Behavior Cloning

Setting: Discounted Infinite Horizon MDP

$$\mathcal{M} = \{ \mathcal{S}, \mathcal{A}, P, r, \gamma \}$$

unknown
unobserved

Expert knows optimal policy π^* and dataset

$$\mathcal{D} = \{ (s_i^*, a_i^*) \}_{i=1}^{M} \sim d_{\mu_0}^{\pi^*}$$

Behavoir Cloning

Reduction to Supervised learning.

Define a policy class, $\pi \in \Pi$

e.g. parametric policies $\Pi = \{ \pi_\theta : \theta \in \mathbb{R}^d \}$

Then we estimate a policy with empirical risk minimization

$$\hat{\pi} = \underset{\pi \in \Pi}{\operatorname{argmin}} \sum_{i=1}^{M} \ell(\pi, s_i^*, a_i^*)$$

Many choices of loss function

- discrete action space: classification
- continuous action space: regression

ex: Negative log likelihood

$$\ell(\pi, s, a) = -\log(\pi(a|s))$$

ex: square loss

$$\ell(\pi, s, a) = \| \pi(s) - a \|_2^2$$

3) BC Analysis

Assumption: Supervised ML is successful

$$\mathbb{E}_{s\sim d^{\pi^*}_{\mu_0}}\left[\mathbb{1}\{\hat{\pi}(s)\neq \pi^*(s)\}\right] \le \varepsilon$$

depend on # datapoints, complexity of Π, whether $\pi^*\in\Pi$

Train (d^{π^*}) and Test ($d^{\hat{\pi}}$) distribution mismatch.

$(1-\gamma)\left[0+\sum_{t=1}^{\infty}\gamma^t 1\right] = (1-\gamma)\frac{\gamma}{1-\gamma} = \gamma$

Recall: $d^{\pi}_{\mu_0}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^{\pi}_t(s;\mu_0)$

Recall: Performance Difference Lemma

$$\underbrace{\mathbb{E}_{s\sim\mu_0}\left[V^{\pi}(s)-V^{\pi'}(s)\right]}_{\text{difference in value (cumulative reward)}} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{\mu_0}}\Big[\underbrace{\mathbb{E}_{a\sim\pi(\cdot|s)}\left[A^{\pi'}(s,a)\right]}_{\text{advantage of }\pi\text{ over }\pi'}\Big]$$

Theorem (BC Performance): Assume $r(s,a)\in[0,1]$ and supervised ML succeeds w/ ε. Then BC returns $\hat{\pi}$

$$\mathbb{E}_{s\sim\mu_0}\left[V^{\pi^*}(s)-V^{\hat{\pi}}(s)\right]\le \frac{2\varepsilon}{(1-\gamma)^2}$$

$\left(\sum_{t=0}^{\infty}\gamma^t r_t \le \frac{1}{1-\gamma}\right)$

Proof

deterministic

$$\mathbb{E}_{s\sim\mu_0}\left[V^{\pi^*}(s)-V^{\hat{\pi}}(s)\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^*}_{\mu}}\left[A^{\hat{\pi}}(s,\pi^*(s))\right]$$

add 0

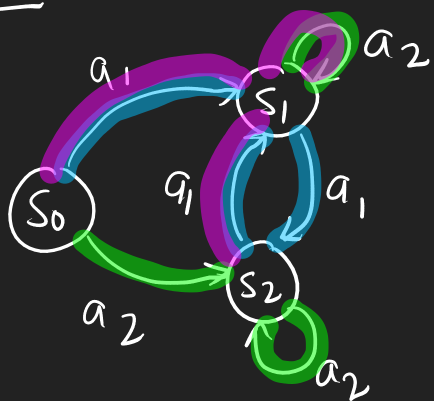
$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^*}_{\mu}}\left[A^{\hat{\pi}}(s,\pi^*(s)) - A^{\hat{\pi}}(s,\hat{\pi}(s))\right]$$

$A \leq \frac{1}{1-\gamma}$

$$\leq \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_\mu^{\pi^*}} \left[\frac{2}{1-\gamma} \mathbb{1}\{\pi^*(s) \neq \hat{\pi}(s)\} \right]}$$

$$\leq \frac{2}{(1-\gamma)^2} \cdot \varepsilon$$

Example: Distribution Shift



s_0 is initial state

$$r(s,a) = \begin{cases} 1 & s = s_1 \\ 0 & \text{otherwise} \end{cases}$$

optimal policy

$$\boxed{d_{s_0}^{\pi^*}(s) = \begin{cases} 1-\gamma & s = s_0 \\ \gamma & s = s_1 \\ 0 & s = s_2 \end{cases}}$$

consider policy

$$\hat{\pi}(s_0) = \begin{cases} a_1 & \text{w.p. } 1 - \frac{\varepsilon}{1-\gamma} \\ a_2 & \text{w.p. } \frac{\varepsilon}{1-\gamma} \end{cases} \qquad \hat{\pi}(s_1) = a_2 \qquad \boxed{\hat{\pi}(s_2) = a_2}$$

Has low SL error

$$\mathbb{E}_{s \sim d_{s_0}^{\pi^*}} \left[\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)} [a \neq \pi^*(s)] \right] = \varepsilon$$

Has quadratic performance error

$$V^{\pi^*}(s_0) = \sum_{t=1}^{\infty} \gamma^t = \frac{\gamma}{1-\gamma} \qquad V^{\hat{\pi}}(s_0) = \left(1 - \frac{\varepsilon}{1-\gamma}\right) \frac{\gamma}{1-\gamma} + \left(\frac{\varepsilon}{1-\gamma}\right) \cdot 0$$

$$V^{\pi^*}(s_0) - V^{\hat{\pi}}(s_0) = \frac{\varepsilon\gamma}{(1-\gamma)^2}$$

Caused by bad policy in $s_2 \rightarrow d_{s_0}^{\pi^*}(s_2) = 0$