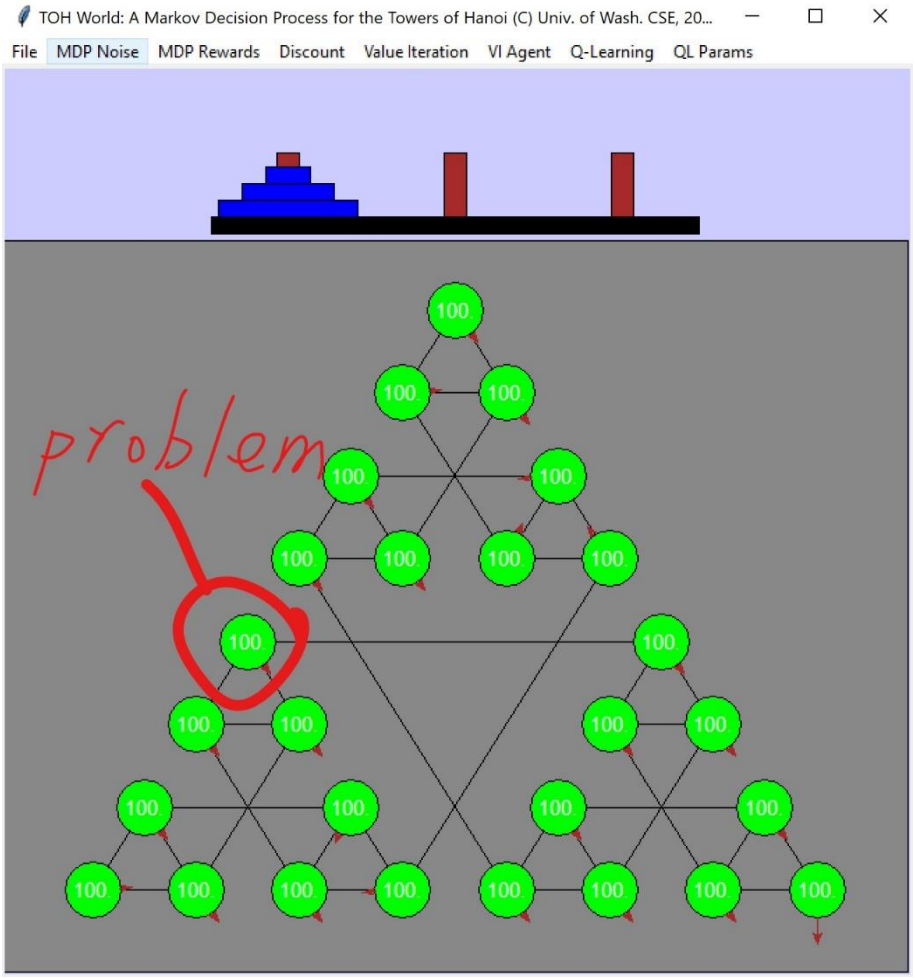
1、

1a) 4 iterations

1b) 8 iterations

1c) It's not a good policy.

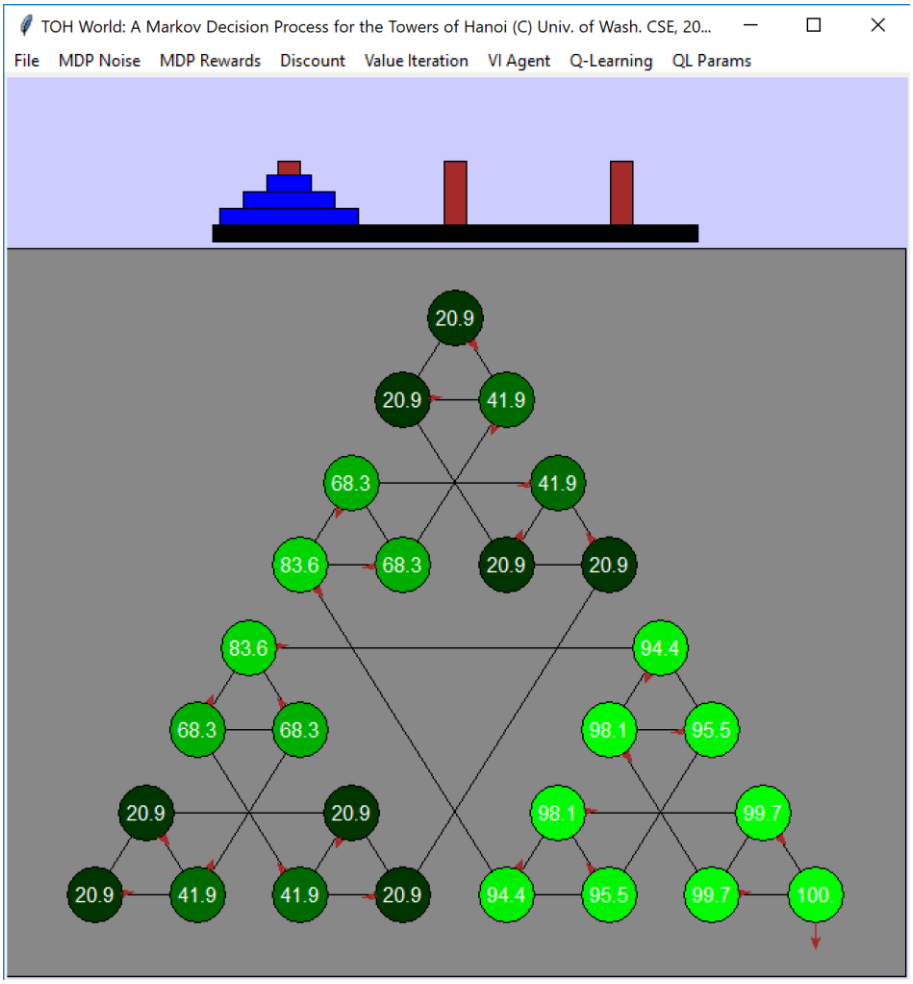


If it's a good policy, at the state I circled, the policy should be going right instead of going down right. Because at each state, it makes no difference for any choice due to the discounting factor is 1 and the deterministic property.

2、

2a) 8 iterations

2c) It is a good policy.



Because of the existence of noise, the V-values of states that are far away from the goal state will be smaller, thus making the states different. As long as the states' values are different, a policy is always good for going from a low-value state to a max high-value state.

2d) 56 iterations

2e) the policy is the same. Since the extra iterations are simply updating V values with respect to the maximum Q value, the policy is not changing.

3、

3a) The policy indicates that with a discount factor of 0.5, at some states, the best move is to go to the second goal with a reward of 10; 0.82

3b) The policy indicates that with a discount factor of 0.9, no states are going to the second goal with a reward of 10; 36.9

4、

4a) 7 out of 10

4b) 4 out of 10

4c) 1,1,2,3,2,1

4d) Yes. The top part of the state space.

5、

5a) It's not essential. Because an optimal could be generated without the v values converged.

5b) It's not necessary for the agent to visit all the states a lot. The agent only needs to visit the states it believes to be good states and re-visit these to confirm that these are good states. The states with low V-values don't needed to be visit a lot for the same reason.