1) Dataset Aggregation w/ DAgger

Setting: Discounted Infinite Horizon MDP
\( \mathcal{M} = \mathbb{S}, \mathbb{A}, \mathbb{P}, \gamma, \mathbb{R} \) possibly unobserved

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

Algorithm: DAgger
Initialize \( \pi^0 \) and dataset \( D = \emptyset \)
For \( t = 0, \ldots, T-1 \)

1) Generate Dataset with \( \pi^t \) & expert
\( D^t = \{ S_i, a_i^* S^\pi_{i+1}, r_i \sim d^\pi_{i+1}, a_i^* = \pi^*(S_i) \} \)

2) Data Aggregation: \( D = D \cup D^t \)

3) Update Policy via SSL:
\( \pi^{t+1} = \arg \min_{\pi \in \mathcal{P}} \mathbb{E} [\mathbb{E} [l(\pi, S, a) | S, a \sim D] \] small if \( \pi(S) \approx a \)

2) Online Learning
Captures idea of learning from additional data over time iterative w/ 2 components
For \( t = 0, 1, \ldots, T-1 \)

1) Learner chooses \( \theta_t \)

2) Suffer the risk \( R_t(\theta_t) = \mathbb{E} [l(\theta_t, z) | z \sim D_t] \) (expected loss)
\[ \frac{1}{T} R(T) = \frac{1}{T} \left[ \sum_{t=0}^{T-1} R_t(\Theta_t) - \min_{\Theta} \sum_{t=0}^{T-1} R_t(\Theta) \right] \]

The baseline is the best parameter in hindsight.

Difference from SL setting:
\[ \Theta_t \text{ (and } R_t \text{) can vary in many ways} \]

Example: in DAgger, we choose \[ \pi^t \]
and suffer \[ \mathbb{E} \left[ l(\pi^t, s, \pi^* s) \right] \]

How should learner choose \[ \Theta_t \]?

Algorithm: Follow the Regularized Leader

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

\[ \Theta_t = \min_{\Theta} \sum_{k=0}^{t-1} \mathbb{E} \left[ l(\Theta_k, z) \right] + \lambda_f(\Theta) \]

where

Data aggregation

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

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

corollary: if \( \ell(T^*, s, \Pi^*(s)) = 0 \), then
\[
\min_{0 \leq t \leq T-1} \mathbb{E} \left[ \ell(T^*_t, s, \Pi_t^*(s)) \right] \leq 0 \text{ (\text{FTL})} = \varepsilon_{\text{FTL}}
\]

Proof: \( \Pi^*_t \) plays role of \( Q_t \), \((s, \Pi^*(s)) \) is \( \varepsilon \), \( d^*_m \) is \( D_t \)

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

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

defined by \( d^t_{\Pi^*} \)

\[
\leq \max \left( R_0, \ldots, R_{T-1} \right) \frac{1}{T} \left( \sum_{t=0}^{T-1} R_t(\Pi_t^*) - \min_{\Pi} \sum_{t=0}^{T-1} R_t(\Pi) \right)
\]

defined by arbitrary dist

\[
\leq O(\sqrt{\varepsilon}) = \varepsilon_{\text{FTL}}
\]

Key fact: accuracy guarantee on \( d^t_{\Pi^*} \) instead

Theorem: if \( \ell(T, s, \Pi^*(s)) \geq 1 \sum_{t} \Pi (s) \neq \Pi^*(s) \)

Then there is \( t \in [0, T-1] \) such that
\[
\mathbb{E} \left[ V_{\Pi^*} (s) - V_{\Pi_t^*} (s) \right] \leq (\max_{s,a} |A^{\Pi^*}_{t}(s,a)|) (1 - \frac{1}{\Delta}) \cdot \varepsilon_{\text{FTL}}
\]
$A^{\pi^*_y}(s, a) \leq 0 \; \forall s, a \; \left[ -Q^{\pi^*_y}(s, a) - V^{\pi^*_y}(s) \right]

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

**Proof** We apply PDL in the other direction.

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

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

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

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

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