Lecture 23: Interactive Imitation Learning

1) Dataset Aggregation with DAgger

Setting: Discounted Infinite Horizon MDP

\[ \mathcal{M} = \left\{ S, A, \mathcal{R}, \gamma, \rho \right\} \]

Expert knows optimal policy \( \pi^* \) and we can query the expert at any state during training.

Algorithm: DAgger

Initialize \( \pi^0 \) and dataset \( \mathcal{D} = \emptyset \)

For \( t = 0, \ldots, T-1 \):

1) Generate Dataset with \( \pi^t \) & Query Expert

\[ \mathcal{D}^t = \left\{ s_i, a^*_i \right\} \text{ where } s_i \sim d^t \text{ and } a^*_i = \pi^*(s_i) \]

2) Data Aggregation:

\[ \mathcal{D} = \mathcal{D} \cup \mathcal{D}^t \]

3) Update policy via SL:

\[ \pi^{t+1} = \arg\min_{\pi \in \Pi} \sum_{s,a} e(\pi, s, a) \]
Last lecture we used an assumption that SL succeeds to reason about the performance of imitation learning. Because DAgger aggregates data, we need to consider a slightly different learning framework.

2) Online learning

The online learning setting captures the idea of learning from additional data over time. It is iterative with two components:

for $t=0,1,\ldots$:

1) Learner chooses $\Theta_t$ (i.e. from past data)

2) Suffer the risk $R_t(\Theta_t) = \mathbb{E}_{z \sim D_t}[l(\Theta_t, z)]$ (expected loss)

We care about average regret

$$\frac{1}{T} R(T) = \frac{1}{T} \left[ \sum_{t=0}^{T-1} R_t(\Theta_t) - \min_{\Theta} \sum_{t=0}^{T-1} R_t(\Theta) \right]$$

The baseline for regret is the best learned parameter in hindsight.

Example: supervised learning with $D_t$ a random sample from $D$. This is like ingesting a large dataset one training example at a time (streaming) and we hope that the performance is similar to batch learning.
Why is this different from the SL setting? \( \Theta_t \) (and thus \( R_t \)) can vary in other ways.

Example: in DAGger, we choose \( \pi_t \) and then suffer \( \ell(t_{pi}^t, s_i, \pi^*(s_i)) \) for \( s_i \sim \mathcal{D}_t \).

In this case \( \mathcal{D}_t \) actually depends on \( \Theta_t \).

How should the learner choose \( \Theta_t \)?

**Algorithm: Follow-the-Regularized Leader**

For \( t = 0, 1, \ldots, T-1 \)

\[
\Theta_t = \min_{\Theta} \sum_{k=0}^{t-1} R_k(\Theta) + \lambda f(\Theta)
\]

\[
= \sum_{k=0}^{t-1} \mathbb{E}_{T_k \sim \mathcal{D}_k} \left[ \ell(\Theta, z_k) \right] = \mathbb{E}_{z_k \sim \mathcal{D}_k} \left[ \sum_{k=0}^{t-1} \ell(\Theta, z_k) \right]
\]

**Theorem (FTL):** If losses are convex and regularizer is strongly convex, then even if risks \( R_t \) (i.e. distributions \( \mathcal{D}_t \)) are chosen adversarially,

\[
\max_{R_0, \ldots, R_{T-1}} \frac{1}{T} \left[ \sum_{t=0}^{T-1} R_t(\Theta_t) - \min_{\Theta} \sum_{t=0}^{T-1} R_t(\Theta) \right] = O\left( \sqrt{\frac{1}{T}} \right)
\]
3) Analysis of DAGger

We can view DAGger as an instance of FTL

**Corollary:** If \( \ell(\Pi^t, s, \Pi^*(s)) = 0 \), then

\[
\min_{0 \leq t \leq T} \mathbb{E}_{s \in \mathcal{D}_t} \left[ \ell(\Pi^t, s, \Pi^*(s)) \right] \leq O(1/\sqrt{T}) = \varepsilon_{\text{FTL}}
\]

**Proof:** \( \Pi_t \) plays the roll of \( \Theta_t \), \( (s, \Pi^*(s)) \) is \( z \) & \( Q_t \) is \( d^\Pi_t \)

\[
\min_{0 \leq t \leq T-1} R_t(\Pi_t) \leq \frac{1}{T} \sum_{t=0}^{T-1} R_t(\Pi_t) \quad (\text{min} \leq \text{avg})
\]

\( (\Pi^* \text{ has 0 loss}) \)

\[
\frac{1}{T} \sum_{t=0}^{T-1} R_t(\Pi_t) - R_t(\Pi^*) = \frac{1}{T} \sum_{t=0}^{T-1} \left( \sum_{t=0}^{T-1} R_t(\Pi_t) - \min_{\Pi^*} \sum_{t=0}^{T-1} R_t(\Pi^*) \right)
\]

\( (\Pi^* \text{ is minimizer}) \)

\[
\leq \frac{1}{T} \left( \sum_{t=0}^{T-1} R_t(\Pi_t) - \min_{\Pi^*} \sum_{t=0}^{T-1} R_t(\Pi^*) \right)
\]

\( (\text{Less than worst-case distributions}) \)

\[
\leq \max_{D_0, D_1, \ldots, D_T} \frac{1}{T} \left( \sum_{t=0}^{T-1} R_t(\Pi_t) - \min_{\Pi^*} \sum_{t=0}^{T-1} R_t(\Pi^*) \right)
\]

\( (\text{FTL theorem}) \)

\[
\leq O(1/\sqrt{T}) = \varepsilon_{\text{FTL}} \quad \square
\]

Notice that this guarantee concerns the performance of \( \Pi_t \) on \( d^\Pi_t \), i.e. on the state distribution that it induces!

(Contrast with supervised ML where we only had guarantees with respect to \( d_{\Pi^*} \)).
Theorem: if \( \ell(\Pi^t, s; \Pi^*(s)) \geq \mathbb{E}[\ell(\Pi^t, s; \Pi^*(s))]^2 \),

There exists \( 0 \leq t \leq T-1 \) such that

\[
\mathbb{E}_{s \sim d_{\Pi^t}} \left[ V_{\Pi^t}^*(s) - V_{\Pi}^*(s) \right] \leq \max_{s, a} \left| A_{\Pi^t}^*(s, a) \right| \frac{1 - \gamma}{1 - \gamma} \epsilon_{\text{FTL}}
\]

Proof: we apply PDL in the other direction

\[
\mathbb{E}_{s \sim d_{\Pi^t}} \left[ V_{\Pi^t}^*(s) - V_{\Pi}^*(s) \right] = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\Pi^t}} \left[ A_{\Pi^t}^* (s, \Pi^t(a)) \right] \quad \text{(PDL)}
\]

\[
(a_{\Pi^t(s; \Pi^t(s)))} = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\Pi^t}} \left[ A_{\Pi^t}^* (s, \Pi^t(a)) - A_{\Pi^t}^* (s, \Pi^*(a)) \right]
\]

\[
\geq \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\Pi^t}} \left[ \max_{s, a} \left| A_{\Pi^t}^* (s, a) \right| \mathbb{E}_{s \sim d_{\Pi^t}} \left[ \mathbb{I}[\ell(\Pi^t, s; \Pi^*(s)) \geq \mathbb{E}[\ell(\Pi^t, s; \Pi^*(s))]^2] \right] \right]
\]

\[
\mathbb{E}_{s \sim d_{\Pi^t}} \left[ V_{\Pi^t}^*(s) - V_{\Pi}^*(s) \right] \leq \frac{1}{1 - \gamma} \max_{s, a} \left| A_{\Pi^t}^* (s, a) \right| \mathbb{E}_{s \sim d_{\Pi^t}} \left[ \mathbb{I}[\ell(\Pi^t, s; \Pi^*(s)) \geq \mathbb{E}[\ell(\Pi^t, s; \Pi^*(s))]^2] \right] \epsilon_{\text{FTL}}
\]
How to interpret $\max_{s,a} |A^{\pi^*}(s,a)|$?

Small if expert $\pi^*$ can quickly recover from mistake. I.e., if we take action $\alpha$ at state $s$ instead of $\pi^*(s)$, it doesn't impact future rewards too much as long as we follow $\pi^*(s)$ going forward. ($Q^{\pi^*}(s,a)$ is not too much smaller than $V^{\pi^*}(s)$)