Last name: Huang First name: Yongjia SID#: 1461069

Collaborators: Donglin Han

# CMPUT 366/609 Assignment 2: Markov Decision Processes 1

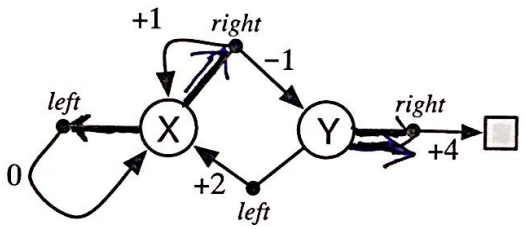
Due: Thursday Sept 28, 11:59pm by gradescope

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, $\pi_1$ and $\pi_2$:

$\pi_1(\text{X}) = \textit{left}$ $\quad$ $\pi_2(\text{X}) = \textit{right}$

$\pi_1(\text{Y}) = \textit{right}$ $\quad$ $\pi_2(\text{Y}) = \textit{right}$

(a) (2 pts.) Show a typical trajectory (sequence of states, actions and rewards) from X for policy $\pi_1$: $X_t, \text{left}_t, 0_t, X_{t+1}, \text{left}_{t+1}, 0_{t+1}, \ldots\ldots$

(b) (2 pts.) Show a typical trajectory (sequence of states, actions and rewards) from X for policy $\pi_2$: $X_t, \text{right}_t, +1_t, X_{t+1}, \text{right}_{t+1}, +1_{t+1}, X_{t+2}, \text{right}_{t+2}, +1_{t+2}, X_{t+3}, \text{right}_{t+3}, -1_{t+3}, Y_{t+4}, \text{right}_{t+4}, \ldots$

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

$G_0 = 1 + 0.5\times 1 + 0.5^2\times 1 + 0.5^3\times(-1) + 0.5^4\times 4 = 1.875$

(d) (2 pts.) Assuming $\gamma = 0.5$, what is the value of state Y under policy $\pi_1$?

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

(e) (2 pts.) Assuming $\gamma = 0.5$, what is the action-value of X,*left* under policy $\pi_1$?

$q_{\pi_1}(\text{X}, \textit{left}) = 0 + 0.5\times 0 + 0.5^2\times 0 \ldots = 0$

(f) (5 pts) Assuming $\gamma = 0.5$, what is the value of state X under policy $\pi_2$?

$v_{\pi_2}(\text{X}) = \frac{3}{4}[1 + 0.5 v_{\pi_2}(X')] + \frac{1}{4}[-1 + 0.5 v_{\pi_2}(Y)] = \frac{8}{5}$

$v_{\pi_2}(Y') = \frac{11}{4}$

**Question 2 [85 points total].** This question has **ten** subparts. The first 9 subparts are questions from SB textbook, second ed. The last subpart (j) is not from SB.

**(a) Exercise 3.1 [6 points]** (Example RL problems).

**(b) Exercise 3.7 [6 points, 3 for each subquestion]** (problem with maze running).

**(c) Exercise 3.8 [6 points]** (computing returns).

**(d) Exercise 3.9 [9 points]** (computing an infinite return).

**(e) Exercise 3.11′ [12 points]** (verify Bellman equation in gridworld example). **(This differs from the textbook.)** The Bellman equation (3.13) must hold for each state for the value function $v_\pi$ shown in Figure 3.3 (see SB text, 2nd ed.). As an example, show numerically that this equation holds for the state just below the center state, valued at -0.4, with respect to its four neighboring states, valued at +0.7, –0.6, –1.2, and -0.4. (These numbers are accurate only to one decimal place.)

**(f) Exercise 3.12 [12 points]** (Bellman equation for action values, $q_\pi$ ).

**(g) Exercise 3.13 [9 points]** (Adding a constant reward in a continuing task).

**(h) Exercise 3.14 [9 points, 3 for each subquestion, 3 for the example]** (Adding a constant reward in an episodic task)

**(i) Exercise 3.15 [8 points, 4 points for each equation]** (half-backup $v_\pi$).

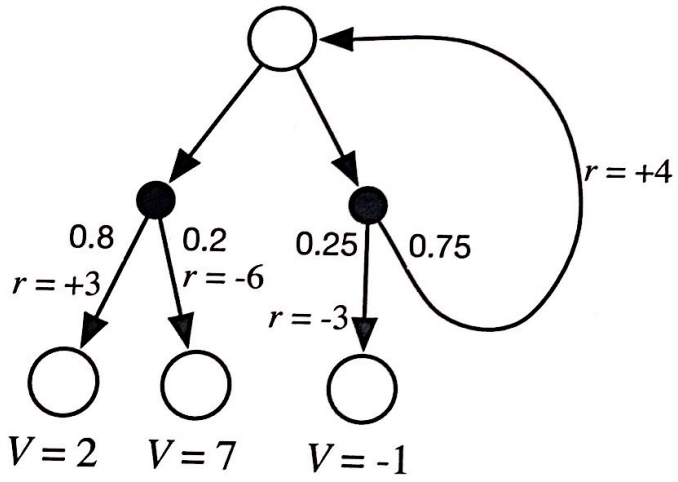
**(j) [8 points, 4 for symbolic form, 4 points for numeric answer]** Figure 3.6 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.7) to express this value symbolically, and then to compute it to three decimal places. Hint: Equation (3.9) is also relevant.

**Bonus Questions [total 15 points available].** There are **two** bonus questions.

**Question 3**: Trajectories, returns, and values (**10 Bonus points**)

Consider the following fragment of an MDP graph. The fractional numbers indicate the world?s transition probabilities and the whole numbers indicate the expected rewards. The three numbers at the bottom indicate what you can take to be the value of the corresponding states. The discount is 0.8. What is the value of the top node for the equiprobable random policy (all actions equally likely) and for the optimal policy? Show your work.

$v_\pi =$ $v_* =$



**Question 4 [5 bonus points].** Complete Exercise 3.6 (episodic pole balancing). See SB textbook, second ed.

Question 2: Exercise 3.1

a). ① Consider a walking robot has two states: leg lift and leg forward. The robot might need to walk on different landscape and in order to learn movements fast and smooth. the agent need to select best action on each states (each actions with different altitude and length). The reward can be +1 if the robot successfully walk forward one step, and −1 if it falls down.

② "Smart light" has two states: day time and night. the actions are different brightness and color, the reward might be 0 if people does not change light mode, and −1 if people change the light mode by hand.

③ ~~The news on t~~

~~"Smart pencil sharpener"~~

Music player has states: play music while studying, when in gym, and when sleeping. the actions are different type of music, the reward can be +2 if user press "like", +1 if user does not do anything, and −1 if user cut the music.

· The limitation of above examples ③ is that the learning agent are easy to get confused since human are easy to change their mood.

6. Exercise 3.7.

1) The equation (3.6) is $G_t = R_{t+1} + R_{t+2} + R_{t+3} + R_T$

~~if the goal is to maximize (3.6), the robot will be focus on the maximum current total goal and it will give up~~

The design of the reward ~~system~~ system confuses the robot's learning agent, since the agent only gets reward +1 when it escape and 0 for all other actions, of course the agent will try different method to escape the maze but it gets no ~~penalty~~ punishment for ~~not~~ wasting time. So, as long as the agent's ~~decision~~ decision helps the robot escape the maze, the total reward is maximized and it is always equals to (+1)

(+1): if there are n available step I can take in the future

2). No, the ~~reward~~ system ~~is~~ needs to be changed to: get reward of (+1000) if it escapes, (-1) ~~if into dead end~~ if into a dead end, and 0 for all others and it needs to be treated as episodic task with discounting, which means the goal is to maximize the expected future total reward (3.7).

c).

Exercise 3.8

$\gamma = 0.5$ , $R_1 = -1$ $R_2 = 2$ $R_3 = 6$ , $R_4 = 3$ $R_5 = 2$ $T = 5$

What are $G_0 - G_5$?

$G_5 = R_6$ ... Since $T = 5$ and there are no time step 6. $G_5 = 0$.

$G_4 = R_5 = \boxed{2}$

$G_3 = R_4 + \gamma R_5 = 3 + 0.5 \cdot 2 = \boxed{4} = R_4 + \gamma G_4$

$G_2 = R_3 + \gamma R_4 + \gamma^2 R_5 = \boxed{8} = R_3 + \gamma G_3$

$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 = \boxed{6} = R_2 + \gamma G_2$

$G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 R_5 = \boxed{2} = R_1 + \gamma G_1$

Exercise 3.9

Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are $G_1$ and $G_0$?

$$G_1 = \frac{7}{1-0.9} = \boxed{70}$$

$$G_0 = R_1 + \gamma \cdot G_1 = 2 + 0.9 \times 70 = \boxed{65}$$

Exercise. 3.11.

By using the diagram for Bellman equation for $(v_{s_1})$.

" It states that the value of the start state must equal the discounted value of the expected next value. plus the reward expected along the way"

$-0.4$ $\quad$ $0.9$ $\quad$ $\frac{0.7+(-0.6)+(-1.2)+(-0.4)}{4}$ $\quad$ $+$ $\quad$ $0\times 4$

$$4\times 0 + 0.9 \cdot \frac{0.7+(-0.6)+(-1.2)+(-0.4)}{4} \approx -0.4$$

Exercise 3.12.

- Bellman equation for $v_\pi$:

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r_a + \gamma v_\pi(s')\right]$$

- Bellman equation for $q_\pi(s,a)$ in terms of the action value, $q_\pi(s',a')$

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

3). Exercise 3.13

Equation 3.7: ~~$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$ (3.7)~~

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

# adding a constant c to all the rewards

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

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \underbrace{\sum_{k=0}^{\infty} \gamma^k \cdot c}_{V_c}$$

Thus, the addition of reward c does not affect the relative values of any states under any policies. only the intervals between them are import

n).

Exercise 3.14.

- Adding a constant c to an episode task would not change the task.

- Since all the reward in an episode are added by c, considering a ε-greedy method, or multi armed bandit problem, the constant c does not matter at all.

By assuming the laugth of episodic task is fixed, thus the sum ~~of the~~ is fixed.

). Exercise 3.15

$V_{\pi}(s) = \sum_{a} E_{\pi}(a|s)$

$\sum_{a} \pi(a|s) \cdot E_{\pi}[R_t | S_t = s, a_t = a]$

$V_{\pi}(s) = \sum_{a} \pi(a|s) \cdot q_{\pi}(s,a)$

j).

(3.7) $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

(3.9) $G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$

**24.418**

Consider $v^*(s')$. $= 16.0$.

$16.0 = 0.9 \cdot 17.8 = 16.02$.

$16.02 \times 0.9 + 10 = \boxed{24.418}$.