

CS188: Exam Practice Session 4

Q1. Wandering Poet

In country B there are N cities. They are all connected by roads in a circular fashion. City 1 is connected with city N and city 2. For $2 \leq i \leq N - 1$, city i is connected with cities $i - 1$ and $i + 1$.

A wandering poet is travelling around the country and staging shows in its different cities.

He can choose to move from a city to a neighboring one by moving East or moving West, or stay in his current location and recite poems to the masses, providing him with a reward of r_i . If he chooses to travel from city i , there is a probability $1 - p_i$ that the roads are closed because of B 's dragon infestation problem and he has to stay in his current location. The reward he is to reap is 0 during any successful travel day, and $r_i/2$ when he fails to travel, because he loses only half of the day.

(a) Let $r_i = 1$ and $p_i = 0.5$ for all i and let $\gamma = 0.5$. For $1 \leq i \leq N$ answer the following questions *with real numbers*:

Hint: Recall that $\sum_{j=0}^{\infty} u^j = \frac{1}{1-u}$ for $u \in (0, 1)$.

(i) What is the value $V^{stay}(i)$ under the policy that the wandering poet always chooses to stay?

(ii) What is the value $V^{west}(i)$ of the policy where the wandering poet always chooses west?

(b) Let N be even, let $p_i = 1$ for all i , and, for all i , let the reward for cities be given as

$$r_i = \begin{cases} a & i \text{ is even} \\ b & i \text{ is odd,} \end{cases}$$

where a and b are constants and $a > b > 0$.

(i) Suppose we start at an even-numbered city. What is the range of values of the discount factor γ such that the optimal policy is to stay at the current city forever? Your answer may depend on a and b .

(ii) Suppose we start at an odd-numbered city. What is the range of values of the discount factor γ such that the optimal policy is to stay at the current city forever? Your answer may depend on a and b .

(iii) Suppose we start at an odd-numbered city and γ does not lie in the range you computed. Describe the optimal policy.

(c) Let N be even, $r_i \geq 0$, and the optimal value of being in city 1 be positive, i.e., $V^*(1) > 0$. Define $V_k(i)$ to be the value of city i after the k th time-step. Letting $V_0(i) = 0$ for all i , what is the largest k for which $V_k(1)$ could still be 0? Be wary of off-by-one errors.

(d) Let $N = 3$, and $[r_1, r_2, r_3] = [0, 2, 3]$ and $p_1 = p_2 = p_3 = 0.5$, and $\gamma = 0.5$. Compute:

(i) $V^*(3)$

(ii) $V^*(1)$

(iii) $Q^*(1, stay)$

Q2. MDPs: Dice Bonanza

A casino is considering adding a new game to their collection, but need to analyze it before releasing it on their floor. They have hired you to execute the analysis. On each round of the game, the player has the option of rolling a fair 6-sided die. That is, the die lands on values 1 through 6 with equal probability. Each roll costs 1 dollar, and the player **must** roll the very first round. Each time the player rolls the die, the player has two possible actions:

1. *Stop*: Stop playing by collecting the dollar value that the die lands on, or
2. *Roll*: Roll again, paying another 1 dollar.

Having taken CS 188, you decide to model this problem using an infinite horizon Markov Decision Process (MDP). The player initially starts in state *Start*, where the player only has one possible action: *Roll*. State s_i denotes the state where the die lands on i . Once a player decides to *Stop*, the game is over, transitioning the player to the *End* state.

- (a) In solving this problem, you consider using policy iteration. Your initial policy π is in the table below. Evaluate the policy at each state, with $\gamma = 1$.

State	s_1	s_2	s_3	s_4	s_5	s_6
$\pi(s)$	<i>Roll</i>	<i>Roll</i>	<i>Stop</i>	<i>Stop</i>	<i>Stop</i>	<i>Stop</i>
$V^\pi(s)$						

- (b) Having determined the values, perform a policy update to find the new policy π' . The table below shows the old policy π and has filled in parts of the updated policy π' for you. If both *Roll* and *Stop* are viable new actions for a state, write down both *Roll/Stop*. In this part as well, we have $\gamma = 1$.

State	s_1	s_2	s_3	s_4	s_5	s_6
$\pi(s)$	<i>Roll</i>	<i>Roll</i>	<i>Stop</i>	<i>Stop</i>	<i>Stop</i>	<i>Stop</i>
$\pi'(s)$	<i>Roll</i>					<i>Stop</i>

(c) Is $\pi(s)$ from part (a) optimal? Explain why or why not.

(d) Suppose that we were now working with some $\gamma \in [0, 1)$ and wanted to run **value iteration**. Select the **one** statement that would hold true at convergence, or write the correct answer next to Other if none of the options are correct.

☐ $V^*(s_i) = \max \left\{ -1 + \frac{i}{6}, \sum_j \gamma V^*(s_j) \right\}$

☐ $V^*(s_i) = \max \left\{ i, \frac{1}{6} \cdot \left[-1 + \sum_j \gamma V^*(s_j) \right] \right\}$

☐ $V^*(s_i) = \max \left\{ -\frac{1}{6} + i, \sum_j \gamma V^*(s_j) \right\}$

☐ $V^*(s_i) = \max \left\{ i, -\frac{1}{6} + \sum_j \gamma V^*(s_j) \right\}$

☐ $V^*(s_i) = \frac{1}{6} \cdot \sum_j \max \{ i, -1 + \gamma V^*(s_j) \}$

☐ $V^*(s_i) = \frac{1}{6} \cdot \sum_j \max \left\{ -1 + i, \sum_k V^*(s_k) \right\}$

☐ $V^*(s_i) = \sum_j \max \left\{ -1 + i, \frac{1}{6} \cdot \gamma V^*(s_j) \right\}$

☐ $V^*(s_i) = \sum_j \max \left\{ \frac{i}{6}, -1 + \gamma V^*(s_j) \right\}$

☐ $V^*(s_i) = \max \left\{ i, -1 + \frac{\gamma}{6} \sum_j V^*(s_j) \right\}$

☐ $V^*(s_i) = \sum_j \max \left\{ i, -\frac{1}{6} + \gamma V^*(s_j) \right\}$

☐ $V^*(s_i) = \sum_j \max \left\{ \frac{-i}{6}, -1 + \gamma V^*(s_j) \right\}$

☐ Other _____