CIS 471/571: Introduction to Artificial Intelligence, Fall 2020

Written Assignment 3: Solution
Deadline: Nov 10th, 2020

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.

Q1. MDPs - Value I

An agent lives in grid world G consisting of grid cells ve into the cells colored black. In this grid world, the agent can take actions to move t when it is not on a numbered square. When the agent is on a numbered square i exit to a terminal state (where it remains), collecting a reward equal to the number written on the square in the process. You decide to run value iteration for grid world G. The value function at

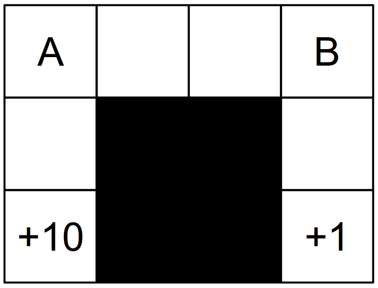


Figure 1: Grid world G

iteration k is $V_k(s)$. The initial value for all grid cells is 0 (that is, $V_0(s) = 0$ for all s). When answering questions about iteration k for $V_k(s)$, either answer with a finite integer or ∞. For all questions, the discount factor is $\gamma = 1$.

Q1.1. (18 points) Consider running value iteration in grid world G. Assume all legal movement actions will always succeed (and so the state transition function is deterministic).

1. What is the smallest iteration k for which $V_k(A) > 0$? For this smallest iteration k, what is the value of $V_k(A)$?

   Answer. $k = 3$, $V_k(A) = 10$

2. What is the smallest iteration k for which $V_k(B) > 0$? For this smallest iteration k, what is the value of $V_k(B)$?

   Answer. $k = 3$, $V_k(B) = 1$

3. What is the smallest iteration k for which $V_k(A) = V^*(A)$? What is the value of $V^*(A)$?

   Answer. $k = 3$, $V^*(A) = 10$

4. What is the smallest iteration k for which $V_k(B) = V^*(B)$? What is the value of $V^*(B)$?

   Answer. $k = 6$, $V^*(B) = 10$

Q1.2. (7 points) **probability** 0.8; with probability 0.2, the actio e iteration in grid world) $= V^*(A)$? What is the value of $V^*(A)$?

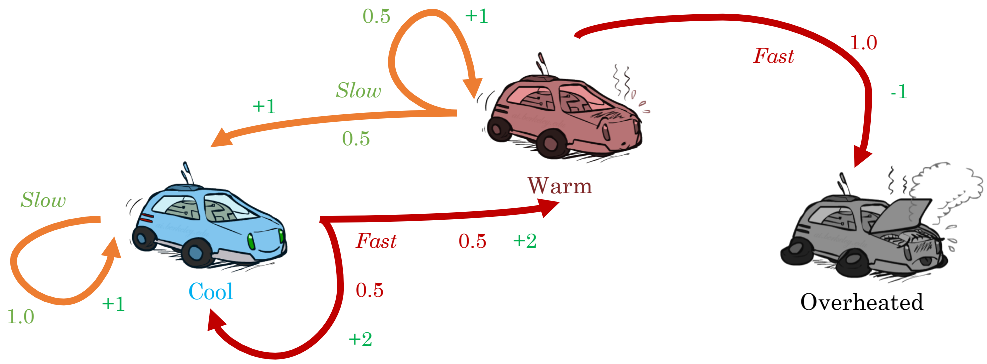
Answer. $k = \infty$, $V^*(A) = 10$. Because $\gamma = 1$ and the cu e optimal policy will move to the exit state with highest reward. This is guar succeed, so the optimal value of state A is 10. However, because the transition is non-deterministic, it's not guaranteed this reward can be collected in 3 steps. It could any number of steps from 3 through infinity, and the values will only have converged after infinitely many iterations.

Note. In practice, we can have a stopping condition, for example, if $|V_{k+1}(A) - V_k(A)| < \epsilon$ for some given ϵ, then we stop the iteration process. If your answer specify an ϵ and find the smallest k according to this ϵ, then your answer is accepted too.

Q2. MDPs - Policy Iteration (25 points)

In a car race, a car robot has three states $\{cool, warm, overheated\}$. At each state, the robot can either move *fast* or *slow*. At state *overheated*, the robot has to exit the race and receive a reward 0. The transition function and reward function are illustrated in the following figure

Consider running policy iteration with the following initial policy: $\pi_0(cool) = fast$, $\pi_0(warm) = fast$ and $\pi_0(overheated) = \emptyset$. The discount factor is $\gamma = 0.1$.



Q2.1. (12.5 points) What are values of states given the fixed policy π_0: $V^{\pi_0}(cool)$, $V^{\pi_0}(warm)$, $V^{\pi_0}(overheated)$? What is the policy π_i at iteration $i = 1$?

Answer. There are two methods to compute utilities of states. The first one is using value iteration-based approach and the second one is using the equation system approach.

Value iteration-based approach. We have $V_0^{\pi_0}(cool) = V_0^{\pi_0}(warm) = 0$. Note that for value iteration-based approach, you can set a stopping condition, for example, if $|V_{k+1}^{\pi_0}(cool) - V_k^{\pi_0}(cool)| < \epsilon$ with $\epsilon = 0.001$, then stop the iteration.

- It can be easily verify t $.1 \times 0] = -1$
- $V_1^{\pi_0}(cool) = 0.5$ $)] = 2$
- $V_2^{\pi_0}(cool) = 0.5 \times [2 + 0.1 \times 2] + 0.5 \times [2 + 0.1$
- $V_3^{\pi_0}(cool) = 0.5 \times [2 + 0.1 \times 2.05] + 0.5 \times [2 + 0.$
- $V_4^{\pi_0}(cool) = 0.5 \times [2 + 0.1 \times 2.0525] + 0.5 \times [2 + 0.1 \times (-1)] = 2.052625$
- Given the stopping condition, then $V^{\pi_0}(cool) \approx 2.052625$ since $2.052625 - 2.0525 < 0.001$

Equation system approach. We have the following equation system

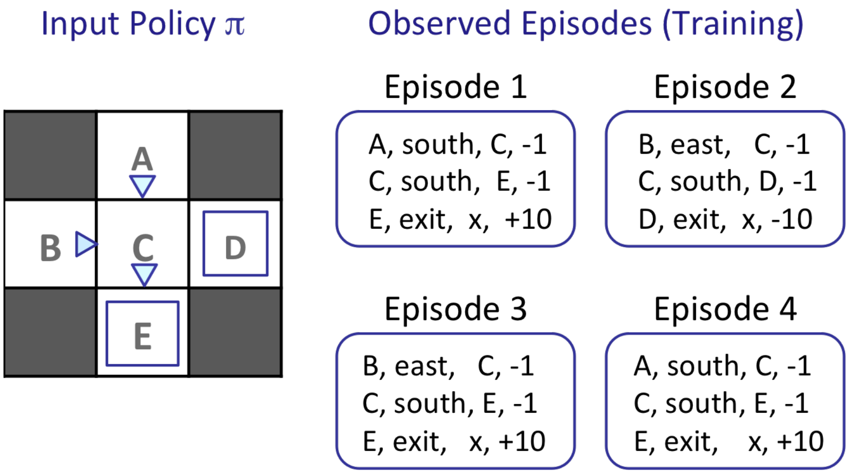
$$
\begin{aligned}
&V^{\pi_0}(warm) = -1.0\\
&V^{\pi_0}(cool) = 0.5 \times [2 + 0.1 \times V^{\pi_0}(cool)] + 0.5 \times [2 + 0.1 \times V^{\pi_0}(warm)]\\
&\implies 0.95 \times V^{\pi_0}(cool) = 1.95 \implies V^{\pi_0}(cool) = \frac{1.95}{0.95} \approx 2.0526.
\end{aligned}
$$

Q2.2. (12.5 points) What is the optimal policy π^* for the robot car? What are the values of $V^*(cool)$,$V^*(warm)$, and $V^*(overheated)$?

Answer. The optimal policy is $\pi^*(cool) = fast$, $\pi^*(warm) = slow$. The values of states are $V^*(warm) = \frac{7}{6} \approx 1.1667$, $V^*(fast) = \frac{13}{6} \approx 2.1667$.

Q3. Model-based RL (10 points)

An agent lives in a grid world as shown in the figure below. The agent tries out a policy π which is indicated by the arrows in the figure. After four trials, the agent observes four episodes.



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. $T(A, sout$ $T(C, south, D) = 0.25.$
In addition, $R(A, sout$ $-1, R(C, south, D) = -1$

Q4. RL - Direct Evaluation (10 points)

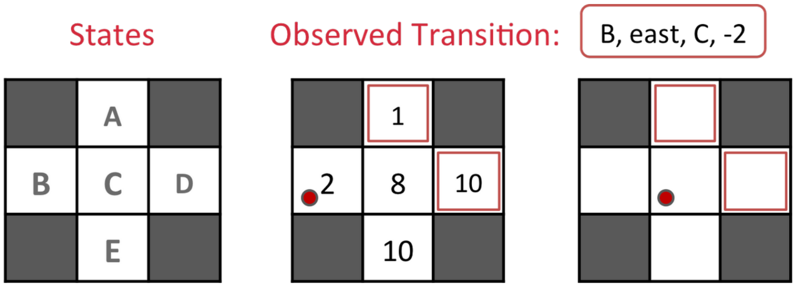
Consider the same problem as in **Q3**. What are the estimat (B), $\hat{V}^\pi(C)$, $\hat{V}^\pi(D)$, and $\hat{V}^\pi(E)$ as obtained by direct evaluation? Assume the discoun 0.8.

Answer.

$$\hat{V}^\pi(A) = -1 + 0.8 \times (-1) + (0.8)^2 \times 10 = 4.6$$
$$\hat{V}^\pi(B) = \frac{1}{2}(-1 + 0.8 \times (-1) + 0.64 \times (-10)) + \frac{1}{2}(-1 + 0.8 \times (-1) + 0.64 \times 10) = -1.8$$
$$\hat{V}^\pi(C) = \frac{3}{4} \times (-1 + 0.8 \times 10) + \frac{1}{4} \times (-1 + 0.8 \times (-10)) = 3$$
$$\hat{V}^\pi(D) = -10$$
$$\hat{V}^\pi(E) = 10$$

Q5. RL - Temporal Difference Learning (6 points)

Consider the grid world 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 $\gamma = 0.8, \alpha = 0.75$, 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.

- $\hat{V}^{\pi}(A) = 1$
- $\hat{V}^{\pi}(B) = 3.8$
- $\hat{V}^{\pi}(C) = 8$
- $\hat{V}^{\pi}(D) = 10$
- $\hat{V}^{\pi}(E) = 10$

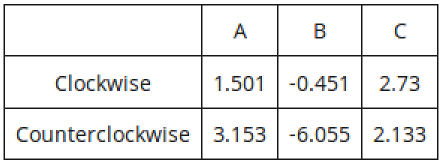
The only value that gets up tarts in state B.

$$\hat{V}^{\pi}(B) = 0.25 \times 2 + 0.75 \times ($$

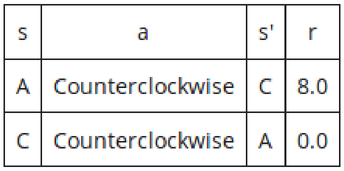
Q6. Model-free Reinforcement Learning (1

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 0.8 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.

Answer.

- $Q(A, clockwise) = 1.501$
- $Q(A, counterclockwise) = 6.6685$
- $Q(B, clockwise) = -.451$
- $Q(B, counterclockwise) = -6.055$
- $Q(C, clockwise) = 2.73$
- $Q(C, counterclockwise) = 3.7339$

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: $Q(A, counterclockwise) = .5 \times 3.153 + .5 \times (8 + .8 \times 2.73) = 6.6685$
Then we update: $Q(C, counterclockwise) = .5 \times 2.133 + .5 \times (0 + .8 \times 6.6685) = 3.7339$.
Because there are only two samples, the other four values stay the same.

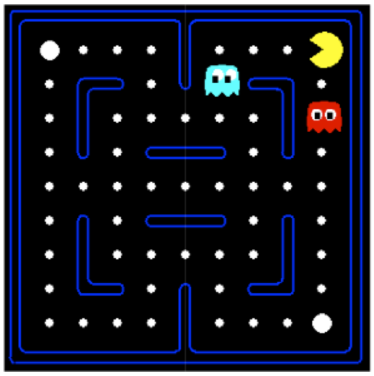
Q7. RL - Feature-based Representation (12 points)

Consider the following feat $Q(s,a) = w_1 f_1(s,a) +$ $w_2 f_2(s,a)$ with:

- $f_1(s,a) = 1/$ (Manhattan distance to nearest dot after ha)
- $f_2(s,a) =$(Manhattan distance to nearest ghost after navi

Q7.1. Assume $w_1 = 3$ 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 $Q(s, south)$.

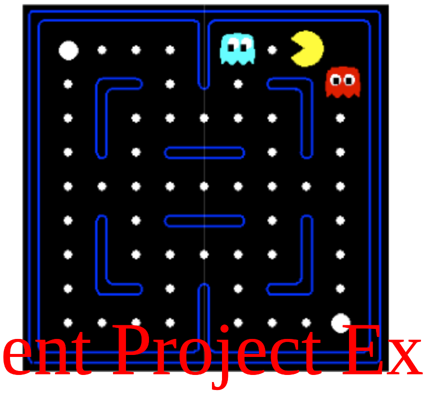
Based on this approximate Q-function, which action would be chosen?



Answer.

- $Q(s, west) = 3 \times 1 + 8 \times 3 = 27$
- $Q(s, south) = 3 \times 1 + 8 \times 1 = 11$
- The chosen action is west since $27 > 11$.

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 suming $\gamma = 0.8$)?

Answer.

- $Q(s', west) = 3 \times 1 + 8 \times 1 = 11$
- $Q(s', east) = 3 \times 1 + 8 \times 1 = 11$
- $sample = [r + \gamma \times \max_{a'} Q(s', a')] = 9 + 0.8 \times 11 = 17.8$

Q7.3. Now provide the update to the weights. Let $\alpha = 0.5$.

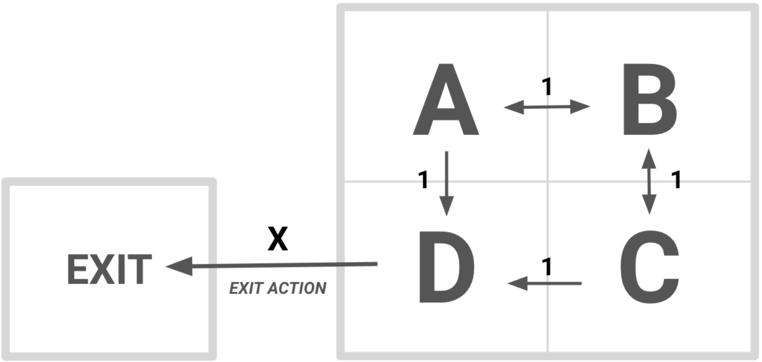
Answer. $w_1 = -4.75$ and $w_2 = -12.25$

Explanation. difference $= [r + \gamma \max_{a'} Q(s', a')] - Q(s, a) = 17.8 - 27 = -9.2$. Therefore,

$$w_1 = w_1 + \alpha(\text{difference})f_1(s, a) = 3 + .5 \times (-9.2) \times 1 = -1.6$$
$$w_2 = w_2 + \alpha(\text{difference})f_2(s, a) = 8 + .5 \times (-9.2) \times 3 = -5.8$$

Q8. Strange MDPs (Graduates Only) (20 points)

In this MDP, the available actions at state A, B, C are LEFT, RIGHT, UP, and DOWN unless there is a wall in that direction. The only action at state D is the EXIT ACTION and gives the agent a reward of x. The reward for non-exit actions is always 1.



Q8.1 (10 points) Let all actions be deterministic. Assume $\gamma = 0.5$. Express the values of states $V^*(D)$, $V^*(A)$, $V^*(C)$, $V^*(B)$ in terms of x.

Answer. We have:

$$V^*(D) = x$$

$$V^*(A) = \max\{1 + 0.5x, 2\}$$

The 2 comes from the utility being an infinite geometric sum of discounted reward $= \frac{1}{1-\frac{1}{2}} = 2$. The idea behind the above equations is to explore the symmetric relatio and C. First, D has only action to take, which is to exit the game. For B, d $Q^*(B, LEFT)$ must be the same as $Q^*(B, DOWN)$. As a result, both LEF n optimal action to take from B. For A and C, if optimal action to take from A is DOWN, then the optimal action to take from C is LEFT. Similarly, if the optimal action to take from A is RIGHT, then the optimal action to take from C is UP. Based on this analysis, there are only two policies which we should consider

- Policy 1: $\pi(D) = exit$, $\pi(A) = DOWN$, $\pi(C) = LEFT$, $\pi(B) = LEFT$ item Policy 2: $\pi(D) = exit$, $\pi(A) = RIGHT$, $\pi(C) = UP$, $\pi(B) = LEFT$

As a result, you only have to compute the values of states according to these policies and take the maximum between the two.

Q8.2 (10 points) Let any non-exit action be successful with probability $= 0.5$. Otherwise, the agent stays in the same state with reward $= 0$. The EXIT ACTION from the state D is still deterministic and will always succeed. Assume that $\gamma = 0.5$. For which value of x does $Q^*(A, DOWN) = Q^*(A, RIGHT)$?

Answer. $Q^*(A, DOWN) = Q^*(A, RIGHT)$ implies $V^*(A) = Q^*(A, DOWN) = Q^*(A, RIGHT)$. Therefore,

$$V^*(A) = Q^*(A, DOWN) = \frac{1}{2}(0 + \frac{1}{2}V^*(A)) + \frac{1}{2}(1 + \frac{1}{2}x) = \frac{1}{2} + \frac{1}{4}V^*(A) + \frac{1}{4}x \quad (1)$$

$$\implies V^*(A) = \frac{2}{3} + \frac{1}{3}x \quad (2)$$

$$V^*(A) = Q^*(A, RIGHT) = \frac{1}{2}(0 + \frac{1}{2}V^*(A)) + \frac{1}{2}(1 + \frac{1}{2}V^*(B)) = \frac{1}{2} + \frac{1}{4}V^*(A) + \frac{1}{4}V^*(B) \quad (3)$$

$$\implies V^*(A) = \frac{2}{3} + \frac{1}{3}V^*(B) \quad (4)$$

Because $Q^*(B, LEFT)$ and $Q^*(B, DOWN)$ are symmetric decisions, $V^*(B) = Q^*(B, LEFT)$. Therefore,

$$V^*(B) = \frac{1}{2}(0 + \frac{1}{2}V^*(B)) + \frac{1}{2}(1 + \frac{1}{2}V^*(A)) = \frac{1}{2} + \frac{1}{4}V^*(B) + \frac{1}{4}V^*(A) \quad (5)$$

$$\implies V^*(B) = \frac{2}{3} + \frac{1}{3}V^*(A) \quad (6)$$

From (4) and (6) we obtain, $V^*(A) = V^*(B) = 1$. From (2), we obtain $x = 1$.