Assignment 6 Report for Part A

Fu Tianyuan (1974487)

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. 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 are needed for turning 1/3 of the states green.

1b. How many iterations of VI are required to get all the states, including the start state, to 100?

> 8 iterations are needed for turning all of the states green.

1c. 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.

> Suggested states are mostly pointing right and right downwards, roughly to the position of the target state in the state space. It is not a good policy since a large portion of the suggested states are illegal, i.e. pointing to a non-existing state. In this case, the policy cannot properly guide the agent to triverse rationally and the agent may end up elsewhere rather than the final state.

2. Repeat the above setup except for 20% noise.

2a. How many iterations are required for the start state to receive a nonzero value.

> 8 iterations are needed for the start state to receive a nonzero value.

2b. At this point, view the policy from VI as before. Is it a good policy? Explain.

> It is relatively a good policy since the suggested state for each state is one with a large expectimax. If referring the policy witht the optimization solution, we can find the direction of policy is matching the solution. Overall, it is a good policy.

2c. Run additional VI steps to find out how many iterations are required for VI to converge. How many is it?

> There will be 48 more iterations (56 in total) needed to converge.

2d. After convergence, examine the computed best policy once again. Has it changed? If so, how? If not, why not? Explain.

> The best policy has not been changed. Before the convergence of the state values, all of the states have been updated with certain values. The values will not converge unless the VI steps furthers up, but the policy is determined by the relative values among the states, and relative values are meaningful by all the

> states have been updated with values. When the VI furthers up, values of all states are converging, but the relative greatness of values keeps the same. That is why the policy did not change over the iteration.

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

3a. 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 that the suggested solution favors the closer goal with the value 10 since the discount value is 0.5, making the discounted 100 reward of the farther goal much and smaller than the discounted value of 10. The start state has the value 0.82.

3b. 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 that the suggested solution favors the farther goal with the value 100 since the discounting value is close to 1. Discounted reward of 100 is greated than the discounted value of 10. The start state has the 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.

4a. In how many of these simulation runs did the agent ever go off the plan?

> In 2 of the 10 simulations, the agent went off the plan.

4b. In how many of these simulation runs did the agent arrive in the goal state (at the end of the golden path)?

> In 8 of the 10 simulations, the agent end up in the goal state.

4c. 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?

> There are 2 runs in total that did not make it to the goal in 10 steps.
>
> Run 1: The final state is [1][2][3]. 2 more steps are needed to achieve the final state.
>
> Run 2: The final state is [][2, 1][3]. 3 more steps are needed to achieve the final state.

4d. Are there parts of the state space that seemed never to be visited by the agent? If so, where (roughly)?

> The upper part in the state diagram seemed unlikey to be visited during the trials. Probability theory can be helpful in explaining this. To achieve a state that located in the upper part, the VI agent has to diverge many times from the policy. Since the divergence in each state is independent from each other, the probability of n time diverging P(diverge for n times) = 0.2^n. Therefore, the agent is unlikely to diverge many times to reach the upper states. With limited number of divergence, the agent can still go back to the suggested path and proceed accordingly.

5. Overall reflections.

5a. 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 may not be essential that state values have to be converged. As we can see from question 2, the policy while states have not been converged may remain unchanged until they are converged. So the optimized policy can be attained before the convergence. But it is still meaningful to get all states converged as long as the computational power is sufficient, as more states are converged, more likely the current policy would be the optimized poicy.

5b. 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 would be important for all the states to be visited since the iterative visiting will leave greater uncertainty to the state values than the VI approach while values are necessary for generating the optimal policy.