Last name:_____ First name:_____ SID#:_____

Collaborators:_____

CMPUT 366/609 Assignment 5: Planning and learning

Due: Tuesday Oct 31st 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 PS 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.

Question 6. **Programming exercise**. **[10 bonus points].** Implement Dyna-Q with Prioritized sweeping described on page 140. Compare your implementation against your Dyna-Q agent from Question 5. Plot the steps per episode vs episodes for 50 episodes (like Figure 8.3) for Dyna-Q and Dyna-Q with Prioritized Sweeping. Setting theta well, can be tricky. Feel free to do some internet research to figure out good values of theta to test. Use 5 planning steps for both methods. Submit all code and your plot.