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

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

i.e., the expected likelihood that the learned policy π_θ disagrees with the expert π^* within the training distribution p_{π^*} of states drawn from random expert trajectories is at most ε.

For convenience, the notation $p_\pi(s_t)$ indicates the state distribution under π at time step t while $p(s)$ indicates the state marginal of π across time steps, unless indicated otherwise.

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

   [Hint: In lecture, we showed a similar inequality under the stronger assumption $\pi_\theta(\overset{a_t}{\cancel{s_t}} \neq \pi^*(s_t) \mid s_t) \leq \varepsilon$ for every $s_t \in \text{supp}(p_{\pi^*})$. Try converting the inequality above into an expectation over p_{π^*} and use a union bound ($\Pr[\bigcup_i E_i] \leq \sum_i \Pr[E_i]$) to get the desired result.]

From hint, we know that $\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) \leq \varepsilon$ for every $s_t \in \text{supp}(p_{\pi^*})$, where ε represent the probability of making mistakes.

Since $p_{\pi^*}(s) \neq p_{\pi_\theta}(s)$, and $p_{\pi^*}(s)$ is training data, thus

$$p_{\pi_\theta}(s_t) = (1-\varepsilon)^t p_{\pi^*}(s_t) + (1-(1-\varepsilon)^t)\, p_{mistake}(s_t)$$

$p_{\pi_\theta}(s_t)$ is the distribution over states at timestep t: sum of probability we made no mistakes and some other probability.

So, $$\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| = (1-(1-\varepsilon)^T)\,|p_{mistake}(s_t) - p_{\pi^*}(s_t)|$$
$$\leq 2(1-(1-\varepsilon)^T)$$
$$\leq 2T\varepsilon$$

where we use identity: $(1-\varepsilon)^T \geq 1 - T\varepsilon$ for $\varepsilon \in [0,1]$

and the fact: worst case of variation divergence is 2 because the worst case is that in one state one probability is 1, the other is 0 and in some other state it's the way around; so the worst possible difference between 2 distributions when you sum over all states is 2.

2. Consider the expected return of the learned policy π_θ for a state-dependent reward $r(s_t)$, where we assume the reward is bounded with $|r(s_t)| \le R_{\max}$:

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

(a) Show that $J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T\varepsilon)$ when the reward only depends on the last state, i.e., $r(s_t) = 0$ for all $t < T$.

(b) Show that $J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T^2\varepsilon)$ for an arbitrary reward.

a) $J(\pi^*) = E_{p_{\pi^*}(s_T)} r(s_T) \le R_{max} \cdot E\, p_{\pi^*}(s_T)$

$J(\pi_\theta) = E_{\pi_\theta}(s_T)\, r(s_T) \le R_{max} \cdot E\, p_{\pi_\theta}(s_T)$

$J(\pi^*) - J(\pi_\theta) \le R_{max} \, | p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T) |$

$\le 2 T R_{max} \varepsilon$

from the conclusion of 1)

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

b) from a) we know that for the last state $J(\pi^*) - J(\pi_\theta) = O(T\varepsilon)$, thus if for all states with arbitrary reward

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

$\le 2T^2 R_{max} \varepsilon$

thus $J(\pi^*) - J(\pi_\theta) = O(T^2 \varepsilon)$