1. Suppose our Q-learning agent, with fixed *alpha*, and discount gamma, was in state 34 did action 7, received reward 3 and ended up in state 65. What value(s) get updated? Give an expression for the new value. (You need to be as specific as possible)

2. In temporal difference learning (e.q. Q-learning), to get the average of a sequence of k values, we let $alpha_k = 1/k$. Explain why it may be advantageous to keep $alpha_k$ fixed in the context of reinforcement learning.

3. Explain what happens in reinforcement learning if the agent always chooses the action that maximizes the Q-value. Suggest two ways that can force the agent to explore.

4. In MDPs and reinforcement learning explain why we often use discounting of future rewards.

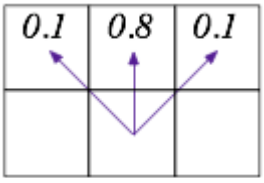
5. What is the main difference between standard value iteration and asynchronous value iteration? Why does asynchronous value iteration often work better than standard value iteration?

6. In feature-based reinforcement learning, what new data point does the experience $<s,a,r,s',a'>$ give for the linear regression? (I.e., what is the new data point for what value?)

7. In learning under uncertainty, when the the EM algorithm used? What is the E-step? What is the M-step?

8. Consider a 5×5 grid game similar to the simple game used in assignments. The agent can be at one of the 25 locations, and there can be a treasure at one of the corners or no treasure.

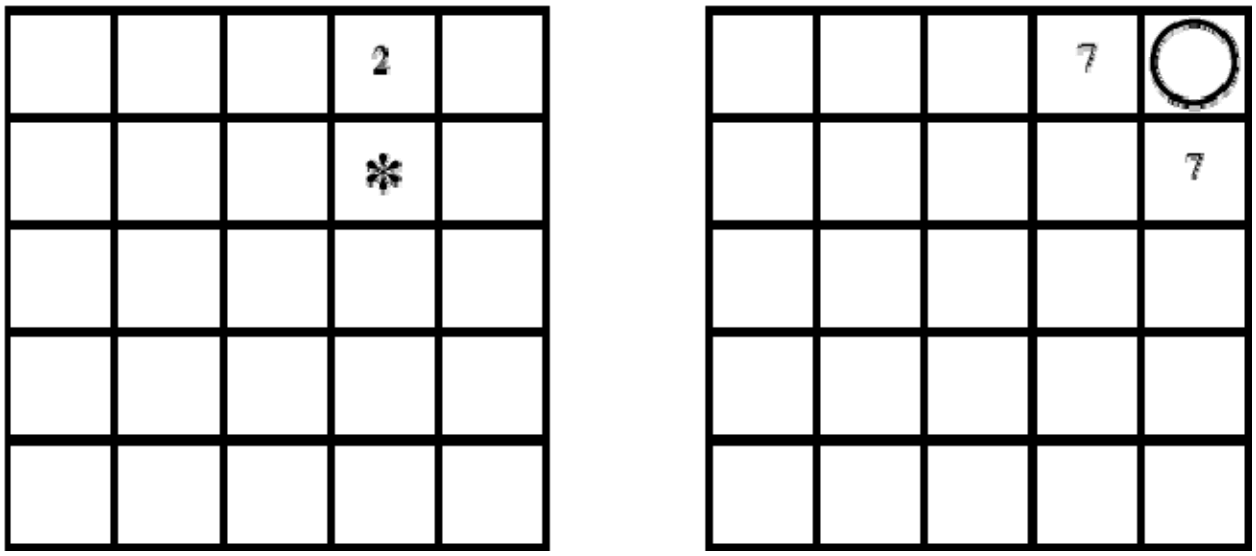
In this game the "up" action has the dynamics given by:



That is the agent goes up with probability 0.8 and goes up-left with probability 0.1 and up-right with probability 0.1.

If there is no treasure, a treasure can appear with probability 0.2. When it appears, it appears randomly at one of the corners, and each corner has an equal probability of treasure appearing. The treasure stays where it is until the agent lands on the square where the treasure is, and the agent gets an immediate reward of *+10*, and the treasure disappears in the next state transition. The agent and the treasure move simultaneously so that if the agent arrives at a square at the same time the treasure appears at the same time, it gets the reward.

Suppose that we have the following utility for each state:



where the left grid shows the utilities for the states where there is no treasure and the right grid shows the values of the states when there is a treasure at the top-right. Assume that states with no number have utility that equals 0.

What is the expected value Q[s*, a_2] of performing action a_2 , which is up, in state s* (marked by a * in the left grid) and then following the optimal policy? You need to show all working, but don't need to do any arithmetic (i.e., leave it as an expression). Explain each terms in your expression.