CS 285

Homework I

Mustafa Suman (ID: 3039767187)

September 12, 2023

1 Analysis

1.1 Assume $\mathbb{E}_{p_{\pi^*}(s)}\pi_\theta(a \neq \pi^*(s) \mid s) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{p_{\pi^*}(s_t)}\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) \leq \varepsilon,$

show that $\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \overset{!}{\leq} 2T\varepsilon.$

Proof: Similary to the lecture, we define

$$p_{\pi_\theta}(s_t) = p\left(\substack{\text{slip off expert policy} \\ \text{at some } t}\right) \cdot p_{\text{mistake}}(s_t) + \left(1 - p\left(\substack{\text{slip off expert policy} \\ \text{at some } t}\right)\right) \cdot p_{\pi^*}(s_t)$$

Slipping off the trajectory at some t is, according to

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ smaller than $P(A) + P(B)$, therefore

$$p\left(\substack{\text{slip off expert policy} \\ \text{at some } t}\right) \leq \sum_t \sum_{s_t} \pi^*(a \neq \pi^*(s_t) \mid s_t) \cdot p_{\pi^*}(s_t)$$

$$\overset{\text{Assumption}}{=} T\varepsilon$$

Putting all things together, we can derive

$$\text{LHS} = \sum_{s_t} p\left(\substack{\text{slip off expert policy.} \\ \text{at some } t}\right) | p_{\text{mistake}}(s_t) - p_{\pi^*}(s_t) | \leq T\varepsilon \sum_{s_t} | p_{\text{mistake}}(s_t) - p_{\pi^*}(s_t) |$$

$$\leq 2T\varepsilon = \text{RHS},$$

since $\sum_{s_t} | p_{\text{mistake}}(s_t) - p_{\pi^*}(s_t) | \leq 2$

$\blacksquare$

1.2.a:

$$J(\pi^*) - J(\pi_\theta) \overset{\text{Def.}}{=} \mathbb{E}_{p_{\pi^*}(s_T)}\left[r(s_T)\right] - \mathbb{E}_{p_{\pi_\theta}(s_t)}\left[r(s_T)\right]$$

$$\overset{\text{Def.}}{=} \sum_{s_T} \left(p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)\right) \cdot r(s_T)$$

$$\leq \sum_{s_T} \left| p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)\right| \cdot \left| r(s_T)\right| \overset{1.1}{\leq} 2T\varepsilon \cdot R_{max} = O(T\varepsilon)$$

1.2.b:

$$J(\pi^*) - J(\pi_\theta) = \sum_{t=1}^{T} \mathbb{E}_{p_{\pi^*}(s_t)}\left[r(s_t)\right] - \mathbb{E}_{p_{\pi_\theta}(s_t)}\left[r(s_t)\right]$$

$$= \sum_{t=1}^{T} \sum_{s_t} \left| p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)\right| \cdot |R_{max}|$$

$$\overset{1.1}{\leq} \sum_{t=1}^{T} 2\varepsilon T \cdot |R_{max}| = 2\varepsilon T^2 \cdot |R_{max}| = O(\varepsilon T^2)$$

3 Behavioral Cloning

3.1 Comparing different environments

3.1.1 Successful environment: Ant-v4

Benchmark Expert Policy:

Metric	Value
Train_AverageReturn	4681.89
Train_StdReturn	30.71
Train_MaxReturn	4712.60
Train_MinReturn	4651.18
Train_AverageEpLen	1000.0

Table 1: Number of rollouts = 2

Trained Policy:

Metric	Value
Eval_AverageReturn	2455.73
Eval_StdReturn	547.68
Eval_MaxReturn	3089.16
Eval_MinReturn	1410.66
Eval_AverageEpLen	945.17

Table 2: Hyperparams are set as num_agent_train_steps_per_iter = 1300, batch_size = 1000, eval_batch_size = 5000, train_batch_size = 100, n_layers = 2, size=64, lr=5e-3

The corresponding `log_file` can be found under the name `q1_bc_ant_Ant-v4_11-09-2023_12-41-32`.

3.1.2 Unsuccessful environment: Hopper-v4

Expert Policy:

Metric	Value
Train_AverageReturn	3717.51
Train_StdReturn	0.35
Train_MaxReturn	3717.87
Train_MinReturn	3717.16
Train_AverageEpLen	1000.00

Table 3: Number of rollouts = 2

Trained Policy:

The corresponding `log_file` can be found under the name `q1_bc_hopper_Hopper-v4_11-09-2023_13-21-22`.

Metric	Value
Eval_AverageReturn	1056.42
Eval_StdReturn	325.34
Eval_MaxReturn	1747.99
Eval_MinReturn	235.93
Eval_AverageEpLen	314.00

Table 4: Hyperparameters are set as in the case before for a fair comparison.

3.2 Hyperparameter Experiment

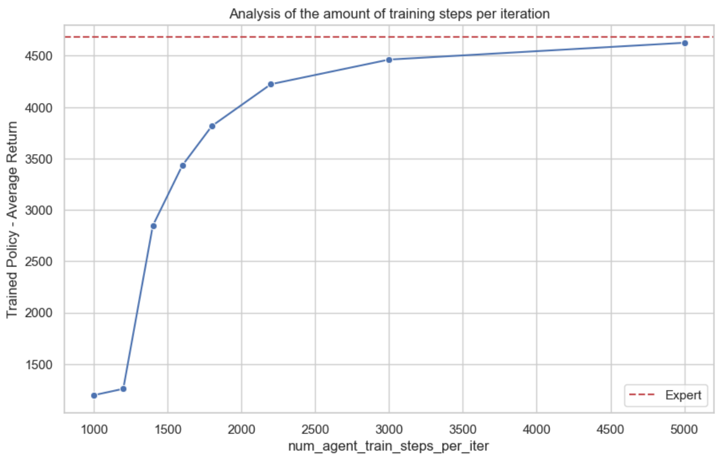

Figure 1: The Hyperparameter we are considering is the about the number of training steps per iteration, i.e. the amount of training in the BC setting.

I chose this parameter, because the impact on the success of the policy is strong as we can see in the figure. With this parameter, I was able to increase the average return in the previous task to achieve a sufficiently high reward.

4 DAGGER

4.1 Ant-v4

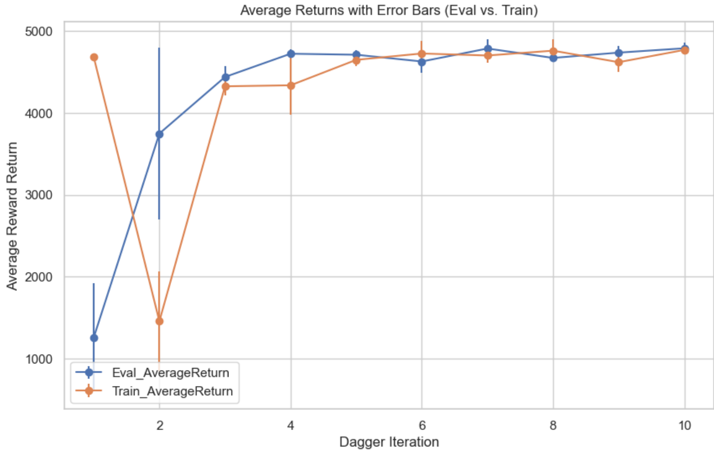

Figure 2: Hyperparams are set as num_agent_train_steps_per_iter = 1200, batch_size = 5000, eval_batch_size = 5000, train_batch_size = 100, n_layers = 2, size=64, lr=5e-3.

The corresponding `log_file` is `q2_dagger_ant_Ant-v4_11-09-2023_17-15-00`.

4.2 Hopper-v4

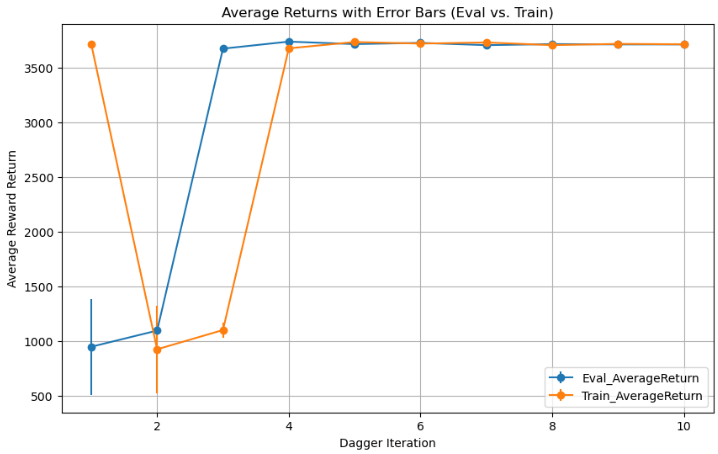

Figure 3: Hyperparams are set as in the figure above.

The corresponding `log_file` is `q2_dagger_hopper_Hopper-v4_11-09-2023_17-53-33`.

6 Discussion

6.1 1. How much time did you spend on each part of this assignment?

Around 9 hours for the theoretical part and ca. 16 hours for the implementation (including setting everything up etc.)

6.2 2. Any additional feedback?

Very nice code and good introduction to RL coding. Creating the plot and figuring out how to store the logs was tedious and took way to much time. Also it is annoying to write a README file when one does not change the funcionality of the program and already writes down the params in the pdf.