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

Collaborators:____

CMPUT 609 Written 4: Temporal-Difference Learning

Due: Tuesday Oct 18 in Gradescope by 2pm (no slip days)

Policy: Can be solved in groups (acknowledge collaborators) but must be written up individually

There are a total of 92 points available on this assignment.

1. [8 points] (greedy w.r.t. a value function)

True or False and justify your answer (8 points total):

(a) **T F** (4 pts) If a policy is greedy with respect to the value function for the equiprobable random policy, then it is an optimal policy.

(b) **T F** (4 pts) If a policy π is greedy with respect to its own value function, v_π, then it is an optimal policy.

2. [22 points] (all about v_π)

Consider the value function v_π for a stochastic policy π and a continuing finite Markov decision process with discounting.

(a) [3 pts] Give an equation *defining* $v_\pi(s)$ in terms of the subsequent rewards $R_{t+1}, R_{t+2}, \ldots$ that would follow if the MDP were in state s at time t. (If you choose to write it in terms of the return, G_t, define your return notation in terms of the underlying rewards.)

(b) [3 pts] Sketch the backup diagram for the dynamic programming algorithm (full backup) for v_π. Find a place to attach the labels s, a, r, and s'.

(c) [4 pts] What is the Bellman equation for v_π? Write it in an explicit form in terms of $p(s', r|s, a)$ so that no expected value notation appears.

(d) [4 pts] Consider the simplest dynamic-programming algorithm for computing v_π. An array $V(s)$ is initialized to zero. Then there are repeated sweeps through the state space, with one update to an array element done for each state. What is that DP update?

(e) [4 pts] Consider the simplest temporal-difference learning method for estimating v_π from experience. An array $V(s)$ is initialized to zero. Then an infinite sequence of experience, $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots$, is processed, with one update to $V(S_t)$ done for each transition. What is the equation for that TD update?

(f) [4 pts] Now consider the simplest Monte Carlo learning method for estimating v_π from *episodic* experience. An array $V(s)$ is initialized to zero. Then an infinite sequence of episodes is experienced, where an individual episode of experience is denoted $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots, R_T, S_T$, where T is the final time step of the episode, S_T is the terminal state, and the value of all terminal states is taken to be zero. When an episode is processed, one update to $V(S_t)$ is made for each time step $t < T$. What is the equation for that Monte-Carlo update? (If you choose to write it in terms of the return, define your return notation in terms of the underlying rewards.)

3. [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
A,2
A,0,B,0,A,2
A,0,B,0,A,0,B,1
B,0,A,2
B,0,A,0,B,1
B,0,A,2
A,0,B,1
B,1
B,1
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.

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.

4. (3 pts) **Multiple choice:** In learning methods, a larger step size, α, usually means

a) more rapidly approaching the final performance level

b) greater error before reaching the final performance level

c) less residual error at the final performance level

d) reduced risk of divergence

e) both a) and d)

5. (3 pts) **Multiple choice:** In TD methods, a larger discount parameter γ, $0 < \gamma < 1$, means

a) a closer approximation to the dynamic-programming solution

b) more concern for immediate rewards relative to later rewards

c) less concern for immediate rewards relative to later rewards

d) both a) and b)

e) both a) and c)

6. [7 points] (weighted average update rule)
Derive the weighted-average update rule (5.7) from (5.6). Follow the pattern of the derivation of the unweighted rule (2.3). Equation numbers are from the SB textbook.
(This is Exercise 5.7 in the SB textbook.)

7. [6 points, 2 for each part] (discuss random-walk value function)

From Figure 6.2 (left) in the SB textbook it appears that the first episode results in a change in only $V(\mathsf{A})$. What does this tell you about what happened on the first episode? Why was only the estimate for this one state changed? By exactly how much was it changed?
(This is Exercise 6.2 in the SB textbook.)

8. [7 points, 2 for each part, 5 for explanation] (could there have been a better alpha?)

The specific results shown in Figure 6.2 (right) in the SB textbook are dependent on the value of the step-size parameter, α. Do you think the conclusions about which algorithm is better would be affected if a wider range of α values were used? Is there a different, fixed value of α at which either algorithm would have performed significantly better than shown? Why or why not?
(This is Exercise 6.3 in the SB textbook.)

9. [6 points] (why Q-learning is off-policy)
Why is Q-learning considered an off-policy control method?
(This is Exercise 6.9 in the SB textbook.)

10. [6 points] (equations of Double Expected Sarsa)
What are the update equations for Double Expected Sarsa with an ε-greedy target policy?
(This is Exercise 6.10 in the SB textbook.)