1) Dataset Aggregation w/ DAgger

Setting: Discounted Infinite Horizon MDP

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

Expert knows the optimal policy π^*

and we 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 & expert

$$\mathcal{D}^t = \{ s_i, a_i^* \}_{i=1}^N \quad s_i \sim d_\mu^{\pi^t} \quad a_i^* = \pi^*(s_i)$$

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

3) Update Policy via SL:

small if $\pi(s) \approx a$

$$\pi^{t+1} = \underset{\pi \in \Pi}{\operatorname{argmin}} \; \mathbb{E}_{s,a \sim \mathcal{D}} \left[\ell(\pi, s, a) \right]$$

2) Online Learning

captures idea of learning from additional data over time

Iterative w/ 2 components

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

1) Learner chooses θ_t

2) Suffer the risk $R_t(\theta_t) = \mathbb{E}_{z \sim \mathcal{D}_t} [\ell(\theta_t, z)]$ (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 is the best parameter in hindsight

Difference from SL setting:

$\mathcal{D}_t$ ($\&$ R_t) can vary in many ways

Example: in DAgger, we choose π^t

and suffer $\mathbb{E}_{s \sim d^{\pi^t}_{M}}\left[\ell(\pi^t, s, \pi^*(s))\right]$

How should Learner choose θ_t?

Algorithm: Follow the Regularized Leader

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

$$\theta_t = \min_{\theta} \underbrace{\sum_{k=0}^{t-1} R_k(\theta)}_{\sum_{k=0}^{t-1} \mathbb{E}_{z\sim\mathcal{D}_k}[\ell(\theta, z)]} + \lambda f(\theta) \quad \text{regularizer}$$

$$\sum_{k=0}^{t-1} \mathbb{E}_{z\sim\mathcal{D}_k}[\ell(\theta, z)] = \mathbb{E}_{z\sim\frac{1}{t}\sum_{k=0}^{T-1}\mathcal{D}_k}[\ell(\theta, z)]$$

data aggregation

Theorem (FTL): if loss functions are convex and regularizer is strongly convex, then

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

3) Analysis of DAgger

e.g. $\|\pi(s)-a\|_2^2$ and $\ell \geq 0$

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

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

Proof: π^t plays role of θ_t, $(s, \pi^*(s))$ is z,

$d_\mu^{\pi^t}$ is D_t

$$\min_{0 \le t \le T-1} R_t(\pi^t) \le \frac{1}{T} \sum_{t=0}^{T-1} R_t(\pi^t)$$

$$= \frac{1}{T} \sum_{t=0}^{T-1} R_t(\pi^t) - \overbrace{R_t(\pi^*)}^{=0}$$

defined by $d_\mu^{\pi^t}$ →

$$= \frac{1}{T} \left(\sum_{t=0}^{T-1} R_t(\pi^t) - \min_{\pi} \sum_{t=0}^{T-1} R_t(\pi) \right)$$

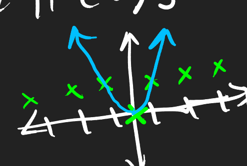
defined by arbitrary dist →

$$\le \max_{R_0, \dots, R_{T-1}} \frac{1}{T} \left(\sum_{t=0}^{T-1} R_t(\pi^t) - \min_{\pi} \sum_{t=0}^{T-1} R_t(\pi) \right)$$

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

Key fact: accuracy guarantee on $d_\mu^{\pi^t}$ instead of $d_\mu^{\pi^*}$!

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



Then there is $t \in \{0, \dots, T-1\}$ such that

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

$A^{\pi^*}(s,a) \leq 0 \quad \forall s,a \quad \left[\mathbb{E}\, Q^{\pi^*}(s,a) - V^{\pi^*}(s) \right]$

$V^{\pi^*}(s)$

$\max_{s,a} |A^{\pi^*}(s,a)|$ is the cost of messing up at one timestep.

<u>Proof</u> We appy PDL in other direction

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

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

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

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

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