1. Analysis

1.

Using the definition of expectation, we can rewrite the condition as:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\pi^*(s_t)}\pi_\theta(a_t \neq \pi^*(s_t)|s_t) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s_t}\pi^*(s_t)\,\pi_\theta(a_t \neq \pi^*(s_t)|s_t) \leq \varepsilon$$

$$\overset{T>0}{\Longrightarrow}\sum_{t=1}^{T}\sum_{s_t}\pi^*(s_t)\,\pi_\theta(a_t \neq \pi^*(s_t)|s_t) \leq T\varepsilon$$

One the other hand, the state distribution under p_{π_θ} at timestep t can be expressed as a mixture of the distribution if no mistakes were made (i.e., the distribution under the expert policy) and the (unknown) distribution resulting from mistakes:

$$p_{\pi_\theta}(s_t) = \left(1 - \mathbb{P}_\theta\left[\bigcup_{t'=1}^{t} a_{t'} \neq \pi^*(s_{t'})|s_{t'}\right]\right)p_{\pi^*}(s_t) + \mathbb{P}_\theta\left[\bigcup_{t'=1}^{t} a_{t'} \neq \pi^*(s_{t'})|s_{t'}\right]p_{mistake}(s_t)$$

Using the above equation and the union bound inequality, we can show that:

$$\begin{aligned}\sum_{s_t}\left|p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)\right| &= \sum_{s_t}\mathbb{P}_\theta\left[\bigcup_{t'=1}^{t} a_{t'} \neq \pi^*(s_{t'})|s_{t'}\right]|p_{mistake}(s_t) - p_{\pi^*}(s_t)| \\ &\leq 2\sum_{s_t}\mathbb{P}_\theta\left[\bigcup_{t'=1}^{t} a_{t'} \neq \pi^*(s_{t'})|s_{t'}\right] \leq 2\sum_{s_t}\sum_{t'=1}^{t}\pi_\theta(a_{t'} \neq \pi^*(s_{t'})|s_{t'}) \\ &\leq 2\sum_{t'=1}^{T}\sum_{s_t}\pi_\theta(a_{t'} \neq \pi^*(s_{t'})|s_{t'}) \leq ?\, 2T\varepsilon\end{aligned}$$

2.a.

$$J(\pi) = \sum_{t=1}^{T}\mathbb{E}_{p_\pi(s_t)}r(s_t) = \sum_{t=1}^{T}\sum_{s_t}p_\pi(s_t)r(s_t)$$

Given that $r(s_t) = 0$ for all $t < T$:

$$J(\pi) = \sum_{s_T}p_\pi(s_T)r(s_T)$$

Combining this with the conclusion from part 1, we obtain:

$$J(\pi^*) - J(\pi_\theta) = \sum_{s_T}\left|p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)\right|r(s_T) \leq R_{max}\sum_{s_T}\left|p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)\right| \leq R_{max}2T\varepsilon \in O(T\varepsilon)$$

2.b.

Similarly, we can show:

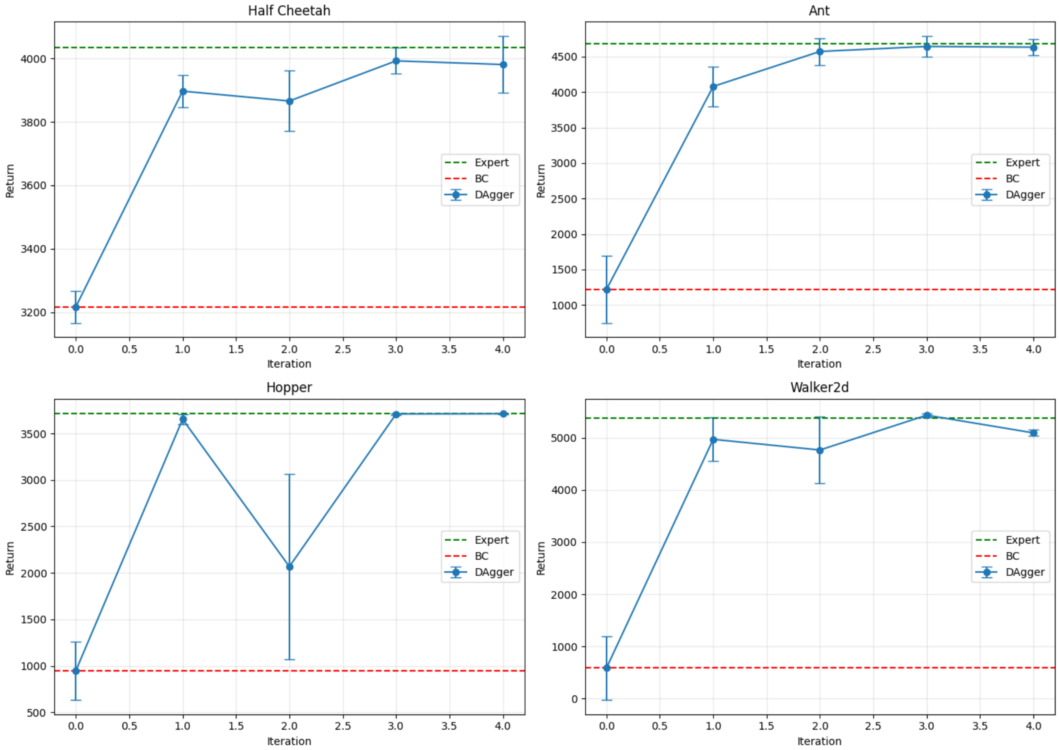
$$J(\pi^*) - J(\pi_\theta) = \sum_{t=1}^{T} \sum_{s_t} \left| p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T) \right| r(s_t) \leq \sum_{t=1}^{T} \sum_{s_t} \left| p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T) \right| R_{max} \leq R_{max} \sum_{t=1}^{T} 2t\varepsilon$$

$$= 2\varepsilon R_{max} \sum_{t=1}^{T} t = 2\varepsilon R_{max} \frac{T(T+1)}{2} \in O(T^2 \varepsilon)$$

2. Behavioural Cloning vs. DAgger

In the Half Cheetah task, Behavioural Cloning performs pretty well, achieving nearly 80% of the expert's performance. However, in the other tasks —particularly Walker2d, which seems to be more complicated than the others— it performs very poorly, failing to achieve even 30% of the expert's average return.

In general, it is evident that DAgger outperforms BC in all tasks by addressing the distributional shift problem.


Half Cheetah
Ant
Hopper
Walker2d

Noteworthy hyperparameters of the model are as follows (consistent across all tasks):

Episode length: 1000
Number of DAgger iterations: 5
Evaluation batch size: 5000
Network architecture (default)

- Number of layers: 2
- Hidden layer size: 64
- Learning rate: 5e-3