Last name: Li First name: Xianhang SID#: 1465904

Collaborators:

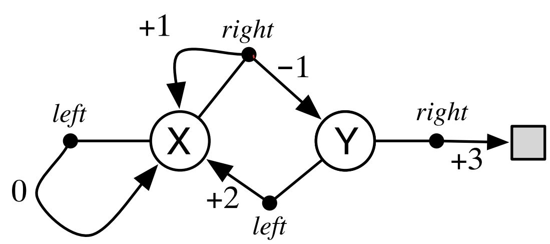
CMPUT 366/609 Assignment 2: Markov Decision Processes 1

Due: Tuesday Oct 2, 11:59pm by Gradescope

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 2/3 (for the transition to X) and 1/3 (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, left, 0 ......

(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, Y, right, 3, . termination.

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

$G_0 = 1 + \gamma \cdot 1 + \gamma^2 \cdot (-1) + \gamma^3 (3) = 1.625$

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

$v_{\pi_1}(\mathsf{Y}) = 1 \cdot 3 = 3$

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

$q_{\pi_1}(\mathsf{X}, left) = 1 \cdot 0 + \gamma v_{\pi_1(X)} = 0$

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

$v_{\pi_2}(\mathsf{X}) = \frac{2}{3}\left[1 \cdot 1 + \gamma V_{\pi_2}(X)\right] + \frac{1}{3}\left[1 \cdot (-1) + \gamma \cdot 3\right]$

$\gamma = 0.5 \quad V_{\pi_2}(X) = \frac{5}{4}$

Question 2 [85 points total]. This question has **ten** subparts. The questions are questions from SB textbook, second ed.

(a) Exercise 3.1 [6 points] (Example RL problems). chess ; maze ; Tower of Hanoi.

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

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

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

(e) Exercise 3.14 [12 points] (verify Bellman equation in gridworld example).

(f) Exercise 3.15 [9 points] (Adding a constant reward in a continuing task).

(g) Exercise 3.16 [9 points, 3 for each subquestion, 3 for the example] (Adding a constant reward in an episodic task)

(h) Exercise 3.17 [12 points] (Bellman equation for action values, q_π).

(i) Exercise 3.18 [8 points, 4 points for each equation]. First write the answer with expected value notation, then replace the expected value with a summation.

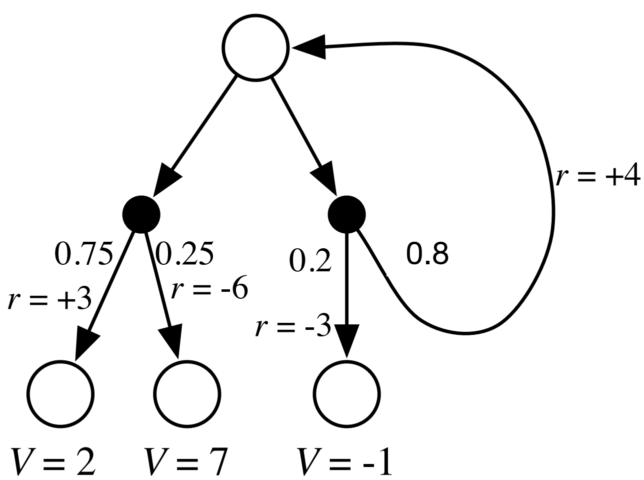
(j) Exercise 3.24 [8 points, 4 for symbolic form, 4 points for numeric answer]

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.

$v_\pi =$ 4.257353 $\qquad v_* =$ 6.77



$V_\pi = 0.5 \cdot q\_black\_left + 0.5 \cdot q\_black\_right$

$= 0.5 \cdot (0.75 \cdot (3+\gamma \cdot 2) + 0.25 \cdot (-6 + \gamma \cdot 7)) + 0.5 \cdot (0.2 \cdot (-3 \cdot \gamma)$

$+ 0.8 \cdot (4 + \gamma v_\pi))$

$= 1.675 + 1.22 + 0.32 \cdot V_\pi$

$V_\pi = 4.257353$

$V_{left} = 3.35$

$V_{right} = 6.77$

$V_* = \max(V_{left}, V_{right}) = 6.77$

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

Exercise 3.7.

$$G_t = \sum_{k=t+1}^{T} R_k.$$

the agent will always get +1 as reward after escaped, no matter how slow or fast. we would like to make the agent get out as fast as possible, so the reward we set was wrong.

we have to tell the agent the defination of time penalty. we can make each step inside maze get −1 reward, or give a discount rate $\gamma = 0.9$.

Exercise 3.8

$$G_n = r_{n+1} + \gamma \cdot G_{n+1}$$

$$G_5 = 0 + 0.5 \cdot G_6 = 0$$

$$G_4 = 2 + 0.5 \cdot 0 = 2$$

$$G_3 = 3 + 0.5 \cdot 2 = 4$$

$$G_2 = 6 + 0.5 \cdot 4 = 8$$

$$G_1 = 2 + 0.5 \cdot 8 = 6$$

$$G_0 = -1 + 0.5 \cdot 6 = 2$$

Exercise 3.9.

$$G_n = \sum_{k=n}^{\infty} \gamma^{k-1} R_k$$

$$G_1 = 7 \sum_{k=2}^{\infty} \gamma^{k-1} = 7 \cdot \frac{1}{1-\gamma} = 70$$

$$G_0 = 2 + 0.9\, G_1 = 65$$

Exercise 3.14

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

$$V_\pi(s) = \sum_a \pi(a|s) \sum 0.25 \cdot \left[0 + \gamma\,(2.3+0.7+0.4-0.4)\right]$$

$$V_\pi(s) = 0.25 \cdot 0.9 \cdot 3$$

$$= 0.675 \approx 0.7$$

Exercise 3.15

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

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

$$V_c = \sum_{k=0}^{\infty} \gamma^k \cdot C$$

$$= C \cdot \sum_{k=0}^{\infty} \gamma^k = C \cdot \frac{1}{1-\gamma}$$

Since C and γ are two constants, V_c is a contant aswell

Exercise 3.16.

for example. the agent will receive a -1 reward while it is in the maze, if we add a contant $c > 1$, the agent will prefer to stay in the maze (reward will alway be a positive number and increase)!

The agent will trying to max the reward, so the agent will stay in the maze forever.

Exercise 3.17

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

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

Exercise 3.18

$$v_\pi(s) = \mathbb{E}_\pi\left[q_\pi(S_t, A_t) \mid S_t = s\right]$$

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

Exercise 3.24

$$v_*(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right]$$

$$= 10 + \gamma 0 + \gamma^2 0 + \gamma^3 0 + \gamma^4 0 + \gamma^5 10 + \gamma^6 0 + \gamma^7 0 + \gamma^8 0 + \gamma^9 0 + \gamma^{10} 10 \ldots\ldots$$

$$= \gamma^0 10 + \gamma^5 10 + \gamma^{10} 10 + \ldots\ldots$$

$$= 10 \sum_{k=0}^{\infty} (\gamma^5)^k$$

$$= 10 \cdot \frac{1}{1-\gamma^5}$$ when $\gamma = 0.9$, $v_*(s) = 24.419$.