CS 188
Fall 2023

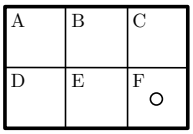
Midterm Review MDPs Solutions

Q1. MDP

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

- Pacman has the standard actions {North, East, South, 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 a 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.



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

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.

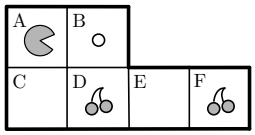
$V^*(A) = 0.25$

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 $V_k(A) = V^*(A)$?

$k = 3$ (see above)

(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?

State	$\pi(state)$
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?

Valid range for the living reward is (-2.5,-1.25).

Let x equal the living reward.

The reward for eating zero cherries {A,B} is $x + 1$ (one step plus food).

The reward for eating exactly one cherry {A,C,D,B} is $3x + 6$ (three steps plus cherry plus food).

The reward for eating two cherries {A,C,D,E,F,E,D,B} is $7x + 11$ (seven steps plus two cherries plus food).

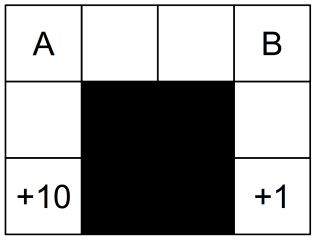
x must be greater than -2.5 to make eating at least one cherry worth it ($3x + 6 > x + 1$).

x must be less than -1.25 to eat less than one cherry ($3x + 6 > 7x + 11$).

Q2. MDPs: Value Iteration

An agent lives in gridworld G consisting of grid cells $s \in S$, and is not allowed to move into the cells colored black. In this gridworld, the agent can take actions to move to neighboring squares, when it is not on a numbered square. When the agent is on a numbered square, it is forced to exit to a terminal state (where it remains), collecting a reward equal to the number written on the square in the process.

Gridworld G



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

(a) Consider running value iteration in gridworld G. Assume all legal movement actions **will always succeed** (and so the state transition function is deterministic).

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

$k =$ 3 $\quad V_k(A) =$ 10

The nearest reward is 10, which is 3 steps away. Because $\gamma = 1$, there is no decay in the reward, so the value propagated is 10.

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

$k =$ 3 $\quad V_k(B) =$ 1

The nearest reward is 1, which is 3 steps away. Because $\gamma = 1$, there is no decay in the reward, so the value propagated is 1.

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

$k =$ 3 $\quad V^*(A) =$ 10

Because $\gamma = 1$, the problem reduces to finding the distance to the highest reward (because there is no living reward). The highest reward is 10, which is 3 steps away.

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

$k =$ 6 $\quad V^*(B) =$ 10

Because $\gamma = 1$, the problem reduces to finding the distance to the highest reward (because there is no living reward). The highest reward is 10, which is 6 steps away.

(b) Now assume all legal movement actions **succeed with probability** 0.8; with probability 0.2, the action fails and the agent remains in the same state.
Consider running value iteration in gridworld G. What is the smallest iteration k for which $V_k(A) =$

$V^*(A)$? What is the value of $V^*(A)$?

$k =$ ∞

$V^*(A) =$ 10

Because $\gamma = 1$ and the only rewards are in the exit states, the optimal policy will move to the exit state with highest reward. This is guaranteed to ultimately 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.