Jianxun Zhou
zhou-jx23@mails.tsinghua.edu.cn

Problem 1.1

Proof. By the defintion of total variation distance, we have

$$\Sigma_{s_t}|p_{\pi_\theta}(s_t) - p_{\pi^\star}(s_t)| = 2 \times d_{TV}(p_{\pi_\theta}, p_{\pi^\star})$$

Let M_i denotes the event the learned policy π_θ makes a mistake at step i and makes no mistake in i-1 steps. Let E_t denotes the event the learned policy π_θ makes at least one mistake in t steps. It follows that

$$Pr(E_t) = Pr(\bigcup_{i=0...t}(M_i)) \leq \bigcup_{i=0...t} Pr(M_i) \leq \bigcup_{i=0...T} Pr(M_i) \leq T\varepsilon$$

By the coupling lemma, the distance of state distributions at time t is bounded by the probability of the two trajectories have diverged by that time:

$$d_{TV}(p_{\pi_\theta}, p_{\pi^\star}) \leq Pr(E_t) \leq T\varepsilon$$

Hence

$$\Sigma_{s_t}|p_{\pi_\theta}(s_t) - p_{\pi^\star}(s_t)| \leq 2T\varepsilon$$

as we desired. □

Problem 1.2.a

Proof. Let S denotes the entire state set .

$$\begin{aligned} |J(\pi^\star) - J(\pi_\theta)| =& |E_{p_{\pi^\star}(s_T)} r(s_T) - E_{p_{\pi_\theta}(s_T)} r(s_T)| \\ =& |\Sigma_{s_T \in S}(p_{\pi^\star}(s_T) - p_{\pi_\theta}(s_T)) \times r(s_T)| \\ \leq& max(r(s_T)) \times |p_{\pi^\star}(s_t) - p_{\pi_\theta}(s_t)| \end{aligned}$$

Recalled that $p_{\pi^\star}(s_t) - p_{\pi_\theta}(s_t) \leq 2T\varepsilon$. It follows that $|J(\pi^\star) - J(\pi_\theta)| \leq R_{max} \times 2T\varepsilon$.

Hence

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

as we desired. □

Problem 1.2.b

Proof.

$$\begin{aligned} |J(\pi^\star) - J(\pi_\theta)| =& |\Sigma_{t=1}^{T}\Sigma(r(s_t) \times (p_{\pi^\star}(s_t) - p_{\pi_\theta}(s_t)))| \\ \leq& \Sigma_{t=1}^{T} R_{max}|p_{\pi^\star}(s_t) - p_{\pi_\theta}(s_t)| \\ \leq& T \times R_{max} \times 2T\varepsilon \end{aligned}$$

Hence

$$J(\pi^\star) - J(\pi_\theta) = \mathcal{O}(T^2\varepsilon)$$

as we desired. $\square$