Last name: Liang First name: Haoming SID#: 1430396

Collaborators: __________

CMPUT 366/609 Assignment 2: Markov Decision Processes 1

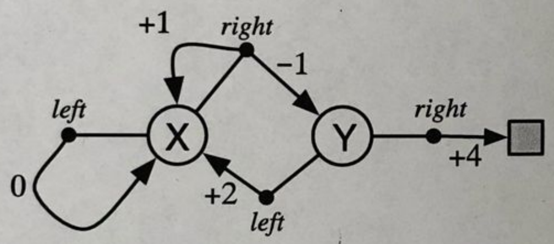
Due: Thursday Sept 28, 11:59pm by gradescope

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

There are a total of 100 points on this assignment, plus 15 extra credit points available.

Be sure to explicitly answer each subquestion posed in each exercise.

Question 1: Trajectories, returns, and values (**15 points total**). This question has six subparts.



Consider the MDP above, in which there are two states, X and Y, two actions, *right* and *left*, and the deterministic rewards on each transition are as indicated by the numbers. Note that if action *right* is taken in state X, then the transition may be either to X with a reward of +1 or to Y with a reward of −1. These two possibilities occur with probabilities 3/4 (for the transition to X) and 1/4 (for the transition to state Y).
Consider two deterministic policies, π_1 and π_2:

$$\pi_1(\mathsf{X}) = \textit{left} \qquad \pi_2(\mathsf{X}) = \textit{right}$$
$$\pi_1(\mathsf{Y}) = \textit{right} \qquad \pi_2(\mathsf{Y}) = \textit{right}$$

(a) (2 pts.) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_1:

X, left, 0, X ...

(b) (2 pts.) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_2:

X, right, +1, X, right, +1, X, right, +1, X, right, −1, Y, right, +4

(c) (2 pts.) Assuming the discount-rate parameter is $\gamma = 0.5$, what is the return from the initial state for the second trajectory?

$G_0 = R_1 + (0.5)R_2 + (0.25)R_3 + (0.125)R_4 + (0.0625)R_5 = 1.875$

(d) (2 pts.) Assuming $\gamma = 0.5$, what is the value of state Y under policy π_1?

$v_{\pi_1}(\mathsf{Y}) = 4$

(e) (2 pts.) Assuming $\gamma = 0.5$, what is the action-value of X,*left* under policy π_1?

$q_{\pi_1}(\mathsf{X}, left) = 0$

(f) (5 pts) Assuming $\gamma = 0.5$, what is the value of state X under policy π_2?

$v_{\pi_2}(\mathsf{X}) = (1+0)\ \frac{3}{4}[1+(0.5)v_{\pi_2}(X)] + \frac{1}{4}[-1+(0.5)\underbrace{v_{\pi_2}(Y)}_{4}] = \frac{3}{4}[1+(0.5)v_{\pi_2}(X)] + \frac{1}{4}(1) = \frac{3}{4} + \frac{3}{8}v_{\pi}(X) + \frac{1}{4}$

$\frac{5}{8}v_{\pi_2}(X) = 1 \rightarrow v_{\pi_2}(X) = \frac{8}{5} = 1.6$

a) Example 1:

Chess game

State: during a game, the location of the chess units

Action: make a move

Reward: win (+1) lose (-1)

The limitation of this example is that simply by assigning rewards to winning and losing, the agent is not learning strategies of winning the game (or by a very slow pace)

Example 2:

Robbery

State: your inventory (full or not)

Action: select an item to rob

Reward: the item itself (the value of the item)

In this example, the burglar (agent) tends to select relatively smaller or lighter items with high values for multiple runs (assuming the burglar is not arrested)

Example 3:

On your way home

State: at home or not

Action: at a crossroad, select a direction to go

Reward: for shorter time spent on going home, the higher the reward

In this example, once you find the shorter path which allows you going home quicker, you will be more likely pick this path for the future

b) Similar to playing a game, due to the complexity of a maze with lots of dead ends, simply by setting a reward at the end does not show any improvement on reaching a goal. The agent escaping the maze will still attempt to try the routes that do not lead to the exit, thus it inefficiently attempts to escape.

To effectively communicate to the agent, a better design of reward would be setting a reward every time the agent moves to a location that leads to the exit, then if the agent tries to escape the maze for the next run (another episode), it will learn to pick the path leading to the rewards.

c) $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \gamma^4 R_{t+5} + \dots$

$= R_{t+1} + \gamma G_{t+1}$

$R_1 = -1 \quad R_2 = 2 \quad R_3 = 6 \quad R_4 = 3 \quad R_5 = 2 \quad \gamma = 0.5$

$G_5 = 0 \quad G_4 = 2 \quad G_3 = 3 + 0.5(2) = 4 \quad G_2 = 6 + 0.5(4) = 8 \quad G_1 = 2 + 0.5(8) = 6$

$G_0 = -1 + 0.5(6) = 2$

d) $R_1 = 2 \quad R_{>1} = 7 \quad \gamma = 0.9$

$G_1 = 7 / (1 – 0.9) = 70 \;\; G_0 = 2 + 0.9(70) = 65$

e) $v_\pi(s)$
$= (1/4)[0 + 0.9(0.7)] + (1/4)[0 + 0.9(-0.6)] + (1/4)[0 + 0.9(-1.2)] + (1/4)[0 + 0.9(-0.4)]$
$= (1/4)[0.9(-1.5)] = (1/4)(-1.35)$ ← round to nearest 10^{-1}
$= (1/4)(-1.4) = -0.35$ ← round to nearest 10^{-1}
$= -0.4$

f) $q_\pi(s,a) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma q_\pi(s',a')]$

g) The signs are very important, since the reward value is the motivation of agent.

$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$
$= \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c)$
$= (\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}) + (\sum_{k=0}^{\infty} \gamma^k c)$

$v_c = \sum_{k=0}^{\infty} \gamma^k c$

h) It will have effect, because if a negative reward value add a constant increased to a positive reward, the agent will incorrectly do learning.

Example:
If a negative reward route leading to a dead end in a maze changes to a positive value due to adding a constant, the agent will keep running into that dead end.

i) $v_\pi(s) = \sum_a \mathbf{E}_\pi[R_t \mid s_t = s, a_t = a]\, \pi(s,a)$

$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a)$

j) Starting from A' = 16

symbol
↓
$v_\pi(s) = 16 / 0.9^4 = 24.387$ to three decimal places
≈ 24.4