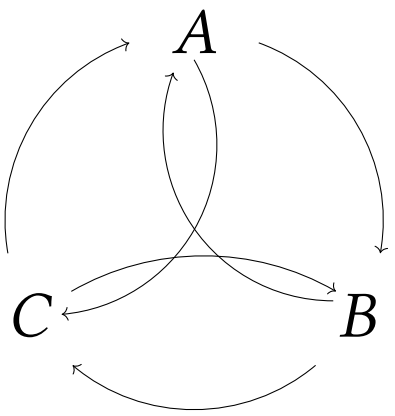
CIS 571 - Assignment 3

Ali Hassani

November 10, 2021

Question 1

1.1 Part 1 - Cycle



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

$\gamma = 0.5$.

P1.1. Suppose after iteration k, we obtain the values below. Provide values at $k+1$.

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

For $s \in \{A, B, C\}$, $V_{k+1}(s) = \max_a Q_{k+1}(s, a)$, and $Q_{k+1}(s, a) = \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V_k(s')]$. Therefore:

	A	B	C
↻	0.70	1.53	2.40
↺	-0.92	1.35	1.30

Table 1: Q_{k+1}

Therefore:

$V_{k+1}(A)$	$V_{k+1}(B)$	$V_{k+1}(C)$
0.70	1.53	2.40

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

$V^\star(A)$	$V^\star(B)$	$V^\star(C)$
0.881	1.761	2.616

For $s \in \{A, B, C\}$, $V^\star(s) = \max_a Q^\star(s, a)$, and $Q^\star(s, a) = \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V^\star(s')]$ (convergence of value iteration). Therefore:

	A	B	C
↻	**0.8805**	**1.7610**	**2.6165**
↺	-0.6920	1.5875	1.5045

Table 2: $Q^\star$

Therefore, the optimal action at every state is moving clockwise.

1.2 Part 2 - Convergence

For all transitions from any state to another, reward is 1. From F, there is no transition to another state, just to itself, the reward for which is 0.

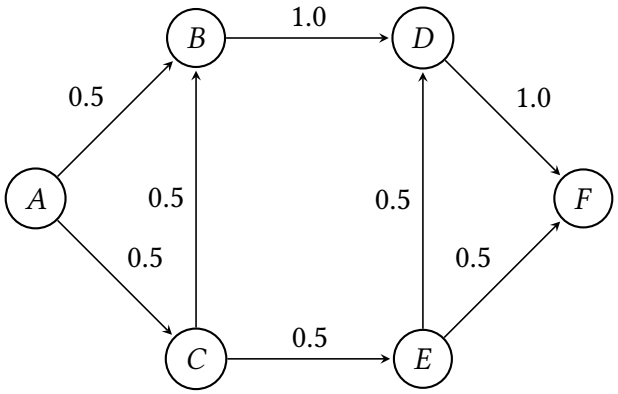


Figure 1: Transition diagram

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.)

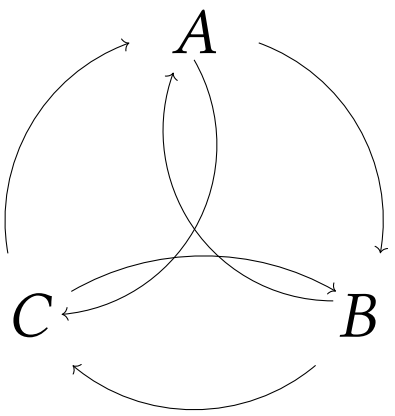
After 2 iterations. To explicitly check: after the first iteration, $V_1(E) = 1.0$, and after the second, $V_2(E) = 1.25$.

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.)

After 4 iterations. To explicitly check:

$V^\star(A)$	$V^\star(B)$	$V^\star(C)$	$V^\star(D)$	$V^\star(E)$	$V^\star(F)$
1.796875	1.500000	1.687500	1.000000	1.250000	0.000000

Question 2



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

Table 3: Transition & reward functions

$\gamma = 0.5$.

2.1 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.

A	B	C
Counterclockwise	Counterclockwise	Counterclockwise

$V^{\pi}_{k+1}(A)$	$V^{\pi}_{k+1}(B)$	$V^{\pi}_{k+1}(C)$
0.000	−0.840	−1.080

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

$V^{\pi}_{k+1}(A)$	$V^{\pi}_{k+1}(B)$	$V^{\pi}_{k+1}(C)$
−2.540	0.584	0.548

2.2 Suppose that policy evaluation converges to the following value function.

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

Provide the values of $Q^{\pi}_{\infty}(A, \circlearrowright)$ and $Q^{\pi}_{\infty}(A, \circlearrowleft)$. What is the updated action for A?

The V_{∞}s above lead to the following Q_{∞}s:

	$Q^{\pi}_{\infty}(A)$	$Q^{\pi}_{\infty}(B)$	$Q^{\pi}_{\infty}(C)$
$\circlearrowright$	−0.56	0.38	1.72
$\circlearrowleft$	−2.63	0.49	0.42

(Note that this contradicts the assumption that the values converged, because at least one policy from this Q should have been approximately equal to the values provided. In other words,)

$Q^{\pi}_{\infty}(A, \circlearrowright) = -0.56$

$Q^{\pi}_{\infty}(A, \circlearrowleft) = -2.63$

Therefore, the updated action for A would be **Clockwise** ($\circlearrowright$).

Question 3

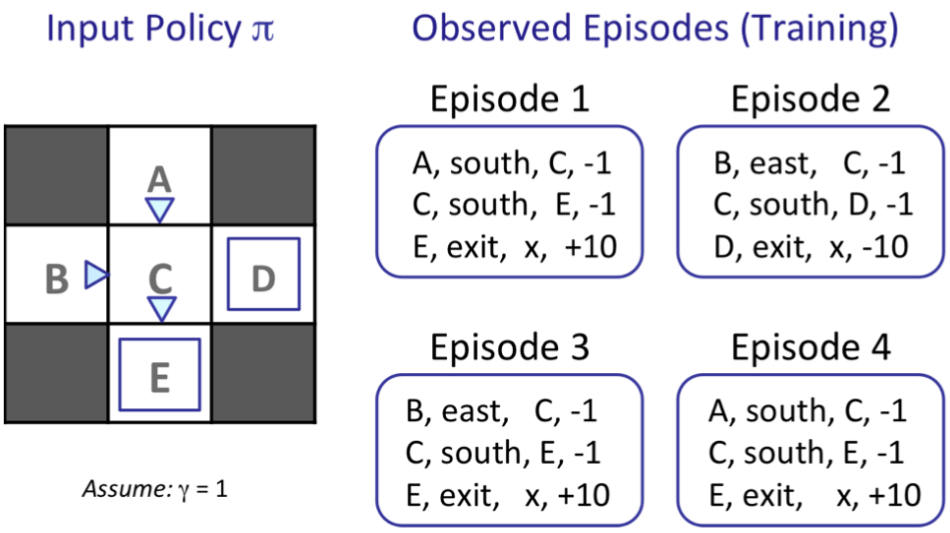


Figure 2: Q3

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

s	a	s'	$T(s, a, s')$	$R(s, a, s')$
A	south	C	1.0	-1
B	east	C	1.0	-1
C	south	D	0.25	-1
C	south	E	0.75	-1
D	exit	x	1.0	-10
E	exit	x	1.0	+10

Table 4: Learned transition & reward functions

Question 4

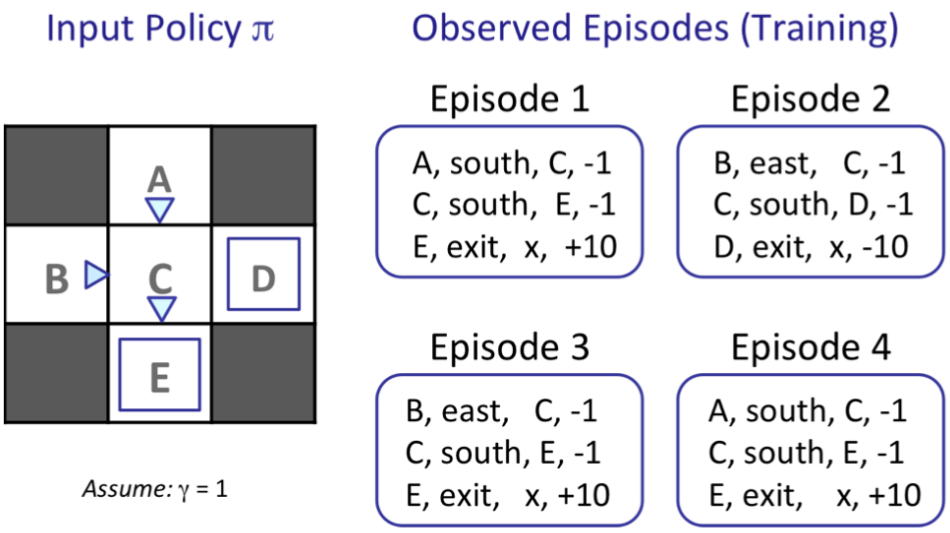


Figure 3: Q4

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

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

Question 5

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 $V^\pi(A)$, $V^\pi(B)$, $V^\pi(C)$, $V^\pi(D)$, and $V^\pi(E)$ after the TD learning update?

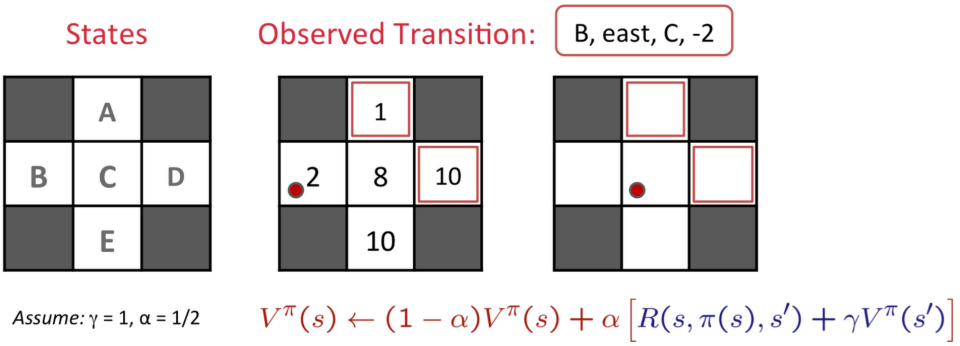


Figure 4: Q5

$V^\pi(A)$	$V^\pi(B)$	$V^\pi(C)$	$V^\pi(D)$	$V^\pi(E)$
1	2	8	10	10

$V^\pi(B) = (1-\alpha)V^\pi(B) + \alpha[-2 + V^\pi(C)] = 4$

Only $V^\pi(B)$ is updated from 2 to 4, the rest of the values stay the same.

Question 6

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
↻	1.501	-0.451	2.730
↺	3.153	-6.055	2.133

s	a	s'	r
A	↻	C	8.0
C	↺	A	0.0

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

Assuming it is a batch-wise update (not sample-wise, because $Q(C, ↺)$ depends on the value of $Q(A, ↺)$ which is being updated):

$$Q(A, ↺) \Leftarrow (1 - \alpha)Q(A, ↺) + \alpha(sample)$$

$$\begin{aligned} Q(A, ↺) &= (0.25 \times 3.153) + (0.75) \times (8.0 + (0.75 \times \max_{a'} Q(C, a'))) \\ &= (0.25 \times 3.153) + (0.75) \times (8.0 + (0.75 \times 2.730)) \\ &= (0.78825) + (0.75) \times (8.0 + 2.0475) \\ &= (0.78825) + (0.75) \times (10.0475) \\ &= 0.78825 + 7.535625 \\ &= 8.323875 \end{aligned}$$

$$\begin{aligned} Q(C, ↺) &= (0.25 \times 2.133) + (0.75) \times (0.0 + (0.75 \times \max_{a'} Q(A, a'))) \\ &= (0.25 \times 2.133) + (0.75) \times ((0.75 \times 3.153)) \\ &= 0.53325 + 1.7735625 \\ &= 2.3068125 \end{aligned}$$

Updated Q-values:

	A	B	C
↻	1.501	-0.451	2.730
↺	**8.323875**	-6.055	**2.3068125**

Assuming it is NOT a batch-wise update (it is sample-wise), $Q(C, \circlearrowleft)$ would be computed as follows:

$$\begin{aligned} Q(C, \circlearrowleft) &= (0.25 \times 2.133) + (0.75) \times (0.0 + (0.75 \times \max_{a'} Q(A, a'))) \\ &= (0.25 \times 2.133) + (0.75) \times ((0.75 \times 8.323875)) \\ &= 0.53325 + 4.6821796875 \\ &= 5.21543 \end{aligned}$$

Updated Q-values:

	A	B	C
$\circlearrowleft$	1.501	-0.451	2.730
$\circlearrowleft$	**8.323875**	-6.055	**5.21543**

Question 7

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:

- $f_1(s, a) = \frac{1}{\text{Manhattan distance to nearest dot after having executed action } a \text{ in state } s)}$
- $f_2(s, a)$ = Manhattan distance to nearest ghost after having executed action a in state s

Assuming $w_1 = 2$ and $w_2 = 8$,

7.1 Assume that the red and blue ghosts are both sitting on top of a dot. Provide the following values. Based on them, which action would be chosen.

$Q(s, west)$ **and** $Q(s, south)$**.**

Given that pacman is currently at state s:

$$f_1(s, west) = \frac{1}{1} = 1$$
$$f_2(s, west) = 3$$
$$f_1(s, south) = \frac{1}{2} = 0.5$$
$$f_2(s, south) = 1$$

Therefore:

$$Q(s, west) = 1 \times 2 + 3 \times 8 = 26$$
$$Q(s, south) = 0.5 \times 2 + 1 \times 8 = 9$$

Therefore, action *west* will be chosen.

7.2 Assume Pac-Man moves West. This results in the state 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 $\gamma = 0.8$)?

$$f_1(s', west) = \frac{1}{1} = 1$$
$$f_2(s', west) = 1$$
$$f_1(s', east) = \frac{1}{2} = 0.5$$
$$f_2(s', east) = 1$$

Therefore:

$$Q(s, west) = 1 \times 2 + 1 \times 8 = 10$$
$$Q(s, east) = 0.5 \times 2 + 1 \times 8 = 9$$

As a result:

$$\begin{aligned} sample &= [r + \gamma \max_{a'} Q(s', a')] \\ &= [9 + 0.8(10)] \\ &= 17 \end{aligned}$$

and:

$$\begin{aligned} difference &= sample - Q(s, west) \\ &= 17 - 26 \\ &= -9 \end{aligned}$$

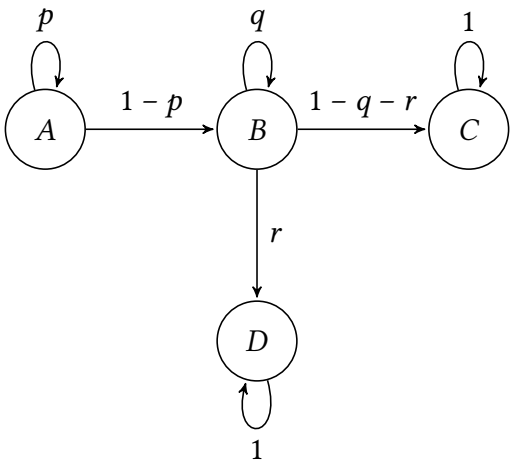
7.3 Now provide the update to the weights.

Let $\alpha = 0.75$.

$$
\begin{aligned}
w_1 &= w_1 + \alpha \times difference \times f_1(s, west) \\
&= 2 + 0.75 \times (-9) \times 1 \\
&= -4.75 \\
w_2 &= w_2 + \alpha \times difference \times f_2(s, west) \\
&= 8 + 0.75 \times (-9) \times 3 \\
&= -12.25
\end{aligned}
$$

Question 8

Pacman is trying to predict the position of a ghost, which 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?
Answer: Because once the ghost leaves A there is no way back:
p^n

b. What is the probability that the ghost first reaches B at time n?
Answer: Because there is no way back to B from its successors C and D, the ghost reaching B at n can only mean the ghost was at A at $n - 1$, therefore:
$(p^{n-1})(1 - p)$ when $n \geq 1$ otherwise 0.

c. What is the probability that the ghost is in B at time n?
Answer: For similar reasons, the ghost being in B at n means it either got to B for the first time, or it reached B at some step less than n and stayed for the remainder:
$\sum_{i=0}^{n-1} (p^i)(1 - p)q^{n-i-1}$ when $n \geq 1$ otherwise 0.

d. What is the probability that the ghost first reaches C at time n?
Answer: It was in B at $n - 1$, and reached C with a probability of $1 - q - r$, therefore:
$[\sum_{i=0}^{n-2} (p^i)(1 - p)q^{n-i-2}](1 - q - r)$ when $n \geq 2$ otherwise 0.

e. What is the probability that the ghost is in C at time n?
Answer: For similar reasons, the ghost being in C at n means it either got to C for the first time, or it reached C at some step less than n and stayed for the remainder with probability 1:
$\sum_{j=2}^{n} \left([\sum_{i=0}^{j-2} (p^{\,i})(1 - p)q^{\,j-i-2}](1 - q - r)\right)$ when $n \geq 2$ otherwise 0.