Analysis

1. State Distribution Divergence Bound

We are given a discrete MDP (Markov Decision Process) and a policy $\pi_\theta(a \mid s)$. The state distribution under the expert policy π^* is $p_{\pi^*}(s)$, and under the learned policy π_θ is $p_{\pi_\theta}(s)$.

We collect expert demonstrations and fit an imitation policy π_θ such that, under the expert's state distribution, the probability that π_θ disagrees with the expert policy π^* at any time step is at most ϵ:

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

We want to show $\displaystyle\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2T\epsilon$

Proof Sketch:

At each time step, the probability that π_θ disagrees with π^* is at most ϵ on average over t. By the union bound, the probability that at least one mistake occurs over T steps is at most $T\epsilon$. Each mistake can cause the state distribution to diverge further from the expert's. Thus, the total variation distance between $p^{\pi_\theta}(s_t)$ and $p^{\pi^*}(s_t)$ summed over t is at most $2T\epsilon$.

Set $e_t = \mathbb{E}_{p^{\pi^*}(s_t)}\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)$

The total variation distance between $p^{\pi_\theta}(s_t)$ and $p^{\pi^*}(s_t)$ is at most $2\sum_{i=1}^{t} e_i$. Summing over all t:

$$\begin{aligned}\sum_{t=1}^{T}\sum_{s_t} |p^{\pi_\theta}(s_t) - p^{\pi^*}(s_t)| \leq\ & 2\sum_{t=1}^{T}\sum_{i=1}^{t} e_i \quad \text{(for each e_i, it appears in the sum for all $t \geq i$, which is $(T-i+1)$ times)} \\ &= 2\sum_{i=1}^{T}(T-i+1)e_i \leq 2T\sum_{i=1}^{T} e_i = 2T^2\epsilon.\end{aligned}$$

2(a). Return Difference: Reward at Last Step Only

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

If the reward only depends on the last state, i.e., $r(s_t) = 0$ for $t < T$, then $J(\pi) = \mathbb{E}_{p_\pi(s_T)} r(s_T)$, where $r(s_T)$ is the reward at the last state, $p_\pi(s_T)$ is the probability that policy π is in state s_T at time T. Thus,

$$|J(\pi^*) - J(\pi_\theta)| = \left|\sum_{s_T}\left(p^{\pi^*}(s_T) - p^{\pi_\theta}(s_T)\right) r(s_T)\right| \leq \sum_{s_T} |p^{\pi^*}(s_T) - p^{\pi_\theta}(s_T)| \cdot |r(s_T)|.$$

Since $|r(s_T)| \leq R_{\max}$ and from part 1, the sum is at most $2T\epsilon$, so

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

2(b). Return Difference: Arbitrary Reward

For arbitrary $r(s_t)$, $J(\pi) = \displaystyle\sum_{t=1}^{T} \mathbb{E}_{p^\pi(s_t)}[r(s_t)]$. The difference at each step is

$$|\mathbb{E}_{p^{\pi^*}(s_t)}[r(s_t)] - \mathbb{E}_{p^{\pi_\theta}(s_t)}[r(s_t)]| \leq \sum_{s_t} |p^{\pi^*}(s_t) - p^{\pi_\theta}(s_t)| \cdot |r(s_t)| \leq 2t\epsilon R_{\max}.$$

Summing over $t = 1$ to T,

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