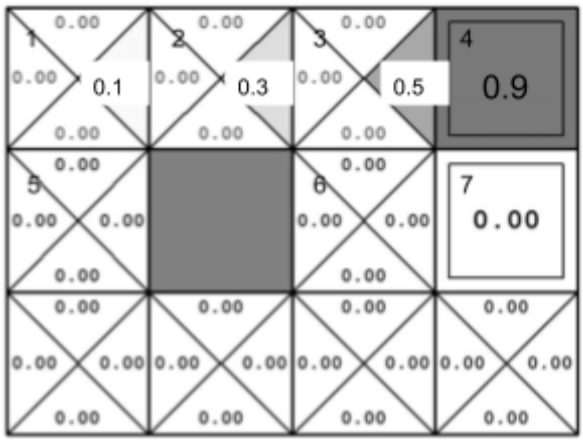
Homework 3

1. (20 pts) When the learning rate is 0.1, and the discount factor is 0.9. Consider the following trajectories.

   (s2, north, s3, r = -0.1); (s3, east, s4, r = -0.1); (s4, north, s4, r = 1.0);



   Use the Q-learning algorithm to help the agent update its Q function: Q(s,a). Please list the Q values that have been changed after each of the three actions. Further, the updated Q value(s) should be used in computing the update of follow-up actions.

2. (20 pts) Answer yes or no to each of the following questions.

   a. RL algorithms assume the availability of an MDP for world modeling

   b. RL algorithms assume the existence of an MDP for world modeling

   c. RL algorithms assume the current world state is fully observable

   d. RL algorithms assume that the reward function is available

   e. RL algorithms assume that the immediate reward is available

3. (10 pts) Very briefly (in a couple of sentences) answer the following two questions

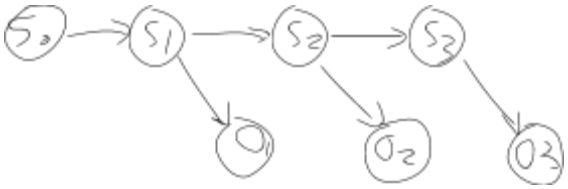
   a. Given an MDP problem and an RL algorithm, what shall we do so that we are able to apply the RL algorithm to the MDP problem?

   b. Given an RL problem and an MDP algorithm, what shall we do so that we are able to apply the MDP algorithm to the RL problem?

4. (30 pts) The following three subfigures show the HMM model, its transition probabilities, and emission probabilities.

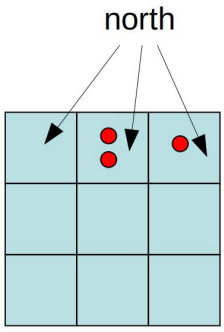
S	O	$P(O\|S)$
A	0	0.8
A	1	0.2
B	0	0.1
B	1	0.9

   a. Compute the stationary distribution of P(S) using the method of solving linear equations introduced in the lecture of "Markov Models."



   b. Compute probability $P(O_1=0, O_2=0, O_3=1)$, where S_0 follows the stationary distribution computed from the last question.

   c. S_0 still follows the stationary distribution. Show the belief change after receiving each observation of O_1, O_2, and O_3, i.e., to compute $P(S_1|O_1=0)$, $P(S_2|O_1=0, O_2=0)$, and $P(S_3|O_1=0, O_2=0, O_3=1)$



5. (20 pts) The right shows the current state of the world, where we use a Particle Filter (PF) to estimate the robot's position. We only maintain three particles, and their positions are: (2,3) (2,3) (3,3)

   We know that the robot is deterministically moving clockwise (i.e., perfect motion control). The robot can correctly observe whether it is in the north area in 0.8 probability.

   a. Describe briefly one complete iteration of the PF's elapse-weight-resample process, i.e., to describe what is changed to the particles after each step.

   b. If we are resampling over and over again, without any motion update. What is going to happen in the limit?