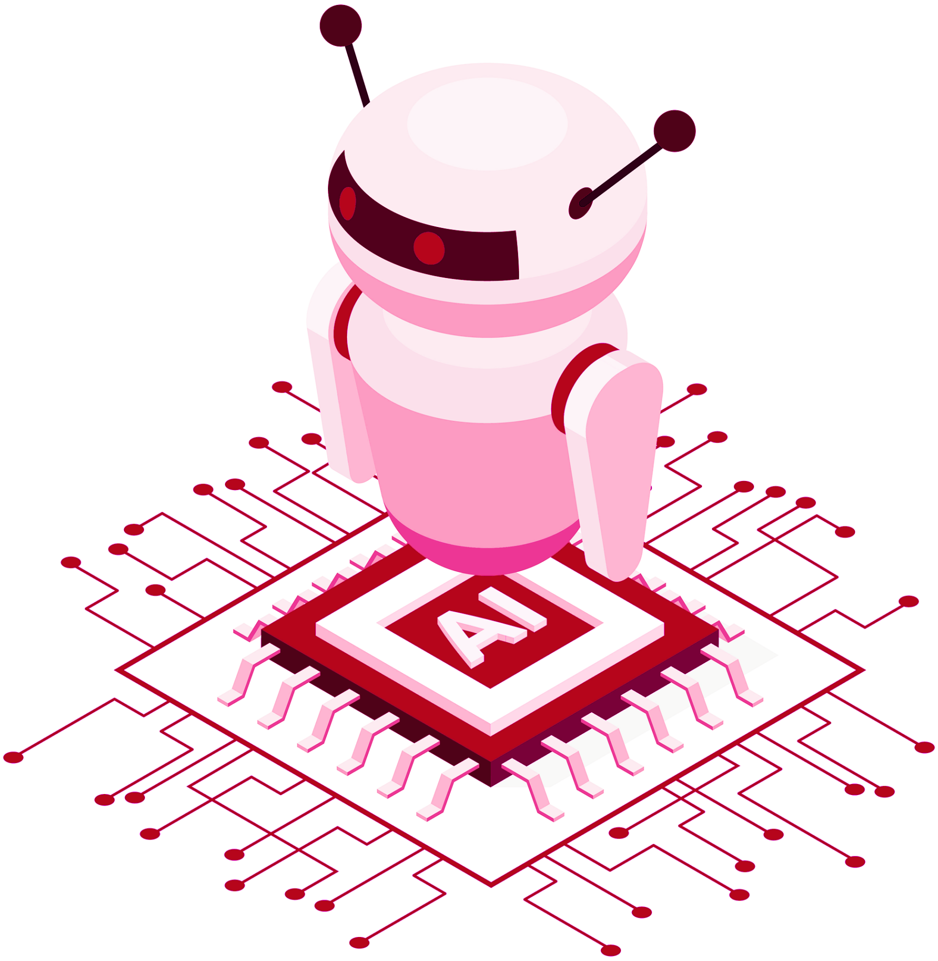

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution for Homework 11:

Imitation Learning and Inverse RL

By:

Payam Taebi

400104867

Spring 2025

Contents

1 Distribution Shift and Performance Bounds

1.1 Task 1: Distribution Shift Bound

Show that the total variation distance between state distributions induced by the learned policy and the expert satisfies:

$$\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2T\varepsilon.$$

Proof. Let

$$\Delta_t \;=\; \sum_{s_t} \bigl|p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)\bigr|$$

be the total variation distance between the state distributions at time t. We will show by induction that

$$\Delta_t \;\leq\; 2t\,\varepsilon.$$

Base case ($t = 0$). By assumption both policies start from the same initial state distribution, so

$$\Delta_0 \;=\; \sum_{s_0} \bigl|p_{\pi_\theta}(s_0) - p_{\pi^*}(s_0)\bigr| = 0 \;\leq\; 0 = 2 \cdot 0 \cdot \varepsilon.$$

Inductive step. Assume $\Delta_{t-1} \leq 2(t-1)\varepsilon$ for some $t \geq 1$. Then

$$p_{\pi_\theta}(s_t) = \sum_{s_{t-1}, a_{t-1}} p_{\pi_\theta}(s_{t-1})\, \pi_\theta(a_{t-1} \mid s_{t-1})\, P(s_t \mid s_{t-1}, a_{t-1}),$$
$$p_{\pi^*}(s_t) = \sum_{s_{t-1}, a_{t-1}} p_{\pi^*}(s_{t-1})\, \pi^*(a_{t-1} \mid s_{t-1})\, P(s_t \mid s_{t-1}, a_{t-1}).$$

Thus

$$\begin{aligned}
\Delta_t &= \sum_{s_t} \Bigl| \sum_{s_{t-1}, a_{t-1}} P(s_t \mid s_{t-1}, a_{t-1}) \Bigl[p_{\pi_\theta}(s_{t-1})\, \pi_\theta(a_{t-1} \mid s_{t-1}) - p_{\pi^*}(s_{t-1})\, \pi^*(a_{t-1} \mid s_{t-1}) \Bigr] \Bigr| \\
&\leq \sum_{s_{t-1}, a_{t-1}} \Bigl| p_{\pi_\theta}(s_{t-1})\, \pi_\theta(a_{t-1} \mid s_{t-1}) - p_{\pi^*}(s_{t-1})\, \pi^*(a_{t-1} \mid s_{t-1}) \Bigr| \quad \text{(by Jensen's inequality, since } P \text{ sums to 1)} \\
&\leq \underbrace{\sum_{s_{t-1}, a_{t-1}} \bigl| p_{\pi_\theta}(s_{t-1}) - p_{\pi^*}(s_{t-1}) \bigr|\, \pi_\theta(a_{t-1} \mid s_{t-1})}_{A} + \underbrace{\sum_{s_{t-1}, a_{t-1}} p_{\pi^*}(s_{t-1})\, \bigl| \pi_\theta(a_{t-1} \mid s_{t-1}) - \pi^*(a_{t-1} \mid s_{t-1}) \bigr|}_{B}.
\end{aligned}$$

We bound each term separately:

$$A = \sum_{s_{t-1}} \bigl| p_{\pi_\theta}(s_{t-1}) - p_{\pi^*}(s_{t-1}) \bigr| \sum_{a_{t-1}} \pi_\theta(a_{t-1} \mid s_{t-1}) = \Delta_{t-1},$$

and using the fact that for each s, $\sum_a |\pi_\theta(a \mid s) - \pi^*(a \mid s)| \leq 2\varepsilon$ by definition of the error ε,

$$B = \sum_{s_{t-1}} p_{\pi^*}(s_{t-1}) \sum_{a_{t-1}} \bigl| \pi_\theta(a_{t-1} \mid s_{t-1}) - \pi^*(a_{t-1} \mid s_{t-1}) \bigr| \leq \sum_{s_{t-1}} p_{\pi^*}(s_{t-1})\, 2\varepsilon = 2\varepsilon.$$

Putting these together,

$$\Delta_t \;\le\; \Delta_{t-1} + 2\varepsilon \;\le\; 2(t-1)\varepsilon + 2\varepsilon \;=\; 2t\,\varepsilon,$$

where the second inequality uses the inductive hypothesis. This completes the induction.

Therefore, for any $t \le T$,

$$\sum_{s_t} \big| p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t) \big| = \Delta_t \le 2t\,\varepsilon \le 2T\,\varepsilon.$$

□

1.2 Task 2: Return Gap for Terminal Rewards

Assume that the reward is only received at the final step (i.e., $r(s_t) = 0$ for all $t < T$). Show that:

$$J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T\varepsilon).$$

Proof. Since rewards are only at the terminal step, the return under policy π is

$$J(\pi) \;=\; \sum_{s_T} p_\pi(s_T)\, r(s_T).$$

Thus the gap in returns between expert and learned policy is

$$J(\pi^*) - J(\pi_\theta) = \sum_{s_T} \big(p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T) \big)\, r(s_T).$$

Taking absolute value and using that $|r(s_T)| \le R_{\max}$, we have

$$\big| J(\pi^*) - J(\pi_\theta) \big| \;\le\; R_{\max} \sum_{s_T} \big| p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T) \big| \;\le\; R_{\max}\, 2T\,\varepsilon \;=\; \mathcal{O}(T\varepsilon),$$

where the last inequality uses the Distribution Shift Bound from Task 1. Hence

$$J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T\varepsilon).$$

□

1.3 Task 3: Return Gap for General Rewards

For a general reward function (i.e., $r(s_t) \ne 0$ for arbitrary t), show that:

$$J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T^2\varepsilon).$$

Proof. The expected return under policy π is

$$J(\pi) = \sum_{t=0}^{T} \sum_{s_t} p_\pi(s_t)\, r(s_t).$$

Hence the gap in returns is

$$
\begin{aligned}
J(\pi^*) - J(\pi_\theta) &= \sum_{t=0}^{T} \sum_{s_t} \big(p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)\big)\, r(s_t) \\
&\le \sum_{t=0}^{T} \sum_{s_t} \big|p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)\big|\, \big|r(s_t)\big|.
\end{aligned}
$$

Let $R_{\max} = \max_{t,s} |r(s_t)|$. Then using the Distribution Shift Bound from Task 1,

$$
\sum_{s_t} \big|p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)\big| \;\le\; 2\,t\,\varepsilon.
$$

Substituting,

$$
\begin{aligned}
J(\pi^*) - J(\pi_\theta) \le \sum_{t=0}^{T} R_{\max}\, 2t\,\varepsilon &= 2R_{\max}\,\varepsilon \sum_{t=0}^{T} t = 2R_{\max}\,\varepsilon\, \frac{T(T+1)}{2} \\
&= R_{\max}\,\varepsilon\, T(T+1) = \mathcal{O}(T^2\varepsilon).
\end{aligned}
$$

Thus

$$
J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T^2\varepsilon).
$$

□

References

[1] Cover image designed by freepik