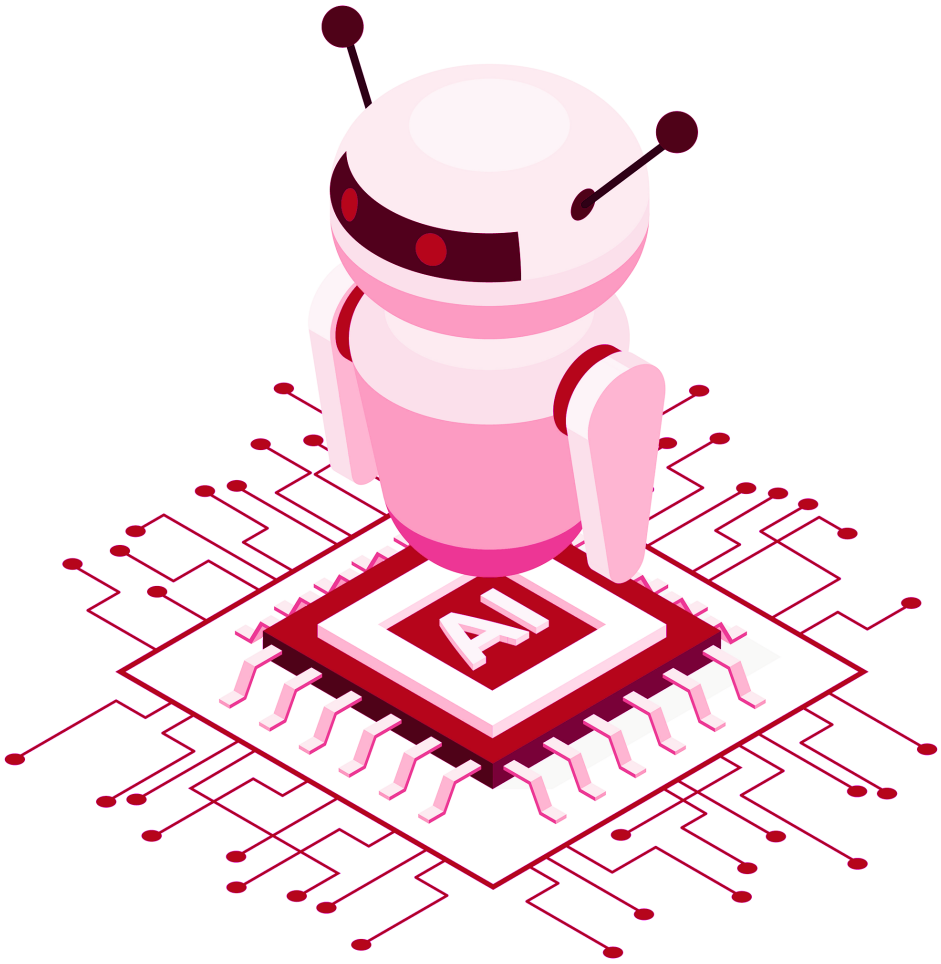

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution for Homework 11:

Imitation Learning and Inverse RL

By:

Asemaneh Nafe

400105285

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:

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

We want to show that the total variation distance between the state distributions induced by the learned policy π_θ and the expert policy π^* satisfies:

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

Setup: Let $d_t^\pi(s)$ denote the state distribution at time t under policy π. Assume the learned policy π_θ satisfies:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{s_t \sim d_t^{\pi^*}} \left[\mathbb{P}_{a\sim\pi_\theta(\cdot|s_t)}(a \neq \pi^*(s_t))\right] \leq \varepsilon,$$

which bounds the average disagreement between the learned and expert policies on the expert's state distribution by ε.

Key Idea: A deviation from the expert policy at step t can affect the distribution over future states from time $t+1$ onward. Hence, errors can accumulate over time due to distributional shift.

From the result of Ross and Bagnell (2010), we know that if the per-step disagreement with the expert is bounded by ε, then the total variation distance between the state distributions at time t satisfies:

$$\|d_t^{\pi_\theta} - d_t^{\pi^*}\|_1 \leq 2t\varepsilon.$$

Summing this over all $t = 1$ to T, we get:

$$\sum_{t=1}^{T} \|d_t^{\pi_\theta} - d_t^{\pi^*}\|_1 \leq \sum_{t=1}^{T} 2\varepsilon = 2T\varepsilon.$$

Conclusion: Therefore, the total variation distance between the state distributions of the learner and expert satisfies:

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

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:

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

Assume that the reward is only received at the final time step, i.e., $r(s_t) = 0$ for all $t < T$, and only $r(s_T) \neq 0$. Under this assumption, the expected return of a policy reduces to:

$$J(\pi) = \mathbb{E}_{s_T \sim p_\pi(s_T)} [r(s_T)] .$$

We aim to show that the return gap between the expert and the learned policy is bounded as:

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

Proof: Let the maximum reward magnitude be bounded as $|r(s_T)| \leq R_{\max}$. Then we can write:

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

Using the triangle inequality:

$$|J(\pi^*) - J(\pi_\theta)| \leq \sum_{s_T} |p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)| \cdot |r(s_T)|.$$

Since $|r(s_T)| \leq R_{\max}$, we have:

$$|J(\pi^*) - J(\pi_\theta)| \leq R_{\max} \sum_{s_T} |p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)| .$$

From Task 1, we know that:

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

Combining the above:

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

Conclusion: Therefore, the return gap is bounded as:

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

1.3 Task 3: Return Gap for General Rewards

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

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

Assume a general reward function where $r(s_t) \neq 0$ for arbitrary t, and the reward at each time step is bounded as $|r(s_t)| \leq R_{\max}$. The expected return of a policy is:

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

We want to show that:

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

Proof: We write the return gap as:

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

This can be rewritten as:

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

Using the triangle inequality and the reward bound:

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

From known results in imitation learning (e.g., Ross and Bagnell, 2010), we have:

$$\|p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)\|_1 \leq 2t\varepsilon.$$

Therefore:

$$|J(\pi^*) - J(\pi_\theta)| \leq R_{\max} \sum_{t=1}^{T} 2t\varepsilon = 2R_{\max}\varepsilon \sum_{t=1}^{T} t = 2R_{\max}\varepsilon \cdot \frac{T(T+1)}{2}.$$

Simplifying:

$$|J(\pi^*) - J(\pi_\theta)| \leq R_{\max}\varepsilon T(T+1) = \mathcal{O}(T^2\varepsilon).$$

Conclusion: The return gap under a general reward function satisfies:

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

References

[1] Cover image designed by freepik