

# Performance Bounded Reinforcement Learning in Strategic Interactions

**Bikramjit Banerjee and Jing Peng**

Dept. of Electrical Engineering & Computer Science  
Tulane University  
New Orleans, LA 70118  
{banerjee, jp}@eecs.tulane.edu

## Abstract

Despite increasing deployment of agent technologies in several business and industry domains, user confidence in fully automated agent driven applications is noticeably lacking. The main reasons for such lack of trust in complete automation are scalability and non-existence of reasonable guarantees in the performance of self-adapting software. In this paper we address the latter issue in the context of learning agents in a Multiagent System (MAS). Performance guarantees for most existing on-line Multiagent Learning (MAL) algorithms are realizable only in the limit, thereby seriously limiting its practical utility. Our goal is to provide certain meaningful guarantees about the performance of a learner in a MAS, *while it is learning*. In particular, we present a novel MAL algorithm that (i) converges to a best response against stationary opponents, (ii) converges to a Nash equilibrium in self-play and (iii) achieves a constant bounded expected regret at any time (no-average-regret asymptotically) in arbitrary sized general-sum games with non-negative payoffs, and against any number of opponents.

## Introduction

Agent technologies are being increasingly deployed in several business and industry domains, including B2B exchanges, supply chain management, car manufacturing etc, and are already providing sustained dramatic benefits. Demands from e-Commerce, particularly on-line auctions, and distributed computing communities are also producing revolutionary ideas in the Agents and MAS research domains. However, user confidence in fully automated agent driven applications is noticeably lacking and a “human in the loop” mind set seems hard to overcome. The main reasons for such lack of trust in complete automation are scalability and non-existence of reasonable guarantees in performance of self-adapting software.

Existing on-line learning algorithms in single agent environments do provide some performance guarantees during learning, but such assurances in multiagent environments are not only lacking but also significantly difficult to provide. In a MAS, a learning agent also forms a part of the

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

environment for every other learning agent. As such, the environments of the learners are usually *non-stationary* which is the key challenge for MAL. It is apparently also a challenge for a learner to learn the moving target of a most beneficial behavior through exploration in such environments, and at the same time ensure “good” performance during the exploration. So much so that the latter property has not been investigated in depth so far. In MAS, either such guarantees are realizable only in the limit, thereby seriously limiting its practical interest, or are realizable in polynomial time (e.g. R-MAX (Brafman & Tennenholtz 2002b)) but only for limited classes of games (e.g., constant-sum games). Our goal is to address this aspect of MAL, i.e., to provide certain meaningful guarantees about the performance of a learner in a MAS, *while it is learning* in general-sum games. We develop a novel algorithm for MAL that maintains the existing policy convergence properties of MAL, viz. *rationality* and *convergence* (Bowling & Veloso 2002), and in addition (i) applies to arbitrary sized general-sum games with non-negative payoffs, (ii) ensures a *bounded regret* at any time such that it is asymptotically *no-regret*. This property guarantees that a learner cannot perform much worse than the best fixed policy at any time while it is learning. The current paper contributes this algorithm with the relevant analyses and is organized as follows: the next section presents definitions from the domain of Multiagent Reinforcement Learning in repeated games, followed by a section on the existing work related to the present endeavor. Thereafter we present the algorithm, the analysis, and conclude with a short summary.

## Multiagent Reinforcement Learning

A Multiagent Reinforcement Learning task is usually modeled (Littman 1994) as a Stochastic Game (SG, also called *Markov Game*), which is a Markov Decision Process with multiple controllers. We focus on stochastic games with a single state, also called repeated games. This refers to a scenario where a matrix game (defined below) is played repeatedly by multiple agents. We shall represent the action space of the *i*th agent as  $A_i$ .

**Definition 1** A matrix game with  $n$  players is given by an  $n$ -tuple of matrices,  $\langle \mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_n \rangle$  where  $\mathbf{R}_i$  is a matrix of dimension  $|A_1| \times |A_2| \dots \times |A_n|$ , such that the payoff of

the  $i$ th agent for the joint action  $(a_1, a_2, \dots, a_n)$  is given by the entry  $R_i(a_1, a_2, \dots, a_n)$ ,  $\forall i$ .

A *constant-sum game* (also called competitive games) is a special matrix game where  $\sum_i R_i(a_1, a_2, \dots, a_n) = c$ ,  $\forall (a_1, a_2, \dots, a_n) \in \prod_k A_k$ ,  $c$  being a constant. If  $c = 0$ , then it is also called a zero-sum game. An example of such a game with 2 players appears in Table 1. This game is called Rock-Scissor-Paper. Here  $A_1 = A_2 = \{R, S, P\}$  and the game payoffs for any joint action sum to 0 as shown in Table 1.

Table 1: Rock-Scissor-Paper Game.  $(a, b)$  in the  $(i, j)$ th cell is the tuple of payoffs for Row agent and Column agent (in that order) for each combination of their actions  $(i, j) \in \{R, S, P\} \times \{R, S, P\}$ .

Actions	Rock (R)	Scissor (S)	Paper (P)
Rock (R)	(0,0)	(1,-1)	(-1,1)
Scissor (S)	(-1,1)	(0,0)	(1,-1)
Paper (P)	(1,-1)	(-1,1)	(0,0)

A *mixed policy*, vector  $\pi_i \in PD(A_i)$  for agent  $i$ , is a probability distribution over  $A_i$ , where  $PD$  is the set of probability distributions over the action space. If the entire probability mass is concentrated on a single action (some actions), it is also called a *pure policy* (*partially mixed policy*). The joint policies of the opponents of the  $i$ th agent will be given by the vector  $\pi_{-i}$ . We shall usually refer to the  $i$ th agent as the learner and the rest of the agents as the opponents. The expected payoff of the learner at any stage in which the policy tuple  $\langle \pi_1, \pi_2, \dots, \pi_n \rangle$  is followed is given by  $V_i(\pi_i, \pi_{-i}) = \sum_{(a_1, \dots, a_n) \in \prod_k A_k} \pi_1(a_1) \dots \pi_n(a_n) R_i(a_1, \dots, a_n)$ .

**Definition 2** For an  $n$ -player matrix game, the best response ( $BR_{\pi_{-i}}^i$ ) of the  $i$ th agent to its opponents' joint policy  $(\pi_{-i})$  is given by  $BR_{\pi_{-i}}^i = \{\pi_i | V_i(\pi_i, \pi_{-i}) \geq V_i(\pi'_i, \pi_{-i}), \forall \pi'_i \in PD(A_i)\}$ .

**Definition 3** A *mixed-policy Nash Equilibrium (NE)* for a matrix game  $(\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_n)$  is a tuple of probability vectors  $\langle \pi_1^*, \pi_2^*, \dots, \pi_n^* \rangle$  (policy profile) such that each is a best response to the rest, i.e.,  $\pi_i^* \in BR_{\pi_{-i}^*}^i \forall i$ . In terms of payoffs, these conditions can be restated as  $V_i(\pi_i^*, \pi_{-i}^*) \geq V_i(\pi_i, \pi_{-i}^*) \forall \pi_i \in PD(A_i), \forall i$ .

No player in this game has any incentive for unilateral deviation from the Nash equilibrium policy, given the others' policy. There always exists at least one such equilibrium profile for an arbitrary finite matrix game (Nash 1951). As an example, the only NE for the 2 player RSP game in Table 1 is  $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$  and  $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$  for the two agents.

**Definition 4** For a given time range  $t = 0 \dots T$ , the regret of a learner (agent  $i$ ),  $Rg_i^T$  is given by  $Rg_i^T = \max_{\pi_i} \sum_{t=1}^{t=T} V_i(\pi_i, \pi_{-i}^t) - \sum_{t=1}^{t=T} V_i(\pi_i^t, \pi_{-i}^t)$ .

This means that if the sum of expected payoffs of the learner in the given time range against the actual unknown policies played by the non-stationary opponent were compared to that of an oracle who *knows* the actual policies to be played

by the opponent *ahead of time* and can statically compute a fixed policy  $\pi_i$  that maximizes  $\sum_{t=1}^{t=T} V_i(\pi_i, \pi_{-i}^t)$  but is *limited* to play only that policy all through the time window  $T$ , then the difference would be the former player's regret. In hindsight (after  $t = T$ ) he finds that always playing  $\pi_i$  instead of the sequence  $\{\pi_i^t\}$  would have yielded a total payoff higher than his actual payoff by  $Rg_i^T$ .

## Related Work

Multiagent Reinforcement Learning has produced primarily two types of algorithms. One type learns some fixed point of the game e.g., NE (Minimax-Q (Littman 1994; Littman & Szepesvari 1996), Nash-Q (Hu & Wellman 1998; 2002), FFQ (Littman 2001)) or correlated equilibrium (CE-Q (Greenwald & Hall 2002)). These algorithms can guarantee a certain minimal expected payoff asymptotically, but it may be possible to guarantee higher payoff in certain situations if the learner is adaptive to the opponents' play, instead of learning the game solution alone. This brings us to the other type of learners that learn a best response to the opponents' actual play e.g., IGA (Singh, Kearns, & Mansour 2000), WoLF-IGA (Bowling & Veloso 2001; 2002), AWESOME (Conitzer & Sandholm 2003a). Since mutual best response is an equilibrium, two similar best responding players (such situations referred to as *self-play*) should be able to converge to an equilibrium. WoLF-IGA achieves this in  $2 \times 2$  games and AWESOME achieves it for arbitrary sized games. Simple Q-learning (Sutton & Burto 1998) is also capable of learning a best response to an arbitrary opponent's policy provided that latter is stationary. Nevertheless, a straightforward application of Q-learning has been shown to perform well in MAL problems (Tan 1993; Sen, Sekaran, & Hale 1994; Sandholm & Crites 1996; Claus & Boutilier 1998). There has also been some work on playing team games (where the game matrices of all agents are identical) (Claus & Boutilier 1998; Wang & Sandholm 2002) with stronger convergence guarantees owing to the correlation of the game matrices.

However, the existing literature in MAL seldom provides any performance guarantees *during* the learning process. One significant line of work with possible impact on MAL in this regard is that on *regret matching* learners. Algorithms have been proposed that achieve  $\lim_{T \rightarrow \infty} \frac{Rg_i^T}{T} = 0$  (called *no-regret* algorithms) but their convergence properties in policies are unknown (Auer *et al.* 1995; Fudenberg & Levine 1995; Freund & Schapire 1999; Littlestone & Warmuth 1994) or at best limited (Jafari *et al.* 2001). Thus we see that existing MAL algorithms either do not provide any performance assurances during learning or they do without any convergence assurances. Our goal is to achieve both. Recent work by Zinkevich (2003) shows that IGA (Singh, Kearns, & Mansour 2000) has a no-regret property (even with extension to larger games) but this algorithm is not guaranteed to converge to Nash policies in self-play. Even WoLF-IGA cannot guarantee convergence to NE in self-play in larger games (Bowling & Veloso 2002). A more recent work (Conitzer & Sandholm 2003b) establishes a bounded loss learning framework (BL-WoLF) in classes of

zero-sum games allowing an agent to learn the minimax solution of such a game with the least amount of cumulative loss against the worst possible opponent (one who knows the game beforehand). Though this does provide performance assurances during learning, the framework does not address general-sum games, or learning adaptive best responses.

There is a fundamentally different notion of equilibrium from a NE of the one-shot game that we address. This is Efficient Learning Equilibrium (ELE) (Brafman & Tennenholtz 2002a), where the players’ learning algorithms are required to reach an equilibrium in polynomial time, but which may not exist under imperfect monitoring settings. Though the computational complexity of a NE is an open problem, its advantage is its guaranteed existence. Also, we do not assume perfect monitoring for our work since the learner does not need to observe the opponents’ payoffs.

### Performance Bounded Multiagent Learning

No-regret property is of great interest for MAL domains since it provides a meaningful performance bound for a learner in a non-stationary environment. We use this notion of performance bound for our learning algorithm. Our goal is to present a MAL algorithm that satisfies the *rationality* and *convergence* criteria of (Bowling & Veloso 2002), i.e.

- Converges to the stationary best response against stationary opponents. This is a base case that ensures consistency of a learning algorithm against non-learners.
- Converges to the NE policy profile of the one-shot game in repeated self-play of a matrix game.

and in addition has the following properties

- Applies to arbitrary sized games, unlike IGA or WoLF-IGA.
- Achieves no-regret payoff against any opponent.

AWESOME (Conitzer & Sandholm 2003a) also has the first 3 properties but the case where the opponents are *neither stationary nor following the same algorithm as the learner* has not been dealt with explicitly. In (Conitzer & Sandholm 2003a), when such agents are encountered the learner will reset repeatedly each time starting off with playing its NE policy and gradually moving toward a best response until the next reset. This does not provide any guarantee that its average payoff will be “good” in such cases. The behavior of WoLF-IGA in such cases is also unspecified. We posit that a “good” payoff in such cases is the *no regret* payoff. We propose to ensure that in situations where the opponents are neither stationary, nor following the learner’s algorithm, at least the learner’s average payoff will approach its no-regret payoff. This property does not conflict with either rationality or convergence; in fact they are in agreement since best response payoff (achieved asymptotically in rational and convergent play) cannot be worse than no-regret payoff. We make the following assumptions for the current work,

1. that *either* the learner knows its own bounded game payoffs (like AWESOME) *or* the payoffs are unknown but non-negative and bounded, i.e., in the range  $[\underline{r}, \bar{r}]$  such that  $\underline{r} \geq 0$ . We call such games *non-negative games*. If

a learner knows its game payoffs but they can be negative, then he can always imaginably transform that game into a non-negative game, then compute policy using the transformed game and use it in playing the actual game. The expected payoffs will differ only by a constant and the no-regret property will remain unchanged.

2. that the agents can observe each other’s instantaneous policies and can use vanishing step sizes for policy improvement (similar to IGA and WoLF-IGA). It can also observe the expected payoffs of all of its actions at any iteration (trivially computable if game payoffs are known).
3. that the agents are given at the start, their portions of an equilibrium policy profile which they converge to if all the agents are following the same prescribed algorithm (similar to WoLF-IGA and AWESOME).

We write the probability of the  $j$ th action of the  $i$ th agent at time  $t$  as  $\pi_i^t(j)$  and the expected payoff of this action against the opponent’s current policy as  $V_i(j, \pi_{-i}^t)$  and note that

$$\sum_j \pi_i^t(j) V_i(j, \pi_{-i}^t) = V_i(\pi_i^t, \pi_{-i}^t). \quad (1)$$

### The ReDVaLeR Algorithm

We propose to use the *Replicator* (Fudenberg & Levine 1998) rule for policy update of our target algorithm with a WoLF-like modification, which we call ReDVaLeR (*Replicator Dynamics with a Variable Learning Rate*).

$$\begin{aligned} \pi_i^{t+1}(j) &= \pi_i^t(j) + \eta \pi_i^t(j) \times [l_i^t(j) V_i(j, \pi_{-i}^t) - \\ &\quad \sum_j l_i^t(j) \pi_i^t(j) V_i(j, \pi_{-i}^t)] \end{aligned} \quad (2)$$

for  $\eta$  being a vanishing step size and base condition:  $\pi_i^0(j) = \frac{1}{|A_i|}$ , learning rates  $l_i^t(j)$  are defined later. As  $\eta \rightarrow 0$ , the above difference equation yields the following differential equation<sup>1</sup> for the dynamics of the  $n$ -player system

$$\begin{aligned} \frac{d}{dt}(\pi_i^t(j)) &= \pi_i^t(j) \times [l_i^t(j) V_i(j, \pi_{-i}^t) - \\ &\quad \sum_j l_i^t(j) \pi_i^t(j) V_i(j, \pi_{-i}^t)] \end{aligned} \quad (3)$$

$j = 1 \dots |A_i|$ ,  $i = 1 \dots n$ . This is similar to the Replicator rule except for the learning rates,  $l$ . The Replicator Dynamics (RD) (Fudenberg & Levine 1998) have been extensively studied in population genetics and evolutionary game theory. It is known that the NE may not be asymptotically stable for RD in many games, e.g., RSP game in Table 1. A recent result in (Hart & Mas-Colell 2003) explains why. They show that if the dynamics are uncoupled (as in RD), i.e., if the agents do not use the opponents’ payoff information in its dynamics, then it is unable to overcome an “information barrier” and consequently unable to guarantee convergence. This also explains why IGA (Singh, Kearns,

<sup>1</sup>The RHS is discontinuous but we do not need to completely solve the system. Our approach is only to show that a chosen fixed point can be made asymptotically stable under the dynamics.

& Mansour 2000) fails to guarantee convergence to NE but WoLF-IGA succeeds. As in WoLF-IGA, the learning rates  $l$  may be one way to provide the necessary coupling, and we define  $l_i^t(j)$  for ReDVaLeR as

$$l_i^t(j) = \begin{cases} 1 & \text{if } \pi_{-i}^t \text{ is fixed} \\ \left\{ \begin{array}{l} 1 + \sigma & \text{if } \pi_i^t(j) < \pi_i^{*j} \\ 1 - \sigma & \text{when } \pi_i^t(j) \geq \pi_i^{*j} \end{array} \right\} & \text{otherwise} \end{cases} \quad (4)$$

where  $\pi_i^*$  is an NE policy. We typically require a constant  $0 < \sigma \ll 1$ . We note that when  $\pi_i^t(j) \geq \pi_i^{*j}$ ,  $\pi_i^t(j)V_i(j, \pi_{-i}^t) \geq \pi_i^{*j}V_i(j, \pi_{-i}^t)$  since  $V_i(j, \pi_{-i}^t) \geq 0 \forall j$ , which is similar to the situation described as *winning* in (Bowling & Veloso 2002), excepting that now it is defined for *each action*. Thus this scheme of variation in the learning rate is in the spirit of WoLF (Win or Learn Fast). Likewise, we call the situation  $\pi_i^t(j) \geq \pi_i^{*j}$  as *winning*, and the situation  $\pi_i^t(j) < \pi_i^{*j}$  as *losing*.

### Analysis of ReDVaLeR

For the sake of brevity, we write  $V_i(j, \pi_{-i}^t)$  simply as  $V_i^j$ . Let  $D_i(\tilde{\pi}_i, \pi_i^t)$  be the Kullback Leibler divergence between the  $i$ th agent's policy at time  $t$  and an arbitrary distribution  $\tilde{\pi}_i$ , given by

$$D_i(\tilde{\pi}_i, \pi_i^t) = \sum_j \tilde{\pi}_i(j) \log \left( \frac{\tilde{\pi}_i(j)}{\pi_i^t(j)} \right) \quad (5)$$

**Lemma 1** *The following holds,*

$$\frac{d}{dt}(D_i(\tilde{\pi}_i, \pi_i^t)) = \sum_j l_i^t(j) \pi_i^t(j) V_i^j - \sum_j l_i^t(j) \tilde{\pi}_i(j) V_i^j$$

**Proof** : Differentiating both sides of equation 5 and using equation 3 we get the result. ■

**Corollary 1** *If  $\pi_i^t$  follows RD instead of ReDVaLeR, then the derivative of the corresponding divergence measure can be given by  $\frac{d}{dt}(D_i^{RD}(\tilde{\pi}_i, \pi_i^t)) = \sum_j (\pi_i^t(j) - \tilde{\pi}_i(j)) V_i^j$*

The following theorem establishes the *rationality* property (Bowling & Veloso 2002) of ReDVaLeR.

**Theorem 1** *If the opponents are playing stationary policies, then the policy sequence of ReDVaLeR converges to the best response to the opponents' policy.*

**Proof** : Let  $\bar{\pi}_i$  be the best response of the  $i$ th agent to the opponents' fixed joint policy, given by  $\pi_{-i}$ . Then putting  $\bar{\pi}_i$  in place of the arbitrary policy in Lemma 1, we have

$$\begin{aligned} \frac{d}{dt}(D_i(\bar{\pi}_i, \pi_i^t)) &= \sum_j l_i^t(j) \pi_i^t(j) V_i^j - \sum_j l_i^t(j) \bar{\pi}_i(j) V_i^j \\ &= \sum_j \pi_i^t(j) V_i^j - \sum_j \bar{\pi}_i(j) V_i^j, \text{ by (4)} \\ &= V_i(\pi_i^t, \pi_{-i}) - V_i(\bar{\pi}_i, \pi_{-i}), \text{ by (1)} \\ &\leq 0, \text{ since } V_i(\pi_i^t, \pi_{-i}) \leq V_i(\bar{\pi}_i, \pi_{-i}) \end{aligned}$$

That  $V_i(\pi_i^t, \pi_{-i}) \leq V_i(\bar{\pi}_i, \pi_{-i})$  is evident since otherwise  $\pi_i^t$  would have been the best response to  $\pi_{-i}$ . Now since

the divergence measure ( $D_i$ ) is not strictly decreasing, it is possible that  $D_i(\bar{\pi}_i, \pi_i^t)$  converges to a non-zero positive value and thus the sequence  $\{\pi_i^t\}$  converges to a policy other than  $\bar{\pi}_i$ , say  $\pi_i'$ . But even in that case  $V_i(\pi_i', \pi_{-i})$  must be equal to  $V_i(\bar{\pi}_i, \pi_{-i})$ , which implies that  $\pi_i'$  must also be a best response. ■

The following theorem establishes that ReDVaLeR produces performance bounded learning.

**Theorem 2** *The following holds,*

$$\int_0^T V_i(\pi_i^t, \pi_{-i}^t) dt \geq \left( \frac{1-\sigma}{1+\sigma} \right) \max_{\pi_i} \int_0^T V_i(\pi_i, \pi_{-i}^t) dt - \frac{\log |A_i|}{(1+\sigma)}$$

*That is, as  $\sigma \rightarrow 0^+$ , the maximum regret of the  $i$ th agent playing against  $n - 1$  arbitrary opponents who play fixed sequences of non-stationary policies in the first  $T$  steps is  $Rg_i^T \leq \log |A_i|$ .*

**Proof** : Note that the integrals exist since  $V_i$  is continuous and bounded. We write  $D_i(\tilde{\pi}_i, \pi_i^0)$  i.e., the initial divergence as  $D_0$ . Integrating the expression in Lemma 1 in the given time range and noting that  $D_i(\tilde{\pi}_i, \pi_i^T) \geq 0$  we have

$$\begin{aligned} -D_0 &\leq \int_0^T \left( \sum_j l_i^t(j) \pi_i^t(j) V_i^j - \sum_j l_i^t(j) \tilde{\pi}_i(j) V_i^j \right) dt \\ &\leq (1+\sigma) \int_0^T V_i(\pi_i^t, \pi_{-i}^t) dt - \\ &\quad (1-\sigma) \int_0^T V_i(\tilde{\pi}_i, \pi_{-i}^t) dt, \text{ by (1), the} \\ &\quad \text{bounds on } l_i^t(j), \text{ and non-negative games} \\ &\leq (1+\sigma) \int_0^T V_i(\pi_i^t, \pi_{-i}^t) dt - \\ &\quad \max_{\pi_i} (1-\sigma) \int_0^T V_i(\pi_i, \pi_{-i}^t) dt \end{aligned}$$

since  $\tilde{\pi}_i$  was arbitrarily chosen. Rearranging and noting that  $D_0 \leq \log |A_i|$  if the initial policy is uniform, we get the result. ■

This result is obtained along the same line as the no-regret property of the multiplicative weight algorithm in (Freund & Schapire 1999). It tells us that the player can ensure a constant bound expected regret at any time provided it uses  $\sigma \rightarrow 0^+$  and that the average regret of the learner is  $\frac{Rg_i^T}{T} \leq \frac{\log |A_i|}{(1+\sigma)T}$ , thus ensuring no-regret asymptotically. Using the technique of Freund & Schapire (1999) we can extend this result to arbitrary adaptive (non-oblivious) opponents in a discrete version of ReDVaLeR (future work).

Finally the following theorem establishes the crucial *convergence* property (Bowling & Veloso 2002) of ReDVaLeR.

**Theorem 3** *When all the  $n$  agents are following ReDVaLeR algorithm, the sequence of their policies converge to their NE in non-negative games of any size, provided they choose their portions of the same equilibrium, and they all choose  $\sigma = 1$  (this requirement may be relaxable).*

**Proof :** The goal is to prove that players using the ReD-VaLeR algorithm on a repeated matrix game can converge to a NE of the one-shot game. That they need to choose their portions of the *same* equilibrium (for rule 4) is not really an extra burden, as discussed in (Conitzer & Sandholm 2003a), for agents sharing a single algorithm. We define the sum of the divergence measures of the  $n$  agents (from equation 5) from their respective equilibria  $\pi_i^*$ , for ReDVaLeR and Replicator Dynamics respectively as

$$S = \sum_{i=1}^n D_i(\pi_i^*, \pi_i^t) \text{ and } S^{RD} = \sum_{i=1}^n D_i^{RD}(\pi_i^*, \pi_i^t)$$

Note that just as  $D_i, D_i^{RD} \geq 0$  and differentiable at all times, so are  $S, S^{RD} \geq 0$ . The general strategy of the proof is to show that  $\frac{d}{dt}(S) < 0$  under appropriate assumptions at any time  $t$ . We accomplish this by comparing  $\frac{d}{dt}(S)$  with  $\frac{d}{dt}(S^{RD})$  at that time had RD been using the same instantaneous policies. Since  $S \geq 0$  and  $S = 0$  holds only when all agents have reached their respective equilibria,  $\frac{d}{dt}(S) < 0$  implies that  $S$  is a Lyapunov function<sup>2</sup> for the dynamic system, i.e., the system converges to the equilibrium. Now at any time  $t$  we consider two distinct cases

**Case 1 :**  $\frac{d}{dt}(S^{RD}) \leq 0$ . Then we have,

$$\begin{aligned} \frac{dS}{dt} &\leq \frac{d}{dt}(S) - \frac{d}{dt}(S^{RD}) \\ &= \sum_i \frac{d}{dt}(D_i(\pi_i^*, \pi_i^t)) - \sum_i \frac{d}{dt}(D_i^{RD}(\pi_i^*, \pi_i^t)) \\ &= \sum_i \sum_j (\pi_i^t(j) - \pi_i^*(j))(l_i^t(j) - 1)V_i^j \\ &= \sum_i \sum_{j, \pi_i^t(j) \geq \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j))(-\sigma)V_i^j \\ &\quad + \sum_i \sum_{j, \pi_i^t(j) < \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j))(\sigma)V_i^j \\ &= \sum_i (< 0) < 0. \end{aligned}$$

It can easily be seen why the  $\sum_j$  terms above must be strictly negative. If  $\pi_i^t \neq \pi_i^*$  then  $\exists j$  s.t.  $0 < \pi_i^t(j) < \pi_i^*(j)$ . The contribution of this term could be zero only if  $V_i^j = 0$ , otherwise this term contributes a strictly negative term to the sum. But note that the opponents are all using ReDVaLeR which allows only mixed policies while learning, and partially mixed or pure policies only in the limit. Given this constraint,  $V_i^j$  can be zero only if all the payoffs for action  $j$  in  $\mathbf{R}_i$  are zeroes. This means the  $j$ th action of the learner must be dominated by all other actions and strictly dominated by at least one other action (unless of course all the payoffs are zero, which is a degenerate case that we ignore). This means the equilibrium probability for the  $j$ th action must be zero, i.e.,  $\pi_i^*(j) = 0$ .

<sup>2</sup>This is true only for internal starting policies since all pure policies are fixed points of (3). So  $S$  cannot be global Lyapunov.

But this contradicts the fact that  $\pi_i^t(j) < \pi_i^*(j)$  for some  $\pi_i^t(j) > 0$ . Hence the contribution of term  $j$  in the  $\sum_j$  must be strictly negative, and consequently the entire sum must be strictly negative. Note that in this case, we do not need to assume  $\sigma = 1$ , but any  $0 < \sigma \ll 1$  will do, thus ensuring asymptotic no-regret payoffs by Theorem 2.

This case actually encompasses all non-negative games where NE is stable under RD but not asymptotically stable, including all non-negative  $2 \times 2$  games with a unique mixed equilibrium. For all other types of non-negative  $2 \times 2$  games, we also have convergence as simplifying the ReDVaLeR update rule (equation 3) for two action games shows that it is identical to a gradient ascent rule (albeit an extra multiplicative term involving the product of both action probabilities, which changes only the magnitude, not the direction of the gradient in such games), which we know converges in policies in these games (Singh, Kearns, & Mansour 2000). That result also extends to ReDVaLeR.

**Case 2 :**  $\frac{d}{dt}(S^{RD}) > 0$ . Then we have,

$$\begin{aligned} \frac{dS^{RD}}{dt} - \frac{dS}{dt} &= \sum_i \frac{d}{dt}(D_i^{RD}(\pi_i^*, \pi_i^t)) - \\ &\quad \sum_i \frac{d}{dt}(D_i(\pi_i^*, \pi_i^t)) \\ &= \sum_{ij} (\pi_i^t(j) - \pi_i^*(j))(1 - l_i^t(j))V_i^j \\ &= \sigma \sum_{ij} |\pi_i^t(j) - \pi_i^*(j)| V_i^j, \text{ by (4)} \\ &> \frac{d}{dt}(S^{RD}), \text{ if } \sigma = 1, \forall i \end{aligned}$$

This implies again that  $\frac{d}{dt}(S) < 0$ . The strict inequality in the last step is explained as in case 1. Thus we see that  $S$  is Lyapunov for the system of ReDVaLeR adaptation, though in case 2, it needs  $\sigma = 1$  which is counterproductive to Theorem 2. ■

When  $\sigma = 1$ , the player essentially *stops* learning (learning rate is 0) when it is winning but continues when losing (learning rate is 2). This principle has been explored before under the title of *Win Stay, Lose Shift* (Posch & Brannath 1997; Nowak & Sigmund 1993). We note that this assignment of learning rates ensures that KL divergence is Lyapunov, i.e., divergence decreases monotonically. However, monotonic convergence is not essential for convergence, and the original schedule of learning rates ( $0 < \sigma \ll 1$ ) may also produce convergence, though KL divergence may sometimes increase in that case. Figure 1 shows that in a non-negative version of the RSP (produced by adding +1 to all payoffs) game, where Case 2 always applies, convergence can be achieved in self-play using  $0 < \sigma \ll 1$ . However, in the Shapley game where no-regret algorithms usually cycle exponentially (Jafari *et al.* 2001), ReDVaLeR has been found to converge (not shown) in self-play only if  $\sigma \geq \frac{1}{3}$ .

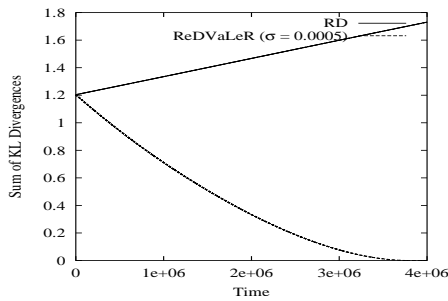


Figure 1: The sum of KL Divergences (S) in the non-negative version of RSP game of Table 1.

## Summary

This paper introduces a novel MAL algorithm, ReDVaLeR, which unlike the previous algorithms has a specific desirable property against opponents that are neither stationary nor using the same algorithm as the learner. It does not attempt to explicitly identify the opponents' algorithms, and as such are unaware of any desirable point of convergence (in policies) against such opponents. Therefore, instead of policy convergence, it achieves constant bounded regret at any time (no-regret payoff asymptotically) against such opponents, while preserving the earlier properties of convergence against stationary opponents and in self-play. We have established the following properties of ReDVaLeR in arbitrary sized general-sum non-negative games: (i) It converges to a stationary best response against an arbitrary number of stationary opponents, (ii) It achieves no-regret payoff against an arbitrary number of opponents playing arbitrary fixed sequences of non-stationary policies, and (iii) If all the players are following ReDVaLeR, then they converge to their portions of a NE, but one case requires a learning rate schedule that conflicts with property (ii). A simple experiment shows that even in this case, the learning rates sometimes can be such that convergence in self-play is maintained and no-regret payoff is achieved.

## References

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 1995. Gambling in a rigged casino: The adversarial multi-arm bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 322 – 331. Milwaukee, WI: IEEE Computer Society Press.

Bowling, M., and Veloso, M. 2001. Rational and convergent learning in stochastic games. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 1021 – 1026.

Bowling, M., and Veloso, M. 2002. Multiagent learning using a variable learning rate. *Artificial Intelligence*.

Brafman, R. I., and Tennenholtz, M. 2002a. Efficient learning equilibrium. In *Proceedings of Neural Information Processing Systems*.

Brafman, R. I., and Tennenholtz, M. 2002b. R-max - A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3:213 – 231.

Claus, C., and Boutilier, C. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the 15th National Conference on Artificial Intelligence*, 746–752. Menlo Park, CA: AAAI Press/MIT Press.

Conitzer, V., and Sandholm, T. 2003a. AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary

opponents. In *Proceedings of the 20th International Conference on Machine Learning*.

Conitzer, V., and Sandholm, T. 2003b. BL-WoLF: A framework for loss-bounded learnability in zero-sum games. In *Proceedings of the 20th International Conference on Machine Learning*.

Freund, Y., and Schapire, R. E. 1999. Adaptive game playing using multiplicative weights. *Games and Economic Behavior* 29:79 – 103.

Fudenberg, D., and Levine, D. 1995. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control* 19:1065 – 1089.

Fudenberg, D., and Levine, K. 1998. *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Greenwald, A., and Hall, K. 2002. Correlated q-learning. In *Proceedings of the AAAI Symposium on Collaborative Learning Agents*.

Hart, S., and Mas-Colell, A. 2003. Uncoupled dynamics do not lead to nash equilibrium. *American Economic Review*.

Hu, J., and Wellman, M. P. 1998. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proc. of the 15th Int. Conf. on Machine Learning (ML'98)*, 242–250. San Francisco, CA: Morgan Kaufmann.

Hu, J., and Wellman, M. 2002. Multiagent Q-learning. *Journal of Machine Learning*.

Jafari, A.; Greenwald, A.; Gondek, D.; and Ercal, G. 2001. On no-regret learning, fictitious play, and nash equilibrium. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 226 – 223.

Littlestone, N., and Warmuth, M. 1994. The weighted majority algorithm. *Information and Computation* 108:212 – 261.

Littman, M. L., and Szepesvari, C. 1996. A generalized reinforcement learning model: Convergence and applications. In *Proceedings of the 13th International Conference on Machine Learning*, 310 – 318.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proc. of the 11th Int. Conf. on Machine Learning*, 157–163. San Mateo, CA: Morgan Kaufmann.

Littman, M. L. 2001. Friend-or-foe Q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Nash, J. F. 1951. Non-cooperative games. *Annals of Mathematics* 54:286 – 295.

Nowak, M., and Sigmund, K. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* 364:56 – 58.

Posch, M., and Brannath, W. 1997. Win-stay, lose-shift. A general learning rule for repeated normal form games. In *Proceedings of the Third International Conference on Computing in Economics and Finance*.

Sandholm, T., and Crites, R. 1996. On multiagent Q-learning in a semi-competitive domain. In Weib, G., and Sen, S., eds., *Adaptation and Learning in Multi-Agent Systems*. Springer-Verlag, 191–205.

Sen, S.; Sekaran, M.; and Hale, J. 1994. Learning to coordinate without sharing information. In *National Conference on Artificial Intelligence*, 426–431. Menlo Park, CA: AAAI Press/MIT Press.

Singh, S.; Kearns, M.; and Mansour, Y. 2000. Nash convergence of gradient dynamics in general-sum games. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 541–548.

Sutton, R., and Burto, A. G. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

Tan, M. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, 330–337.

Wang, X., and Sandholm, T. 2002. Reinforcement learning to play an optimal nash equilibrium in team markov games. In *Advances in Neural Information Processing Systems 15, NIPS*.

Zinkevich, M. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*.