Assignment 1: Imitation learning

I. Analysis

1. $E_{p_{\pi^*}(s)} \pi_\theta(a \neq \pi^*(s)|s) \leq \varepsilon$

$\Rightarrow \sum_s p_{\pi^*}(s)\, \pi_\theta(a \neq \pi^*(s)|s) \leq \varepsilon$

Now we consider that the event when the learned policy disagree with the expert policy at each time step is E

$\Rightarrow$ The probability that at least one mistake occur over the horizon T is $Pr[U_T E_T]$

Using the inequality: $Pr[U_i E_i] \leq \sum_i Pr[E_i]$

$$\Rightarrow Pr[U_T E_T] \leq \sum_{t=1}^{T} Pr[E_t]$$

However: $Pr[E] \leq \varepsilon$.

$$\Rightarrow Pr[U_T E_T] \leq \sum_{t=1}^{T} Pr[E_t] \leq \varepsilon T$$

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

2.

a/ $$J(\pi^*) - J(\pi_\theta) = \sum_{t=1}^{T}\left(E_{p_{\pi^*}(s_t)} r(s_t) - E_{p_{\pi_\theta}(s_t)} r(s_t)\right)$$

$$= \sum_{t=1}^{T}\sum_{s_t}\left(p_{\pi^*}(s_t)\, r(s_t) - p_{\pi_\theta}(s_t)\, r(s_t)\right)$$

$$= \sum_{t=1}^{T}\sum_{s_t} r(s_t)\left(p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)\right)$$

$$\leq \sum_{t=1}^{T} |r(s_t)|\,|p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)|$$. Proving in lecture: $|p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)| \leq 2\varepsilon t$

$$\leq \sum_{t=1}^{T} |r(s_t)|\, 2\varepsilon t$$

With reward only exist in the last state

$$\Rightarrow J(\pi^*) - J(\pi_\theta) \leq \sum_{t=1}^{T} |r(s_t)|\, 2\varepsilon t$$

$$= 2\varepsilon T$$

$$= O(\varepsilon T)$$ "Proved!!!"

b/ Similar to a/, We have: $J(\pi^*) - J(\pi_\theta) \leq \sum_{t=1}^{T} |r(s_t)|\, 2\varepsilon T$. Since this question ask for an arbitrary reward

$$\leq \sum_{t=1}^{T} R_{max}\, 2\varepsilon T = 2\varepsilon T^2 R_{max} = O(\varepsilon T^2)$$ "Proved!!!"