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	
B	
C	
D	
E	

(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)

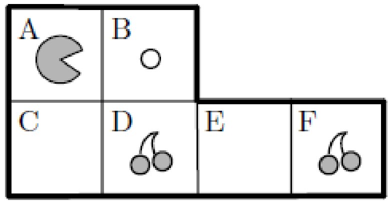
$V^*(A) =$

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

(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)

$k =$

(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.

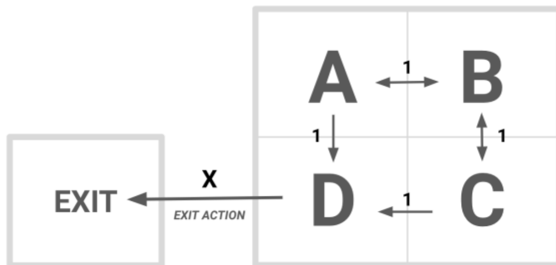


(i) With no discount ($\gamma = 1$) and a living reward of -1, what is the optimal policy for the states in this level's state space? (1-point)

State (*hint: three-element tuple*)	$\pi(state)$
A,	
A,	
A,	
A,	
C,	
C,	
C,	
C,	
D,	
D,	
E,	
E,	
E,	
E,	
F,	
F,	

(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)

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)

$V^*(D) =$ $\qquad\qquad\qquad$ $V^*(C) =$

$V^*(A) =$ $\qquad\qquad\qquad$ $V^*(B) =$

(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)

(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)