CS 4/5789 20 April 2022 Prof Sarah Dean

Lecture 23: Interactive Imitation Learning

1) Dataset Aggregation with DAgger

Setting: Discounted Infinite Horizon MDP

$\mathcal{M} = \{ \mathcal{S}, \mathcal{A}, P, r, \gamma \}$ unknown. possibly unobserved

Expert knows optimal policy π^* and we can query the expert at any state during training.

Algorithm: DAgger

Initialize π^0 and dataset $\mathcal{D} = \phi$

For $t = 0, \dots, T-1$:

1) Generate Dataset with π^t & Query Expert

$\mathcal{D}^t = \{ s_i, a_i^* \}$ where $s_i \sim d_\mu^{\pi^t}$ and $a_i^* = \pi^*(s_i)$

2) Data Aggregation: $\mathcal{D} = \mathcal{D} \cup \mathcal{D}^t$

3) Update policy via SL:

$$\pi^{t+1} = \underset{\pi \in \Pi}{\operatorname{argmin}} \sum_{s,a \in \mathcal{D}} \ell(\pi, s, a)$$

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 θ_t (i.e. from past data)
2) Suffer the risk $R_t(\theta_t) = \mathbb{E}_{z \sim D_t}\left[\ell(\theta_t, z)\right]$
   (expected loss)

We care about average regret

$$\frac{1}{T} R(T) = \frac{1}{T}\left[\sum_{i=0}^{T-1} R_t(\theta_t) - \min_{\theta} \sum_{i=0}^{T-1} R_t(\theta)\right]$$

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

<u>ex</u> - 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?
$\mathcal{D}_t$ (and thus R_t) can vary in other ways

Example: in DAgger, we choose π^t and then suffer $\ell(\pi^t, s_i, \pi^*(s_i))$ for $s_i \sim d_\mu^{\pi^t}$

In this case $\mathcal{D}_t$ actually depends on θ_t

How should the learner choose θ_t?

Algorithm: Follow-the-Regularized Leader

For $t = 0, 1, \dots, T-1$

regularizer

$$\theta_t = \min_\theta \underbrace{\sum_{k=0}^{t-1} R_k(\theta)}_{} + \lambda f(\theta)$$

data aggregation!

$$= \sum_{k=0}^{t-1} \mathbb{E}_{z \sim D_k}[\ell(\theta, z)] = \mathbb{E}_{z_k \sim D_k}\left[\sum_{k=0}^{t-1} \ell(\theta, z_k)\right]$$

Theorem (FTL): If losses are convex and regularizer is strongly convex, then even if risks R_t (ie. distributions D_t) are chosen adversarially,

$$\max_{R_0, \dots, R_{T-1}} \frac{1}{T}\left[\sum_{t=0}^{T-1} R_t(\theta_t) - \min_\theta \sum_{t=0}^{T-1} R_t(\theta)\right] = O\left(1/\sqrt{T}\right)$$

3) Analysis of DAgger

We can veiw DAgger as an instance of FTL

<u>Corollary</u>: if $\ell(\pi^*, s, \pi^*(s))=0$, then

$$\min_{0\le t\le T-1} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\ell(\pi^t, s, \pi^*(s))\right] \le O(1/\sqrt{T}) = \varepsilon_{FTL}$$

<u>Proof</u>: π_t plays the roll of θ_t, & $(s, \pi^*(s))$ is z & D_t is $d_\mu^{\pi^t}$

$$\min_{0\le t\le T-1} R_t(\pi_t) \le \frac{1}{T}\sum_{t=0}^{T-1} R_t(\pi_t) \quad \text{(min} \le \text{avg)}$$

(π^* has 0 loss)
$$= \frac{1}{T}\sum_{t=0}^{T-1} R_t(\pi_t) - R_t(\pi^*)$$

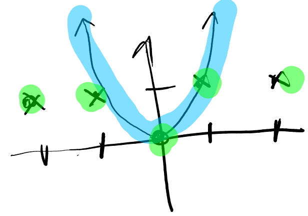
(π^* is minimizer)
$$\le \frac{1}{T}\left(\sum_{t=0}^{T-1} R_t(\pi_t) - \min_{\pi}\sum_{t=0}^{T-1} R_t(\pi^*)\right)$$

(Less than worst-case distributions)
$$\le \max_{D_0, D_{T-1}} \frac{1}{T}\left(\sum_{t=0}^{T-1} R_t(\pi_t) - \min_{\pi}\sum_{t=0}^{T-1} R_t(\pi^*)\right)$$

(FTL theorem)
$$\le O(1/\sqrt{T}) = \varepsilon_{FTL} \quad \square$$

Notice that this guarantee concerns the performance of π_t on $d_\mu^{\pi^t}$, i.e. on the state distribution that it induces!

(Contrast with supervised ML where we only had guarantees with respect to $d_\mu^{\pi^*}$!)



Theorem: if $\ell(\pi^t, s, \pi^*(s)) \geq \mathbb{1}\{\pi^t(s) \neq \pi^*(s)\}$,

There exists $0 \leq t \leq T-1$ such that

$$\mathbb{E}_{s\sim\mu_0}\left[V^{\pi^*}(s) - V^{\pi^t}(s)\right] \leq \frac{\max_{s,a} |A^{\pi^*}(s,a)|}{1-\gamma} \cdot \varepsilon_{FTL}$$

Proof: We apply PDL in the other direction

$$\mathbb{E}_{s\sim\mu_0}\left[V^{\pi^t}(s) - V^{\pi^*}(s)\right] = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^*}(s, \pi^t(a))\right] \quad \text{(PDL)}$$

$$(A^{\pi}(s,\pi(s))=0) \qquad = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^*}(s,\pi^t(a)) - A^{\pi^*}(s,\pi^*(a))\right]$$

$$\geq \frac{-1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\max_{s,a} |A^{\pi^*}(s,a)| \, \mathbb{1}\{\pi^t(s) \neq \pi^*(s)\}\right]$$

$$\mathbb{E}_{s\sim\mu_0}\left[V^{\pi^*}(s) - V^{\pi^t}(s)\right] \leq \frac{1}{1-\gamma} \max_{s,a} |A^{\pi^*}(s,a)| \, \mathbb{E}_{s\sim d_\mu^{\pi^t}} \mathbb{1}\{\pi^t(s) \neq \pi^*(s)\}$$

(assumption on loss)

$$\leq \frac{1}{1-\gamma} \max_{s,a} |A^{\pi^*}(s,a)| \underbrace{R_t(\pi^t)}_{\varepsilon_{FTL}}$$

How to interperet $\max_{s,a} |A^{\pi^*}(s,a)|$?

Small if expert π^* can quickly recover from mistake. I.e. if we take action a at state s instead of $\pi^*(s)$, it doesn't impact futeure rewards too much as long as we follow $\pi^*(s)$ going forward.

($Q^{\pi^*}(s,a)$ is not too much smaller than $V^{\pi^*}(s)$)