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

Collaborators:_____

CMPUT 499 Written 5: Planning and Exam Practice

Due: Tuesday Oct 25 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 69 points available on this assignment.

The first two questions are exercises are from the Sutton and Barto textbook, 2nd edition in progress.

8.2 [6 points]
Why did the Dyna agent with exploration bonus, Dyna-Q+, perform better in the first phase as well as in the second phase of the blocking and shortcut experiments?

8.3 [6 points]
Careful inspection of Figure 8.8 reveals that the difference between Dyna-Q+ and Dyna-Q narrowed slightly over the first part of the experiment. What is the reason for this?

1. [6 points] What are the backup diagrams for the following methods?

   a. [3pts] TD(0)

   b. [3pts] DP for q_*

2. [6 points] Why is Dyna-Q considered a *planning* method? A good way to answer this is to say in words what makes something a planning method, and then say, in words, why Dyna-Q is such a something. (Do not say that something is a planning method because it plans.)

3. [22 points] Consider the following MDP. There are three states, A, B, and C, and two actions, right and left. From A, if action left is taken, then there is a fifty-fifty chance of going to the two other states, B and C, with a reward of +1 in either case. If the right action is taken from state A, then there is a fifty-fifty chance of going to states A and C, with a reward of 0 in each case. From state B, the left action always leads to state A with a reward of 0, and the right action always leads to state C with a reward of 1. From state C, the left action half the time leads to B with a reward of 1 and half the time leads to C with a reward of +3, and the right action always leads back to C with a reward of 1.

   a. [1pt] What is the state set? $\mathcal{S}$=

   b. [1pt] What is the action set? $\mathcal{A}$=

   c. [1pt] What is the reward set? $\mathcal{R}$=

   d. [4pts] What is the dynamics function, p? Make a table with five columns with headings across the top for s, a, s', r, and $p(s',r|s,a)$. Add a row to the table for any values of the first four for which the last, $p(s',r|s,a)$, is non-zero.

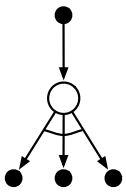
e. [4pts] Draw a picture of the states and actions of the MDP (open circles for states, solid dark ones for actions). Use the convention that the action leaving a state from its left side is the left action and the action leaving a state from its right side is the right action. You may also use the convention that when there are two arcs leaving an action node, then the two possibilities occur with equal probability. Label the states A, B, and C. Label the arcs from actions to states with the appropriate reward.

f. [5pts] Suppose the discount-rate parameter γ is 0.5. What is the optimal deterministic policy?

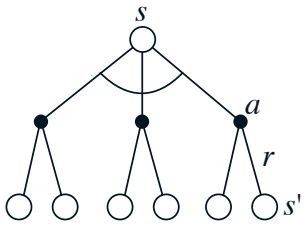
g. [6pts] What is the state-value function for the optimal policy? Give the numerical values of the three states. Show your work.

4. [8 points - 2 points each] What are the names of the algorithms with these backup diagrams?

a.



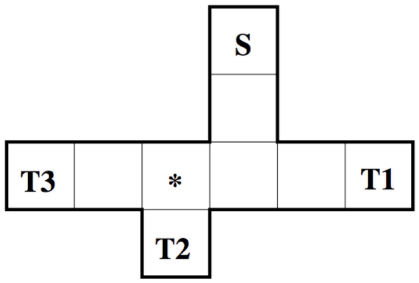
b.



c.

d.

5. (15 points total) Consider the gridworld shown below. There is a start state S and three terminal states, T1, T2, and T3. The task is episodic and discounted, with each episode starting in S and ending in one of the terminal states. Actions move one square at a time step (no diagonal moves). Terminal states T1, T2, and T3 respectively deliver rewards of 2, 4, and 6. Moving into the state marked * delivers a reward of -1. All other rewards are 0. The optimal policy depends on the value of the discount rate γ, $0 \leq \gamma \leq 1$.



(a) (5 pts) For what range of γ values does an optimal policy take the agent to T1?

(b) (5 pts) For what range of γ values does an optimal policy take the agent to T2?

(c) (5 pts) For what range of γ values does an optimal policy take the agent to T3?