COMPSCI 402 Artificial Intelligence

Assignment 2 – MDP   Total points: 8-point

Q1. Pacman is using MDPs to maximize his expected utility. In each environment:

- Pacman has the standard actions $\{\boldsymbol{North}, \boldsymbol{East}, \boldsymbol{South}, \boldsymbol{West}\}$ unless blocked by an outer wall
- There is a reward of 1 point when eating the dot (for example, in the grid below, $R(C;\ South;\ F) = 1$)
- The game ends when the dot is eaten

(a) Consider the following grid where there is a single food pellet in the bottom right corner (F). The discount factor is 0.5. There is no living reward. The states are simply the grid locations.

A	B	C
D	E	F ○

(i) What is the optimal policy for each state? (1-point)

State	$\pi(state)$
A	east or south
B	east or south
C	south
D	east
E	east

(ii) What is the optimal value for the state of being in the upper left corner (A)? Reminder: the discount factor is 0.5. (1-point)

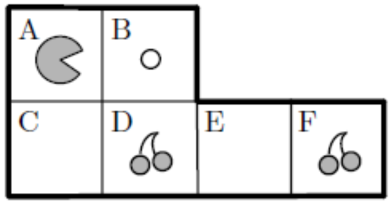
$V^*(A) =$

k	V(A)	V(B)	V(C)	V(D)	V(E)	V(F)
0	0	0	0	0	0	0
1	0	0	1	0	1	0
2	0	0.5	1	0.5	1	0
3	0.25	0.5	1	0.5	1	0
4	0.25	0.5	1	0.5	1	0

(iii) Using value iteration with the value of all states equal to zero at k=0, for which iteration k will $V_k(A) = V^*(A)$? (1-point)

$k =$ 3

(b) Consider a new Pacman level that begins with cherries in locations D and F. Landing on a grid position with cherries is worth 5 points and then the cherries at that position **disappear**. There is still one dot, worth 1 point. The game still only ends when the dot is eaten.



(i) With no discount ($\gamma = 1$) and a living reward of -1, what is the optimal policy for the states in this level's state space? (1-point)

State (*hint: three-element tuple*)	$\pi(state)$
A, D_{cherry}=true, F_{cherry} = true	South
A, D_{cherry}=true, F_{cherry} = false	South
A, D_{cherry}=false, F_{cherry} = true	East
A, D_{cherry}=false, F_{cherry} = false	East
C, D_{cherry}=true, F_{cherry} = true	East
C, D_{cherry}=true, F_{cherry} = false	East
C, D_{cherry}=false, F_{cherry} = true	East
C, D_{cherry}=false, F_{cherry} = false	North / East
D, D_{cherry}=false, F_{cherry} = true	East
D, D_{cherry}=false, F_{cherry} = false	North
E, D_{cherry}=true, F_{cherry} = true	East
E, D_{cherry}=true, F_{cherry} = false	West
E, D_{cherry}=false, F_{cherry} = true	East
E, D_{cherry}=false, F_{cherry} = false	West
F, D_{cherry}=true, F_{cherry} = false	West
F, D_{cherry}=false, F_{cherry} = false	West

(ii) With no discount ($\gamma = 1$), what is the range of living reward values such that Pacman eats exactly one cherry when starting at position A? (1-point)

Suppose x is the living reward

when eating zero cherry $\{A, B\}$, reward is $x+1$

when eating two cherries $\{A, C, D, E, F, E, D, B\}$, reward is $7x+11$

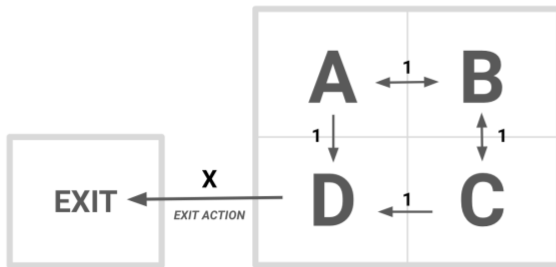
and for eating one cherry. $\{A, C, D, B\}$, reward is $3x+6$

$\because \quad 7x+11 > 3x+6 > x+1$

$\therefore \quad -2.5 < x < -1.25$

so range of living reward is $(-2.5, -1.25)$

Q2. 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**.



(a) Let all actions be deterministic. Assume $\gamma = \frac{1}{2}$. Express the following in terms of x. (1-point)

$V^*(D) = x$ $\qquad$ $V^*(C) = \max(1+0.5x, 2)$

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

(b) Let any non-exit action be successful with **probability**$= \frac{1}{2}$. 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 $\gamma = \frac{1}{2}$. For which value of x does $Q^*(A; DOWN) = Q^*(A; RIGHT)$? Box your answer and justify/show your work. (1-point)

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

$= Q^*(A, DOWN) = Q^*(A, RIGHT)$

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

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

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

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

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

$\because V^*(B) = Q^*(B, LEFT)$

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

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

$\therefore x = 1$

(c) We now add one more layer of complexity. Turns out that the reward function is not guaranteed to give a particular reward when the agent takes an action. Every time an agent transitions from one state to another, once the agent reaches the new state s', a fair 6-sided dice is rolled. If the dices lands with value x, the agent receives the reward $R(s, a, s') + x$. The sides of dice have value 1, 2, 3, 4, 5, and 6. Write down the new bellman update equation for $V_{k+1}(s)$ in terms of $T(s, a, s'), R(s, a, s'), V_k(s'), and\ \gamma$. (1-point)

$$V_{k+1}(s) = \max_a \sum_{s'} T(s,a,s') \left[\frac{1}{6}\left(\sum_{i=1}^{6} R(s,a,s') + i\right) + \gamma V_k(s')\right]$$

$$= \max_a \sum_{s'} T(s,a,s') \left(R(s,a,s') + 3.5 + \gamma V_k(s')\right)$$