CS 188 Fall 2024 Introduction to Artificial Intelligence Written HW4

Due: Tuesday 10/1 at 11:59pm.

Policy: Can be solved in groups (acknowledge collaborators) but must be submitted individually.

Make sure to show all your work and justify your answers.

Note: This is a typical exam-level question. On the exam, you would be under time pressure, and have to complete this question on your own. We strongly encourage you to first try this on your own to help you understand where you currently stand. Then feel free to have some discussion about the question with other students and/or staff, before independently writing up your solution.

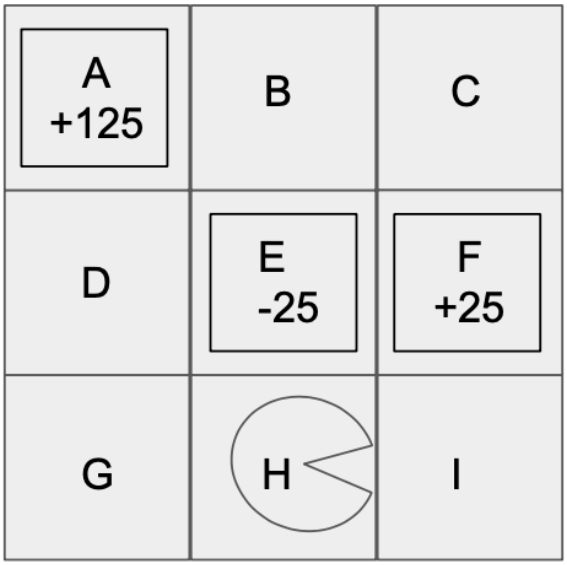
Note: Leave the self-assessment sections blank for the original submission of your homework. After the homework deadline passes, we will release the solutions. At that time, you will review the solutions, self-assess your initial response, and complete the self-assessment sections below. The deadline for the self-assessment is 1 week after the original submission deadline.

Your submission on Gradescope should be a PDF that matches this template. Each page of the PDF should align with the corresponding page of the template (page 1 has name/collaborators, question begins on page 2.). **Do not reorder, split, combine, or add extra pages**. The intention is that you print out the template, write on the page in pen/pencil, and then scan or take pictures of the pages to make your submission. You may also fill out this template digitally (e.g. using a tablet.)

First name	
Last name	
SID	
Collaborators	

Q1. [15 pts] MDP

Pacman is in the following grid, starting at state H. He can take any one of four actions: Up, Down, Left and Right. If he tries to take an action that would move him towards a wall, then he stays in place. For example, if he goes down from H, he will remain at H. A, E, and F are exit states. Once Pacman enters an exit state, he will exit immediately and receive the appropriate award shown in the diagram and listed below.



Hence, for any state s and action a:

$$R(s,a,s') = \begin{cases} 125 & \text{if } s' = A, \\ -25 & \text{if } s' = E, \\ 25 & \text{if } s' = F, \\ 0 & \text{otherwise.} \end{cases}$$

Assume that all actions are deterministic unless otherwise specified.

(a) Fill in the optimal action(s) for H for the given discount factors.

(i) [1 pt]

$\gamma = 1$ ☐ Up ■ Left ■ Right

Using a discount factor of 1, there is no decay in reward and with deterministic actions, the optimal policy will always be to take exit A. Performing value iteration, we can see that after 5 iterations, all states will have a value of 125, so the optimal policy would be moving to any non-terminal state other than A.

(ii) [1 pt]

$\gamma = \frac{1}{5}$ ☐ Up ■ Left ■ Right

Using a discount factor of $\frac{1}{5}$, we can perform value iteration and solve for the optimal values of G and I to be $V^*(G) = \left(\frac{1}{5}\right)^2 \cdot 125 = 5$ and $V^*(I) = \frac{1}{5} \cdot 25 = 5$. Since the optimal value of I and G are equal, both left and right are optimal actions. Taking action up will enter exit state E and receive a reward of -25.

(iii) [1 pt]

$\gamma = \frac{1}{10}$ ☐ Up ☐ Left ■ Right

Using a discount factor of $\frac{1}{10}$, we can perform value iteration and solve for the optimal values of I and G to be $V^*(G) = \left(\frac{1}{10}\right)^2 \cdot 125 = 1.25$ and $V^*(I) = \frac{1}{10} \cdot 25 = 2.5$. Taking action up will enter exit state E and receive a reward of -25. So the optimal policy will be to go right toward state I.

(b) Assume for this part only that the transition function for the MDP is deterministic for all transitions except the following:

$T(H, \text{ Right}, I) = p$ $\quad$ $T(H, \text{ Left}, E) = p$
$T(H, \text{ Right}, E) = 1 - p$ $\quad$ $T(H, \text{ Left}, G) = 1 - p$

What values of p for the given discount factor γ would ensure that Pacman strictly prefers taking action Right from H:

(i) [2 pts]

$\gamma = 1$ ☐ 0 ☐ 0.1 ☐ 0.5 ■ 0.9 ■ 1 ☐ None of the above

Using a discount factor of 1, there is no decay in reward so running value iteration after 3 iterations will give $V^*(G) = 125$ and $V^*(I) = 25$. As before, taking the action up will guarantee a reward of -25. Now considering the non-determinism of actions left and right for state H, we can calculate the associated Q-values.
$Q(H, \text{ Left}) = p \cdot (-25) + (1 - p) \cdot 125 = 125 - 150p$
$Q(H, \text{ Right}) = p \cdot 25 + (1 - p) \cdot (-25) = 50p - 25$
To solve for the value of p that makes the Right action preferable, we want $Q(H, \text{ Right}) > Q(H, \text{ Left})$. This means $50p - 25 > 125 - 150p$ which can be reduced to $p > 0.5$.

(ii) [2 pts]

$\gamma = \frac{1}{10}$ ☐ 0 ☐ 0.1 ■ 0.5 ■ 0.9 ■ 1 ☐ None of the above

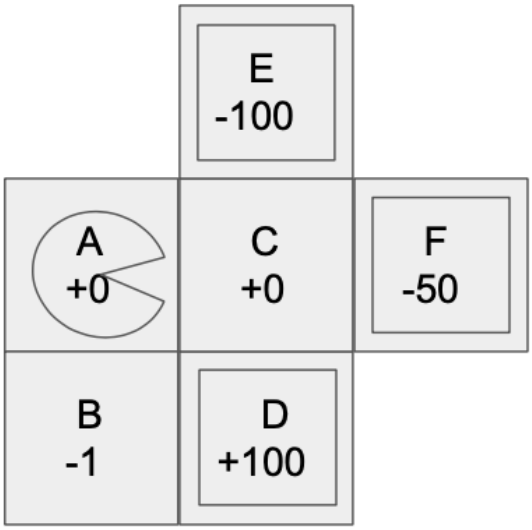
With discounted rewards, the optimal values become $V^*(G) = 1.25$ and $V^*(I) = 2.5$ as calculated in part a. As before, taking the action up will guarantee a reward of -25. Now considering the non-determinism of actions left and right for state H, we can calculate the associated Q-values.
$Q(H, \text{ Left}) = p \cdot (-25) + (1 - p) \cdot 1.25 = 1.25 - 26.5p$
$Q(H, \text{ Right}) = p \cdot 2.5 + (1 - p) \cdot (-25) = 27.5p - 25$
To solve for the value of p that makes the Right action preferable, we want $Q(H, \text{ Right}) > Q(H, \text{ Left})$. This means $27.5p - 25 > 1.25 - 26.5p$ which can be reduced to $p > 0.486$.

Q1(a-b) Self-Assessment - leave this section blank for your original submission. We will release the solutions to this problem after the deadline for this assignment has passed. After reviewing the solutions for this problem, assess your initial response by checking one of the following options:

- ◯ I fully solved the problem correctly, including fully correct logic and sufficient work (if applicable).
- ◯ I got part or all of the question incorrect.

If you selected the second option, explain the mistake(s) you made and why your initial reasoning was incorrect (do not re-iterate the solution. Instead, reflect on the errors in your original submission). Approximately 2-3 sentences for *each* incorrect sub-question.

(c) Consider the following new grid:



Here, D, E and F are exit states.

Pacman starts in state A. The reward for entering each state is reflected in the grid. Assume that discount factor $\gamma = 1$.

(i) [3 pts] Write the optimal values $V^*(s)$ for $s = A$ and $s = C$ and the optimal policy $\pi^*(s)$ for $s = A$.

$V^*(A) =$ 100
$V^*(C) =$ 100

$\pi^*(A) =$ ○ Up ○ Down ○ Left ● Right

Since the discount factor is 1 and actions are deterministic, all values will eventually converge to the optimal exit value of 100. The optimal action from A is to move right to state C since $Q(A, \text{Right}) = 100$ and $Q(A, \text{Down}) = 99$.

(ii) [2 pts] Now, instead of Pacman, Pacbaby is travelling in this grid. Pacbaby has a more limited set of actions than Pacman and can never go left. Hence, Pacbaby has to choose between actions: Up, Down and Right.

Pacman is rational, but Pacbaby is indecisive. If Pacbaby enters state C, Pacbaby finds the two best actions and randomly, with equal probability, chooses between the two. Let $\pi^*(s)$ represent the optimal policy for **Pacman**. Let $V(s)$ be the values under the policy where Pacbaby acts according to $\pi^*(s)$ for all $s \neq C$, and follows the indecisive policy when at state C. What are the values $V(s)$ for $s = A$ and $s = C$?

$V(A) =$ 25
$V(C) =$ 25

We first calculated $V(C)$. Under the indecisive policy, Pacbaby will choose between actions down and right (exit states D and F) with equal probability so $V(C) = \frac{1}{2} \cdot 100 + \frac{1}{2} \cdot (-50) = 25$.
For $V(A)$, Pacbaby is following Pacman's optimal policy which we calculated in the previous part to be the Right action. Since the discount factor is 1 and actions are deterministic, the value of A is $V(A) = 25$.

(iii) [3 pts] Now Pacman knows that Pacbaby is going to be indecisive when at state C and he decides to recompute the optimal policy for Pacbaby at all other states, anticipating his indecisiveness at C. What is Pacbaby's new policy $\pi(s)$ and new value $V(s)$ for $s = A$?

$V(A) =$ 99

$\pi(A) =$ ○ Up ● Down ○ Left ○ Right

From the previous part, we know that $Q(A, \text{ Right}) = 25$. We also know from part a(i) that $Q(A, \text{ Down}) = 99$. Therefore, the value of A is the max over all Q-values which would be 99 by taking action Down.

Q1(c) Self-Assessment - leave this section blank for your original submission. We will release the solutions to this problem after the deadline for this assignment has passed. After reviewing the solutions for this problem, assess your initial response by checking one of the following options:

- ◯ I fully solved the problem correctly, including fully correct logic and sufficient work (if applicable).
- ◯ I got part or all of the question incorrect.

If you selected the second option, explain the mistake(s) you made and why your initial reasoning was incorrect (do not re-iterate the solution. Instead, reflect on the errors in your original submission). Approximately 2-3 sentences for *each* incorrect sub-question.