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

Collaborators:_____

CMPUT 499 Written 3: Solving MDPs

Due: Tuesday Oct 4 in Gradescope by 2pm (no slip days)

Policy: Can be solved in groups (acknowledge collaborators) but must be written up individually. Be sure to explicitly answer each subquestion posed in each exercise.

There are a total of 90 points available on this assignment.

Question 1 [12 points, 6 for each part]
In Figure 3.8 in the SB textbook, Let us call the state to the right of state A and to the left of state B as state C. The figure gives the optimal value of state C as 22.0, to one decimal place. Use your knowledge of the optimal policy and Equation 3.2 in the SB textbook to express this value symbolically, and then to compute it to three decimal places.
(This is a modified form of Exercise 3.16 in the SB textbook.)

Question 2 [4 points]
Give an equation for v_* in terms of q_*. Break it down into a form with no expectations.

Question 3 [4 points]
Give an equation for q_* in terms of v_*. Break it down into a form with no expectations.

Question 4 [4 points]
Give an equation for π_* in terms of q_*. You can assume $\pi*$ is deterministic. (no expectations.)

Question 5 [4 points]
Give an equation for π_* in terms of v_*. You can assume $\pi*$ is deterministic. (no expectations.)

Question 6 [6 points, 3 for each part]
In Example 4.1 (from the SB textbook), if π is the equiprobable random policy, what is $q_\pi(4,\texttt{up})$?

What is $q_\pi(5,\texttt{left})$?

Question 7 [10 points, 5 for each part]
In Example 4.1 (from the SB textbook), suppose a new state 15 is added to the Gridworld just to the right of state 11, and actions up, left, and right take the agent to states 7, 11, and 10, respectively. Action down, takes the agent to a terminal state. Assume that the transitions from the original states are unchanged. What, then, is $v\pi(15)$ for the equiprobable random policy? In the next step, suppose the dynamics of state 11 are also changed, such that action right from state 11 takes the agent to the new state 15. What is $v\pi(15)$ for the equiprobable random policy in this case?

Now suppose the dynamics of state 8 are also changed, such that action `left` from state 8 takes the agent to the new state 15. What is $v_\pi(15)$ for the equiprobable random policy in this case?

Question 8 [9 points, 3 for each part]
What are the equations analogous to (4.3), (4.4), and (4.5) in the SB textbook for the action-value function q_π and its successive approximation by a sequence of functions $q_0, q_1, q_2, \ldots$?
(This is Exercise 4.3 in the SB textbook.)

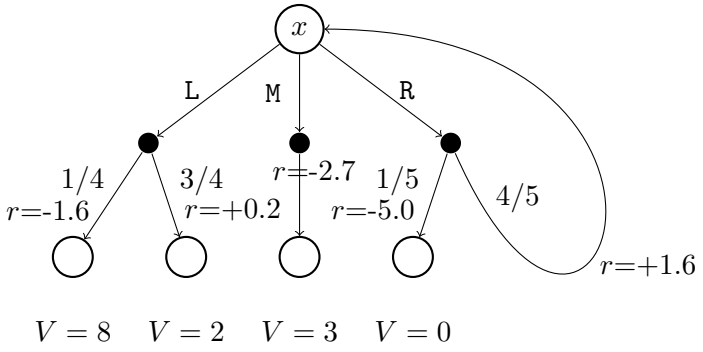
Question 9 [8 points]
How would policy iteration be defined for action values? Give a complete algorithm for computing q_*, analogous to that on page 85 in the SB textbook for computing v_*. Please pay special attention to this exercise, because the ideas involved will be used throughout the rest of the course.
(This is Exercise 4.6 in the SB textbook.)

Question 10 [5 points]
What is the analog of the value iteration backup, (4.10) in the SB textbook for action values, $q_{k+1}(s, a)$?
(This is Exercise 4.10 in the SB textbook.)

Question 11 [6 points, 2 for each part]
Consider the diagrams on the right in Figure 5.1 in the SB textbook. Why does the estimated value function jump up for the last two rows in the rear? Why does it drop off for the whole last row on the left? Why are the frontmost values higher in the upper diagrams than in the lower?
(This is Exercise 5.1 in the SB textbook.)

Question 12 [6 points]
What is the backup diagram for Monte Carlo estimation of q_π?
(This is Exercise 5.2 in the SB textbook.)

Question 13 [12 points, 6 points for each part]
Consider the following fragment of an MDP graph. The fractional numbers indicate the world's transition probabilities and the whole numbers indicate expected rewards. The three numbers at the bottom indicate what you can take to be the value of the corresponding states. The discount rate is 0.9. Let π be the equiprobable random policy (all actions equally likely). What is the value of the top node for this policy and for the optimal policy? That is, what are $v_\pi(x)$ and $v_*(x)$? Show your work.



$v_\pi(x)$ =

$v_*(x)$ =