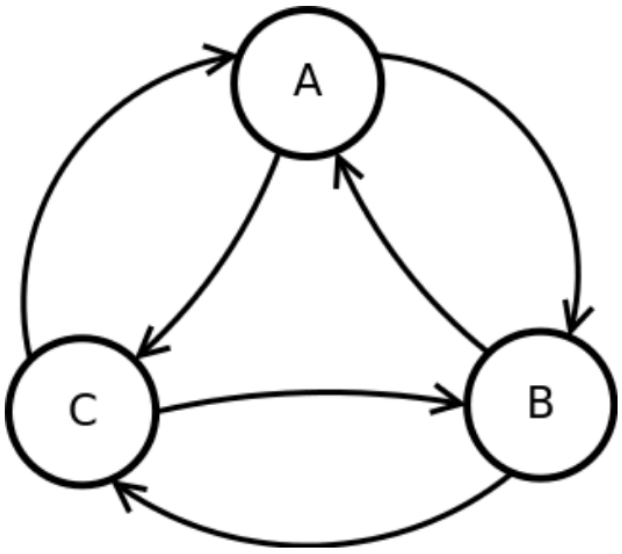
Written Assignment 3: Solution

Deadline: November 08th, 2023

Instruction: You may discuss these problems with classmates, but please complete the write-ups individually. Remember the collaboration guidelines set forth in class: you may meet to discuss problems with classmates, but you may not take any written notes (or electronic notes, or photos, etc.) away from the meeting. Your answers must be **typewritten**, except for figures or diagrams, which may be hand-drawn. Please submit your answers (pdf format only) on **Canvas**.

Q1. MDPs - Value Iteration (30 points)

Part 1 - Cycle. (15 points) Consider the following transition diagram, transition function and reward function for an MDP.



Discount Factor, $\gamma = 0.5$

s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	1.0	0.0
A	Counterclockwise	C	1.0	-2.0
B	Clockwise	A	0.4	-1.0
B	Clockwise	C	0.6	2.0
B	Counterclockwise	A	0.6	2.0
B	Counterclockwise	C	0.4	-1.0
C	Clockwise	A	0.6	2.0
C	Clockwise	B	0.4	2.0
C	Counterclockwise	A	0.4	2.0
C	Counterclockwise	B	0.6	0.0

P1.1. Suppose that after iteration k of value iteration, we obtain the following values for V_k:

$V_k(A)$	$V_k(B)$	$V_k(C)$
0.400	1.400	2.160

Provide the value of $V_{k+1}(A)$, $V_{k+1}(B)$, and $V_{k+1}(C)$.

Answer. Note that: $V_{k+1}(s) = \max_a Q_{k+1}(s, a)$

$Q_{k+1}(A, \texttt{clockwise}) = R(A, \texttt{clockwise}, B) + \gamma V_k(B) = 0.0 + 0.5 \times 1.4 = 0.7$
$Q_{k+1}(A, \texttt{counterclockwise}) = R(A, \texttt{counterclockwise}, C) + \gamma V_k(C) = -2.0 + 0.5 \times 2.16 = -0.92$
$V_{k+1}(A) = \max(Q_{k+1}(A, \texttt{clockwise}), Q_{k+1}(A, \texttt{counterclockwise})) = 0.7$

Similarly, we have:

$Q_{k+1}(B, \texttt{clockwise}) = 0.4 \times (-1.0 + 0.5 \times 0.4) + 0.6 \times (2.0 + 0.5 \times 2.16) = 1.528$
$Q_{k+1}(B, \texttt{counterclockwise}) = 0.6 \times (2.0 + 0.5 \times 0.4) + 0.4 \times (-1.0 + 0.5 \times 2.16) = 1.352$
$V_{k+1}(B) = \max(1.528, 1.352) = 1.528$
$Q_{k+1}(C, \texttt{clockwise}) = 0.6 \times (2.0 + 0.5 \times 0.4) + 0.4 \times (2.0 + 0.5 \times 1.4) = 2.4$
$Q_{k+1}(C, \texttt{counterclockwise}) = 0.4 \times (2.0 + 0.5 \times 0.4) + 0.6 \times (0.0 + 0.5 \times 1.4) = 1.3$
$V_{k+1}(C) = 2.4$

P1.2. Suppose that we ran value iteration to completion and found the following value function, V^*. What are the optimal actions from states A, B, and C, respectively?

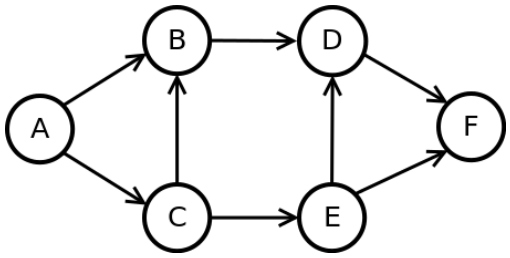
$V^*(A)$	$V^*(B)$	$V^*(C)$
0.881	1.761	2.616

Answer. Similar to **P1.1**, we have:

$Q^*(A, \texttt{clockwise}) = 0.0 + 0.5 \times 1.761 = 0.8805$
$Q^*(A, \texttt{counterclockwise}) = -2.0 + 0.5 \times 2.616 = -0.692$
Therefore, $\pi^*(A) = \texttt{clockwise}$
$Q^*(B, \texttt{clockwise}) = 0.4 \times (-1.0 + 0.5 \times 0.881) + 0.6 \times (2.0 + 0.5 \times 2.616) = 1.761$
$Q^*(B, \texttt{counterclockwise}) = 0.6 \times (2.0 + 0.5 \times 0.881) + 0.4 \times (-1.0 + 0.5 \times 2.616) = 1.5875$
Therefore, $\pi^*(B) = \texttt{clockwise}$
$Q^*(C, \texttt{clockwise}) = 0.6 \times (2.0 + 0.5 \times 0.881) + 0.4 \times (2.0 + 0.5 \times 1.761) = 2.6165$
$Q^*(C, \texttt{counterclockwise}) = 0.4 \times (2.0 + 0.5 \times 0.881) + 0.6 \times (0.0 + 0.5 \times 1.761) = 1.5045$
Therefore, $\pi^*(C) = \texttt{clockwise}$

Part 2 - Convergence. (15 points) We will consider a simple MDP that has six states, A, B, C, D, E, and F. Each state has a single action, `go`. An arrow from a state x to a state y indicates that it is possible to transition from state x to next state y when `go` is taken. If there are multiple arrows leaving a state x, transitioning to each of the next states is equally likely. The state F has no outgoing arrows: once you arrive in F, you stay in F for all future times. The reward is one for

all transitions, with one exception: staying in F gets a reward of zero. Assume a discount factor = 0.5. We assume that we initialize the value of each state to 0. (Note: you should not need to explicitly run value iteration to solve this problem.)



P2.1. After how many iterations of value iteration will the value for state E have become exactly equal to the true optimum? (Enter inf if the values will never become equal to the true optimal but only converge to the true optimal.)

Answer. 2

P2.2. How many iterations of value iteration will it take for the values of all states to converge to the true optimal values? (Enter inf if the values will never become equal to the true optimal but only converge to the true optimal.)

Answer. 4

Explanation. Because there are no moves from state F, we have the optimal value of F upon initializing. Since all the rewards are earned from transitions, finding the optimal value of a state amounts to finding the longest path from that state to F. For example, state D, whose longest path to F is only length 1, will find its optimal value after only one iteration.

$$V^*(D) = V_1(D) = R(D, go, F) + \gamma V^*(F) = 1.$$

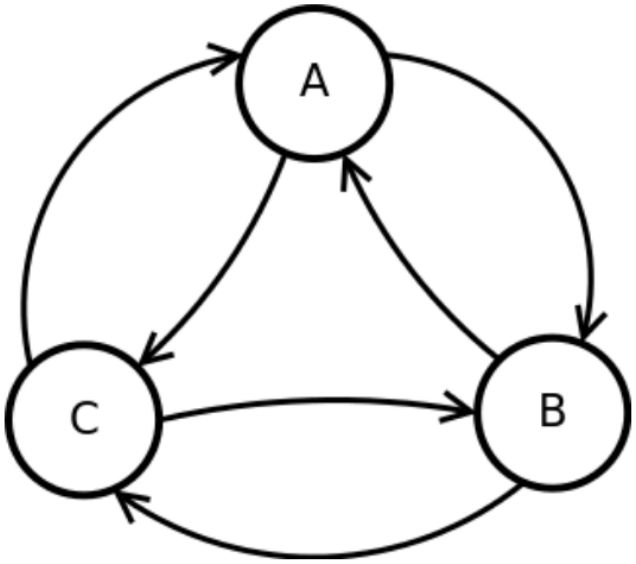
Similarly, the state A will find its optimal value after four iterations, because it will find out about its length 4 path to F after four iterations. Because A's length 4 path is the longest of the graph, it will take four iterations for all states to converge to their optimal values.

Q2. MDPs - Policy Iteration (20 points)

Consider the following transition diagram, transition function and reward function for an MDP.

Q2.1. (10 points) Suppose we are doing policy evaluation, by following the policy given by the left-hand side table below. Our current estimates (at the end of some iteration of policy evaluation) of the value of states when following the current policy is given in the right-hand side table.

Provide the value of $V^{\pi}_{k+1}(A)$, $V^{\pi}_{k+1}(B)$, and $V^{\pi}_{k+1}(C)$



Discount Factor, $\gamma = 0.5$

s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	1.0	0.0
A	Counterclockwise	C	1.0	-2.0
B	Clockwise	A	0.4	-1.0
B	Clockwise	C	0.6	2.0
B	Counterclockwise	A	0.6	2.0
B	Counterclockwise	C	0.4	-1.0
C	Clockwise	A	0.6	2.0
C	Clockwise	B	0.4	2.0
C	Counterclockwise	A	0.4	2.0
C	Counterclockwise	B	0.6	0.0

A	B	C
Counterclockwise	Counterclockwise	Counterclockwise

$V_k^\pi(A)$	$V_k^\pi(B)$	$V_k^\pi(C)$
0.000	-0.840	-1.080

Answer. Note that: $V_{k+1}^\pi(s) = \sum_{s'} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V_k^\pi(s')]$. Therefore, we have:

$$V_{k+1}^\pi(A) = 1.0 \times (R(A, \texttt{counterclockwise}, C) + \gamma V_k^\pi(C)) = -2.0 + 0.5 \times (-1.08) = -2.54$$
$$V_{k+1}^\pi(B) = 0.6 \times (2.0 + 0.5 \times 0.0) + 0.4 \times (-1.0 + 0.5 \times (-1.08)) = 0.584$$
$$V_{k+1}^\pi(C) = 0.4 \times (2.0 + 0.5 \times 0.0) + 0.6 \times (0.0 + 0.5 \times (-0.84)) = 0.548$$

Q2.2. (10 points) Suppose that policy evaluation converges to the following value function, V_∞^π. Provide the values of $Q_\infty^\pi(A, clockwise)$ and $Q_\infty^\pi(A, counterclockwise)$. What is the updated action for A?

$V_\infty^\pi(A)$	$V_\infty^\pi(B)$	$V_\infty^\pi(C)$
-0.203	-1.114	-1.266

Answer. $Q_\infty^\pi(A, clockwise) = -0.557$ and $Q_\infty^\pi(A, counterclockwise) = -2.633$.
The updated action for A is clockwise.

Explanation.

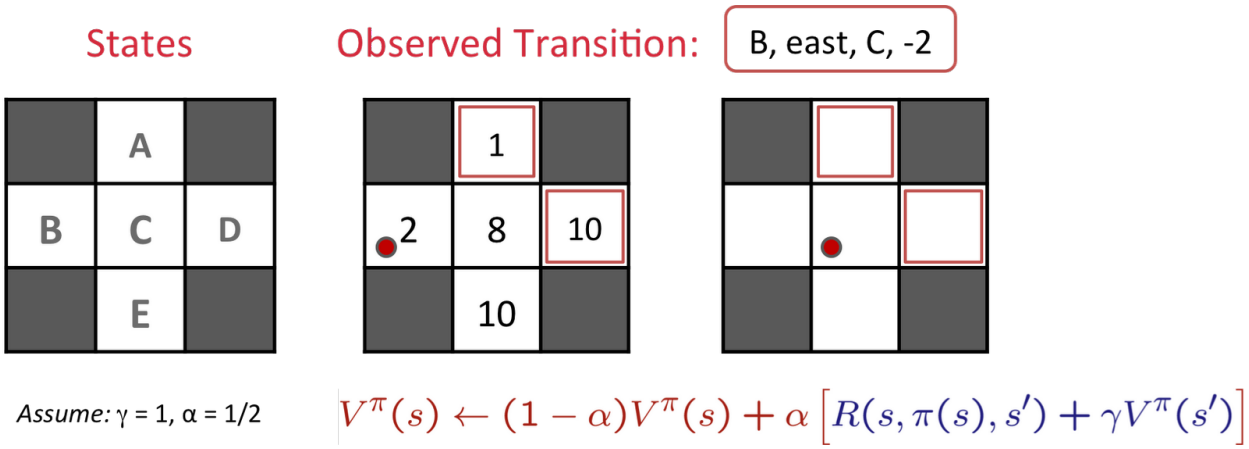
$$Q^{\pi}_{\infty}(A,clockwise) = T(A, clockwise, B)[R(A, clockwise, B) + \gamma V^{\pi}_{\infty}(B)] + T(A, clockwise, C)[R(A, clockwise, C) + \gamma V^{\pi}_{\infty}(C)] = -0.557$$

$$Q^{\pi}_{\infty}(A,counterclockwise) = T(A, counterclockwise, B)[R(A, counterclockwise, B) + \gamma V^{\pi}_{\infty}(B)] + T(A, counterclockwise, C)[R(A, counterclockwise, C) + \gamma V^{\pi}_{\infty}(C)] = -2.633$$

The updated action for state A will be the action that results in the higher Q^{π}_{∞}

Q3. Temporal Difference Learning (10 points)

Consider the gridworld shown below. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function V^{π} for each state. A transition is observed, that takes the agent from state B through taking action east into state C, and the agent receives a reward of -2. Assuming $\gamma = 1, \alpha = 0.5$, what are the value estimates of $\hat{V}^{\pi}(A), \hat{V}^{\pi}(B), \hat{V}^{\pi}(C), \hat{V}^{\pi}(D)$, and $\hat{V}^{\pi}(E)$ after the TD learning update? (note: the value will change for one of the states only)



Answer.

- $\hat{V}^{\pi}(A) = 1$
- $\hat{V}^{\pi}(B) = 4$
- $\hat{V}^{\pi}(C) = 8$
- $\hat{V}^{\pi}(D) = 10$
- $\hat{V}^{\pi}(E) = 10$

The only value that gets updated is $\hat{V}^{\pi}(B)$, because the only transition observed starts in state B.

$$\hat{V}^{\pi}(B) = 0.5 \times 2 + 0.5 \times (-2 + 8) = 4.$$

Q4. Active Reinforcement Learning (40 points)

Q4.1. Q-learning (20 points) Pacman is in an unknown MDP where there are four states [A, B, C, D] and two actions [Left, Right]. We are given the following samples generated from taking actions in the unknown MDP. For the following problems, assume $\gamma = 0.8$ and $\alpha = 0.75$.

We run Q-learning on the following samples:

s	a	s'	r
A	Left	B	2.0
B	Right	D	-1.0
D	Left	C	3.0
C	Left	A	-2.0
A	Right	D	1.0

What are the estimates for the Q-values $Q(C, Left)$ and $Q(A, Right)$ as obtained by Q-learning? All Q-values are initialized to 0. **Answer.** We obtain the following:

- First sample, update $Q(A, Left) = 0.25 \times 0 + 0.75 \times 2.0 = 1.5$
- Second sample, update $Q(B, Right) = 0.25 \times 0 + 0.75 \times (-1.0) = -0.75$
- Third sample, update $Q(D, Left) = 0.25 \times 0 + 0.75 \times 3.0 = 2.25$
- Fourth sample, update $Q(C, Left) = 0.25 \times 0 + 0.75 \times (-2.0 + 0.8 \times 1.5) = -0.6$
- Fifth sample, update $Q(A, Right) = 0.25 \times 0 + 0.75 \times (1.0 + 0.8 \times 2.25) = 2.1$

Q4.2. Approximate Q-learning (14 points) For this part, we will switch to a feature based representation. We will use two features:

- $f_1(s, a) = 1$.
- $f_2(s, a) = \begin{cases} -1 & \text{if } a = \text{Left} \\ 1 & \text{if } a = \text{Right} \end{cases}$

Starting from initial weights of 0, we are going to use the first two samples in the above table to update the weights:

1. What are the weights after the first update? (using the first sample)
   Answer. Since difference = 2.0, thus:
   - $w_1 = 0 + 0.75 \times 2.0 \times 1 = 1.5$
   - $w_2 = 0 + 0.75 \times 2.0 \times (-1) = -1.5$
2. What are the weights after the second update? (using the second sample)
   Answer. Since difference = -1.0, thus:
   - $w_1 = 1.5 + 0.75 \times (-1.0) \times 1 = 0.75$
   - $w_2 = -1.5 + 0.75 \times (-1.0) \times 1 = -2.25$

Q4.3. Exploration. (6 points) In Q-learning, we can modify the reward original reward function $R(s, a, s')$ to visit more states and choose new actions. $N(s, a)$ refers to the number of times that you have visited state s and taken action a in your samples.

1. Which of the following rewards would encourage the agent to visit unseen states and actions **(Yes/No)**?
   - $R(s, a, s') + \frac{1}{1+N(s,a)}$ **(Yes)**
   - $R(s, a, s') + N^2(s, a)$ **(No)**
   - $-\exp(N(s, a) + 1)$ **(Yes)**
2. Which of the following modified rewards will eventually converge to the optimal policy with respect to the original reward function $R(s, a, s')$ **(Yes/No)**?
   - $R(s, a, s') + \frac{1}{1+N(s,a)}$ **(Yes)**
   - $R(s, a, s') + N^2(s, a)$ **(No)**
   - $-\exp(N(s, a) + 1)$ **(No)**

Q5. Properties of MDPs (Grads Only) (10 points)

Consider an MDP $(\mathbf{S}, \mathbf{A}, T, R)$ with a finite state space $\mathbf{S}$, finite action space $\mathbf{A}$, the transition function $T(s, a, s')$, a reward function $R(s, a, s')$, and a discount factor $\gamma \in (0, 1)$. The reward $R(s, a, s') \geq 1$, for all (s, a, s'). Denote by $V_k(s)$ the value of state s after k iterations regarding the value iteration method and $V^*(s)$ the optimal value of state s.

Initially, $V_0(s) = 1$ for all s. Prove that $V^*(s) \geq V_k(s)$, for all k.
Hint: First prove $V_{k+1}(s) \geq V_k(s)$, for all k using induction.

Answer. When $k = 0$, we have: $V_1(s) = \max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V_0(s')] \geq 1 = V_0(s)$. Suppose that $V_{k'+1}(s) \geq V_{k'}(s)$ for all $k' < k$, we are going to show that $V_{k+1}(s) \geq V_k(s)$. Indeed,

$$V_{k+1}(s) = \max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V_k(s')]$$
$$\geq \max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V_{k-1}(s')] = V_k(s)$$

Therefore, $V_{k+1}(s) \geq V_k(s)$ for all k. Since value iteration converges to the optimal $V^*(s)$, we obtain $V^*(s) \geq V_k(s)$ for all k.