Last name: Lu First name: Brian (Boyuan) SID#: 1371773

Collaborators:

CMPUT 366/609 Assignment 2: Markov Decision Processes 1

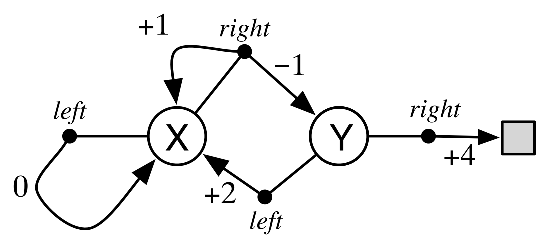
Due: Thursday Sept 28, 11:59pm by gradescope

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:

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

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

$G_0 =$

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

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

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

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

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

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

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).

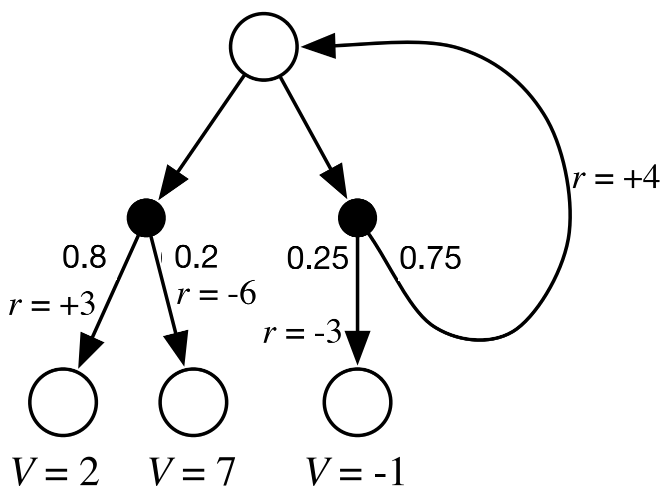
(j) [8 points, 4 for symbolic form, 4 points for numeric answer] Figure 3.6 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.7) to express this value symbolically, and then to compute it to three decimal places.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

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.

CMPUT 366

Brian (Boyuan) Lu

Q1

(a).

state	actions	reward	transition
X	left	+0	X
X	left	+0	X
X	left	+0	Y
X	left	+0	Y
⋮	⋮	⋮	⋮

(b)

state	action	reward	transition
X	right	+1	X
X	right	+1	X
X	right	+1	X
Y	right	+4	Special absorbing state (SOS)
SOS	—	0	SOS
SOS	—	0	SOS
⋮	⋮	⋮	⋮

(c) $G_0 = \sum_{k=0}^{T} \gamma^k R_{k+1}$ $\gamma = 0.5$

$= 1 + 0.5 \times 1 + 0.5^2 \times 1 + 0.5^3 \times 4 + 0 \cdots$

$= 1 + 0.5 + 0.25 + 0.5$

$= 2.25$

(d). $\gamma = 0.5$. value of state ?

$v_{\pi_1}(Y) = E_{\pi_1}[G_t \mid S_t = Y]$

$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} [S_t = s]$ for all $s \in S$,

For state Y in policy π_1 it always perform "right" action.

Thus, R_1 is +4 and all other reward is 0 since it goes to a Special absorbing state $V_{\pi_1}(Y) = 4$

(e) $\gamma = 0.5$.

$q_{\pi_1}(X, left) = E_{\pi_1}[G_t \mid S_t = X, A_t = left]$

~~reward of~~ $\pi_1(X)$ is always left, reward of it is always 0.

Thus. $q_{\pi_1}(X, left) = E_{\pi_1}[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = X, A_t = left]$

$= E_{\pi_1}[\sum_{k=0}^{\infty} 0.5^k \cdot 0] = 0$

(f). $\gamma = 0.5$, $V_{\pi_2}(X) = ?$

Using Bellman equation

$$V_{\pi_2}(X) = E_{\pi_2}[G_t \mid S_t = s]$$
$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s,a)[r + \gamma E_\pi[G_{t+1}|S_{t+1}=s']]$$
$$= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')] \text{ for all } s \in S$$

for action on state X at policy π_2 it is always "right"

Thus $\sum_a \pi(a|s) = 1$.

There are two ~~two~~ possible transition of right action on state X.

Thus

$$\sum_{s',r} p(s',r|X, a_{right})[r + \gamma V_{\pi_2}(s')]$$
$$= 0.75 \times [1 + 0.5 V_{\pi_2}(s')] + 0.25 \times [-1 + 0.5 \underbrace{V_{\pi_2}(s')}_{4}]$$
$$= 0.75 \times [1 + 0.5 V_{\pi_2}(s')] + 0.25 \times (-1+2).$$

$\therefore V_{\pi_2}(X) = 1 \times (0.75 \times (1 + 0.5 V_{\pi_2}(s')) + 0.25)$

$V_{\pi_2}(X) = 0.75 + 0.25 + 0.375 V_{\pi_2}(s')$

$V_{\pi_2}(X) = 1 + 0.375 V_{\pi_2}(s')$

If goes to infinit time step

$$V_{\pi_2}(X) = 1 + 0.375 \times (0.375 V_{\pi_2}(s'') + 1).$$
$$= 1 + 0.375 \times 0.375 V_{\pi_2}(s'') + 0.375$$

This term goes to very small and neglectable when $k \to \infty$

Thus $V_{\pi_2}(X) = \sum_{k=0}^{\infty} 1 \times 0.375^k$

$$= \frac{1}{1-0.375} = \frac{1}{0.625} = 1.6$$

Question 2.

(a) Exercise 3.1

1. Self-driving Car. the sensor will detect nearby environment and transfer to readable data, these data are evaluated by the Environment. then input the environment reflection to Agent. Agent performing estimation upon the updated data sets. The state evaluate the car states and environment state, these are also input to the Agent.

2. Map GPS. It has a environment to check the car's position and traffic condition, then feed to agent, agent evaluate the map condition and car position to provide the best path to the GPS.

3. Black-Jack AI, the AI playing games with other players or AIs for many times. Environment monitor the strategies when playing the cards. agent perform the estimate on each play and deep search the next few possible plays and evaluate on each path of them. Last, come out with the optimal play

(b) Exercise 3.7

- what is going wrong?
  - It gives a reward on each successful episode, which means doesn't matter how many time-step this episode take. if it reaches the goal. it will have a +1 reward. This is a problem because it doesn't count for the number of time-steps. Thus it can't improve the running. It should aim to find episode with minimum time-steps

- How to effectively communicate to the agent?
  - This case should introduce the discounting with γ^k, where k is the time-steps each episode takes. Thus, when time step is small. as γ is smaller than 1. when ~~few~~ time-steps is large reward $\times \gamma^k = 1 \times \gamma^k$ become smaller as k become larger. Thus, when time-steps is small. can gives a larger discounting reward. In this, it can be improved by seeking larger reward (smaller time-steps).

(c) Exercise 3.8

- G_5 is the terminating step, thus $G_5 = R_5 = 2$

- $G_4 = R_{t+1} + \gamma G_{t+1}$
  $= R_5 + 0.5 \times \cancel{2} G_5$
  $= 2 + 0.5 \times 2$
  $= 3$
- $G_3 = R_4 + \gamma \times G_4$
  $= 3 + 0.5 \times 3$
  $= 4.5$
- $G_2 = R_3 + \gamma \times G_3$
  $= 6 + 0.5 \times 4.5$
  $= 8.25$
- $G_1 = R_2 + \gamma \times G_2$
  $= 2 + 0.5 \times 8.25$
  $= 6.125$

$G_0 = R_1 + \gamma \times G_1$
$G_0 = -1 + 0.5 \times 6.125$
$G_0 = 2.0625$

(d) Exercise 3.9.

$$G_1 = \sum_{k=2}^{\infty} \gamma^k \times 7 = \frac{7}{1-\gamma} = \frac{7}{1-0.9} = \frac{7}{0.1} = \underline{70}.$$

$$\begin{aligned} G_0 &= R_1 + \gamma \times G_1 \\ &= 2 + 0.9 \times 70 \\ &= 2 + 63 \\ &= \underline{65} \end{aligned}$$

(e). Exercise 3.11

Assume the point below the center is point C.

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

There are four possible actions for point C, South, North, East, West
and each has a $\pi(a|s) = 0.25$ (S) (N) (E) (W)

$$\begin{aligned} \text{Thus. } V_\pi(c) = \; & 0.25 \times \sum_{s',r} p(s',r|s,\text{South})[r + \gamma V_\pi(s')] \\ & + 0.25 \times \sum_{s',r} p(s',r|s,\text{North})[r + \gamma V_\pi(s')] \\ & + 0.25 \times \sum_{s',r} p(s',r|s,\text{East})[r + \gamma V_\pi(s')] \\ & + 0.25 \times \sum_{s',r} p(s',r|s.\text{West})[r + \gamma V_\pi(s')] \end{aligned}$$

For $p(s',r|s,\text{south})$, $p(s',r|s,\text{North})$, $p(s',r|s,\text{East})$, $p(s',r|s,\text{west})$
each of them is 1 since state transition from point C to a nearby cell is absolute as long as the corresponding action is taken

$$\begin{aligned} \text{Thus } V_\pi(c) = \; & 0.25 \times [r_{South} + \gamma V_\pi(s'_{South})] \\ & + 0.25 \times [r_{North} + \gamma V_\pi(s'_{North})] \\ & + 0.25 \times [r_{east} + \gamma V_\pi(s'_{east})] \\ & + 0.25 \times [r_{west} + \gamma V_\pi(s'_{west})] \end{aligned}$$

reward is 0 for all 4 directions, $\gamma = 0.9$. $V_\pi(s'_{south}) = -1.2$
$V_\pi(s'_{North}) = 0.7$, $V_\pi(s'_{east}) = -0.6$, $V_\pi(s'_{west}) = -0.4$.

$$\begin{aligned} \Rightarrow V_\pi(C) &= 0.25 \times (0 + 0.9 \times 0.7) + 0.25 \times (0 + 0.9 \times -1.2) + 0.25 \times 0.9 \times -0.6 + 0.25 \times 0.9 \times -0.4 \\ &= 0.1575 - 0.27 - 0.135 - 0.09 = -0.3375 \approx \boxed{-0.34 \therefore \text{close to } -0.4} \end{aligned}$$

(f) Exercise 3.12:

$$q_\pi(s,a) = E[G_t \mid S_t = s, A_t = a]$$
$$= E[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a]$$

The expected value is getting from a fixed r plus the $\gamma \cdot G_{t+1}$, which is $\gamma \times$ all possible action values will be take in the future times each of their $\pi(a'|s')$.

Thus $\quad R_{t+1} + \gamma G_{t+1} = [r + \gamma \cdot \sum_{a'} \pi(a'|s') \cdot q_\pi(s', a')]$

And for different choices of states, each has a possible ratio of $p(s', r \mid s, a)$

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

(g) Exercise 3.13.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

If adding a constand c to all the reward gives

$$V_c(s) = R_{t+1} + c + \gamma(R_{t+2}+c) + \gamma^2(R_{t+3}+c) \dots$$

$$V_c(s) = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \dots + c + \gamma c + \gamma^2 c + \gamma^3 c + \dots$$

$$\therefore V_c(s) = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k \cdot c$$

$\sum_{k=0}^{\infty} \gamma^k \cdot c = \frac{c}{1-\gamma}$ $\quad\therefore$ every $V_c(s)$ $s \in S$, they are added with a constant $\frac{c}{1-\gamma}$; therefore it won't affect their relative values.

since c is a constant, γ is a constant Then V_c is a constant. If adding this constant to all states, it won't affact the relative valhes under any states and policy.

(h).

- Would it leave the task unchanged?

No. it will change the task as each episode will have a more different result compare to before

- why?

Because when adding a constant to each reward in the episode, each episode will have a different total reward in the end. Before, each episode will have a -1 reward if fail, 0 for success, thus all successful path will have the same reward. This can't improve the algorithm to find the optimal path. When adding a constant to each reward (negative constant), each episode will have different number of time-steps, therefore different number of rewards As a result, when add a negative constant to each reward ~~on~~ in a episode, will make episodes having different total rewards. Longer the path, more time steps, more rewards steps, and thus lower the total rewards in this episodes. Therefore, the optimal path will be the episode has largest ~~rew~~ total reward

- Example:
For example Episode 1 has 8 time-steps with a successful escape
Episode 2 has 10 time-steps with successful escape.

- If Not adding a constant ~~to~~ each reward:
Episode 1 and episode 2 will have equal total reward which is 0.
And can't identify which one is ~~better~~ better from their total rewards

- If adding a negative constant to each reward:
$C = -1$
Episode 1 ~~have~~ has total reward $= 8 \times -1 = -8$
Episode 2 has total reward $= 10 \times -1 = -10$
Episode 1 > Episode 2
Therefore picking episode 1 will give a higher return

(i) Exercise 3.15
Equation 1:
$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$
$$= E_\pi[\sum_a q_\pi(s,a) \mid S_t = s, A_t = a]$$

Equation 2:
$$V_\pi(s) = \sum_a \pi(a|s) \cdot q_\pi(s,a)$$
$$= \pi(a_1|s) \times q_\pi(s,a_1) + \pi(a_2|s) \times q_\pi(s,a_2) + \pi(a_3|s) \times q_\pi(s,a_3)$$

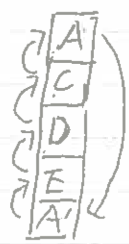
(j) Eq 3.7 : $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$ –

Eq 3.9 : $G_t = \sum_{k=0}^{\infty} a\gamma^k = \frac{a}{1-\gamma}$

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

- Reasoning: The optimal value of the best state of the gridworld is 24.4 since from the optimal policies grid. All other state cells tends to move towards state A.
  For state A, it always transfer to A' which is the second state cell of last row. And then A' intends to move up until reach state A, and come back to A' again. Thus, state transition of A to A' is a circle :

$$A \to A' \xrightarrow[\substack{\text{move up through}\\ \text{all other states}\\ \text{until reach state A}}]{} A \to A' \longrightarrow A \to A'$$



$\therefore$

$$V_*(A) = G_t = R_{t+1} + \gamma R_{t+2} + \gamma R_{t+3} + \dots$$

$$V_*(A) = V_*(A') + \gamma V_*(E) + \gamma^2 V_*(D) + \gamma^3 V_*(C) + \gamma^4 V_*(A) + \gamma^5 V_*(A') \dots$$

since state transition reward except $\underset{10}{\underline{A \to A'}}$ and $\underset{5}{\underline{B \to B'}}$ are 0,

Then has: $V_*(A) = \gamma^0 \times 10 + 0 + 0 + 0 + 0$
$+ \gamma^5 \times 10 + 0 + 0 + 0 + 0$
$+ \gamma^{10} \times 10 + 0 + 0 \dots$
$+ \gamma^{15} \times 10 \dots \infty$

Then $V_*(A) = \sum_{k=0}^{\infty} \gamma^{5k} \times 10$

$$= \sum_{k=0}^{\infty} \gamma^{5k} \cdot 10 = 10 \times (\gamma^5)^k = \frac{10}{1-0.9^5} = \frac{10}{0.40951} = 24.419.$$

$$\therefore \Rightarrow V_*(A) = -\frac{1}{\ln(0.9)} \times \frac{1}{5} \times 10 = \frac{2}{\ln(0.9)} = -(-18.982) = 18.982$$