285 HW1

Charles Xu

September 2023

1 Question 1

1.1 Question 1.1

Let E_i be the event that the policy π_θ makes a mistake and the of a mistake is bounded by ϵ Then the probability of making at least 1 mistake over t steps is $Pr\left[\bigcup_{i=1}^{t} E_i\right]$.

Using the union bound,

$$Pr\left[\bigcup_{i=1}^{t} E_i\right] \leq \sum_{i=1}^{t} Pr\left[E_i\right] \leq t\epsilon$$

We can write the probably of being at state s_t as

$$p_\theta\left(s_t\right) = \left(1 - Pr\left[\bigcup_{i=1}^{t} E_i\right]\right) p_{\pi^*}\left(s_t\right) + Pr\left[\bigcup_{i=1}^{t} E_i\right] p_{mistake}\left(s_t\right)$$

where $p_{\pi^*}\left(s_t\right)$ is the probably of being at state s_t following the optimal policy, and $p_{mistake}$ is the probably of being at state s_t given some mistake

Following similar step as in lecture, we can rearrange and get

$$\left|p_\theta\left(s_t\right) - p_{\pi^*}\left(s_t\right)\right| = Pr\left[\bigcup_{i=1}^{t} E_i\right] \left|p_{mistake}\left(s_t\right) - p_{\pi^*}\left(s_t\right)\right| \leq 2Pr\left[\bigcup_{i=1}^{t} E_i\right] = 2t\epsilon$$

$$\Sigma_{s_t}\left|p_\theta\left(s_t\right) - p_{\pi^*}\left(s_t\right)\right| \leq 2T\epsilon$$

.

1.2 Question 1.2

1.2.1 Question 1.2a

$$J\left(\pi^*\right) = E_{p_{\pi^*}\left(s_T\right)}\left[r\left(s_T\right)\right]$$

$$J\left(\pi_\theta\right) = E_{p_{\pi_\theta}\left(s_T\right)}\left[r\left(s_T\right)\right]$$

Difference in reward:

$$J(\pi^*) - J(\pi_\theta) = E_{p_{\pi^*}(s_T)}\left[r(s_T)\right] - E_{p_{\pi_\theta}(s_T)}\left[r(s_T)\right]$$

If the reward only depends on the last state, we can use the probably $p_\theta(s_T) \leq 2T\epsilon$. So in the worst case, the policy makes a mistake along the trajectory and ends in a state resulting in no reward with probably at most $2T\epsilon$.

$$J(\pi^*) - J(\pi_\theta) \leq J(\pi^*) - ((1 - (2T\epsilon))J(\pi^*) = 2T\epsilon J(\pi^*) \leq 2T\epsilon R_{max}$$

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

1.2.2 Question 1.2b

For an arbitrary reward, the difference in reward is the sum of difference in reward from timestep 1 to T, given by:

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

$$J(\pi^*) - J(\pi_\theta) \leq \sum_{t=1}^{T} 2t\epsilon R_{max}$$

$$J(\pi^*) - J(\pi_\theta) \leq 2T^2\epsilon R_{max}$$

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

2 BC

2.1 Question 3.1

Table 1: BC Evaluation Results

Environment	Eval Average Return	Eval Std Return
Ant-v4	1019	190
Hopper-V4	214	169

Table 2: Hidden size [64, 64]. 1e6 training iterations.

2.2 Question 3.2

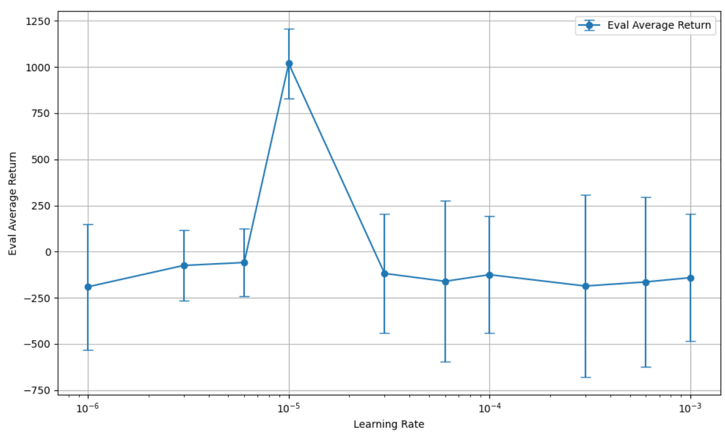

Figure 1: Effect of learning rate on average return in teh Ant-v4 task. I chose to sweep the learning rate because I found that this was a big factor on how well my model trained. With the default value, the loss quickly diverges. With too small learning rate, it takes more steps to converge.

3 Dagger

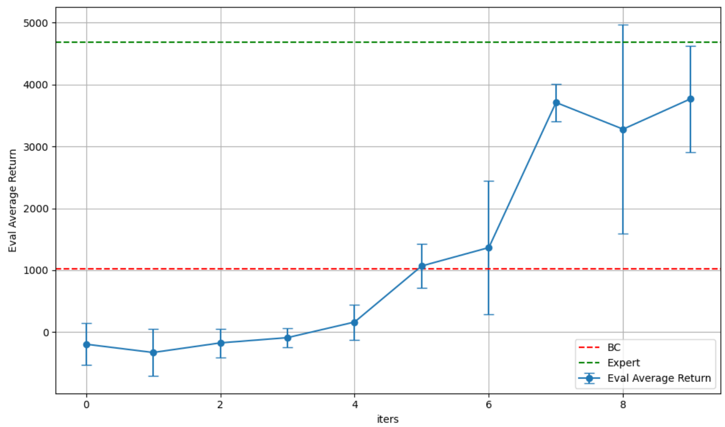

Figure 2: Dagger average return compared to the expert and flat BC policies across iterations on the Ant-v4 task.

4 Discussion

1. I spend 6 hours on this assignment
2. No