Assignment 6 Report for Part A

Wayne Wang

1. Using the menu commands, set up the MDP for TOH with 3 disks, no noise, one goal, and living reward=0. The agent will use discount factor 1. From the Value Iteration menu select "Show state values (V) from VI", and then select "Reset state values (V) and Q values for VI to 0".
   Use the menu command "1 step of VI" as many times as needed to answer these questions:

   1a. 4 iterations.

   1b. 8 iterations.

   1c. This is not a good policy, after each state get green, the value for each value is 100. This is due to the discount factor of 1 and living reward of 0. Since there is no penalty for taking any actions, each state would have a value of 100 and any action at any state would be considered as optimized action.

2. Repeat the above setup except for 20% noise.
   2a. How many iterations are required for the start state to receive a nonzero value.

   2a. 8 iterations.

   2c. This is a good policy. With the 20% noise, each state will be generated with different value due to the probability of each action in each state. There are no illegal suggested actions and the action for each state is optimized.

   2d. 56 iterations.

2e. The policy doesn't change, because after convergence, there is already an optimized policy. More iterations will make very little change to the state value but these changes are too small to make any changes to the policy.

3. Repeat the above setup, including 20% noise but with 2 goals and discount = 0.5.

   3a.

   - **This policy indicates that when the discount is 0.5, it would be more optimized to go to the goal state with reward of 10 rather than the goal state with reward of 100.**
   - **Start state value is 0.82**

   3b.

   - **This policy indicates that when the discount is 0.9, it would be more optimized to go the the goal state with the reward of 100 rather than the goal state with reward of 10.**
   - **Start state value is 36.9**

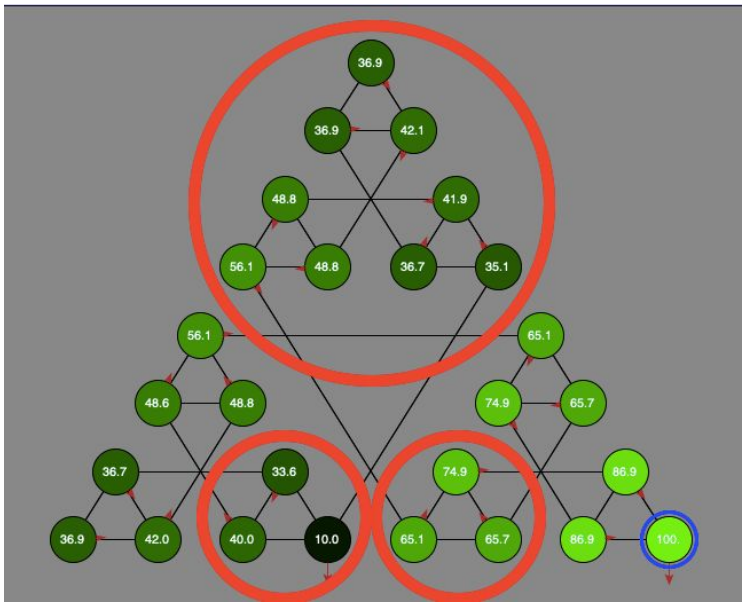
4. Now try simulating the agent following the computed policy. Using the "VI Agent" menu, select "Reset state to s0". Then select "Perform 10 actions". The software should show the motion of the agent taking the actions shown in the policy. Since the current setup has 20% noise, you may see the agent deviate from the implied plan. Run this simulation 10 times, observing the agent closely.

   4a. 2 simulations.

   4b. 8 simulations

   4c. Trial #3 : 2 steps away; Trial #7 : 1 step away

4d. Yes, the 9 states in the upper triangle seemed never to be visited. The 3 states "triangle" on the right bottom of the left 9 states "triangle", The 3 states "triangle" on the left bottom of the right 9 states "triangle" seemed never to be visited as well.



5. Overall reflections.

5a. No, to have a good policy does not require the values of each states have converged. However, it does require enough number of iterations so the change between kVPlus1[s] and kV[s] is not big enough to cause a change in the policy.

5b. It is very important for all the states to be visited a lot. As states being visited, the value for each state will be changed, especially at the first several times of visits. When a state's value is changed, it is likely other states around would be affected by the changed state value. Therefore, enough time of revisit in each state will help to determine the optimized actions from states to states → to a good policy.