Assignment 5 Report for Part A

Nikola Dancejic

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".
    a. 4 iterations
    b. 8 iterations
    c. The policy seems to mainly point to the right which in many cases is an illegal action. This does not seem like a good policy because it seems to know the general direction that it needs to go but not necessarily the right path. This is likely because the rewards are not discounted so there is no motivation to get to the solution in the smallest number of steps.
2. Repeat the above setup except for 20% noise.
    a. 8 iterations
    b. Yes, this is a much better policy, now it shows the optimal path from the start but also works from any point.
    c. It takes 56 iterations to converge
    d. The policy has not changed because it found a good policy and more iterations have just confirmed it’s effective.
3. Repeat the above setup, including 20% noise but with 2 goals and discount = 0.5.
    a. The policy indicates that it is more efficient to go for the smaller reward from the start state. The start state is 0.82
    b. The policy indicates that it is always more efficient to go for the higher reward than the lower reward. The start state 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.
    a. 8
    b. 7
    c. The first time it was 1 state away. The second time it was 4 states away the third time it was 1 state away
    d. The Agent rarely goes to the top of the triangle.
5. Overall reflections.
    a. Not necessarily, as we saw in question 2, the correct policy was found after 8 iterations and just reinforced.
    b. The states would have to be visited enough to extract a decent policy, The agent might not even have to visit every state as long as it finds a path that works most of the time.