Homework 1

1 Analysis 1

Proof 1.1 记前 T 个时刻，均采取对应策略的概率为 p_{correct}，可得：

$$p_{\text{correct}} = \left(1 - \Pr\left(\cup_{\mathrm{t=1}}^{\mathrm{T}} \left(\pi_\theta(\mathrm{a_t}) \neq \pi_*(\mathrm{a_t})\right)\right)\right). \tag{1}$$

以此可以仿照 *tutorial*，写出在 π_θ 下得到 s_t 的概率：

$$p_{\pi_\theta}(s_t) = p_{\text{correct}} p_{\pi^*}(s_t) + (1 - p_{\text{correct}}) p_{\text{wrong}}(s_t), \tag{2}$$

其中，p_{wrong} 是比较复杂的分布，我们不去考虑它。进行变形，得到：

$$|p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| = (1 - p_{\text{correct}}) \left|p_{\text{wrong}}(s_t) - p_{\pi^*}(s_t)\right|. \tag{3}$$

利用题干中提到的不等式：

$$1 - p_{\text{correct}} \leq \sum_{t=1}^{T} \Pr\left(\pi_\theta(\mathrm{a_t}) \neq \pi_*(\mathrm{a_t})\right) = \epsilon \mathrm{T}. \tag{4}$$

代入式 *3*，并考虑所有情况下的 s_t：

$$\begin{aligned}\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| &\leq \sum_{s_t} \epsilon T \left|p_{\text{wrong}}(s_t) - p_{\pi^*}(s_t)\right| \\ &= \sum_{s_t} \left|p_{\text{wrong}}(s_t) - p_{\pi^*}(s_t)\right| \epsilon T \\ &\leq 2\epsilon T,\end{aligned} \tag{5}$$

从第二行到第三行的放缩使用了全变分距离的性质。

2 Analysis 2

2.1 Question 1

Proof 2.1

$$\begin{aligned}J(\pi^*) - J(\pi_\theta) &= \sum_t \mathbb{E}_{s_t \sim p_{\pi^*}}[r(s_t)] - \sum_t \mathbb{E}_{s_t \sim p_{\pi_\theta}}[r(s_t)] \\ &= \mathbb{E}_{s_T \sim p_{\pi^*}}[r(s_T)] - \mathbb{E}_{s_T \sim p_{\pi_\theta}}[r(s_T)] \\ &= \sum_{s_T} p_{\pi^*}(s_T) r(s_T) - \sum_{s_T} p_{\pi_\theta}(s_T) r(s_T) \\ &= \sum_{s_T} r(s_T) \left(p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)\right) \\ &\leq R_{\max} \sum_{s_T} |p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)| \\ &\leq 2 R_{\max} \epsilon T = \mathcal{O}(\epsilon T).\end{aligned} \tag{6}$$

2.2 Question 2

Proof 2.2

$$
\begin{aligned}
J(\pi^*) - J(\pi_\theta) &= \sum_t \mathbb{E}_{s_t \sim p_{\pi^*}}[r(s_t)] - \sum_t \mathbb{E}_{s_t \sim p_{\pi_\theta}}[r(s_t)] \\
&= \sum_t \sum_{s_T} p_{\pi^*}(s_t) r(s_t) - \sum_t \sum_{s_T} p_{\pi_\theta}(s_t) r(s_t) \\
&= \sum_t \sum_{s_t} r(s_t) \left(p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)\right) \\
&\leq T R_{\max} \sum_{s_T} |p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)| \\
&\leq 2 R_{\max} \epsilon T^2 = \mathcal{O}(\epsilon T^2).
\end{aligned}
\tag{7}
$$