AI 3000 : Reinforcement Learning

Assignment № 1

Due Date : 10/09/2025

Easwar Subramanian, IIT Hyderabad 25/08/2025

Problem 1 : Markov Reward Process

Consider the following snake and ladders game as depicted in the figure below.

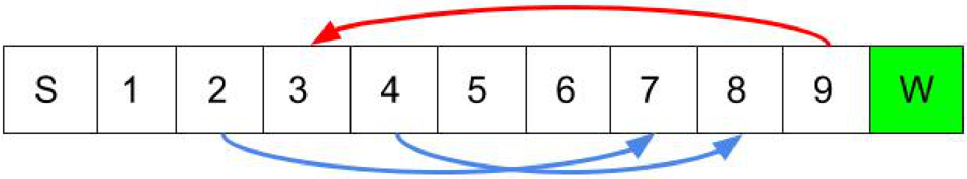

Figure 1: A Simple Snake and Ladder Problem

- Initial state is S and a fair four sided dice is used to decide the next state at each time
- Player must land exactly on state W to win
- Dice throw that take you further than state W leave the state unchanged

(a) Identify the states, transition matrix of this Markov process. (2 points)

(b) Construct a suitable reward function, discount factor and use the Bellman equation for the Markov reward process to compute how long does it take "on average" (the expected number of dice throws) to reach the state W from any other state. (3 points)

(a) The states of the Markov process is given by, $\mathcal{S} = \{S, 1, 3, 5, 6, 7, 8, W\}$. Positions 2 and 4 of the grid are same as positions 7 and 8 respectively.

The transition matrix is given by,

$$P = \begin{pmatrix} 0 & 0.25 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.25 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0 & 0.25 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0.25 & 0 & 0 & 0 & 0.5 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(b) The state W is an absorbing state since if the Markov process reach the state W there are no further state transitions possible apart from staying at state W.

(c) Following are the suitable reward functions and discount factor γ.

- The suitable discount factor is $\gamma = 1$ as we are estimating "average" number of steps to reach W.
- The reward for any state could be $R(s) = -1$ for $s \neq W$ and $R(s) = 0$ for $s = W$. Then $V(s) = 0$ for $s = W$.
- The Bellman evaluation equation for an MRP given by $V = (I - \gamma P)^{-1}R$ which when solved for $V(s)$ would give the "average number" of die throws required to reach state W from state s. The matrix $(I - \gamma P)$ becomes invertible if we set $V(s) = 0$ for $s = W$. One may find the inverse of the matrix $(I - \gamma P_{7\times7})$ to compute the average die throws from other seven states. Upon solving, the vector $V(s)$ is given by,

$$V(s) = \{7.0833, 7, 6.6667, 6.6667, 5.3333, 5.3333, 5.3333\}$$

Problem 2 : Effect of Noise and Discounting

Consider the grid world problem shown in Figure 2. The grid has two terminal states with positive

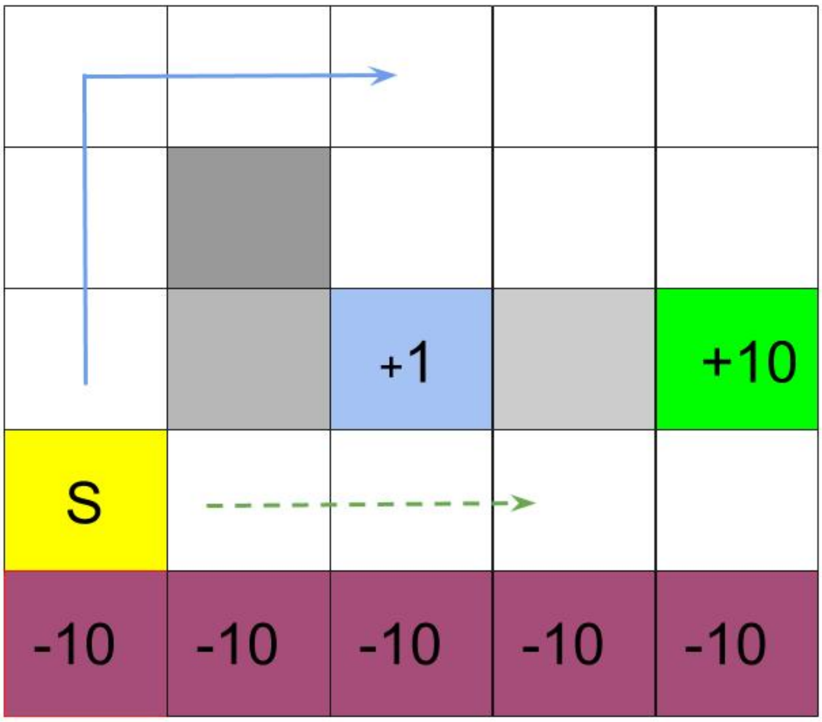

Figure 2: Modified Grid World

payoff (+1 and +10). The bottom row is a cliff where each state is a terminal state with negative payoff (-10). The greyed squares in the grid are walls. The agent starts from the yellow state

S. As usual, the agent has four actions $\mathcal{A}$ = (Left, Right, Up, Down) to choose from any non-terminal state and the actions that take the agent off the grid leaves the state unchanged. Notice that, if agent follows the dashed path, it needs to be careful not to step into any terminal state at the bottom row that has negative payoff. There are four possible (optimal) paths that an agent can take.

- Prefer the close exit (state with reward +1) but risk the cliff (dashed path to +1)
- Prefer the distant exit (state with reward +10) but risk the cliff (dashed path to +10)
- Prefer the close exit (state with reward +1) by avoiding the cliff (solid path to +1)
- Prefer the distant exit (state with reward +10) by avoiding the cliff (solid path to +10)

There are two free parameters to this problem. One is the discount factor γ and the other is the noise factor (η) in the environment. Noise makes the environment stochastic. For example, a noise of 0.2 would mean the action of the agent is successful only 80 % of the times. The rest 20 % of the time, the agent may end up in an unintended state after having chosen an action.

(a) Identify what values of γ and η lead to each of the optimal paths listed above with reasoning. If necessary, you could implement the value iteration algorithm on this environment and observe the optimal paths for various choices of γ and η. (5 Points)

[**Hint** : For the discount factor, try high and low γ values like 0.9 and 0.1 respectively. For noise, consider deterministic and stochastic environment with noise level η being 0 or 0.5 respectively]

1. When γ is low, RL agent is 'short sighted' and better rewards available in the distant future is not given importance. Further, when noise is zero in the environment, there is no danger of tripping to the cliff. Therefore, for low γ and low η, the agent would prefer the close exit and risk the cliff.
2. When γ is low, RL agent is 'short sighted' and better rewards available in the distant future is not given importance. Further, when noise is high or moderate in the environment, there is danger of tripping to the cliff. Therefore, for low γ and low η, the agent would prefer the close exit and not risk the cliff.
3. When γ is high, RL agent is 'far sighted' and better rewards available in the distant future is given importance. Further, when noise is low or zero in the environment, there is less or no danger of tripping to the cliff. Therefore, for high γ and low η, the agent would prefer the distant exit and risk the cliff.
4. When γ is high, RL agent is 'far sighted' and better rewards available in the distant future is given importance. Further, when noise is high or medium in the environment, there is danger of tripping to the cliff. Therefore, for high γ and high η, the agent would prefer the distant exit and not risk the cliff.

Problem 3 : Markov Decision Process

Let M be an infinite horizon MDP given by $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$ with $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$ and $\gamma \in [0, 1)$. Suppose that the reward function $\mathcal{R}(s, a, s')$ for any successor states $s, s' \in \mathcal{S}$ and action $a \in \mathcal{A}$ is non-negative and bounded.

(a) Let $\hat{M} =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \hat{\mathcal{R}}, \gamma >$ be another infinite horizon MDP with a modified reward function $\hat{\mathcal{R}}$ such that

$$\mathcal{R}(s, a, s') - \hat{\mathcal{R}}(s, a, s') = \varepsilon$$

where ε is a constant independent of $s \in \mathcal{S}$ or $a \in \mathcal{A}$. Given a policy π, let V^π and $\hat{V}^\pi$ be value functions of policy π for MDPs M and $\hat{M}$ respectively. Derive an expression that relates $V^\pi(s)$ to $\hat{V}^\pi(s)$ for any state $s \in \mathcal{S}$ of the MDP. (3 Points)

Considering the definition of $V^\pi(s)$, the state value function under policy π, we have

$$V^\pi(s) = \mathbb{E}_\pi \left(\sum_{k=0}^{\infty} \gamma^k r^{t+k+1} \right)$$

We assume that each reward has a constant added to it. That is we consider the reward $\hat{r}_{t+k+1}$ in terms of r_{t+k+1} by

$$\hat{r}_{t+k+1} = r_{t+k+1} + \varepsilon$$

Then, the state value function for this new sequence of rewards is given by,

$$\begin{aligned}
\hat{V}^\pi(s) &= \mathbb{E}_\pi \left(\sum_{k=0}^{\infty} \gamma^k \hat{r}^{t+k+1} \right) \\
&= \mathbb{E}_\pi \left(\sum_{k=0}^{\infty} \gamma^k \left(r_{t+k+1} + \varepsilon \right) \right) \\
&= \mathbb{E}_\pi \left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \right) + \mathbb{E}_\pi \left(\gamma^k \epsilon \right) \\
&= V^\pi(s) + \mathbb{E}_\pi \left(\gamma^k \varepsilon \right) = V^\pi(s) + \varepsilon \sum_{k=0}^{\infty} \gamma \\
&= V^\pi(s) + \frac{\varepsilon}{1 - \gamma}
\end{aligned}$$

The alternate relation

$$\hat{V}^\pi = V^\pi (I - \gamma P)^{-1} \varepsilon$$

is dependent on the model of the MDP.

(b) Does M and $\hat{M}$ have the same optimal policy ? Explain. (3 Points)

The MDPs M and $\hat{M}$ will have the same optimal policy as :

$$\arg\max_a \left[\hat{r}(s, a, s') + \gamma \sum_s^{\prime} P(s'|s, a) \hat{V}_*(s') \right] = \arg\max_a \left[r(s, a, s') + \varepsilon + \gamma \sum_s^{\prime} P(s'|s, a) V_*(s') + \frac{\varepsilon}{1 - \gamma} \right]$$

$$\arg\max_a \left[r(s,a,s') + \gamma \sum_s^{\prime} P(s'|s,a)V_*(s') + \varepsilon + \frac{\varepsilon}{1-\gamma} \right] = \arg\max_a \left[r(s,a,s') + \gamma \sum_s^{\prime} P(s'|s,a)V_*(s') \right]$$

(c) State and prove an analogous result for the sub-question (a) for the case when M and $\hat{M}$ are finite horizon MDPs with horizon length $H < \infty$. (4 Points)

A derivation similar to part(a) will yield the answer

$$\hat{V}^{\pi}(s) = V^{\pi}(s) + \frac{\varepsilon(1-\gamma^H)}{1-\gamma}$$

(d) Now, consider an indefinite MDP or a stochastic shortest path MDP where the horizon length H can vary. A subset of the state space $S_{\text{term}} \subset \mathcal{S}$ is considered terminal if a trajectory of the form $s_0, a_0, r_1, s_1, a_1, r_2, \cdots$, keeps rolling out until a terminal state $S_H \in S_{\text{term}}$ is visited. In general, the length of the episode H is a random variable. Does the analogous result of sub-question (a) hold when M and $\hat{M}$ are indefinite MDPs ? Explain. (4 Points)

The simple answer is no. It is not easy to come up with similar relationships as in part(a). Either a counter-example or a derivation similar to part(e) with the final expression having an expectation over H is accepted as an answer..

(e) For this sub-question let $\hat{M} =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \hat{\mathcal{R}}, \gamma >$ be a infinite horizon MDP with a modified reward function $\hat{\mathcal{R}}$ such that

$$\left| \mathcal{R}(s,a,s') - \hat{\mathcal{R}}(s,a,s') \right| \le \varepsilon$$

where ε is a constant independent of s and a . Derive an expression that relates the optimal value functions $V_*(s)$ and $\hat{V}_*(s)$. Would M and $\hat{M}$ have the same optimal policy ? Explain. (6 Points)

From

$$\left| \mathcal{R}(s,a,s') - \hat{\mathcal{R}}(s,a,s') \right| \le \varepsilon$$

we can write

$$R(s,a,s') - \varepsilon \le \hat{R}(s,a,s') \le R(s,a,s') + \varepsilon$$

Now using definitions of V^{π} and $\hat{V}(\pi)$, we can derive a relation

$$\left| V^{\pi}(s) - \hat{V}^{\pi}(s) \right| \le \frac{\varepsilon}{1-\gamma}$$

from which a similar relationship between V_* and $\hat{V}_*$ can be derived. That M and $\hat{M}$ will not have the same optimal policy can be argued either using the argmax similar to part (c) or through counter example.

Problem 4 : Programming Value and Policy Iteration

Implement value and policy iteration algorithm and test it on **'Frozen Lake'** environment in openAI gym. **'Frozen Lake'** is a grid-world like environment available in gym. The purpose of this exercise is to to help you get hands on with using gym and to understand the implementation details of value and policy iteration algorithm(s)

This question will not be graded but **will** still come in handy for future assignments.

ALL THE BEST