Homework 1

Problem Set

Date: 06 December 2024

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

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

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

   Hint 1: In lecture, we showed a similar inequality under the stronger assumption $\pi_\theta(s_t \neq \pi^*(s_t) \mid s_t) \leq \epsilon$ 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, $\Pr\left[\bigcup_i E_i\right] \leq \sum_i \Pr[E_i]$.

   Solution:

   Let E_t be the event that the learned policy π_θ makes an error at time step t; that is, π_θ takes an action different from the expert policy π^* at time step t. Mathematically,

   $$E_t = \{\pi_\theta(a_t \neq \pi^*(s_t) | s_t)\}.$$

   The probability of at least one error up to time step $t-1$ is bounded by the sum of the probabilities of each error event:

   $$\Pr\left[\bigcup_{k=1}^{t-1} E_k\right] \leq \sum_{k=1}^{t-1} \Pr[E_k].$$

   Note:

- The state distribution at time t under policy π_θ is determined by the sequence of actions taken from time 1 to $t-1$.

The expected error at Time t is:

$$\mathbb{E}_{p_{\pi^*}(s_t)}\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) = \delta_t$$

By definition, $\frac{1}{T}\sum_{t=1}^{T}\delta_t \leq \epsilon$. Therefore, the total expected errors up to time step $t-1$ is bounded by:

$$\sum_{k=1}^{t-1}\mathbb{E}_{p_{\pi^*}(s_t)}\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) \leq \sum_{k=1}^{t-1}\delta_k \leq (t-1)\epsilon.$$

By the total variation distance, we have:

$$\sum_{s_t}|p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2\mathrm{Pr}\left[\bigcup_{k=1}^{t-1}E_k\right] \leq 2(t-1)\epsilon.$$

Note:

- The total variation distance is defined as $d_{TV}(p,q) = \frac{1}{2}\sum_s|p(s) - q(s)|$.
- Errors lead to a difference in the state distribution ($d_{TV}(p,q)$).

Since $t \leq T$, we have:

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

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

Solution:

$$J(\pi^*) - J(\pi_\theta) = \mathbb{E}_{p_{\pi^*}(s_T)}[r(s_T)] - \mathbb{E}_{p_{\pi_\theta}(s_T)}[r(s_T)]$$

Since $|r(s_T)| \leq R_{\max}$, we have:

$$|J(\pi^*) - J(\pi_\theta)| \leq \sup_{s_T}|r(s_T)|d_{TV}(p_{\pi^*}, p_{\pi_\theta}) \leq R_{\max}d_{TV}(p_{\pi^*}, p_{\pi_\theta}) = R_{\max}\cdot T\epsilon$$

Note that:

- For any bounded function f, the difference in expectation under two distributions is bounded by: $|\mathbb{E}_p[f] - \mathbb{E}_q[f]| \leq \sup_x|f(x)|d_{TV}(p,q)$, where $\sup_x$

denotes the supremum (least upper bound) over all possible values of x. In other words, $\sup_x |f(x)|$ is the maximum absolute value that the function f can take.

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

Solution:

$$J(\pi^*) - J(\pi_\theta) = \sum_{t=1}^{T} \left(\mathbb{E}_{p_{\pi^*}(s_t)}[r(s_t)] - \mathbb{E}_{p_{\pi_\theta}(s_t)}[r(s_t)]\right) = R_{\max} \sum_{t=1}^{T} \sum_{s_t} |p_{\pi^*}(s_t) - p_{\pi_\theta}(s_t)| \leq R_{\max} \sum_{t=1}^{T} 2(t$$

Note:

- $\sum_{t=1}^{T} 2(t-1)$ is the sum of an arithmetic series. Let $k = t-1$, then $\sum_{t=1}^{T} 2(t-1) = 2\sum_{k=0}^{T-1} k = 2 \cdot \frac{(T-1)T}{2} = T^2 - T$.

Knowledge Points

Mathematical Formulas

- Union bound: $\Pr\left[\bigcup_i E_i\right] \leq \sum_i \Pr[E_i]$.
  - When considering errors up to time $t-1$, it is essential to sum the error probabilities up to that point, not including time t, because the state at time t depends on actions up to time $t-1$.
- Markov's Inequality: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$.
- Total variation distance: $d_{TV}(p,q) = \frac{1}{2}\sum_s |p(s) - q(s)|$.
- For any bounded function f, $|\mathbb{E}_p[f] - \mathbb{E}_q[f]| \leq \sup_x |f(x)| d_{TV}(p,q)$.
- Sum of an arithmetic series: $\sum_{k=0}^{n} k = \frac{n(n+1)}{2}$.

Moral Behind the Questions

- Error propagation in sequential decision making:
  - Even small discrepancies between a learned policy and an expert policy can lead to significant differences in state distributions over time.
  - Errors made at early time steps can propagate and amplify as the agent continues to make decisions, especially in sequential settings like MDPs.