

CS 285 Homework 1: Imitation Learning

1 Analysis

1. Let us define

$$\Pr[\text{mistake}] := \Pr[\text{make a mistake at some time step } \leq t].$$

Then, following the same idea as in lecture we have

$$\begin{aligned} p_{\pi_\theta}(s_t) &= (1 - \Pr[\text{mistake}]) \cdot p_{\pi^*}(s_t) + \Pr[\text{mistake}] \cdot p_{\text{mistake}}(s_t) \\ \implies \sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| &= \Pr[\text{mistake}] \cdot \sum_{s_t} |p_{\text{mistake}}(s_t) - p_{\pi^*}(s_t)| \\ &\leq 2 \Pr[\text{mistake}] \end{aligned}$$

Now, it suffices to show that $\Pr[\text{mistake}] \leq T\epsilon$:

$$\begin{aligned} \Pr[\text{mistake}] &= \Pr(\cup_{0 \leq i \leq t} \text{first mistake occurs at time step } i) \\ &\leq \sum_{i \leq t} \Pr(\text{first mistake at time step } i) \\ &= \sum_{i \leq t} \Pr(\text{matches expert policy until time step } i) \\ &\leq \sum_{i \leq t} \sum_{s_i} \Pr[\text{make a mistake at } s_i] \cdot p_{\pi^*}(s_i) \\ &= \sum_{i \leq t} \mathbb{E}_{p_{\pi^*}(s_i)}[\pi_\theta(a_i \neq \pi^*(s_i) \mid s_i)] \\ &\leq \sum_{i \leq T} \mathbb{E}_{p_{\pi^*}(s_i)}[\pi_\theta(a_i \neq \pi^*(s_i) \mid s_i)] \\ &\leq T\epsilon \end{aligned}$$

as desired. Note that we applied the union bound between the 1st and 2nd lines.

2. (a) When the reward only depends on the last state, we have

$$\begin{aligned} J(\pi^*) - J(\pi_\theta) &= \sum_{t=1}^T \mathbb{E}_{p_{\pi^*}(s_t)}[r(s_t)] - \sum_{t=1}^T \mathbb{E}_{p_{\pi_\theta}(s_t)}[r(s_t)] \\ &= \mathbb{E}_{p_{\pi^*}(s_T)}[r(s_T)] - \mathbb{E}_{p_{\pi_\theta}(s_T)}[r(s_T)] \\ &= \sum_{s_T} r(s_T) \cdot p_{\pi^*}(s_T) - \sum_{s_T} r(s_T) \cdot p_{\pi_\theta}(s_T) \\ &\leq R_{\max} \cdot \sum_{s_T} |p_{\pi_\theta}(s_T) - p_{\pi^*}(s_T)| \\ &\leq 2R_{\max} \cdot T\epsilon \\ &= \mathcal{O}(T\epsilon) \end{aligned}$$

(b) For any arbitrary reward, we have

$$\begin{aligned} J(\pi^*) - J(\pi_\theta) &= \sum_{t=1}^T \mathbb{E}_{p_{\pi^*}(s_t)}[r(s_t)] - \sum_{t=1}^T \mathbb{E}_{p_{\pi_\theta}(s_t)}[r(s_t)] \\ &= \sum_{t=1}^T \sum_{s_t} r(s_t) \cdot (p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)) \\ &\leq R_{\max} \sum_{t=1}^T \sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \\ &\leq R_{\max} \sum_{t=1}^T 2T\epsilon \\ &= 2R_{\max} \cdot T^2\epsilon \\ &= \mathcal{O}(T^2\epsilon) \end{aligned}$$

2 Behavioral Cloning

3 DAGGER

4 Extra Credit: SWITCHDAGGER

5 Discussion