Last name:_____LI_____ First name:_____Xianhang_____ SID#:_____1465904_____

Collaborators:_____

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

Due: Wednesday Oct 31 by gradescope

There are a total of **90 points** available on this assignment. There are **5 bonus points** available from one bonus questions.

Question 1. [66 points total] This question is comprised of exercises from the SB textbook. It contains **four** parts.

(a) Exercise 5.2 **[4 points]** (*first-visit v every-visit 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 from assignment #3. Implement the windy gridworld as an environment program, the Sarsa agent (an agent program), and the experiment program. 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 embedded in Example 6.5 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 in Example 6.5.

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

Exercise 5.2.

To find the state-value function for this policy by a Monte Carlo approach, one simulates many blackjack games using the policy and average the returns following each state. Note that in this task the same state never recurs within one.

Exercise 6.6

1. We could use DP. DP come up with a model(environment), so it could review one ahead value and one before value.
2. We could also use MC. MC would generate a lot of potential moves for each state and then using fraction to end it.

DP is the most suitable method for this situation. DP can converge small problems very efficient like in this problem. However, MC needs many episodes to compute, and it can only converge approximately .

Exercise 6.11

Q-learning is about greedy choice policy. It is independently being followed by agent in each policy, and it becomes the optimal policy at the end. It learned about one policy by following another policy in the same time, so it is an off-policy method.

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

Suppose you observe the following 12 episodes generated by an unknown Markov reward process, where A and B are states and the numbers are rewards:

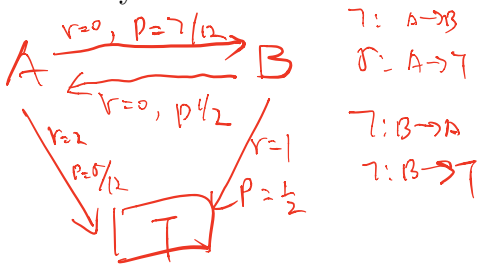
A,0,B,1	B,0,A,2	B,1
A,2	B,0,A,0,B,1	B,1
A,0,B,0,A,2	B,0,A,2	B,1
A,0,B,0,A,0,B,1	A,0,B,1	B,0,A,0,B,0,A,2

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. Explain how you arrived at your answer.

$A = [(0+1) + 2 + (0+0+2) + (0+0+0+1) + 2 + (0+1) + (2) + (0+1) + (0+0+2)] \div 9 = \frac{14}{9}$

$B = [1 + (0+2) + (0+0+1) + (0+2) + (0+0+1) + (0+2) + 1 + 1 + 1 + (0+0+0+2)] \div 11 = \frac{15}{11}$

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)? Explain how you arrived at your answer.



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

$$\begin{cases} V_A = \frac{5}{12} \times (2 + V_t) + \frac{7}{12}(0 + V_B) \\ V_B = \frac{1}{2} \times (1 + V_t) + \frac{1}{2}(0 + V_A) \end{cases}$$

$V_A = \frac{27}{17}$

$V_B = \frac{22}{17}$

Bonus Question.

Question 4. Programming exercise. [5 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 (including the agent program for n-step Sarsa) and plots.