1. Introduction

Reproducing results published in scientific journals is an important part of participating in the research community. This project attempts to reproduce the results seen in Figures 3a-d from Greenwald and Hall (2003).

2. Description of Experiments

The experiment that was reproduced was that of the soccer game in Figure 1. This formulation of the game comes from Greenwald and Hall (2003), which the authors took from Littman (1994). See these papers for a complete description of this game. In short, there is a grid world with two agents who have goals at either end. If the player with the ball enters the goal, the player that owns that goal scores +100 points while the other player scores -100 points. The ball changes ownership when one player attempts to move into the cell of another player that is not moving. This is a two-player zero sum game that has no deterministic equilibrium policies.

Several different Q learning techniques were used to find policies and to see whether the Q-values converged. These were utilitarian correlated Q, Friend Q, and Foe Q, along with traditional Q learning.

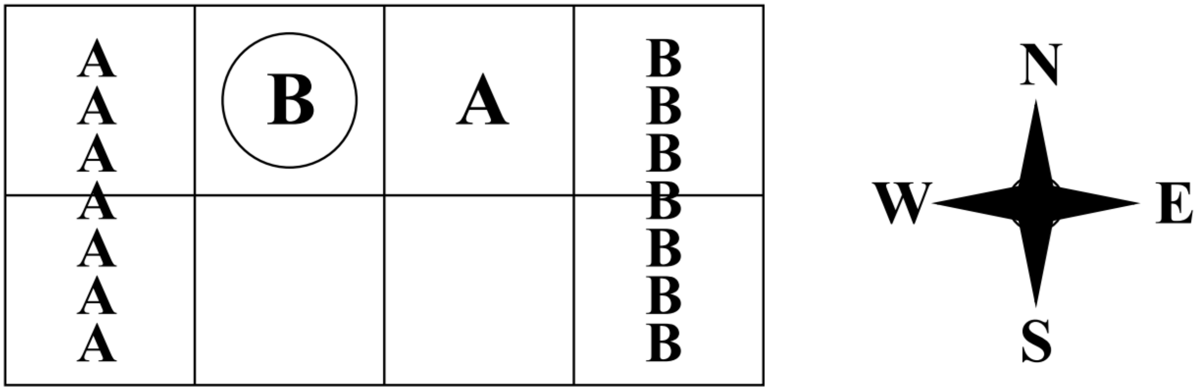

Figure 1: Soccer game (Greenwald and Hall)

3. Implementation Details

BURLAP (http://burlap.cs.brown.edu/) was used to replicate these results. While BURLAP had implemented many of the Q-learning techniques necessary for this experiment, constructing the soccer game was challenging. BURLAP had a GridGame class that could be used with stochastic multiplayer games, but to include the mechanics of this game required large portions of this class and its counterparts to be rewritten. As such, several new classes were created that took the behavior of GridGame and specialized it for this soccer grid game.

The correlated Q, Friend Q, and Foe Q algorithms supplied by BURLAP were used in these experiments. For traditional Q learning the naïve Q-learner was used.

4. Results and Discussions

As can be seen in Figure 2-Figure 7 below, correlated Q, Friend Q, and Foe Q all converge, confirming the results of Greenwald and Hall (2003). However, the graphs below do not quite match Greenwald and Hall's results. For all the figures below, the data was first "cleaned" by removing all the zero entries. For Figure 2, Figure 4, and Figure 6 a constant learning rate was used. This produced results more similar to

Greenwald and Hall for correlated Q; the Q values converge, but do not do so smoothly. The Foe Q values converge more like the Friend Q values from Greenwald and Hall while the Friend Q values in Figure 4 converge more like Foe Q in Greenwald and Hall. It is not know why this is the case.

The Greenwald and Hall (2003) paper does not explicitly indicate how the learning rate was used, but it seems to imply that it was decayed rather than held constant. As such, Figure 3, Figure 5, and Figure 7 show the results using an exponential decay learning rate. Here the plots agree with Greenwald and Hall in that they all converge, but none of them have the same pattern except for Friend Q.

Furthermore, the Q values for the tests conducted here converge much faster than in Greenwald and Hall. This is likely due to differences in discount factor, initial Q values, and the fraction of the time a random action was taken. Attempts were made to match the results as closely as possible.

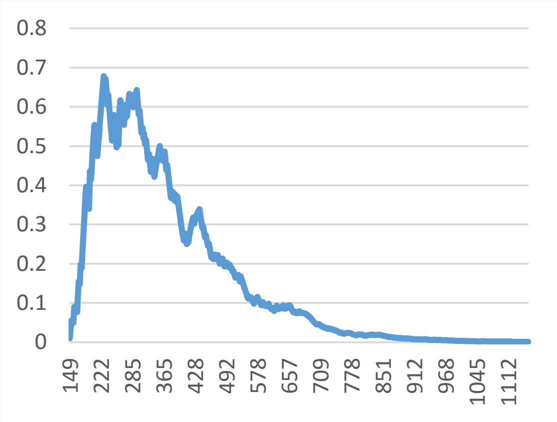

Figure 2: Correlated Q error – constant learning rate

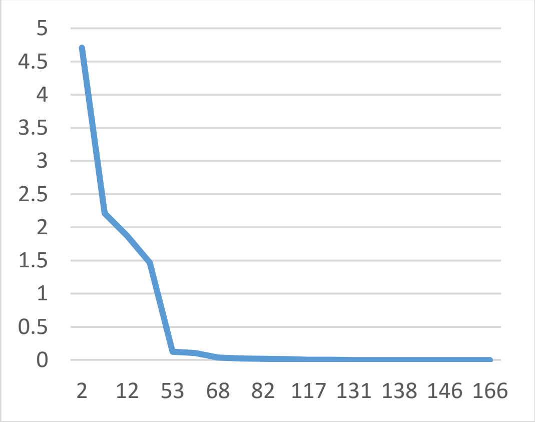

Figure 3: Correlated Q error – decayed learning rate

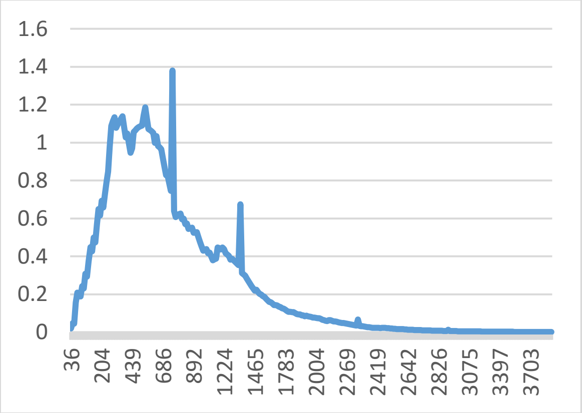

Figure 4: Friend Q error – constant learning rate

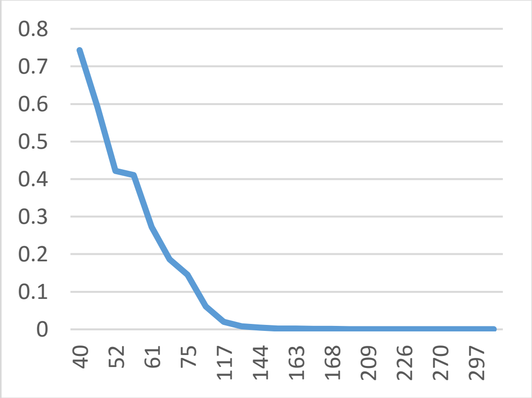

Figure 5: Friend Q error – decayed learning rate

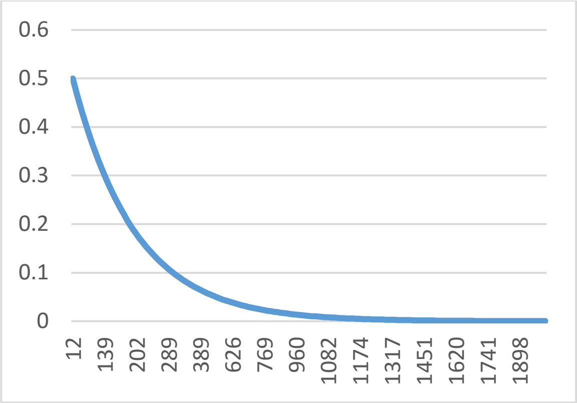

Figure 6: Foe Q error – constant learning rate

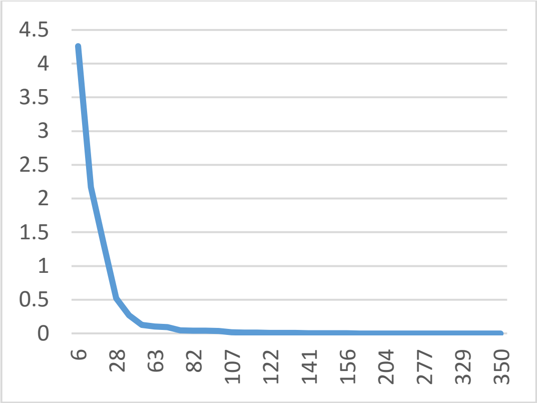

Figure 7: Foe Q error – decayed learning rate

For the naïve Q-learner, the Q values did not converge. With a constant learning rate, Figure 8 shows that the Q values are constantly oscillating with absolutely no decay. In Figure 9 we see the error tapering, but this is only because the learning rate decays exponentially. These results agree with Greenwald and Hall except in the rate of decay in Figure 9 – the error decays much faster than in Greenwald and Hall.

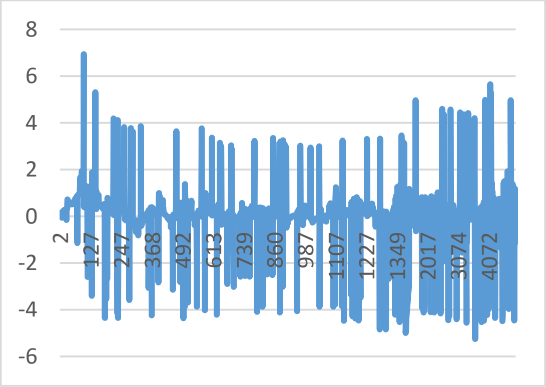

Figure 8: Naive Q - constant learning rate

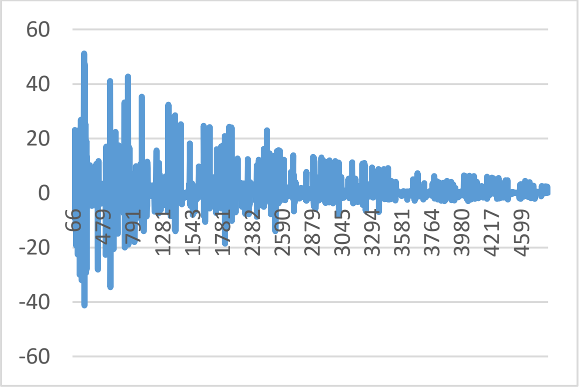

Figure 9: Naive Q - decayed learning rate

5. Conclusions and Future Work

The major results from Greenwald and Hall were confirmed. Correlated Q, Friend Q, and Foe Q all converge, albeit not quite in the same manner as the original paper. Meanwhile naïve Q does not converge because there are no deterministic optimal policies.

There are a few items that need future work. First, the results produced here do not match as closely to the published results as could be desired. More testing with variations in learning rates, discount factors, initial Q values, and the percentage of times that a random action is selected (epsilon). There was not adequate time to fully test these combinations and parse the data. Additionally, a way to visualize the game as it is being playing and to view Q values as they are being updated would be useful to aid in both understanding the strategies of the various learners and debugging.