

# Learning Correlated Equilibria in Noncooperative Games with Cluster Structure

Omid Namvar Gharehshiran and Vikram Krishnamurthy

University of British Columbia  
Department of Electrical and Computer Engineering  
2332 Main Mall, Vancouver, BC V6T 1Z4, Canada  
{omidn,vikramk}@ece.ubc.ca

**Abstract.** We consider learning correlated equilibria in noncooperative repeated games where players form clusters. In each cluster, players observe the action profile of cluster members and receive local payoffs, associated to performing localized tasks within clusters. Players also acquire global payoffs due to global interaction with players outside cluster, however, are oblivious to actions of those players. A novel adaptive learning algorithm is presented which generates trajectories of empirical frequency of joint plays that converge almost surely to the set of correlated  $\varepsilon$ -equilibria. Thus, sophisticated rational global behavior is achieved by individual player's simple local behavior.

**Keywords:** Adaptive learning, correlated equilibrium, differential inclusions, stochastic approximation.

## 1 Introduction

Consider a noncooperative repeated game with a set of players comprising multiple non-overlapping clusters. Clusters are characterized by the subset of players that perform the same task locally and share information of their actions with each other. However, clusters do not disclose their action profile to other clusters. In fact, players inside clusters are even oblivious to the existence of other clusters or players. Players repeatedly take actions to which two payoffs are associated: *i) local payoffs*: due to performing localized tasks within clusters, *ii) global payoffs*: due to global interaction with players outside clusters. The incremental information that players acquire at the end of each period then comprises: i) the realized payoff, delivered by a third party (e.g. network controller in sensor networks), and ii) observation of action profile of cluster members. Players then utilize this information and continuously update their strategies – via the proposed regret-based learning algorithm – to maximize their expected payoff. The question we tackle in this paper is: Given this simple local behavior of individual agents, can the clustered network of players achieve sophisticated global behavior? Similar problem have been studied in the Economics literature. For seminal works, the reader is referred to [1,2,3].

**Main Results:** Regret-matching as a strategy of play in long-run interactions has been introduced in [1,2]. In [1] the authors prove that when all players share action information and follow the proposed regret-based learning procedure, under general conditions, the global behavior converges to the set of correlated equilibrium. A milder assumption is that players only observe the outcome, namely, stage payoffs. A regret-based reinforcement learning algorithm is proposed in [2] whereby players build statistics of their past experience and infer how their payoff would have improved based on the history of realized payoffs. Our model differs from the above works as it incorporates cluster structure where action information is *only* locally shared. The main result of this paper is that if every player follows the proposed adaptive regret-based learning algorithm, the global behavior of the network converges to the set of correlated  $\varepsilon$ -equilibria [4]. The presented learning procedure can be simply regarded as a non-linear adaptive filtering algorithm. In addition, we show (via empirical numerical studies) that, taking advantage of the excess information disclosed within clusters, an order of magnitude faster convergence to the set of correlated  $\varepsilon$ -equilibria can be achieved as compared to the regret-based reinforcement learning algorithm in [2].

Correlated equilibrium is a generalization of Nash equilibrium and describes a condition of competitive optimality. It is, however, more preferable for on-line adaptive learning in distributed systems with tight computation/energy constraints (e.g. wireless sensor networks [5,6]) due to the following reasons: i) Structural Simplicity: it is a convex polytope, whereas the Nash equilibria are isolated points at the extrema of this set [7], ii) Computational Simplicity: computing correlated equilibrium requires solving a linear program (that can be solved in polynomial time), whereas computing Nash equilibrium necessitates finding fixed points. iii) Coordination Capability: it directly takes into account the ability of players to coordinate their actions. Indeed, Hart and Mas-Colell observe in [2] that for most simple adaptive procedures, "...there is a natural coordination device: the common history, observed by all players. It is thus reasonable to expect that, at the end, independence among players will not obtain." This coordination leads to potentially higher payoffs than if players take their actions independently as required by Nash equilibrium [4].

**Context:** The motivation for such formulation stems from multi-agent networks that require some sort of cluster structure such as intruder monitoring in sensor networks. Consider a multiple-target localization scenario in an unattended ground sensor network [5,6]. Depending on their locations, sensors form clusters each responsible for localizing a particular target. Sensors receive two payoffs: i) local payoffs, based on the importance and accuracy of the information provided about the local phenomena, ii) global payoffs, for communicating the collected data to the sink through the communication channel, which is globally shared amongst all sensors. Consideration of the potential local interaction among sensors leads to a more realistic modeling, hence, more sophisticated design of reconfigurable networked sensors.

## 2 Regret-Based Learning with Cluster Structure

### 2.1 Game Model

Consider the *finite* repeated strategic-form noncooperative game:

$$\mathbf{G} = \left( \mathcal{K}, (C_m)_{m \in \mathcal{M}}, (\mathcal{A}^k)_{k \in \mathcal{K}}, (U^k)_{k \in \mathcal{K}}, (\boldsymbol{\sigma}^k)_{k \in \mathcal{K}} \right), \quad (1)$$

where each component is described as follows:

- 1) *Set of Players*:  $\mathcal{K} = \{1, 2, \dots, K\}$ . Individual players are denoted by  $k \in \mathcal{K}$ .
- 2) *Local Clusters*  $C_m$ : Set  $\mathcal{K}$  is partitioned into  $M$  non-overlapping clusters  $C_m \subset \mathcal{K}$ ,  $m \in \mathcal{M} = \{1, \dots, M\}$ . We make the *cluster monitoring* assumption:  $k, k' \in C_m$  if and only if  $k$  knows  $a_n^{k'}$  and  $k'$  knows  $a_n^k$  at the end of period  $n$ . Note that isolated players, which do not belong to any cluster, are formulated as singleton clusters.
- 3) *Action Set*:  $\mathcal{A}^k = \{1, 2, \dots, A^k\}$  denotes the set of action indices for each player  $k$ , where  $|\mathcal{A}^k| = A^k$ .
- 4) *Payoff Function*:  $U^k : \mathcal{A}^{\mathcal{K}} \rightarrow \mathbb{R}$  denotes the payoff function for each player  $k$ . Here,  $\mathcal{A}^{\mathcal{K}} = \times_{k \in \mathcal{K}} \mathcal{A}^k$  represents the set of  $K$ -tuple of *action profiles*. A generic element of  $\mathcal{A}^{\mathcal{K}}$  is denoted by  $\mathbf{a} = (a^k, \dots, a^K)$  and can be rearranged as  $(a^k, \mathbf{a}^{-k})$  for any player  $k$ , where  $\mathbf{a}^{-k} \in \times_{\substack{k' \in \mathcal{K} \\ k' \neq k}} \mathcal{A}^{k'}$ .

The payoff for each player  $k \in \mathcal{K}$  is formulated as:

$$U^k(a^k, \mathbf{a}^{-k}) = U_l^k(a^k, \mathbf{a}^{C_m}) + U_g^k(a^k, \mathbf{a}^{-C_m}). \quad (2)$$

Here,  $\mathbf{a}^{C_m} \in \times_{\substack{k' \in C_m \\ k' \neq k}} \mathcal{A}^{k'}$  and  $\mathbf{a}^{-C_m} \in \times_{\substack{k' \in \mathcal{K} \\ k' \notin C_m}} \mathcal{A}^{k'}$  denote the joint action profile of cluster  $C_m$  (to which player  $k$  belongs) excluding player  $k$  and the joint action profile of all players excluding cluster  $C_m$ , respectively. In addition,  $U_l^k(a^k, \mathbf{a}^{C_m^{-k}}) = 0$  if cluster  $C_m$  is singleton.

Time is discrete  $n = 1, 2, \dots$ . Each player  $k$  takes an action  $a_n^k$  at time instant  $n$  and receives a payoff  $U_n^k(a_n^k)$ . Each player is assumed to know its local payoff function  $U_l^k(\cdot)$ ; Hence, taking action  $a_n^k$  and knowing  $\mathbf{a}_n^{C_m}$ , is capable of evaluating its stage local payoff. Players do *not* know the global payoff function  $U_g^k(\cdot)$ . However, they can compute their *realized* global payoffs as follows:

$$U_{g,n}^k(a_n^k) = U_n^k(a_n^k) - U_l^k(a_n^k, \mathbf{a}_n^{C_m}). \quad (3)$$

Note that, even if players knew  $U_g^k(\cdot)$ , they could not compute stage global payoffs as they are unaware of the actions taken by players outside cluster, namely,  $\mathbf{a}_n^{-C_m}$ .

5) *Strategy*  $\boldsymbol{\sigma}^k$ : At period  $n$ , each player  $k$  selects actions according to a randomized strategy  $\boldsymbol{\sigma}^k \in \Delta \mathcal{A}^k = \{\mathbf{p}^k \in \mathbb{R}^{A^k}; p^k(a) \geq 0, \sum_{a \in \mathcal{A}^k} p^k(a) = 1\}$ . The learning algorithm is an adaptive procedure whereby obtaining relatively

high payoff by a given action  $i$  at period  $n$  increases the probability of choosing that action  $\sigma_{n+1}^k(i)$  in the following period.

## 2.2 Learning Correlated $\varepsilon$ -equilibria

The game  $\mathbf{G}$ , defined in (1), is played repeatedly in discrete time  $n = 1, 2, \dots$ . Each player  $k$  generates two average regret matrices and update their elements over time: (i)  $\bar{\alpha}_{A^k \times A^k}^k$ , which records *average local-regrets*, and (ii)  $\bar{\beta}_{A^k \times A^k}^k$ , which is an unbiased estimator of the *average global-regrets*. Each element  $\bar{\alpha}_n^k(i, j)$ ,  $i, j \in \mathcal{A}^k$ , gives the time-average regret, in terms of gains and losses in local payoff values, had the player selected action  $j$  every time he played action  $i$  in the past. However, players are not capable of computing their global payoffs and only receive the realized values. Each element  $\bar{\beta}_n^k(i, j)$ ,  $i, j \in \mathcal{A}^k$ , thus provides an unbiased *estimate* (based on the realized global payoffs) of the average regrets for replacing action  $j$  every time  $i$  was played in the past.

Positive overall-regrets (sum of local- and global-regrets) imply the opportunity to gain higher payoffs by switching action. Therefore, agents take only positive regrets  $|\bar{\alpha}_n^k(i, j) + \bar{\beta}_n^k(i, j)|^+$  into account to determine switching probabilities  $\sigma_n^k$ . Here,  $|x|^+ = \max\{0, x\}$ . The more positive the regret for not choosing an action, the higher is the probability that the player picks that action. At each period, with probability  $1 - \delta$ , player  $k$  chooses its consecutive action according to  $|\bar{\alpha}_n^k(i, j) + \bar{\beta}_n^k(i, j)|^+$ . With the remaining probability  $\delta$ , player  $k$  randomizes amongst the actions  $\mathcal{A}^k$  according to a uniform distribution. This can be interpreted as “exploration” which is essential as players continuously learn their global payoff functions. Exploration forces all actions to be chosen with a minimum frequency, hence, rules out actions being rarely chosen.

The adaptive regret-based learning algorithm can then be summarized as follows:

**Algorithm 1:** *Adaptive Regret-based Learning with Partial Local Information*

0) **Initialization:** Set  $0 < \delta < 1$ . Initialize  $\psi_0^k(i) = 1/A^k$ , for all  $i \in \mathcal{A}^k$ ,  $\bar{\alpha}_0^k = \mathbf{0}_{A^k \times A^k}$  and  $\bar{\beta}_0^k = \mathbf{0}_{A^k \times A^k}$ .

For  $n = 1, 2, \dots$  repeat the following steps:

1) **Strategy Update and Action Selection:** Select action  $a_n^k = j$  according to the following distribution

$$\sigma_n^k = (1 - \delta) \boldsymbol{\mu}_n^k + \frac{\delta}{A^k} \cdot \mathbf{1}_{A^k}, \quad (4)$$

where  $\mathbf{1}_{A^k} = [1, 1, \dots, 1]_{A^k \times 1}$  and  $\boldsymbol{\mu}_n^k$  denotes an invariant measure for the following transition probabilities:

$$\psi_n^k(i) = \begin{cases} \frac{1}{\xi^k} |\bar{\alpha}_{n-1}^k(a_{n-1}^k, i) + \bar{\beta}_{n-1}^k(a_{n-1}^k, i)|^+, & i \neq a_{n-1}^k, \\ 1 - \sum_{\substack{j \in \mathcal{A}^k \\ j \neq i}} \psi_n^k(j), & i = a_{n-1}^k. \end{cases} \quad (5)$$

Here,  $\xi^k$  is chosen such that  $\xi^k > \sum_{j \in \mathcal{A}^k - \{a_{n-1}^k\}} \psi_n^k(j)$ .

2) **Local Information Exchange:** Player  $k$ : i) broadcasts  $a_n^k$  to the cluster members, ii) receives actions of cluster members and forms the profile  $\mathbf{a}_n^{C_m}$ .

3) **Regret Update:**

3.1: Local Regret Update

$$\bar{\alpha}_n^k(i, j) = \bar{\alpha}_{n-1}^k(i, j) + \epsilon_n \left[ \left( U_l^k(j, \mathbf{a}_n^{C_m}) - U_l^k(a_n^k, \mathbf{a}_n^{C_m}) \right) \mathbb{I}\{a_n^k = i\} - \bar{\alpha}_{n-1}^k(i, j) \right]. \quad (6)$$

3.2: Global Regret Update

$$\bar{\beta}_n^k(i, j) = \bar{\beta}_{n-1}^k(i, j) + \epsilon_n \left[ \frac{\sigma_n^k(i)}{\sigma_n^k(j)} U_{g,n}^k(a_n^k) \mathbb{I}\{a_n^k = j\} - U_{g,n}^k(a_n^k) \mathbb{I}\{a_n^k = i\} - \bar{\beta}_{n-1}^k(i, j) \right]. \quad (7)$$

Here,  $\mathbb{I}\{\cdot\}$  denotes the indicator function and the step-size is selected as  $\epsilon_n = 1/(n+1)$  (in static games) or  $\epsilon_n = \bar{\epsilon}$ ,  $0 < \bar{\epsilon} \ll 1$ , (in slowly time-varying games).

4) **Recursion:** Set  $n \leftarrow n + 1$  and go to Step 1.

*Remark 1.* The game model may evolve with time due to: i) players joining/leaving the game, ii) players appending/shrinking the set of choices, iii) changes in players' incentives, and iv) changes in cluster membership agreements. In these cases, to keep players responsive to the changes, a constant step-size  $\epsilon_n = \bar{\epsilon}$  is required in (6) and (7). Algorithm 1 cannot respond to multiple successive changes in the game as players' strategies are functions of the time-averaged regrets.

## 3 Global Behavior and Convergence Analysis

### 3.1 Global Behavior and Correlated $\epsilon$ -equilibrium

Consider game  $\mathbf{G}$ , defined in (1), and suppose each player employs Algorithm 1 to pick action for the next period. The global behavior, denoted by  $\bar{\mathbf{z}}_n$ , is defined as the (discounted) *empirical frequency of joint play* of all players. Formally,

$$\bar{\mathbf{z}}_n = \begin{cases} \frac{1}{n} \sum_{\tau \leq n} \mathbf{e}_{\mathbf{a}_\tau}, & \text{if } \epsilon_n = \frac{1}{n}, \\ \bar{\epsilon} \sum_{\tau \leq n} (1 - \bar{\epsilon})^{n-\tau} \mathbf{e}_{\mathbf{a}_\tau}, & \text{if } \epsilon_n = \bar{\epsilon}, \end{cases} \quad (8)$$

where  $\mathbf{e}_{\mathbf{a}_\tau}$  denotes the  $(\prod_{k \in \mathcal{K}} A^k)$ -dimensional unit vector with the element corresponding to  $\mathbf{a}_\tau$  being equal to one. The second line in (8) is a *discounted* version of the first line and will be used in slowly evolving games. Note that  $\bar{\mathbf{z}}_n$  is only used for the global convergence analysis of Algorithm 1 – it does not need to be computed by the players. However, in multi-agent systems such as sensor networks, a network controller can monitor  $\bar{\mathbf{z}}_n$  and use it to adjust sensors' parameters, thereby changing the equilibrium set in novel ways.

Before proceeding with the main theorem of this paper, we provide the definition of the correlated  $\varepsilon$ -equilibrium  $\mathcal{C}_\varepsilon$ .

**Definition 1.** Let  $\boldsymbol{\pi}$  denote a joint distribution on  $\mathcal{A}^\mathcal{K}$ , where  $\pi(\mathbf{a}) \geq 0$  for all  $\mathbf{a} \in \mathcal{A}^\mathcal{K}$  and  $\sum_{\mathbf{a} \in \mathcal{A}^\mathcal{K}} \pi(\mathbf{a}) = 1$ . The set of correlated  $\varepsilon$ -equilibrium, denoted by  $\mathcal{C}_\varepsilon$ , is the convex set [4]

$$\mathcal{C}_\varepsilon = \left\{ \boldsymbol{\pi} : \sum_{\mathbf{a}^{-k}} \pi^k(i, \mathbf{a}^{-k}) \left[ U^k(j, \mathbf{a}^{-k}) - U^k(i, \mathbf{a}^{-k}) \right] \leq \varepsilon, \forall i, j \in \mathcal{A}^k, \forall k \in \mathcal{K} \right\}. \quad (9)$$

For  $\varepsilon = 0$  in (9),  $\mathcal{C}_0$  is called the set of correlated equilibria.

In (9),  $\pi^k(i, \mathbf{a}^{-k})$  denotes the probability of player  $k$  choosing action  $i$  and the rest playing  $\mathbf{a}^{-k}$ . Definition 1 simply states that when the recommended signal  $\mathbf{a}$ , chosen according to the distribution  $\boldsymbol{\pi}$ , allocates positive probability to playing action  $i$  by player  $k$ , choosing  $j \in \mathcal{A}^k - \{i\}$  (instead of  $i$ ) does not lead to a higher expected payoff.

### 3.2 Convergence to Correlated $\varepsilon$ -equilibrium

The following theorem states the main result of this paper:

**Theorem 1.** Suppose each player  $k \in \mathcal{K}$  employs the learning procedure in Algorithm 1. Then, for each  $\varepsilon > 0$ , there exists  $\hat{\delta}(\varepsilon)$  such that if  $\delta < \hat{\delta}(\varepsilon)$  in Algorithm 1, the global behavior  $\bar{\mathbf{z}}_n$  converges almost surely (for  $\epsilon_n = 1/n$ ) to the set of correlated  $\varepsilon$ -equilibria in the following sense:

$$\bar{\mathbf{z}}_n \xrightarrow{\text{a.s.}} \mathcal{C}_\varepsilon \text{ as } n \rightarrow \infty \quad \text{iff} \quad d(\bar{\mathbf{z}}_n, \mathcal{C}_\varepsilon) = \inf_{\mathbf{z} \in \mathcal{C}_\varepsilon} |\bar{\mathbf{z}}_n - \mathbf{z}| \xrightarrow{\text{a.s.}} 0 \text{ as } n \rightarrow \infty. \quad (10)$$

For constant step-size  $\epsilon_n = \bar{\varepsilon}$ ,  $\bar{\mathbf{z}}_n$  weakly tracks  $\mathcal{C}_\varepsilon$ .

The above theorem implies that, for constant step-size  $\epsilon_n = 1/n$ , the stochastic process  $\bar{\mathbf{z}}_n$  enters and stays in the correlated  $\varepsilon$ -equilibrium set  $\mathcal{C}_\varepsilon$  forever with probability one. In other words, for any  $\varepsilon > 0$ , there exists  $N(\varepsilon) > 0$  with probability one such that for  $n > N(\varepsilon)$ , one can find a correlated equilibrium  $\boldsymbol{\pi} \in \mathcal{C}_0$  at the most  $\varepsilon$ -distance of  $\bar{\mathbf{z}}_n$ . In addition, if the game evolves with time slowly enough, Algorithm 1 can properly track the time-varying set of correlated  $\varepsilon$ -equilibria.

*Remark 2.* If one replaces  $\delta$  in Algorithm 1 with  $\delta_n$ , such that  $\delta_n \rightarrow 0$  slowly enough as  $n \rightarrow \infty$ , convergence to the set of correlated equilibria  $\mathcal{C}_0$  (instead of  $\varepsilon$ -equilibria  $\mathcal{C}_\varepsilon$ ) can be achieved in static games. This result cannot be expanded to the games slowly evolving with time.

*Proof.* The proof uses concepts in stochastic averaging theory [8] and Lyapunov stability of differential inclusions [9]. In what follows, a sketch of the proof will be presented:

1) *Asymptotics of the Discrete-time Dynamics*: Trajectories of the piecewise constant continuous-time interpolation of the stochastic processes  $\bar{\alpha}_n^k$  and  $\bar{\beta}_n^k$  converges almost surely to (for  $\epsilon_n = 1/n$ ), as  $n \rightarrow \infty$ , or weakly tracks (for  $\epsilon_n = \bar{\epsilon}$ ), as  $\bar{\epsilon} \rightarrow 0$ , trajectories  $\bar{\alpha}^k(t)$  and  $\bar{\beta}^k(t)$  evolving according to the system of inter-connected differential inclusion-equation:

$$\begin{cases} \frac{d\bar{\alpha}^k}{dt} \in \mathcal{L}^k(\bar{\alpha}^k, \bar{\beta}^k) - \bar{\alpha}^k, \\ \frac{d\bar{\beta}^k}{dt} = \mathcal{G}^k(\bar{\alpha}^k, \bar{\beta}^k) - \bar{\beta}^k, \end{cases} \quad (11)$$

where elements of the set-valued matrix  $\mathcal{L}^k(\bar{\alpha}^k, \bar{\beta}^k)$  and matrix  $\mathcal{G}^k(\bar{\alpha}^k, \bar{\beta}^k)$  are given by:

$$\mathcal{L}_{ij}^k(\bar{\alpha}^k, \bar{\beta}^k) = \left\{ [U_l^k(j, \nu^{C_m}) - U_l^k(i, \nu^{C_m})] \sigma^k(i); \nu^{C_m} \in \Delta \mathcal{A}^{C_m - \{k\}} \right\}, \quad (12)$$

$$\mathcal{G}_{ij}^k(\bar{\alpha}^k, \bar{\beta}^k) = [U_{g,t}^k(j) - U_{g,t}^k(i)] \sigma^k(i), \quad (13)$$

for some bounded measurable process  $U_{g,t}^k(\cdot)$ . Here,

$$U_l^k(a^k, \nu^{C_m}) = \int_{\mathcal{A}^{C_m - \{k\}}} U_l^k(a^k, \mathbf{a}^{C_m}) d\nu^{C_m}(\mathbf{a}^{C_m}), \quad (14)$$

In addition,  $\Delta \mathcal{A}^{C_m - \{k\}}$  denotes the simplex of probability measures over  $\mathcal{A}^{C_m - \{k\}}$ . The proof for the case of slowly time-varying game includes mean square error bounds and weak convergence analysis.

Furthermore, if (11) is Lyapunov stable, trajectories of the continuous-time interpolation of the stochastic processes  $\bar{\alpha}_n^k$  and  $\bar{\beta}_n^k$  converges almost surely to (for  $\epsilon_n = 1/n$ ), as  $n \rightarrow \infty$ , or weakly tracks (for  $\epsilon_n = \bar{\epsilon}$ ), as  $\bar{\epsilon} \rightarrow 0$ , the set of global attractors of (11).

2) The coupled system of differential inclusion-equation (11) is Lyapunov stable and the set of global attractors is characterized by  $|\bar{\alpha}^k(i, j) + \bar{\beta}^k(i, j)|^+$  being confined *within* an  $\varepsilon$ -distance of  $\mathbb{R}^-$ , for all  $i, j \in \mathcal{A}^k$ . Formally, for almost every solution to (11),

$$\lim_{t \rightarrow \infty} |\bar{\alpha}_t^k(i, j) + \bar{\beta}_t^k(i, j)|^+ \leq \varepsilon, \quad \forall i, j \in \mathcal{A}. \quad (15)$$

This, together with step 1, proves that if player  $k$  employs the learning procedure in Algorithm 1,  $\forall \varepsilon \geq 0$ , there exists  $\hat{\delta}(\varepsilon) \geq 0$  such that if  $\delta \leq \hat{\delta}(\varepsilon)$  in Algorithm 1:

**Table 1.** Agents' Payoff Matrix

		$2 : x_1$		$2 : x_2$	
Local: $(U_l^1, U_l^2)$	$1 : x_1$	(3, 5)	(2, 3)		
	$1 : x_2$	(3, 3)	(5, 4)		

		$2 : x_1$		$2 : x_2$		$2 : x_1$		$2 : x_2$	
Global: $(U_g^1, U_g^2, U_g^3)$	$1 : x_1$	(-1, 3, 1)	(2, -1, 3)	(1, -1, 3)	(0, 3, 1)				
	$1 : x_2$	(1, -1, 3)	(1, 4, 1)	(3, 3, 1)	(-1, 0, 3)				
		$3 : y_1$			$3 : y_2$				

$$\limsup_{n \rightarrow \infty} \left| \bar{\alpha}_n^k(i, j) + \bar{\beta}_n^k(i, j) \right|^+ \leq \varepsilon \quad \text{w.p. 1, } \forall i, j \in \mathcal{A}^k. \tag{16}$$

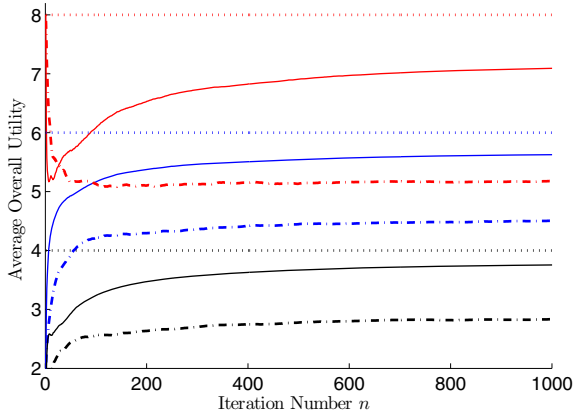
3) The global behavior  $\bar{\mathbf{z}}_n$  converges to  $\mathcal{C}_\varepsilon$  if and only if (16) holds for all players  $k \in \mathcal{K}$ . Thus, if every player  $k$  follows Algorithm 1,  $\bar{\mathbf{z}}_n$  converges almost surely (in static games) or weakly tracks (in slowly evolving games) the set of correlated  $\varepsilon$ -equilibrium  $\mathcal{C}_\varepsilon$ . □

## 4 Numerical Example

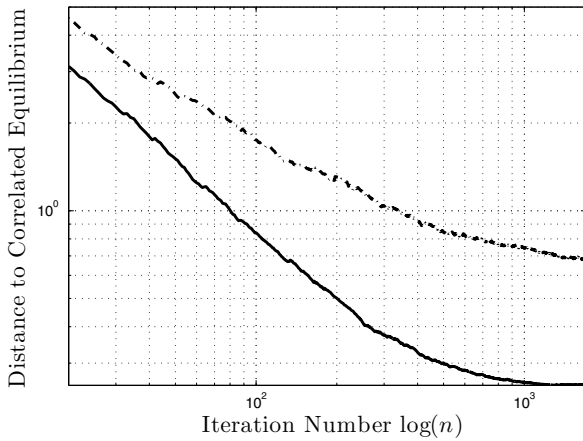
In this section we study a small hypothetical multi-agent network comprising three agents  $\mathcal{K} = \{1, 2, 3\}$ . Agents 1 and 2 are allocated the same task, hence, form the cluster  $\mathcal{C} = \{1, 2\}$  and share action information. Agent 3 forms a singleton cluster, hence, neither observes the action profile of  $\mathcal{C}$ , nor discloses its action to agents 1 and 2. Agents 1 and 2 repeatedly take action from the same action set  $\mathcal{A}^1 = \mathcal{A}^2 = \{x_1, x_2\}$ . Agent 3, due to performing a different task, chooses from a different action set  $\mathcal{A}^3 = \{y_1, y_2\}$ . Table 1 gives the payoffs in normal form. The set of correlated equilibrium is singleton (a pure strategy), where probability one is assigned to  $\mathbf{a}^* = (x_2, x_2, y_1)$  and zero to others.

In numerical studies, we set  $\epsilon_n = 1/n$  and  $\delta = 0.1$ . Figure 1 illustrates the behavior of Algorithm 1 and compares its performance to the reinforcement learning algorithm proposed in [2]. The sample paths shown in Fig. 1 are averaged over 50 independent runs of the algorithms starting with the same initial conditions  $\mathbf{a}_1 = (x_1, x_1, y_1)$ . Note that Theorem 1 proves convergence to the set of correlated  $\varepsilon$ -equilibrium. Therefore, although the average utilities increases with the number of iterations in Fig. 1(a), it only reaches an  $\varepsilon$ -distance of the values achievable in correlated equilibrium depending on the choice of exploration parameter  $\delta$  in Algorithm 1. Comparing the slopes of the lines in Fig. 1(b),  $m_1 = -0.182$  (for regret-based reinforcement learning [2]) and  $m_2 = -0.346$  (for Algorithm 1) numerically verifies that exploiting local action information results in an order of magnitude faster convergence to the set of correlated  $\varepsilon$ -equilibria.





(a) Average overall utility



(b) Distance to correlated equilibrium

**Fig. 1.** Performance Comparison: The solid and dashed lines represent the results from Algorithm 1 and the reinforcement learning algorithm in [2], respectively. In (a), the blue, red and black lines illustrate the sample paths of average payoffs of agents 1, 2 and 3, respectively. The dotted lines also represent the payoffs achievable in the correlated equilibrium.

## 5 Conclusion

We considered noncooperative repeated games with cluster structure and presented a simple regret-based adaptive learning algorithm that ensured convergence of global behavior to the set of correlated  $\varepsilon$ -equilibria. Noting that reaching correlated equilibrium can be conceived as consensus formation in actions amongst players, the proposed learning algorithm could have significant

applications in frameworks where coordination is sought among “players” in a distributed fashion, e.g. smart sensor systems and cognitive radio. It was numerically verified that utilizing the excess information shared/observed within clusters could lead to an order of magnitude faster convergence results.

## References

1. Hart, S., Mas-Colell, A.: A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68, 1127–1150 (2000)
2. Hart, S., Mas-Colell, A.: A reinforcement procedure leading to correlated equilibrium. In: *Economic Essays: A Festschrift for Werner Hildenbrand*, pp. 181–200 (2001)
3. Hart, S., Mas-Colell, A.: A general class of adaptive strategies. *Journal of Economic Theory* 98, 26–54 (2001)
4. Aumann, R.J.: Correlated equilibrium as an expression of Bayesian rationality. *Econometrica: Journal of the Econometric Society* 55, 1–18 (1987)
5. Krishnamurthy, V., Maskery, M., Yin, G.: Decentralized adaptive filtering algorithms for sensor activation in an unattended ground sensor network. *IEEE Transactions on Signal Processing* 56, 6086–6101 (2008)
6. Gharehshiran, O.N., Krishnamurthy, V.: Coalition formation for bearings-only localization in sensor networks – a cooperative game approach 58, 4322–4338 (2010)
7. Nau, R., Canovas, S.G., Hansen, P.: On the geometry of nash equilibria and correlated equilibria. *International Journal of Game Theory* 32, 443–453 (2004)
8. Kushner, H.J., Yin, G.: *Stochastic Approximation Algorithms and Applications*, 2nd edn. Springer, New York (2003)
9. Benaïm, M., Hofbauer, J., Sorin, S.: Stochastic approximations and differential inclusions; Part II: Applications. *Mathematics of Operations Research* 31, 673–695 (2006)