# 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** $'4'$]

## 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 $\varepsilon$ is a constant independent of $s \in \mathcal{S}$ or $a \in \mathcal{A}$. Given a policy $\pi$, let $V^\pi$ and $\hat{V}^\pi$ be value functions of policy $\pi$ 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)

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

(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)

(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)

(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)

(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 $\varepsilon$ 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)

(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)

ALL THE BEST