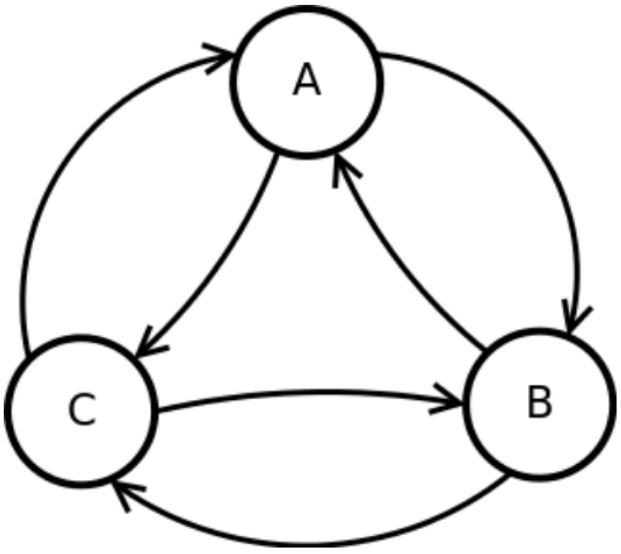
Written Assignment 3

Deadline: February 20th, 2020

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. 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?

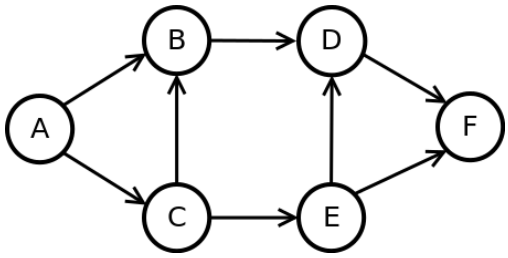
$V^*(A)$	$V^*(B)$	$V^*(C)$
0.881	1.761	2.616

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. 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

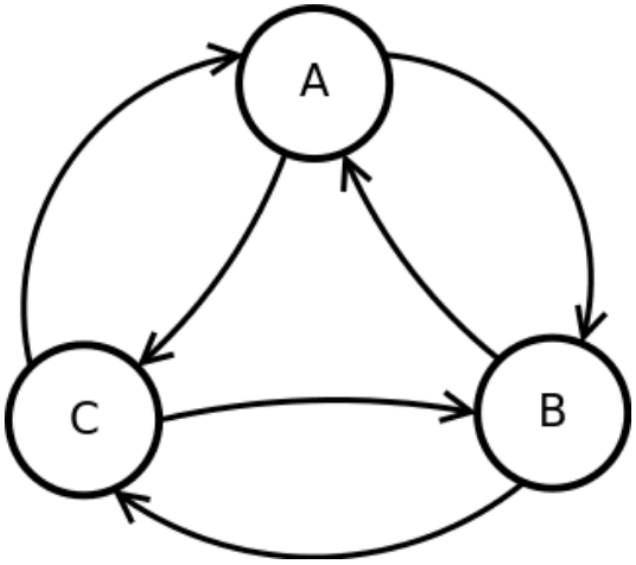
Explanation. Because there are no moves from state F, we have the optimal value of F upon initializing. Since all the rewards are earned from transitions, finding the optimal value of a state amounts to finding the longest path from that state to F. For example, state D, whose longest path to F is only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



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

Q2.1. 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 $V^{\pi}_{k+1}(C)$

A	B	C
Counterclockwise	Counterclockwise	Counterclockwise

$V^{\pi}_{k}(A)$	$V^{\pi}_{k}(B)$	$V^{\pi}_{k}(C)$
0.000	-0.840	-1.080

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

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

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

Answer. $Q^{\pi}_{\infty}(A, clockwise) = -.1722$ and $Q^{\pi}_{\infty}(A, counterclockwise) = -.2026$.

The updated action for A is clockwise.

$V^{\pi}_{\infty}(A)$	$V^{\pi}_{\infty}(B)$	$V^{\pi}_{\infty}(C)$
-0.203	-1.114	-1.266

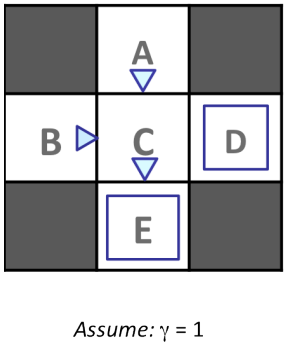
Explanation.

$$Q^{\pi}_{\infty}(A,clockwise) = T(A, clockwise, B)[R(A, clockwise, B) + \gamma V^{\pi}_{\infty}(B)] + T(A, clockwise, C)[R(A, clockwise, C) + \gamma V^{\pi}_{\infty}(C)] = -.1722$$

$$Q^{\pi}_{\infty}(A,counterclockwise) = T(A, counterclockwise, B)[R(A, counterclockwise, B) + \gamma V^{\pi}_{\infty}(B)] + T(A, counterclockwise, C)[R(A, counterclockwise, C) + \gamma V^{\pi}_{\infty}(C)] = -.2026$$

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

Q3. Model-based Reinforcement Learning (10 points)



Observed Episodes (Training)

Episode 1
A, south, C, -1
C, south, E, -1
E, exit, x, +10

Episode 2
B, east, C, -1
C, south, D, -1
D, exit, x, -10

Episode 3
B, east, C, -1
C, south, E, -1
E, exit, x, +10

Episode 4
A, south, C, -1
C, south, E, -1
E, exit, x, +10

What model would be learned from the above observed episodes (transition/reward functions)?

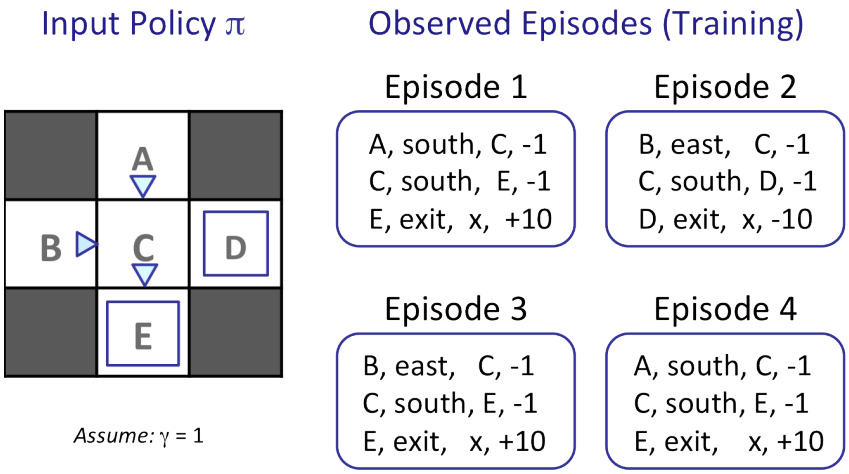
Answer. $T(A, south, C) = 1$, $T(C, south, E) = 0.75$, $T(B, east, C) = 1$, $T(C, south, D) = 0.25$. In addition, $R(A, south, C) = -1$, $R(C, south, E) = -1$, $R(B, east, C) = -1$, $R(C, south, D) = -1$

Q4. RL - Direct Evaluation (10 points)

What are the estimates for $\hat{V}^{\pi}(A)$, $\hat{V}^{\pi}(B)$, $\hat{V}^{\pi}(C)$, $\hat{V}^{\pi}(D)$, $\hat{V}^{\pi}(E)$ as obtained by direct evaluation?

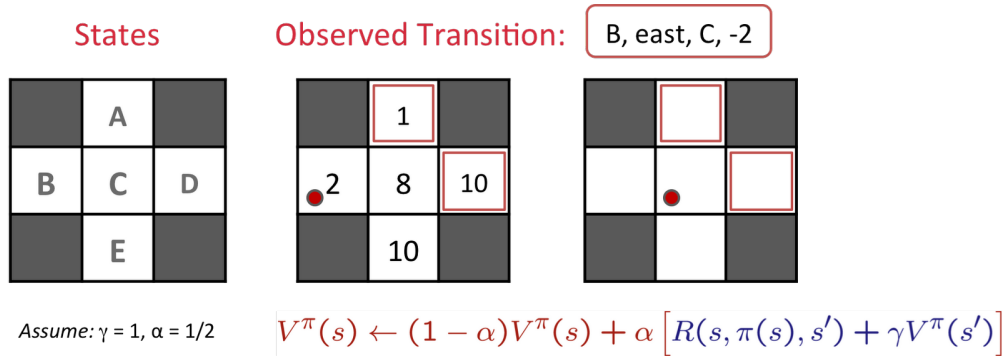
Answer. We have:

$$\hat{V}^{\pi}(A) = 8, \hat{V}^{\pi}(B) = -2, \hat{V}^{\pi}(C) = 4, \hat{V}^{\pi}(D) = -10, \hat{V}^{\pi}(E) = 10$$



Q5. RL - Temporal Difference Learning (6 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?



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.

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

Q6. Model-free Reinforcement Learning (12 points)

Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given with samples of what an agent actually experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, instead of first estimating the transition and reward functions, we will directly estimate the Q function using Q-learning. Assume, the discount factor, γ is 0.75 and the step size for Q-learning, α is 0.75.

Our current Q function, $Q(s, a)$, is shown in the left figure. The agent encounters the samples shown in the right figure:

	A	B	C
Clockwise	1.501	-0.451	2.73
Counterclockwise	3.153	-6.055	2.133

s	a	s'	r
A	Counterclockwise	C	8.0
C	Counterclockwise	A	0.0

Provide the Q-values for all pairs of (state, action) after both samples have been accounted for.

Answer.

- $Q(A, clockwise) = 1.501$
- $Q(A, counterclockwise) = 8.323875$
- $Q(B, clockwise) = -.451$
- $Q(B, counterclockwise) = -6.055$
- $Q(C, clockwise) = 2.73$
- $Q(C, counterclockwise) = 5.2154296875$

For each (s, a, s', r) transition sample, you update the Q value function as follows:

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(R(s, a, s') + \gamma \max_{a'} Q(s', a'))$$

First, we update: $Q(A, counterclockwise) = .25 \times 3.153 + .75 \times (8 + .75 \times 2.73) = 8.323875$
Then we update: $Q(C, counterclockwise) = .25 \times 2.133 + .75 \times (0 + .75 \times 8.323875) = 5.2154296875$.
Because there are only two samples, the other four values stay the same.

Q7. RL - Feature-based Representation (12 points)

Consider the following feature based representation of the Q-function: $Q(s, a) = w_1 f_1(s, a) + w_2 f_2(s, a)$ with:

- $f_1(s, a) = 1/$ (Manhattan distance to nearest dot after having executed action a in state s)
- $f_2(s, a) =$(Manhattan distance to nearest ghost after having executed action a in state s)

Q7.1. Assume $w_1 = 2$ and $w_2 = 8$. Assume that the red and blue ghosts are both sitting on top of a dot. Provide the values of $Q(s, west)$ and $Q(s, south)$.

Based on this approximate Q-function, which action would be chosen?

Answer.

- $Q(s, west) = 2 \times 1 + 8 \times 3 = 26$
- $Q(s, south) = 2 \times 1 + 8 \times 1 = 10$
- The chosen action is west since $26 > 10$.

Q7.2. Assume Pac-Man moves West. This results in the state s' shown below. Pac-Man receives reward 9 (10 for eating a dot and -1 living penalty).

Provide the values of $Q(s', west)$ and $Q(s', east)$. What is the sample value (assuming $\gamma = 0.8$)?

Answer.

- $Q(s', west) = 2 \times 1 + 8 \times 1 = 10$
- $Q(s', east) = 2 \times 1 + 8 \times 1 = 10$
- $sample = [r + \gamma \times \max_{a'} Q(s', a')] = 9 + 0.8 \times 10 = 17$

Q7.3. Now provide the update to the weights. Let $\alpha = 0.75$.

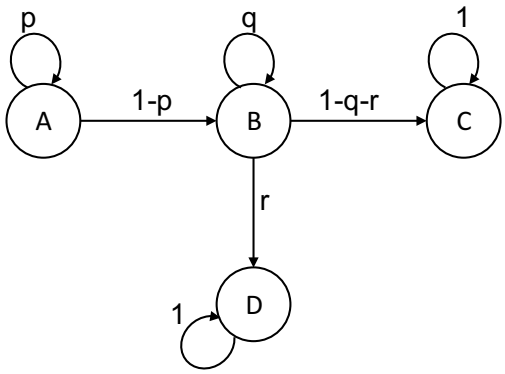
Answer. $w_1 = -4.75$ and $w_2 = -12.25$

Explanation. difference $= [r + \gamma \max_{a'} Q(s', a')] - Q(s, a) = 17 - 26 = -9$. Therefore,

$$w_1 = w_1 + \alpha(\text{difference}) f_1(s, a) = 2 + .75 \times (-9) \times 1 = -4.75$$
$$w_2 = w_2 + \alpha(\text{difference}) f_2(s, a) = 8 + .75 \times (-9) \times 3 = -12.25$$

Q8. A Not So Random Walk (Grads Only) (20 points)

Pacman is trying to predict the position of a ghost, which he knows has the following transition graph:



Here, $0 < p < 1$, $0 < q < 1$, and $0 < r < 1$ are arbitrary probabilities. It is known that the ghost always starts in state A. For this problem, we consider time to begin at 0. For example, at time 0, the ghost is in A with probability 1, and at time 1, the ghost is in A with probability p or in B with probability $1 - p$.

(a) What is the probability that the ghost is in A at time n?
For the ghost to be in A at time n, it must have stayed in A for n steps, which occurs with probability p^n

(b) What is the probability that the ghost first reaches B at time n?
For the ghost to first reach B at time n, it must have stayed in A for $n - 1$ steps, then transitioned to B. This occurs with probability $p^{n-1}(1 - p)$.

(c) What is the probability that the ghost is in B at time n?
For the ghost to be in B at time n, it must have first reached B at time i for some $1 \leq i \leq n$, then stayed there for $n - i$ steps. Summing over all values of i gives:

$$\sum_{i=1}^{n} p^{i-1}(1-p)q^{n-i} = \frac{(1-p)q^n}{p} \sum_{i=1}^{n} \left(\frac{p}{q}\right)^i = \frac{(1-p)q^n}{p} \times \frac{p}{q} \times \frac{1 - (\frac{p}{q})^n}{1 - \frac{p}{q}} = (1-p)\frac{q^n - p^n}{q - p}$$

(d) What is the probability that the ghost first reaches C at time n?
For the ghost to first reach C at time n, it must have been in B at time $n-1$ then transitioned to C. This occurs with probability: $(1-p)\frac{q^{n-1}-p^{n-1}}{q-p}(1-q-r)$

(e) What is the probability that the ghost is in C at time n?
For the ghost to be in C at time n, it must have first reached C at time i for some $2 \leq i \leq n$, then stayed there for $n-i$ steps. Summing over all i gives:

$$\sum_{i=2}^{n}(1-p)\frac{q^{i-1}-p^{i-1}}{q-p}(1-q-r) = \frac{(1-p)(1-q-r)}{q-p}\left(\frac{1-q^n}{1-q} - \frac{1-p^n}{1-p}\right)$$