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

No-regret algorithms have been proposed to control a wide variety of multi-agent systems. The appeal of no-regret algorithms is that they are easily implementable in large scale multi-agent systems because players make decisions using only retrospective or "regret based" information. Furthermore, there are existing results proving that the collective 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 no-regret algorithms. In this paper, we introduce a modification of the traditional no-regret algorithms by (i) exponentially discounting the memory and (ii) bringing in a notion of inertia in players' decision process. We show how these modifications can lead to an entire class of regret based algorithms that provide *almost sure* convergence to a pure Nash equilibrium in any weakly acyclic game.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence — *Intelligent Agents; Multiagent systems*

General Terms

Algorithms, Design, Theory

Keywords

Game Theory; Learning Algorithms; Cooperative distributed problem solving; Multiagent learning; Emergent behavior

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

AAMAS'07, May 14–18, 2007, Honolulu, Hawai'i, USA.

1. INTRODUCTION

The applicability of regret based algorithms for multi-agent learning has been studied in several papers [7, 4, 14, 2, 8, 1]. 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 [19, 13]. 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 coarse correlated equilibrium) in any game [28]. 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 [3] to structured prediction problems [7].

In the more general regret based algorithms, each player makes a decision using *only* information regarding the regret for each of his possible actions. If an 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 [9]. 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 asymptotically to a set of points of no-regret [9]. Traditionally, a point of no-regret has been viewed 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 [14]. 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 [13], Jafari et al. showed through simulations that no-regret algorithms provide convergence to a Nash equilibrium in dominance solvable, constant-sum,

978-81-904262-7-5 (RPS)

and general sum 2×2 games. In [4], 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 [3], 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 wide variety of multi-agent systems.

It turns out that weakly acyclic games, which generalize potential games [20], are closely related to multi-agent systems [17]. The connection can be seen by recognizing that in any multi-agent system there is a global objective. 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.

An open question is whether no-regret algorithms converge to a Nash equilibrium in n-player weakly acyclic games. In this paper, we introduce a modification of the traditional no-regret algorithms that (i) exponentially discounts the memory and (ii) brings in a notion of inertia in players' decision process. We show how these modifications can lead to an *entire class* of regret based algorithms 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 the 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.

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

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

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

2. SETUP

2.1 Finite Strategic-Form Games

A finite strategic-form game [6] 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 $U_i : Y \to \mathbf{R}$ 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 $y := (y_1, \ldots, y_n)$. Before introducing the definition of Nash equilibrium, 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 [26, 25, 15]. In an engineered multi-agent system, there may exists a global objective function $\phi : Y \to \mathbf{R}$ that a global planner is seeking to maximize. Furthermore, players' local utility functions need to be somewhat aligned with the global objective function. It is precisely this alignment that guarantees the existence of at least one pure Nash equilibrium in such systems [17].

We will now introduce three classes of games that fit within the multi-agent systems framework. Roughly speaking, the difference between the classes is the degree to which the players' local utility functions are aligned with the global objective function.

2.2 Identical Interest Games

The first class of games that we will consider is identical interest games. In such a game, the players' utility functions $\{U_i\}_{i=1}^n$ are chosen to be the same. That is, for some function $\phi : Y \to \mathbf{R}$,

$$U_i(y) = \phi(y),$$

for every $\mathcal{P}_i \in \mathcal{P}$ and for every $y \in Y$. It is easy to verify that all identical interest games have at least one pure Nash equilibrium, namely any action profile y that maximizes $\phi(y)$.

2.3 Potential Games

A significant generalization of an identical interest game is a potential game [20]. In a potential game, the change in a player's utility that results from a unilateral change in strategy equals the change in the "potential". Specifically, there is a function $\phi : Y \to \mathbf{R}$ such that for every player $\mathcal{P}_i \in \mathcal{P}$, for every $y_{-i} \in Y_{-i}$, and for every $y_i', y_i'' \in Y_i$,

$$U_i(y_i', y_{-i}) - U_i(y_i'', y_{-i}) = \phi(y_i', y_{-i}) - \phi(y_i'', y_{-i}).$$

When this condition is satisfied, the game is called a potential game with the potential function ϕ. It is easy to see that in potential games, any action profile maximizing the potential function is a pure Nash equilibrium, hence every potential game possesses at least one such equilibrium.

2.4 Weakly Acyclic Games

Consider any finite game G with a set Y of action profiles. A *better reply path* is a sequence of action profiles $y^1, y^2, ..., y^L$ such that, for every $1 \leq \ell \leq L-1$, there is exactly one player $\mathcal{P}_{i_\ell}$ such that i) $y^\ell_{i_\ell} \neq y^{\ell+1}_{i_\ell}$, ii) $y^\ell_{-i_\ell} = y^{\ell+1}_{-i_\ell}$, and iii) $U_{i_\ell}(y^\ell) < U_{i_\ell}(y^{\ell+1})$. In other words, one player moves at a time, and each time a player moves he increases his own utility.

Suppose now that G is a potential game with potential function ϕ. Starting from an arbitrary action profile $y \in Y$, construct a better reply path $y = y^1, y^2, ..., y^L$ until it can no longer be extended. Note first that such a path cannot cycle back on itself, because ϕ is strictly increasing along the path. Since Y is finite, the path cannot be extended indefinitely. Hence, the last element in a maximal better reply path from any joint action, y, must be a Nash equilibrium of G.

This idea may be generalized as follows. The game G is *weakly acyclic* if for any $y \in Y$, there exists a better reply path starting at y and ending at some pure Nash equilibrium of G [27, 28]. Potential games are special cases of weakly acyclic games.

Similarly to potential games, a weakly acyclic game can also be interpreted through the following global objective condition: a game G is weakly acyclic if there exists a global objective function $\phi : Y \to \mathbf{R}$ such that for any action profile $y \in Y$ that is *not* a Nash equilibrium, there exists at least one player $\mathcal{P}_i \in \mathcal{P}$ with an action $y'_i \in Y_i$ such that

$$U_i(y'_i, y_{-i}) > U_i(y) \text{ and } \phi(y'_i, y_{-i}) > \phi(y).$$

Roughly speaking, a weakly acyclic game requires that for any action profile at least one player's local utility function is aligned with the global objective.

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 [17].

2.5 Repeated Games

Here, we deal with the issue of how players can learn to play a pure Nash equilibrium through repeated interactions; see [5, 28, 27, 12, 24, 22]. 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, the general strategy adjustment mechanism of player $\mathcal{P}_i$ takes the form

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

A strategy adjustment mechanism of this form has complete information, meaning that each player is (i) able to observe the complete action profile at every time step and (ii) is aware of the structural form of his utility function. In this paper, we will consider strategy adjustment mechanism with significantly less informational requirements.

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 has 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 [9], in which 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

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

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 the actual utility received at time $k-1$, $U_i(y(k-1))$) 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 $k-1$.

In regret matching, 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

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

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 will asymptotically vanish (similar results hold for different regret based adaptive dynamics); see [9, 10, 11]. This implies 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 $\mathcal{P}_i \in \{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$, where

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

see Theorem 3.1 in [28].

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.

	L	R
U	2	−1
D	1	−2

M_1

	L	R
U	0	0
D	0	0

M_2

	L	R
U	−2	1
D	−1	2

M_3

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 2 to all players. The set of coarse correlated equilibria contains 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

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

subject to $z(ULM_2) = z(DRM_2)$ and $z(URM_2) = z(DLM_2)$. Any coarse correlated equilibrium of this form yields a utility of 0 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 [28] 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 easily 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.

[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 $\Delta(|Y|)$.

4. REGRET BASED DYNAMICS WITH FADING MEMORY AND INERTIA

To enable convergence to a pure Nash equilibrium in weakly acyclic games, we will modify the conventional regret based dynamics 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 $RB_i : \mathbf{R}^{|Y_i|} \to \Delta(|Y_i|)$ is any continuous function (on $\{x \in \mathbf{R}^{|Y_i|} : [x]^+ \neq 0\}$) 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

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

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 [16]. 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]. A motivation for the use of inertia is to instill a degree of hesitation into the decision making process to ensure that players do not overreact to various situations. We will assume that no player is indifferent between distinct strategies.

ASSUMPTION 4.2. *Player utilities satisfy*

$$U_i(y_i^1, y_{-i}) \neq U_i(y_i^2, y_{-i}), \forall\ y_i^1, y_i^2 \in Y_i,$$
$$y_i^1 \neq y_i^2,\ \forall\ y_{-i} \in Y_{-i},\ \forall\ i \in \{1, ..., n\}.$$

The following theorem establishes the convergence of regret based dynamics 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 regret based dynamics 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 regret based dynamics 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

5.1 Three Player Identical Interest Game

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 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).

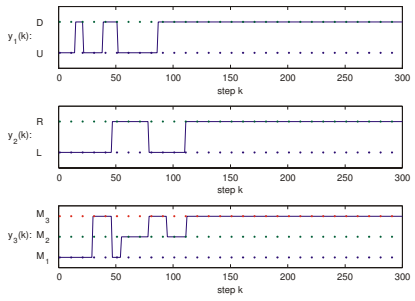

Figure 2: Evolution of the actions of players using RB.

5.2 Distributed Traffic Routing

5.2.1 Congestion Game Setup

Congestion games are a specific class of games in which player utility functions have a special structure.

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

In order to define a congestion game, we must specify the action set, Y_i, and utility function, $U_i(\cdot)$, of each player. Towards this end, let $\mathcal{R}$ denote a finite set of "resources". For each resource $r \in \mathcal{R}$, there is an associated "congestion function"

$$c_r : \{0, 1, 2, ...\} \rightarrow \mathbf{R}$$

that reflects the cost of using the resource as a function of the number of players using that resource.

The action set, Y_i, of each player, $\mathcal{P}_i$, is defined as the set of resources available to player $\mathcal{P}_i$, i.e.,

$$Y_i \subset 2^{\mathcal{R}},$$

where $2^{\mathcal{R}}$ denotes the set of subsets of $\mathcal{R}$. Accordingly, an action, $y_i \in Y_i$, reflects a selection of (multiple) resources, $y_i \subset \mathcal{R}$. A player is "using" resource r if $r \in y_i$. For an action profile $y \in Y_1 \times \cdots \times Y_n$, let $\sigma_r(y)$ denote the total number of players using resource r, i.e., $|\{i : r \in y_i\}|$. In a congestion game, the utility of player $\mathcal{P}_i$ using resources indicated by y_i depends only on the total number of players using the same resources. More precisely, the utility of player $\mathcal{P}_i$ is defined as

$$U_i(y) = -\sum_{r \in y_i} c_r(\sigma_r(y)). \tag{4}$$

The negative sign stems from $c_r(\cdot)$ reflecting the cost of using a resource and $U_i(\cdot)$ reflecting a utility or reward function. Any congestion game with utility functions as in (4) is a potential game [21, 20] with the potential function

$$\phi(y) = -\sum_{r \in R} \sum_{k=1}^{\sigma_r(y)} c_r(k).$$

5.2.2 Distributed Traffic Routing Example

We consider a simple scenario with 100 players seeking to traverse from node A to node B along 5 different parallel roads as illustrated in Figure 3. Each player can select any road as a possible route. In terms of congestion games, the set of resources is the set of roads, $\mathcal{R}$, and each player can select one road, i.e., $Y_i = \mathcal{R}$.

We will assume that each road has a linear cost function

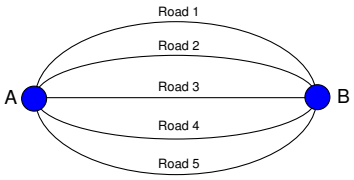

Figure 3: Network Topology for a Congestion Game

with positive (randomly chosen) coefficients,

$$c_{r_i}(k) = a_i k + b_i, \ i = 1, ..., 5,$$

where k represent the number of vehicles on that particular road. This cost function may represent the delay incurred by a driver as a function of the number of other drivers sharing the same road. The actual coefficients or structural form of the cost function are unimportant as we are just using this example as an opportunity to illustrate the convergence properties of the proposed regret based algorithms.

We simulated a case where drivers choose their initial routes randomly, and every day thereafter, adjusted their routes using the regret based dynamics with the RB_i function given in (3) with inertia factor $\alpha = 0.85$ and discount factor $\rho = 0.1$. The number of vehicles on each road fluctuates initially and then stabilizes as illustrated in Figure 4. Figure 5 illustrates the evolution of the congestion cost on each road. One can observe that the congestion cost on each road converges approximately to the same value, which is consistent with a Nash equilibrium with large number of drivers. This behavior resembles an approximate "Wardrop equilibrium" [23], which represents a steady-state situation in which the congestion cost on each road is equal due to the fact that, as the number of drivers increases, the effect of an individual driver on the traffic conditions becomes negligible.

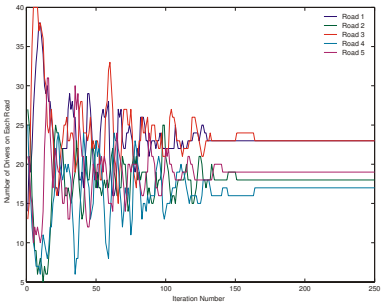

Figure 4: Evolution of Number of Vehicles on Each Route

We would like to note that the simplistic nature of this example was solely for illustrative purposes. Regret based dynamics could be employed on any congestion game with arbitrary network topology and congestion functions. Furthermore, well known learning algorithms such as fictitious play [19] could not be implemented even on this very simple congestion game. A driver using fictitious play would need to track the empirical frequencies of the choices of the 99 other drivers and compute an expected utility evaluated over a probability space of dimension 5^{99}.

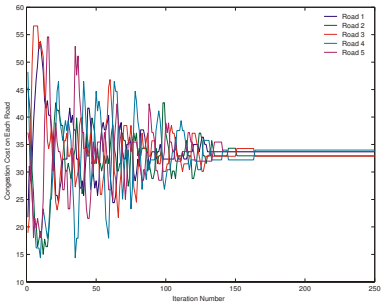

Figure 5: Evolution of Congestion Cost on Each Route

We would also like to note that in a congestion game, it may be unrealistic to assume that players are aware of the congestion function on each road. This implies that each driver is unaware of his own utility function. However, even in this setting, regret based dynamics can be effectively employed under the condition that each player can evaluate congestion levels on alternative routes. On the other hand, if a player is only aware of the congestion experienced, then one would need to examine the applicability of payoff based algorithms [18].

6. CONCLUSIONS

In this paper we analyzed the applicability of regret based algorithms on multi-agent systems. 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 traditional no-regret algorithms that (i) exponentially discounts the memory and (ii) 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. Research supported in part by ARO grant #W911NF0410316 and NSF grant #ECS-0501394.

8. REFERENCES

[1] G. Arslan, J. R. Marden, and J. S. Shamma. Autonomous vehicle-target assignment: A game theoretical formulation. *ASME Journal of Dynamic Systems, Measurement and Control*, 2006. submitted to.

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

[3] A. Blum, E. Evan-Dar, and K. Ligett. Routing without regret: On convergence to Nash equilibria of regret-minimizing algorithms in routing games. In *the Proceedings of the 25th Annual ACM Symposium on Principles of Distributed Computing*, pages 45–52, 2006.

[4] M. Bowling. Convergence and no-regret in multiagent learning. In *Advances in Neural Information Processing Systems*, page 209, 2005.

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

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

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

[8] A. Greenwald and A. Jafari. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Conference on Learning Theory*, pages 2–12, 2003.

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

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

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

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

[13] A. Jafari, A. Greenwald, D. Gondek, 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.

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

[15] S. Mannor and J. Shamma. Multi-agent learning for engineers. 2006. to appear in special issue of *Artificial Intelligence*.

[16] 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*.

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

[18] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma. Payoff based dynamics for multi-player weakly acyclic games. *submitted to SIAM Journal of Control and Optimization*, 2007.

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

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

[21] R. W. Rosenthal. A class of games possessing pure-strategy Nash equilibria. *Int. J. Game Theory*, **2**:65–67, 1973.

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

[23] J. Wardrop. Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers*, volume I, pt. II, pages 325–378, London, Dec. 1952.

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

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

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

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

[28] H. P. Young. *Strategic Learning and its Limits*. Oxford University Press, 2005.

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

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

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 $\tilde{R}_i^{y_i^0}(k_0) \leq 0$. 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 $\tilde{R}_i^{y_i}(k_0+N) \leq 0$ 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 2.4 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. □