Assignment 1: Imitation Learning

Yasin Sonmez

September 11, 2023

1 Analysis

1. Defining $P_{no-mistake}(t)$ as the probability of the policy not making any mistake for the first t timesteps. Similar to what we did in the lecture with the more strict bounds, we can write:

$$p_{\pi_\theta}(s_t) = P_{no-mistake}(t)p_{train}(s_t) + (1 - P_{no-mistake}(t))p_{mistake}(s_t)$$

$$p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t) = (1 - P_{no-mistake}(t))(p_{mistake}(s_t) - p_{train}(s_t))$$

$$|p_{\pi_\theta}(s) - p_{\pi^*}(s)| = (1 - P_{no-mistake}(t))|(p_{mistake}(s) - p_{train}(s))|$$

$$|p_{\pi_\theta}(s) - p_{\pi^*}(s)| = \sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2(1 - P_{no-mistake}(t))$$

The last inequality is- due to the fact that TDV can have at most value 2. Now we need to find a bound on $P_{no-mistake}(t)$. Let e_t and $e_t^{'}$ be the probability of making a mistake at time step t conditioned on no mistake made before timestep t and at least one step made before timestep t. Therefore we have:

$$P_{no-mistake}(t) = (1 - e_t)P_{no-mistake}(t-1) \tag{1}$$

We also have

$$\mathbb{E}_{p_{\pi^*(s_t)}}\pi_\theta(a \neq \pi^*(s_t) \mid s_t) = P_{no-mistake}(t-1)e_t + (1 - P_{no-mistake}(t-1))e_t^{'}$$

$$\implies P_{no-mistake}(t-1)e_t \leq \mathbb{E}_{p_{\pi^*(s_t)}}\pi_\theta(a \neq \pi^*(s_t) \mid s_t) \tag{2}$$

Combining eq. 1 and inequality 2, we have:

$$P_{no-mistake}(t) = (1 - e_t)P_{no-mistake}(t-1) \geq P_{no-mistake}(t-1) - \mathbb{E}_{p_{\pi^*(s_t)}}\pi_\theta(a \neq \pi^*(s_t) \mid s_t)$$

If we continue the above inequality by expanding the terms similarly we get:

$$P_{no-mistake}(t) \geq P_{no-mistake}(0) - \sum_{t=1}^{T} \mathbb{E}_{p_{\pi^*(s_t)}}\pi_\theta(a \neq \pi^*(s_t) \mid s_t) = 1 - \sum_{t=1}^{T} \mathbb{E}_{p_{\pi^*(s_t)}}\pi_\theta(a \neq \pi^*(s_t) \mid s_t)$$

$$\implies 1 - P_{no-mistake}(t) \leq \sum_{t=1}^{T} \mathbb{E}_{p_{\pi^*(s_t)}}\pi_\theta(a \neq \pi^*(s_t) \mid s_t) \leq \epsilon T$$

Therefore we have:

$$|p_{\pi_\theta}(s) - p_{\pi^*}(s)| = \sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2(1 - P_{no-mistake}(t)) \leq 2\epsilon T$$

2.

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

$$\leq \sum_{t=1}^{T} \sum_{s_t} (p_{\pi_\theta} + (p_{\pi^*} - p_{\pi_\theta}))r(s_t) \leq \sum_{t=1}^{T} \sum_{s_t} p_{\pi_\theta} r(s_t) + \sum_{t=1}^{T} \sum_{s_t} |p_{\pi^*} - p_{\pi_\theta}| r(s_t) \tag{3}$$

(a) From the definition of expected return and using the results from the above problem and the fact that reward is only at the last step:

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

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

(b) From the definition of expected return and using the results from the above problem:

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

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

2 Editing Code

3 Behavioral Cloning

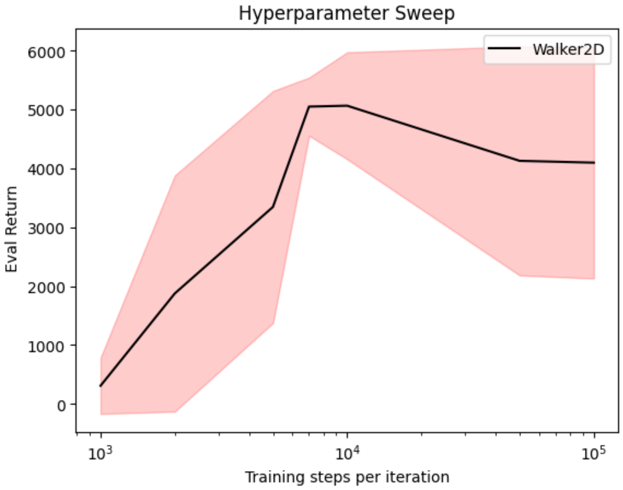
1. In table 1 we have results for two different environments with the default parameters.

Environment Name	Expert Result	BC
Ant-v4	4682	4536 ± 596
Walker2D	5383	309 ± 477

Table 1: ep_len=10^3, eval_batch_size= 10^5. Mean and standard deviation were reported. All the hyperparameters are the default parameters

2. In fig. 1 we have hyperparameter sweep results in a plot.

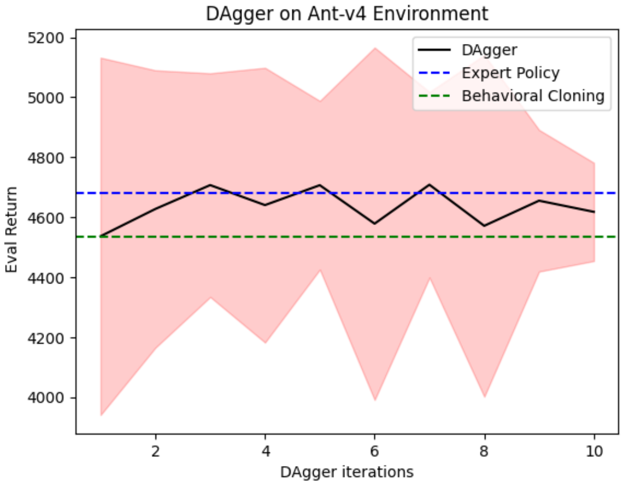
Figure 1: Hyperparameter sweep for the hyperparameter "τ = num_agent_train_steps_per_iter". ep_len=10^3, eval_batch_size= 10^5. This parameter was chosen since the agent performed poorly with the default hyperparameters and since the environment is more complex than good performing environments in the previous question it made sense to train for more steps per iteration to learn the underlying structure better. For all the other hyperparameters, the default hyperparameters are used



4 DAgger

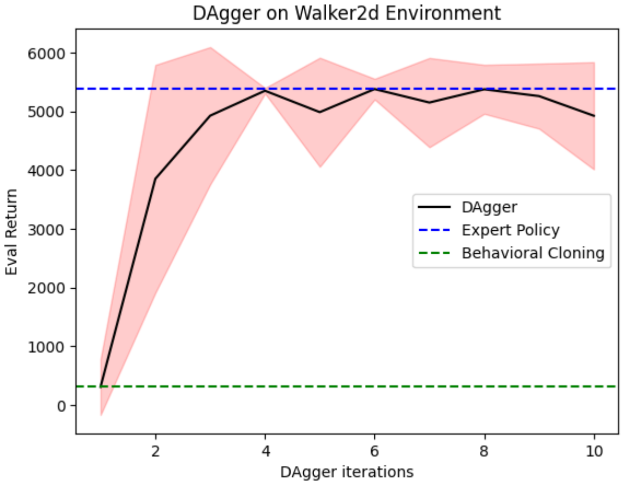
In fig. 2 we have results for the DAgger algorithm on "Ant-v4" environment.

Figure 2: DAgger result for the "Ant-v4" environment. ep_len=10^3, eval_batch_size= 10^5, For all other hyperparameters, the default hyperparameters are used. Shaded regions are the standard deviation



In fig. 3 we have results for the DAgger algorithm on "Walker2D" environment

Figure 3: DAgger result for the "Walker2D" environment. ep_len=10^3, eval_batch_size= 10^5, For all other hyperparameters, the default hyperparameters are used. Shaded regions are the standard deviation



5 Discussion

1.
   - Part 1.1: 3 hours
   - Part 1.2: 1 hour
   - Part 2: 2 hours
   - Part 3: 2 hours
   - Part 4: 2 hours