Correlated-Q Learning - Greenwald and Hall 2003

Hermann Hans
Department of Computer Science, Georgia Institute of Technology

April 20, 2017

Abstract

This report attempts to reproduce a subset of the findings that were presented in the Amy Greenwald and Keith Hall's paper, 'Correlated Q-Learning' (2003) [1]. Their paper covers four variants of Correlated-Q (CE-Q) learning algorithms and their analysis on several grid games. Our work focuses on the reconstruction of the experiments on soccer [2], describing theory, implementation, and a comparison between the results of this report and Greendwald's paper. A video with a quick overview of the reconstruction of these findings is available at`TODO`

1 Introduction

An interesting and theoretically grounded extension to the reinforcement framework is its application to learning optimal strategies for Markov games. From a game theoretic perspective, this is motivated by a desire for multi-agent game playing systems which converge to equilibrium policies. Nash Equilibria in games are defined as solutions induced by agent policies in which no single player may increase its expected utility by unilaterally changing its policy. This concept is further extended by Littman in [3] to include adversarial and coordination equilibria. An adversarial equilibrium has the standard Nash property and additionally no agent i is hurt by any change of other players' policies. Conversely, coordination equilibria define solutions in which the game yields maximal utility for all players equivalent to reward maximization over the joint space of agent actions; That is, a version of the game in which a controller selects all agent actions to maximize cumulative reward. Not all games possess an adversarial equilibrium, however in two player zero-sum games (as in the soccer game explored in the present work, initilly proposed in [2] and further in [1]) *all* Nash equilibria are adversarial.

1.1 Markov Games

As discussed in [1], stochastic games are a generalization of Markov Decision Processes (MDPs) and repeated games. Greenwald offers an excellent summary of the properties of Markov games and we will briefly sketch her notation here for the purposes of clarity.

A stochastic game is a tuple $\langle I, S, A, P, R\rangle$. Where I is a set of n players, S is the set of states, $A = \{A_i(s) : s \in S, 1 \leq i \leq n\}$ where $A_i(s)$ is the ith player's set of actions in state s, P is the state transition function (as in MDPs) that is a probability distribution of observing a succeeding state s' conditioned on the prior state s and the joint over player actions. The reward function R is similarly defined over the joint of player actions $R_i(s, \vec{a})$ which denotes the ith player's reward for state $s \in S$ and a joint action of all the players. We may then define a state-action value function for each player as:

$$Q_i(s, \vec{a} = (1-\gamma)R_i(s, \vec{a}) + \gamma \sum_{s'} P[s'|s, \vec{a}]V_i(s') \tag{1}$$

Greenwald observes that of central importance to a rigorous definition of Markov games is an appropriate selection of the value function $V(s)$. In the context of our soccer playing reproduction, we identify four important variants. Q-learning (not shown here), foe-Q, friend-Q and Correlated-Q.

As proposed by Littman [2] the foe value function for two player zero-sum Markov games which incorporates von Neumann's minimax function in place of the usual max operation as per. Bellman's Equation. We adopt Greenwald's notation of using $\Sigma_i(s)$ to indicate a probabalistic action space of a player i in a state s.

$$V_1(s) = \max_{\sigma_1 \in \Sigma_1(s)} \min_{a_2 \in A_2(s)} Q_1(s, \sigma_1, a_2) = -V_2(s) \tag{2}$$

Littman [2] also proposes a similar value function which will cause Q-learning to converge to a collaborative equilibria (if one exists). It is useful (in terms of implementation) to recognize that this is simply the usual Q-learning value function over the joint action space of both agents.

$$V_1(s) = \max_{\vec{a} \in A(s)} Q_i(s, \vec{a}) \tag{3}$$

$$V_1(s) = \max_{a \in A_i(s)} Q_i(s, a) \tag{4}$$

Finally we consider the Correlated-Q equilibrium selection function. For our purposes we only consider the utilitarian form (uCE-Q). Let CE be a Correlated Equilibrium, that is, a probability distribution over the joint action space of both agents in which all agents optimize with respect to all other agents' probabilities conditioned on their own.

$$\sigma \in \operatorname*{argmax}_{\sigma \in CE} \sum_{i \in I} \sum_{\vec{a} \in A} \sigma(\vec{a}) Q_i(s, \vec{a}) \tag{5}$$

Assuming the utilitarian correlated equilibrium selection function, we may define uCE-Q as: $CE_i(\vec{Q}(s)) = \{\sum_{\vec{a} \in A} \sigma(\vec{a}) Q_i(s, \vec{a})\}$

Soccer

We explored the version of the two-player zero sum Markov game soccer used in [1], our implementation follows the same definitions for S, A and R; our transition function T is defined to implement the collision and ball possession transfer dynamics that are described in [1]. The size of the state space was computed combinatorially by $|S| = |I| \times n \times (n-1)$ where I is the number of agents and thus the number of ball possession sub-states, and n is the grid size. We were able map Greenwald's description of the soccer random action execution rules by observing that each $s \in S$ will have at most 2 efferent edges (in the case of a collision state). Therefore, we may construct T by iterating over its first two dimensions $S \times A$ and computing a probability distribution over $s' \in S$ according to the rule set. A state transition matrix marginalized over the joint action space of both players will be included in the slides associated with this report. We validated our soccer model by letting the agents play random games and observing the outcome statistics.

Results

As in [1], four experiments were conducted. Significant assumptions regarding the learning rates, action selection and other parameters had to be made as they were not stated explicitly in [1]. In all experiments, 10^6 iterations of some Q-learning variant were run on the soccer game.

Although Greenwald does not explicitly indicate the learning rate decay schedule, we can infer from the algorithms' behavior in her figures that it is likely something of the form $\alpha = \dfrac{1}{\beta T + 1}$ where

T is the iteration and β is a hyperparamter. Since Greenwald indicates that $\alpha \to 0.001$ in these experiments, we tweaked β to make $\alpha \approx 0.001$ when $T = 10^6$. The temporal discount factor γ is also not designated for the soccer game, we assume $\gamma = .9$ is in the earlier grid experiments. Finally the exploration parameter ϵ which parametrizes the ϵ-greedy action selection behavior of the algorithm was estimated to follow a similar schedule to α; this is similarly consistent with the earlier gridworld experiments in [1]. Each game was initialized in the s_0 state, we inferred that an appropriate choice would be the state discussed at length in [1] with B in possession in the upper-left corner of the play field and A immediately to the east.

Q-Learning

In all experiments, we initialize one Q function for each agent in the case of standard Q-learning, $Q_A(s, a), s \in S, a \in A_A(s)$ and $Q_B(s, a), s \in S, a \in A_A(s)$. On each timestep, each player selects its action ϵ-greedily wrt. ϵ and its Q_i. These actions are composed into a joint action $\vec{a}$ and rewards r_a, r_b and s' are observed according to T and R. Both Q functions are then updated according to:

$$Q_i(s, a_i) \leftarrow Q_i(s, a_i) + \alpha[r_i + \gamma \max_{a'} Q_i(s', a') - Q_i(s, a_i)] \tag{6}$$

Figure 1. contains the resulting $|Q_i^t(s, \vec{a}) - Q_i^{t-1}(s, \vec{a})|$ for agent A taking action a_0 in state s_0. As in [1] we observe that Q-learning does not converge for this problem and instead produces random oscillations in its utility estimations, constrained in magnitude by the learning rate decay schedule. As observed by Greenwald, this is due to Q learning's inability to select a single adversarial equilibrium; instead the algorithm erratically attempts to minimize competing objects and fails to converge to a stable strategy in the interval considered.

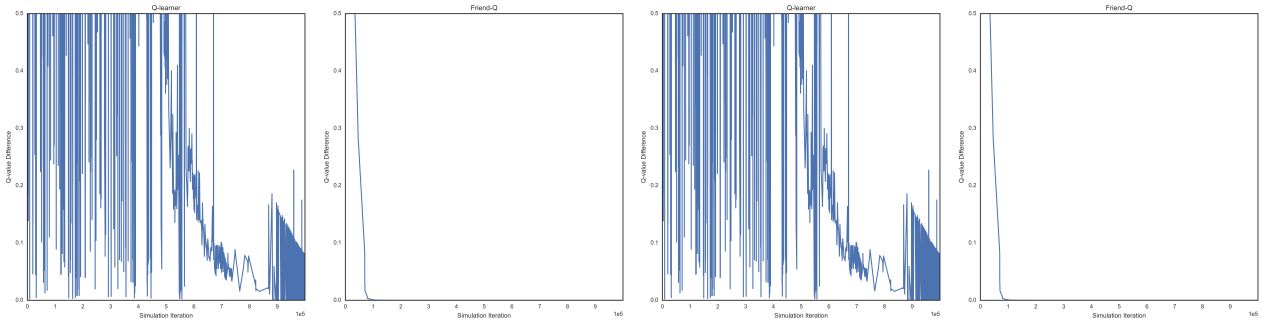

Figure 1: Difference in $Q_A(s_0, a_0)$ between iterations for the initial state s_0 for a particular action a_0 for agent A trained with Q-learning

Friend-Q

We implemented friend-Q similarly but with two distinct differences. Both players Q functions are $Q_i(s, \vec{a}$; Note that unlike Q-learning, the action set is now defined as the joint over both players. Secondly, although each player learns separate action-value functions, and receives distinct rewards, in the action selection step their value estimates are compared and the larger is selected.

$$\vec{a_t} = \underset{\vec{a} \in A(s), Q_i \in \vec{Q}}{\text{argmax}} \; Q_i(s, \vec{a}) \tag{7}$$

This formulation implies a search in the joint action space of both players for a collaborative equilibria. Since no rational collaborative equilibria exist, friend-Q converges to a deterministic policy for B which affords A a non-sensical advantage in this state. This somewhat spurious result is difficult to reason about and discuss theoretically.

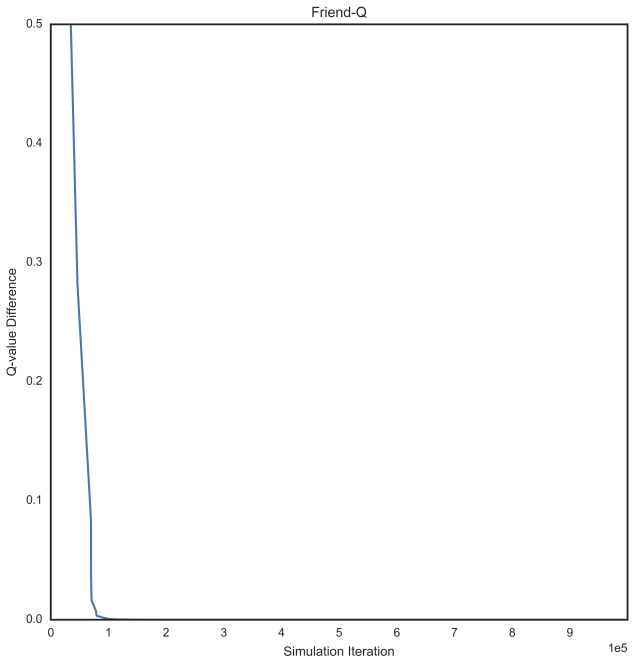

Figure 2: Difference in $Q_A(s_0, a_0)$ between iterations for the initial state s_0 for a particular action a_0 for agent A trained with friend-Q

Foe-Q and Correlated-Q

Unfortunately we are still in the process of debugging the linear program necessary to implement the two last algorithms in [1]. It is our hope that we will have time remaining to pursue this further, however we anticipate this will need to be the subject of follow-work.

In lieu of empirical analysis, we offer an extended theoretical discussion. The mimimax or foe-Q algorithm implements possibly mixed strategies by sampling from a probability distribution $\Sigma(s)$ over actions conditioned on the state. In this paradigm, we compute expected values in terms of the maximal expected reward given $\sigma \in \Sigma(s)$ and minimized over the possible action selections of the opponent. In the update step of this algorithm, we must solve the linear program which imposes the probability constraints (must ensure Σ is a proper probability distribution) and rationality constraints (maximize the minimum value obtainable wrt. the action space of the opponent).

In zero-sum games, minimax strategies will coincide with the adversarial Nash equilibria of the game. Using this insight, the Nash-Q definition of the value function discussed in [1] generalizes this notion; however, as we have already observed, there may be more than one adversarial equilibrium. Correlated-Q learning primarily answers the question of what criteria to use in order to select target equilibria. With an appropriate CE selection function, a probability distribution over the joint action space of both agents is computable via. similar linear programming constraints described for foe-Q.

Conclusion

One of the primary challenges in this work was interpreting the results under an unknown learning rate decay schedule. We observed that the measured difference in Q was strongly impacted by our choice of α schedule and the β parameter. Without knowing the precise method used in [1], it is difficult to compare our results and make strong claims as to their validity (or invalidity!). Secondly, it is apparent from the plots in figure 3 of [1] that some form of moving average or subsample was made over the values. Again, without knowing how this procedure was conducted, our results will look qualitatively different. A final difficulty in this implementation was the number of external resources

that needed to be referred to in order to understand the four algorithms compared; this is compounded by drastically different notation and algorithm formulations.

References

[1] Amy Greenwald and Keith Hall. Correlated-q learning. In *In AAAI Spring Symposium*, pages 242–249. AAAI Press, 2003.

[2] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *IN PROCEEDINGS OF THE ELEVENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, pages 157–163. Morgan Kaufmann, 1994.

[3] Michael L. Littman. Friend-or-foe q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 322–328, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.