AI 3000 / CS 5500 : Reinforcement Learning

Assignment № 1

Due Date : 04/09/2023

Couse Instructor : Easwar Subramanian 23/08/2023

Problem 1 : Markov Reward Process

Consider a fair four sided dice with faces marked as $\{'1', '2', '3', '4'\}$. The dice is tossed repeatedly and independently. By formulating a suitable Markov reward process (MRP) and using Bellman equation for MRP, find the expected number of tosses required for the pattern $'1234'$ to appear. Specifically, answer the following questions.

(a) Identify the states, transition probablities and terminal states (if any) of the MRP (3 Points)

(b) Construct a suitable reward function, discount factor and use the Bellman equation for MRP to find the 'average' number of tosses required for the pattern $'1234'$ to appear. (7 Points)

[**Explanation : For the target pattern to occur, four consequective tosses of the dice should result in different faces of the dice being on the top, in the specific order** $'1$**,** $'2'$**,** $'3'$ **and** $'4'$]

Answer

Call 1234 our target. Consider a chain that starts from a state called nothing (denote by $\emptyset$) and is eventually absorbed at 1234. If we first toss 1 then we move to state 1 because this is the first letter of our target. If we toss any other facethen we move back to $\emptyset$ having expended 1 unit of time. Being in state 1 we either move to a new state 12 if we toss 2 and we are 1 step closer to the target or, if we toss 1 we move back to state 1. If any other face shows up, we move back to $\emptyset$: we have expended 1 more unit of time. We can construct the state sequence similarly and the transition diagram looks like below.

Now we can write down the states of the MRP $< \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma >$ as follows.

- The set of states $\mathcal{S} = \{\emptyset, 1, 12, 123, 1234\}$
- The transition matrix $\mathcal{P}$ is given by,

$$
\begin{array}{c} \\ \emptyset \\ 1 \\ 12 \\ 123 \\ 1234 \end{array}
\begin{array}{c}
\begin{array}{ccccc} \emptyset & 1 & 12 & 123 & 1234 \end{array} \\
\begin{pmatrix}
0.75 & 0.25 & 0 & 0 & 0 \\
0.5 & 0.25 & 0.25 & 0 & 0 \\
0.5 & 0.25 & 0 & 0.25 & 0 \\
0.5 & 0.25 & 0 & 0 & 0.25 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}
\end{array}
$$

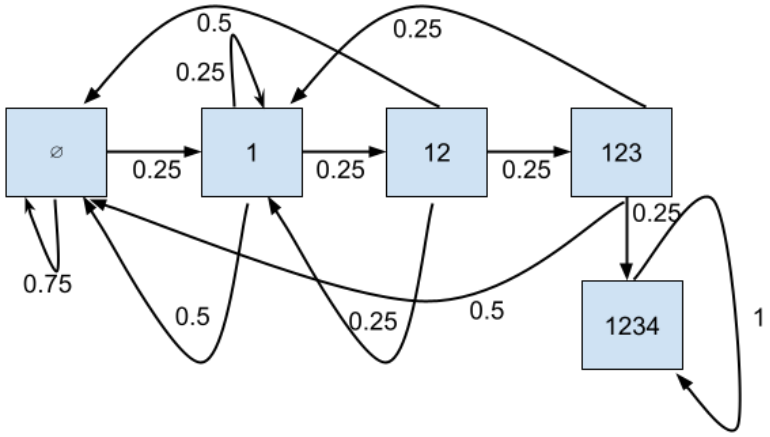

Figure 1: Suitable Markov Reward Process

- The absorbing state is 1234 and this MRP is very similar to the snake and ladder problem discussed in the class. So, every time we toss the dice, we get a reward of -1 and when we reach the absorbing state we get a reward of 0. So, $\mathcal{R}(s) = -1 \quad \text{for} \quad s \in \{\phi, 1, 12, 123\}$ and $\mathcal{R}(1234) = 0$.
- The discount factor $\gamma = 1$.

The Bellman evaluation equation for an MRP is given by $V = (I - \gamma\mathcal{P})^{-1}\mathcal{R}$ which when solved for $V(s)$ would give the "expected number" of dice throws required to reach state 1234 from any other state s of the MRP. The matrix $(I - \gamma P)$ becomes invertible if we set $V(s) = 0$ for $s = 1234$. One may find the inverse of the matrix $(I - \gamma\mathcal{P}_{4\times4})$ and multiply with $\mathcal{R}_{4\times1}$ to compute the expected dice throws from any given state of the MRP. Specifically, we are interested from state $\emptyset$. Upon solving one can find that the expected number of coin tosses from state $\emptyset$ to reach 1234 is 256.

Problem 2 : Markov Decision Process

(a) 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, what is the lower and upper bound on the discounted sum of rewards ? (3 Points)

(b) 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.

(c) 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 :

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

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

(d) From sub-question (b) can one argue that the assumption that the MDP M in sub-question (a) has non-negative and bounded reward is without loss in generality ? What if the MDP M is allowed to have negative but bounded rewards ? (3 Points)

Yes, it is WLOG. The max of the negative reward can be added to all rewards and from part (a) and (b), one can argue that the optimal policies are same.

(e) State and prove an analogous result for the sub-question (b) 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

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

We have accepted various versions of this answer (like $H+1$, H, $H-t$ etc..)

(f) 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 (b) 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..

(g) 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

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

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

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

we can write

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

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

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

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.

(h) Now consider the MDP M of sub-question (a). Does scaling the discount factor by a constant $\kappa \in (0,1)$ alter the optimal policy ? Explain. (4 Points)

An example to show that the optimal policies will be differnt if the discount factor is scaled would be sufficient. One such example is in Assignment 2 (gridworld with two goal states)

ALL THE BEST