

$RV_{\sigma(t)}$: A unifying approach to performance and convergence in online multiagent learning

Bikramjit Banerjee

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

Jing Peng

Department of Electrical Engineering &
Computer Science
Tulane University
New Orleans, LA 70118. USA
<http://www.eecs.tulane.edu/Peng>
jp@eecs.tulane.edu

ABSTRACT

We present a new multiagent learning algorithm ($RV_{\sigma(t)}$) that can guarantee both no-regret performance (all games) and policy convergence (some games of arbitrary size). Unlike its predecessor ReDVaLeR, it (1) does not need to distinguish whether its opponents are self-play or otherwise non-stationary, (2) is allowed to know its portion of *any* equilibrium that, we argue, leads to convergence in some games in addition to no-regret. Although the regret of $RV_{\sigma(t)}$ is analyzed in continuous time, we show that it grows slower than in other no-regret techniques like GIGA and GIGA-WoLF. We show that $RV_{\sigma(t)}$ can converge to coordinated behavior in coordination games, while GIGA, GIGA-WoLF may converge to poorly coordinated (mixed) behaviors.

Categories and Subject Descriptors

I.2.11 [Distributed Artificial Intelligence]: Multiagent Systems;
I.2.6 [Artificial Intelligence]: Learning

General Terms

Algorithms, Theory, Performance

Keywords

Multiagent Learning, Game Theory

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*. We follow this line of research to present a new MAL algorithm for repeated games, with the general philosophy of policy convergence against some classes of opponents but otherwise ensuring high payoffs.

Most MAL algorithms are guaranteed to learn the best response strategies against (eventually) stationary opponents, i.e., those that play stationary strategies after an unknown of rounds. Even single agent learning techniques like Q-learning can achieve this. Building on this property, some MAL algorithms (e.g., WoLF-IGA [4], AWESOME [5]) ensure, in addition, convergence to a Nash equilibrium strategy in *self-play*, i.e., when all agents are following the *same* learning algorithm. However, there is a significantly large class of opponents that is not covered by either of these two cases, i.e., opponents that are neither stationary, nor self-play. In order to fill this void, we proposed [1] that a MAL algorithm should additionally ensure that if the opponent belongs to neither of the above classes, then it will achieve *no-regret* (defined in section 2) payoffs - something that can be construed as reasonably good performance when facing uncertain opponents. Although we had presented an algorithm, ReDVaLeR [1], that achieved all three properties, there were some limiting knowledge requirements such as

- The ReDVaLeR agent needed to know whether its opponents are self-play, or otherwise non-stationary. In this paper, we study a variant of ReDVaLeR, called $RV_{\sigma(t)}$ [2], that eliminates this assumption.
- In case of self-play, all ReDVaLeR agents need to know their portions of the *same* Nash equilibrium. $RV_{\sigma(t)}$ eliminates this requirement for some classes of games. Specifically, $RV_{\sigma(t)}$ agents only need to know their portions of *any* equilibrium in any 2×2 game, and in a game of any size with a unique mixed equilibrium (established before in [2]). In this paper, we argue theoretically and empirically, that this also holds in some coordination games with multiple equilibria.

We also show that the expression (in continuous time) for regret of $RV_{\sigma(t)}$ can have a slightly better form than those (discrete time) of GIGA [6] and GIGA-WoLF [3]. In a class of coordination games, GIGA, GIGA-WoLF are found to converge to mixed equilibria that make miscoordination highly likely besides yielding poor expected payoff to both agents, while $RV_{\sigma(t)}$ is found to always achieve strict coordination with Pareto-dominating payoffs.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

AAMAS'06 May 8–12 2006, Hakodate, Hokkaido, Japan.
Copyright 2006 ACM 1-59593-303-4/06/0005 ...\$5.00.

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, since we are more interested in dealing with uncertainties due to the other learners' changing behaviors. This refers to a scenario where a matrix game – given by an n -tuple of matrices, $(\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_n)$ where $\mathbf{R}_i$ is a payoff 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 $R_i(a_1, a_2, \dots, a_n)$, $\forall i$ – is played repeatedly by multiple agents. We represent the action space of the i th agent as A_i . Table 1 shows examples of bimatrix games of a special class that we call *Simplified Coordination Games*. It is “simplified” because only coordinated action profiles have non-zero payoffs. This game has n pure equilibria (plus some mixed equilibria) given by the matrix diagonals, but players may prefer different ones (here row player prefers 3rd row, column player prefers 1st column).

Table 1: Simplified Coordination Games

Actions	C_1	C_2	C_3
R_1	(2,100)	(0,0)	(0,0)
R_2	(0,0)	(51,51)	(0,0)
R_3	(0,0)	(0,0)	(100,2)

A mixed policy, vector $\pi_i \in \Delta(A_i)$ for agent i , is a probability distribution over A_i . We write the probability of the j th action of the i th agent at time t as $\pi_i^t(j)$. 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 π_{-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 j th action of the learner against the opponent's current policy, π_{-i}^t is written as $V_i(j, \pi_{-i}^t)$.

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

3. OUR APPROACH: ReDVaLeR WITH VARIABLE σ

We make the usual assumptions about the boundedness of the game payoffs (only non-negative in our case); the learner's ability to use vanishing step sizes ($\eta \rightarrow 0$), and that the learner can observe its payoff vector $\mathbf{V}_i$, i.e., $V_i(j, \pi_{-i}^t)$ for all actions j at any round. It *does not* need to observe the opponents' payoffs or policies/actions.

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

$$\pi_i^{t+1}(j) = \frac{\pi_i^t(j) + \eta \pi_i^t(j) l_i^t(j) V_i(j, \pi_{-i}^t)}{1 + \eta \sum_j l_i^t(j) \pi_i^t(j) V_i(j, \pi_{-i}^t)} \quad (1)$$

for any j , η being a small step size and initial condition: $\pi_i^0(j) > 0$.

In contrast with [1], the learning rates ($l_i^t(j)$) for our algorithm, $RV_{\sigma(t)}$, are defined as

$$l_i^t(j) = \begin{cases} 1 + \sigma(t) & \text{if } \pi_i^t(j) < \pi_i^*(j) \\ 1 - \sigma(t) & \text{if } \pi_i^t(j) \geq \pi_i^*(j) \end{cases} \quad (2)$$

for a suitable σ -schedule as defined below, and π_i^* being the learner's portion of a Nash equilibrium [2] policy.

DEFINITION 1 (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. $\sigma(t) \geq \sigma(t'), \forall t' \geq t$, and
3. $\sigma(t) \rightarrow 0$ as $t \rightarrow \infty$.

3.1 Properties of $RV_{\sigma(t)}$

The following theorems (proofs can be found in [2]) summarize the main properties of $RV_{\sigma(t)}$ for $\eta \rightarrow 0$. We call games of arbitrary size having unique mixed Nash equilibrium as UME games.

THEOREM 1 (EVENTUALLY STATIONARY OPPONENTS). If there exists a time τ_1 (unknown to the learner) such that all other agents play stationary policies at all times $t > \tau_1$, then for a given $\epsilon > 0$, there exists a time $\tau > \tau_1$, such that after τ an $RV_{\sigma(t)}$ agent i using σ -schedule in Definition 1, is guaranteed to converge to the set of ϵ -best response policies (i.e. payoffs are within ϵ of the best response payoff).

THEOREM 2 (NO-REGRET). If an $RV_{\sigma(t)}$ agent i uses the decaying σ -schedule of Definition 1 and $\pi_i^0(j) = 1/|A_i|$, then $\lim_{T \rightarrow \infty} \frac{Rg_i^T}{T} \leq 0$, i.e., the algorithm is asymptotically no-regret.

THEOREM 3 (SELF-PLAY IN 2×2 GAMES). In any 2×2 game, there exists a σ -schedule satisfying Definition 1, such that $RV_{\sigma(t)}$ will converge to a Nash equilibrium profile in self-play, even if they are given their portions of different equilibria.

THEOREM 4 (SELF-PLAY IN UME GAMES). There exists a σ -schedule satisfying Definition 1, which when followed by n $RV_{\sigma(t)}$ agents guarantees the convergence of their policies to the equilibrium profile in all UME games of any size, provided each agent knows appropriate loose-bounds on (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, $\min_i V_i^*$, and (4) the total number of agents, n .

3.2 Regret growth of $RV_{\sigma(t)}$

It has been argued in [2] that the σ -schedule $\sigma(t) = \frac{1}{1+\beta\sqrt{t}}$ for a small constant β (proportional to $1/(\bar{r}_i |A_i|)$) satisfies Definition 1 and all of the theorems simultaneously. This allows us to deduce a more direct expression for the regret in terms of the time variable, using the following lemma. Note that this lemma immediately leads to Theorem 2 for any σ -schedule satisfying Definition 1, and particularly for $\sigma(t) = \frac{1}{1+\beta\sqrt{t}}$.

LEMMA 5. The regret of an $RV_{\sigma(t)}$ agent (for $\eta \rightarrow 0$) is given by $Rg_i^T \leq 2\bar{r}_i \int_0^T \sigma(t) dt + \log |A_i|$.

Thus our unifying choice of $\sigma(t) = \frac{1}{1+\beta\sqrt{t}}$ yields a regret bound of

$$Rg_i^T \leq \left(\frac{4\bar{r}_i}{\beta} \right) \left(\sqrt{T} - \frac{\log(1 + \beta\sqrt{T})}{\beta} \right) + \log |A_i|,$$

whereas GIGA-WoLF has

$$Rg_i^T \leq (2 + \bar{r}_i^2 |A_i|) \sqrt{T} - |A_i| \bar{r}_i^2$$

and half that for GIGA. The presence of the log term in our expression for Rg_i^T might make it grow slower, though our analysis is in continuous time as opposed to discrete time analysis of GIGA and GIGA-WoLF. In order to make a more direct comparison, we experiment with all these methods in *discrete time*.

4. SIMPLIFIED COORDINATION GAMES

The following two lemmas (proofs omitted) and the experimental results suggest that $RV_{\sigma(t)}$ in self-play converges to playing coordinated actions even if they were given different pure equilibrium policies (e.g., the ones they prefer) at start. GIGA and GIGA-WoLF on the other hand can converge to poorly coordinated behavior with poor payoffs. Although the definition of this class of games and the following results are presented for two agent scenarios, it can be trivially extended to the multi-player case.

LEMMA 6. *Two $RV_{\sigma(t)}$ agents cannot converge asymptotically to uncoordinated pure policies in the SCGs.*

LEMMA 7. *If at a large enough time in the SCG, both $RV_{\sigma(t)}$ agents assign their highest probabilities to the same action (j), then the probability of both players playing j increases.*

Our experiments in this class of games used the game shown in Table 1 and various completely mixed starting policy profiles. Each $RV_{\sigma(t)}$ player was given its portion of the equilibrium policy that it would prefer, e.g., player 1 was given $[0, 0, 1]$ and player 2 was given $[1, 0, 0]$.

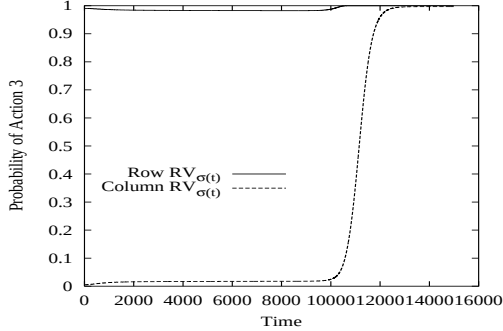


Figure 1: The probabilities of R_3 and C_3 of two $RV_{\sigma(t)}$ agents starting at $\langle [0.005, 0.005, 0.99], [0.99, 0.005, 0.005] \rangle$, in the coordination game in Table 1. Each agent used $\beta = 0.2$ and $\eta = 2 \times 10^{-3}$.

Two $RV_{\sigma(t)}$ agents always converged to a coordinated action profile, sometimes even to action profiles $\langle R_1, C_1 \rangle$ and $\langle R_3, C_3 \rangle$ depending on the starting point, in both games. Figure 1 show both $RV_{\sigma(t)}$ agents converging to their 3^{rd} actions (i.e., R_3, C_3) even when each starts close to its desired policy and uses its desired equilibrium. Essentially, one agent learns to compromise while the other maximizes its payoff, reaching a coordinated profile. Which agent compromises depends more on the starting policies than the learning dynamics.

GIGA and GIGA-WoLF converge to either coordinated action profiles or to mixed equilibria e.g., $\langle [1/51, 0, 50/51], [50/51, 0, 1/51] \rangle$. This happens when they start close to their preferred policies like the $RV_{\sigma(t)}$ agents. Figure 2 show the sum

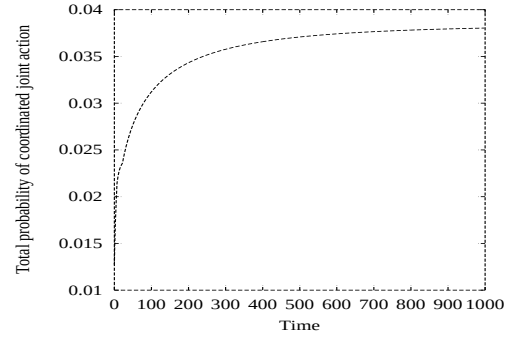


Figure 2: The sum of the joint probabilities of diagonal actions when two GIGA-WoLF agents start at $\langle [0.005, 0.005, 0.99], [0.99, 0.005, 0.005] \rangle$, in the coordination game in Table 1. Each agent used $\eta_1 = 2 \times 10^{-3}$.

of the coordinated joint actions of two GIGA-WoLF agents, i.e., $\sum_s \pi_1^t(s) \pi_2^t(s)$, which converges to the inner product of the mixed equilibrium profile that the agents converge to (in this case 0.038446751). If the agents had converged to a purely coordinated profile, then $\sum_s \pi_1^t(s) \pi_2^t(s)$ would have converged to 1. Thus figure 2 also shows the low probability of coordination of the resulting profile (3.8%). This not only means that the agents would play uncoordinated actions most of the times, but also that even the expected payoffs of *both* agents (≈ 1.96 for both) are lower than at any coordinated action profile. Note that GIGA-WoLF converges to this same undesired equilibrium whether it uses $\eta_1 = 1$ or $\eta_1 = 2 \times 10^{-3}$ (like $RV_{\sigma(t)}$), where $\eta_t = \eta_1 / \sqrt{t}$.

5. ACKNOWLEDGEMENT

We thank Mike Bowling for helping with the code of GIGA-WoLF. This work was supported in part by Grants from Army Research Office W911NF-05-1-0377 and Louisiana BOR LEQSF(2002-05)-RD-A-29.

6. REFERENCES

- [1] B. Banerjee and J. Peng. Performance bounded reinforcement learning in strategic interactions. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04)*, pages 2 – 7, San Jose, CA, 2004. AAAI Press.
- [2] B. Banerjee and J. Peng. Convergence of no-regret learning in multiagent systems. In *Proceedings of the First International Workshop on Learning and Adaptation in Multiagent Systems (LAMAS)*, Utrecht, The Netherlands, 2005. Held in conjunction with AAMAS-05.
- [3] M. Bowling. Convergence and no-regret in multiagent learning. In *Proceedings of NIPS 2004/5*, 2005.
- [4] M. Bowling and M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136:215 – 250, 2002.
- [5] V. Conitzer and T. Sandholm. 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.
- [6] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, Washington DC, 2003.