Regret Based Dynamics: Convergence in Weakly Acyclic Games

Jason R. Marden *
Department of Mechanical and Aerospace Engineering
University of California
Los Angeles, CA 90095
marden@ucla.edu

Gürdal Arslan
Department of Electrical Engineering
University of Hawaii, Manoa
Honolulu, HI 96822
gurdal@hawaii.edu

Jeff S. Shamma
Department of Mechanical and Aerospace Engineering
University of California
Los Angeles, CA 90095
shamma@ucla.edu

ABSTRACT

Regret based algorithms have been proposed to control a wide variety of multi-agent systems. The appeal of regret-based algorithms is that (1) these algorithms are easily implementable in large scale multi-agent systems and (2) there are existing results proving that the behavior will asymptotically converge to a set of points of "no-regret" in any game. We illustrate, through a simple example, that no-regret points need not reflect desirable operating conditions for a multi-agent system. Multi-agent systems often exhibit an additional structure (i.e. being "weakly acyclic") that has not been exploited in the context of regret based algorithms. In this paper, we introduce a modification of regret based algorithms by (1) exponentially discounting the memory and (2) bringing in a notion of inertia in players' decision process. We show how these modifications can lead to an entire class of regret based algorithm that provide *almost sure* convergence to a pure Nash equilibrium in any weakly acyclic game.

Categories and Subject Descriptors

[**Cooperative distributed problem solving:**]: Multiagent learning; Emergent behavior; Coordination, cooperation, and teamwork.

1. INTRODUCTION

The applicability of regret-based algorithms for multi-agent learning has been studied in several papers [6, 3, 13, 1, 7]. The appeal of regret-based algorithms is two fold. First of all, regret-based algorithms are easily implementable in large scale multi-agent systems when compared with other learning algorithms such as fictitious play [16, 12]. Secondly, there is a wide range of algorithms, called "no-regret" algorithms, that guarantee that the collective behavior will asymptotically converge to a set of points of no-regret (also referred to as correlated or coarse correlated equilibrium) in any game. A point of no-regret characterizes a situation for which the average utility that a player actually received is as high as the average utility that the player "would have" received had that player used a different fixed strategy at all previous time steps. No-regret algorithms have been proposed in a variety of settings ranging from network routing problems [2] to structured prediction problems [6].

In regret-based algorithms, each player makes a decision using *only* information regarding the regret for each of his possible actions. If the regret-based algorithm guarantees that a player's maximum regret asymptotically approaches zero then the algorithm is referred to as a no-regret algorithm. The most common no-regret algorithm is regret matching [8]. In regret matching, at each time step, each player plays a strategy where the probability of playing an action is proportional to the positive part of his regret for that action. In a multi-agent system, if all players adhere to a no-regret learning algorithm, such as regret matching, then the group behavior will converge to a set of points of no-regret asymptotically. In general, one may view a point of no-regret as a desirable or efficient operating condition because each player's average utility is as good as the average utility that any other action would have yielded [13]. However, a point of no-regret says little about the performance; hence knowing that the collective behavior of a multi-agent system will converge to a set of points of no-regret in general does not guarantee an efficient operation.

There have been attempts to further strengthen the convergence results of no-regret algorithms for special classes of games. For example, in [12], Jafari et al. showed that no-regret algorithms provide convergence to a Nash equilibrium in dominance solvable, constant-sum, and general sum 2×2 games. In [3], Bowling introduced a gradient based regret algorithm that guarantees that players' strategies converge to a Nash equilibrium in any 2 player 2 action repeated game. In [2], Blum et al. analyzed the convergence of no-regret algorithms in routing games and proved that behavior will approach a Nash equilibrium in various settings. However, the classes of games considered here can not fully model a

*Jason R. Marden is a Ph.D. student in the Department of Mechanical and Aerospace Engineering, University of California, Los Angeles.

wide variety of multi-agent systems.

It turns out that weakly acyclic games, which is a generalized form of potential games, are closely related to multi-agent systems [15]. The connection can be seen by recognizing that in any multi-agent system there is a global objective that may or may not be known to the players. In the case that the players are aware of the global objective and furthermore the player's utility is set as the global utility then we have an identical interest game. In the more general case, each player is assigned a local utility function that is appropriately aligned with the global objective. It is precisely this alignment that connects the realms of multi-agent systems and weakly acyclic games.

It remains as an open question as to whether no-regret algorithms converge to a Nash equilibrium in n-player weakly acyclic games. In this paper, we introduce a modification of regret-based algorithms that (1) exponentially discounts the memory and (2) brings in a notion of inertia in players' decision process. We show how these modifications can lead to an entire class of regret based algorithm that provide almost sure convergence to a pure Nash equilibrium in any weakly acyclic game. It is important to note that convergence to a Nash equilibrium also implies convergence to a no-regret point.

In Section 2 we review the game theoretic setting. In Section 3 we discuss a no-regret algorithm, regret matching, and illustrate the performance issues involved with no-regret points in a simple 3 player identical interest game. In Section 4 we introduce a new class of learning dynamics referred to as regret based dynamics with fading memory and inertia. In Section 5 we present some simulation results. Section 6 presents some concluding remarks.

Notation

— For a finite set A, $|A|$ denotes the number of elements in A.

— $I\{\cdot\}$ denotes the indicator function, i.e., $I\{S\} = 1$ if the statement S is true; otherwise, it is zero.

— $\mathcal{R}^n$ denotes the n dimensional Euclidian space.

— $\Delta(n)$ denotes the simplex in $\mathcal{R}^n$, i.e., the set of n dimensional probability distributions.

— For $x \in \mathcal{R}^n$, $[x]^+ \in \mathcal{R}^n$ denotes the vector whose $i-$th entry equals $\max(x_i, 0)$.

2. SETUP

2.1 Finite Strategic-Form Games

A finite strategic-form game [5] consists of an n-player set $\mathcal{P} := \{\mathcal{P}_1, ..., \mathcal{P}_n\}$, a finite action set Y_i for each player $\mathcal{P}_i \in \mathcal{P}$, and a utility function $U_i : Y \to \mathcal{R}$ for each player $\mathcal{P}_i \in \mathcal{P}$, where $Y := Y_1 \times ... \times Y_n$. We will henceforth use the term "game" to refer to such a finite strategic-form game.

In a one-stage version of a game, each player $\mathcal{P}_i \in \mathcal{P}$ simultaneously chooses an action $y_i \in Y_i$, and as a result receives a utility $U_i(y)$ depending on the action profile $y := (y_1, \ldots, y_n)$. Players are assumed to be selfish, i.e., each player $\mathcal{P}_i$ is interested in maximizing its own utility $U_i(y)$ by choosing its own action $y_i \in Y_i$.

In such a setting, players are to negotiate an action profile $y^* \in Y$ that is player by player optimal, i.e., no player has an incentive to unilaterally deviate from y^*. This leads us to the well-known notion of Nash equilibrium. Before introducing the notion of Nash equilibrium in more precise terms, we will introduce some notation. Let y_{-i} denote the collection of the actions of players *other than* player $\mathcal{P}_i$, i.e.,

$$y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n),$$

and let $Y_{-i} := Y_1 \times \ldots Y_{i-1} \times Y_{i+1} \ldots Y_n$. With this notation, we will sometimes write an assignment profile y as (y_i, y_{-i}). Similarly, we may write $U_i(y)$ as $U_i(y_i, y_{-i})$. Using the above notation, an action profile y^* is called a *pure Nash equilibrium* if, for all players $\mathcal{P}_i \in \mathcal{P}$,

$$U_i(y_i^*, y_{-i}^*) = \max_{y_i \in Y_i} U_i(y_i, y_{-i}^*). \quad (1)$$

Furthermore, if the above condition is satisfied with a unique maximizer for every player $\mathcal{P}_i \in \mathcal{P}$, then y^* is called a strict (Nash) equilibrium.

In general, a pure Nash equilibrium may not exist for an arbitrary game. However, we are interested in engineered multi-agent systems where agent objectives are designed to achieve an overall objective [21, 20]. In such systems, at least one pure Nash equilibrium exists by design.

2.2 Potential Games

Consider the class of games [17] where player utilities $\{U_i\}_{i=1}^n$ are aligned with a global utility $\phi : Y \mapsto \mathcal{R}$ in the following sense: For every player, $\mathcal{P}_i \in \mathcal{P}$, for every $y_{-i} \in Y_{-i}$, and for every $\bar{y}_i, y_i'' \in Y_i$,

$$U_i(\bar{y}_i, y_{-i}) - U_i(y_i'', y_{-i}) > 0$$
$$\Rightarrow$$
$$\phi(\bar{y}_i, y_{-i}) - \phi(y_i'', y_{-i}) > 0.$$

In other words, an improvement in utility U_i received by any player $\mathcal{P}_i \in \mathcal{P}$ for its two different actions $\bar{y}_i, y_i'' \in Y_i$, when the actions of other players are fixed at $y_{-i} \in Y_{-i}$, always results in an improvement in the global utility ϕ. When player utilities satisfy the above alignment condition, the corresponding game is called a (generalized ordinal) potential game with the potential function ϕ. It is easy to see that, in potential games, an action profile maximizing the potential function is always a pure Nash equilibrium, i.e., at least one pure Nash equilibrium exists in potential games.

2.3 Weakly Acyclic Games

A weakly acyclic game [22, 23] is defined by the following condition: For any action profile $y \in Y$, there exists a finite of sequence of action profiles $y^1, y^2, ..., y^L$ such that

1. $y^1 = y$, and
2. For any $1 \leq \ell \leq L-1$, there is one player $\mathcal{P}_{i_\ell} \in \{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$ such that

$$y_{-i_\ell}^{\ell} = y_{-i_\ell}^{\ell+1} \text{ and } U_{i_\ell}(y^\ell) < U_{i_\ell}(y^{\ell+1}).$$

3. y^L is a pure Nash equilibrium.

It is straightforward to verify that any potential game is a weakly acyclic game. In the rest of the paper, we will focus on weakly acyclic games (for which at least one pure Nash equilibrium exists by definition) because we believe that most engineered multi-agent systems will lead to weakly acyclic games.

2.4 Repeated Games

Here, we deal with the issue of how players can learn to play a pure Nash equilibrium through repeated interactions; see [4, 23, 22, 11, 19, 18]. We assume that each player has access to its own utility function but not to the utility functions of other players. This private utilities assumption is motivated in multi-agent systems by the requirement that each agent has access to local information only.

We now consider a repeated game where, at each step $k \in \{1, 2, \dots\}$, each player $\mathcal{P}_i \in \mathcal{P}$ simultaneously chooses an action $y_i(k) \in Y_i$ and receives the utility $U_i(y(k))$ where $y(k) := (y_1(k), \dots, y_n(k))$. Each player $\mathcal{P}_i \in \mathcal{P}$ chooses its action $y_i(k)$ at step k according to a probability distribution $p_i(k)$ which we will specify later. Also, at step k before choosing action $y_i(k)$, each player $\mathcal{P}_i \in \mathcal{P}$ adjusts his strategy $p_i(k)$ based on the previous action profiles $y(1), \dots, y(k-1)$ which are accessible by all players at step k. More formally, strategy adjustment mechanism of player $\mathcal{P}_i$ takes the form

$$p_i(k) = F_i \ y(1), ..., y(k-1); U_i \ .$$

There are many important considerations in the choice of F_i such as computational burden on each player as well as the players' long-term behavior. In the next section, we will review a particular strategy update mechanism, namely regret matching, which have reasonable computational burden on each player. We will then go on to prove that a class of regret based dynamics including regret matching are convergent to a pure Nash equilibrium in all weakly acyclic games whenever players exponentially discount their past information and use some inertia in the decision making process.

3. REGRET MATCHING

We introduce regret matching, from [8], whose main distinction is that players choose their actions based on their *regret* for not choosing particular actions in the past steps.

Define the average regret of player $\mathcal{P}_i$ for an action $\bar{y}_i \in Y_i$ as

$$R_i^{\bar{y}_i}(k) := \frac{1}{k-1} \sum_{m=1}^{k-1} \left(U_i(\bar{y}_i, y_{-i}(m)) - U(y(m)) \right).$$

In other words, the player $\mathcal{P}_i$'s average regret for $\bar{y}_i \in Y_i$ would represent the average improvement in his utility if he had chosen $\bar{y}_i \in Y_i$ in all past steps and all other players' actions had remained unaltered.

Each player $\mathcal{P}_i$ using regret matching computes $R_i^{\bar{y}_i}(k)$ for every action $\bar{y}_i \in Y_i$ using the recursion

$$R_i^{\bar{y}_i}(k+1) = \frac{k-1}{k} R_i^{\bar{y}_i}(k) + \frac{1}{k} \left(U_i(\bar{y}_i, y_{-i}(k)) - U_i(y(k)) \right).$$

Note that, at every step $k > 1$, player $\mathcal{P}_i$ updates all entries in his average regret vector $R_i(k) := \left. R_i^{\bar{y}_i}(k) \right._{\bar{y}_i \in Y_i}$. To update his average regret vector at step k, it is sufficient for player $\mathcal{P}_i$ to observe (in addition to $y_i(k-1)$) his hypothetical utilities $U_i(\bar{y}_i, y_{-i}(k-1))$, for all $\bar{y}_i \in Y_i$, that would have been received if he had chosen $\bar{y}_i$ (instead of $y_i(k-1)$) and all other player actions $y_{-i}(k-1)$ had remained unchanged at step $k-1$.

Once player $\mathcal{P}_i$ computes his average regret vector, $R_i(k)$, he chooses an action $y_i(k)$, $k > 1$, according to the probability distribution $p_i(k)$ defined as

$$p_i^{\bar{y}_i}(k) = \text{Prob}(y_i(k) = \bar{y}_i) = \frac{R_i^{\bar{y}_i}(k)^{\ +}}{\sum_{\bar{y}_i \in Y_i} \left[R_i^{\bar{y}_i}(k) \right]^+},$$

for any $\bar{y}_i \in Y_i$, provided that the denominator above is positive; otherwise, $p_i(k)$ is the uniform distribution over Y_i ($p_i(1) \in \Delta(|Y_i|)$ is always arbitrary). Roughly speaking, a player using regret matching chooses a particular action at any step with probability proportional to the average regret for not choosing that particular action in the past steps. It turns out that the average regret of a player using regret matching would asymptotically vanish (similar results hold for different regret based adaptive dynamics); see [8, 9, 10]. This would imply that the empirical frequencies of the action profiles $y(k)$ would almost surely converge to the set of *coarse correlated equilibria*[1], where a coarse correlated equilibrium is any probability distribution $z \in \Delta(|Y|)$ satisfying

$$\sum_{y_{-i} \in Y_{-i}} U_i(y_i, y_{-i}) z_{-i}(y_{-i}) \leq \sum_{y \in Y} U_i(y) z(y),$$

for all $y_i \in Y_i$ and for all $\mathcal{P}_i \in \{\mathcal{P}_1, \dots, \mathcal{P}_n\}$, where

$$z_{-i}(y_{-i}) := \sum_{\bar{y}_i \in Y_i} z(\bar{y}_i, y_{-i}),$$

see Theorem 3.1 in [23].

In general, the set of Nash equilibria is a proper subset of the set of coarse correlated equilibria. Consider for example the following 3−player identical interest game characterized by the player utilities shown in Figure 1.

M_1

	L	R
U	1	0
D	0	−1

M_2

	L	R
U	0	0
D	0	0

M_3

	L	R
U	−1	0
D	0	1

Figure 1: A 3−player Identical Interest Game.

Player $\mathcal{P}_1$ chooses a row U or D, Player $\mathcal{P}_2$ chooses a column L or R, Player $\mathcal{P}_3$ chooses a matrix M_1, or M_2, or M_3. There are two pure Nash equilibria (U, L, M_1) and (D, R, M_3) both of which yield maximum utility 1 to all players. The set of coarse correlated equilibria contains

[1]This does not mean that the action profiles $y(k)$ will converge, nor does it mean that the empirical frequencies of $y(k)$ will converge to a point in $\Delta(|Y|)$.

these two pure Nash equilibria as the extremum points of $\Delta(|Y|)$ as well as many other probability distributions in $\Delta(|Y|)$. In particular, the set of coarse correlated equilibria contains all those $z \in \Delta(|Y|)$ satisfying

$$\sum_{y \in Y : y_3 = M_2} z(y) = 1, \text{ and } z(ULM_2) = z(DRM_2),$$

which yield 0 utility to all players. Clearly, one of the two pure Nash equilibria would be more desirable to all players then any other outcome including the above coarse correlated equilibria. However, the existing results at the time of writing this paper such as Theorem 3.1 in [23] only guarantee that regret matching will lead players to the set of coarse correlated equilibria and not necessarily to a pure Nash equilibrium. While this example is simplistic in nature, one must believe that situations like this could arise in more general weakly acyclic games.

We should emphasize that regret matching could indeed be convergent to a pure Nash equilibrium in weakly acyclic games; however, to the best of authors' knowledge, no proof for such a statement exists. The existing results characterize the long-term behavior of regret matching in general games as convergence to the set of coarse correlated equilibria, whereas we are interested in proving that the action profiles $y(k)$ generated by regret matching will converge to a pure Nash equilibrium when player utilities constitute a weakly acyclic game, an objective which we will pursue in the next section.

4. REGRET BASED DYNAMICS WITH FADING MEMORY AND INERTIA

To enable convergence to a pure Nash equilibrium in weakly acyclic games, we will modify regret matching in two ways. First, we will assume that each player has a fading memory, that is, each player exponentially discounts the influence of its past regret in the computation of its average regret vector. More precisely, each player computes a discounted average regret vector according to the recursion

$$\tilde{R}_i^{\bar{y}_i}(k+1) = (1-\rho)\tilde{R}_i^{\bar{y}_i}(k) + \rho\left(U_i(\bar{y}_i, y_{-i}(k)) - U_i(y(k))\right),$$

for all $\bar{y}_i \in Y_i$, where $\rho \in (0,1]$ is a parameter with $1-\rho$ being the discount factor, and $\tilde{R}_i^{\bar{y}_i}(1) = 0$.

Second, we will assume that each player chooses an action based on its discounted average regret using some inertia. Therefore, each player $\mathcal{P}_i$ chooses an action $y_i(k)$, at step $k > 1$, according to the probability distribution

$$\alpha_i(k) RB_i(\tilde{R}_i(k)) + (1-\alpha_i(k))\mathbf{v}^{y_i(k-1)},$$

where $\alpha_i(k)$ is a parameter representing player $\mathcal{P}_i$'s willingness to optimize at time k, $\mathbf{v}^{y_i(k-1)}$ is the vertex of $\Delta(|Y_i|)$ corresponding to the action $y_i(k-1)$ chosen by player $\mathcal{P}_i$ at step $k-1$, and $RB_i : \mathcal{R}^{|Y_i|} \to \Delta(|Y_i|)$ is any continuous function (on $\{x \in \mathcal{R}^{|Y_i|} : [x]^+ \neq 0\}$) satisfying

$$\begin{gathered} x^\ell > 0 \Leftrightarrow RB_i^\ell(x) > 0 \\ \text{and} \\ [x]^+ = 0 \Rightarrow RB_i^\ell(x) = \tfrac{1}{|Y_i|}, \ \forall \ell, \end{gathered} \tag{2}$$

where x^ℓ and $RB_i^\ell(x)$ are the ℓ-th components of x and $RB_i(x)$ respectively.

We will call the above dynamics regret based dynamics (RB) with fading memory and inertia. One particular choice for the function RB_i is

$$RB_i^\ell(x) = \frac{x^{\ell\ +}}{\sum_{m=1}^{|Y_i|}[x^m]^+}, \text{ (when } [x]^+ \neq 0) \tag{3}$$

which leads to regret matching with fading memory and inertia. Another particular choice is

$$RB_i^\ell(x) = \frac{e^{\frac{1}{\tau}x^\ell}}{\sum_{x^m>0} e^{\frac{1}{\tau}x^m}} I\{x^\ell > 0\}, \text{ (when } [x]^+ \neq 0),$$

where $\tau > 0$ is a parameter. Note that, for small values of τ, player $\mathcal{P}_i$ would choose, with high probability, the action corresponding to the maximum regret. This choice leads to a stochastic variant of an algorithm called Joint Strategy Fictitious Play (with fading memory and inertia); see [14]. Also, note that, for large values of τ, player $\mathcal{P}_i$ would choose any action having positive regret with equal probability.

According to these rules, player $\mathcal{P}_i$ will stay with his previous action $y_i(k-1)$ with probability $1-\alpha_i(k)$ regardless of his regret. We make the following standing assumption on the players' willingness to optimize.

ASSUMPTION 4.1. *There exist constants $\underline{\varepsilon}$ and $\bar{\varepsilon}$ such that*

$$0 < \underline{\varepsilon} < \alpha_i(k) < \bar{\varepsilon} < 1$$

for all steps $k > 1$ and for all $i \in \{1, ..., n\}$.

This assumption implies that players are always willing to optimize with some nonzero inertia[2] We make the following assumption on the Nash equilibria of the game.

ASSUMPTION 4.2. *All pure Nash equilibria are strict.*

The following theorem establishes the convergence of RB with fading memory and inertia to a pure Nash equilibrium.

THEOREM 4.1. *In any weakly acyclic game satisfying Assumption 4.2, the action profiles $y(t)$ generated by RB with fading memory and inertia satisfying Assumption 4.1 converge to a pure Nash equilibrium almost surely.*

We provide a complete proof for the above result in the Appendix. We note that, in contrast to the existing weak convergence results for regret matching in general games, the above result characterizes the long-term behavior of RB with fading memory and inertia, in a strong sense, albeit in a restricted class of games. We next numerically verify our theoretical result through some simulations.

5. SIMULATIONS

We extensively simulated the RB iterations for the game considered at the end of Section 3. We used the RB_i function given in (3) with inertia factor $\alpha = 0.5$ and discount

[2]This assumption can be relaxed to holding for sufficiently large k, as opposed to all k.

factor $\rho = 0.1$. In all cases, player action profiles $y(k)$ converged to one of the pure Nash equilibria as predicted by our main theoretical result. A typical simulation run shown in Figure 2 illustrates the convergence of RB iterations to the pure Nash equilibrium (D, R, M_3).

6. CONCLUSIONS

In this paper we analyzed the applicability of no-regret algorithms on multi-agent systems from a worst-case perspective. We demonstrated that a point of no-regret may not necessarily be a desirable operating condition. Furthermore, the existing results on regret-based algorithms do not preclude these inferior operating points. Therefore, we introduced a modification of the regret-based algorithms that (1) exponentially discounts the memory and (2) brings in a notion of inertia in players' decision process. We showed how these modifications can lead to an entire class of regret based algorithms that provide convergence to a pure Nash equilibrium in any weakly acyclic game. The authors believe that similar results hold for no-regret algorithms without fading memory and inertia but thus far the proofs have been elusive.

7. ACKNOWLEDGMENTS

The first and second authors gratefully acknowledge Lockheed Martin for funding support of this work, Subcontract No. 22MS13658.

8. REFERENCES

[1] B. Banerjee and J. Peng. Efficient no-regret multiagent learning. In *The 20th National Conference on Artificial Intelligence (AAAI-05)*, 2005.

[2] A. Blum, E. Evan-Dar, and K. Ligett. On convergence to nash equilibria of regret-minimizing algorithms in routing games. In *PODC*, 2006.

[3] M. Bowling. Convergence and no-regret in multiagent learning. In *NIPS*, 2004.

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

[5] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, Cambridge, MA, 1991.

[6] G. J. Gordon. No-regret algorithms for structured prediction problems. Technical Report CMU-CALD-05-112, Department of Machine Learning at Carnegie Mellon.

[7] A. Greenwald and A. Jafari. A general class of no-regret learning algorithms and game-theoretic equilibria. In *COLT*, pages 2–12, 2003.

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

[9] S. Hart and A. Mas-Colell. A general class of adaptative strategies. *Journal of Economic Theory*, **98**:26–54, 2001.

[10] S. Hart and A. Mas-Colell. Regret based continuous-time dynamics. *Games and Economic Behavior*, **45**:375–394, 2003.

[11] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, UK, 1998.

[12] A. Jafari, A. Greenwald, D., and G. Ercal. On no-regret learning, fictitious play, and nash equilibrium. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 226–233, 2001.

[13] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[14] J. R. Marden, G. Arslan, and J. S. Shamma. Joint strategy fictitious play with inertia for potential games. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 6692–6697, December 2005. Submitted to *IEEE Transactions on Automatic Control*.

[15] J. R. Marden, G. Arslan, and J. S. Shamma. Connections between cooperative control and potential games illustrated on the consensus problem. July 2007. Submitted to *Proceedings of the European Control Conference*.

[16] D. Monderer and L. Shapley. Fictitious play property for games with identical interests. *Journal of Economic Theory*, **68**:258–265, 1996.

[17] D. Monderer and L. Shapley. Potential games. *Games and Economic Behavior*, **14**:124–143, 1996.

[18] L. Samuelson. *Evolutionary Games and Equilibrium Selection*. MIT Press, Cambridge, MA, 1997.

[19] J. Weibull. *Evolutionary Game Theory*. MIT Press, Cambridge, MA, 1995.

[20] D. Wolpert and K. Tumor. An overview of collective intelligence. In J. M. Bradshaw, editor, *Handbook of Agent Technology*. AAAI Press/MIT Press, 1999.

[21] D. H. Wolpert and K. Tumor. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, **4**(2&3):265–279, 2001.

[22] H. P. Young. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton University Press, New Jersey, 1998.

[23] H. P. Young. *Strategic Learning and Its Limits*. Oxford University Press Inc., New York, US, 2004.

APPENDIX

A. PROOF OF THEOREM 4.1

We will first state and prove a series of claims.

CLAIM A.1. *Fix any $k_0 > 1$. Then,*

$$\tilde{R}_i^{y_i(k_0)}(k_0) > 0 \Rightarrow \tilde{R}_i^{y_i(k)}(k) > 0$$

for all $k > k_0$.

PROOF. Suppose $\tilde{R}_i^{y_i(k_0)}(k_0) > 0$. We have

$$\tilde{R}_i^{y_i(k_0)}(k_0+1) = (1-\rho)\tilde{R}_i^{y_i(k_0)}(k_0) > 0.$$

If $y_i(k_0+1) = y_i(k_0)$, then

$$\tilde{R}_i^{y_i(k_0+1)}(k_0+1) = \tilde{R}_i^{y_i(k_0)}(k_0+1) > 0.$$

If $y_i(k_0+1) \neq y_i(k_0)$, then

$$\tilde{R}_i^{y_i(k_0+1)}(k_0+1) > 0.$$

The argument can be repeated to show that $\tilde{R}_i^{y_i(k)}(k) > 0$, for all $k > k_0$. □

Define

$$\begin{aligned}
M_u &:= \max\{U_i(y) : y \in Y, \mathcal{P}_i \in \mathcal{P}\},\\
m_u &:= \min\{U_i(y) : y \in Y, \mathcal{P}_i \in \mathcal{P}\},\\
\delta &:= \min\{|U_i(y^1) - U_i(y^2)| > 0 :\\
&\qquad y^1, y^2 \in Y, y^1_{-i} = y^2_{-i}, \mathcal{P}_i \in \mathcal{P}\},\\
N &:= \min\{n \in \{1,2,...\} :\\
&\qquad (1-(1-\rho)^n)\delta - (1-\rho)^n(M_u - m_u) > \delta/2\},\\
f &:= \min\{RB_i^m(x) : |x^\ell| \leq M_u - m_u, \forall \ell,\\
&\qquad x^m \geq \delta/2, \text{ for one } m, \ \forall \mathcal{P}_i \in \mathcal{P}\}.
\end{aligned}$$

Note that $\delta, f > 0$, and $|\tilde{R}_i^{y_i}(k)| \leq M_u - m_u$, for all $\mathcal{P}_i \in \mathcal{P}$, $y_i \in Y_i$, $k > 1$.

CLAIM A.2. *Fix $k_0 > 1$. Assume*

1. *$y(k_0)$ is a strict Nash equilibrium, and*
2. *$\tilde{R}_i^{y_i(k_0)}(k_0) > 0$ for all $\mathcal{P}_i \in \mathcal{P}$, and*
3. *$y(k_0) = y(k_0+1) = ... = y(k_0+N-1)$.*

Then, $y(k) = y(k_0)$, for all $k \geq k_0$.

PROOF. For any $\mathcal{P}_i \in \mathcal{P}$ and any $y_i \in Y_i$, we have

$$\begin{aligned}
\tilde{R}_i^{y_i}(k_0+N) &= (1-\rho)^N \tilde{R}_i^{y_i}(k_0)\\
&\quad + 1-(1-\rho)^N \quad U_i(y_i, y_{-i}(k_0))\\
&\quad -U_i(y_i(k_0), y_{-i}(k_0)) \ .
\end{aligned}$$

Since $y(k_0)$ is a strict Nash equilibrium, for any $\mathcal{P}_i \in \mathcal{P}$ and any $y_i \in Y_i$, $y_i \neq y_i(k_0)$, we have

$$U_i(y_i, y_{-i}(k_0)) - U_i(y_i(k_0), y_{-i}(k_0)) \leq -\delta.$$

Therefore, for any $\mathcal{P}_i \in \mathcal{P}$ and any $y_i \in Y_i$, $y_i \neq y_i(k_0)$,

$$\begin{aligned}
\tilde{R}_i^{y_i}(k_0+N) &\leq (1-\rho)^N(M_u - m_u) - (1-(1-\rho)^N)\delta\\
&< -\delta/2 < 0.
\end{aligned}$$

We also know that, for all $\mathcal{P}_i \in \mathcal{P}$,

$$\tilde{R}_i^{y_i(k_0)}(k_0+N) = (1-\rho)^N \tilde{R}_i^{y_i(k_0)}(k_0) > 0.$$

This proves the claim. □

CLAIM A.3. *Fix $k_0 > 1$. Assume*

1. *$y(k_0)$ is not a Nash equilibrium, and*
2. *$y(k_0) = y(k_0+1) = ... = y(k_0+N-1)$*

Let $y^* = (y_i^*, y_{-i}(k_0))$ be such that

$$U_i(y_i^*, y_{-i}(k_0)) > U_i(y_i(k_0), y_{-i}(k_0)),$$

for some $\mathcal{P}_i \in \mathcal{P}$ and some $y_i^* \in Y_i$. Then, $\tilde{R}_i^{y_i^*}(k_0+N) > \delta/2$, and y^* will be chosen at step k_0+N with at least probability $\gamma := (1-\bar{\epsilon})^{n-1}\underline{\epsilon} f$.

PROOF. We have

$$\begin{aligned}
\tilde{R}_i^{y_i^*}(k_0+N) &\geq -(1-\rho)^N(M_u - m_u) + (1-(1-\rho)^N)\delta\\
&> \delta/2.
\end{aligned}$$

Therefore, the probability of player $\mathcal{P}_i$ choosing y_i^* at step k_0+N is at least $\underline{\epsilon} f$. Because of players' inertia, all other players will repeat their actions at step k_0+N with probability at least $(1-\bar{\epsilon})^{n-1}$. This means that the action profile y^* will be chosen at step k_0+N with probability at least $(1-\bar{\epsilon})^{n-1}\underline{\epsilon} f$. □

CLAIM A.4. *Fix $k_0 > 1$. We have $\tilde{R}_i^{y_i(k)}(k) > 0$ for all $k \geq k_0 + 2Nn$ and for all $\mathcal{P}_i \in \mathcal{P}$ with probability at least*

$$\prod_{i=1}^{n} \frac{1}{|Y_i|}\gamma(1-\bar{\epsilon})^{2Nn}.$$

PROOF. Let $y^0 := y(k_0)$. Suppose $\tilde{R}_i^{y_i^0}(k_0) < 0$. Furthermore, suppose that y^0 is repeated N consecutive times, i.e. $y(k_0) = ... = y(k_0+N-1) = y^0$, which occurs with at least probability at least $(1-\bar{\epsilon})^{n(N-1)}$.

If there exists a $y^* = (y_i^*, y_{-i}^0)$ such that $U_i(y^*) > U_i(y^0)$, then, by Claim A.3, $\tilde{R}_i^{y_i^*}(k_0+N) > \delta/2$ and y^* will be chosen at step k_0+N with at least probability γ. Conditioned on this, we know from Claim A.1 that $\tilde{R}_i^{y_i(k)}(k) > 0$ for all $k \geq k_0+N$.

If there does not exist such an action y^*, then $\tilde{R}_i^{y_i}(k_0+N) < 0$ for all $y_i \in Y_i$. An action profile (y_i^w, y_{-i}^0) with $U_i(y_i^w, y_{-i}^0) < U_i(y^0)$ will be chosen at step k_0+N with at least probability $\frac{1}{|Y_i|}(1-\bar{\epsilon})^{n-1}$. If $y(k_0+N) = (y_i^w, y_{-i}^0)$, and if furthermore (y_i^w, y_{-i}^0) is repeated N consecutive times, i.e., $y(k_0+N) = ... = y(k_0+2N-1)$, which happens with probability at least $(1-\bar{\epsilon})^{n(N-1)}$, then, by Claim A.3, $\tilde{R}_i^{y_i^0}(k_0+2N) > \delta/2$ and the action profile y^0 will be chosen at step (k_0+2N) with at least probability γ. Conditioned on this, we know from Claim A.1 that $\tilde{R}_i^{y_i(k)}(k) > 0$ for all $k \geq k_0+2N$.

In summary, $\tilde{R}_i^{y_i(k)}(k) > 0$ for all $k \geq k_0 + 2N$ with at least probability

$$\frac{1}{|Y_i|}\gamma(1-\bar{\epsilon})^{2Nn}.$$

We can repeat this argument for each player to show that $\tilde{R}_i^{y_i(k)}(k) > 0$ for all times $k \geq k_0 + 2Nn$ and for all $\mathcal{P}_i \in \mathcal{P}$ with probability at least

$$\prod_{i=1}^{n} \frac{1}{|Y_i|}\gamma(1-\bar{\epsilon})^{2Nn}.$$

FINAL STEP: Establishing convergence to a strict Nash equilibrium:

Fix $k_0 > 1$. Define $k_1 := k_0 + 2Nn$. Let $y^1, y^2, \ldots, y^L$ be a finite sequence of action profiles satisfying the conditions given in Subsection 2.3 with $y^1 := y(k_1)$.

Suppose $\tilde{R}_i^{y_i(k)}(k) > 0$ for all $k \geq k_1$ and for all $\mathcal{P}_i \in \mathcal{P}$, which, by Claim A.4, occurs with probability at least

$$\prod_{i=1}^{n} \frac{1}{|Y_i|}\gamma(1-\bar{\epsilon})^{2Nn}.$$

Suppose further that $y(k_1) = ... = y(k_1 + N - 1) = y^1$ which occurs with at least probability $(1-\bar{\epsilon})^{n(N-1)}$. According to Claim A.3 the action profile y^2 will be played at step $k_2 := k_1 + N$ with at least probability γ. Suppose now $y(k_2) = ... = y(k_2 + N - 1) = y^2$, which occurs with at least probability $(1-\bar{\epsilon})^{n(N-1)}$. According to Claim A.3, the action profile y^3 will be played at step $k_3 := k_2 + N$ with at least probability γ.

We can repeat the above arguments until we reach the strict Nash equilibrium y^L at step k_L (recursively defined as above) and stay at y^L for N consecutive steps. From Claim 2, this would mean that the action profile would stay at y^L for all $k \geq k_L$.

Therefore, given $k_0 > 1$, there exists constants $\tilde{\epsilon} > 0$ and $\tilde{T} > 0$, both of which are independent of k_0, and a strict Nash equilibrium y^*, such that the following event happens with at least probability $\tilde{\epsilon}$: $y(k) = y^*$ for all $k \geq k_0 + \tilde{T}$. This proves Theorem 4.1.

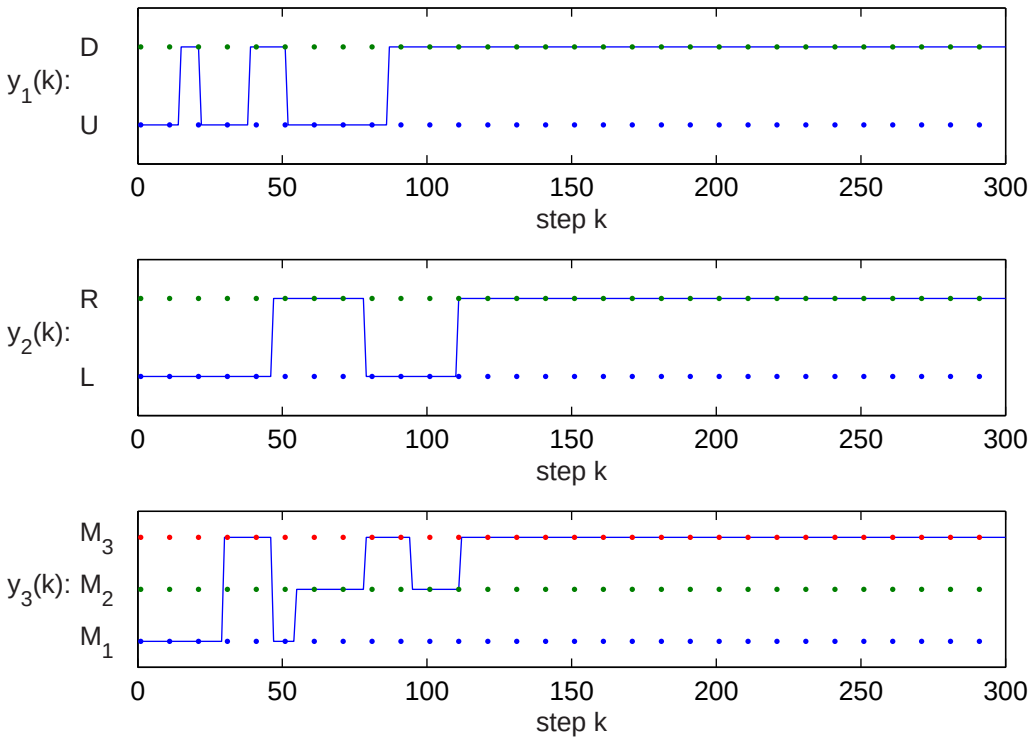

Figure 2: Evolution of the actions of players using RB.