Last name: ZHU First name: HAOTIAN SID#: 1467741

Collaborators:

CMPUT 366/609 Assignment 2: Markov Decision Processes 1

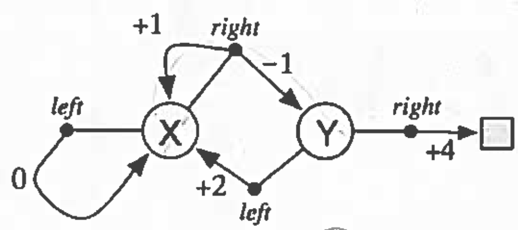
Due: Thursday Sept 28, 11:59pm by gradescope

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:

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

$\pi_2(\mathsf{X}) = right$
$\pi_2(\mathsf{Y}) = right$

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

State	X	X	X	X
action	left	left	left	left
reward	0	0	0	0

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

state	X	X	X	X	Y
action	right	right	right	right	right
reward	+1	+1	+1	−1	+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?

$G_0 = \frac{15}{8}$

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

$v_{\pi_1}(\mathsf{Y}) = \mathbb{E}_{\pi_1}[G_0 \mid S = Y] = 4$

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

$q_{\pi_1}(\mathsf{X}, left) = \mathbb{E}[G_0 \mid A = \text{left}, S = X] = 0$

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

$v_{\pi_2}(\mathsf{X}) = \mathbb{E}_{\pi_2}[G_0 \mid S = X] = \frac{3}{4} \times [1 + \frac{1}{2} V_{\pi_2}(X)] + \frac{1}{4} \times [-1 + 4]$

$\Rightarrow V_{\pi_2}(x) = 1.6$

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_π 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_π).

(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_π).

(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.

(a) ① reply text machine : human send a message to machine and machine replies messages. Action : Send message. state : different group of people. reward : get from people, people rate the message.

② Shot basketball : state : different distance from basket. Action : choose angle and power. reward get positive reward when score points. limit : machine can not distinguish bank shot and shot

③ Find quickest drive way : -assume we need find a quickest way from A to B. Action : choose a way. State : different time, like morning peak, weekday. weekend and so on. reward time cost.

(b) since we use epsodes task, the $G_t = \sum_{i=t+1}^{T} R_i$,
When we choose the reward $= 0$ for any step and reward $= 1$ for escaping maze. The reward is not relative to steps. Computer maximizes goal by escaping maze instead of shortest way to escape maze. To fix it, we need set reward $= -1$ for each step, get reward $= 1$ if escape

(c) $\gamma = 0.5$ $\quad R_1 = -1 \quad R_2 = 2 \quad R_3 = 6 \quad R_4 = 3 \quad R_5 = 2. \quad T = 5$

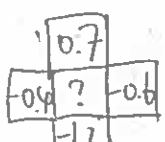
$$G_5 = 0, \quad G_4 = \tfrac{1}{2} \times G_5 + R_5 = 2$$

$$G_3 = 2 \times \tfrac{1}{2} + 3 = 4, \quad G_2 = \tfrac{1}{2} \times 4 + 6 = 8$$

$$G_1 = \tfrac{1}{2} \times 8 + 2 = 6 \qquad G_0 = \tfrac{1}{2} \times 6 + (-1) = 2$$

(d)
$$G_0 = 2 + 7 \cdot \gamma + 7 \cdot \gamma^2 + \cdots$$
$$= 7 - 5 + 7 \cdot \gamma + 7 \cdot \gamma^2 + \cdots$$
$$= -5 + 7 \cdot \left(\frac{1}{1-0.9}\right)$$
$$= -5 + 70 = 65$$
$$G_1 = 7 + 7 \cdot \gamma^2 + 7 \cdot \gamma^3 + \cdots$$
$$= 7 \cdot \frac{1}{1-0.9} = 70$$

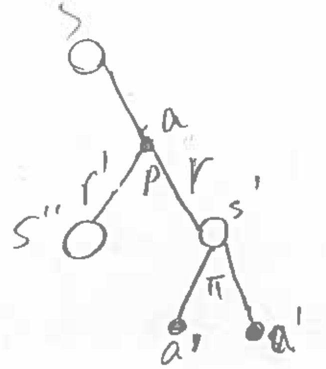
(e) $V_\pi(s) = \mathbb{E}[G_t \mid S_t = s]$, we now at $(2,2)$



choose the direction of next move randomly
$\Rightarrow P(\text{north/south/west/East}) = \frac{1}{4}$ & reward =

$$V_\pi(s) = \tfrac{1}{4}(-0.6 + 0.7 - 0.4 - 1.2) \times 0.9 = -0.3375$$

(f)



$q_\pi(s,a)$

$$= \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

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

(g) $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$ becomes $G_t' = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + C)$

$$\Rightarrow G_t' = \sum_{k=0}^{\infty} (\gamma^k R_{t+k+1} + \gamma^k \cdot c)$$

$$= G_t + \sum_{k=0}^{\infty} \gamma^k \cdot c = G_t + \frac{c}{1-\gamma}$$

$\frac{c}{1-\gamma}$ is a constant

$$V'(s) = \mathbb{E}\left[G_t + \frac{c}{1-\gamma} \,\middle|\, S_t = s\right] = \mathbb{E}[G_t \mid S_t = s] + \mathbb{E}\left[\frac{c}{1-\gamma} \,\middle|\, S_t = s\right]$$

$$= V(s) + \frac{c}{1-\gamma}$$

therefore only signs of the rewards are Important.

(h)

YES, for example, maze runing,

let each step's reward is -1, when computer escapes from maze, gets reward=(

$G_t = \sum_{i=t+1}^{T} R_i$; to max G_t, computer needs use less step.

Now we add a c (constant) to all R_i

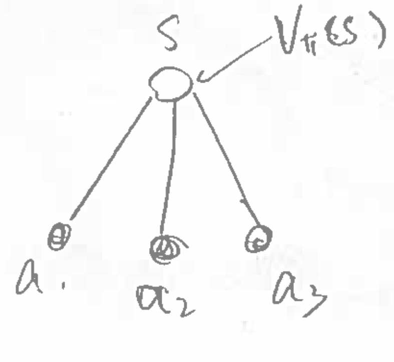
$$\Rightarrow G'_t = \sum_{i=t+1}^{T} (R_i + c)$$

if $c = 2$ then $R_i + c$ alway > 0

then G_t max when computer takes more steps.

Therefore it may effect episodic task.

(i)

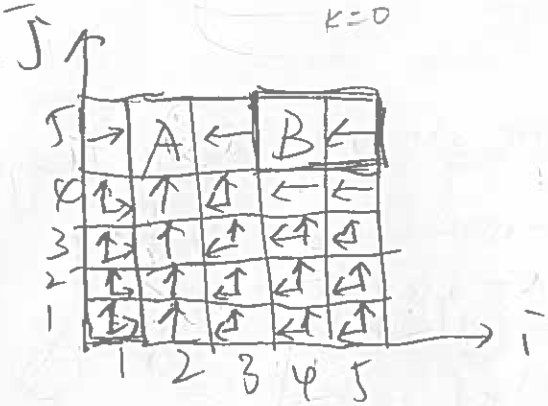


$$V_\pi(s) = \pi(a_1|s) \cdot q(s,a_1) + \pi(a_2|s) \cdot q(s,a_2) + \pi(a_3|s) \cdot q(s,a_3)$$

(1)

(3.9) $G_t = \frac{1}{1-\gamma}$

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



the max $q_{\pi *}(s,a)$ in $S=(2,5)$ always go down

$$V_*(s=(2,5)) = \max q_{\pi *}(s,a)$$
$$= \max_a \sum_{s',r} p(s',r|s,a)\left[r+\gamma V_*(s')\right]$$
$$= 10+0+0+0+0+10\cdot(0.9)^5$$
$$+0+0+0+0+10\cdot(0.9)^{10}+\cdots+\cdots$$

each 5 steps get $10\cdot(0.9)^k$

$$\Rightarrow \quad V_*(s=(2,5)) = 10+(0.9)^5\cdot 10$$
$$+(0.9)^{10}\cdot 10+\cdots+$$

let $q = (0.9)^5$

$$\Rightarrow \quad V_*(s=(2,5)) = 10+10\cdot q+10\cdot q^2+\cdots+q^{\cdot}$$
$$= \frac{10}{1-q} = \frac{10}{1-(0.9)^5}$$
$$\approx 24.419$$