CS285: Deep Reinforcement Learning
Assignment 1
Written Report

Alan Sorani

March 19, 2025

1 Analysis

Consider the problem of imitation learning within a discrete MDP with horizon T and an expert policy π^*. We gather expert demonstrations from π^* and fit an imitation policy π_θ to these trajectories so that

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

The notation $p_\pi\left(s_t\right)$ indicates the state distribution under a policy π at time step t, while $p_\pi\left(s\right)$ indicates the state marginal of π across time steps, unless indicated otherwise.

1. TODO
2. We have

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

   (a) Assuming $r\left(s_t\right) = 0$ for all $t < T$, we get that

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

   and by the triangle inequality and Item 1 we get that

$$\begin{aligned} \left|J\left(\pi^*\right) - J\left(\pi_\theta\right)\right| &\leq \sum_{s_T}\left|p_{\pi^*}\left(s_T\right) - p_{\pi_\theta}\left(s_T\right)\right|\left|r\left(s_T\right)\right| \\ &\leq R_{\max}\sum_{s_T}\left|p_{\pi^*}\left(s_T\right) - p_{\pi_\theta(s_T)}\right| \\ &\leq 2T\varepsilon R_{\max}. \end{aligned}$$

   Hence $J\left(\pi^*\right) - J\left(\pi_\theta\right) = \mathrm{O}\left(T\varepsilon\right)$.

(b) Without the assumption on the reward, we get

$$\begin{aligned}
\left|J\left(\pi^*\right)-J\left(\pi_\theta\right)\right| &\leq \sum_{t\in T}\sum_{s_t}\left|p_{\pi^*}\left(s_t\right)-p_{\pi_\theta}\left(s_t\right)\right|\left|r\left(s_t\right)\right| \\
&\leq R_{\max}\sum_{t\in T}\sum_{s_t}\left|p_{\pi^*}\left(s_t\right)-p_{\pi_\theta}\left(s_t\right)\right| \\
&\leq R_{\max}\sum_{t\in T}2t\epsilon \\
&= 2R_{\max}\varepsilon\frac{T\left(T+1\right)}{2} \\
&= \mathrm{O}\left(T^2\varepsilon\right),
\end{aligned}$$

as required.

3 Behavioral Cloning

4 DAgger

5 Discussion