Chanjoon Park
Feb 10, 2024

CS 285: Deep Reinforcement Learning, Decision Making, and Control
Assignment 1. Imitation Learning

1. Analysis

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

   Hint 1: in lecture, we showed a similar inequality under the stronger assumption $\pi_\theta(s_t \neq \pi^*(s_t) \mid s_t) \leq \varepsilon$ for every $s_t \in \mathrm{supp}(p_{\pi^*})$. Try converting the inequality above into an expectation over p_{π^*}

   Hint 2: use the union bound inequality: for a set of events E_i, $\Pr[\bigcup_i E_i] \leq \sum_i \Pr[E_i]$

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.

Solutions.

1.1.

Consider a discrete MDP, horizon T, an expert policy π_E, and a policy π that is parameterized by θ.

First, we have given boundaries of the expected likelihood as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{p^*_\pi(s_t)}\pi_\theta(a_t \neq \pi^*(s)|s) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s_t}p^*_\pi(s_t)\pi_\theta(a \neq \pi^*(s)) \leq \epsilon \tag{1}$$

And using the union bound inequality, we can derive the following:

$$\sum_{s_t}p^*_\pi(s_t)\pi_\theta(a \neq \pi^*(s)) \leq \epsilon \tag{2}$$

Second, derive the probability of the policy π_θ like Lecture 2.

$$p_{\pi_\theta}(s_t) = (1-\epsilon)^t p_{\pi^*}(s_t) + (1-(1-\epsilon)^t)p_{\pi_{\text{mistake}}}(s_t) \tag{3}$$

Finally, show that $\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \le 2T\epsilon$.

$$\begin{aligned}
\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| &= \sum_{s_t} |(1-\epsilon)^t p_{\pi^*}(s_t) + (1-(1-\epsilon)^t)p_{\pi_{\text{mistake}}}(s_t) - p_{\pi^*}(s_t)| \\
&= \sum_{s_t} |(1-(1-\epsilon)^t)p_{\pi_{\text{mistake}}}(s_t)| \\
&\le \sum_{s_t} (1-(1-\epsilon t))|p_{\pi_{\text{mistake}}}(s_t) - p_{\pi^*}(s_t)| \quad (\because (1-\epsilon)^t \ge 1-\epsilon t \text{ for } \epsilon \in [0,1]) \\
&\le \sum_{s_t} 2(1-(1-\epsilon t)) = \sum_{s_t} 2\epsilon t = 2T\epsilon \quad (\because \sum_{s_t} t = T)
\end{aligned}$$

Note: I assumed that sum of t over s_t is same as T.

1.2.

(a) Show that $J(\pi^*) - J(\pi_\theta) = O(T\epsilon)$ when $r(s_t) = 0$ for all $t < T$

$$\begin{aligned}
J(\pi^*) - J(\pi_\theta) &= \sum_{t=1}^{T} \mathbb{E}_{p_\pi^*}(s_t)r(s_t) - \sum_{t=1}^{T} \mathbb{E}_{p_{\pi_\theta}}(s_t)r(s_t) \\
&= \sum_{t=1}^{T} \Big(\sum_{s_t} (p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t))r(s_t) \Big) \\
&= (p_{\pi^*}(s_T) - p_{\pi_\theta}(s_T))r(s_T) \le 2T\epsilon R_{\max}
\end{aligned}$$

Thus, $J(\pi^*) - J(\pi_\theta) = O(T\epsilon)$

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

$$\begin{aligned}
J(\pi^*) - J(\pi_\theta) &= \sum_{t=1}^{T} \mathbb{E}_{p_\pi^*}(s_t)r(s_t) - \sum_{t=1}^{T} \mathbb{E}_{p_{\pi_\theta}}(s_t)r(s_t) \\
&= \sum_{t=1}^{T} \Big(\sum_{s_t} (p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t))r(s_t) \Big) \\
&\le \sum_{t=1}^{T} \Big(\sum_{s_t} |p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)| \cdot R_{max} \Big) \\
&\le \sum_{t=1}^{T} 2T\epsilon R_{max} = 2T^2\epsilon R_{max}
\end{aligned}$$

Thus, $J(\pi^*) - J(\pi_\theta) = O(T^2\epsilon)$