Reinforcement Learning: Project 3

Liangliang Yang (lyang338)

I. Overview

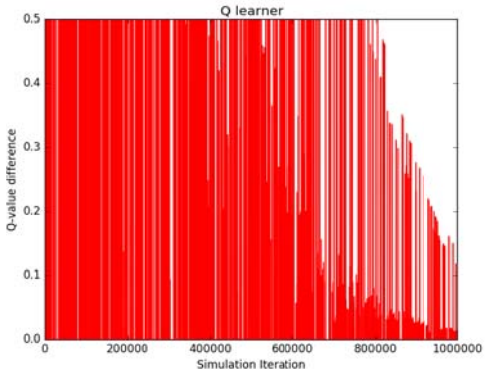
In this project we will try to replicate several multi-agent Q learning algorithms that learn equilibrium policies described in the Greenwald (2003) paper. We will focus on the "soccer game" described in the paper, which is a zero-sum two-player Markov game with no deterministic equilibrium policies. The four algorithms we will mainly discussed in this report is normal Q, Friend-Q, Foe-Q and Correlated-Q, and also there will be four figures shown below. The goal is to gain a better understanding of how these algorithms work in detail.

II. Q-learner

First we will talk about the normal Q-learning algorithm. We have already implemented this algorithm in the previous homework. Here we can assume there is only one agent, and he will only consider its own actions and rewards, hence every step this agent will just try to maximize its reward. The value function in this situation is

$$V_i(s) = \max_{a \in A_i(s)} Q_i(s,a)$$

The Q-value difference vs. number of iterations result is as below. As we can see, the Q-value difference will not converge, since the agent ignores the opponent's action.



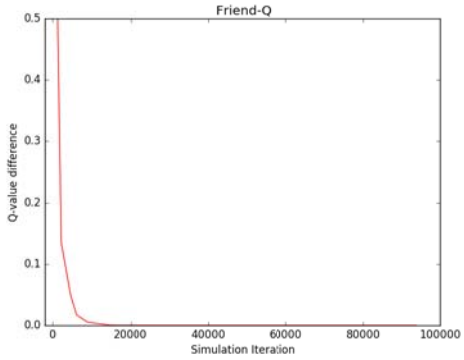
III. Friend-Q

Just like the name, in the situation of Friend-Q learning, the agent will assume its opponent is a friend. In particular, during the soccer game, the agent A will pass the ball to its opponent B, and B will score for him. Hence, the objective function and Q update rule for Friend-Q learning algorithm is as below. The joint actions are *A1(s) x A2(s).*

$$V_i(s) = \max_{\vec{a} \in A(s)} Q_i(s, \vec{a})$$

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

Similarly, we will implement the experiment and plot the Q-vale difference as below. As we can see, the Friend-Q converges very fast, since the learned policy is not reasonable, since it ignores that fact that each agent should try to defeat the other one.

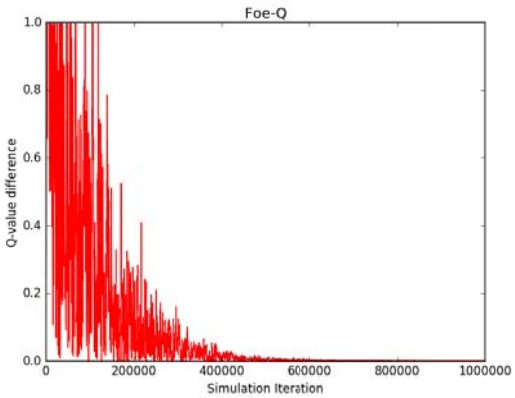


IV. Foe-Q

Unlike the previous two learning algorithms, the Foe-Q will consider there are two agents in the game, and they will try to defeat each other. Under the Foe-Q, the agent will try to maximize its own reward while minimize its opponent's reward. The objective function is:

$$V(s) = \max_{\pi \in \Pi(A_1)} \min_{a_2 \in A_2} \sum_{a_1 \in A_1} \pi(a_1) Q_1[s, a_1, a_2]$$

This minimax strategy can be converted to a linear program, and then we can solve it by using solvers.lp() function in CVXOPT library (we can also use scipy.optimize.linprog(), but it is slower). The experiment result is plotted as below. As we can see, the result is similar to the plot in the paper.

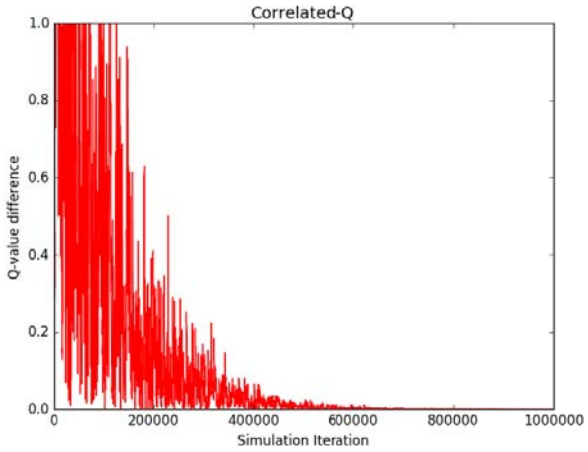


V. Correlated-Q

The last part we will discuss about result of correlated-Q. The CE-Q is used to solve the problem where multiple agents need to select among multiple equilibria. In the paper, the authors discussed four different variants of solution, and here we will focus on the utilitarian correlated-Q learning. The objective function for uCEQ is:

$$\sigma \in \arg\max_{\sigma \in CE} \sum_{i \in I} \sum_{\vec{a} \in A} \sigma(\vec{a}) Q_i(s, \vec{a})$$

Again we will use linear programming to solve the problem. Compared to the Feo-Q, there will be more constraint functions (there will be 25 actions combinations). The experimental result for Q difference is plotted as below. Again we can see that the Q-difference converges afer about 600000 iterations. Also during the experiment I see that the CEQ is much slower than Foe-Q, which I think is due to the higher complexity of linear programming.



VI. Summary

In this project, I have successfully replicated the four figures regarding the soccer game in the paper. All plots are similar to the figures in the original paper. By following the implementation and experiments, I have gained a better understanding about the difference between the four Q learning (normal-Q, Friend-Q, Foe-Q and Correlated-Q) algorithms. Overall, this is a very challenging and interesting projects.

Youtube presentation link: https://youtu.be/4HhKpojXews