CS 4700 Final Review

1. Policy iteration
   1. There are number of policies in an environment, where is the number of actions and is the number of states.

We know that the optimal policy converges here with and , so solve and , which gives us .

1. Q-Learning
   1. Then the update formula would be . Since , then the updated Q-value will **increase**.
   2. No, it will never be higher or lower, but equal to to cancel out with it because it converged already. This is only in the deterministic setting. In the stochastic setting, higher or lower than is still possible.