MDPs and reinforcement learning
Due Dec 3rd In class

1. Consider an undiscounted MDP having three states (1, 2, 3), with rewards (-1, -2, 0) respectively. State 3 is a terminal state. In state 1 and 2 there are two possible actions: a and b. The transition model is as follows:

- In state 1, action **a** moves the agent to state 2 with prob. 0.8, and makes the agent stay put with prob. 0.2
- In state 2, action **a** moves the agent to state 1 with prob. 0.8, and makes the agent stay put with prob. 0.2
- In either state 1 or 2, action b moves the agent to state 3 with probability 0.1 and makes the agent stay put with probability 0.9

Answer the following questions:
a) Apply policy iteration, showing each step in full, to determine the optimal policy and the values of states 1 and2. Assume that the initial policy has action b in both states.
b) What happens to policy iteration if the initial policy has action a in both states? Does discounting help? Does the optimal policy depend on the discount factor?

2. Consider playing Tic-Tac-Toe against an opponent who plays randomly. In particular, assume the opponent chooses with uniform probability any open space, unless there is a forced move(in which it makes the obvious correct move).
a) Formulate the problem of learning an optimal Tic-Tac-Toe strategy in this case as a Q-learning task. What are the states, transitions, and rewards in this nondeterministic MDP? (Note that the opponent is considered to be part of the world)
b) Suppose we represent the Q function as a table, how many entries would it have and why?
c) Will you program succeed if the opponent plays optimally rather than randomly? Why?