
Extending Implicit Negotiation to Repeated Grid Games

Robin Carnow

Department of Computer Science, Rutgers University, New Brunswick, NJ 08901 USA

RCRS@PAUL.RUTGERS.EDU

Abstract

Leader-follower relationships have been shown to be successful in implicit negotiations between autonomous agents via their interactions in repeated, bimatrix, general sum games. This paper evaluates the fixed leadership strategies, Bully and Godfather, in the new, more complex venue of grid games. Extended versions of best-response Q-learners as follower agents were used, to show that implicit negotiations can coordinate autonomous agents and yield greater utility, even in an environment of greater complexity. Furthermore, I will show the inherent advantages of signaling in the coordination of agents. The testbed is a stochastic grid game with many similarities to “Chicken”. The merit of each leader and follower was evaluated by looking at the convergence of Q-learners, the distribution of rewards for each agent, and the move efficiency of game play.

1. Introduction

Previous research by Littman and Stone (2001) showed that a Nash equilibrium can be established by autonomous agents with the use of implicit threats in the context of repeated, bimatrix games (two-player, general-sum). In their research, they used leader agents implementing a fixed policy, and follower agents using Q-learning, to learn an optimum policy given the leader’s actions. They evaluated two different leadership strategies. The first was a greedy leader called Bully, which takes actions that yield the highest reward for itself under the assumption that the follower will choose a best response. The second leader, called Godfather (GF), is a two-state, generalization of the tit-for-tat strategy (Axelrod, 1984) that allows the follower to take the action that will earn a greater reward for both agents. If the follower takes the wrong action, GF uses a greedy strategy that forces the follower to only earn a small reward (called its *security level*). Through this conditioning, GF and a follower, were able to establish a high yielding, Nash equilibrium. The follower learned what actions would cause GF to use a greedy strategy, and the highest expected reward in that case (i.e. its security level). At that point the follower’s policy has no incentive to change since any alternative

policy would be suboptimal. By definition, this is a Nash equilibrium (Nash, 1951).

This paper evaluates extended versions of these two leadership strategies, as well as two versions of follower agents, in the new venue of Markov games. The game in this evaluation is called Grid Game 2 (GG2) (Hu & Wellman, 2000), which is a stochastic version of “Chicken”, that originally had two deterministic asymmetric equilibrium policies, and a stochastic symmetric equilibrium. With some alterations, GG2 has a new deterministic *symmetric* equilibrium policy. If both agents coordinate and wait for each other to position properly, both can enter the goal state at the same time, earning substantially larger payoffs. This combination of policies is called a *targetable pair*.

In this game Godfather’s implicit threat is “Fulfill your half of the targetable pair or else”. The GF agent makes use of the new ability of expressing its *mood*, trusting or angry, during game play. I will show that this added information dramatically improves the coordination between leader and follower.

The follower agents will use Q-learning to find a best-response to the leaders’ actions. The difference between Q-learners is that one can perceive Godfather’s mood and the other cannot. Using this disparity, we will be able to see the difference that GF’s signal makes in coordination.

The purpose of this evaluation is to explore the advantages of using implicit negotiations in Markov games and to unmask the integral nature of signals in the coordination of a learner and a multi strategy leader.

2. Grid Game 2

The grid game used in this paper is an instance of a Markov game, which is a tuple $(I, S, A_i(s), P, R_i)$, where I is a set of n players, S is a finite set of states, $A_i(s)$ is the i^{th} player’s set of actions at state $s \in S$, P is a probability distribution over transitions conditioned on the current state and joint actions taken, and $R_i(s, \vec{a})$ is the i^{th} player’s reward for state s and joint actions $\vec{a} \in A(s) = A_1(s) * \dots * A_n(s)$.

The source of experimental results is an altered version of Grid Game 2 (GG2) from Hu & Wellman (2000). The grid game is made up of a three cell by three cell square, where the initial states for agent A and agent B are the

lower right and left corners respectively. There is only one goal for both agents in the center cell at the top row of the grid. There are also two barriers in the upper side of each agent's initial cell position. If an agent tries to move through the barrier, the move fails with probability .5. Figure 1 depicts the grid world as described.

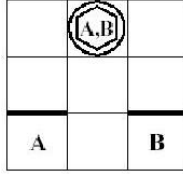


Figure 1. Grid Game 2 in its initial state with player A in cell 0,0 and B in cell 0,2. The goal state is marked in top-center cell.

Actions are executed simultaneously. The action set is North, South, East, West, and StayPut. If the two agents move into the same cell they each receive a reward of -1 and neither agent changes position. If an agent reaches a goal by itself it scores 100 points. The game is over as soon as any agent enters the goal. If both agents enter the goal simultaneously they are each awarded 200 points. In all other cases, moves have reward of 0.

2.1 Equilibrium Policies

GG2 originally resembled the game, “Chicken”, where each agent has the opportunity to make a bold move and earn more rewards by passing between the barriers. If both agents take this bold action, they will collide and both will receive -1 for their rewards. GG2 allowed for exactly two deterministic equilibrium policies, where one agent travels through the barrier and the other travels up the middle or vice versa. These are called an asymmetric equilibria since the agent that moves up the middle will earn an average score of $(100+200)/2 = 150$, and the agent that moves up the side, going through the barrier, will score $(0+200)/2 = 100$ on average.

The addition of the action StayPut allows for new deterministic *symmetric* policies, where an agent can wait for the other, so that both can enter the goal together. This would earn an average reward of 200 for both agents, which is the highest score. One way this can happen is if the leader takes actions that will condition the follower to wait for it. In the following section, I discuss the algorithms needed to produce this end result.

3. Leader and Follower Algorithms

The two leader agents will use simple algorithms that cause the follower agents, assumed to be implementing a best-response strategy, to converge to a coordination point. Each follower chooses actions to maximize its reward, given the follower's experience in previous games.

3.1 Q-learning

The follower agents, Q_0 and Q_1 , execute Q-learning. The difference between Q_0 and Q_1 is that Q_1 is given a signal indicating when the GF is “angry” and Q_0 is not.

Each Q-learner incorporates the actions of other agents in a Markov game (Littman, 1994). The equation to calculate a Q-value for the i^{th} player, given a state s , and joint actions of all players \bar{a} is given by:

$$Q_i(s, \bar{a}) = (1-\gamma) R_i(s, \bar{a}) + \gamma \sum_{s'} \Pr[s'|s, \bar{a}] V_i(s') \quad (1)$$

where $V_i(s') = \max_{\bar{a} \in A(s)} Q_i(s', \bar{a})$ is the state value of the future state s' calculated by finding the future action that maximizes the Q-value at s' given the joint action of all players. This is called Friend-Q, because the learner chooses the optimal action assuming that the other agents will choose their next action to be one that maximizes the reward for the learner (Littman, 2001). $\Pr[s'|s, \bar{a}]$ is the transition probability of having s' as the next state given the current state and joint actions taken. The discount factor $0 \leq \gamma < 1$, controls how much weight is given to present reward versus future reward. This creates a horizon of considered future states which solves the problem of infinite sums by reducing the weight of future reward by a factor of γ for each step farther away from the current step. $R_i(s, \bar{a})$ is the i^{th} player's reward obtained for state s and joint action \bar{a} .

The update equation to improve Q-value estimates after a transition from state s to state s' is:

$$Q_i(s, \bar{a}) = (1 - \alpha) Q_i(s, \bar{a}) + \alpha [(1-\gamma) R_i(s, \bar{a}) + \gamma V_i(s')] \quad (2)$$

where α is the learning rate, $0 < \alpha < 1$, which dictates how much weight is given to new versus old experience.

Both agents will choose actions according to the ϵ -greedy policy. This chooses a random action with probability ϵ or the action that maximizes $Q(s, \bar{a})$ with probability $1 - \epsilon$.

3.2 Bully Leader Strategy

The Bully leader strategy is deterministic and has one state. Bully takes the action which will maximize its own reward in a particular state assuming the follower will learn a best response to that action after revisiting the state a sufficient number of times. For example, in GG2, the first move of Bully will be West (i.e. the bold move) since that would be the move that gets the Bully to the goal. The follower takes on the “chicken” role and moves North, through the barrier, once it has learned Bully's first move.

Since Bully is assuming that the follower is adaptive, and will learn that colliding with Bully is not the optimal policy, Bully chooses actions that maximize its rewards, without taking into account the negative rewards of collisions. This means that the Bully will keep taking the action that will bring it to the goal deterministically, no matter what action the other agent takes. Because of the fixed nature of Bully's policy the follower agent will react to Bully's next board position in the same way it would if

the Bully’s path was just another feature of the grid world. As Littman and Stone (2001) pointed out, this relationship between leader and follower is “Nash-like”. The follower is maximizing its payoffs assuming Bully stays fixed, and Bully has chosen to behave in a way that optimizes its payoffs assuming the other agent is a best-response learner. Note that this Nash-like relationship is critical between leader and follower. To prove this to yourself imagine a game where two Bully agents play. They would collide forever since both would expect the other to learn how to stay out of its way.

3.3 Godfather Leader Strategy

The Godfather leader strategy consists of a two state strategy, “trust” and “angry”. Depending on the conformity of the follower agent to fulfill its half of the targetable pair, the GF will transfer from happy to angry or vice versa.

GF initially starts out using the happy strategy and gives the Q-learner an opportunity to “do the right thing”. In GG2 the “right thing” would be for the follower to move to a cell that is adjacent to the goal and wait until GF is in the center cell; as a result, both players can move into the goal simultaneously. If GF is using a happy strategy it stays put for one move (to allow the follower to move East if it chooses to). GF then moves West (since the follower should move north on its second move), proceeding North to the center cell on the third move. At this point GF stays put for one move if the follower is not in one of the upper corners. Finally, GF moves North to enter the goal.

If the follower fulfills its half of the targetable pair, then GF will keep using the happy strategy and both agents will be rewarded with very high reward. On the other hand, if the follower gives into the temptation of entering the goal state before GF, GF will set an angry flag and it will transfer to its angry strategy. GF’s angry strategy is the same as Bully’s strategy. Using its angry strategy, GF will force the follower to only earn its security level of reward. An agent’s security level in GG2, is the reward an agent can expect to receive playing an opponent that is greedily taking actions to maximize its own reward. After GF has transferred to the angry strategy, it continues in an angry state during four games, then it transfers back to the trusting strategy. This way the Q-learner is deterred from angering the GF.

4. Experiments

In the evaluation of the above mentioned leaders and followers, measurements were recorded to capture: convergence of Q values, move efficiency, distribution of rewards to each agent, and average probability of a follower angering GF.

There were 100 experiments each with 1 million moves. An experiment was run for each of the following pairs: Bully vs. Q₀, GF vs. Q₀, GF vs. Q₁, and Q₀ vs. Q₀. In the

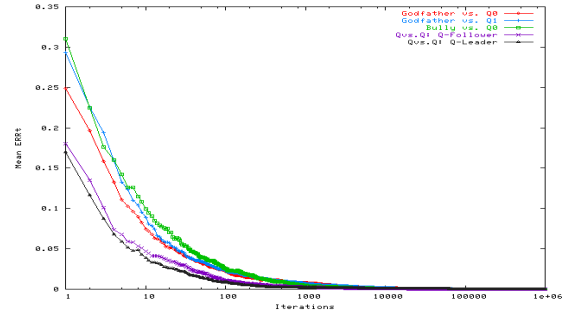


Figure 2. Average convergence of Q-learners in grid game 2.

experimental runs both Q₀ and Q₁ will have the following parameter settings: $\gamma = 0.9$, $\alpha = 0.1$, and $\epsilon = 0.1$.

4.1 Convergence Measures

The error at time t for agent i is the difference between $Q_i(s, a)$ at time t and $Q_i(s, a)$ at time $t-1$:

$$ERR_i^t = |Q_i^t(s, a) - Q_i^{t-1}(s, a)|.$$

Figure 2 plots the means of ERR_i^t over time. The x-axis represents the move count (in log scale), and the corresponding y-values are the means of ERR_i^t for all $t = 0, \dots, x$. Each point is the average of the means of ERR_i^t over the 100 experiments.

The plots show Q values converging to zero for Q₀ and Q₁ in each match. Although this shows faster convergence in matches that are less likely to coordinate, the important questions to ask are: “What policies do these agents converge to?” and “What are the values of these coordination points?”

4.2 Utility of Equilibrium Policies Reached

To compare the equilibrium policies reached by each agent pair, the mean and standard deviation of rewards and games accomplished were calculated. These measures, shown in Table 1, were taken over the last 30,000 moves of each experiment.

As shown, both GF matches were effective in not only earning more reward than Bully, but also yielding greater symmetry of reward between leader and follower. This is to be expected since GF is encouraging a Nash equilibrium through conditioning its followers, while Bully is maximizing reward for itself, causing the follower to earn only its security level. In this game, Bully also earns less reward in the long run since it will only score the larger reward if the follower is lucky enough to surpass the barrier (i.e. 50% of the time). Bully does have the characteristic of greatest games finished. This is because of Bully’s use of only one policy. It is this fixed policy that allows the Q-learner to converge more quickly and stay in a stable asymmetric equilibrium.

Q₀ vs. Q₀ both have a continuum of different policies, which contributes significantly to the large standard deviations in their rewards. They are very effective in completing games in the least amount of moves. After

Table 1. Averages and standard deviations (in parenthesis) of rewards and games accomplished.

Experiments	Leader Reward	Follower Reward	Game Count
Bully vs. Q_0	129.458 (11.349)	59.051 (22.623)	9577.33 (231.807)
Godfather vs. Q_0	158.125 (20.258)	131.24 (30.243)	7085.24 (681.866)
Godfather vs. Q_1	178.244 (5.395)	156.742 (10.753)	6565.68 (319.427)
Q_0 vs. Q_0	117.549 (46.065)	136.188 (55.586)	9371.08 (348.526)

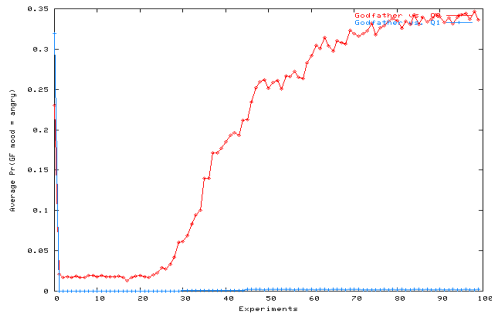


Figure 3. Change in average probability of each Q-learner angering GF over 100 experiments.

some closer inspection, I found that the Q-learners also found a symmetric equilibrium by using the StayPut action, and waiting for an agent if it was unable to make it through the barrier. This was due to the larger payoff for both agents entering the goal at the same time. If the difference between non-cooperative and cooperative goal was less, the agents might have greedily entered the goal and not waited.

Godfather vs. Q_1 had the greatest yielding equilibrium: with the greatest score, best symmetry, and least standard deviation. These results reaffirm that a Tit-for-tat type strategy will earn higher average reward in comparison with a greedy strategy.

4.3 Advantages of Signals in Coordination

To show the vital importance of signal use in the conditioning of a learner by a multi strategy leader, statistics were gathered to show the change in average probability of each Q-learner angering Godfather over 100 experiments. Figure 3 shows Q_1 quickly learning how to keep GF happy where as Q_0 actually angers the GF more each experiment. This phenomenon is caused by the lack of knowledge Q_0 has of GF's mood. Q_0 does not see a difference between a state where GF is angry, and the same state where GF is trusting. Therefore, as Q_0 angers GF more, Q_0 's policy converges to the optimum policy of reacting to a greedy leader (just as it did with Bully). By doing this Q_0 's policy becomes increasingly greedy which causes GF to be angry more of the time. The end result of this defect is that GF's conditioning of Q_0 is ineffective in teaching the value of coordination.

5. Conclusion

This research contributes further evidence to the argument

that best-response agents can be led to stable, high yielding, Nash equilibria by the use of implicit threats. This was shown by introducing two fixed leadership strategies that were extended to the more complex venue of Markov games. Experimentation with a version of GG2, demonstrated that two followers are not as effective in finding a stable equilibrium when compared with agents in a leader-follower relationship. Furthermore, it was shown that Godfather, a generalization of tit-for-tat, was able to stabilize a symmetric Nash equilibrium that earned a greater, more consistent score than a greedy leader, called Bully.

This paper also brought to light the indispensable need for signals to condition learners when the leader implements a multi-state strategy. In fact, these results show that a best-response learner will converge to a suboptimal asymmetric equilibrium as a result of the lack of signals. This was due to a chain reaction where, after the imperceptive agent angered Godfather, it took the angry strategy to be a new characteristic in the environment. This caused it to converge to the Bully-led policy, forcing GF to stay angry. In future research dealing with coordination, it is expected that signals will be used between adaptive agents to quickly facilitate stable strategy transitions.

Acknowledgements

I thank Michael Littman for all of his helpful and inspiring discussions.

References

- Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.
- J. Hu and M. Wellman. Experimental results of multiagent reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 407-414, July 2000.
- Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of Eleventh International Conference on Machine Learning*, pages 157-163. Morgan Kaufmann, 1994.
- Michael L. Littman. Friend or foe Q-learning in general-sum Markov games. In *Proceedings of Eighteenth International Conference on Machine Learning*, pages 322-328, June 2001.
- Michael L. Littman and Peter Stone. Implicit negotiation in repeated games. In *Proceedings of The Eighth International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, August 2001.
- J.F. Nash. Non-cooperative games. *Annals of Mathematics*, 54:286-295, 1951.