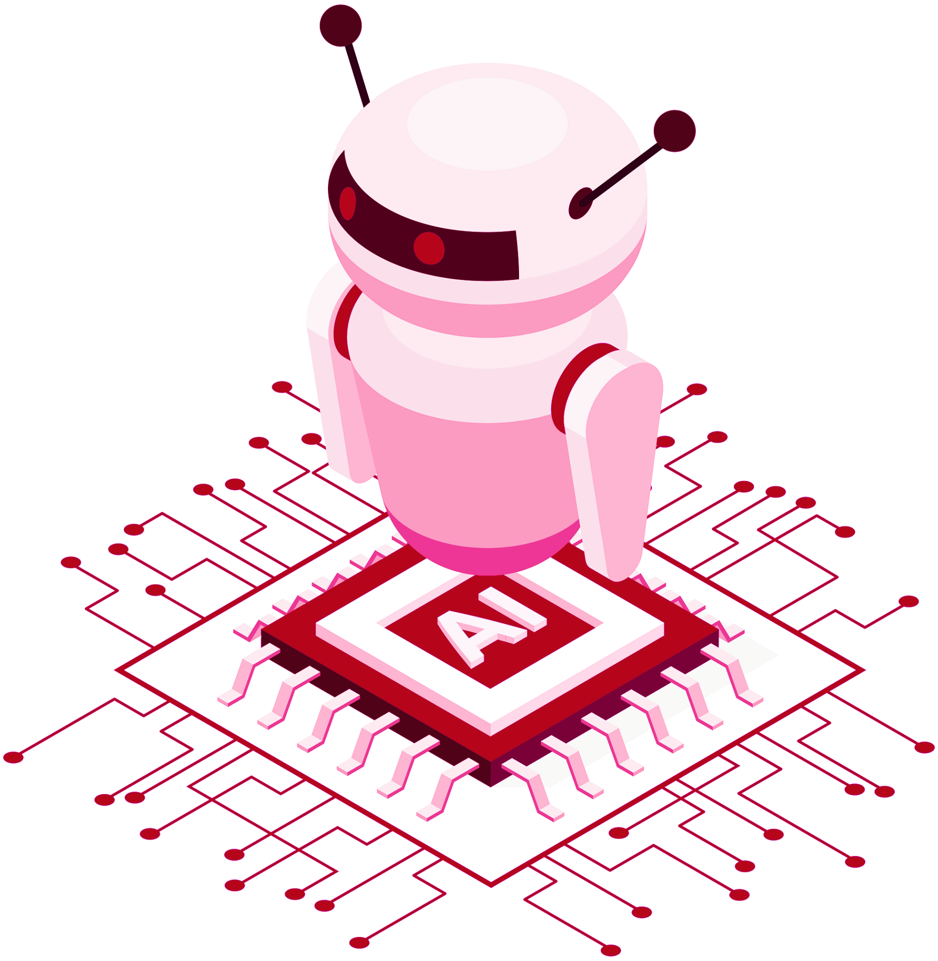

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution for Homework 11:

Imitation Learning and Inverse RL

By:

Amir Kooshan Fattah Hesari
401102191

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:

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

Answer

Based on the description, we have the following bound on the expected of the times that the learned policy differs from the expert policy :

$$\mathbb{E}_{p_{\pi^*}(s)}\left[\pi_\theta(a \neq \pi^*(s)|s)\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{p_{\pi^*}(s_t)}\left[\pi_\theta(a_t \neq \pi^*(s_t)|s_t)\right] \leq \epsilon$$

we can treat the given expectation on the times that the action chosen by the learned policy differs from the expert's policy as the expectation of an indicator function under the distirubtion induced by following the expert's policy :

$$\mathbb{E}_{p_{\pi^*}(s)}\left[\mathbb{I}(g)\right] = \int \mathbb{I}(g)p_{\pi^*}(s)ds = \Pr(g)$$
$$\mathbb{E}_{p_{\pi^*}(s)}\left[\mathbb{I}(\pi_\theta(a_t|s_t) \neq \pi^*(a_t|s_t))\right] = \Pr\left[\pi_\theta(a_t|s_t) \neq \pi^*(a_t|s_t)\right] = q_t$$
$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{p_{\pi^*}(s_t)}\left[\pi_\theta(a_t \neq \pi^*(s_t)|s_t)\right] = \frac{1}{T}\sum_{t=1}^{T} q_t \leq \epsilon \Rightarrow \sum_{t=1}^{T} q_t \leq T\epsilon$$

Now we introduce the coupling inequality.

$$\begin{aligned}\underbrace{\mu(A) - v(A)}_{\text{where } \mu \text{ and } v \text{ are probability measures}} &= \Pr(X \in A) - \Pr(Y \in A) \\ &= \Pr(X \in A, X \neq Y) + \Pr(X \in A, X = Y) \\ &- \Pr(Y \in A, X = A) - \Pr(Y \in A, X \neq Y)\end{aligned}$$

$$\Rightarrow \Pr(X \in A) - \Pr(Y \in A) = \Pr(X \in A, X \neq Y) - \Pr(Y \in A, X \neq Y)$$
$$\Pr(X \in A, X \neq Y) - \Pr(Y \in A, X \neq Y) \leq \Pr(X \in A, X \neq Y) \leq \Pr(X \neq Y)$$
$$\Rightarrow \mu(A) - v(A) \leq \Pr(X \neq Y)$$

and by symmetry :

$$\Rightarrow v(A) - \mu(A) \leq \Pr(X \neq Y) \Rightarrow |\mu(A) - v(A)| \leq \Pr(X \neq Y)$$

And if we take a supremum over all sets A in our measure space , we have the following identity :

$$\sup_{A\in\mathcal{S}} |\mu(A) - v(A)| = ||\mu - v||_{TV} \leq \Pr\left[X \neq Y\right]$$

we have proved an important lemma that we are going to use. Now for the proof , consider the event E_t and A_t such that

$$E_t = \{a_t \neq \pi^*(s_t)\}$$
$$A_t = \cup_{\tau=1}^{t} E_\tau$$

meaning that until time t at least one action differs from the expert's policy. Therefore the difference can be written in the following foramt :

$$\sum_{s_t} |p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)| =$$
$$\sum_{s_t} |p_{\pi^*}(s_t|A_t)\Pr(A_t) - p_{\pi_\theta}(s_t|A_t)\Pr(A_t) + \underbrace{p_{\pi^*}(s_t|\sim A_t)\Pr(\sim A_t) - p_{\pi_\theta}(s_t|\sim A_t)\Pr(\sim A_t)}_{\text{0 because } \sim A_t \text{ means that our actions and expert's actions are same}} |$$
$$= \sum_{s_t} |p_{\pi^*}(s_t|A_t)\Pr(A_t) - p_{\pi_\theta}(s_t|A_t)\Pr(A_t)| \;= \sum_{s_t} |p_{\pi^*}(s_t|A_t) - p_{\pi_\theta}(s_t|A_t)|\Pr(A_t)$$
$$\Rightarrow \sum_{s_t} |p_{\pi^*}(s_t|A_t) - p_{\pi_\theta}(s_t|A_t)| \leq 2, \quad \text{when the distributions differ maximum of TV is 2}$$
$$\Rightarrow \sum_{s_t} |p_{\pi^*}(s_t|A_t) - p_{\pi_\theta}(s_t|A_t)|\Pr(A_t) \leq 2\Pr(A_t)$$
$$\Rightarrow \Pr(A_t) \leq \sum_{\tau} \Pr(E_\tau)$$

and from the definition provided we have :

$$\Pr(E_\tau) = \mathbb{E}_{s_\tau \sim p_{\pi^*}}[\pi_\theta(a_\tau \neq \pi^*(s_\tau)|s_\tau)]\,,\text{based on the identity } \mathbb{E}[\mathbb{I}(g)] = \Pr(g)$$
$$\sum_{\tau} \mathbb{E}_{s_\tau \sim p_{\pi^*}}[\pi_\theta(a_\tau \neq \pi^*(s_\tau)|s_\tau)] \leq T\epsilon \Rightarrow \Pr(A_T) \leq \sum_{\tau=1}^{T} \Pr(E_\tau) \leq T\epsilon$$
$$\Rightarrow \sum_{s_t} |p_{\pi^*}(s_t|A_t) - p_{\pi_\theta}(s_t|A_t)| \leq 2\Pr(A_t) \leq 2T\epsilon$$

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:

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

Answer

From the given definition we have that :

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

$$r(s_t) = 0; \quad \forall t < T \Rightarrow J(\pi^*) - J(\pi_\theta) = \mathbb{E}_{p_{\pi^*}(s_T)}\left[r(s_T)\right] - \mathbb{E}_{p_{\pi_\theta}(s_T)}\left[r(s_T)\right]$$

$$J(\pi^*) - J(\pi_\theta) = \sum_{s_T} r(s_T)p_{\pi^*}(s_T) - r(s_T)p_{\pi_\theta}(s_T) \leq R_{max} \underbrace{\sum_{s_T} p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)}_{\leq T\epsilon}$$

$$\Rightarrow J(\pi^*) - J(\pi_\theta) \leq R_{max}T\epsilon \approx \mathcal{O}(T\epsilon)$$

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:

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

Answer

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

$$\begin{aligned} J(\pi^*) - J(\pi_\theta) = \sum_{t=1}^{T} \mathbb{E}_{p_{\pi^*}(s_t)}\left[r(s_t)\right] - \mathbb{E}_{p_{\pi_\theta}(s_t)}\left[r(s_t)\right] &= \sum_{t=1}^{T}\sum_{s_t} p_{\pi^*}(s_t)r(s_t) - p_{\pi_\theta}(s_t)r(s_t) \\ &\leq \sum_{t=1}^{T} R_{max} \underbrace{\sum_{s_t} p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)}_{\leq T\epsilon} \end{aligned}$$

$$\Rightarrow J(\pi^*) - J(\pi_\theta) \leq \textstyle\sum_{t=1}^{T} R_{max}T\epsilon \approx R_{max}T^2\epsilon \approx \mathcal{O}(T^2\epsilon)$$

References

[1] Cover image designed by freepik