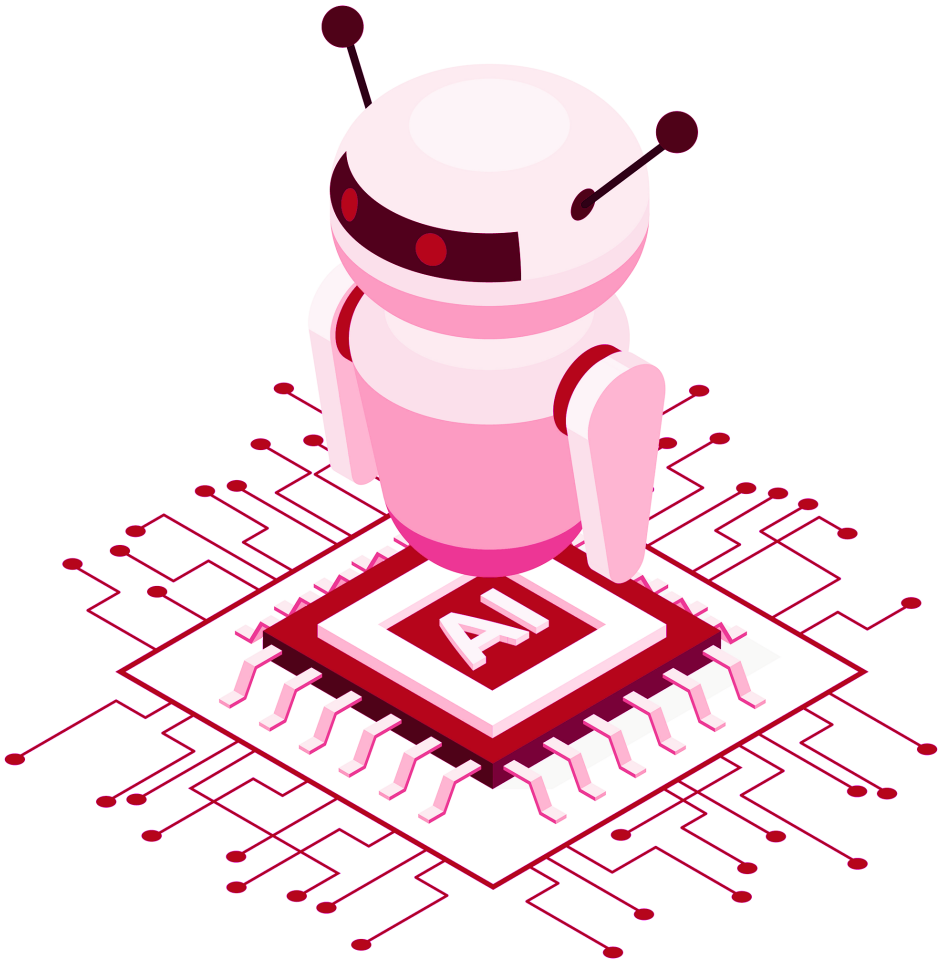

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution for Homework 11:

Imitation Learning and Inverse RL

By:

Ashkan Majidi

400109984

RIML

Deep RL Course

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:

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

The core assumption of behavioral cloning is that the learned policy mimics the expert's actions. We define ε as the maximum one-step error probability. More formally, let $\epsilon_s = \frac{1}{2}\sum_a |\pi_\theta(a|s) - \pi^*(a|s)|$ be the total variation distance between the action distributions at state s. We assume a uniform bound $\varepsilon = \sup_s \epsilon_s$. This implies $\sum_a |\pi_\theta(a|s) - \pi^*(a|s)| \leq 2\varepsilon$ for all s.

We will prove by induction that $\sum_s |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2t\varepsilon$. The desired bound follows from this, as $t \leq T$.

Base Case (t=0): At $t = 0$, the state distributions are identical by definition.

$$\sum_{s_0} |p_{\pi_\theta}(s_0) - p_{\pi^*}(s_0)| = \sum_{s_0} |p_0(s_0) - p_0(s_0)| = 0.$$

The inequality $0 \leq 2(0)\varepsilon$ holds.

Inductive Hypothesis: Assume for some timestep $t \geq 0$ that $\sum_s |p_{\pi_\theta}(s_t = s) - p_{\pi^*}(s_t = s)| \leq 2t\varepsilon$.

Inductive Step: We want to show this holds for $t+1$. Let $d_t(s) = p_{\pi_\theta}(s_t = s)$ and $d_t^*(s) = p_{\pi^*}(s_t = s)$. The state distribution at $t+1$ is given by:

$$d_{t+1}(s') = \sum_s \sum_a d_t(s)\pi_\theta(a|s)P(s'|s,a).$$

We analyze the L1 distance at $t+1$:

$$\sum_{s'} |d_{t+1}(s') - d^*_{t+1}(s')| = \sum_{s'} \left| \sum_{s,a} \left(d_t(s)\pi_\theta(a|s) - d_t^*(s)\pi^*(a|s)\right) P(s'|s,a) \right|.$$

We add and subtract the term $d_t(s)\pi^*(a|s)$:

$$\sum_{s'} \left| \sum_{s,a} \left(d_t(s)\pi_\theta(a|s) - d_t(s)\pi^*(a|s) + d_t(s)\pi^*(a|s) - d_t^*(s)\pi^*(a|s)\right) P(s'|s,a) \right|$$

$$\leq \sum_{s'} \left| \sum_{s,a} d_t(s)(\pi_\theta(a|s) - \pi^*(a|s))P(s'|s,a) \right| \quad \text{(Term A)}$$

$$+ \sum_{s'} \left| \sum_{s,a} (d_t(s) - d_t^*(s))\pi^*(a|s)P(s'|s,a) \right| \quad \text{(Term B)}$$

by the triangle inequality.

Bounding Term A (Error from policy mismatch):

$$\begin{aligned}\text{Term A} &\leq \sum_{s'}\sum_{s,a} d_t(s)|\pi_\theta(a|s) - \pi^*(a|s)|P(s'|s,a) \\ &= \sum_{s,a} d_t(s)|\pi_\theta(a|s) - \pi^*(a|s)| \sum_{s'} P(s'|s,a) \\ &= \sum_{s} d_t(s) \sum_{a} |\pi_\theta(a|s) - \pi^*(a|s)| \\ &\leq \sum_{s} d_t(s)(2\varepsilon) = 2\varepsilon \sum_{s} d_t(s) = 2\varepsilon.\end{aligned}$$

Bounding Term B (Error propagation): Let's define the transition operator under policy π^* as $(T_{\pi^*} f)(s') = \sum_{s,a} f(s)\pi^*(a|s)P(s'|s,a)$. Term B is $\sum_{s'} |(T_{\pi^*} d_t)(s') - (T_{\pi^*} d_t^*)(s')|$. Since a Markov transition kernel is a contraction on the L1 norm:

$$\begin{aligned}\text{Term B} &\leq \sum_{s} |d_t(s) - d_t^*(s)| \\ &\leq 2t\varepsilon \quad \text{(by the inductive hypothesis).}\end{aligned}$$

Combining the terms:

$$\sum_{s'} |d_{t+1}(s') - d_{t+1}^*(s')| \leq \text{Term A} + \text{Term B} \leq 2\varepsilon + 2t\varepsilon = 2(t+1)\varepsilon.$$

By induction, the statement holds for all t. Since $t \leq T$ for any step in the horizon, we can use a looser bound:

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

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 (or cost) for a policy π is $J(\pi) = _{\tau\sim\pi}\left[\sum_{t=0}^{T-1} r(s_t)\right]$. Given that reward is only non-zero at the final timestep T (let's assume the horizon is $T+1$ and the last reward is at s_T), the return simplifies to:

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

The difference in performance between the expert and the learned policy is:

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

Let's assume the reward function is bounded, i.e., $|r(s)| \leq R_{max}$ for all s. Taking the absolute value of the difference:

$$\begin{aligned}|J(\pi^*) - J(\pi_\theta)| &= \left|\sum_{s_T}(p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T))r(s_T)\right| \\ &\leq \sum_{s_T}|p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)||r(s_T)| \quad \text{(by triangle inequality)} \\ &\leq R_{max}\sum_{s_T}|p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)|.\end{aligned}$$

From Task 1, we have the bound $\sum_{s_t}|p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)| \leq 2t\varepsilon$. For the final timestep $t = T$, this gives:

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

Substituting this into our inequality for the performance gap:

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

Since R_{max} is a constant, the performance gap is of the order $\mathcal{O}(T\varepsilon)$.

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

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:

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

For a general reward function, the expected return is the sum of expected rewards at each step:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T-1} r(s_t)\right] = \sum_{t=0}^{T-1}\mathbb{E}_{s_t\sim p_\pi(s_t)}[r(s_t)] = \sum_{t=0}^{T-1}\sum_{s_t} p_\pi(s_t)r(s_t).$$

The performance gap is:

$$\begin{aligned}J(\pi^*) - J(\pi_\theta) &= \sum_{t=0}^{T-1}\left(\sum_{s_t} p_{\pi^*}(s_t)r(s_t) - \sum_{s_t} p_{\pi_\theta}(s_t)r(s_t)\right) \\ &= \sum_{t=0}^{T-1}\sum_{s_t}(p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t))r(s_t).\end{aligned}$$

Again, we assume $|r(s)| \leq R_{max}$ and take the absolute value:

$$\begin{aligned}|J(\pi^*) - J(\pi_\theta)| &\leq \sum_{t=0}^{T-1}\left|\sum_{s_t}(p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t))r(s_t)\right| \\ &\leq \sum_{t=0}^{T-1}\sum_{s_t}|p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)||r(s_t)| \\ &\leq \sum_{t=0}^{T-1} R_{max}\sum_{s_t}|p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)|.\end{aligned}$$

From the proof of Task 1, we have the tighter, step-dependent bound $\sum_{s_t} |p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)| \leq 2t\varepsilon$. Substituting this into the sum:

$$\begin{aligned} |J(\pi^*) - J(\pi_\theta)| &\leq \sum_{t=0}^{T-1} R_{max}(2t\varepsilon) \\ &= 2R_{max}\varepsilon \sum_{t=0}^{T-1} t. \end{aligned}$$

The sum is an arithmetic series: $\sum_{t=0}^{T-1} t = \frac{(T-1)T}{2}$.

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

The term $T(T-1) = T^2 - T$ is dominated by T^2 for large T. Therefore, the performance gap is of the order $\mathcal{O}(T^2\varepsilon)$.

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

References

[1] Cover image designed by freepik