Homework 1 Solutions

Viswesh Nagaswamy Rajesh
CS285: Deep Reinforcement Learning

1 Problem 1

Consider the problem of imitation learning within a discrete MDP with horizon T and an expert policy π^*. We gather expert demonstrations from π^* and fit an imitation policy π_θ to these trajectories so that

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

i.e., the expected likelihood that the learned policy π_θ disagrees with the expert π^* within the training distribution p_{π^*} of states drawn from random expert trajectories is at most ε.

For convenience, the notation $p_\pi(s_t)$ indicates the state distribution under π at time step t while $p(s)$ indicates the state marginal of π across time steps, unless indicated otherwise.

1.1 Question 1.1

Show that $\sum_t |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2T\varepsilon$.

Solution

We aim to show the total variation distance between the state distributions under the learned policy π_θ and the expert policy π^*. Given that $\mathbb{E}_{p_{\pi^*}(s_t)}\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) \leq \varepsilon$ for all $s_t \sim p_{\pi^*}(s_t)$.

First, consider the probability of making a mistake at each time step t. By the definition, the expected likelihood that the learned policy π_θ disagrees with the expert policy π^* is at most ε:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{p_{\pi^*}(s_t)}\left[\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)\right] \leq \varepsilon.$$

Using the stronger assumption that $\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) \leq \varepsilon$ for every $s_t \in \mathrm{supp}(p_{\pi^*})$, we can sum this over all time steps:

$$\sum_{t=1}^{T}\sum_{s_t} p_{\pi^*}(s_t)\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) \leq T\varepsilon.$$

Applying the union bound, the total variation distance between the distributions $p_{\pi_\theta}(s_t)$ and $p_{\pi^*}(s_t)$ is given by:

$$\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2 \sum_{t=1}^{T} \sum_{s_t} p_{\pi^*}(s_t) \pi_\theta(a_t \neq \pi^*(s_t) \mid s_t).$$

Substituting the upper bound from the previous step:

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

Thus, we have shown that:

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

1.2 Question 1.2

Consider the expected return of the learned policy π_θ for a state-dependent reward $r(s_t)$, where we assume the reward is bounded with $|r(s_t)| \leq R_{\max}$:

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

(a) Show that $J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T\varepsilon)$ when the reward only depends on the last state, i.e., $r(s_t) = 0$ for all $t < T$.

(b) Show that $J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T^2\varepsilon)$ for an arbitrary reward.

Solution

(a) When $r(s_t) = 0$ for all $t < T$, the reward only depends on the last state:

$$J(\pi) = \mathbb{E}_{p_\pi(s_T)} r(s_T).$$

Therefore:

$$J(\pi^*) = \mathbb{E}_{p_{\pi^*}(s_T)} r(s_T),$$
$$J(\pi_\theta) = \mathbb{E}_{p_{\pi_\theta}(s_T)} r(s_T).$$

The difference in returns:

$$J(\pi^*) - J(\pi_\theta) = \left| \mathbb{E}_{p_{\pi^*}(s_T)} r(s_T) - \mathbb{E}_{p_{\pi_\theta}(s_T)} r(s_T) \right|.$$

Since $|r(s_T)| \leq R_{\max}$ and using the bound on the state distribution discrepancy:

$$\left| \mathbb{E}_{p_{\pi^*}(s_T)} r(s_T) - \mathbb{E}_{p_{\pi_\theta}(s_T)} r(s_T) \right| \leq R_{\max} \sum_{s_T} |p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T)|.$$

From Question 1.1, we know that:

$$\sum_{s_T} |p_{\pi_\theta}(s_T) - p_{\pi^*}(s_T)| \leq 2T\varepsilon.$$

Therefore,

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

(b) For an arbitrary reward, we have:

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

The difference in returns:

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

Using the bound on $|r(s_t)| \leq R_{\max}$:

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

From Question 1.1, we know that:

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

Therefore,

$$|J(\pi^*) - J(\pi_\theta)| \leq R_{\max} \cdot 2T\varepsilon \cdot T = \mathcal{O}(T^2\varepsilon).$$

2 Table 1: Mean and Standard Deviation of Returns for Training and Evaluation

Train Steps	Eval Mean Return	Eval Std Return	Train Mean Return	Train Std Return
500	4033.012	186.307	4681.892	30.709
1000	4511.322	113.388	4681.892	30.709
1500	4582.842	44.019	4681.892	30.709
2000	4630.633	154.208	4681.892	30.709

Table 1: Mean and standard deviation of returns for training and evaluation across different train steps.

3 Figures

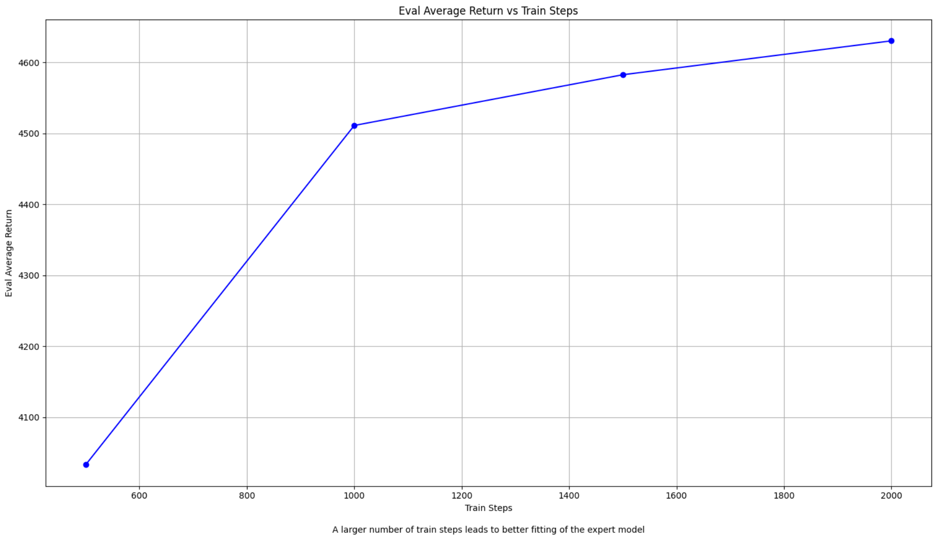

Figure 1: Plot of Evaluation and Training Returns for Experiment 2

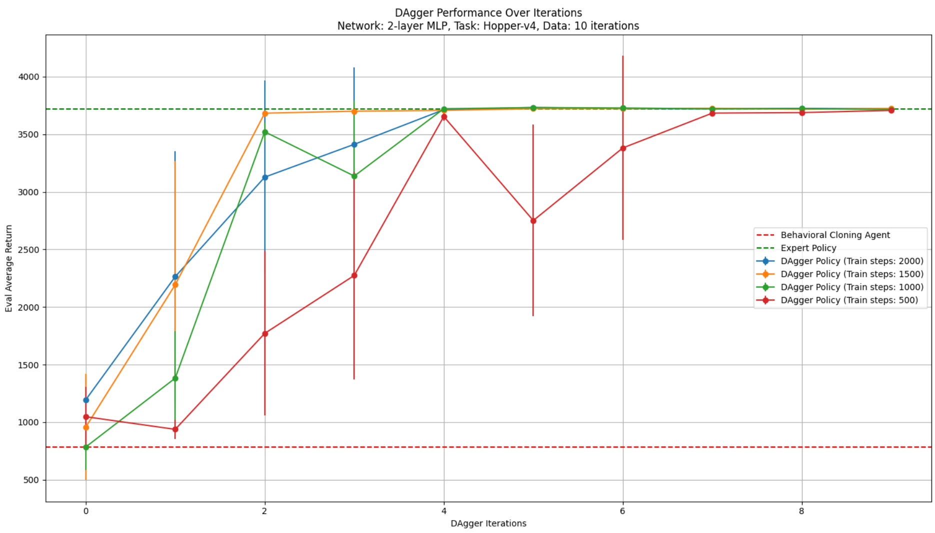

Figure 2: Plot of Evaluation and Training Returns for Experiment 3