Last name: Li First name: Minghan SID#: 156/234

Collaborators: _____

CMPUT 366/609 Assignment 5: Planning and learning

Due: Thursday Nov 2nd by gradescope

There are a total of 84 points available on this assignment, as well as 10 bonus points

The first 4 questions are exercises are from the Sutton and Barto textbook, 2nd edition:

Question 1: Exercise 6.2 [6 points] (*TD vs MC; driving home example*)

Question 2: Exercise 6.3 [6 points] (*first episode in chain problem*)

Question 3. Exercise 6.4 [6 points] (*impact of step-size parameter*)

Question 4. Exercise 8.3 [6 points] (*Dyna-Q+ vs Dyna-Q*)

Question 5. **Programming Exercise [60 points]**.

Part 1. [40 points] Implement Dyna-Q on the grid world described in Example 8.1 of the SB textbook. You can modify your windy grid world environment program from the previous assignment. Your task is to recreate Figure 8.3 (SB). Your learning curves should be averaged over 10 runs, with the random seed controlled appropriately such that the number of steps per episode during the first episode is the same for all three parameterizations of Dyna-Q. Note your results are not guaranteed to be exactly the same.

Part 2. [20 points]. Experiment with one of the key parameters of your Dyna-Q agent. Perform a systematic parameter sweep of the alpha parameter. You will test 6 different alpha values in:
{0.03125, 0.0625, 0.125, 0.25, 0.5, 1.0},
recording the performance of your agent for each setting of alpha.
Then you will plot the performance of your agent for each value of alpha in the set. Specifically plot alpha-value on the x-axis, and the average number of steps per episode over the first 50 episodes (averaged over 10 runs) on the y-axis. That is, each point on the graph reports the performance of Dyna-Q, for one specific setting of alpha, averaged over 10 independent runs.

In order to get good performance you may have to experiment with the exploration rate parameter (epsilon). Start with the values used in the book. You are not required for systematically sweep epsilon, but you are free to do so. Use number of planning steps equal to 5.

This exercise requires implementing one agent program (Dyna-Q), an environment program (implementing the gridworld), and two experiment programs—the first for generating your version of Figure 8.3, and the second to run your parameter sweep of alpha. Please submit all code, including scripts and plotting code, and all plots.

Q1.

Scenario: Suppose the man in the driving home example has been driving home from work for many years. One day his wife asks him to buy a cake on the way home cause it's his son's birthday. The man doesn't want to be late so he has to estimate how much time would be left after buying the cake. TD will immediately give the man a rough estimate no matter where he chooses to buy the cake on the same route; but MC can only gives him the answer after he gets home, which might cause him to be late. And TD also seems closer to how we actually make decision in this scenario

Q2.

(1) It tells us that the agent ended up at the left terminal in the first episode.

(2) Because the TD learning rule for every ~~non-terminal~~ state with non-terminal next-state is $V(s) := V(s) + \alpha(r + \gamma V(s') - V(s))$ In the first episode every state value is initialized to 0.5, and r is 0, so there would be zero update for these value estimates; as for the state A, because the agent ends up at its left terminal the TD learning rule will be: $V(s) := V(s) + \alpha(r - V(s))$. In this case r is 0, so the update would be $-\alpha V(A)$ which cause this decrease at value of state A

(3) $-\alpha V(A) = -0.1 \times 0.5 = -0.05$

Q3. (1) No. Because the learning rates chosen in Figure 6.2 is enough and reasonable, which would lead to appropriate empirical question. Of course, a wide range of α may show more results, e.g. TD with $\alpha = 0.09$ is slower than MC with $\alpha = 0.03$. But it would be meaningless to compare them since they don't reflect the actual "gap" between these two algorithms. The parameters chose in the graph are representative enough to show TD is better MC, so even with a wide range of α won't change this conclusion

(2) No. Because both the learning rules of TD and MC show certain consistency in α. We can already see the trend of the results from the chosen α, which suggests a vague bound of the error with varying α, so there would not be any signific changes with any other α value; Plus, "better" is a vague word, it depends on how you define it, e.g. asymtotic error, convergence rate and so on.

Q4. Because what's happening in Dyna Q^+ is the agent actually learns a "suboptimal" policy: it will visit the states that haven't been visited for a long time, which is caused by the extra "exploration bonus", so the increment of the accumulative reward will be smaller over time since once a while the agent will take some suboptimal routes.