Assignment 1: Imitation Learning

Due September 11, 11:59 pm

1 Analysis Solutions

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 $p_\pi(s)$ 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 $\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) \leq \varepsilon$ for every $s_t \in \text{supp}(p_{\pi^*})$. 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, $\Pr[\bigcup_i E_i] \leq \sum_i \Pr[E_i]$

   Solution.

   Define Pr(mistake in t timesteps) as the probability that our policy π_θ disagreed with the optimal policy π^* at least once in the first t timesteps. We can bound,

   $$\mathsf{Pr}(\text{mistake in } t \text{ timesteps}) = \mathsf{Pr}\left(\cup_{i=1}^{t}\text{first mistake at timestep } i\right)$$

   Now, let $E_i = \{\text{first mistake at timestep } i\}$ be the event that π_θ behaves optimally for the first $i-1$ timesteps, but makes a mistake at timestep i. Then, we can use the union bound to get,

   $$\begin{aligned}\mathsf{Pr}(\text{mistake in } t \text{ timesteps}) &= \mathsf{Pr}\left(\cup_{i=1}^{t}\text{first mistake at timestep } i\right)\\ &\leq \sum_{i=1}^{t}\mathsf{Pr}\left(\text{first mistake at timestep } i\right)\\ &= \sum_{i=1}^{t}\sum_{s_t} p_{\pi^*}(s_t)\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)\\ &= \sum_{i=1}^{t}\mathbb{E}_{p_{\pi^*}(s_t)}\left[\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t)\right]\\ &\leq \varepsilon T\,,\end{aligned}$$

   where we first marginalize over all states where the mistake could occur, then utilize the fact that the state occurs with probability $p_{\pi_{\pi^*}}(s_t)$ because π_θ has not made any mistakes yet. The last inequality arrives from using the assumption and that $t \leq T$.

   Now, we can rewrite,

   $$p_{\pi_\theta}(s_t) = \mathsf{Pr}(\text{mistake in } t \text{ timesteps})\tilde{p}(s_t) + (1 - \mathsf{Pr}(\text{mistake in } t \text{ timesteps}))p_{\pi^*}(s_t)\,,$$

   where $\tilde{p}$ is any arbitrary distribution over states. Finally, we have,

   $$\sum_{s_t}|p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq \mathsf{Pr}(\text{mistake in } t \text{ timesteps})\sum_{s_t}|p_{\pi^*}(s_t) - \tilde{p}(s_t)| \leq 2T\varepsilon\,,$$

   as desired.

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) Note that

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

in the case where $r(s_t) = 0$ for all $t < T$. We can write the regret as

$$\begin{aligned} J(\pi^*) - J(\pi_\theta) &= \mathbb{E}_{p_{\pi^*}(s_T)} r(s_T) - \mathbb{E}_{p_\pi(s_T)} r(s_T) \\ &= \sum_{s_T} (p_{\pi^*}(s_T) - p_\pi(s_T)) r(s_T) \\ &\leq R_{\max} \sum_{s_T} (p_{\pi^*}(s_T) - p_\pi(s_T)) \\ &\leq 2 R_{\max} T \varepsilon \end{aligned}$$

as desired.

(b) We can write the regret as

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

as desired.