Classes of Multiagent Q-learning Dynamics with ϵ-greedy Exploration

Michael Wunder MWUNDER@CS.RUTGERS.EDU
Michael Littman MLITTMAN@CS.RUTGERS.EDU
Monica Babes BABES@CS.RUTGERS.EDU
RL^3 Laboratory, Department of Computer Science, Rutgers University, Piscataway, NJ USA 08854

Abstract

Q-learning in single-agent environments is known to converge in the limit given sufficient exploration. The same algorithm has been applied, with some success, in multiagent environments, where traditional analysis techniques break down. Using established dynamical systems methods, we derive and study an idealization of Q-learning in 2-player 2-action repeated general-sum games. In particular, we address the discontinuous case of ϵ-greedy exploration and use it as a proxy for value-based algorithms to highlight a contrast with existing results in policy search. Analogously to previous results for gradient ascent algorithms, we provide a complete catalog of the convergence behavior of the ϵ-greedy Q-learning algorithm by introducing new subclasses of these games. We identify two subclasses of Prisoner's Dilemma-like games where the application of Q-learning with ϵ-greedy exploration results in higher-than-Nash average payoffs for some initial conditions.

1. Introduction

Q-learning (Watkins & Dayan, 1992) was developed as a reinforcement-learning (RL) algorithm to maximize long-term expected reward in multistate environments. It is known to converge to optimal values in environments that can be formulated as Markov decision processes (Tsitsiklis, 1994). Its elegance and simplicity make Q-learning a natural candidate for application to multiplayer general-sum games, leading to questions about its asymptotic behavior in this context. While the study of simultaneous learning agents has generated much interest, characterization of their behavior is still incomplete. Algorithms such as Q-learning that use accumulated data about the values of actions are of interest beyond RL, as similar mechanisms are hypothesized to exist in mammalian brains (Dayan & Niv, 2008).

In this work, we examine the behavior of two players executing ϵ-greedy Q-learning in a repeated general-sum game. Although some applications of Q-learning have used state representations that include recent history (Littman & Stone, 2001), we focus on a simpler representation consisting of just a single state. Our idealized algorithm is one that consists of infinitely small learning steps making it possible to apply ideas from dynamical systems theory directly to the algorithm (Tuyls et al., 2003). We map the varied behavior of this algorithm using much of the same terminology and methods as has been applied to other multiagent dynamical approaches.

As opposed to a purely *value-based* approach like Q-learning, past work using dynamical systems to analyze multiagent learners has centered on *policy-search* algorithms (Singh et al., 2000) or a mix of the two (Bowling & Veloso, 2001). In cases where learning is equivalent to or resembles policy-gradient algorithms, researchers have found that adaptive methods tend to converge to a Nash equilibrium (Tuyls et al., 2003) or "orbit" a Nash equilibrium (Singh et al., 2000). In the mold of this earlier work, the present paper fully describes the long-run convergence behavior of ϵ-greedy Q-learning—a commonly used algorithm that has not yet been analyzed in this way. A surprising finding of this paper is that when Q-learning is applied to games, a pure greedy value-based approach causes Q-learning to endlessly "flail" in some games instead of converging. For the first time, we provide a detailed picture of the behavior of Q-learning with ϵ-greedy exploration across the full spectrum of 2-player 2-action games. While many games finally converge to a Nash equilibrium, some significant games induce chaotic behavior that averages higher reward than any

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

Nash equilibrium of the game. Since some of these games have a dominant action, this outcome is somewhat counterintuitive. Nonetheless, we show how this behavior is not merely an empirical quirk but a fundamental property of this algorithm, which holds potentially profound implications. To our knowledge, this phenomenon has not been reported elsewhere.

Section 2 derives local learning dynamics for the algorithms we consider. Using a dynamical systems approach in Section 3, we describe the types of asymptotic behavior corresponding to different classes of games. Section 4 goes into greater detail about previously undescribed non-convergent behavior of ϵ-greedy Q-learning in a specific subclass of games. Section 5 compares two learning algorithms.

2. Learning as a Dynamical System

This section introduces IQL-ϵ and summarizes IGA.

2.1. ϵ-Greedy Infinitesimal Q-learning (IQL-ϵ)

The ϵ-greedy Q-learning algorithm selects its highest valued (greedy) action with some fixed probability $(1 - \frac{\epsilon(k-1)}{k})$ and randomly selects among all other $k-1$ actions with probability $\frac{\epsilon}{k}$. Earlier papers have demonstrated superior performance of this algorithm in games (Sandholm & Crites, 1995; Zawadzki et al., November 2008) relative to similar learners and carried out dynamical systems analysis (Gomes & Kowalczyk, 2009) as a model for ordinary Q-learning. However, none has systematically documented the resulting range of outcomes of the dynamics model itself. In particular, we are the first to fully describe the behavior of a deterministic model of the algorithm for all possible games within the 2-person 2-action space.

The (one-state) Q-learning update rule when an agent takes action a and receives reward R

$$Q(a) = Q(a) + \alpha(R + \gamma \max_{a'} Q(a') - Q(a))$$

becomes $\frac{\partial Q(a)}{\partial t} = R + \gamma \max_{a'} Q(a') - Q(a)$ when $\alpha \to 0$. We call this algorithm IQL-ϵ for Infinitesimal Q-learning with ϵ-greedy exploration. The discount rate is γ. We write

- Q_{a_i} and Q_{b_j} for the Q-values of action i for row player RP, and action j for column player CP,
- $\dot{Q}_{a_i}$ for $\frac{\partial Q(a_i)}{\partial t}$, the update of action i for RP and $\dot{Q}_{b_j}$ for the update of action j for CP,
- (r_{ij}, c_{ij}) for the respective payoffs, or rewards, for RP and CP when RP plays i and CP j.

Due to the fact that there can be only one greedy action at a time, IQL-ϵ's updates lead to semi-continuous dynamics best classified as a piecewise-smooth, or hybrid, dynamical system. A *general hybrid dynamical system* (GHDS) is a system $H = [P, \Sigma, J]$ with the following parts (Di Bernardo et al., 2008):

- P is the set of index states or discrete dynamical system states;
- $\Sigma = \bigcup_{p \in P} \Sigma_p$ is the set of ordinary differential equations (or *flows*) for index state p;
- $J = J_{p \in P}$ is the set of jump transition maps.

The simple IQL-ϵ GHDS can be represented as an automaton whose nodes are four complete and separate dynamical system flows and where transitions between the nodes, or index states, must be taken when certain conditions along them are met. When the values for one players actions change ordering, the system jumps to the index state containing the dynamics corresponding to the new greedy policies. For this paper, only one state exists in the agents' environment—all state transitions are jump transitions in this model.

Using this notation, we examine the following equations for a combination of possible greedy actions for the two players. Consider what happens when a_1^* and b_1^* are greedy. RP chooses a_1^* $1 - \frac{\epsilon}{2}$ of the time and $\hat{a_2}$ $\frac{\epsilon}{2}$ of the time, making its expected reward $R_{11} = r_{11}(1 - \frac{\epsilon}{2}) + r_{12}\frac{\epsilon}{2}$ where ϵ is the exploration rate. In this case, RP will update Q_{a_1} according to:

$$\begin{aligned} \dot{Q}_{a_1} &= r_{11}(1 - \frac{\epsilon}{2}) + r_{12}\frac{\epsilon}{2} + (\gamma - 1)Q_{a_1} \\ &= R_{11} + \gamma \max_{a'} Q_{a'} - Q_{a_1}. \end{aligned}$$

However, this equation only describes the rate of update *when the value of a_1 is updated.* To capture the exact rate, consider that the greedy action is taken a fraction $(1 - \frac{\epsilon}{2})$ of the time. In contrast, the non-greedy action is taken $\frac{\epsilon}{2}$ often. Weighting the updates appropriately, when Action a_1 is greedy for both players, the four Q-values obey the following system of differential equations (Gomes & Kowalczyk, 2009), $\Sigma_{a_1^* b_1^*}$:

$$\begin{aligned} \dot{Q}_{a_1} &= (R_{11} + Q_{a_1}(\gamma - 1))(1 - \frac{\epsilon}{2}), \\ \dot{Q}_{a_2} &= (R_{21} + Q_{a_1}\gamma - Q_{a_2})\frac{\epsilon}{2}, \\ \dot{Q}_{b_1} &= (C_{11} + Q_{b_1}(\gamma - 1))(1 - \frac{\epsilon}{2}), \\ \dot{Q}_{b_2} &= (C_{12} + Q_{b_1}\gamma - Q_{b_2})\frac{\epsilon}{2}. \end{aligned}$$

We can find the solutions for the above equations using linear dynamical systems theory (Di Bernardo et al.,

2008). While the solutions define a single dynamical flow $\Sigma_{a_1^* b_1^*}$ where a_1^* and b_1^* are the greedy actions for RP and CP, similar equations can be defined for the other three joint greedy policies. Note that because the system can switch flows, the values may not converge to the end point dictated by this flow alone. We say that the learning algorithm has *converged* if the ratio of strategies in a small period of time stays within a fixed range. See Figure 1 for examples of converging and non-converging policies. Also, note that the equations are deterministic, in spite of the random exploration, because of the infinitesimal learning rate.

2.2. One-player Sliding Greedy Update

In cases in which the convergence points of the flows lie within the index state of a single flow, the above IQL-ϵ analysis is sufficient to disclose the final destination of the algorithm's values. If there is disagreement, the IQL-ϵ GHDS can end up with dynamics that slide along the boundary between two or more index states. An investigation of the resulting dynamics, known as a *Filippov sliding system* (Di Bernardo et al., 2008), is crucial for analyzing these more complex situations.

When one player has two equal Q-values and both sums of discounted rewards are lower than the current value, this player has a *sliding* greedy action. The values may change in lockstep, although the two actions are selected at different rates. Consider what happens when CP has one clear greedy action. Figure 2(inset) shows an illustrated example of this update dynamics. Here, the two actions for RP have the same value and the Q-values for both players drop until CP's greedy action switches. The term "greedy" does not fully capture this type of dynamics for RP because, essentially, its greedy action alternates infinitely often over a given interval so it has no particular greedy action. Instead, define the current *favored action* to be the action f with the higher expected reward during a sliding update (let $\bar{f}$ be the other action). It turns out that f also has a higher probability of play than $\bar{f}$ when both values are dropping. Therefore, f is the action played more often. Define ϕ_f to be the fraction of time where RP plays f. The updates $\dot{Q}_{\bar{f}}$ and $\dot{Q}_f$, taken from the definition of Q-learning, capture the change of respective Q-values over continuous time, observed separately. The formula for ϕ_{fb^*} is the ratio of the non-favored action's update rate to the total update rate while CP's greedy action is b^* and its non-greedy action is $\hat{b}$:

$$\phi_{fb^*} = \frac{\dot{Q}_{\bar{f}}}{\dot{Q}_{\bar{f}} + \dot{Q}_f}$$

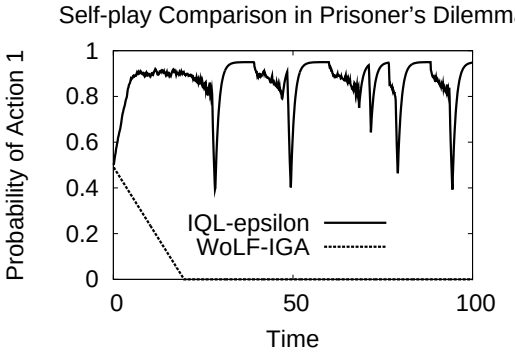

Figure 1. Probabilities for Action 1 for RP in two self-play scenarios both in the Prisoner's Dilemma game. WoLF-IGA is seen to converge to the defection action (Nash), while IQL-ϵ oscillates around a mix of both actions, mostly cooperation. See Figure 2 and Section 5 for more details.

$$= \frac{r_{\bar{f}b^*}(1-\frac{\epsilon}{2}) + r_{\bar{f}\hat{b}}\frac{\epsilon}{2} + Q_{\bar{f}}(\gamma-1)}{(r_{\bar{f}b^*}+r_{fb^*})(1-\frac{\epsilon}{2}) + (r_{\bar{f}\hat{b}}+r_{f\hat{b}})\frac{\epsilon}{2} + (Q_{\bar{f}}+Q_f)(\gamma-1)}.$$

There is a natural intuition behind the ratio ϕ_{fb^*}. Ignoring exploration, if each update is different and negative, the algorithm more often selects the one that decreases more slowly because it is more often the greedy action. In fact, the ratio selected is identical to the ratio of the other value's standalone rate of decrease to the combined rates for both actions. If RP plays f with this proportion, then both values actually decrease at the same overall rate as the faster one is selected less frequently. As a result, the update rates for CP depend on this fraction ϕ_{fb^*}:

$$\begin{aligned} \dot{Q}_{b^*} &= ((1-\phi_{fb^*})c_{\bar{f}b^*} + \phi_{fb^*}c_{fb^*})(1-\frac{\epsilon}{2}), \\ \dot{Q}_{\hat{b}} &= ((1-\phi_{fb^*})c_{\bar{f}\hat{b}} + \phi_{fb^*}c_{f\hat{b}})\frac{\epsilon}{2}. \end{aligned}$$

This reasoning only applies to falling values. If values rise, the arbitrarily chosen greedy one will be updated more rapidly resulting in a positive feedback loop.

2.3. Two-player Sliding Greedy Update

At times, if both players have Q-values at parity, the GHDS may comprise a *dual sliding system*. In the language of hybrid dynamical systems, this situation is equivalent to very low thresholds of switching between index states, meaning that no single flow describes the behavior in this regime. While some definable patterns show up during these periods, researchers in this field acknowledge the potential for unpredictable or chaotic behavior as $\alpha \to 0$ (Di Bernardo et al., 2008).

In some instances, the close distance between values can mean that decisions regarding how to implement continuous estimation can also affect long-run convergence, even for $\alpha \to 0$. There are several ways to define the idealized continuous version of Q-learning

in this situation. For the rest of the analysis, we follow the convention of assuming discrete updates, but keep $\alpha \to 0$. This definition is consistent with the original setup and does not require new assumptions. It also recognizes that the two updates are always separate, even if values are equal. As a result of multiple sliding updates, a solution is no longer provided by solving a single hybrid system of differential equations, thereby complicating exact prediction of behavior. Fortunately, we are still able to clearly predict whether the system moves into a steady attractor for this particular GHDS (Sections 3 and 4).

2.4. Infinitesimal Gradient Ascent (IGA)

IGA (Singh et al., 2000) defines the joint strategy of the players by a pair (p, q), the probabilities of the first action for both players. Strategies are updated in the direction of the gradient of the reward at time t:

$$p_{t+1} = p_t + \alpha \frac{\partial V_r(p,q)}{\partial p} \qquad q_{t+1} = q_t + \alpha \frac{\partial V_c(p,q)}{\partial q}.$$

It has been shown that IGA either leads the strategy pair to a Nash equilibrium or an orbit yielding an average reward equal to the Nash equilibrium. A modified version, WoLF-IGA, always converges to the Nash in these games (Bowling & Veloso, 2001).

3. Classes of 2-player 2-action games

The space of games can be divided according to characterizations of their equilibria. We show how IQL-ϵ behaves in each of these classes. For simplicity, we assume all reward values are distinct. (Ties can make games belong to multiple subclasses, complicating exposition.) Table 1 gives payoff matrices for some of the games we mention. The main results of this section and the paper are summed up by Table 2.

Subclass 1 covers games that only have a single mixed Nash equilibrium, meaning that the players play their actions with probabilities p and q such that $0 < p, q < 1$. The space includes games that meet all of the following conditions: $(R_{11} - R_{21})(R_{12} - R_{22}) < 0$, $(C_{11} - C_{12})(C_{21} - C_{22}) < 0$, and $(R_{11} - R_{21})(C_{11} - C_{12}) < 0$. Zero-sum games like Matching Pennies (MP) are in this category, as is the new Spoiled Child (SC) game. Subclass 2 contains games that have two pure Nashes and one mixed Nash. These games satisfy the following conditions: $(R_{11} - R_{21})(R_{12} - R_{22}) < 0$, $(C_{11} - C_{12})(C_{21} - C_{22}) < 0$, and $(R_{11} - R_{21})(C_{11} - C_{12}) > 0$. Examples of games in this category include Bach/Stravinsky (B/S), Chicken (CH), and some Coordination games. Subclass 3 is the set of games in which at least one of the players has a pure dominant strategy, if $(R_{11} - R_{21})(R_{12} - R_{22}) > 0$ or $(C_{11} - C_{12})(C_{21} - C_{22}) > 0$. Examples in this class include all variants of Prisoner's Dilemma (PD) and Deadlock (DL).

Table 1. Payoffs for representative games in each subclass. RP's rewards are listed first in each pair.

Subclass 1a — Matching Pennies

MP	H	T
H	1, 0	0, 1
T	0, 1	1, 0

Subclass 1b — Spoiled Child

SC	B	M
S	1, 2	0, 3
P	0, 1	2, 0

Subclass 2a — Bach/Stravinsky

B/S	B	S
B	1, 2	0, 0
S	0, 0	2, 1

Subclass 2b — Chicken

CH	D	H
D	15, 15	1, 20
H	20, 1	0, 0

Subclass 3a — Deadlock

DL	b_1	b_2
a_1	1, 1	0, 3
a_2	3, 0	2, 2

Subclass 3b — Prisoner's Dilemma

PD	C	D
C	3, 3	0, 4
D	4, 0	1, 1

Our results explicate the behavior of IQL-ϵ in these various classes, taking exploration into account. For clarity, our analyses generally assume that the initial Q-values are their maximum possible values given the payoff matrix. It is common practice (Sutton & Barto, 1998) to initialize learning algorithms this way and it ensures that the algorithms play all actions greedily for some amount of time before settling down. IGA has its own classes based on the level of payoff sensitivity to the other player's strategy (Singh et al., 2000).

In each of these subclasses, IQL-ϵ further divides the space according to a simple rule so that on one side the algorithm always converges while on the other it is not guaranteed. Define Subclasses 1b, 2b, and 3b such that $\exists i, j \;\; R_{ij} > R_N$ and $C_{ij} > C_N$, where R_N and C_N are either the unique expected Nash payoffs for RP and CP, or the lowest value Nash payoffs (in Subclass 2). Thus, there is some pure non-Nash strategy combination that is a higher payoff than the Nash equilibrium value for both players, much like the cooperative payoff in PD. While IGA gradually alters its strategy toward a best response, IQL-ϵ, in contrast, switches its greedy action suddenly, starving one value of updates. As a result, sometimes an action retains a relatively high value even when not a best response.

Theorem 1 *IQL-ϵ converges to the mixed Nash equilibrium when playing any game in Subclass 1a.*

Proof (sketch) It is clear that no pure strategies will ultimately be successful because the other player can

Table 2. A summary of the behavior of IQL-ϵ and a taxonomy of games.

Subclass	**1a**	**1b**	**2a**	**2b**	**3a**	**3b**
# Pure Nashes	0	0	2	2	1	1
# Mixed Nashes	1	1	1	1	0	0
RP Action 2 is dominant?	No	No	No	No	Yes	Yes
$\exists i,j\ R_{ij} > R_N$ & $C_{ij} > C_N$?	No	Yes	No	Yes	No	Yes
Example game	MP	SC	B/S	CH	DL	PD
IQL-ϵ converges?	Yes	No	Yes	Y/N	Yes	Y/N

exploit any non-Nash strategy. IQL's Q-values converge to tied values where the ratio of actions selected matches that of the Nash equilibrium. In Q-learning, however, the agents can only approximate a mixed strategy by continually shifting their greedy actions.

During learning, the values and strategies seem to cycle repeatedly at points removed from the mixed Nash, only to slowly approach it in the long term. As the greedy strategy combination loops around the payoff table, exploration causes a slight convergence between the greedy and non-greedy value. On the next iteration, the values are closer to the Nash by some factor that depends on ϵ. In the limit, therefore, the policies close in on this equilibrium. □

Theorem 2 *In all Subclass 1b games, IQL-ϵ never converges to a single strategy, pure or mixed.*

Proof (sketch) This result arises because the "cooperative" joint strategy that has higher values than the Nash equilibrium acts temporarily as a better attractor than the Nash whenever both of those actions are greedy. The metaphor of a Spoiled Child (SC) illustrates the dynamics of this subclass, where the parent is RP and the child is CP. There is no pure Nash equilibrium in this class of games, so the IQL-ϵ players first drift towards playing the actions with a ratio resembling the mixed Nash strategy. As the values drop toward the value of this equilibrium, parent and child eventually greedily play the cooperative actions (Spoil and Behave, respectively). These values rise above the lower Nash values toward the cooperative payoff. However, this new attractor is not stable either, because the child would rather Misbehave, at which point the parent prefers the harmful Punish action. Thus, the system switches to a new set of dynamics and the cycle repeats. Unlike Subclass 1a, both greedy actions move away from Nash during cooperation and therefore prevent convergence to a mixed strategy. □

Theorem 3 *IQL-ϵ converges to one of the two pure Nash equilibria in all games of Subclass 2a.*

Proof (sketch) Consider the behavior of the dynamics once all Q-values have dropped below or are at the level of one of the Nash equilibria. At any point, the values move toward the payoffs resulting from the current greedy actions. This payoff either represents one of the pure Nashes or it does not. If it does, the greedy actions remain greedy, as neither player get a higher value by exploring. If the greedy actions do not correspond to a pure Nash, then at some point one player switches greedy actions. The new combination is necessarily a Nash equilibrium by the payoff structure of this class. In addition, the mixed Nash is unstable because the dynamics of the payoff structure deviate from any mixed strategy to one of the pure strategies, returning to the earlier argument for convergence. □

Theorem 4 *IQL-ϵ may or may not converge to one of the two pure Nash equilibria in Subclass 2b.*

Proof (sketch) Some values lead the dynamics to one of the stable pure Nashes, while others cycle much like Subclass 1b (Wunder et al., 2010). □

4. Analysis of Convergence in Dominant Action Games

This section delves into the convergence behavior for Subclass 3 games, which have a dominant action for at least one player. Intuitively, this class seems the simplest—no matter what actions its opponent plays, a player always has an unchanging preferred response. In fact, IGA behaves this way by increasing its dominant action probability until it reaches a Nash strategy. In the IQL-ϵ system, dominant actions can be unstable and lead to sudden shifts of fortune, or even chaotic behavior. The PD time series in Figures 1 and 2 show the strange, non-repeating pattern of updates possible in Subclass 3b, which persists at all learning rates and is therefore an intrinsic property of Q-learning.

4.1. Dominant Action Games: Subclass 3a

Call RP's action a_2 *dominant* when $R_{11} < R_{21}$ and $R_{12} < R_{22}$. If $\neg(\exists i,j\ R_{ij} > R_N$ and $C_{ij} > C_N)$, the game is a member of Subclass 3a.

Theorem 5 *In Subclass 3a games, IQL-ϵ converges to the pure Nash equilibrium identified by one player's dominant action and the other player's best response.*

Proof (sketch) If there is no payoff preferable to the Nash for both players involving RP's non-dominant action a_1, it simply plays a_2. Once RP's Q-values drop below $\frac{R_{21}}{1-\gamma}$ or $\frac{R_{22}}{1-\gamma}$, no other payoff can attract the algorithm to play a_1. At that point, CP is faced with

a static optimization problem and its values inevitably converge to $\frac{C_{21}}{1-\gamma}$ and $\frac{C_{22}}{1-\gamma}$. Therefore, IQL-ϵ converges to the Nash equilibrium by definition. □

4.2. Conditions for Convergence in Subclass 3b

Define Subclass 3b as the remaining dominant action games, those for which $\exists i, j\ R_{ij} > R_N$ and $C_{ij} > C_N$. Prisoner's Dilemma resides in this class, where unexpected dynamics proceed according to a predictable sequence of update *phases* (Figure 2 and Table 3). Each phase represents a temporarily stable combination of index states and flows that persist until conditions change. The phases arise from a conflict between the selfish incentive of the dominant action and the mutual benefit of cooperative action. In a pattern similar to human conflict, the flows transition from I, *peaceful cooperation*, to II, *aggressive war*, to III, *domination*, to IV, *rebellion*, then either back to I or to DD, total war. *Aggressive war* breaks the peace so that one player's values gain advantage, while the *rebellion* results from the reaction of the long-dominated player. These repeated phases form a chaotic attractor due to the dual sliding update (Figure 2(right)).

For the dynamics to converge to the Nash equilibrium, one of the players (CP, for instance) must sustain the dominant action greedily against *both* actions of RP so RP's values can fall. Figure 2(inset) shows an example of this condition in phase IV. To keep the values decreasing to the Nash point, the players must switch roles before both cooperative actions become greedy again, thereby perpetuating phase IV. Once one of the players can defect greedily against the other's greedy defection, convergence to Nash is assured. The value below which mutual defection (DD) is inevitable is the following threshold Q_{DD}, found when (non-greedy update) $\dot{Q}_{\hat{b_1}}$ is less than (greedy update) $\dot{Q}_{b_2^*}$:

$$\frac{\epsilon}{2}(C_{21}+(\gamma-1)Q_{DD}) < (1-\frac{\epsilon}{2})(C_{22}+(\gamma-1)Q_{DD}) \quad (1)$$

$$Q_{DD} < \frac{-\frac{\epsilon}{2}C_{21} + (1-\frac{\epsilon}{2})C_{22}}{(1-\gamma)(1-\epsilon)}. \quad (2)$$

As CP's values decrease during phase III, they drop below a defection threshold (DT) where exploring $Q_{\hat{b_1}}$ drops faster than greedy $Q_{b_2^*}$. In this case b_2^*, D, is greedy in response to mixed actions of RP. Say $C_{\phi 22}$ is the D reward against RP's sliding update. Like above, Q_{DT} is defined by the inequality $\dot{Q}_{\hat{b_1}} < \dot{Q}_{b_2^*}$:

$$Q_{DT} \quad < \quad \frac{-\frac{\epsilon}{2}C_{\phi 21} + (1-\frac{\epsilon}{2})C_{\phi 22}}{(1-\gamma)(1-\epsilon)}.$$

These dynamics from Section 2.2 imply ϕ_{22} is equivalent to the percentage of time that RP spends playing a_2 when its dropping values are equal and CP is playing b_2 greedily. In general, ϕ_{22} rises as values decrease.

Table 3. Properties of phase dynamics during repeated Prisoner's Dilemma by phase. These phases repeat in the order identified in Figure 2. (Arrows denote transitions.)

Comparison	I	II	III	IV
RP Value Q_{a_1} ? Q_{a_2}	$>$	$<$	$=$	$=$
CP Value Q_{b_1} ? Q_{b_2}	$>\rightarrow<$	$<\rightarrow=$	$=$	$<$
RP Update $\dot{Q}_{a_1}$? $\dot{Q}_{a_2}$	$>\rightarrow<$	$<\rightarrow>$	$\leq$	$=$
CP Update $\dot{Q}_{b_1}$? $\dot{Q}_{b_2}$	$>$	$<\rightarrow=$	$=$	$<\rightarrow>$

An important cooperation threshold (CT), Q_{CT}, relates to the level where $\dot{Q}_{\hat{a_2}} > \dot{Q}_{a_1^*}$. Essentially, if both of a player's values are very close and above Q_{CT}, it cannot cooperate for long before $Q_{\hat{a_2}}$ overtakes $Q_{a_1^*}$:

$$\frac{\epsilon}{2}((\gamma-1)Q_{CT}+R_{21}) \geq (1-\frac{\epsilon}{2})((\gamma-1)Q_{CT}+R_{11}) \quad (3)$$

$$Q_{CT} \geq \frac{(1-\frac{\epsilon}{2})R_{11} - \frac{\epsilon}{2}R_{21}}{(1-\epsilon)(1-\gamma)}. \quad (4)$$

As long as $Q_{CT} \leq Q_{DT}$ for some player, then convergence to the Nash equilibrium is assured because it has nothing to lose by defecting. If this condition is true for long enough, the other player may be in a position to trigger a chain of defections leading to the Nash.

Phase IV, observed in the closeup Figure 2(inset), either leads to convergence or back to I, depending on its length and its values when it commences. It begins when CP plays D greedily against greedy C below the threshold Q_{DT}, thereby dropping RP's values. If phase IV begins with Q_{b_2} just below Q_{DT}, then it will be too short and convergence cannot happen, as the flow returns to peaceful cooperation and the cycle restarts. However, phase IV might not begin as soon as Q_{b_2} crosses the threshold if the possibility of transitioning to the crucial index state is zero, regardless of the continuity of the updates. Delaying phase IV makes CP eventually defect for longer periods, increasing the likelihood of convergence to Nash. In the case of PD, this question is settled during phase III, the dual sliding update. Essentially, CP must first erase RP's gains made when RP defected against its C with two Ds. After two defections, RP cooperates, but now so does CP, so phase III continues. Some games prevent the onset of IV below defection threshold Q_{DT} until ϕ_{22} rises above its own threshold. Once it is known where phase IV must begin as $\alpha \to 0$, one iteration is enough to show whether the system converges. We have mapped the region of symmetric games where uniform IQL-ϵ does or does not converge to the Nash equilibrium (Wunder et al., 2010).

Theorem 6 *In Subclass 3b games, certain starting values guarantee the IQL-ϵ dynamics converge to the*

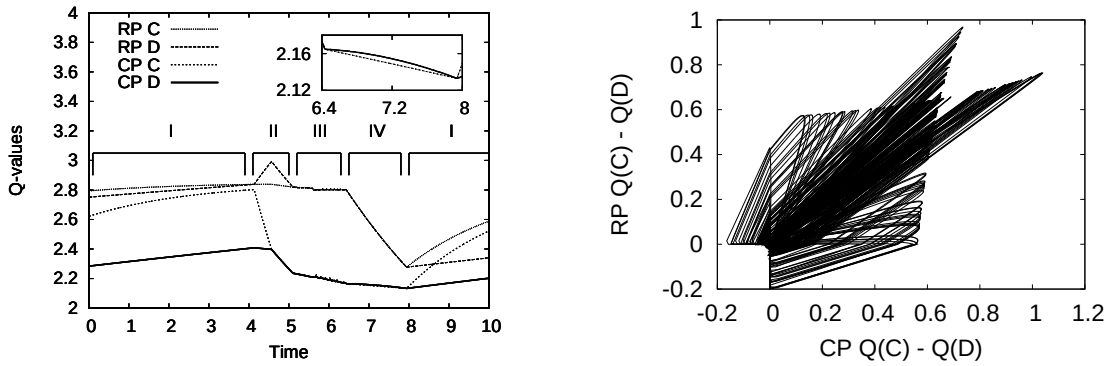

Figure 2. (left) The sequence of phases I-IV during PD with IQL-ϵ agents. The value ordering is documented in Table 3. Some of these phases exist entirely within a single index state of the GHDS (I), while others rotate between all four index states (III). In the **peaceful cooperation** phase I, both agents cooperate greedily. Eventually, via exploration, the defection value appeals to one of the players, RP, leading to **aggressive war** (II). In II, RP forces both of CP's values to drop until neither player has a clear greedy action. Phase III, **domination**, is the dual sliding update, so that the algorithm alternates between mutual cooperation and one player defecting. When CP's values drop below the Q_{DT} threshold, it becomes profitable to defect against both actions of the other player, initiating **rebellion** (IV). After this final phase, both players re-enter peaceful phase I, thereby renewing the cycle. (inset) Close-up of phase IV of the cycle. (right) A 2-D projection of the chaotic attractor over many cycles. The system switches flows along the x=0 and y=0 axes. Note how at these values the system avoids the unstable mutual defection flow in the lower left quadrant.

pure Nash. For other values, the dynamics do not permanently converge to any particular strategy, but average rewards of both players are higher than the Nash.

Proof (sketch) There exists a DD threshold Q_{DD} below which two-player greedy defection ($\Sigma_{a_2^* b_2^*}$) is a sink and does not jump to another index state. Starting values that meet this condition, or lead later to values that do, converge to the Nash. In addition, a high R_{21} value that delays phase IV while $Q_{DT} \geq Q_{CT}$ encourages convergence as RP can defect freely.

Other starting values enter a series of phases. If phase IV always occurs immediately after the Q-values drop below Q_{DT}, mutual cooperation ($\Sigma_{a_1^* b_1^*}$) temporarily attracts the cooperation Q-values away from dominant action a_2 and convergence does not result. Delayed onset of phase IV, meanwhile, can lead to sustained greedy defection and convergence. If neither players' values ever drop below the threshold Q_{DD}, by the construction of Q-learning the players must be receiving higher average values than the Nash values. □

Although IQL-ϵ does not converge in the usual sense in some games, certain strategies are clearly prevalent for low ϵ. In games where IQL-ϵ does not reach a Nash equilibrium, the system remains in the cooperative outcome for time proportional to $\frac{1}{\epsilon}$. Therefore, as $\epsilon \to 0$, the strategies played by IQL-ϵ converge to C.

5. Empirical Comparison

We ran IGA and IQL-ϵ in a representative game from each class, using the payoffs in Table 1, for 100 simulated units of continuous time. We approximated the solutions numerically ($\alpha = 0.0001$) and used parameters of $\gamma = 0$ and $\epsilon = 0.1$. We chose starting Q-values distinct from the Nash values to allow the algorithms to demonstrate their full behavior. Figure 3 provides a time-series plot of the Q-values for representative games. Larger values of α show the same patterns, but with more noise (Gomes & Kowalczyk, 2009).

In MP, the two algorithms essentially behave the same way, ultimately converging to Nash. DL (not shown) converges simply and similarly for both algorithms.

Both algorithms converge in B/S but identical starting points may lead IQL-ϵ and IGA to find different equilibria (coordinating on B vs. S). IGA converges to a pure Nash in CH (not shown), and IQL-ϵ sometimes converges. In the case of cyclic activity, it manages a minimum average reward of 5.7, higher than either the mixed Nash (3.3) or lower pure Nash (1).

IQL-ϵ never converges in SC, but IGA will converge to the mixed Nash. Once again, we see IQL-ϵ attaining higher reward than IGA; around 2 and 1.2 for the two players instead of 1.5 and 1. These observations provide clues about the diverse and sometimes beneficial nature of non-convergence as well as important similarities within classes. In contrast to PD, games in this subclass reach a periodic attractor.

Finally, the PD series (Figure 1) compares the policies of IQL-ϵ with IGA. IGA converges to the Nash (DD). While low initial values will lead IQL-ϵ to DD, here IQL-ϵ does not converge for the chosen starting values. IQL-ϵ meets all conditions that describe a chaotic pat-

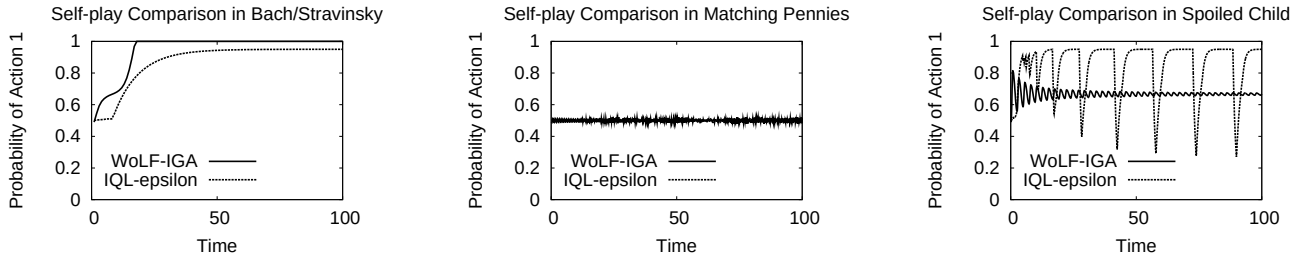

Figure 3. Action probabilities for RP in two self-playing algorithms in representative games (Table 1). The policies of WoLF-IGA converge, while the IQL-ϵ dynamics do not for some starting values in the PD or SC games. Both agents converge to one of the pure Nashes in B/S and the mixed Nash in MP. In SC, the IQL-ϵ players oscillate periodically while WoLF-IGA reaches the fixed point defined by the mixed Nash of the game. See Section 5 for more details.

tern. The average reward obtained by the IQL-ϵs is around 2.75 and 2.45, exceeding IGA's value of 1.0.

6. Conclusion

Motivated by the important role of Q-learning in multiagent RL, we set out to catalog the dynamics of a continuous-time variant. We documented a wide range of outcomes in 2-player, 2-action games, varying from rapid convergence to Nash to unceasing oscillations above the Nash. Of particular interest is the complex behavior of Q-learning with ϵ-greedy exploration in Prisoner-Dilemma-like games, since the algorithm is able to achieve higher-than-Nash outcomes in this previously undiscovered chaotic system. The increasing prevalence of mutually cooperative non-Nash strategies as exploration is decreased to zero is itself worthy of investigation. We see no reason to eliminate the possibility that this result would arise in games with more players or actions.

Future work in this setting will investigate the impact of irregular oscillatory behavior in larger games or real-world systems. The current work may provide insight into important related settings like games with more players or actions as well as Q-learning in games with history states and in stochastic games.

7. Acknowledgements

The authors are supported via NSF HSD-0624191 and thank S. Singh and M. Kaisers for discussions.

References

Bowling, Michael and Veloso, Manuela. Rational and convergent learning in stochastic games. *Proceedings of the Seventeenth IJCAI*, 2001.

Dayan, Peter and Niv, Yael. Reinforcement learning and the brain: The good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185–196, 2008.

Di Bernardo, Mario, Budd, Chris, and Champneys, A. R. *Piecewise-smooth Dynamical systems: Theory and Applications*. Springer-Verlag, 2008.

Gomes, Eduardo Rodrigues and Kowalczyk, Ryszard. Dynamic analysis of multiagent Q-learning with e-greedy exploration. *Proceedings of the 2009 International Conference on Machine Learning*, 2009.

Littman, Michael L. and Stone, Peter. Implicit negotiation in repeated games. In *Eighth International ATAL Workshop (ATAL-2001)*, pp. 393–404, 2001.

Sandholm, Tuomas W. and Crites, Robert H. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37:144–166, 1995.

Singh, Satinder, Kearns, Michael, and Mansour, Yishay. Nash convergence of gradient dynamics in general-sum games. In *Proceedings of UAI*, 2000.

Sutton, Richard S. and Barto, Andrew G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Tsitsiklis, John N. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3): 185–202, 1994.

Tuyls, Karl, Verbeeck, K., and Lenaerts, T. A selection-mutation model for Q-learning in multi-agent systems. *2nd International AAMAS*, pp. 693–700, 2003.

Watkins, Christopher J. C. H. and Dayan, Peter. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

Wunder, M., Littman, M., and Babes, M. Classes of multiagent Q-learning dynamics with e-greedy exploration. Technical report, Rutgers University DCS-tr-670, 2010.

Zawadzki, E., Lipson, A., and Leyton-Brown, K. Empirically evaluating multiagent learning algorithms. *Working Paper*, November 2008.