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

Problem 1.1

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

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

Let M_i denote the event the learned policy π_θ makes a mistake at step i and makes no mistake in i-1 steps. Let E_t denote 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 that the two trajectories have diverged by that time:

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

Hence

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

as we desired. □

Problem 1.2.a

Proof. Let S denote the entire state space .

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

Recall 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.

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

Hence

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

as we desired. □

Problem 2

Not applicable.

Problem 3.1

Both **Ant-v4** and **Hopper-v4** are trained with `n-layers=2,net-size=64,eval-batch-size=10000 and others parameters are kept at their defaults`. 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

Varying **training batch size** from 100 to 1000 with step size 100, Figure 1 illustrates the performance of **Hopper-v4** as a function of training batch size.

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 Varying Training Batch Size On Hopper-v4

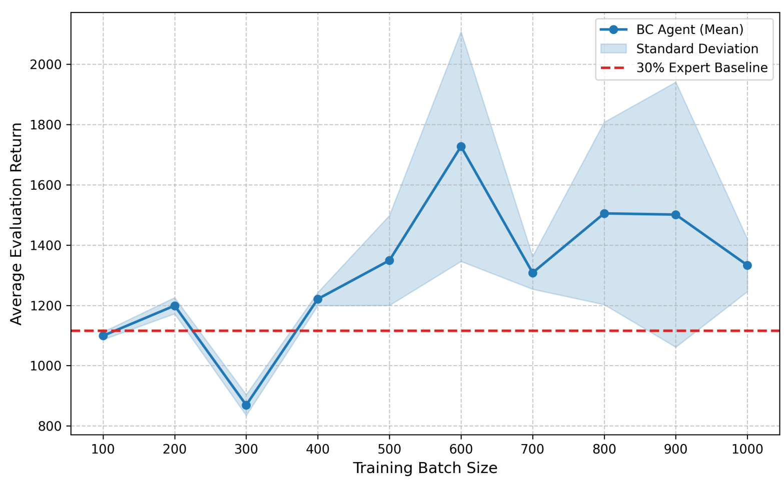
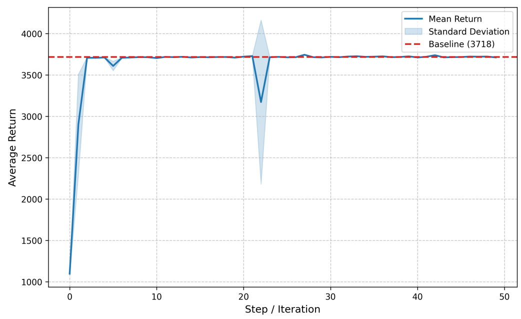
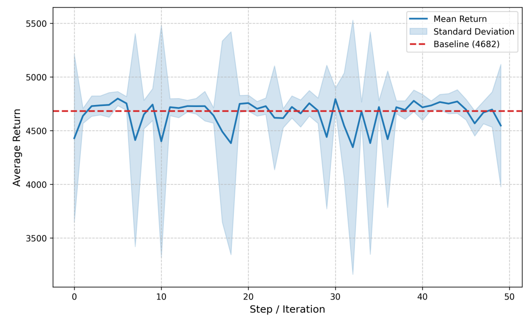


Figure 1: **Mean Return** vs. **Training Batch Size** On **Hopper-v4**
(size=64,n_layers=2,eval_batch_size=10000)

Problem 4.1



(a) **Mean Return** vs. **Training Step** On **Hopper-v4**(size=64,n_layers=2,eval_batch_size=10000)



(b) **Mean Return** vs. **Training Step** On **Ant-v4**(size=64,n_layers=2,eval_batch_size=10000)

Problem 5.1

Proof. Let $C_{t,n}$ denote $C(\tilde{\pi}^n)$ with horizon T. $C_{0,n} = 0 \le A(0,n) = 0$; $C_{T,0} = C(\pi^\star) \le A(T,0) = 0$.

Assume that for all $t + n \le k$ we have $C_{t,n} \le A(t,n)$. For $t_1 + n_1 = k + 1$, consider $\tilde{\pi}^n = S_{X_n}(\hat{\pi}^n, \tilde{\pi}^{n-1})$ where $X_n + 1$ $Geom(1-\alpha)$:

1. $X_n = 0(\Pr = 1-\alpha)$: The policy immediately switches to $\tilde{\pi}^{n-1}$ with $C_{t,n-1}$. By our assumption, the cost is
$$\mathbb{E}[C_{t,n}|X_n = 0] = (1-\alpha) \times A(t, n-1)$$
2. $X_n \ge 1(\Pr = \alpha)$: The policy acts at the first step with ε probability of failing.

   If the policy fails the first step, the error for the entire trajectory is bounded by
$$\mathbb{E}[C_{t,n}|X_n \ge 1, \text{fails on step 1}] \le \alpha\varepsilon T$$
   If the policy succeeds on the first step, it matches the expert's action at step 1. For the remaining t-1 steps, the policy becomes $S^{X_n-1}(\hat{\pi}^n, \tilde{\pi}^{n-1})$ because the memoryless property of Geometric Distribution: the Distribution of $X_n - 1$ is identical to X_n given $X_n \ge 1$. So the cost is
$$\mathbb{E}[C_{t,n}|X_n \ge 1, \text{succeeds on step 1}] \le \alpha \times (1-\varepsilon)A(t-1,n)$$

Adding up all the costs we got:
$$C_{t,n} \le \alpha\varepsilon T + \alpha(1-\varepsilon)A(t-1,n) + (1-\alpha)A(t,n-1) = A(t,n)$$
Setting $T = t$, we concluded:
$$C(\tilde{\pi}^n) \le A(T,n)$$
□

Problem 5.2

Proof. We prove by induction.

1. Base cases:
   1. $t = 0$: $A(0, n) = 0 \leq 0 \times n\alpha\varepsilon$
   2. $n = 0$: $A(t, 0) = 0 \leq 0 \times t\alpha\varepsilon$
2. Inductive hypothesis:
   For any $t + n \leq k$, $A(t, n) \leq Tn\alpha\varepsilon$.
3. Induction:
   For any $t + n \leq k + 1$,

$$\begin{aligned} C_{t,n} &\leq A(t, n) \\ &= \alpha\varepsilon t + \alpha(1-\varepsilon)A(t-1, n) + (1-\alpha)A(t, n-1) \end{aligned}$$

Apply the inductive hypothesis:

$$\begin{aligned} A(t, n) &\leq \alpha\varepsilon t + \alpha(1-\varepsilon)(t-1)n\alpha\varepsilon + (1-\alpha)t(n-1)\alpha\varepsilon \\ &= \alpha\varepsilon \times (t + n\alpha(t - 1 - \varepsilon t + \varepsilon) + t(n - 1 - \alpha n + \alpha)) \\ &= \alpha\varepsilon \times (t + tn\alpha - n\alpha - tn\alpha\varepsilon + n\alpha\varepsilon + tn - t - tn\alpha + t\alpha) \\ &= \alpha\varepsilon \times (n\alpha\varepsilon - n\alpha - tn\alpha\varepsilon + tn + t\alpha) \end{aligned}$$

Ignore the scale term $\alpha\varepsilon$, we get

$$\begin{aligned} & n\alpha\varepsilon - n\alpha - tn\alpha\varepsilon + tn + t\alpha \\ =& tn + \alpha \times ((t - n) + (1 - t)n\varepsilon) \end{aligned}$$

In switchDagger, we let $n \geq t \geq 1$. And then $t - n \leq 0$, $1 - t \leq 0$. Hence, we conclude:

$$C(\widetilde{\pi}^n) = C_{T,n} \leq A(T, n) \leq Tn\alpha\varepsilon$$

as we desired.

□

Problem 5.3

Proof. First, for any policy π, $C(\pi) \leq \sum_{t=1}^{T} \max \mathbb{E}_{s_t \sim p_\pi} \Pr[\pi(s_t) \neq \pi^\star(s_t)] \leq T$. π^n is policy that transfers control from $\widetilde{\pi}^n$ to expert policy $\pi^\star$ at step $X^\star$. If $X^\star \geq T$, $\pi^n = \widetilde{\pi}^n$. Hence,

$$\begin{aligned} C(\pi^n) - C(\widetilde{\pi}^n) &\leq \Pr[X^\star \leq T] \times T \\ &= e^{\frac{-n}{(1-\alpha)T}} \times T \end{aligned}$$

It follows that

$$C(\pi^n) \leq C(\widetilde{\pi}^n) + e^{\frac{-n}{(1-\alpha)T}}T$$

as we desired. □

Problem 5.4

In summary we get the upper bound of policy:

$$C(\pi^N) \leq TN\alpha\varepsilon + e^{\frac{-n}{(1-\alpha)T}}T$$

Let $\alpha = 1/T$ and $N = Tlog(1/\varepsilon)$.Substituting into the formula:

$$C(\pi^N) \leq T\varepsilon log(1/\varepsilon) + T\varepsilon^{T/(T-1)} = \mathcal{O}(T\varepsilon(log(1/\varepsilon + 1)))$$

Therefore:

$$C(\pi^N) = \mathcal{O}(T\varepsilon log(1/\varepsilon))$$

as we desired.