Assignment 1

PENG Qiheng

Student ID 225040065

October 16, 2025

Problem 1.

1. There are two possible policies π_1 and π_2, while:

$$\pi_1(s_0) = a_1, \pi_1(s_1) = a_0, \pi_1(s_2) = a_0, \pi_1(s_3) = a_0$$
$$\pi_2(s_0) = a_2, \pi_2(s_1) = a_0, \pi_2(s_2) = a_0, \pi_2(s_3) = a_0$$

2. The optimal value function for each state is:

$$V^*(s_0) = \max\{\gamma V^*(s_1), \gamma V^*(s_2)\}$$
$$V^*(s_1) = \gamma(pV^*(s_3) + (1-p)V^*(s_1))$$
$$V^*(s_2) = 1 + \gamma(qV^*(s_3) + (1-q)V^*(s_0))$$
$$V^*(s_3) = 10 + \gamma V^*(s_0)$$

3. **Yes**. When $p = 0$, we have:

$$V^*(s_1) = \gamma V^*(s_1) \Rightarrow V^*(s_1) = 0$$
$$V^*(s_2) = 1 + \gamma(qV^*(s_3) + (1-q)V^*(s_0)) \geq 1$$
$$V^*(s_0) = \max\{\gamma V^*(s_1), \gamma V^*(s_2)\} = \gamma V^*(s_2)$$

   Thus, in this case, $\forall \gamma \in [0, 1)$ and $q \in [0, 1], \pi^*(s_0) = a_2$

4. **No**. When $p = 1$ and $\gamma < \frac{1}{V^*(s_3)}$, we have:

$$V^*(s_1) = \gamma V^*(s_3) < 1$$
$$V^*(s_2) = 1 + \gamma(qV^*(s_3) + (1-q)V^*(s_0)) \geq 1$$
$$V^*(s_0) = \max\{\gamma V^*(s_1), \gamma V^*(s_2)\} = \gamma V^*(s_2)$$

   Thus, in this case, $\forall q \in [0, 1], \pi^*(s_0) = a_2$.

Problem 2.

1. The discount factor γ balances the short-term and long-term rewards. And with $\gamma < 1$, it can guarantee that an optimal value function $V^*(s)$ exists and is finite for all states $s \in \mathcal{S}$

2. While SSPs exists a finite path from s to s_G and the solution policy in SSPs should instead minimize the expected cost to reach a goal state, it does not need a discount factor to ensure that an optimal value function exists and is finite for all states.

3. The optimal value function for SSPs is:

$$V^*(s) = \begin{cases} \min_{a \in A} \left\{ C(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s') \right\}, & s \notin G \\ 0, & s \in G \end{cases}$$

4. With an MDP $\mathcal{M} =< S, s_0, A, C, P >$ with discount factor $\gamma \in [0, 1)$, we can construct an SSP $\mathcal{S} =< S', s_0, G, A, C', P' >$ as follows:

$$S' = S \cup G$$

$s_0 \in S$ is the initial state

G is the set of goal states

A is the finite action space

$$C' = \begin{cases} C(s,a), & s \notin G \\ 0, & s \in G \end{cases}$$

$$P'(s'|s,a) = \begin{cases} \alpha P(s'|s,a), & s \notin G, s' \notin G \\ 1 - \alpha, & s \notin G, s' \in G \\ 0, & s \in G, s' \notin G \\ 1, & s \in G, s' \in G \end{cases}$$

Problem 3.

1. A_2 **and** A_4. It's easy to calculate that:

$$\hat{\mu}_0(1) = 0, \hat{\mu}_0(2) = 0, \hat{\mu}_0(3) = 0, \hat{\mu}_0(4) = 0$$
$$\hat{\mu}_1(1) = 3, \hat{\mu}_1(2) = 0, \hat{\mu}_1(3) = 0, \hat{\mu}_1(4) = 0$$
$$\hat{\mu}_2(1) = 3, \hat{\mu}_2(2) = 2, \hat{\mu}_2(3) = 0, \hat{\mu}_2(4) = 0$$
$$\hat{\mu}_3(1) = 2, \hat{\mu}_3(2) = 2, \hat{\mu}_3(3) = 0, \hat{\mu}_3(4) = 0$$
$$\hat{\mu}_4(1) = 2, \hat{\mu}_4(2) = 2, \hat{\mu}_4(3) = 1, \hat{\mu}_4(4) = 0$$

So we have:

$$\text{argmax}\mu_0 = \{1, 2, 3, 4\}, \quad \text{argmax}\mu_1 = 1,$$
$$\text{argmax}\mu_2 = 1, \quad \text{argmax}\mu_3 \in \{1, 2\}, \quad \text{argmax}\mu_4 \in \{1, 2\}$$
$$A_2 = 2 \neq \text{argmax}\mu_1, \quad A_4 = 3 \neq \text{argmax}\mu_3$$

Thus, A_2 and A_4 are definitely exploratory.

2. As the calculation above, we have:

$$A_1 = 1 = \text{argmax}\mu_0, \quad A_3 = 1 = \text{argmax}\mu_2, \quad A_5 = 2 = \text{argmax}\mu_4$$

Thus, A_1, A_3 and A_5 are possibly exploratory.

Problem 4.

1. Follow the hint, we have:

$$|\hat{\mu}_1(k-1) - \hat{\mu}_2(k-1)| \leq 2w(k-1) = 2\sqrt{\frac{2\log(T)}{k-1}}$$

   By using the reverse triangle inequality, we have:

$$\begin{aligned}
\Delta &= |\mu_1 - \mu_2| \\
&\leq |\hat{\mu}_1(k-1) - \hat{\mu}_2(k-1)| + |\hat{\mu}_1(k-1) - \mu_1| + |\hat{\mu}_2(k-1) - \mu_2| \\
&\leq 2\sqrt{\frac{2\log(T)}{k-1}} + |\hat{\mu}_1(k-1) - \mu_1| + |\hat{\mu}_2(k-1) - \mu_2|
\end{aligned}$$

   By using the *Chebyshev's Inequality*, we have:

$$P\left(|\hat{\mu}_1(k-1) - \mu_1| \leq \sqrt{\frac{2\log(T)}{k-1}}\right) \geq 1 - \frac{1}{2\log(T)}$$
$$P\left(|\hat{\mu}_2(k-1) - \mu_2| \leq \sqrt{\frac{2\log(T)}{k-1}}\right) \geq 1 - \frac{1}{2\log(T)}$$

   Thus, with the probability of $1 - \frac{1}{2\log(T)}$ (with a large T, the probability will be very high), we have:

$$\begin{gathered}
\Delta \leq 2\sqrt{\frac{2\log(T)}{k-1}} + 2\sqrt{\frac{2\log(T)}{k-1}} = 4\sqrt{\frac{2\log(T)}{k-1}} \\
\Rightarrow \Delta \leq 4\sqrt{\frac{2\log(T)}{k-1}} \\
\Rightarrow k \leq \frac{32\log(T)}{\Delta^2} + 1
\end{gathered}$$

2. Using result from part (1), we have:

$$\bar{R}_T = K \cdot \Delta \leq \Delta + \frac{32\log(T)}{\Delta}$$