Last name:_______________________ First name:______________ SID#:____________________

Collaborators:____

CMPUT 366/609 Assignment 4: Monte Carlo and Temporal-Difference methods

Due: Thursday Oct 19 by gradescope

There are a total of **92 points** available on this assignment. There are **15 bonus points** available from two bonus questions.

Question 1. [68 points total] This question is comprised of exercises from the SB textbook. It contains **four** parts.

(a) Exercise 5.1 **[6 points]** (*value function for blackjack*)

(b) Exercise 6.6 **[6 points]** (*how to compute v_π for the chain*)

(c) Exercise 6.9' **[50 points]**. **Windy Gridworld with King's Moves.** Re-solve the windy gridworld task assuming eight possible actions, including the diagonal moves, rather than the usual four. (1) How much better can you do with the extra actions? (2) Can you do even better by including a ninth action that causes no movement at all other than that caused by the wind? Be sure to answer the two questions posed above and submit evidence for your answers in the form of additional plots. **In addition describe the parameter settings used in your experiment (alpha & epsilon).**

You are required to use RL-Glue for this exercise. Use the rl_glue.py and utils.py code from assignment #3. You can either take gambler_exp.py, mc_agent.py, and gambler_env.py and modify them to implement the windy gridworld, the Sarsa agent, and the appropriate experiment respectively. Or you can write your own environment, agent and experiment programs from scratch. Use whatever plotting software that is convenient for you.

Please submit your **agent** (one-step Sarsa)**, environment** (windy-gridworld with king's moves)**,** and **experiment program** and any additional scripts and graphing code. You will submit at least two plots. The first plot shows the performance of your Sarsa agent with eight actions. This will be a a learning curve like figure 6.4 in the book: Episodes vs time steps. The second plot will be the performance of your Sarsa agent with nine actions; again a learning curve like Figure 6.4

(d) Exercise 6.11 **[6 points]**. (*off-policy Q-learning*)

Question 2. [24 points] (episodic example of TD and MC)

Suppose you observe the following 9 episodes generated by an unknown Markov reward process, where A and B are states and the numbers are rewards:

A,0,B,6
A,3
A,2,B,0,A,3
B,0,A,2,B,2
B,0,A,3
B,2
B,6
B,2
B,6

1. (8 pts) Give the values for states A and B that would be obtained by the batch first-visit Monte-Carlo method using this data set (assuming no discounting). You may express your answer using fractions. Briefly explain how you arrived at your answer.

2. (8 pts) If you were to form a maximum-likelihood model of a Markov reward process on the basis of these episodes (and these episodes alone), what would it be? (sketch its state-transition diagram with transition probabilities and expected rewards.)

3. (8 pts) Give the values for states A and B that would be obtained by the batch TD method. Briefly explain how you arrived at your answer. You may express your answer using fractions.

Bonus Questions.

Question 3. [5 bonus points]. Exercise 5.6 from SB textbook 2nd Edition. (*modified MC policy evaluation algorithm*). Please justify your update rule in a similar (but not exactly the same way) as equation 2.3 in SB textbook.

Question 4. Programming exercise. [10 bonus points]. Resolve the windy grid world with king's moves described in Question 2 using n-step Sarsa. What values of n work best ?(test several values of n, and submit plots for each) Can you get your implementation to outperform one-step Sarsa? (plot the performance of one-step Sarsa vs n-step Sarsa) What is involved in making a fair comparison? (explain what is involved in making fair empirical comparisons in RL) Please submit all code and plots.