CS 440 - Homework 3

Matthew Liu - mliu56

1.

- $\alpha =, \gamma =, \varepsilon =$
- $\alpha =, \gamma =, \varepsilon =$
- $\alpha =, \gamma =, \varepsilon =$

2.

$$\alpha =, \gamma =, \varepsilon =$$

3.

Overall, it would increase the amount of exploration done as it can randomize positions that would otherwise never be explored. However the final behavior would also be significantly less table and have more erratic actions.

4.

- Our pong environment is not a true Markov decision process as a true process is a series of states with edges leading to other states. The sum of the probability of all outward edges from a state is 1.0, and those values determine the likelihood of that edge being taken (Hence a weighted random action). Our Pong environment takes 2 different approaches, if a random number is below the epsilon value, then take a completely random action/edge (all outward edges thus weighed equally), otherwise use the future Q-values and pick the best one (Greedy, means whichever edge leads to the best option has weight 1.0). Even so, a Q learner in a non-MDP environment would face major issues in balancing out how to choose a action to take. The entire design would be drastically different and I imagine significantly less efficient.
-

5.

- There is no systematic relationship $\rightarrow$ 4, similar to (c), (d), the randomization comes into play and different randomizations result in significantly different results.
- The differences are likely to be very small $\rightarrow$ 0, a combination of exploration & greediness is the optimal way and thus will converge the soonest
- $t1 > t2 > t3 > t4 \rightarrow 0$, it depends on the "luckiness", a totally random action can be more efficient than the best learning program, but at the same time an "unlucky" randomization can be significantly worse than the worst learning program.

- $t1 > t2 > t3 > t4 \rightarrow 0$, same reasons as above
- It is likely that another relation holds $\rightarrow 0$, same as (a), (c), (d)