CS 188 Spring 2025 — Introduction to Artificial Intelligence — Written HW5

Due: Wednesday 3/5 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. [6 pts] Exploring the World

In this question, our CS188 agent is stuck in a maze. We use Q-learning with an epsilon greedy strategy to solve the task. There are 4 actions available: north (N), east (E), south (S), and west (W).

(a) [2 pts] What is the probability of each action if the agent is following an epsilon greedy strategy and the best action in state s under the current policy is N? Given that we are following an epsilon-greedy algorithm, we have a value ϵ. Use this value ϵ in your answer. $p(a_i|s)$ is the probability of taking action a_i in state s.

$p(N|s)$ $\boxed{(1 - \epsilon) + 0.25\epsilon}$ $p(E|s)$ $\boxed{0.25\ \epsilon}$

$p(S|s)$ $\boxed{0.25\ \epsilon}$ $p(W|s)$ $\boxed{0.25\ \epsilon}$

The solution should place equal probabilities on E, S, and W, and the rest should be placed on N. The probability for N should decrease linearly with ϵ.

(b) [2 pts] We also modify the reward original reward function $R(s, a, s')$ to visit more states and choose new actions. Which of the following rewards would encourage the agent to visit unseen states and actions?

$N(s, a)$ refers to the number of times that you have visited state s and taken action a in your samples.

- □ $R(s, a, s') + \sqrt{N(s, a)}$
- ■ $R(s, a, s') + \sqrt{\frac{1}{N(s,a)+1}}$
- ■ $\sqrt{\frac{1}{N(s,a)+1}}$
- ■ $R(s, a, s') - \sqrt{N(s, a)}$
- □ $-\sqrt{\frac{1}{N(s,a)+1}}$
- ■ $\exp(R(s, a, s') - N(s, a))$

The modified reward should be a monotonically decreasing function of N.

(c) [2 pts] Which of the following modified rewards will eventually converge to the optimal policy with respect to the original reward function $R(s, a, s')$? $N(s, a)$ is the same as defined in part (b).

- □ $R(s, a, s') + \sqrt{N(s, a)}$
- ■ $R(s, a, s') + \sqrt{\frac{1}{N(s,a)+1}}$
- □ $\sqrt{\frac{1}{N(s,a)+1}}$
- □ $R(s, a, s') - \sqrt{N(s, a)}$
- □ $-\sqrt{\frac{1}{N(s,a)+1}}$

□ $\exp(R(s, a, s') - N(s, a))$

The modified reward should converge to the original reward as N increases.

Q1(a-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.

Q2. [14 pts] Square World

In this question we will consider the following gridworld:

0	**0**	**0**	**0**	**0**	**0**
+1	**A**	**B**	**C**	**D**	**+100**
0	**0**	**0**	**0**	**0**	**0**

Every grid square corresponds to a state. If a state is annotated with a number, it means that after entering this state only one action is available, the Exit action, and this will result in the reward indicated by the number and the episode will end. There is no reward otherwise. For the other 4 states, named A, B, C, D, two actions are available: Left and Right. Assume $\gamma = 1$ in this problem.

(a) [4 pts] Assume that the failure probability for the actions Left and Right is 0.5, and in case of failure the agents moves up or down, and the episode is terminated after the Exit action.

What are the optimal values?

States	A	B	C	D
$V^*(s)$	6.25	12.5	25	50

The optimal policy is to go to the right from state A, B, C, and D. Thus,

$V^*(D) = 0.5 * 0 + 0.5 * 100 = 50$

$V^*(C) = 0.5 * 0 + 0.5 * V(D) = 25$

$V^*(B) = 0.5 * 0 + 0.5 * V(C) = 12.5$

$V^*(A) = 0.5 * 0 + 0.5 * V(B) = 6.25$

(b) [4 pts] Still assume the failure probability in the previous part. Now assume further that there is an *integer* living reward r when the episode is not terminated. Is there a value of r that would make the optimal policy **only** decide Left at state D? If so, what's the minimum value of r?

Answer: 51

Let X0 represent each of the grid that has number 0. We make no distinction between them because $V(X0)$ is the same for all $X0$. Let X100 and X1 represent the states that have number 100 and 1, respectively.

At convergence we have the optimal value function V^* such that

$V^*(D) = \max\{r + 0.5 * V(X100) + 0.5 * V(X0), r + 0.5 * V(C) + 0.5 * V(X0)\}$

$V^*(C) = \max\{r + 0.5 * V(B) + 0.5 * V(X0), r + 0.5 * V(D) + 0.5 * V(X0)\}$

$V^*(B) = \max\{r + 0.5 * V(A) + 0.5 * V(X0), r + 0.5 * V(C) + 0.5 * V(X0)\}$

$V^*(A) = \max\{r + 0.5 * V(X1) + 0.5 * V(X0), r + 0.5 * V(B) + 0.5 * V(X0)\}$

Because the optimal policy decides only Left at D, we have $V(C) > V(X100)$. We can show that $V^*(x) = 2r$ for all $x \in A, B, C, D$. Thus, we have $2r > 100$ and the smallest integer is 51.

Another interpretation is to give the living reward when taking the Exit action from state $X0$ and $X100$. In this case, $V^*(x) = 3r$ for all $x \in A, B, C, D$, and $3r > r + 100$. So, the answer is still 51.

(c) [4 pts] Assume we collected the following episodes of experiences in the form (state, action, next state, reward): (we use $X1$ and $X100$ to denote the leftmost and rightmost states in the middle row and Done to indicate the terminal state after an Exit action).

$$(B, \text{Left}, A, 0), (A, \text{Left}, X1, 0), (X1, \text{Exit}, Done, +1)$$
$$(B, \text{Right}, C, 0), (C, \text{Right}, D, 0), (D, \text{Right}, X100, 0), (X100, \text{Exit}, Done, +100)$$

If we run Q-learning initializing all Q-values equal to 0, and with appropriate stepsizing, replaying each of the above episodes infinitely often till convergence, what will be the resulting values for:

(State, Action)	(B, Left)	(B, Right)	(C, Left)	(C, Right)
$Q^*(s,a)$	1	100	0	100

At convergence, $Q^*(B, Left) = Q^*(A, Left) = Q^*(Exit, Done) = 1$ and $Q^*(B, Right) = Q^*(C, Right) = Q^*(D, Right) = Q^*(X100, Exit) = 100$. Other state-action pairs will have 0 value because we have not seen them.

(d) [2 pts] Now we are trying to do feature-based Q-learning. Answer the below True or False question.

There exists a set of features that are functions of state only such that approximate Q-learning will converge to the optimal Q-values.

○ True ● False

False, because optimal Q values in this gridworld depend on actions.

Q2(a-d) 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.