Convergence of Gradient Dynamics with a Variable Learning Rate

Michael Bowling MHB@CS.CMU.EDU
Manuela Veloso MMV@CS.CMU.EDU
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA

Abstract

As multiagent environments become more prevalent we need to understand how this changes the agent-based paradigm. One aspect that is heavily affected by the presence of multiple agents is learning. Traditional learning algorithms have core assumptions, such as Markovian transitions, which are violated in these environments. Yet, understanding the behavior of learning algorithms in these domains is critical. Singh, Kearns, and Mansour (2000) examine gradient ascent learning, specifically within a restricted class of repeated matrix games. They prove that when using this technique the average of expected payoffs over time converges. On the other hand, they also show that neither the players' strategies nor their expected payoffs themselves are guaranteed to converge. In this paper we introduce a variable learning rate for gradient ascent, along with the WoLF ("Win or Learn Fast") principle for regulating the learning rate. We then prove that this modification to gradient ascent has the stronger notion of convergence, that is, strategies and payoffs converge to a Nash equilibrium.

1. Introduction

Environments involving multiple agents are becoming more prevalent. Information and electronic commerce agents are opening new arenas of multiagent systems. Also, the decreasing cost of robotic agents is creating many new environments involving collaboration, competition, and other interactions. Learning is a key element to agent-based systems both to improve the agent's performance and to adapt to a potentially dynamic environment. In multiagent systems, learning is even more critical since the agent needs to adapt to the behavior of the other agents. If the other agents are also adapting then this becomes quite a challenging problem. Game theory studies strategic interactions among multiple players and provides a framework for analyzing these interactions. In this paper, we examine the multiagent learning problem in the game theoretic framework of a repeated matrix game.

Learning in repeated games have been studied extensively in game theory (Fudenberg & Levine, 1999). Fictitious play (Robinson, 1951) is one mechanism where the learner assumes the other players are following some Markovian strategy, which is estimated from their historical play. The player then uses these empirically estimated strategies to select its optimal action. For a class of repeated games, fictitious play in self-play (i.e. when all agents use fictitious play) has the property that the empirical averages of the strategies played will approach a Nash equilibrium. On the other hand, their actual strategies do not necessarily converge, nor does the expected payoff at the current time step.

Another learning approach is that of model learning. An example of this is where the other agent is modelled as a finite automata, and optimal play is then calculated using the learned automata (Carmel & Markovitch, 1996). This has very desirable properties when the other players can indeed be modelled as small fixed finite automata, but in situations of simultaneous adaptation or even self-play the behavior is unclear. This also assumes that the other players are computationally inferior since it assumes their automaton is strictly smaller in order to be learned and exploited.

A different approach is to examine a simple learning technique that does not make hindering assumptions about the other players. Singh, Kearns, and Mansour (2000) examined the use of a simple gradient ascent learner. Specifically, they examined the use of gradient ascent in the space of mixed strategies, where players sought to maximize their expected payoffs. They analyzed this technique in the class of two-player, two-action, general-sum repeated matrix games. As they pointed out, examining gradient ascent in this setting is necessary to understanding its use in complex environments. These problems often necessitate biased learning using a parameterized solution space, making gradient ascent very appealing.

Singh and colleagues proved that gradient ascent in self-play displays a *weak* notion of convergence. Specifically, they prove the players' strategies are guaranteed to converge to a Nash equilibrium *or* the payoffs to the players over time would average to the equilibrium. There is no guarantee of convergence of the players' expected payoffs or even a non-trivial bound. Instead, they give an example of a class of games where the expected payoffs do not converge (see Section 4.3).

In this paper we look at the effects of applying a *variable learning rate* to gradient ascent, specifically using the *WoLF principle* ("Win or Learn Fast") to regulate the learning rate. We had observed empirically that this had strong effects on the convergence properties of a learning algorithm (Bowling & Veloso, 2001). Building on the results of Singh and colleagues, we prove theoretically that this modification does in fact exhibit the *strong* notion of convergence. Specifically, we prove that the players strategies and expected payoffs converge to a Nash equilibrium in two-player, two-action general-sum repeated games. We will begin by reviewing the original results for gradient ascent in Section 2, and then introduce the concepts of a variable learning rate and the WoLF principle in Section 3. In Section 4 we prove the convergence properties of this modified gradient ascent, and a short discussion follows before concluding.

2. Gradient Ascent

We begin with a very brief overview of repeated matrix games and the concept of Nash equilibrium. We will then discuss the previous work examining gradient ascent dynamics, highlighting what is necessary for our analysis.

2.1 Repeated Matrix Games

We will examine infinitely repeated matrix (or normal-form) games. (Osborne & Rubinstein, 1994) The players simultaneously select an action from their available set, and the joint action of the players determine their payoffs according to their payoff matrix. This process is then repeated indefinitely. Players may also select actions stochastically using some probability distribution over their available actions. This is said to be a mixed strategy. This will be made more concrete when we examine two-player, two-action matrix games below.

A Nash equilibrium (Nash, Jr., 1950) of a matrix game is a set of strategies, one for each player, where no player can increase its expected payoff by deviating from this equilibrium strategy. This is a powerful concept, and in classical game theory equilibria are considered to be the set of "rational" solutions. We will examine Nash equilibria not as a goal in their own right, but rather as strategies with zero gradient, and therefore possible convergence points for gradient ascent learners. This is stated formally in Lemma 2.

2.2 Learning using Gradient Ascent

Singh, Kearns, and Mansour (2000) examined the dynamics of using gradient ascent in two-player, two-action, iterated matrix games. We can represent this problem as two matrices,

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \quad C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}.$$

Each player selects an action from $\{1, 2\}$ which determines the payoffs to the players. If the *row* player selects action i and the *column* player selects action j, then the row player receives a payoff r_{ij} and the column player receives the payoff c_{ij}.

Since this is a two-action game, a strategy (i.e. a probability distribution over the two available actions) can be represented as a single value. Let $\alpha \in [0, 1]$ be a strategy for the row player, where α corresponds to the probability the player selects the first action and $(1 - \alpha)$ is the probability the player selects the second action. Similarly, let β be a strategy for the column player. We can consider the joint strategy (α, β) as a point in $\mathbb{R}^2$ constrained to the unit square.

For any pair of strategies (α, β), we can write the expected payoffs the row and column player will receive. Let $V_r(\alpha, \beta)$ and $V_c(\alpha, \beta)$ be these expected payoffs, respectively. Then,

$$\begin{aligned} V_r(\alpha, \beta) &= \alpha\beta r_{11} + \alpha(1-\beta)r_{12} + \\ &\quad (1-\alpha)\beta r_{21} + (1-\alpha)(1-\beta)r_{22} \\ &= u\alpha\beta + \alpha(r_{12} - r_{22}) + \\ &\quad \beta(r_{21} - r_{22}) + r_{22} \end{aligned} \tag{1}$$

$$\begin{aligned} V_c(\alpha, \beta) &= \alpha\beta c_{11} + \alpha(1-\beta)c_{12} + \\ &\quad (1-\alpha)\beta c_{21} + (1-\alpha)(1-\beta)c_{22} \\ &= u'\alpha\beta + \alpha(c_{12} - c_{22}) + \\ &\quad \beta(c_{21} - c_{22}) + c_{22} \end{aligned} \tag{2}$$

where,

$$\begin{aligned} u &= r_{11} - r_{12} - r_{21} + r_{22} \\ u' &= c_{11} - c_{12} - c_{21} + c_{22}. \end{aligned}$$

A player can now consider the effect of changing its strategy on its expected payoff. This can be computed as just the partial derivative of its expected payoff with respect to its strategy,

$$\frac{\partial V_r(\alpha, \beta)}{\partial \alpha} = \beta u + (r_{12} - r_{22}) \tag{3}$$

$$\frac{\partial V_c(\alpha, \beta)}{\partial \beta} = \alpha u' + (c_{21} - c_{22}). \tag{4}$$

In the gradient ascent algorithm a player will adjust its strategy after each iteration so as to increase its expected payoffs. This means the player will move its strategy in the direction of the current gradient with some step size, η. If (α_k, β_k) are the strategies on the kth iteration, and both players are using gradient ascent then the new strategies will be,

$$\begin{aligned}\alpha_{k+1} &= \alpha_k + \eta \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \alpha_k} \\ \beta_{k+1} &= \beta_k + \eta \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \beta_k}.\end{aligned}$$

If the gradient will move the strategy out of the valid probability space (i.e. the unit square) then the gradient is projected back on to the probability space. This will only occur on the boundaries of the probability space. The question to consider then is what can we expect will happen if both players are using gradient ascent to update their strategies.

The analysis, by Singh and colleagues, of gradient ascent examines the dynamics of the learners in the case of an infinitesimal step size ($\lim_{\eta \to 0}$). They call this algorithm Infinitesimal Gradient Ascent (IGA). They observe later that an algorithm with an appropriately decreasing step size will have the same properties as IGA. In the next section we will briefly outline their analysis.

2.3 Analysis of IGA

The main conclusion of Singh, Kearns, and Mansour (2000) is the following theorem.

Theorem 1 *If both players follow Infinitesimal Gradient Ascent (IGA), where $\eta \to 0$, then their strategies will converge to a Nash equilibrium* OR *the average payoffs over time will converge in the limit to the expected payoffs of a Nash equilibrium.*

Their proof of this theorem proceeds by examining the dynamics of the strategy pair, (α, β). This is an affine dynamical system in $\mathbb{R}^2$ where the dynamics are defined by the differential equation,

$$\begin{bmatrix} \frac{\partial \alpha}{\partial t} \\ \frac{\partial \beta}{\partial t} \end{bmatrix} = \begin{bmatrix} 0 & u \\ u' & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} (r_{12} - r_{22}) \\ (c_{21} - c_{22}) \end{bmatrix}.$$

If we define U to be the multiplicative matrix term above with off-diagonal values u and u', then we can classify the dynamics of the system based on properties of U. From dynamical systems theory, if U is invertible then there are only two qualitative forms for the dynamics of the system, depending on whether U has purely real or purely imaginary eigenvalues. This results in three cases: U is not invertible, U has purely real eigenvalues, or U has purely imaginary eigenvalues. The qualitative forms of these different cases are shown in Figure 1. Their analysis then proceeded by examining each case geometrically. One important consideration is that the basic forms above are for the unconstrained dynamics not the dynamics that projects the gradient onto the unit square. Basically, this requires considering all possible positions of the unit square relative to the dynamics shown in Figure 1.

One crucial aspect to their analysis were points of zero-gradient in the constrained dynamics, which they show to correspond to Nash equilibria. In the unconstrained dynamics, there exist at most one point of zero-gradient, which is called the center and denoted (α^*, β^*). This point can be found mathematically by setting equations 3 and 4 to zero and solving,

$$(\alpha^*, \beta^*) = \left(\frac{(c_{22} - c_{21})}{u'}, \frac{(r_{22} - r_{12})}{u} \right).$$

Notice that the center may not even be inside the unit square. In addition, if U is not invertible then there is no point of zero gradient in the unconstrained dynamics. But in the constrained dynamics, where gradients on the boundaries of the unit square are projected onto the unit square, additional points of zero gradient may exist. When IGA converges it will be to one of these points with zero gradient.

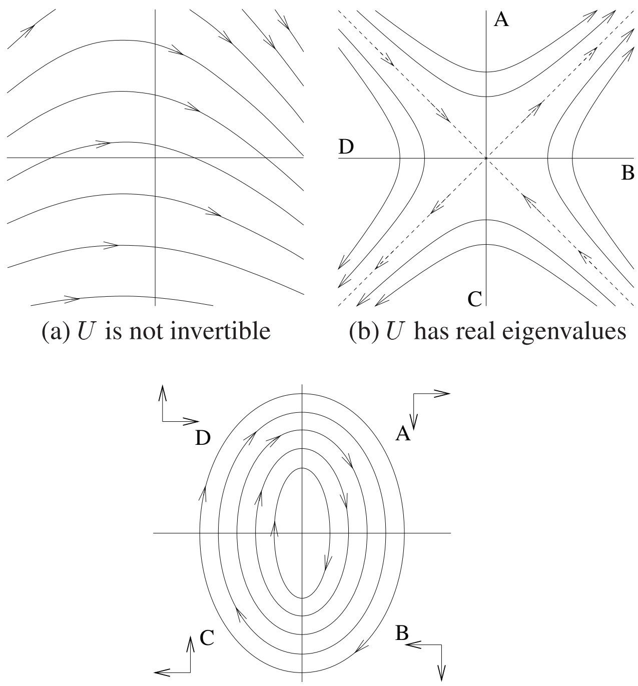
This theorem is an exciting result since it is one of the first convergence results for a payoff-maximizing multiagent learning algorithm. The notion of convergence, though, is rather weak. In fact, not only may the players' policies not converge when playing gradient ascent but the expected payoffs may not converge either. Furthermore, at any moment in time the expected payoff of a player could be arbitrarily poor.[1] Not only does this make it difficult to evaluate a learner, it also could be potentially disastrous when applied with temporal differencing for multiple state stochastic games, which assumes that expected payoffs in the past predict expected payoffs in the future.

In the next section we will examine a method for addressing this convergence problem. We will then prove that this new method has the stronger notion of convergence, i.e. players will *always* converge to a Nash equilibrium.

3. Variable Learning Rate and the WoLF Principle

We now introduce the concept and study the impact of a variable learning rate. In the gradient ascent algorithm pre-

[1]The idea that average payoffs converge only means that if there's a period of arbitrarily low expected payoffs there must be some corresponding period in the past or in the future of arbitrarily high expected payoffs.



(a) U is not invertible (b) U has real eigenvalues

(c) U has imaginary eigenvalues

Figure 1. Qualitative forms of the IGA dynamics.

sented above the steps taken in the direction of the gradient were constant. We will now allow them to vary over time, thus changing the update rules to,

$$\begin{aligned}\alpha_{k+1} &= \alpha_k + \eta \ell_k^r \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \alpha} \\ \beta_{k+1} &= \beta_k + \eta \ell_k^c \frac{\partial V_r(\alpha_k, \beta_k)}{\partial \beta}\end{aligned}$$

where,

$$\ell_k^{r,c} \in [\ell_{\min}, \ell_{\max}] > 0$$

At the kth iteration the algorithm takes a step of size $\eta \ell_k$ in the direction of the gradient. Notice the restrictions on ℓ_k require that it be strictly positive and bounded, thus bounding the step sizes as well.

The specific method for varying the learning rate that we are contributing is the WoLF ("Win or Learn Fast") principle. The essence of this method is to learn quickly when losing, and cautiously when winning. The intuition is that a learner should adapt quickly when it is doing more poorly than expected. When it is doing better than expected, it should be cautious since the other players are likely to change their policy. The heart of the algorithm is how to determine whether a player is winning or losing. For the analysis in this section each player will select a Nash equilibrium and compare their expected payoff with the payoff they would receive if they played according to the selected equilibrium strategy. Let α^e be the equilibrium strategy selected by the row player, and β^e be the equilibrium strategy selected by the column player. Notice that no requirement is made that the players choose the same equilibrium (i.e. the strategy pair (α^e, β^e) may not be a Nash equilibrium). Formally,

$$\ell_k^r = \begin{cases} \ell_{\min} & \text{if } V_r(\alpha_k, \beta_k) > V_r(\alpha^e, \beta_k) & \text{WINNING} \\ \ell_{\max} & \text{otherwise} & \text{LOSING} \end{cases}$$

$$\ell_k^c = \begin{cases} \ell_{\min} & \text{if } V_c(\alpha_k, \beta_k) > V_c(\alpha_k, \beta^e) & \text{WINNING} \\ \ell_{\max} & \text{otherwise} & \text{LOSING} \end{cases}$$

With a variable learning rate such as this we can still consider the case of an infinitesimal step size ($\lim_{\eta \to 0}$). We will call this algorithm WoLF-IGA and in the next section show that the WoLF adjustment has a very interesting effect on the convergence of the algorithm.

4. Analysis of WoLF-IGA

We will prove the following result.

Theorem 2 *If in a two-person, two-action, iterated general-sum game, both players follow the WoLF-IGA algorithm (with $\ell_{\max} > \ell_{\min}$), then their strategies will converge to a Nash equilibrium.*

Notice that this is the more standard notion of convergence and strictly stronger than what is true for basic IGA.

The proof of this theorem will follow closely with the proof of Theorem 1 from Singh and colleagues, by examining the possible cases for the dynamics of the learners. First, let us write down the differential equations that define the system with an infinitesimal step size,

$$\begin{bmatrix} \frac{\partial \alpha}{\partial t} \\ \frac{\partial \beta}{\partial t} \end{bmatrix} = \begin{bmatrix} 0 & \ell^r(t) u \\ \ell^c(t) u' & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \ell^r(t)(r_{12} - r_{22}) \\ \ell^c(t)(c_{21} - c_{22}) \end{bmatrix}.$$

We will call the multiplicative matrix with off-diagonal entries $U(t)$ since it now depends on the learning rates at time t, $\ell^r(t)$ and $\ell^c(t)$. At time t, the qualitative form of the dynamics is determined by the $U(t)$ matrix and can be summarized into three general cases,

- $U(t)$ is not invertible,
- $U(t)$ has purely real eigenvalues, or
- $U(t)$ has purely imaginary eigenvalues.

The first thing to note is that the above cases do not depend on t. The following lemma is even stronger.

Lemma 1 *$U(t)$ is invertible if and only if U (as defined for IGA in Section 2.3) is invertible. $U(t)$ has purely imaginary* *eigenvalues if and only if U has purely imaginary eigenvalues. $U(t)$ has purely real eigenvalues if and only if U has purely real eigenvalues.*

Proof. Since $\ell^{r,c}(t)$ is positive, it is trivial to see that $(\ell^r(t)u)(\ell^c(t)u') = (\ell^r(t)\ell^c(t))uu'$ is greater-than, less-than, or equal-to zero, if and only if uu' is greater-than, less-than, or equal-to zero, respectively. Since these are the exact conditions of invertibility and purely real/imaginary eigenvalues the lemma is true. □

So $U(t)$ will always satisfy the same case (and therefore have the same general dynamics) as IGA *without* a variable learning rate. In the sections that follow we will be examining each of these cases separately. The proofs of most of the cases will proceed similarly to the proof for IGA. In fact most of the proof will not rely on any particular learning rate adjustment at all. Only in the final sub-case of the final case will we be forced to deviate from their arguments. This is due to the fact that variable learning rates in general do not change the overall direction of the gradient (i.e. the sign of the partial derivatives). Since most of the proof of IGA's convergence only depends on the signs of the derivatives, we can use the same arguments. For these cases we will present only an abbreviated proof of convergence to illustrate that the variable learning rate does not affect their arguments. We recommend the IGA analysis for a more thorough examination including helpful diagrams. In the remaining sub-case, where IGA is shown not to converge, we will show that in this case WoLF-IGA will converge to a Nash equilibrium.

We will make liberal use of a crucial lemma from their proof for IGA. This lemma implies that if the algorithms converge then what the strategies converge to must be a Nash equilibrium.

Lemma 2 *If, in following IGA or WoLF-IGA, $\lim_{t\to\infty}(\alpha(t), \beta(t)) = (\bar{\alpha}, \bar{\beta})$, then $(\bar{\alpha}, \bar{\beta})$ is a Nash equilibrium.*

Proof. The proof for IGA is given in (Singh et al., 2000), and shows that the algorithm converges if and only if the projected gradient is zero, and such strategy pairs must be a Nash equilibrium. For WoLF-IGA notice also that the algorithm converges if and only if the projected gradient is zero, which is true if and only if the projected gradient in IGA is zero. Therefore that point must be a Nash equilibrium. □

Now we will examine the individual cases.

4.1 $U(t)$ is Not Invertible

In this case the dynamics of the strategy pair has the qualitative form shown in Figure 1(a).

Lemma 3 *When U(t) is not invertible, IGA with any learning rate adjustment leads the strategy pair to converge to a point on the boundary that is a Nash equilibrium.*

Proof. Notice that $U(t)$ is not invertible if and only if u or u' is zero. Without loss of generality, assume u is zero, then the gradient for the column player is constant. The column player's strategy, β, will converge to either zero or one (depending on whether the gradient was positive or negative). At this point, the row player's gradient becomes constant and therefore must also converge to zero or one, depending on the sign of the gradient. The joint strategy therefore converges to some corner, which by Lemma 2 is a Nash equilibrium. □

4.2 $U(t)$ has Real Eigenvalues

In this case the dynamics of the strategy pair has the qualitative form shown in Figure 1(b).

Lemma 4 *When U(t) has real eigenvalues, IGA with any learning rate adjustment leads the strategy pair to converge to a point that is a Nash equilibrium.*

Proof. Without loss of generality, assume that $u, u' > 0$. This is the dynamics shown in Figure 1(b). Consider the case where the center is inside the unit square. Notice that if the strategy pair is in quadrant A, the gradient is always up and right. Therefore, any strategy pair in this region will eventually converge to the upper-right corner of the unit square. Likewise, strategies in quadrant C will always converge to the bottom-left corner. Now consider a strategy pair in quadrant B. The gradient is always up and left, and therefore the strategy will eventually exit this quadrant, entering quadrant A or C, or possibly hitting the center. At the center the gradient is zero, and so it has converged. If it enters one of quadrants A or C then we've already shown it will converge to the upper-right or lower-left corner. Therefore, the strategies always converge and by Lemma 2 the point must be a Nash equilibrium. Cases where the center is not within the unit square or is on the boundary of the unit square can also be shown to converge by a similar analysis, and is discussed in (Singh et al., 2000). □

4.3 $U(t)$ has Imaginary Eigenvalues

In this case the dynamics of the strategy pair has the qualitative form shown in Figure 1(c). This case can be further broken down into sub-cases depending where the unit square is in relation to the center.

Center is Not Inside the Unit Square. In this case we still can use the same argument as for IGA.

Lemma 5 *When U(t) has imaginary eigenvalues and the center, (α^*, β^*), is not inside the unit square, IGA with any* *learning rate adjustment leads the strategy pair to converge to a point on the boundary that is a Nash equilibrium.*

Proof. There are three cases to consider. The first is the unit square lies entirely within a single quadrant. In this case the direction of the gradient will be constant (e.g. down-and-right in quadrant A). Therefore the strategies will converge to the appropriate corner (e.g. bottom-right corner in quadrant A). The second case is the unit square is entirely within two neighboring quadrants. Consider the case that it lies entirely within quadrants A and D. The gradient always points to the right and therefore the strategy will eventually hit the right boundary at which point it will be in quadrant A and the gradient will be pointing downward. Therefore in this case it will converge to the bottom right corner. We can similarly show convergence for other pairs of quadrants. The third and final case is when the center is on the boundary of the unit square. In this case some points along the boundary will have a projected gradient of zero. By similar arguments to those above, any strategy will converge to one of these boundary points. See (Singh et al., 2000) for a diagram and further explanation. Since in all cases the strategy pairs converge, by Lemma 2 they must have converged to a Nash equilibrium. □

Center is Inside the Unit Square. This is the final sub-case and is the point where the dynamics of IGA and WoLF-IGA qualitatively differ. We will show that, although IGA will not converge in this case, WoLF-IGA will. The proof will identify the areas of the strategy space where the players are "winning" and "losing" and show that the trajectories are actually piecewise elliptical in such a way that they spiral towards the center. All of the lemmas in this subsection implicitly assume that $U(t)$ has imaginary eigenvalues and the center is inside the unit square. We begin with the following lemma that considers the dynamics for fixed learning rates.

Lemma 6 *If the learning rates, ℓ^r and ℓ^c, remain constant, then the trajectory of the strategy pair is an elliptical orbit around the center, (α^*, β^*), and the axes of this ellipse are,*

$$\left[\begin{array}{c} 0 \\ \sqrt{\frac{\ell^c|u|}{\ell^r|u'|}} \end{array}\right], \quad \left[\begin{array}{c} 1 \\ 0 \end{array}\right].$$

Proof. This is just a result from dynamical systems theory (Reinhard, 1987) as mentioned in (Singh et al., 2000) when $U(t)$ has imaginary eigenvalues. □

We now need the critical lemma that identifies the areas of strategy space where the players are using a constant learning rate. Notice that this corresponds to the areas where the players are "winning" or "losing".

Lemma 7 *The player is "winning" if and only if that player's strategy is moving away from the center.*

Proof. Notice that in this sub-case where $U(t)$ has imaginary eigenvalues and the center is within the unit square, the game has a single Nash equilibrium, which is the center. So, the players' selected equilibrium strategies for the WoLF principle must be the center, i.e. $(\alpha^e, \beta^e) = (\alpha^*, \beta^*)$. Now, consider the row player. The player is "winning" when its current expected payoffs is larger than the expected payoffs if it were to play its selected equilibrium. This can be written as,

$$V_r(\alpha, \beta) - V_r(\alpha^e, \beta) > 0.$$

We can write out the left hand side by using equation 1.

$$\begin{array}{c}(\alpha\beta u + \alpha(r_{12} - r_{22}) + \beta(r_{21} - r_{22}) + r_{22}) - \\ (\alpha^e\beta u + \alpha^e(r_{12} - r_{22}) + \beta(r_{21} - r_{22}) + r_{22}),\end{array}$$

and then simplify substituting the center for the equilibrium strategies,

$$\begin{array}{c}(\alpha - \alpha^*)\beta u + (\alpha - \alpha^*)(r_{12} - r_{22}) = \\ (\alpha - \alpha^*)(\beta u - (r_{12} - r_{22})) = \\ (\alpha - \alpha^*)\frac{\partial V_r(\alpha,\beta)}{\partial \alpha}.\end{array}$$

Notice that this is positive if and only if the two factors have the same sign. This is true if and only if the player's strategy, α, is greater than the center, α^*, and it is increasing, or it is smaller than the center and decreasing. So the player is winning if and only if its strategy is moving away from the center. The same can be shown for the column player. □

Corollary 1 *Throughout any one quadrant, the learning rate is constant.*

Combining Lemmas 6 and 7, we find that the trajectories will be piece-wise elliptical orbits around the center, where the pieces correspond to the quadrants defined by the center. We can now prove convergence for a limited number of starting strategy pairs. We will then use this lemma to prove convergence for any initial strategy pairs.

Lemma 8 *If $\ell_{\max} > \ell_{\min}$ then for any initial strategy pair, (α^*, β_0) or (α_0, β^*), that is "sufficiently close" to the center, the strategy pair will converge to the center. "Sufficiently close" here means that the elliptical trajectory from this point defined when both players use 1 as their learning rate lies entirely within the unit square.*

Proof. Without loss of generality assume $u > 0$ and $u' < 0$. This is the case shown in Figure 1(c). Let $l = \sqrt{\frac{\ell_{\min}}{\ell_{\max}}} < 1.0$, and $r = \sqrt{\frac{|u'|}{|u|}}$. Consider an initial strategy (α^*, β_0) with $\beta_0 > \beta^*$.

For any fixed learning rates for the players, by Lemma 6, the trajectory forms an ellipse centered at (α^*, β^*) and with

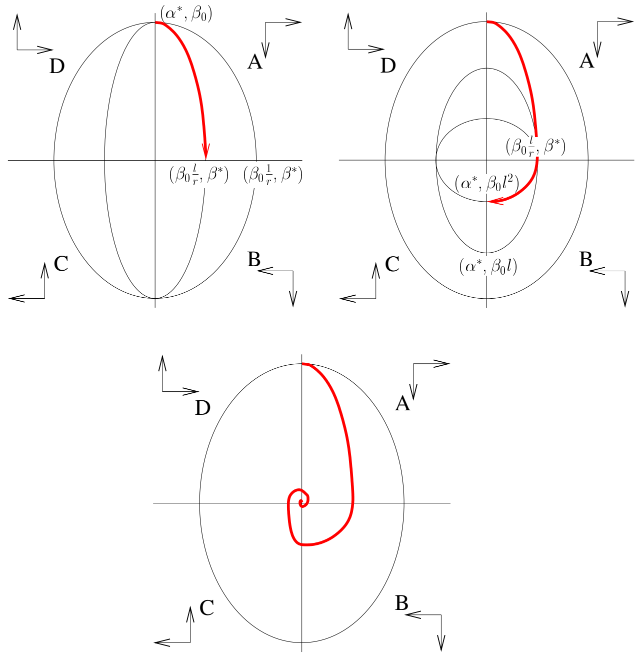


Figure 2. The trajectory of learning rates using WoLF-IGA when $U(t)$ has imaginary eigenvalues and the center is inside the unit square.

the ratio of its y-radius to its x-radius equal to,

$$\sqrt{\frac{\ell^c}{\ell^r}}r.$$

Since the trajectory is piecewise elliptical we can consider the ellipse that the trajectory follows while in each quadrant. This is shown graphically in Figure 2. As the trajectory travels through quadrant A, by Lemma 7, we can observe that the row player is "winning" and the column player is "losing". Therefore, $\ell^r = \ell_{\min}$ and $\ell^c = \ell_{\max}$, so the ratio of the ellipse's axes will be r/l, and this ellipse will cross into quadrant B at the point $(\beta_0 \frac{l}{r}, \beta^*)$. Similarly, in quadrant B, the row player is "losing" and the column player is "winning" therefore the ratio of the ellipse's axes will be rl and the ellipse will cross into quadrant C at the point $(\alpha^*, \beta_0 l^2)$.

We can continue this to return to the axis where the trajectory began. The strategy pair at that point will be $(\alpha^*, \beta_0 l^4)$. So, for each orbit around the center we decrease the distance to the center by a factor of $l^4 < 1.0$, and therefore the trajectory will converge to the center. We can reason identically for any other sufficiently close initial strategies on the axes. $\square$

Lemma 9 *When $U(t)$ has imaginary eigenvalues and the center, (α^*, β^*), is inside the unit square, WoLF-IGA leads the strategy pair to converge to the center, and therefore to a Nash equilibrium.*

Proof. The proof is just an extension of Lemma 8. Consider the largest ellipse when both players learning rates are one that fits entirely within the unit square. This ellipse will touch the boundary of the unit square and do so at the boundary of two quadrants. Now consider any initial strategy pair. The strategy pair will follow piecewise elliptical orbits or move along the unit square boundary and eventually will hit the boundary between those same quadrants. At this point it is on or inside the largest ellipse defined when players have a learning rate of one, and therefore we can apply Lemma 8 and so the trajectory will converge to the center. So, from any initial strategy pair the trajectory will converge to the center, which is a Nash equilibrium. $\square$

Lemmas 3, 4, 5, and 9 combine to prove Theorem 2. In summary the WoLF principle strengthens the IGA convergence result. In self-play with WoLF-IGA, players' strategies and their expected payoffs converge to Nash equilibrium strategies and payoffs of the matrix game. This result can be generalized beyond self-play in the following corollary, which we state without proof.

Corollary 2 *If in a two-person, two-action, iterated general-sum game, both players follow the WoLF-IGA algorithm but with different $\ell_{\min}$ and $\ell_{\max}$, then their strategies will converge to a Nash equilibrium if,*

$$\frac{\ell^r_{\min}\ \ell^c_{\min}}{\ell^r_{\max}\ \ell^c_{\max}} < 1.$$

Specifically, WoLF-IGA (with $\ell_{\max} > \ell_{\min}$) versus IGA ($\ell_{\max} = \ell_{\min}$) will converge to a Nash equilibrium.

5. Discussion

There are some final points to be made about this result. First, we will present some further justification for the WoLF principle as it has been used in other learning related problems. Second, we will present a short discussion on determining when a player is "winning".

5.1 Why WoLF?

Apart from this theoretical result the WoLF principle may appear to be just an unfounded heuristic. But actually it has been studied in some form in other areas, notably when considering an adversary. In evolutionary game theory the *adjusted replicator dynamics* (Weibull, 1995) scales the individual's growth rate by the inverse of the overall success of the population. This will cause the population's composition to change more quickly when the population as a whole is performing poorly. A form of this also appears as a modification to the *randomized weighted majority* algorithm (Blum & Burch, 1997). In this algorithm, when an expert makes a mistake, a portion of its weight loss is redistributed among the other experts. If the algorithm is placing large weights on mistaken experts (i.e. the algorithm is "losing"), then a larger portion of the weights are redistributed (i.e. the algorithm adapts more quickly.)

In addition, a variable learning rate and the WoLF principle has been applied to a reinforcement algorithm, policy hill-climbing (Bowling & Veloso, 2001). WoLF policy hill-climbing was empirically demonstrated to converge to equilibria in self-play in both repeated matrix games, as well as more complex multiple state stochastic games.

5.2 Defining "Winning"

The WoLF principle for adjusting the learning rate is to learn faster when losing, more slowly when winning. This places a great deal of emphasis on how to determine that a player is winning. In the description of WoLF-IGA above, the row-player was considered winning when,

$$V_r(\alpha_k, \beta_k) > V_r(\alpha^e, \beta_k).$$

Essentially, the player was winning if he'd prefer his current strategy to that of playing some equilibrium strategy against the other player's current strategy.

Another possible choice of determining when a player is winning might be if its expected payoff is currently larger than the value of the game's equilibrium (or some equilibrium if multiple exist). It is interesting to note that in zero-sum games with mixed strategy equilibria these two rules are actually identical.

In general-sum games, though, this is not necessarily the case. There exist games with points in the strategy space where the player is receiving a lower expected payoff than the equilibrium value, but the equilibrium strategy would not do any better.[2] Essentially, the player is not doing poor because of its strategy, but rather because of the play of the other player. It is at this point that the gradient is likely to be moving the strategy away from the equilibrium, and so learning quickly would discourage convergence.

6. Conclusion

This paper examines the effects of using a variable learning rate, specifically the WoLF principle, on gradient ascent learning. We show that this modification to gradient ascent has a surprising affect on the dynamics of learning in two-player, two-action, general-sum repeated matrix games. We prove a strong notion of convergence, that is not true of basic gradient ascent, such that both player's strategies and expected payoffs converge to those of a Nash equilibrium. This result will simplify the application of gradient ascent techniques to more rich policy and problem spaces. The convergence of expected payoffs makes evaluation of policies and hence the credit assignment problem in multiple state problems considerably easier.

[2] An example of such a matrix game is,

$$R = \begin{bmatrix} 0 & 3 \\ 1 & 2 \end{bmatrix} \quad C = \begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix},$$

with the strategy point, $(\alpha, \beta) = (0, 0.9)$. The only Nash equilibrium is $(0.5, 0.5)$. Hence, $V_r(\alpha^*, \beta) = 0.7 < V_r(\alpha, \beta) = 1.1 < V_r(\alpha^*, \beta^*) = 2$, and so the two rules would disagree.

Acknowledgements. This research was sponsored by the United States Air Force under Grants Nos F30602-00-2-0549 and F30602-98-2-0135. The content of this publication does not necessarily reflect the position or the policy of the sponsors and no official endorsement should be inferred.

References

Blum, A., & Burch, C. (1997). On-line learning and the metrical task system problem. *Tenth Annual Conference on Computational Learning Theory*. Nashville, TN.

Bowling, M., & Veloso, M. (2001). Rational and convergent learning in stochastic games. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. Seattle, WA. To Appear.

Carmel, D., & Markovitch, S. (1996). Learning models of intelligent agents. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Fudenberg, D., & Levine, D. K. (1999). *The theory of learning in games*. The MIT Press.

Kuhn, H. W. (Ed.). (1997). *Classics in game theory*. Princeton University Press.

Nash, Jr., J. F. (1950). Equilibrium points in n-person games. *PNAS*, *36*, 48–49. Reprinted in (Kuhn, 1997).

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. The MIT Press.

Reinhard, H. (1987). *Differential equations: Foundations and applications*. McGraw Hill Text.

Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics*, *54*, 296–301. Reprinted in (Kuhn, 1997).

Singh, S., Kearns, M., & Mansour, Y. (2000). Nash convergence of gradient dynamics in general-sum games. *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence* (pp. 541–548). Morgan Kaufman.

Weibull, J. W. (1995). *Evolutionary game theory*. The MIT Press.