Assignment-1

Pradeep Mundlik (AI21BTECH11022)

September 5, 2023

Problem 1: Markov Reward Process

a. **Solution:**
States:

- State 0: No progress towards the target pattern.
- State 1: Rolled '1' so far.
- State 2: Rolled '12' so far.
- State 3: Rolled '123' so far.
- State 4: Pattern '1234' achieved. (Terminal State)

Transition probabilities between states:

- From State 0:
  - Transitions to State 1 with probability $\frac{1}{4}$ (rolling '1').
  - Transitions to State 0 with probability $\frac{3}{4}$ (rolling '2', '3', or '4').
- From States 1, 2, and 3:
  - Transitions to the next state with probability $\frac{1}{4}$.
- From State 4:
  - Remains in State 4 with probability 1 (pattern is already achieved).

b. **Solution:**
Reward Function: We will use a reward of 0 for the terminal state (State 4), and a reward of 1 for all other states.

$$R(s) = \begin{cases} 0 & \text{s = 4} \\ 1 & \text{s = 0,1,2,3} \end{cases}$$

Discount Factor: Since there's no mention of discounting, we'll assume a discount factor of 1. ($\gamma = 1$)

Bellman Equation: The Bellman equation for each state s can be written as follows:

$$V(s) = R(s) + \sum_{s'} P(s'|s) \cdot V(s')$$

Where $R(s)$ is reward for s, $P(s'|s)$ is transition probability from state s to s', and $V(s)$ is the expected cumulative reward starting from state s.

Solving the Bellman Equations: Let's calculate the expected number of tosses starting from State 0 (V(0)) using the Bellman equation:

$$\begin{aligned}
V(0) &= 1 + \frac{3}{4}V(0) + \frac{1}{4}V(1) \\
V(1) &= 1 + \frac{1}{2}V(0) + \frac{1}{4}V(1) + \frac{1}{4}V(2) \\
V(2) &= 1 + \frac{1}{2}V(0) + \frac{1}{4}V(1) + \frac{1}{4}V(3) \\
V(3) &= 1 + \frac{1}{2}V(0) + \frac{1}{4}V(1) + \frac{1}{4}V(4) \\
V(4) &= 0
\end{aligned}$$

Solving above system of equations gives us, $V(0) = 64$.
Expected number of tosses required for the pattern '1234' to appear is 64.

Problem 2: Markov Decision Process

a. **Solution:**
In an infinite horizon Markov Decision Process (MDP), the discounted sum of rewards is calculated using the following formula:

$$V(s) = E\left[\left(\sum \gamma^t R(s_t, a_t, s_{t+1})\right) | s_0 = s\right]$$

Where γ is discount factor, R is reward for transition from s_t to s_{t+1} by taking action a_t.
Since the reward function $R(s, a, s0)$ is non-negative and bounded, let's denote its upper bound as M and lower bound as m.
Lower Bound:

$$V_{Lower} = E\left[\left(\sum \gamma^t m\right) | s_0 = s\right] = \frac{m}{1-\gamma}$$

Upper Bound:

$$V_{Upper} = E\left[\left(\sum \gamma^t M\right) | s_0 = s\right] = \frac{M}{1-\gamma}$$

b. **Solution:**
We have two infinite horizon MDPs: $M =< S, A, P, R, \gamma >$ and $\hat{M} =< S, A, P, \hat{R}, \gamma >$, with reward functions R and $\hat{R}$, respectively. The relationship between these reward functions is given as:

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

The value funtion $V_n^{\pi}(s)$ for M can be defined as:

$$V_n^{\pi}(s) = E\left[\left(\sum \gamma^t R(s_t, a_t, s_{t+1})\right) | s_0 = s\right]$$

The value funtion $\hat{V}_n^{\pi}(s)$ for M can be defined as:

$$\hat{V}_n^{\pi}(s) = E\left[\left(\sum \gamma^t \hat{R}(s_t, a_t, s_{t+1})\right) | s_0 = s\right]$$

Now, we can factor out ϵ,

$$\begin{aligned}
V_n^{\pi}(s) &= E\left[\left(\sum \gamma^t R(s_t, a_t, s_{t+1})\right) | s_0 = s\right] \\
V_n^{\pi}(s) &= E\left[\left(\sum \gamma^t (\hat{R}(s_t, a_t, s_{t+1}) + \epsilon)\right) | s_0 = s\right] \\
V_n^{\pi}(s) &= E\left[\left(\sum \gamma^t \hat{R}(s_t, a_t, s_{t+1})\right) | s_0 = s\right] + \sum \gamma^t \epsilon \\
V_n^{\pi}(s) &= \hat{V}_n^{\pi}(s) + \frac{\epsilon}{1-\gamma}
\end{aligned}$$

This expression shows that the difference between the value functions $V_n^{\pi}(s)$ and $\hat{V}_n^{\pi}(s)$ is proportional to the constant ϵ and inversely proportional to $(1-\gamma)$, the discount factor.

c. **Solution:**
From result in above question, we have

$$\Delta V = V_n^{\pi}(s) - \hat{V}_n^{\pi}(s) = \frac{\epsilon}{1-\gamma}$$

Here we can see that ϵ and γ are constants. Therefore, ΔV is also constant.

Now consider, $\pi*$ is optimal policy for M.

$$\implies V_n^{\pi*}(s) \geq V_n^{\pi}(s) \text{ for all } s \text{ and policies}$$

Adding ΔV in above inequality,

$$\begin{aligned}
V_n^{\pi*}(s) + \Delta V &\geq V_n^{\pi}(s) + \Delta V \\
\hat{V}_n^{\pi*}(s) &\geq \hat{V}_n^{\pi}(s)
\end{aligned}$$

Above inequality holds for all s and policies, therefore $\pi*$ is optimal policy for $\hat{M}$ as well.

So, we can conclude that M and $\hat{M}$ have same optimal policies (for constant ϵ).

d. **Solution:**
If the MDP M is allowed to have negative but bounded rewards, the same derivation can still be applied, and the expression relating the value functions $V^{\pi}(s)$ and $\hat{V}^{\pi}(s)$ remains valid. The constant ϵ added to the reward function in $\hat{M}$ still plays the same role in the analysis. The key point is that the ϵ modification to the reward function is independent of whether the original rewards are non-negative or include negative values, as long as they are bounded. The presence of negative rewards can make the optimization problem more challenging, but the relationship between $V^{\pi}(s)$ and $\hat{V}^{\pi}(s)$ as derived in sub-question (b) remains applicable.
So, it is not always necessary, and the analysis can be extended to MDPs with negative but bounded rewards.

e. **Solution:**
Let's consider the case of finite horizon MDPs with length H and from question (b), we have

$$V_n^\pi(s) = E\left[\left(\sum_{t=0}^{t=H-1} \gamma^t R(s_t, a_t, s_{t+1})\right)|s_0 = s\right]$$

$$\hat{V}_n^\pi(s) = E\left[\left(\sum_{t=0}^{t=H-1} \gamma^t \hat{R}(s_t, a_t, s_{t+1})\right)|s_0 = s\right]$$

$$\hat{V}_n^\pi(s) = E\left[\left(\sum_{t=0}^{t=H-1} \gamma^t (\hat{R}(s_t, a_t, s_{t+1}) - \epsilon)\right)|s_0 = s\right]$$

$$\hat{V}_n^\pi(s) = E\left[\left(\sum_{t=0}^{t=H-1} \gamma^t R(s_t, a_t, s_{t+1})\right)|s_0 = s\right] - E\left[\sum_{t=0}^{t=H-1} \gamma^t * \epsilon|s_0 = s\right]$$

$$\hat{V}_n^\pi(s) = V_n^\pi(s) - E\left[\sum_{t=0}^{t=H-1} \gamma^t * \epsilon|s_0 = s\right]$$

$$\Delta V_n^\pi(s) = E\left[\sum_{t=0}^{t=H-1} \gamma^t * \epsilon|s_0 = s\right]$$

$$\Delta V_n^\pi(s) = \epsilon * \left(\frac{1-\gamma^H}{1-\gamma}\right)$$

f. **Solution:**
In indefinite MDPs or stochastic shortest path MDPs where the length of the episode (horizon length H) can vary, the analogous result of sub-question (b) may not hold in the same straightforward manner as in the finite or infinite horizon cases. In these MDPs, the length of the episode (H) is not fixed but rather a random variable. The duration of an episode can vary based on the stochastic nature of the environment or problem. The episodes can have different lengths depending on when a terminal state is encountered.

g. **Solution:**
Bellman equation of optimal value funtion for M is,

$$V_*(s) = \max_a Q_*(s, a)$$

$$V_*(s) = \max_a \left[\sum_{s'} P_{ss'}^a \left(R_{ss'}^a + \gamma V_*(s')\right)\right]$$

for $\hat{M}$ is,

$$\hat{V}_*(s) = \max_a \left[\sum_{s'} P_{ss'}^a \left(\hat{R}_{ss'}^a + \gamma \hat{V}_*(s')\right)\right]$$

Let's take the absolute difference between these two equations,

$$\left|V_*(s) - \hat{V}_*(s)\right| = \left|\max_a \left[\sum_{s'} P^a_{ss'}\left(R^a_{ss'} + \gamma V_*(s')\right)\right] - \max_a \left[\sum_{s'} P^a_{ss'}\left(\hat{R}^a_{ss'} + \gamma \hat{V}_*(s')\right)\right]\right|$$

$$\left|V_*(s) - \hat{V}_*(s)\right| = \left|\max_a \left[\sum_{s'} P^a_{ss'}\left(\left(R^a_{ss'} - \hat{R}^a_{ss'}\right) + \gamma\left(V_*(s') - \hat{V}_*(s')\right)\right)\right]\right|$$

$$\left|V_*(s) - \hat{V}_*(s)\right| \leq \max_a \sum_{s'} P^a_{ss'}\left[\left|R^a_{ss'} - \hat{R}^a_{ss'}\right| + \gamma\left|V_*(s') - \hat{V}_*(s')\right|\right]$$

Now, we can use fact that optimal functions are bounded, let U is upper bound fot both $V_*(s)$ and $\hat{V}_*(s)$ for all states.
Also, we given that $R(s, a, s_0) - \hat{R}(s, a, s_0) \leq \epsilon$,

$$\left|V_*(s) - \hat{V}_*(s)\right| \leq \max_a \sum_{s'} P^a_{ss'}\left(\epsilon + \gamma * 2 * U\right)$$

$$\left|V_*(s) - \hat{V}_*(s)\right| \leq \max_a \sum_{s'} P^a_{ss'}\left(\epsilon + 2\gamma U\right)$$

The difference between the optimal value functions depends on ϵ, the bound U, the discount factor γ, and the transition probabilities. Policies are derived from the value functions, so optimal policies for the MDPs will depend on how large or small above factors are.

h. **Solution:**
As, we are given that $0 < \kappa < 1$,
So, $\kappa\gamma < \gamma$,

In an MDP, discount factor(γ) represents importance of future rewards vs immediate rewards. A smaller γ places more emphasis on immediate rewards, while a larger γ places more emphasis on long-term rewards. Therefore, changing the discount factor from γ to $\kappa\gamma$ effectively changes the balance between immediate and future rewards. This will lead to less emphasis on future rewards and more emphasis on immediate rewards. The agent will become more "impatient" and may prioritize actions that provide immediate rewards, even if they result in lower long-term cumulative rewards. The optimal policy may change in favor of actions that have better immediate rewards, as the agent discounts the future more heavily.
So, scaling the discount factor in an MDP does alter the optimal policy because it changes the trade-off between immediate and future rewards.