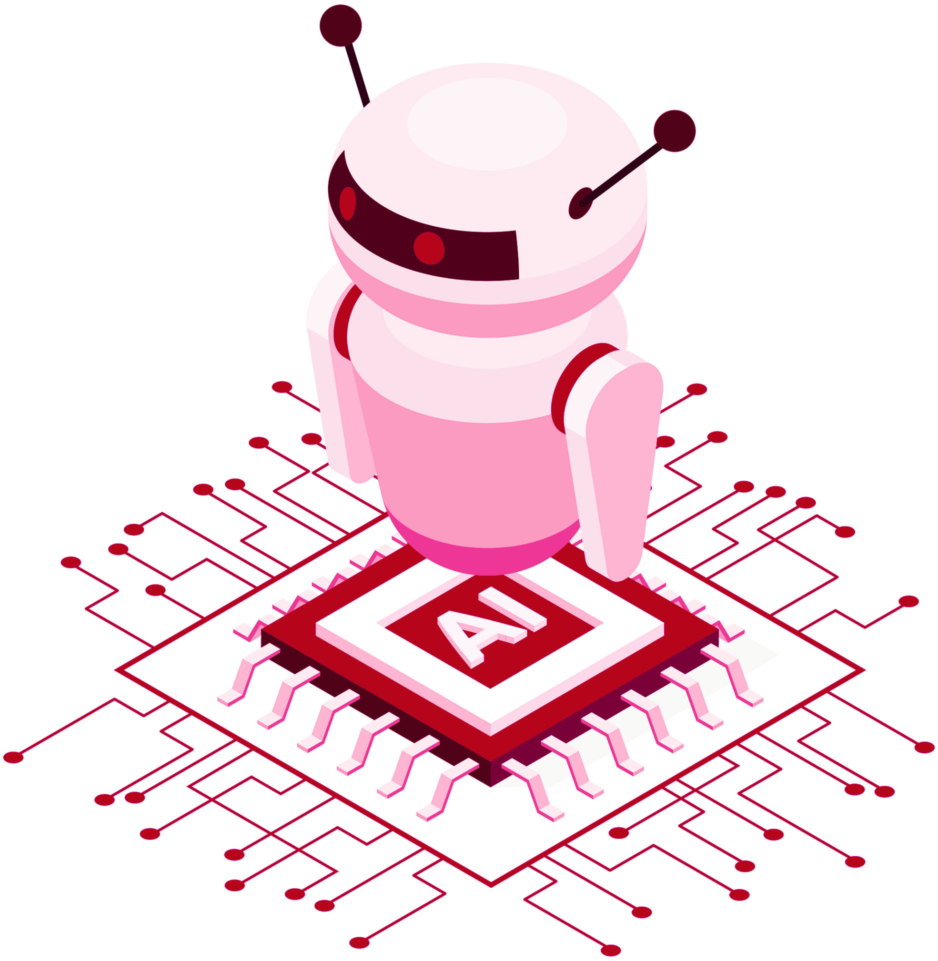

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution for Homework 11:

Imitation Learning and Inverse RL

By:

Ayeen Poostforoushan

401105742

RIML

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:

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

Define R_t as the event where the learner's action disagrees with the expert at time t:

$$R_t = \{a_t \neq \pi^*(s_t)\}.$$

Define A_t as the event where at least one disagreement occurs by time t:

$$A_t = \bigcup_{\tau=1}^{t} R_\tau.$$

The total variation distance decomposes as:

$$\begin{aligned}
\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| = \sum_{s_t} \Bigg| & \underbrace{(p_{\pi_\theta}(s_t \mid \neg A_t)\Pr(\neg A_t) - p_{\pi^*}(s_t \mid \neg A_t)\Pr(\neg A_t))}_{\text{0 when } \neg A_t} \\
& + \underbrace{(p_{\pi_\theta}(s_t \mid A_t) - p_{\pi^*}(s_t \mid A_t))\Pr(A_t)}_{\text{difference under } A_t} \Bigg| \\
& \leq \sum_{s_t} 0 + \sum_{s_t} |p_{\pi_\theta}(s_t \mid A_t) - p_{\pi^*}(s_t \mid A_t)| \Pr(A_t) \\
& \leq 2\Pr(A_t),
\end{aligned}$$

where the last inequality is because the total variation distance between two distributions is at most 2.

By the union bound:

$$\Pr(A_t) \leq \sum_{\tau=1}^{t} \Pr(R_\tau).$$

Each $\Pr(R_\tau)$ is the disagreement probability under the expert's state distribution:

$$\Pr(R_\tau) = \mathbb{E}_{s_\tau \sim p_{\pi^*}} \left[\pi_\theta(a_\tau \neq \pi^*(s_\tau) \mid s_\tau)\right].$$

From the imitation error bound:

$$\sum_{\tau=1}^{T} \mathbb{E}_{s_\tau \sim p_{\pi^*}} \left[\pi_\theta(a_\tau \neq \pi^*(s_\tau) \mid s_\tau)\right] \leq T\varepsilon.$$

Since $t \leq T$, we have:

$$\sum_{\tau=1}^{t} \Pr(R_\tau) \leq \sum_{\tau=1}^{T} \Pr(R_\tau) \leq T\varepsilon.$$

So when we combine all the inequalities we have:

$$\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2\Pr(A_t) \leq 2\sum_{\tau=1}^{t} \Pr(R_\tau) \leq 2T\varepsilon.$$

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:

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

The expected return for a policy π is:

$$J(\pi) = \sum_{t=1}^{T} \mathbb{E}_{p_\pi(s_t)}[r(s_t)] = \mathbb{E}_{p_\pi(s_T)}[r(s_T)],$$

since rewards are zero except at $t = T$. The return gap is:

$$\begin{aligned}\Delta J &= J(\pi^*) - J(\pi_\theta) = \mathbb{E}_{p_{\pi^*}(s_T)}[r(s_T)] - \mathbb{E}_{p_{\pi_\theta}(s_T)}[r(s_T)] \\ &= \sum_{s_T} r(s_T)p_{\pi^*}(s_T) - \sum_{s_T} r(s_T)p_{\pi_\theta}(s_T) = \sum_{s_T} r(s_T)\left(p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)\right)\end{aligned}$$

Since $|r(s_T)| \leq R_{\max}$, we bound the absolute gap:

$$|\Delta J| = \left|\sum_{s_T} r(s_T)\left(p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)\right)\right| \leq \sum_{s_T} |r(s_T)|\,|p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)| \leq R_{\max}\sum_{s_T} |p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)|\,.$$

From part 1, the total variation distance at time T is bounded. So we finally have:

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

1.3 Task 3: Return Gap for General Rewards

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

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

The return gap is:

$$\Delta J = J(\pi^*) - J(\pi_\theta) = \sum_{t=1}^{T} \left(\mathbb{E}_{p_{\pi^*}(s_t)}[r(s_t)] - \mathbb{E}_{p_{\pi_\theta}(s_t)}[r(s_t)]\right).$$

Expanding each expectation over states:

$$\Delta J = \sum_{t=1}^{T}\left(\sum_{s_t} r(s_t)p_{\pi^*}(s_t) - \sum_{s_t} r(s_t)p_{\pi_\theta}(s_t)\right) = \sum_{t=1}^{T}\sum_{s_t} r(s_t)\left(p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)\right).$$

Since $|r(s_t)| \le R_{\max}$ for all t, we bound the absolute gap:

$$|\Delta J| \le \sum_{t=1}^{T} \sum_{s_t} |r(s_t)| \cdot |p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)| \le R_{\max} \sum_{t=1}^{T} \sum_{s_t} |p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)| \,.$$

From Task 1, for each time step t, the total variation distance is bounded by:

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

Now putting this bound for every t:

$$|\Delta J| \le R_{\max} \sum_{t=1}^{T} 2T\varepsilon = R_{\max} \cdot 2T\varepsilon \cdot T = 2R_{\max}T^2\varepsilon.$$

So the return gap satisfies the requested $\mathcal{O}(T^2\varepsilon)$

References

[1] Cover image designed by freepik