as the transition matrix for some nominal Markov process, and let P^ε be a perturbed version of this nominal process where the size of the perturbation is $\varepsilon > 0$. Throughout this paper, 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^ε 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

$$0 < \lim_{\varepsilon\to 0} \frac{P^\varepsilon_{x\to x'}}{\varepsilon^{r(x\to x')}} < \infty. \tag{24}$$

The constant $r(x \to x')$ is referred to as the *resistance* of the 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^ε. The theory of resistance trees presented in [34] 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* [10] if $\lim_{\varepsilon \to 0^+} \mu_x^\varepsilon > 0$, where μ^ε is the stationary distribution corresponding to P^ε.

In this paper, we adopt the technique provided in [34] for identifying the stochastically stable states through a graph theoretic analysis over the recurrent classes of the unperturbed process P^0. To that end, let $Y_0, Y_1, \ldots, Y_m$ 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 $p = \{(x_1, x_2), (x_2, x_3), \ldots, (x_{k-1}, x_k)\}$ 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

$$r(Y_i, Y_j) = \min_{p \in \mathcal{P}_{ij}} \sum_{(x,x') \in p} r(x, x'). \tag{25}$$

The recurrent classes $Y_0, Y_1, \ldots, Y_m$ satisfy the following properties: (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 identify the resistance between the various recurrent classes. The second step focuses on analyzing spanning trees of the weighted, directed graph $\mathcal{G}$ whose vertices are recurrent classes of the process P^0, and whose edge weights are given by the resistances between classes in (25). Denote $\mathcal{T}_i$ to be the set of all spanning trees of $\mathcal{G}$ rooted at recurrent class Y_i. Next, we compute the stochastic potential of each recurrent class which is defined as follows:

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 3 ([34]) *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^ε be a regular perturbation of P^0 with a unique stationary distribution μ^ε. Then:*

(1) *As $\varepsilon \to 0$, μ^ε converges to a stationary distribution μ^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 main results associated with Theorem 1 (setting $\delta = \varepsilon$) using the terminology defined in 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.,

$$q(s) \in \underset{q \in q(S) \cap \text{CCE}}{\arg\max} \sum_{i \in N} \sum_{a \in \mathcal{A}} U_i(a) q^a. \tag{26}$$

- 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.,

$$q(s) \in \underset{q \in q(S)}{\arg\max} \sum_{i \in N} \sum_{a \in \mathcal{A}} U_i(a) q^a. \tag{27}$$

For convenience, and with an abuse of notation, define

$$U_i(s) := \sum_{a \in \mathcal{A}} U_i(a) q^a(s) \tag{28}$$

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:

(i) Define the unperturbed process, P^0.
(ii) Determine the recurrent classes of process P^0.
(iii) Establish transition probabilities of process P^ε.
(iv) Determine the stochastically stable states of P^ε using Theorem 3.

Part 1: Defining the unperturbed process

The unperturbed process P^0 is effectively the process identified in Sect. 3 where $\varepsilon = 0$. Rather than dictate the entire process as done previously, here we highlight the main attributes of the unperturbed process that may not be obvious upon initial inspection.

- 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 (22).
- The baseline utility u_i^b in (8) associated with joint baseline strategy s^b is now of the form

$$u_i^b = U_i(s^b). \tag{29}$$

  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 (29).

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.*

Proof: We begin by showing that any state $x \in C^0$ is a recurrent class of the unperturbed process. According to P^0, if the system reaches state x, then it remains at x with certainty for all future time. Hence, each $x \in C^0$ is a recurrent class of P^0. Next, we show that

D^0 constitutes a single recurrent class. Consider any two states $x, y \in D^0$. According to the unperturbed process, P^0, the probability of transitioning from x to y is strictly positive $\left(\geq \prod_{i\in N} 1/|S_i|\right)$; hence, the resistance of the transition $x \to y$ is 0. Further note that the probability of transitioning to any state not in D^0 is zero. Hence, D^0 forms a single recurrent class of P^0.

The last part of the proof involves proving that any state $x = \{[s_i^b, m_i]\}_{i\in N} \notin C^0 \cup D^0$ is not recurrent in P^0. Since $x \notin C^0 \cup D^0$, it consists of both content and discontent players. Denote the set of discontent players by $J = \{i \in N : m_i = D\} \neq \emptyset$. We will show that the discontent players J will play a sequence of strategies with positive probability that drives at least one content player to become discontent. Repeating this argument at most n times shows that any state x of the above form will eventually transition to the all discontent state, proving that x is not recurrent.

To that end, let $x(1) = x$ be the state at the beginning of the 1-st period. According to the unperturbed process P^0, each discontent player randomly selects a strategy $s_i \in S_i$ which becomes part of the player's state at the ensuing stage. Suppose each discontent agent selects a trial strategy $s_i = (a_i^1, \ldots, a_i^w) \in \mathcal{A}_i^w \subset S_i$ during the 1-st period, i.e., the discontent players select strategies of the finest granularization. Note that each agent selects a strategy with probability $\geq 1/|S_i|$. Here, the trial payoff for each player $i \in N$ associated with the joint strategies $s = (\{s_i^b\}_{i\notin J}, \{s_i\}_{i\in J})$ is

$$u_i^t(s) = \int_0^1 U_i(s(z))\mathrm{d}z \tag{30}$$

$$= \frac{1}{w}U_i(a) + \int_w^1 U_i(s'(z))\mathrm{d}z, \tag{31}$$

for some $a \in \mathcal{A}$ as $s_i(z) = s_i(z')$ for any $z, z' \in [0, 1/w]$ for any agent $i \in N$. If $u_i^t < u_i^b$ for any agent $i \notin J$, agent i becomes discontent in the next stage and we are done.

For the remainder of the proof suppose $u_i^t(s) \geq u_i^b(s^b)$ for all agents $i \notin J$. This implies all agents $N\backslash J$ will be content at the beginning of the second stage. By interdependence, there exists a collective action $\tilde{a}_J \in \prod_{j\in J} \mathcal{A}_j$ and an agent $i \notin J$ such that $U_i(a) \neq U_i(\tilde{a}_J, a_{N\backslash J})$. Suppose each discontent agent selects a trial strategy $s_i' = (\tilde{a}_i^1, a_i^2, \ldots, a_i^w) \in \mathcal{A}_i^w \subset S_i$ during the second period, i.e., only the first component of the strategy changed. The trial payoff for each player $i \in N$ associated with the joint strategies $s' = (\{s_i^b\}_{i\notin J}, \{s_i'\}_{i\in J})$ is

$$\begin{aligned} u_i^t(s') &= \int_0^1 U_i(s'(z))\mathrm{d}z \\ &= \frac{1}{w}U_i(\tilde{a}_J, a_{N\backslash J}) + \int_w^1 U_i(s'(z))\mathrm{d}z \\ &\neq u_i^t(s) \end{aligned}$$

If $u_i^t(s') < u_i^t(s)$, agent i will become discontent at the ensuing stage and we are done. Otherwise, agent i will stay content at the ensuing stage. However, if each discontent agent selects a trial strategy $s_i'' = (a_i^1, a_i^2, \ldots, a_i^w) \in \mathcal{A}_i^w \subset S_i$ during the third period, we know $u_i^t(s'') < u_i^t(s')$, where $s'' = (\{s_i^b\}_{i\notin J}, \{s_i''\}_{i\in J})$. Hence, agent i will become discontent at the beginning of period 4. This argument can be repeated at most n times, completing the proof. □

Part 3: Transition probabilities of process P^ε

Here, we establish the transition probability $P^\varepsilon_{x\to x^+}$ for a pair of arbitrary states, $x, x^+ \in X$. Let $x_i = [s_i, m_i]$, $x_i^+ = [s_i^+, m_i^+]$ for $i \in N$, $s = (s_1, s_2, \ldots, s_n)$, and $s^+ = (s_1^+, s_2^+, \ldots, s_n^+)$. Then,

$$P^\varepsilon_{x\to x^+} = \sum_{\tilde{s}^t\in S}\sum_{\tilde{s}^a\in S}\Big(\Pr[x^+ \mid s^t = \tilde{s}^t,\ s^a = \tilde{s}^a] \times \Pr[s^a = \tilde{s}^a \mid s^t = \tilde{s}^t]\Pr[s^t = \tilde{s}^t]\Big). \tag{32}$$

Note that the strategy selections and state transitions are conditioned on state x; for notational brevity we do not explicitly write this dependence. Here, s^t and s^a represent the joint trial and acceptance strategies during the period before the transition to x^+. The double summation in (32) is over all possible trial actions, $\tilde{s}^t \in S$, and acceptance strategies, $\tilde{s}^a \in S$. However, recall from (14) to (17) that, when transitioning from x to x^+, not all strategies can serve as intermediate trial and acceptance strategies. In particular, transitioning to state x^+ requires that $s^a = s^+$; hence if $\tilde{s}^a \neq s^+$, then $\Pr[x^+ \mid s^t = \tilde{s}^t,\ s^a = \tilde{s}^a] = 0$, so we can rewrite (32) as:

$$P^\varepsilon_{x\to x^+} = \sum_{\tilde{s}^t\in S}\Big(\Pr[x^+ \mid s^t = \tilde{s}^t,\ s^a = s^+] \times \Pr[s^a = s^+ \mid s^t = \tilde{s}^t]\Pr[s^t = \tilde{s}^t]\Big) \tag{33}$$

There are three cases for the transition probabilities in (33). Before proceeding, we make the following observations. The last term in (33), $\Pr[s^t = \tilde{s}^t]$, is defined in Sect. 3; we will not repeat the definition here. For the first two terms, agents' state transition and strategy selection probabilities are independent when conditioned state x and on the joint trial and acceptance strategy selections. Hence, we can write the first term as:

$$\Pr[x^+ \mid s^t = \tilde{s}^t, s^a = s^+] = \prod_{i\in N}\Pr[x_i^+ \mid s^t = \tilde{s}^t, s^a = s^+] \tag{34}$$

and the second term as:

$$\Pr[s^a = s^+ \mid s^t = \tilde{s}^t] = \prod_{i\in N}\Pr[s_i^a = s_i^+ \mid s^t = \tilde{s}^t]. \tag{35}$$

The following three cases specify individual agents' probability of choosing the acceptance strategy s_i^a in (35) and transitioning to state x_i^+ in (34).

Case (i) agent i is content in state x, i.e., $m_i = C$, and did not experiment, $s_i^t = s_i$:

For (35), since $s_i^a \in \{s_i^t, s_i\}$ we know that

$$\Pr[s_i^a = s_i^+ \mid s^t = \tilde{s}^t] = \begin{cases} 1 & \text{if } s_i^+ = s_i \\ 0 & \text{otherwise} \end{cases}.$$

In (34), for any trial strategy $s^t = \tilde{s}^t$, the probability of transitioning to a state x_i^+ depends on realized average payoffs u_i^b and u_i^a. In particular, if $x_i^+ = [s_i^+, C]$, then we must have that $u_i^a \geq u_i^b - \varepsilon$, so

$$\Pr\left[x_i^+ = [s_i^+, C] \mid s^a = s^+, s^t = \tilde{s}^t\right] = \int_0^1 \Pr[u_i^b = \eta] \int_{\eta-\varepsilon}^1 \Pr[u_i^a = \nu \mid s^t = \tilde{s}^t, s^a = s^+] \mathrm{d}\nu \mathrm{d}\eta.$$

Then, the probability that $x_i^+ = [s_i^+, D]$ is

$$1 - \Pr\left[x_i^+ = [s_i^+, C] \mid s^a = s^+, s^t = \tilde{s}^t\right].$$

Case (ii) agent i is content and experimented, $s_i^t \neq s_i$: For (35), agent i's acceptance strategy depends on its average baseline and trial payoffs, u_i^b and u_i^t. Recall, if $u_i^t \geq u_i^b + \varepsilon$, then $s_i^a = s_i$, i.e., agent i's acceptance strategy is simply its baseline strategy from state x. Otherwise $s_i^a = s_i^t$. Utilities u_i^b and u_i^t depend on joint strategies s and s^t and on the common random signals sent during the corresponding phases. Therefore,

$$\Pr[s_i^a = s_i^+ \mid s^t = \tilde{s}^t \neq s] = \int_0^1 \int_0^1 \Pr[s_i^a = s_i^+ \mid u_i^b = \eta, u_i^t = \nu, s_i^t = s_i] \times \Pr[u_i^b = \eta] \Pr[u_i^t = \nu \mid s^t = \tilde{s}^t] \mathrm{d}\eta \mathrm{d}\nu$$

In (34), since agent i remains content and sticks with its acceptance strategy from the previous period,

$$\Pr[x_i^+ \mid s^a = s^+, s^t = \tilde{s}^t] = \begin{cases} 1 & \text{if } s_i^+ = s_i^a \\ 0 & \text{otherwise} \end{cases}.$$

Case (iii) agent i is discontent:
For (35),

$$\Pr[s_i^a = s_i^+ \mid s^t = \tilde{s}^t] = \begin{cases} 1 & \text{if } s_i^+ = s_i^t \\ 0 & \text{otherwise} \end{cases}.$$

In (34), agent i's probability of becoming content depends only on its received payoff during the acceptance phase; it becomes content with probability $\varepsilon^{1-u_i^a}$ and remains discontent with probability $1 - \varepsilon^{1-u_i^a}$. Hence, if $x_i^+ = [s_i^+, C]$,

$$\Pr\left[x_i^+ = [s_i^+, C] \mid s^a = s^+, s^t = \tilde{s}^t\right] = \int_0^1 \varepsilon^{1-\eta} \Pr[u_i^a = \eta \mid s^a = s^+, s^t = \tilde{s}^t] \mathrm{d}\eta.$$

Then,

$$\Pr\left[x_i^+ = [s_i^+, D] \mid s^a = s^+, s^t = \tilde{s}^t\right] = 1 - \Pr\left[x_i^+ = [s_i^+, C] \mid s^a = s^+, s^t = \tilde{s}^t\right]$$

Now that we have established transition probabilities for process P^ε, we may state the following lemma.

Lemma 2 *The process P^ε is a regular perturbation of P^0.*

It is straightforward to see that P^ε satisfies the first two conditions of Definition 2 with respect to P^0. The fact that transition probabilities satisfy the third condition, Eq. (24), follows from the fact that the dominant terms in $P^\varepsilon_{x\to y}$ are polynomial in ε. This is immediately clear in all but the incorporation of realized utilities into the transition probabilities, as in (33). However, for any joint strategy, s, and associated average payoff u_i, since

$$\mathbb{E}[u_i] = \mathbb{E}\left[\frac{1}{\bar{p}}\sum_{\tau=\ell}^{\ell+\bar{p}-1} U_i(s(z(\tau)))\right] = U_i(s).$$

for any time period of length $\bar{p}$ in which joint strategy s is played throughout the entire period. Moreover, $\mathrm{Var}\big[U_i(s(z(\tau)))\big] \leq 1$. Therefore, we may use Chebyshev's inequality and the fact that $\bar{p} = \lceil 1 / \varepsilon^{nc+2} \rceil$ to see that

$$\Pr\Big[\big|u_i - U_i(s)\big| \geq \varepsilon\Big] \leq \frac{\mathrm{Var}\big[U_i(s(z(\tau)))\big]}{\bar{p}\varepsilon^2} \leq \varepsilon^{nc}. \tag{36}$$

Note that this applies for all average utilities, u_i^b, u_i^t, and u_i^a in the aforementioned state transition probabilities.

Part 4: Determining the stochastically stable states

We begin by defining

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

Here, we show that, if $C^\star$ is non-empty, then a state x is stochastically stable if and only if $q(s)$ satisfies (26). The fact that $q(s)$ must satisfy (27) 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

We summarize resistances between recurrent classes in the following claim.

Claim: 1 Resistances between recurrent classes satisfy:
For $x \in C^0$ with corresponding joint strategy s,

$$r(D^0 \to x) = \sum_{i\in N}(1 - U_i(s)). \tag{37}$$

For a transition of the form $x \to y$, where $x \in C^\star$ and $y \in (C^0 \cup D^0)\backslash\{x\}$,

$$r(x \to y) \geq 2c. \tag{38}$$

For a transition of the form $x \to y$ where $x \in C^0\backslash C^\star$ and $y \in (C^0 \cup D^0)\backslash\{x\}$,

$$r(x \to y) \geq c. \tag{39}$$

For every $x \in C^0\backslash C^\star$, there exists a path $x = x^0 \to x^1 \to \cdots \to x^m \in C^\star \cup D^0$ with resistance

$$r(x^j \to x^{j+1}) = c,\ \forall j \in \{0, 1, \ldots, m-1\}. \tag{40}$$

These resistances are computed in a similar manner to the proof establishing resistances in [23]; however, care must be taken due to the fact that there is a small probability that average received utilities fall outside of the window $U_i(s) \pm \varepsilon$ during a phase in which joint strategy s is played. We illustrate this by proving (37) in detail; the proofs are omitted for other types of transitions for brevity.

Proof: Let $x \in D^0$, $x^+ \in C^0$ with $x_i = [s_i, D]$ and $x_i^+ = [s_i^+, C]$ for each $i \in N$. Again, for notational brevity, we drop the dependence on state x in the following probabilities. Note that all agents must select $s^t = s_i^+$ in order to transition to state $x_i = [s_i^+, C]$; otherwise the transition probability is 0. we have

$$\begin{aligned} P^{\varepsilon}_{x\to x^+} &\overset{(a)}{=} \Pr[x^+ \mid s^a = s^+, s^t = s^+] \\ &\quad \times \Pr[s^a = s^+ \mid s^t = s^+] \Pr[s^t = s^+] \\ &\overset{(b)}{=} \Pr[x^+ \mid s^a = s^+, s^t = s^+] \Pr[s^t = s^+] \\ &\overset{(c)}{=} \Pr[x^+ \mid s^a = s^+, s^t = s^+] \prod_{i\in N} 1 / |S_i| \\ &= \prod_{i\in N} \frac{1}{|S_i|} \Pr[x_i^+ \mid s^a = s^+, s^t = s^+] \end{aligned}$$

where: (a) follows from the fact that $s_i^a = s_i^t$ since $m_i = D$ in state x for all $i \in N$, (b) $\Pr[s^a = s^+ \mid s^t = s^+] = 1$ since all agents are discontent and hence commit to their trial strategies during the acceptance period, and (c) $\Pr[s^t = s^+] = \prod_{i\in N} 1 / |S_i|$ since each discontent agent selects its trial strategy uniformly at random from S_i.

We now show that

$$0 < \lim_{\varepsilon\to 0^+} \frac{P^{\varepsilon}_{x\to x^+}}{\varepsilon^{\sum_{i\in N} 1-U_i(s^+)}} < \infty \tag{41}$$

satisfying (24). For notational simplicity, we define

$$\begin{aligned} U_i^+ &:= U_i(s^+) + \varepsilon, \\ U_i^- &:= U_i(s^+) - \varepsilon. \end{aligned} \tag{42}$$

We first lower bound $P^{\varepsilon}_{x\to x^+}$:

$$\begin{aligned} P^{\varepsilon}_{x\to x^+} &= \prod_{i\in N} \frac{1}{|S_i|} \Pr[x_i^+ \mid s^a = s^+, s^t = s^+] \\ &= \prod_{i\in N} \frac{1}{|S_i|} \int_0^1 \Pr[u_i^a = \eta \mid s^a = s^+, s^t = s^+] \varepsilon^{1-\eta} \mathrm{d}\eta \\ &\geq \prod_{i\in N} \frac{1}{|S_i|} \int_{U_i^-}^{U_i^+} \Pr[u_i^a = \eta \mid s^a = s^+, s^t = s^+] \varepsilon^{1-\eta} \mathrm{d}\eta \\ &\overset{(a)}{\geq} \prod_{i\in N} \frac{\varepsilon^{1-U_i^-}}{|S_i|} \int_{U_i^-}^{U_i^+} \Pr[u_i^a = \eta \mid s^a = s^+, s^t = s^+] \mathrm{d}\eta \\ &\overset{(b)}{\geq} \prod_{i\in N} \frac{\varepsilon^{1-U_i^-}}{|S_i|} (1 - \varepsilon^{nc}) \end{aligned}$$

$$= \frac{\varepsilon^{\sum_{i\in N} 1-U_i^-} + O(\varepsilon^{nc})}{\prod_{i\in N} |S_i|} \tag{43}$$

where (a) is from the fact that $\varepsilon^{1-\eta}$ is continuous and increasing in η for $\varepsilon \in (0, 1)$, and (b) follows from (36). Continuing in a similar fashion, it is straightforward to show

$$P^{\varepsilon}_{x\to x^+} \leq \varepsilon^{\sum_{i\in N}(1-U_i^+)} + O(\varepsilon^{nc}). \tag{44}$$

Given (43) and (44), and the fact that U_i^+ and U_i^- satisfy (42), we have that $P^{\varepsilon}_{x\to x^+}$ satisfies (24) with resistance $\sum_{i\in N}\left(1 - U_i(s^+)\right)$ as desired. □

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 [23]. The proof is repeated below for completeness.

Lemma 3 *Let $x \in C^0\backslash C^\star$ 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:*

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

Proof: In order to establish the stochastic potentials for each recurrent class, we will lower and upper bound them.

Lower bounding the stochastic potentials: To lower bound the stochastic potentials of each recurrent class, we determine the lowest possible resistance that a tree rooted at each of these classes may have.

(1) Lower bounding $\gamma(D^0)$:

$$\gamma(D^0) \geq c|C^0\backslash C^\star| + 2c|C^\star|$$

In a tree rooted at D^0, each state in C^0 must have an exiting edge. In order to exit a state in $C^0\backslash C^\star$, only a single agent must experiment, contributing resistance c. To exit a state in $C^\star$, at least two agents must experiment, contributing resistance $2c$.

(2) Lower bounding $\gamma(x)$, $x \in C^0\backslash C^\star$:

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

Here, each state in $C^0\backslash\{x\}$ must have an exiting edge, which contributes resistance $\left(|C^0\backslash C^\star| - 1\right)c + 2c|C^\star|$. The recurrent class D^0 must also have an exiting edge, contributing at least resistance $\sum_{i\in N}(1 - U_i(s))$.

(3) Lower bounding $\gamma(x^\star)$, $x^\star \in C^\star$:

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

Again, each state in $C^0 \backslash \{x^\star\}$ must have an exiting edge, which contributes resistance $\left(|C^0 \backslash C^\star| - 1\right) c + 2c|C^\star|$. The recurrent class D^0 must also have an exiting edge, contributing resistance at least $\sum_{i \in N}(1 - U_i(s^\star))$.
Upper bounding the stochastic potentials: In order to upper bound the stochastic potentials, we construct trees rooted at each recurrent class which have precisely the resistances established above.

(1) Upper bounding $\gamma(D^0)$:

$$\gamma(D^0) \leq c|C^0 \backslash C^\star| + 2c|C^\star|$$

Begin with an empty graph with vertices X. For each state $x \in C^0 \backslash C^\star$, add a path ending in $C^\star \cup D^0$ so that each edge has resistance c. This is possible due to Claim 1. Now eliminate redundant edges; this contributes resistance at most $c|C^0 \backslash C^\star|$ since each state in $C^0 \backslash C^\star$ has exactly one outgoing edge. Finally, add an edge $x^\star \to D^0$ for each $x^\star \in C^0$; this contributes resistance $2c|C^\star|$.

(2) Upper bounding $\gamma(x)$, $x \in C^0 \backslash C^\star$:

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

This follows by a similar argument to the previous upper bound, except here we add an edge $D^0 \to x$ which contributes resistance $\sum_{i \in N}(1 - U_i(s))$.

(3) Upper bounding $\gamma(x^\star)$, $x^\star \in C^\star$:

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

This follows from an identical argument to the previous bound. □

We now use Lemma 3 to complete the proof of Theorem 1. For the first part, suppose $C^\star$ is non-empty, and let

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

where joint strategy s corresponds to state x. Then,

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

For $x \in C^0$,

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

For $x \in C^\star$ with

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

Applying Theorem 3, $x^\star$ 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 $C^\star = \emptyset$. □

References

1. Alos-Ferrer C, Netzer N (2010) The logit-response dynamics. Games Econ Behav 68:413–427
2. Alpcan T, Basar T (2010) Network security: a decision and game-theoretic approach, 1st edn. Cambridge University Press, Cambridge
3. Altman E, Bonneau N, Debbah M (2006) Correlated equilibrium in access control for wireless communications. In: 5th international conference on networking
4. Arieli I, Babichenko Y (2012) Average testing and the efficient boundary. J Econ Theory 147:2376–2398
5. Aumann R (1987) Correlated equilibrium as an expression of bayesian rationality. Econometrica 55(1):1–18
6. Barman S, Ligett K (2015) Finding any nontrivial coarse correlated equilibrium is hard. SIGecom Exch 14:76–79
7. Blume L (1993) The statistical mechanics of strategic interaction. Games Econ Behav 5:387–424
8. Boussaton O, Cohen J (2012) On the distributed learning of Nash equilibria with minimal information. In: 6th international conference on network games, control, and optimization
9. Foster DP, Vohra R (1997) Calibrated learning and correlated equilibrium. Games Econ Behav 21:40–55
10. Foster DP, Young HP (1990) Stochastic evolutionary game dynamics. Theor Popul Biol 38:219–232
11. Foster DP, Young HP (2006) Regret testing: learning to play Nash equilibrium without knowing you have an opponent. Theor Econ 1:341–367
12. Frihauf P, Krstic M, Basar T (2012) Nash equilibrium seeking in noncooperative games. IEEE Trans Autom Control 57(5):1192–1207
13. Germano F, Lugosi G (2007) Global Nash convergence of Foster and Young's regret testing. Games Econ Behav 60:135–154
14. Gharesifard B, Cortes J (2012) Distributed convergence to Nash equilibria by adversarial networks with directed topologies. In: 51st IEEE conference on decision and control
15. Han Z, Niyato D, Saad W, Baar T, Hjrungnes A (2012) Game theory in wireless and communication networks: theory, models, and applications, 1st edn. Cambridge University Press, Cambridge
16. Hart S (2005) Adaptive heuristics. Econometrica 73(5):1401–1430
17. Hart S, Mas-Colell A (2000) A simple adaptive procedure leading to correlated equilibrium. Econometrica 68:1127–1150
18. Hart S, Mas-Colell A (2003) Uncoupled dynamics do not lead to nash equilibrium. Am Econ Rev 93(5):1830–1836
19. Ho YC, Sun FK (1974) Value of information in two-team zero-sum problems. J Optim Theory Appl 14(5):557–571
20. Jiang A, Leyton-Brown K (2011) Polynomial-time computation of exact correlated equilibrium in compact games. In: Proceedings of the 12th ACM electronic commerce conference (ACM-EC)

21. Lasauce S, Tembine H (2011) Game theory and learning for wireless networks, 1st edn. Elsevier, Amsterdam
22. MacKenzie A, DaSilva L (2006) Game theory for wireless engineers, 1st edn. Morgan & Claypool Publishers, San Rafael
23. Marden JR (2017) Selecting efficient correlated equilibria through distributed learning. Games Econ Behav 106:114–133
24. Marden JR, Shamma JS (2012) Revisiting log-linear learning: asynchrony, completeness and payoff-based implementation. Games Econ Behav 75(2):788–808
25. Marden JR, Shamma JS (2014) Game theory and distributed control. In: Young HP, Zamir S (eds) Handbook of game theory, vol 4. Elsevier, Amsterdam
26. Marden JR, Young HP, Arslan G, Shamma JS (2009) Payoff based dynamics for multi-player weakly acyclic games. SIAM J Control Optim 48(1):373–396
27. Marden JR, Young HP, Pao LY (2014) Achieving Pareto optimality through distributed learning. SIAM J Control Optim 52:2753–2770
28. Menache I, Ozdaglar A (2011) Network games: theory, models, and dynamics, 1st edn. Morgan & Claypool Publishers, San Rafael
29. Papadimitriou C (2005) Computing correlated equilibria in multiplayer games. In: Proceedings of the annual ACM symposium on theory of computing
30. Papadimitriou C, Roughgarden T (2008) Computing correlated equilibria in multi-player games. J ACM 55:1–29
31. Poveda J, Quijano N (2013) Distributed extremum seeking for real-time resource allocation. In: American control conference
32. Pradelski B, Young HP (2012) Learning efficient Nash equilibria in distributed systems. Games Econ Behav 75:882–897
33. Wang B, Han Z, Liu K (2009) Peer-to-peer file sharing game using correlated equilibrium. In: 43rd annual conference on information sciences and systems, 2009. CISS 2009, pp 729–734
34. Young HP (1993) The evolution of conventions. Econometrica 61(1):57–84
35. Young HP (2009) Learning by trial and error. Games Econ Behav 65:626–643