HW6 Yutong Liu

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:
    a. How many iterations of VI are required to turn 1/3 of the states green? (i.e., get their expected utility values to 100).
    4 iterations
    b. How many iterations of VI are required to get all the states, including the start state, to 100?
    8 iterations
    c. From the Value Iteration menu, select "Show Policy from VI". (The policy at each state is indicated by the outgoing red arrowhead. If the suggested action is illegal, there could still be a legal state transition due to noise, but the action could also result in no change of state.) Describe this policy. Is it a good policy? Explain.
    For this case, it is not a good policy since the most of policy are illegal move and does not show the path to the final state. The state value would not change due to the value of discount number is 1.

2. Repeat the above setup except for 20% noise.
    a. How many iterations are required for the start state to receive a nonzero value.
    8 iterations
    b. At this point, view the policy from VI as before. Is it a good policy? Explain.
    It is a good policy since it shows the path to the final state.
    c. Run additional VI steps to find out how many iterations are required for VI to converge. How many is it?
    25 iterations
    d. After convergence, examine the computed best policy once again. Has it changed? If so, how? If not, why not? Explain.
    It does not change since the optimal path be founded in 8 iterations.

3. Repeat the above setup, including 20% noise but with 2 goals and discount = 0.5
    a. Run Value Iteration until convergence. What does the policy indicate? What value does the start state have? (start state value should be 0.82)
    The policy indicates the final state is small reward.
    Start value is 0.82.
    b. Reset the values to 0, change the discount to 0.9 and rerun Value Iteration until convergence. What does the policy indicate now? What value does the start state have? (start state value should be 36.9)
    The policy indicates the final state is the large reward.
    Start value 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.
    a. In how many of these simulation runs did the agent ever go off the plan?
    2
    b. In how many of these simulation runs did the agent arrive in the goal state (at the end of the golden path)?
    8
    c. For each run in which the agent did not make it to the goal in 10 steps, how many steps away from the goal was it?
    2 steps away
    d. Are there parts of the state space that seemed never to be visited by the agent? If so, where (roughly)?
    The agent does not visit the state on top triangle.
5. Overall reflections.
    a. Since it is having a good policy that is most important to the agent, is it essential that the values of the states have converged?
    It is not necessary to converge since the optimal pass would be generated before it converge.

    b. If the agent were to have to learn the values of states by exploring the space, rather than computing with the Value Iteration algorithm, and if getting accurate values requires re-visiting states a lot, how important would it be that all states be visited a lot?
    It is not necessary to visit all state since the optimal path may generated(or converge) before visit all state.