Written Assignment 3

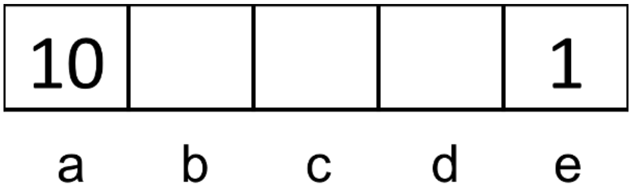
Deadline: November 13th, 2019

Instruction: You may discuss these problems with classmates, but please complete the write-ups individually. (This applies to BOTH undergraduates and graduate students.) 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. Value Iteration (20 points)

Consider the gridworld where Left and Right actions are successful 100% of the time.

Specifically, the available actions in each state are to move to the neighboring grid squares. From state a, there is also an exit action available, which results in going to the terminal state and collecting a reward of 10. Similarly, in state e, the reward for the exit action is 1. Exit actions are successful 100% of the time.

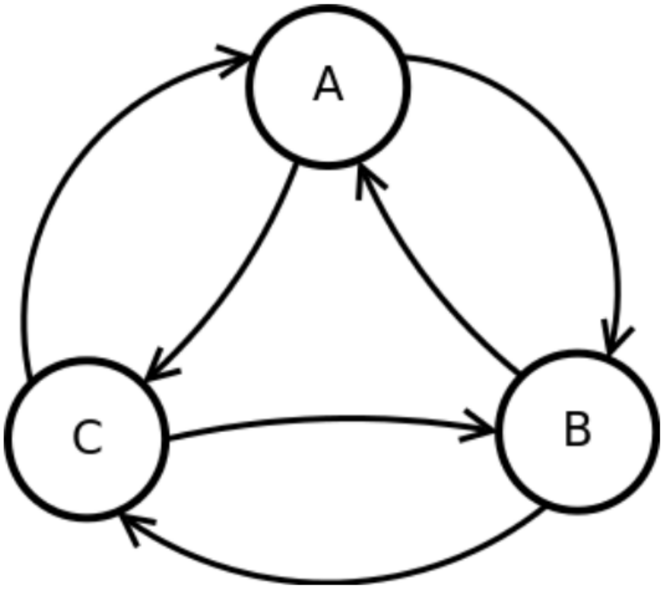


Q1.1. Let the discount factor $\gamma = 1.0$. Provide the value of $V_5(d)$.

Q1.2. Let the discount factor $\gamma = 0.2$. Provide the values of $V^*(a), V^*(b), V^*(c), V^*(d), V^*(e)$.

Q2. Policy Iteration: Cycle (20 points)

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



Discount Factor, γ = 0.5

s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	0.8	0.0
A	Clockwise	C	0.2	2.0
A	Counterclockwise	B	0.4	1.0
A	Counterclockwise	C	0.6	0.0
B	Clockwise	C	1.0	-1.0
B	Counterclockwise	A	0.6	-2.0
B	Counterclockwise	C	0.4	1.0
C	Clockwise	A	1.0	-2.0
C	Counterclockwise	A	0.2	0.0
C	Counterclockwise	B	0.8	-1.0

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 $V_{k+1}^{\pi}(A)$.

A	B	C
Counterclockwise	Counterclockwise	Counterclockwise

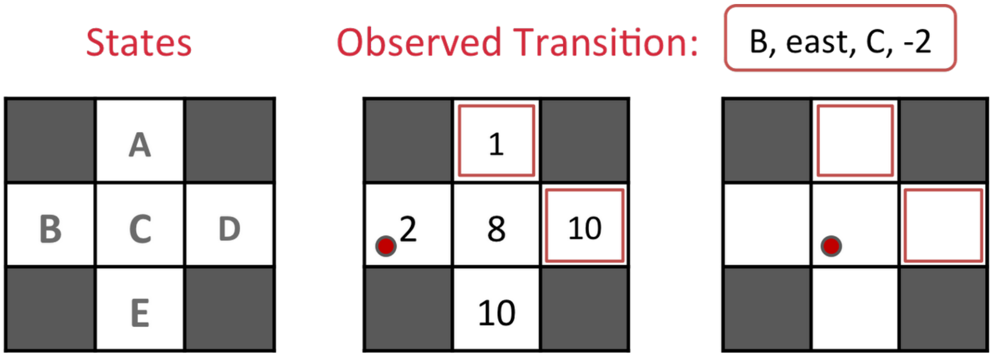
$V_k^{\pi}(A)$	$V_k^{\pi}(B)$	$V_k^{\pi}(C)$
0.000	-0.840	-1.080

Q2.2. 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

Q3. Temporal Difference Learning (20 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)



Assume: $\gamma = 1$, $\alpha = 1/2$

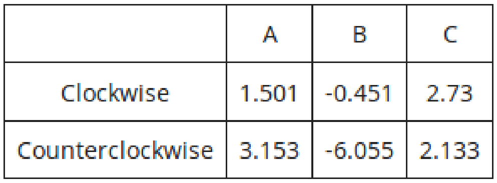
$$V^\pi(s) \leftarrow (1-\alpha)V^\pi(s) + \alpha\left[R(s,\pi(s),s') + \gamma V^\pi(s')\right]$$

Q4. Model-free Reinforcement Learning: Cycle (20 points)

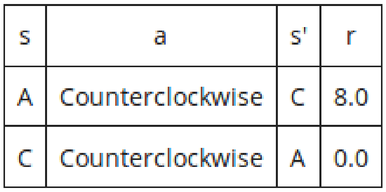
Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given with samples of what an agent actually experiences when it interacts with the environment (although, we do know 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 0.5 and the step size for Q-learning, α is 0.5.

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.

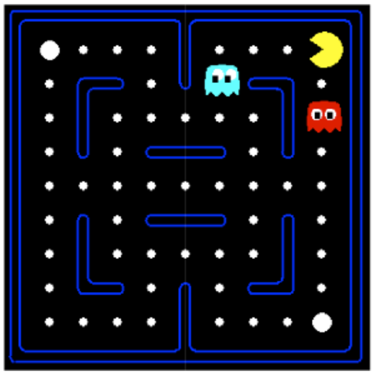
Q5. Feature-based Representation (20 points)

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:

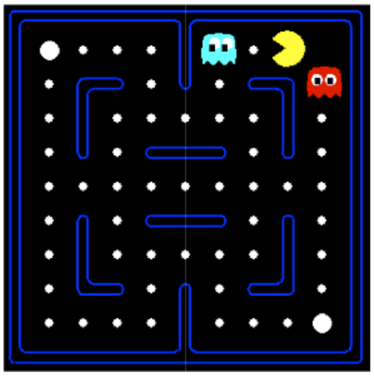
- $f_1(s,a) = 1/$ (Manhattan distance to nearest dot after having executed action a in state s)
- $f_2(s,a) =$(Manhattan distance to nearest ghost after having executed action a in state s)

Q5.1. Assume $w_1 = 1$ and $w_2 = 10$. 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?



Q5.2. 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 $\gamma = 1$)?

Q5.3. Now provide the update to the weights. Let $\alpha = 0.5$.

Q6. Properties of MDPs (Grads only) (20 points)

Consider two MDPs $(\mathbf{S}, \mathbf{A}, T, R)$ and $(\mathbf{S}, \mathbf{A}, T, \hat{R})$ with the same state space, action space, transition function, and the discount factor $\gamma \in (0, 1)$. The reward function $R(s, a, s') = m \times \hat{R}(s, a, s') + n$ for all (s, a, s') where (m, n) are constants. Denote by $\{V^*(s), \pi^*(s)\}$ and $\{\hat{V}^*(s), \hat{\pi}^*(s)\}$ the corresponding optimal values and policies of these two MDPs.

Let's assume that $(\hat{V}^*, \hat{\pi}^*)$ is known. Compute (V^*, π^*) based on $(\hat{V}^*, \hat{\pi}^*)$, γ, m, and n.
Hint: Use the Bellman equation.