Last name: Li First name: Minghan SID#: 1561234

Collaborators:

CMPUT 366/609 Assignment 2: Markov Decision Processes 1

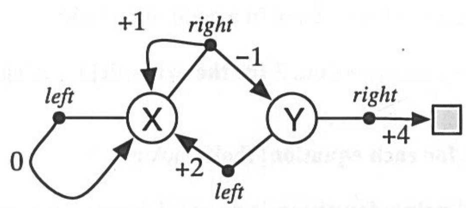
Due: Thursday Sept 28, 11:59pm by gradescope

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, 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, Y, right, +4, terminal

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

$G_0 =$ 1.5

(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?

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

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

$v_{\pi_2}(\mathsf{X}) =$ 1.6

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)

i) Farming: The task is to plant some crops in Spring and harvest as much as possible in Autumn.

- The states will be the health condition of the crops
- The actions will be watering, fertilizing and weeding.
- The rewards will be how much you can harvest at the end of the year (positive)

ii) Skiing: The task is to learn how to ski.

- The States will be the physical positions of your body. e.g. angle of the joints, and the speed of skiing.
- The actions will be how you move your joints and limbs
- The rewards will be ~~how fast can you ski and~~ how many times you fall over (negative).

iii) Paper writing: The task is to write a qualified academic paper.

- The states will be the content of your paper.
- The actions will be how you refine your paper.
- The rewards will be the feedback from advisors, editors or other people.

(b)

i) The problem is that the rewards do not tell the robot to escape the maze as soon as possible before it receives the terminal reward +1. Because the rewards are zero everywhere, so the value function will be zero for all states before it escapes, therefore the actions will be random in the maze.

ii) No. The rewards tell the agent that "staying in the maze is ok" rather than "staying in the maze is bad and escaping it as soon as possible is good". Therefore we should set the rewards to be 0 at the exit and -1 everywhere else. In this way the agent would know that staying in the maze is bad and avoid wandering around some dead-ends.

(c) $G_5 = 0$, $G_4 = 2 + 0.5 \times G_5 = 2$,

$G_3 = 3 + 0.5 \cdot G_4 = 4$, $G_2 = 6 + 0.5 \cdot G_3 = 8$,

$G_1 = 2 + 0.5 \cdot G_2 = 6$, $G_0 = -1 + 0.5 G_1 = 2$

(d) $G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \cdots$

$= \sum_{k=0}^{\infty} \gamma^k R_{k+2} = 7 \cdot \frac{1}{1-\gamma} = 70$

$G_0 = R_1 + \gamma G_1$

$= 2 + 0.9 \times 70 = 65$

(e)

$V(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) (r + \gamma V(s'))$

$= \frac{1}{4} \times (0.9 \times ((-0.6) + (-0.4) + (-1.2) + 0.7)$

$\simeq 0.3375$

which is closed to the value 0.4 (in one decimal place)

(f)

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

(g) Proof : we assume $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

$V'(s) = E[G_t' | s] = E[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) | s]$

$= E[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + c \sum_{k=0}^{\infty} \gamma^k | s]$

$= E[G_t | s] + \frac{c}{1-\gamma}$

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

$= V(s) + V_c(s)$

$\therefore V_c = \frac{c}{1-\gamma}$

(h) It's different from continuing task.

In episodic case, we assume $G_t = \sum_{k=0}^{T} \gamma^k R_{t+k+1}$

$V'(s) = E[G_t' | s] = E[\sum_{k=0}^{T} \gamma^k (R_{t+k+1} + c) | s]$

$= E[\sum_{k=0}^{T} \gamma^k R_{t+k+1} + c \sum_{k=0}^{T} \gamma^k | s]$

$= E[G_t | s] + c \sum_{k=0}^{T} \gamma^k$

$= V(s) + c \cdot \sum_{k=0}^{T} \gamma^k$

We can see the return is different. In continuing case we add a constant in reward, which gives a value function plus a constant. In episodic task V_c is not a consta[nt] but depends on T, the timestep on which the episode ends

(i)

1) $V_\pi(s) = E_\pi[q_\pi(s,a)]$

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

(j) According to the optimal policy, we have

$V_*(s) = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

$= 10 + 0.9 \times 0 + 0.9^2 \times 0 + 0.9^3 \times 0 + 0.9^4 \times 0 + 0.9^5 \times 10 + \cdots$

$= (1 + 0.9^5 + 0.9^{10} + \cdots + (0.9^5)^n) \times 10$

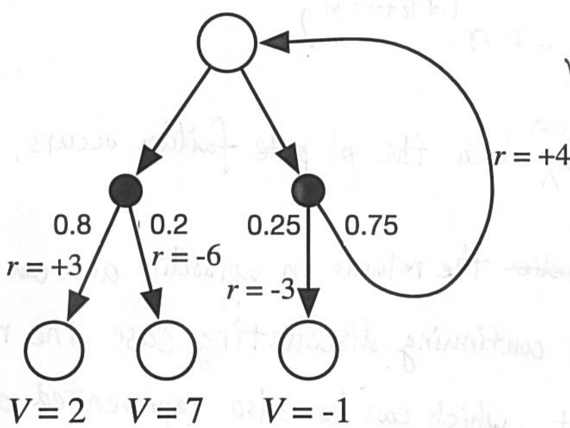
$= \frac{1}{1-0.9^5} \times 10$

$\simeq 24.4$

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.



$v_\pi =$ 4.15

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) \cdot (r + \gamma V(s'))$$

$$= \frac{1}{2} \times 0.8 \times [3 + 0.8 \times 2] + \frac{1}{2} \times 0.2 \times [-6 + 0.8 \times 7] + \frac{1}{2} \times 0.25 \times [-3 + 0.8 \times (-1)] + \frac{1}{2} \times 0.75 \times [4 + 0.8 V_\pi(s)]$$

$$= 2.905 + 0.3 V_\pi(s)$$

$$V_\pi(s) = 4.15$$

$v_* =$ 5.125

$$V_* = \max_a \sum_{s',r} p(s',r|s,a)(r + \gamma V(s'))$$

$$= \max_a q(s,a)$$

$$q(s, left) = \sum_{s'r} p(s',r|s,a)(r + \gamma V(s')) = 3.6$$

$$q(s, right) = 5.125$$

$$\therefore V_* = \max_a q(s,a) = 5.125$$

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

i) The return at each time would be:

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k = -\sum_{k=t+1}^{T} \gamma^{k-t-1} = -(1+\gamma+\gamma^2+\cdots+\gamma^T)$$

where T is the timestep that the episode ends.

ii) For the continuing, discounted task the return would be:

$$G_t = -(\gamma^{T_1} + \gamma^{T_1+T_2} + \cdots + \gamma^{T_1+T_2+\cdots+T_n})$$

where T_i is the time-step on which the ~~pl~~ ~~pole~~ failure occurs.

The difference between ~~episodic~~ the returns in episodic and continuing, discounted case is, in continuing discounting case the return will converge to a constant, which can be also represented as "reward rate" for each time-step; but for episodic case this is not guaranteed. And for episodic case the Return will have higher variance than in the continuing case, for it recieves "-1" ~~at~~ at every time-step rather than zero.