Homework 3

Kaizhao Liang(kl2)

1. $\{\alpha = 0.4, \gamma = 0.95, \epsilon = 0.04\}$ average score: 13.2
   This is the default set of values.

   $\{\alpha = 0.4, \gamma = 0.95, \epsilon = 0.1\}$ average score: 9.2
   By changing the ϵ to 0.1, I allow the learner to explore the states space more. With determined number of tranning games, this set could inform me the ϵ's influence on the convergence of the learner.

   $\{\alpha = 0.4, \gamma = 0.5, \epsilon = 0.04\}$ average score: 5.8
   By changing the γ to 0.5, I reduce the affect of the future state to the current state. With determined number of tranning games, this set could inform me the γ's influence on the convergence of the learner.

   $\{\alpha = 0.01, \gamma = 0.95, \epsilon = 0.04\}$ average score: 4.2
   By changing the α to 0.01, I slower the process of learning. With determined number of tranning games, this set could inform me the α's influence on the convergence of the learner.

2. The best set I found is $\{\alpha = 0.4, \gamma = 0.95, \epsilon = 0.04\}$average score: 13.2

3. if the initial spot is randomized, the trainning is going to take longer or more games. However, once it converges, the final behaviour would be better than the one with the set initialized values, since it's forced to explore more states and learn more.

4. a. Because our pong environment is continuous. The position and velocity of the ball as well as the possible position of the paddle are not discrete. Thus both the state set and the action set are not finite, which by the definition of the MDP should be finite. There are a few major problems faced by a Q-learning agent. Firstly, when discretizing the state space to generate a finite set, some of the information about the state is lost, which otherwise could influence the result of the later decision making. Secondly, the MDP's exploration is based on probability, which means that in some of the cases, the optimal decision might not be explored.

   b. If the Pong environment was a real Markov Decision Process, the converging time or the number of training games required would be fewer than if it is not. In addition, the average performance on the future games will be higher and more stable, since there is less uncertainty both in training and the actual game.

5. a. My rating for this statement is 0. The ϵ decides how much the algorithm is going to explore or how random the decision is made in the training. Normally, the time spent should be mostly porportional to the range of the exploration. THe t and the ϵ do have systematic relationship.However, the distribution of the world influences the converging time, too.

   b. My rating for this statement is 5. There is a chance that all the decisions are optimal or their rewards are similar. So the converging time for the $\epsilon = 1$ and $\epsilon = 0.01$ learners could be similar.

   c. My rating for this statement is 5. This is natural and what would happen most of the time, when the rules of the world are not complicated and the greedy approaches are straightforwardly the best.

   d. My rating for this statement is 0. The converging time of the $\epsilon = 1$ learner should be the upperbound for all the running time for different ϵ values, since it's essentially a BFS. So the t_1 is not possible to be the smallest one.

   e. My rating for this statement is 5. Depending on the properties of the game, it's possible that the $t_1 \approx t_4 > t_2 \approx t_3$, especially when greedy approaches lead to local maximas and breath first is too blind. The converging time of the Markov decision process depends on the distribution of the world.