Unifying Convergence and No-regret in Multiagent Learning

Bikramjit Banerjee and Jing Peng

Department of Electrical Engineering & Computer Science
Tulane University
New Orleans, LA 70118
{banerjee, jp}@eecs.tulane.edu
http://www.eecs.tulane.edu/Banerjee

Abstract. We present a new multiagent learning algorithm, $RV_{\sigma(t)}$, that builds on an earlier version, ReDVaLeR . ReDVaLeR could guarantee (a) convergence to best response against stationary opponents and *either* (b) constant bounded regret against arbitrary opponents, *or* (c) convergence to Nash equilibrium policies in self-play. But it makes two strong assumptions: (1) that it can distinguish between self-play and otherwise non-stationary agents and (2) that all agents know their portions of the *same* equilibrium in self-play. We show that the adaptive learnng rate of $RV_{\sigma(t)}$that is explicitly dependent on time can overcome both of these assumptions. Consequently, $RV_{\sigma(t)}$theoretically achieves (a') convergence to near-best response against *eventually* stationary opponents, (b') no-regret payoff against arbitrary opponents *and* (c') convergence to some Nash equilibrium policy in some classes of games, in self-play. Each agent now needs to know its portion of *any* equilibrium, and does not need to distinguish among non-stationary opponent types. This is also the first successful attempt (to our knowledge) at convergence of a no-regret algorithm in the Shapley game.

1 Introduction

Multiagent learning (MAL) in a reinforcement learning setting has been an active field of study recently. The problem is simply of multiple controllers trying to learn individually "optimal" control policies in a shared Markov Decision Process (MDP), often called a *stochastic game* or a *Markov game*. The difficulty arises from the fact that all agents are learning simultaneously, which means the MDP faced by each agent is essentially *non-stationary*. Hence, the concept of "optimal" policy becomes ill-defined, and depends on the collective behavior of the other agents. Previous research has attempted to tackle this problem by considering various *opponent classes* such that there is a well-defined "optimal policy" for the learner *for each class of opponents*. Typically the research contributions in this aspect has been to design learning algorithms that learn the appropriate behaviors for the corresponding class of opponents, *without any access to the class information*.

The present paper follows this line of research and makes several fundamental contributions. In particular, we point out that the class taxonomy of the opponents suggested so far is incomplete. We then fill the void and present the first algorithm that can tackle

all opponent classes. More specifically, we present a new multiagent learning algorithm for repeated games, with the general philosophy of policy convergence against some classes of opponents but otherwise ensuring high payoffs. We build on our previous algorithm, ReDVaLeR [1], that we proved to guarantee (a) convergence to best response against stationary opponents and either (b) constant bounded regret against arbitrary opponents or (c) convergence to Nash equilibrium policies in self-play. It was shown to achieve both (b) and (c) empirically but needed to assume that all agents must know their portions of the *same* equilibrium. In this paper we present a new technique extending ReDVaLeR , called $RV_{\sigma(t)}$, that theoretically achieves (a') convergence to near-best response against *eventually* stationary opponents, (b') no-regret payoff against arbitrary opponents and (c') convergence to Nash equilibrium policies in some classes of games, in self-play. Each agent now needs to know only its portion of *any* equilibrium, besides the other assumptions made in ReDVaLeR . Additionally, since $RV_{\sigma(t)}$can achieve both (b') and (c') simultaneously, it does not need to distinguish between a self-play agent and an otherwise non-stationary agent.

No-regret has been an attractive property for a learner facing unknown opponents - the case that precludes any meaningful definition of a "desirable behavior" even for the agent designer. In such cases, no-regret stipulates a specific behavior sequence that achieves "safe" play in terms of payoffs, but otherwise does not attempt convergence to any specific behavior. However, depending on the opponents, we may want a no-regret learner to indeed converge to some policy, e.g., we would want it to converge to Nash equilibrium policy in self-play. Previous research [2] has empirically shown that in some games (such as the Shapley game in Table 1), no-regret learners are unable to converge in self-play. A major consequence of our theoretical results is that $RV_{\sigma(t)}$is *both no-regret and convergent* in self-play in some classes of games that includes the Shapley game. The rest of the paper is organized as follows: sections 2 and 3 present the background and the related work respectively. In section 4 we present the $RV_{\sigma(t)}$technique and in section 5, its analysis. We present our conclusions in section 6.

2 Multiagent Reinforcement Learning

A Multiagent Reinforcement Learning task is usually modeled 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 ith agent as A_i.

Definition 1 *A matrix game with n players is given by an n-tuple of matrices, $\langle \mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_n \rangle$ where $\mathbf{R}_i$ is a matrix of dimension $|A_1| \times |A_2| \ldots \times |A_n|$, such that the payoff of the ith agent for the joint action $(a_1, a_2, \ldots, a_n)$ is given by the entry $R_i(a_1, a_2, \ldots, a_n)$, $\forall i$.*

As is usual, we assume that payoffs are bounded, $R_i(a_1, a_2, \ldots, a_n) \in [\underline{r}_i, \bar{r}_i]$, for real $\underline{r}_i, \bar{r}_i$. Table 1 shows an example game of 2 players with 3 actions per player, called the Shapley game.

Table 1. The Shapley Game.

$$\mathbf{R}_1=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}, \qquad \mathbf{R}_2=\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix}$$

A *mixed policy*, vector $\boldsymbol{\pi}_i \in \Delta(A_i)$ for agent i, is a probability distribution over A_i. 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 ith agent will be given by the vector $\boldsymbol{\pi}_{-i}$. We shall usually refer to the ith 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 $\langle \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n \rangle$ is followed is given by $V_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}) = \sum_{(a_1,\ldots,a_n)\in\prod_k A_k} \pi_1(a_1)\ldots\pi_n(a_n)R_i(a_1,\ldots,a_n)$.

Definition 2 *For an n-player matrix game, an ϵ-best response ($BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}$) of the ith agent to its opponents' joint policy ($\boldsymbol{\pi}_{-i}$), for some $\epsilon \geq 0$, is given by*

$$BR^i_{\epsilon,\boldsymbol{\pi}_{-i}} = \{\boldsymbol{\pi}_i | V_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}) \geq V_i(\boldsymbol{\pi}'_i, \boldsymbol{\pi}_{-i}) - \epsilon,\ \forall \boldsymbol{\pi}'_i \in \Delta(A_i)\}$$

Definition 3 *A mixed-policy Nash Equilibrium (NE) for a matrix game $\langle \mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_n \rangle$ is a tuple of probability vectors $\langle \boldsymbol{\pi}^*_1, \boldsymbol{\pi}^*_2, \ldots, \boldsymbol{\pi}^*_n \rangle$ (policy profile) such that each is a best response to the rest, i.e., $\boldsymbol{\pi}^*_i \in BR^i_{\boldsymbol{\pi}^*_{-i}}\ \forall i$. In terms of payoffs, these conditions can be restated as*

$$V_i(\boldsymbol{\pi}^*_i, \boldsymbol{\pi}^*_{-i}) \geq V_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}^*_{-i})\ \forall \boldsymbol{\pi}_i \in \Delta(A_i)\ , \forall i$$

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 [3]. As an example, the only NE of the 2 player Shapley game in Table 1 is $\langle [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}], [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}] \rangle$.

Definition 4 *For a given time range $t = 0 \ldots T$, the regret of a learner (agent i), Rg^T_i is given by $Rg^T_i = \max_{\boldsymbol{\pi}_i} \sum_{t=1}^{t=T} V_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}^t_{-i}) - \sum_{t=1}^{t=T} V_i(\boldsymbol{\pi}^t_i, \boldsymbol{\pi}^t_{-i})$.*

This compares the total payoff of the actual sequence of policies of the learner with the best response to the empirical distribution of the opponent.

3 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 [4, 5], Nash-Q [6], FFQ [7]) or correlated equilibrium (CE-Q [8]). 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 [9], WoLF-IGA [10, 11], AWESOME [12]. 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 2×2 games (assuming it knows its portion of any equilibrium) and AWESOME achieves it for arbitrary sized games. But an AWESOME agent needs to know an entire equilibirum profile, meaning that it not only knows the others' equilibrium policy, but also that all agents agree on which equilibrium they know in games with multiple equilibria.

Performance guarantees *during* the learning process are provided by *regret matching* learners. These are algorithms that achieve $\lim_{T\to\infty} \frac{Rg_i^T}{T} \leq 0$ (called *no-regret* algorithms) but their convergence properties in policies are unknown [13–16] or at best limited [2]. A generalized version of IGA, called GIGA, was shown to be no-regret [17] but its convergence property is unknown. Clearly, there was a need for a MAL algorithm that could address (eventually) stationary opponents and self-play *as well as other types of opponents.* Our previous work on ReDVaLeR [1] filled this void and allowed a learner to be no-regret against this large class of opponents, in addition to satisfying the base cases againt stationary and self-play opponents. Subsequently, the WoLF version of GIGA [18] was shown to be also no-regret but convergent to NE only in 2×2 games against GIGA. Our previous algorithm, ReDVaLeR [1] also had limitations. It was shown to achieve both convergence and no-regret in arbitrary sized games, but for conflicting settings of a parameter σ. Each ReDVaLeR agent was also assumed to know its portion of *some* equilibrium, i.e., there was agreement on equilibrium selection. Our present work builds on ReDVaLeR and uses a single time dependent σ that achieves both convergence and no-regret properties simultaneously. More importantly, we relax the assumption that agents agree on which equilibrium they know their portions of. Another recent work proposed a similar set of properties for a MAL algorithm, with a greater focus on payoff [19]. This algorithm achieves near best response against stationary players (in contrast we guarantee near best response against the larger set of eventually stationary opponents), at least non-Pareto dominated (by another equilibrium) equilibrium payoff in self-play (in contrast we provide convergence to some equilibrium policy), and at least the minimax payoff against all other players (in contrast, we guarantee the stronger property of no-regret payoff that could be greater than the minimax payoff depending on the opponents) in polynomial time. However, this algorithm needs to know the game matrix of *all* agents which is stronger than our 2 assumptions combined that the learner knows only its own game matrix and its portion of any equilibrium policy.

4 Our Approach: ReDVaLeR with variable σ

We make the following assumptions for the current work,

1. that the learner knows its own bounded game payoffs (like AWESOME)
2. that the agents can observe each other's instantaneous policies [1] and can use vanishing step sizes for policy improvement (similar to IGA and WoLF-IGA).

[1] This assumption is only used in Theorem 10 but it is dispensable. It is possible to collapse this and the previous assumption into a single assumption that the learner can *only* observe its payoff vector at every round, as in [18]. We shall show the details in a consolidated version.

3. that the agents are given at the start, their portions of *any* equilibrium policy profile (like WoLF-IGA). They might see their portions of different equilibria in games with multiple equilibria.

We write the probability of the jth action of the ith agent at time t as $\pi_i^t(j)$ and the expected payoff of this action against the opponent's current policy as $V_i(j, \boldsymbol{\pi}^t_{-i})$ and note that $\sum_j \pi_i^t(j) V_i(j, \boldsymbol{\pi}^t_{-i}) = V_i(\boldsymbol{\pi}^t_i, \boldsymbol{\pi}^t_{-i})$.

We use the ReDVaLeR algorithm (see [1] for details) with a time-varying schedule for σ in place of a constant. The discrete form of the algorithm (slightly different from [1]) is

$$\pi_i^{t+1}(j) = \frac{\pi_i^t(j) + \eta \pi_i^t(j) l_i^t(j) V_i(j, \boldsymbol{\pi}^t_{-i})}{1 + \eta \sum_j l_i^t(j) \pi_i^t(j) V_i(j, \boldsymbol{\pi}^t_{-i})} \tag{1}$$

for η being a small step size and initial condition: $\pi_i^0(j) = \frac{1}{|A_i|}$. Note that the probability values generated above are automatically bounded in the range $[0, 1]$ if $\underline{r}_i \geq 0$. Also the distribution is indeed a probability distribution (i.e., sum 1) without the need of a projection operation unlike GIGA or GIGA-WoLF.

In continuous time, i.e., as $\eta \to 0$, the above equation yields the same differential equation as ReDVaLeR for the dynamics of the n-player system

$$\frac{d}{dt}(\pi_i^t(j)) = \pi_i^t(j) \times [l_i^t(j) V_i(j, \boldsymbol{\pi}^t_{-i}) - \sum_j l_i^t(j) \pi_i^t(j) V_i(j, \boldsymbol{\pi}^t_{-i})] \tag{2}$$

$j = 1 \ldots |A_i|,\ i = 1 \ldots n$. In contrast with [1], the learning rates ($l_i^t(j)$) for this algorithm, $RV_{\sigma(t)}$, are defined as

$$l_i^t(j) = \begin{cases} 1 + \sigma(t) \text{ if } \pi_i^t(j) < \pi_i^*(j) \\ 1 - \sigma(t) \text{ if } \pi_i^t(j) \geq \pi_i^*(j) \end{cases} \tag{3}$$

for a suitable σ-schedule as defined below.

Definition 5 (σ-Schedule) *A time decaying schedule for $\sigma(t)$ is defined by the 3 conditions:*

1. *$\sigma(t)$ is continuous and $1 \geq \sigma(t) \geq 0, \forall t$,*
2. *$\sigma(t) \geq \sigma(t'), \forall t' \geq t$,*
3. *$\sigma(t) \to 0$ as $t \to \infty$.*

5 Analysis of $RV_{\sigma(t)}$

In the following analysis, we shall use the symbol $\|\mathbf{x}\|$ to mean the $\mathcal{L}_\infty$ norm of a vector, i.e., $\|\mathbf{x}\| = \max_i |x_i|$ and the symbol $\mathbf{1}$ to mean a vector of all 1's. We also assume that the game payoffs are all positive (i.e., $\underline{r}_i \geq 0$), which is merely a technical assumption, since if the agent knows its game payoffs it can easily make an affine transformation to satisfy this assumption. The new game is strategically unchanged and the no-regret property also holds in the original game.

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

$$D_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}^t_i) = \sum_j \tilde{\pi}_i(j) \log \left(\frac{\tilde{\pi}_i(j)}{\pi^t_i(j)} \right) \tag{4}$$

With a slight abuse of notation, we will refer to $\dot{D}_i$ or $\frac{dD_i}{dt}$ as the projection of the gradient of the function D_i (equation 4) along the solution trajectory of (2) for a given initial policy profile. When the trajectory follows the unmodified Replicator Dynamics, we write the same as $\frac{dD^{RD}_i}{dt}$. The following result is crucial to all subsequent analyses.

Lemma 1 [1] *For any fixed policy $\tilde{\boldsymbol{\pi}}_i$,*

$$\frac{d}{dt}(D_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}^t_i)) = \sum_j l^t_i(j)\pi^t_i(j)V^j_i - \sum_j l^t_i(j)\tilde{\pi}_i(j)V^j_i$$

$$\frac{d}{dt}(D^{RD}_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}^t_i)) = V_i(\boldsymbol{\pi}^t_i, \boldsymbol{\pi}^t_{-i}) - V_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}^t_{-i})$$

5.1 Convergence against eventually stationary opponents

Here we establish that $RV_{\sigma(t)}$with σ-schedule of definition 5 converges to the set of ϵ-best responses against stationary opponents from which follow the convergence against eventually stationary opponents. The following lemma is used and is straightforward to prove.

Lemma 2 (Payoff-continuity) *If $\boldsymbol{\pi}_{i1}$ and $\boldsymbol{\pi}_{i2}$ are two policy vectors of agent i against the stationary joint policy of the opponents $\boldsymbol{\pi}_{-i}$ and if $\|\boldsymbol{\pi}_{i1} - \boldsymbol{\pi}_{i2}\| \leq \alpha$ for some $\alpha > 0$, then*

$$|V_i(\boldsymbol{\pi}_{i1}, \boldsymbol{\pi}_{-i}) - V_i(\boldsymbol{\pi}_{i2}, \boldsymbol{\pi}_{-i})| \leq \alpha |A_i| \bar{r}_i$$

In other words, if two policies are close then so are their payoffs against a given joint policy of the opponents.

The following Theorem establishes that $RV_{\sigma(t)}$with a non-stationary σ (Definition 5) converges to the set of ϵ-best responses against stationary opponents.

Theorem 3 *For a given $\epsilon > 0$, there exists a time τ, such that after τ a $RV_{\sigma(t)}$agent i using σ-schedule in Definition 5 against $n-1$ stationary agents, is guaranteed to converge to the set of ϵ-best response policies, $BR^i_{\epsilon, \boldsymbol{\pi}_{-i}}$.*

Proof : Suppose the opponents' joint stationary policy is given by $\boldsymbol{\pi}_{-i}$, and let us consider some $\bar{\boldsymbol{\pi}}_i \in BR^i_{0, \boldsymbol{\pi}_{-i}}$. Clearly the payoff of all policies in $BR^i_{0, \boldsymbol{\pi}_{-i}}$ have the same value and let this value be $\bar{V}_i$. At any given time t we consider the following two cases:

Case 1 : $\boldsymbol{\pi}_i^t \notin BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}$

This means

$$V(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}) < \bar{V}_i - \epsilon \tag{5}$$

Now substituting $\bar{\boldsymbol{\pi}}_i$ in place of the arbitrary policy in Lemma 1, we get

$$\begin{aligned}
\frac{d}{dt}(D_i(\bar{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_i^t)) &= \sum_j l_i^t(j)\pi_i^t(j)V_i^j - \sum_j l_i^t(j)\bar{\boldsymbol{\pi}}_i(j)V_i^j \\
&\leq (1+\sigma(t))V_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}) - (1-\sigma(t))V_i(\bar{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i}) \\
&= V_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}) - V_i(\bar{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i}) + \sigma(t)[V_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}) + V_i(\bar{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i})] \\
&< -\epsilon + \sigma(t)[V_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}) + V_i(\bar{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i})] \text{ , by (5)} \\
&\leq -\epsilon + 2\sigma(t)\bar{V}_i
\end{aligned}$$

Now according to Definition 5 there exists a time (τ) such that for all $t' > \tau$, $\sigma(t') < \frac{\epsilon}{2\bar{V}_i}$. Thus at all such times $\frac{d}{dt}(D_i(\bar{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_i^{t'})) < 0$ whenever $\boldsymbol{\pi}_i^{t'} \notin BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}$. This means, the policy approaches a best response at such times. By Lemma 2, the policy cannot approach a best response without its payoff approaching $\bar{V}_i$. Thus at some point t', the value of the policy will exceed $\bar{V}_i - \epsilon$ and so, $\boldsymbol{\pi}_i^{t'} \in BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}$. This brings us to the 2nd case.

Case 2 : $\boldsymbol{\pi}_i^t \in BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}$. Also $t \geq \tau$

If $\boldsymbol{\pi}_i^{t'} \in BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}, \forall t' > t$ then we are done. Otherwise, there exists a time $t' > t$ such that $\boldsymbol{\pi}_i^{t'-\eta} \in BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}$ and $\boldsymbol{\pi}_i^{t'} \notin BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}$, where η is the time step size used in Equation 1. So $V_i(\boldsymbol{\pi}_i^{t'-\eta}, \boldsymbol{\pi}_{-i}) \geq V_i(\boldsymbol{\pi}_i^{t'}, \boldsymbol{\pi}_{-i})$. Also

$$\begin{aligned}
\|\boldsymbol{\pi}_i^{t'-\eta} - \boldsymbol{\pi}_i^{t'}\| &\leq \eta(1+\sigma(t'-\eta))\bar{r}_i \\
&\leq 2\eta\bar{r}_i
\end{aligned}$$

Then by Lemma 2,

$$V_i(\boldsymbol{\pi}_i^{t'-\eta}, \boldsymbol{\pi}_{-i}) - V_i(\boldsymbol{\pi}_i^{t'}, \boldsymbol{\pi}_{-i}) \leq 2\eta|A_i|\bar{r}_i^2$$

that is

$$V_i(\boldsymbol{\pi}_i^{t'}, \boldsymbol{\pi}_{-i}) \geq \bar{V}_i - (\epsilon + 2\eta|A_i|\bar{r}_i^2)$$

So even though $\boldsymbol{\pi}_i^{t'}$ is not an ϵ-best response, it is an $(\epsilon + 2\eta|A_i|\bar{r}_i^2)$-best response. Also from time t' case 1 applies and both of the policy and the payoff approach that of a strict best response. Thus after τ, the payoff never falls below $\bar{V}_i - (\epsilon + 2\eta|A_i|\bar{r}_i^2) = \bar{V}_i - \epsilon$, since $\eta \to 0$. Lastly, $\boldsymbol{\pi}_i^t$ may not converge to any specific policy in $BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}$, only stay in this set asymptotically. □

An immediate corollary of Theorem 3 is that $RV_{\sigma(t)}$will converge to an ϵ-best response even if the opponents do not always play stationary policies, as long as they *settle down to a stationary profile at some finite time*, τ_1. This is justified by replacing τ in the proof of Theorem 3 by $\max\{\tau, \tau_1\}$. We state this result as the following corollary.

Corollary 4 *If there exists a time τ_1 such that all other agents play stationary policies at all times $t > \tau_1$, then for a given $\epsilon > 0$, there exists a time τ, such that after $\max\{\tau_1, \tau\}$ an $RV_{\sigma(t)}$agent i using σ-schedule in definition 5, is guaranteed to converge to the set of ϵ-best response policies, $BR^i_{\epsilon,\boldsymbol{\pi}_{-i}}$.*

Note that this does not require all of the opponents to start playing a stationary profile simultaneously, only that the last opponent to settle down should do so at some finite time point τ_1. Also note that this notion of eventually stationary opponent profile is a stronger condition than the non-stationary opponent policies *with a limit* considered in [20]. In the latter the opponents may never actually settle down but continue with an ever decreasing distance from a limiting profile.

5.2 No-regret property

Here we prove the no-regret property of $RV_{\sigma(t)}$. Compared to ReDVaLeR , now the regret is no longer constant bounded but can grow with time. However, with the help of the following lemma (stated without proof) we can show that the average regret goes to 0.

Lemma 5 (Vanishing average) *Given definition 5 for $\sigma(t)$, we have* $\lim_{T\to\infty} \frac{\int_0^T \sigma(t)dt}{T} = 0$.

Theorem 6 *If a $RV_{\sigma(t)}$agent i uses the decaying σ-schedule of definition 5, then*

$$\lim_{T\to\infty} \frac{Rg_i^T}{T} \leq 0$$

i.e., the algorithm is asymptotically no-regret.

Proof : As in Theorem 2 in [1], we have

$$\begin{aligned}
-D_0 &\leq \int_0^T \left(\sum_j l_i^t(j)\pi_i^t(j)V_i^j - \sum_j l_i^t(j)\tilde{\pi}_i(j)V_i^j \right) dt \\
&\leq \int_0^T (1+\sigma(t))V_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}^t)dt - \int_0^T (1-\sigma(t))V_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i}^t)dt \\
&= \int_0^T V_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}^t)dt - \int_0^T V_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i}^t)dt \\
&\quad + \int_0^T \sigma(t) \left[V_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}^t) + V_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i}^t) \right] dt
\end{aligned}$$

Rearranging and again noting that $D_0 \leq \log|A_i|$ and that $\tilde{\boldsymbol{\pi}}_i$ was chosen arbitrarily, we have

$$\int_0^T V_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}^t)dt \geq \max_{\tilde{\boldsymbol{\pi}}_i} \int_0^T V_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i}^t)dt - \log|A_i|$$

$$
\begin{aligned}
&- \int_0^T \sigma(t) \left[V_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}^t) + V_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i}^t) \right] dt \\
&\geq \max_{\tilde{\boldsymbol{\pi}}_i} \int_0^T V_i(\tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i}^t) dt - 2\bar{r}_i \int_0^T \sigma(t) dt - \log |A_i|
\end{aligned}
$$

Thus the regret of the ith agent is bounded by

$$
Rg_i^T \leq 2\bar{r}_i \int_0^T \sigma(t) dt + \log |A_i|
$$

The result now follows from Lemma 5. □

We postpone the choice of actual form of $\sigma(t)$ till the end of section 5.3 in order to satisfy both convergence and no-regret, whereby we also compare the emerging expression of regret with those from GIGA, GIGA-WoLF.

5.3 Convergence in Self-play

Since we do not assume any coordination in the choice of the equilibrium, for games with multiple equilibria, the agents may be given their portions of different equilibria. Although this is not difficult to handle in 2×2 games [11], in larger games this becomes a daunting task. In this paper we show that the variable learning rate is useful for this purpose, in games *of any size* but with a unique mixed equilibrium. Even though coordination in equilibrium selection is by default in such games, it has proven to be a hard case for convergence in self-play beyond 2×2 games. $RV_{\sigma(t)}$is the first algorithm that extends this property to such games of arbitrary size, and this is also experimentally validated in two such games, viz., the Shapley game (Table 1) and the game in Table 2 with a unique partially mixed equilibrium. This is addition to $RV_{\sigma(t)}$being convergent in *all* 2×2 games with possibly multiple equilibria, which we show next.

2×2 games In all 2×2 games, the $RV_{\sigma(t)}$algorithm can be shown to be equivalent to WoLF-IGA. In cases where IGA converges (in policy) in self-play, only the direction of the gradient matters and this remains same for $RV_{\sigma(t)}$. In the special case where WoLF-IGA (but not IGA) converges in policy, the learning rate change in $RV_{\sigma(t)}$turns out to be the same as WoLF-IGA thus guaranteeing convergence like WoLF-IGA. Hence $RV_{\sigma(t)}$always converges to an equilibrium policy in all 2×2 games, when given its portion of *any* equilibrium, similar to WoLF-IGA.

Games with unique mixed equilibrium Here we prove that a σ-schedule can be designed satisfying definition 5 such that convergence to equilibrium can be achieved in these games. We make another technical assumption, that the minimum game payoff of i is strictly positive, $\underline{r}_i > 0$, for all i. Again this is easy to satisfy in self-play without changing the game strategically. The following lemmas will be used in the proof of the final Theorem for convergence of $RV_{\sigma(t)}$in self-play.

As a first step we show that the requirement on the value of σ (i.e., $\sigma = 1$; Theorem 3 in [1]) from the perspective of any learner i can be relaxed in two ways. The first is a direct but minor relaxation given by the lemma below.

Lemma 7 *If the policy of i is not ϵ_i-close to its equilibrium, i.e., $\min_j |\pi_i^t(j) - \pi_i^*(j)| > \epsilon_i$, for some $\epsilon_i > 0$, then $\frac{d}{dt} D_i(\boldsymbol{\pi}_i^*, \boldsymbol{\pi}_i^t) < -\alpha$ for some $0 < \alpha < \epsilon_i \underline{r}_i$, if i uses*

$$\sigma > \frac{1}{1 + \frac{\epsilon_i \underline{r}_i - \alpha}{\bar{r}_i + \alpha}}.$$

Proof : The proof closely follows Theorem 3 in [1]. In case 1 of that proof, σ only needs to be positive. It is really case 2 that needs to be relaxed. It is easily seen that the proof of case 2 in that theorem can be stated for individual agents as well. Consequently when $\frac{dD_i^{RD}}{dt} > 0$, we have

$$\frac{dD_i^{RD}}{dt} - \frac{dD_i}{dt} = \sigma \sum_j \left| \pi_i^t(j) - \pi_i^*(j) \right| V_i^j \tag{6}$$

Now since $\min_j |\pi_i^t(j) - \pi_i^*(j)| > \epsilon_i$, there is at least one action, say k, such that $\pi_i^t(k) < \pi_i^*(k)$. Therefore, $|\pi_i^t(k) - \pi_i^*(k)| V_i^k \geq \epsilon_i \underline{r}_i$. So,

$$\begin{aligned}
\frac{dD_i^{RD}}{dt} &= \sum_{j \neq k} (\pi_i^t(j) - \pi_i^*(j)) V_i^j + (\pi_i^t(k) - \pi_i^*(k)) V_i^k \\
&= \sum_{j \neq k} (\pi_i^t(j) - \pi_i^*(j)) V_i^j - |\pi_i^t(k) - \pi_i^*(k)| V_i^k \\
&\leq \sum_{j \neq k} (\pi_i^t(j) - \pi_i^*(j)) V_i^j - \epsilon_i \underline{r}_i
\end{aligned}$$

Hence,

$$\sum_{j \neq k} (\pi_i^t(j) - \pi_i^*(j)) V_i^j \geq \frac{dD_i^{RD}}{dt} + \epsilon_i \underline{r}_i \tag{7}$$

Equation 6 gives us $\frac{dD_i^{RD}}{dt} - \frac{dD_i}{dt} \geq \sigma \sum_{j \neq k} (\pi_i^t(j) - \pi_i^*(j)) V_i^j$. Substituting from Equation 7, we have $\frac{dD_i^{RD}}{dt} - \frac{dD_i}{dt} \geq \sigma (\frac{dD_i^{RD}}{dt} + \epsilon_i \underline{r}_i)$. The result follows noting that $\frac{dD_i^{RD}}{dt} \leq \bar{r}_i$. □

To illustrate the nature of this relaxation, if $\epsilon_i = 2 \times 10^{-3}$, $\bar{r}_i = 2$, $\underline{r}_i = 1$, and $\alpha = 10^{-3}$, then we have $\sigma > 0.9995$. The main Theorem on convergence of $RV_{\sigma(t)}$in self-play, however, depends on this to be maintained for only a finite time ($O(\frac{1}{\alpha})$).

The following lemma allows σ to approach 0 in self-play, but applies only when the others are *sufficiently close* to their portions of the equilibrium. First we define "sufficiently close". Let us write the vector of ith agents payoff, V_i^j over index j ($j \in A_i$), as $\mathbf{V}_i$. Also from Game Theory [21] we know that for a non-negative game with a unique completely mixed equilibrium there is a constant $V_i^* > 0$ for each i such that, $V_i(j, \boldsymbol{\pi}_{-i}^*) = V_i^*, \ \forall j$. Clearly when the opponents' policies are close to their respective equilibria, $V_i^j = V_i(j, \boldsymbol{\pi}_{-i}^t)$ is also close to $V_i(j, \boldsymbol{\pi}_{-i}^*) = V_i^*$, since payoffs are bounded.

Definition 6 *The opponents of agent i are said to be* sufficiently close *to their equilibria if $\|\mathbf{V}_i - V_i^*\mathbf{1}\| < V_i^*$, and this distance does not exceed V_i^* at any future time.*

Now the following Lemma relaxes the the value of σ from 1 when the opponents are *sufficiently close* to their equilibria.

Lemma 8 *In self-play in non-negative games with a unique completely mixed equilirium, when the opponents of agent i are sufficiently close to their equilibria, the value of σ used by i need only satisfy $\sigma(t) > \frac{\|\mathbf{V}_i - V_i^*\mathbf{1}\|}{V_i^*}$ to ensure convergence of $\boldsymbol{\pi}_i^t$ to $\boldsymbol{\pi}_i^*$.*

Proof : Let us call $c = \|\mathbf{V}_i - V_i^*\mathbf{1}\|$. Note that under the given conditions, $V_i^* - c \leq V_i^j \leq V_i^* + c, \forall j$. Then the rate of variation in $D_i(\boldsymbol{\pi}_i^*, \boldsymbol{\pi}_i^t)$ can be given as before by,

$$
\begin{aligned}
\frac{dD_i}{dt} &= \sum_j (\pi_i^t(j) - \pi_i^*(j)) l_i^t(j) V_i^j \\
&= \sum_{j:\pi_i^t(j) \geq \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j))(1-\sigma(t))V_i^j + \sum_{j:\pi_i^t(j) < \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j))(1+\sigma(t))V_i^j \\
&\leq \sum_{j:\pi_i^t(j) \geq \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j))(1-\sigma(t))(V_i^* + c) \\
&\quad + \sum_{j:\pi_i^t(j) < \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j))(1+\sigma(t))(V_i^* - c) \\
&= (1-\sigma(t))(V_i^* + c) \sum_{j:\pi_i^t(j) \geq \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j)) \\
&\quad + (1+\sigma(t))(V_i^* - c) \sum_{j:\pi_i^t(j) < \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j)) \\
&= (1-\sigma(t))(V_i^* + c) \sum_{j:\pi_i^t(j) \geq \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j)) \\
&\quad - (1+\sigma(t))(V_i^* - c) \sum_{j:\pi_i^t(j) \geq \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j)) \\
&= \left[\sum_{j:\pi_i^t(j) \geq \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j)) \right] [(1-\sigma(t))(V_i^* + c) - (1+\sigma(t))(V_i^* - c)]
\end{aligned}
$$

The equality of the last but one step follows from the fact that $\sum_{j:\pi_i^t(j) \geq \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j)) + \sum_{j:\pi_i^t(j) < \pi_i^*(j)} (\pi_i^t(j) - \pi_i^*(j)) = 0$. Since the factor in the first square braces in the last step is strictly positive, the only situation when this is strictly negative is when $\sigma(t) > \frac{c}{V_i^*}$. This makes D_i Lyapunov implying convergence to $\boldsymbol{\pi}_i^*$. □

Interestingly, if all agents are *sufficiently close* to their equilibria and all use the σ as in Lemma 8, then all of them will converge to their respective equilibria. This means for each i, $\|\mathbf{V}_i - V_i^*\mathbf{1}\|$ will decrease and that agent will be able to further decrease its σ with time while satisfying Lemma 8. The key is to get them *sufficiently close* to their equilibria. We show how in the next Theorem but before that we state one last necessary Lemma.

Lemma 9 (KLD-$\mathcal{L}_\infty$ correspondence) *For any two probability distributions,* $\mathbf{x}, \mathbf{y}$*, we have* $\|\mathbf{x} - \mathbf{y}\| \leq \epsilon$ *for some* $1 > \epsilon > 0$ *if* $D(\mathbf{x}, \mathbf{y}) \leq \frac{2\epsilon^2}{\log 2}$.

The following Theorem establishes the convergence of $RV_{\sigma(t)}$to Nash equilibrium in self-play under appropriate assumptions.

Theorem 10 *There exists a* σ*-schedule satisfying definition 5, which when followed by* n $RV_{\sigma(t)}$*agents guarantees the convergence of their policies to the unique completely mixed equilibrium profile of the strictly positive game, provided each agent knows*

1. *the maximum game payoff of any agent,* $R_{\max} = \max_i \bar{r}_i$*,*
2. *the maximum size of action space among all agents,* $\max_i |A_i|$*,*
3. *the minimum equilibrium payoff among all agents,* $\min_i V_i^*$
4. *the total number of agents,* n*.*

Proof : The proof is stated in two steps. In step 1, we establish how agents can get *sufficiently close* to their equilibria. In step 2, we show how they can continue to approach their equilibria in self-play satisfying condition 3 of Definition 5.

Step 1: For each i, we need the opponents $(-i)$ to be *sufficiently close* to their equilibria. Now any agent p can make $\|\boldsymbol{\pi}_p^t - \boldsymbol{\pi}_p^*\| \leq \delta_p$ for some δ_p by using $\sigma_p > \frac{1}{1+\frac{\epsilon_p \underline{r}_p - \alpha_p}{\bar{r}_p + \alpha_p}}$ (Lemma 7) for sufficiently long (say τ) to bring $D_p(\boldsymbol{\pi}_p^*, \boldsymbol{\pi}_p^t)$ down from initial value $D_p(\boldsymbol{\pi}_p^*, \boldsymbol{\pi}_p^0) = \log |A_p| + \sum_j \pi_p^*(j) \log \pi_p^*(j)$ to $D_p(\boldsymbol{\pi}_p^*, \boldsymbol{\pi}_p^\tau) \leq \frac{2\delta_p^2}{\log 2}$ (Lemma 9) at the rate of α_p (Lemma 7). Therefore,

$$\tau \geq \frac{D_p(\boldsymbol{\pi}_p^*, \boldsymbol{\pi}_p^0) - D_p(\boldsymbol{\pi}_p^*, \boldsymbol{\pi}_p^\tau)}{\alpha_p}$$

and this can be easily computed. Note that agent p can also compute appropriate ϵ_p and α_p since it has the knowledge of the necessary policies, $\boldsymbol{\pi}_p^t$ and $\boldsymbol{\pi}_p^*$.
Now if $\|\boldsymbol{\pi}_p^t - \boldsymbol{\pi}_p^*\| \leq \delta_p \; \forall p \in \{-i\}$, then $\|\boldsymbol{\pi}_{-i}^t - \boldsymbol{\pi}_{-i}^*\| \leq \sum_p \delta_p$ approximately (ignoring the terms in second and higher powers of δ). As a consequence, i's opponents will be *sufficiently close* to their equilibria if

$$\begin{aligned} \|\mathbf{V}_i - V_i^* \mathbf{1}\| &\leq \max_j |V_i^j - V_i^*| \\ &\leq |A_i| \bar{r}_i \|\boldsymbol{\pi}_{-i}^t - \boldsymbol{\pi}_{-i}^*\| \\ &\leq |A_i| \bar{r}_i \sum_p \delta_p \end{aligned}$$

is less than V_i^*. This can be ensured for all agents p, by forcing

$$\delta_p \leq \frac{\min_i V_i^*}{n R_{\max} \max_i |A_i|}$$

Hence the conditions in the Theorem statement. Thus all agents can be brought *sufficiently close* to their equilibria by some σ-schedule following Definition 5.

Step 2: After τ, each agent i must always satisfy Lemma 8. Since the starting value ($\sigma(\tau)$) has been specified in **Step 1**, i only needs to know an appropriate $\frac{d\sigma}{dt}$ to keep changing its σ satisfying Lemma 8. It is easy to see that a suitable $\frac{d\sigma}{dt}$ is

$$0 > \frac{d\sigma}{dt} > \left(\frac{-1}{V_i^*}\right) \max_j \left| \sum_{a_{-i}} R_i(j, a_{-i}) \frac{d}{dt} \left(\pi_{-i}^t(a_{-i})\right) \right| \tag{8}$$

where a_{-i} is a joint action played by i's opponents. The appropriate rate in (8) can be computed from i's observation of its opponents' policies at all times. Also since (8) requires $\frac{d\sigma}{dt}$ be always negative after τ, Definition 5 is satisfied. This completes the proof. □

Note that while in self-play (8) will lead $\frac{d\sigma}{dt}$ to approach 0 from below as $t \to \infty$, if the opponents are not self-play $\frac{d\sigma}{dt}$ may not approach 0. But since $\frac{d\sigma}{dt}$ is negative, the no-regret property (Theorem 6) will be preserved if we make $\left|\frac{d\sigma}{dt}\right|$ explicitly decay with time. A sample schedule that does this and satisfies (8) is (for $t \geq 1$)

$$\frac{d\sigma}{dt} = \left(\frac{-1}{V_i^* \sqrt{t}}\right) \max_j \left| \sum_{a_{-i}} R_i(j, a_{-i}) \frac{d}{dt} \left(\pi_{-i}^t(a_{-i})\right) \right| \tag{9}$$

Thus a $RV_{\sigma(t)}$agent can use the above σ-schedule for convergence to equilibrium in self-play while being oblivious of the nature of the others. In case the others are not self-play agents, the same schedule will guarantee the results of Theorem 3, Corollary 4 and Theorem 6.

Table 2. A 3 actions game with lone mixed equilibrium.

$$\mathbf{R}_1 = \begin{pmatrix} 1 & 3 & 1 \\ 1 & 10 & 1 \\ 5 & 1 & 2 \end{pmatrix}, \qquad \mathbf{R}_2 = \begin{pmatrix} 7 & 1 & 1 \\ 1 & 0 & 1 \\ 10 & 15 & 1 \end{pmatrix}$$

Theorem 10 basically says that if σ decays slow enough, then monotonic convergence of the sum of KL divergences can be achieved in self-play. For the following experiments we use a σ-schedule $\sigma(t) = \frac{1}{1+\beta\sqrt{t}}$ (of the form of (9)) and show the results for various values of β in Figure 1 corresponding to the Shapley game (Table 1) and the game in Table 2 respectively. We used $\eta = 2 \times 10^{-4}$ and the starting policies were selected close to the edges of the probability simplex since these are the policies that make convergence most difficult in RD. Note that the game in Table 1 is not strictly positive, and that in Table 2 is not strictly positive for the column agent. Also note that σ does not really need to be close to 1 as long as step 1 of Theorem 10 requires. In both experiments, in just 2000 iterations σ climbs down to less than 75% for the middle values of β as shown in Figure 1.

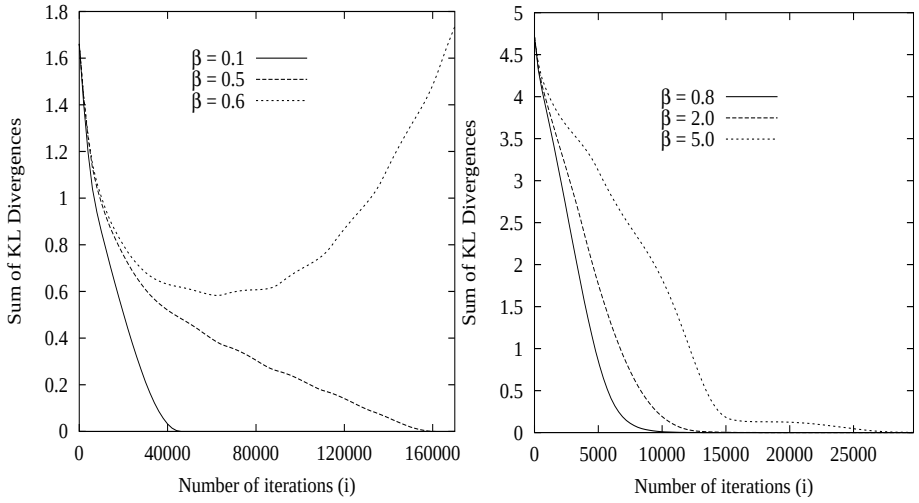

Fig. 1. The Sum of KL Divergences in the Shapley game (left) and in the game in Table 2 (right), with $\sigma = \frac{1}{1+\beta\sqrt{t}}$.

6 Conclusion

We have presented a modification of ReDVaLeR that could guarantee (a) convergence to best response against stationary opponents and *either* (b) constant bounded regret against arbitrary opponents, *or* (c) convergence to Nash equilibrium policies in self-play. The original ReDVaLeR algorithm was shown to achieve both (b) and (c) empirically but assumed that all agents must know their portions of the *same* equilibrium. The new algorithm, $RV_{\sigma(t)}$, theoretically achieves (a') convergence to near-best response against *eventually* stationary opponents, (b') no-regret payoff against arbitrary opponents *and* (c') convergence to some Nash equilibrium policy in some classes of games, in self-play. Each agent now needs to know only its portion of *any* equilibrium. Although we have shown property c' in games with unique mixed equilibrium only, we have also found it to hold in some other classes of games, like coordination games (omitted here). Future work include further generalization and discrete analysis. We also intend to experiment further with learning rate schedules identical to GIGA to directly compare their regret growth rates.

References

1. Banerjee, B., Peng, J.: Performance bounded reinforcement learning in strategic intercations. In: Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04), San Jose, CA, AAAI Press (2004) 2 – 7
2. Jafari, A., Greenwald, A., Gondek, D., Ercal, G.: On no-regret learning, fictitious play, and Nash equilibrium. In: Proceedings of the 18th International Conference on Machine Learning. (2001) 216 – 223
3. Nash, J.F.: Non-cooperative games. Annals of Mathematics **54** (1951) 286 – 295
4. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: Proc. of the 11th Int. Conf. on Machine Learning, San Mateo, CA, Morgan Kaufmann (1994) 157–163

5. Littman, M., Szepesvári, C.: A generalized reinforcement learning model: Convergence and applications. In: Proceedings of the 13th International Conference on Machine Learning. (1996) 310 – 318
6. Hu, J., Wellman, M.P.: Nash Q-learning for general-sum stochastic games. Journal of Machine Learning Research **4** (2003) 1039 – 1069
7. Littman, M.L.: Friend-or-foe Q-learning in general-sum games. In: Proceedings of the Eighteenth International Conference on Machine Learnig, Williams College, MA, USA (2001)
8. Greenwald, A., Hall, K.: Correlated Q-learning. In: Proceedings of AAAI Symposium on Collaborative Learning Agents. (2002)
9. Singh, S., Kearns, M., Mansour, Y.: Nash convergence of gradient dynamics in general-sum games. In: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence. (2000) 541–548
10. Bowling, M., Veloso, M.: Rational and convergent learning in stochastic games. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence, Seattle, WA (2001) 1021 – 1026
11. Bowling, M., Veloso, M.: Multiagent learning using a variable learning rate. Artificial Intelligence **136** (2002) 215 – 250
12. Conitzer, V., Sandholm, T.: 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. (2003)
13. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: Gambling in a rigged casino: The adversarial multi-arm bandit problem. In: Proceedings of the 36th Annual Symposium on Foundations of Compter Science, Milwaukee, WI, IEEE Computer Society Press (1995) 322 – 331
14. Fudenberg, D., Levine, D.K.: Consistency and cautious fictitious play. Journal of Economic Dynamics and Control **19** (1995) 1065 – 1089
15. Freund, Y., Schapire, R.E.: Adaptive game playing using multiplicative weights. Games and Economic Behavior **29** (1999) 79 – 103
16. Littlestone, N., Warmuth, M.: The weighted majority algorithm. Information and Computation **108** (1994) 212 – 261
17. Zinkevich, M.: Online convex programming and generalized infinitesimal gradient ascent. In: Proceedings of the 20th International Conference on Machine Learning, Washington DC (2003)
18. Bowling, M.: Convergence and no-regret in multiagent learning. In: Proceedings of NIPS 2004/5. (2005)
19. Powers, R., Shoham, Y.: New criteria and a new algorithm for learning in multi-agent systems. In: Proceedings of NIPS 2004/5. (2005)
20. Weinberg, M., Rosenschein, J.S.: Best-response multiagent learning in non-stationary environments. In: Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS). Volume 2., New York, NY, ACM (2004) 506 – 513
21. Owen, G.: Game Theory. Academic Press, UK (1995)