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. □

Problem 2

No Problem.

Problem 3.1

Both **Ant-v4** and **Hopper-v4** are train with n-layers=2&net-size=64&eval-batch-size=10000 and others kept default. Their performance ratios compared to expert are as follows.

environment	avg.ret	std.ret	avg.ret.exp	perf ratio
Ant-v4	4430	780.6	4682	94.62%
Hopper-v4	1098	7.242	3718	29.53%

Table 1: **Ant-v4** vs. **Hopper-v4**

Problem 3.2

Varing **training batch size** from 100 to 1000 with step size 100, we get the performance-batch_size graph[6] of environment **Hopper-v4**.

train_batch_size	avg. ret	std. ret	avg. ret. exp	perf ratio
100	1099	12.6	3718	29.56%
200	1199	26.82	3718	32.25%
300	869.3	34.59	3718	23.38%
400	1221	21.77	3718	32.84%
500	1349	149.1	3718	36.28%
600	1727	381.1	3718	46.45%
700	1308	54.33	3718	35.18%
800	1505	302.5	3718	40.48%
900	1501	439.6	3718	40.37%
1000	1333	87.06	3718	35.85%

Table 2: Behavioral Cloning Performance About Training Batch Size On Hopper-v4

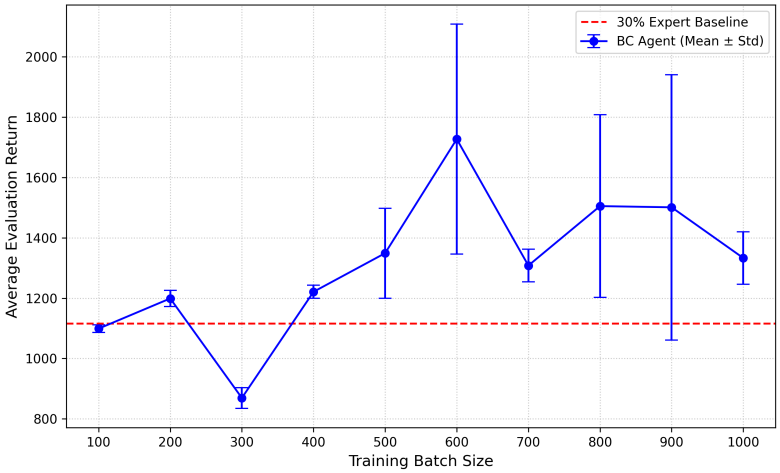

Figure 1: Performance About **Training Batch Size** On Hopper-v4

Problem 4.1