AI3000/CS5500: Reinforcement Learning
Assignment 1

Taha Adeel Mohammed

CS20BTECH11052

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)

We can formulate a Markov Process with the states

$$\mathcal{S} = \{S_0, 1, 12, 123, 1234\},$$

where S_0 is the initial state and 1234 is the terminal state. $1, 12, 123, 1234$ represent having gotten the sequence $1, 12, 123, 1234$ on consecutive die rolls respectively. The states with their transition probabilities are shown in the figure below.

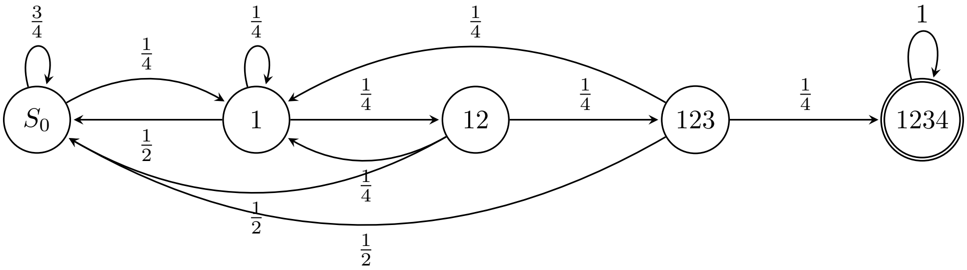


Figure 1: Markov Chain

The transition probabilities are also shown in the matrix below:

$$\mathcal{P} = \begin{array}{c} \\ \text{S0} \\ 1 \\ 12 \\ 123 \\ 1234 \end{array} \begin{array}{c} \begin{array}{ccccc} \text{S0} & 1 & 12 & 123 & 1234 \end{array} \\ \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ \frac{1}{2} & \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \tag{1}$$

(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)

We assign a reward of -1 to each of the states $S_0, 1, 12, 123$ and a reward of 0 to the terminal state 1234. i.e

$$\mathcal{R}(s) = \begin{cases} -1 & \text{if } s \in \{S_0, 1, 12, 123\} \\ 0 & \text{if } s = 1234 \end{cases} \tag{2}$$

We also set the discount factor $\gamma = 1$. This way, -$V(s)$ would represent the expected number of tosses required to reach the terminal state 1234 from state s. Hence our required answer of average number of tosses required for the pattern $'1234'$ to appear is $-V(S_0)$.

The Bellman equation for MRP is given by:

$$V(s) = \mathcal{R}(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s, s')V(s') \tag{3}$$

Therefore, we have:

$$\begin{align} V(S_0) &= -1 + \frac{3}{4}V(S_0) + \frac{1}{4}V(1) \tag{4} \\ V(1) &= -1 + \frac{1}{2}V(S_0) + \frac{1}{4}V(1) + \frac{1}{4}V(12) \tag{5} \\ V(12) &= -1 + \frac{1}{2}V(S_0) + \frac{1}{4}V(1) + \frac{1}{4}V(123) \tag{6} \\ V(123) &= -1 + \frac{1}{2}V(S_0) + \frac{1}{4}V(1) + \frac{1}{4}V(1234) \tag{7} \\ V(1234) &= 0 \tag{8} \end{align}$$

Solving the above equations, we get:

$$V(S_0) = -256, \quad V(1) = -252, \quad V(12) = -240, \quad V(123) = -192, \quad V(1234) = 0$$

Therefore, the average number of tosses required for the pattern $'1234'$ to appear is given by $-V(S_0) = \boxed{\text{256 tosses.}}$

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 $\mathcal{R}^a_{ss'}$ 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)

The discounted sum of rewards is given by:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{9}$$

Hence the least possible G_t theoretically obtainable, i.e. a lower bound on G_t is when for all $k \geq 0$, $r_{t+k+1} = r_{min} = min(\mathcal{R}^a_{ss'})$ $\forall$ $a \in \mathcal{A}$ and $s, s' \in \mathcal{S}$, which exists because the reward function is non-negative (and hence lower-bounded by 0). Therefore, we have:

$$\boxed{G_t \geq \sum_{k=0}^{\infty} \gamma^k r_{min} = \frac{r_{min}}{1 - \gamma}} \tag{10}$$

Similarly, the upper bound on G_t occurs when $\forall\ k \geq 0$, $r_{t+k+1} = r_{max} = max(\mathcal{R}^a_{ss'})$ $\forall\ a \in \mathcal{A}$ and $s, s' \in \mathcal{S}$, which exists because the reward function is bounded. Therefore, we have:

$$\boxed{G_t \leq \sum_{k=0}^{\infty} \gamma^k r_{max} = \frac{r_{max}}{1-\gamma}} \tag{11}$$

(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

$$\mathcal{R}^a_{ss'} - \hat{\mathcal{R}}^a_{ss'} = \epsilon,$$

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)

We know that the value function $V^\pi(s)$ is given by:

$$V^\pi(s) = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s \right] \tag{12}$$

Since the reward function $\mathcal{R}^a_{ss'}$ is modified by a constant ϵ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, using linearity of expectation, we have:

$$\begin{aligned} V^\pi(s) &= E_\pi \left[\sum_{k=0}^{\infty} \gamma^k \left(\hat{\mathcal{R}}^a_{ss'} + \epsilon \right) \middle| s_t = s \right] && (13) \\ &= E_\pi \left[\sum_{k=0}^{\infty} \gamma^k \hat{\mathcal{R}}^a_{ss'} \middle| s_t = s \right] + E_\pi \left[\sum_{k=0}^{\infty} \gamma^k \epsilon \middle| s_t = s \right] && (14) \\ &= E_\pi \left[\sum_{k=0}^{\infty} \gamma^k \hat{\mathcal{R}}^a_{ss'} \middle| s_t = s \right] + \epsilon \sum_{k=0}^{\infty} \gamma^k && (15) \\ &= \hat{V}^\pi(s) + \frac{\epsilon}{1-\gamma} && (16) \end{aligned}$$

Therefore we have:

$$\boxed{\implies V^\pi(s) - \hat{V}^\pi(s) = \frac{\epsilon}{1-\gamma}} \tag{17}$$

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

Yes, M and $\hat{M}$ have the same optimal policy. This is because the optimal policy π^* is the one that maximizes the value function $V^{\pi^*}(s)$ for all $s \in \mathcal{S}$. Since we have shown in sub-question (b) that $V^{\pi^*}(s)$ and $\hat{V}^{\pi^*}(s)$ are related by a constant, therefore the optimal policy is the same, as the same action at each state would maximize both value functions $V^{\pi^*}(s)$ and $\hat{V}^{\pi^*}(s)$.

Mathematically, we can see this using the policy iteration method. We start with an initial policy π_0 and iteratively update it until we arrive at the optimal policy π^*. In each update for M, we have:

$$\pi_{k+1}(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V^{\pi_k}(s')\right] \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

Similarly, for $\hat{M}$, we have the update:

$$\hat{\pi}_{k+1}(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left[\hat{\mathcal{R}}^a_{ss'} + \gamma \hat{V}^{\pi_k}(s')\right] \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

$$\implies \hat{\pi}_{k+1}(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma (V^{\pi_k}(s') - \frac{\epsilon}{1-\gamma})\right] \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

$$\implies \hat{\pi}_{k+1}(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V^{\pi_k}(s')\right] \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

$$\implies \hat{\pi}_{k+1}(a|s) = \pi_{k+1}(a|s) \tag{22}$$

Hence we can see that after each iteration of the policy iteration method, the policy π_{k+1} for M and $\hat{\pi}_{k+1}$ for $\hat{M}$ are the same. Therefore, the optimal policy π^* for M and $\hat{\pi}^*$ for $\hat{M}$ are also the same.

(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, the assumption that the MDP M in sub-question (a) has non-negative and bounded reward is without loss in generality, as we have shown in sub-question (b) that the optimal policy is the same for both MDPs M and $\hat{M}$, even if the reward function of M is modified by a constant ϵ. Hence for reward functions with negative but bounded rewards, we can simply add a constant ϵ to the reward function to make it non-negative and bounded, and the optimal policy would still be the 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)

The value function for a finite horizon MDP M is given by:

$$V^\pi(s) = E_\pi \left[\sum_{k=0}^{H} \gamma^k r_{t+k+1} \middle| s_t = s \right] \tag{23}$$

Similarly to sub-question (b), since the reward function $\mathcal{R}^a_{ss'}$ is modified by a constant ϵ

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have:

$$V^{\pi}(s) = E_{\pi}\left[\sum_{k=0}^{H} \gamma^k \left(\hat{\mathcal{R}}^a_{ss'} + \epsilon\right) \middle| s_t = s\right] \tag{24}$$

$$= E_{\pi}\left[\sum_{k=0}^{H} \gamma^k \hat{\mathcal{R}}^a_{ss'} \middle| s_t = s\right] + E_{\pi}\left[\sum_{k=0}^{H} \gamma^k \epsilon \middle| s_t = s\right] \tag{25}$$

$$= E_{\pi}\left[\sum_{k=0}^{H} \gamma^k \hat{\mathcal{R}}^a_{ss'} \middle| s_t = s\right] + \epsilon \sum_{k=0}^{H} \gamma^k \tag{26}$$

$$= \hat{V}^{\pi}(s) + \frac{\epsilon\left(1 - \gamma^{H+1}\right)}{1 - \gamma} \tag{27}$$

Therefore we have:

$$\boxed{\implies V^{\pi}(s) - \hat{V}^{\pi}(s) = \frac{\epsilon\left(1 - \gamma^{H+1}\right)}{1 - \gamma}} \tag{28}$$

(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 $S_{term} \subset \mathcal{S}$ 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 $S_H \in \mathcal{S}$ 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)

For an indefinite MDP, with a varying horizon length H, the value function is given by:

$$V^{\pi}(s) = E_{\pi}\left[\sum_{k=0}^{H} \gamma^k r_{t+k+1} \middle| s_t = s\right] \tag{29}$$

Since the reward function $\mathcal{R}^a_{ss'}$ is modified by a constant ϵ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have:

$$V^{\pi}(s) = E_{\pi}\left[\sum_{k=0}^{H} \gamma^k \left(\hat{\mathcal{R}}^a_{ss'} + \epsilon\right) \middle| s_t = s\right] \tag{30}$$

$$= E_{\pi}\left[\sum_{k=0}^{H} \gamma^k \hat{\mathcal{R}}^a_{ss'} \middle| s_t = s\right] + E_{\pi}\left[\sum_{k=0}^{H} \gamma^k \epsilon \middle| s_t = s\right] \tag{31}$$

$$= E_{\pi}\left[\sum_{k=0}^{H} \gamma^k \hat{\mathcal{R}}^a_{ss'} \middle| s_t = s\right] + \epsilon \sum_{k=0}^{H} \gamma^k \tag{32}$$

$$= \hat{V}^{\pi}(s) + \frac{\epsilon\left(1 - \gamma^{H+1}\right)}{1 - \gamma} \tag{33}$$

Therefore we have:

$$\boxed{\implies V^{\pi}(s) - \hat{V}^{\pi}(s) = \frac{\epsilon\left(1 - \gamma^{H+1}\right)}{1 - \gamma}} \tag{34}$$

Therefore no, the analogous result of sub-question (b) does not hold when M and $\hat{M}$ are indefinite MDPs, as the horizon length H is not fixed, and hence the value functions $V^{\pi}(s)$ and $\hat{V}^{\pi}(s)$ are not related by a constant.

(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**

$$\mathcal{R}^a_{ss'} - \hat{\mathcal{R}}^a_{ss'} \leq \epsilon$$

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

Similar to sub-question (b), we claim that $V_*(s)$ and $\hat{V}_*(s)$ are related by:

$$V_*(s) - \hat{V}_*(s) \leq \frac{\epsilon}{1-\gamma} \tag{35}$$

Proof by Induction: (Value Iteration Method)
We start with an initial value function $V_0(s) = 0$ and $\hat{V}_0(s) = 0 \ \forall \ s \in \mathcal{S}$. Then we iteratively update these value functions using the Value Iteration Method until we arrive at the optimal value functions $V_*(s)$ and $\hat{V}_*(s)$.

Base Case: $V_0(s) - \hat{V}_0(s) = 0 \leq \frac{\epsilon}{1-\gamma} \ \forall \ s \in \mathcal{S}$.

Induction Hypothesis: $V_k(s) - \hat{V}_k(s) \leq \frac{\epsilon}{1-\gamma} \ \forall \ s \in \mathcal{S}$, where V_k and $\hat{V}_k$ are the value functions after k iterations of the Value Iteration Method.

Induction Step: We know that the value functions $V_{k+1}(s)$ and $\hat{V}_{k+1}(s)$ are given by:

$$V_{k+1}(s) = \max_{a\in\mathcal{A}} \left[\sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma V_k(s') \right) \right] \tag{36}$$

$$\hat{V}_{k+1}(s) = \max_{a\in\mathcal{A}} \left[\sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\hat{\mathcal{R}}^a_{ss'} + \gamma \hat{V}_k(s') \right) \right] \tag{37}$$

Note that these two equations might be maximized by different actions $a_1, a_2 \in \mathcal{A}$. However

$$\max_{a\in\mathcal{A}} f(a) - \max_{a\in\mathcal{A}} g(a) \leq \max_{a\in\mathcal{A}} \left[f(a) - g(a) \right] \tag{38}$$

Therefore by subracting the two equations, and utilising above inequality and our induction hypothesis, we get:

$$
\begin{aligned}
V_{k+1}(s) - \hat{V}_{k+1}(s) &\leq \max_{a\in\mathcal{A}} \left[\sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma V_k(s')\right) \right] - \max_{a\in\mathcal{A}} \left[\sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\hat{\mathcal{R}}^a_{ss'} + \gamma \hat{V}_k(s')\right) \right] \\
&\leq \max_{a\in\mathcal{A}} \left[\sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma V_k(s')\right) - \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\hat{\mathcal{R}}^a_{ss'} + \gamma \hat{V}_k(s')\right) \right] \\
&\leq \max_{a\in\mathcal{A}} \left[\sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} - \hat{\mathcal{R}}^a_{ss'} + \gamma \left(V_k(s') - \hat{V}_k(s')\right)\right) \right] \\
&\leq \max_{a\in\mathcal{A}} \left[\sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\epsilon + \gamma \left(V_k(s') - \hat{V}_k(s')\right)\right) \right] \\
&\leq \max_{a\in\mathcal{A}} \left[\sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\epsilon + \gamma \frac{\epsilon}{1-\gamma}\right) \right] \\
&\leq \max_{a\in\mathcal{A}} \left[\sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'} \left(\frac{\epsilon}{1-\gamma}\right) \right] = \frac{\epsilon}{1-\gamma}
\end{aligned}
$$

Therefore we have:

$$\boxed{\implies V_{k+1}(s) - \hat{V}_{k+1}(s) \leq \frac{\epsilon}{1-\gamma}} \tag{39}$$

Hence by by the principle of mathematical induction, we have:

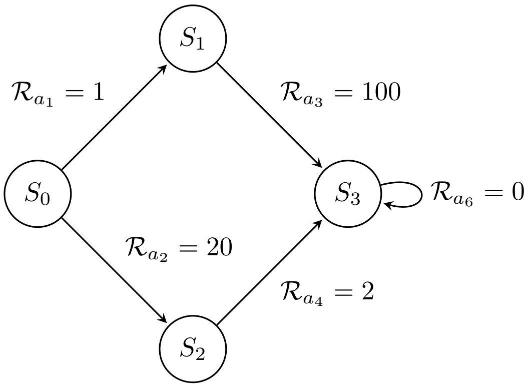
$$\boxed{\implies V_*(s) - \hat{V}_*(s) \leq \frac{\epsilon}{1-\gamma}} \tag{40}$$

Optimal Policy for M and $\hat{M}$:
M and $\hat{M}$ need **not** have the same optimal policy. The modification in the reward function may affect the optimal policy, especially if the original optimal policy was sensitive to the exact reward values or the ϵ value is huge. Also in above calculations, we have seen that the optimal value functions can have different actions $a_1, a_2 \in \mathcal{A}$ that maximize them. Hence the optimal policy for M and $\hat{M}$ may be different.

(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)

Yes, scaling the discount factor by a constant $\kappa \in (0, 1)$ might alter the optimal policy. For example, it could convert an agent from a far-sighted agent to a myopic agent, or vice-versa.

For example, consider the below infinite horizon MDP M, with the original discount factor $\gamma = 0.5$. The image below shows the states of the MDP, with each arrow representing an action that will deterministically take them to the next state, and the reward $\mathcal{R}(s, a, s')$ associated with the transition shown on the arrow.



For $\gamma = 0.5$, for policy $\pi_1 = \{a_1, a_3, a_6, a_6, \cdots\}$ and $\pi_2 = \{a_2, a_4, a_6, a_6, \cdots\}$, we have:

$$\begin{aligned} Q^{\pi_1}(S_0, a_1) &= 1 + 0.5 \times 100 + 0 + \ldots = 51 \\ Q^{\pi_1}(S_0, a_2) &= 20 + 0.5 \times 2 + 0 + \ldots = 21 \\ \therefore \pi^* = \pi_1 &= \underset{a \in \mathcal{A}}{\operatorname{argmax}}\ Q^{\pi_1}(S_0, a) = a_1 \end{aligned}$$

However, if we scale gamma by $\kappa = 0^+$, i.e. $\gamma' \to 0$, then for policy $\pi_1 = \{a_1, a_3, a_6, a_6, \cdots\}$ and $\pi_2 = \{a_2, a_4, a_6, a_6, \cdots\}$, we have:

$$\begin{aligned} Q^{\pi_1}(S_0, a_1) &= 1 + 0 \times 100 + 0 + \ldots = 1 \\ Q^{\pi_1}(S_0, a_2) &= 20 + 0 \times 2 + 0 + \ldots = 20 \\ \therefore \pi^* = \pi_2 &= \underset{a \in \mathcal{A}}{\operatorname{argmax}}\ Q^{\pi_2}(S_0, a) = a_2 \end{aligned}$$

Therefore, scaling the discount factor by a constant $\kappa \in (0, 1)$ can alter the optimal policy.