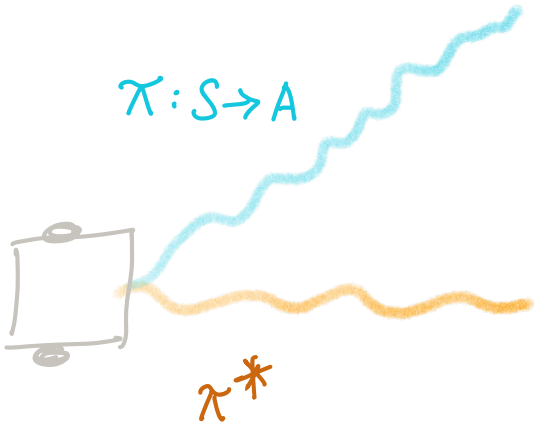


How much worse is policy π over expert π^*?

$$J(\pi) - J(\pi^*)$$

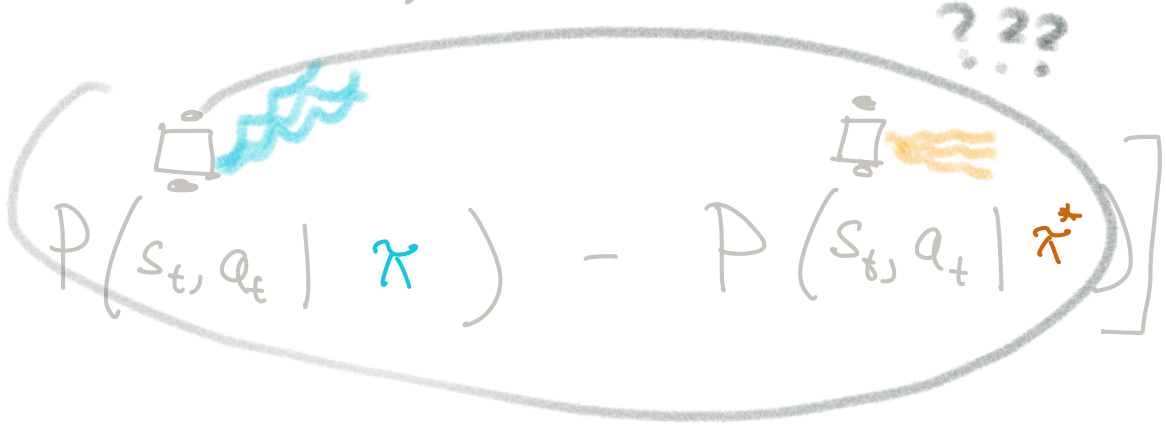
$J(\pi)$: Performance (Total cost)

WHY IS THIS HARD?

$$J(\pi) - J(\pi^*)$$

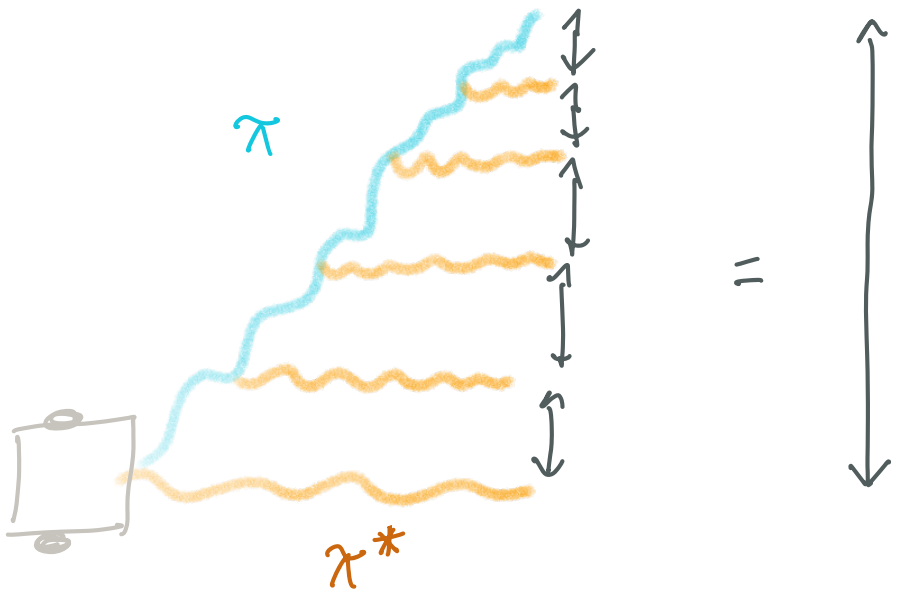
$$= \sum_{t=0}^{T-1} \mathop{E}_{\substack{s_t \sim d_t^{\pi} \\ a_t \sim \pi}} \left[c(s_t, a_t) \right] - \sum_{t=0}^{T-1} \mathop{E}_{\substack{s_t \sim d_t^{\pi^*} \\ a_t \sim \pi^*(s_t)}} \left[c(s_t, a_t) \right]$$

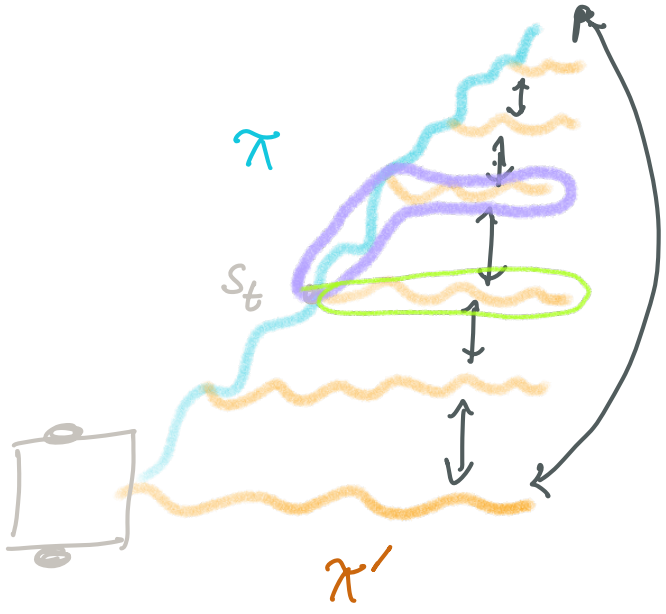
$$= \sum_{t=0}^{T-1} \sum_{s_t, a_t} c(s_t, a_t) \left[P(s_t, a_t \mid \pi) - P(s_t, a_t \mid \pi^*) \right]$$



CHALLENGE: DIFFERENCE IN DISTRIBUTIONS HARD TO ANALYZE

$J(\pi) - J(\pi^*)$





Performance Difference Lemma

$$J(\pi) - J(\pi^*)$$

$$= \sum_{t=0}^{T-1} \mathop{E}_{s_t \sim d_t^{\pi}} \left[Q^{\pi^*}(s_t, \pi) - Q^{\pi^*}(s_t, \pi^*) \right]$$

Rollout Learner till timestep t

Difference in Q^*-value of Learner − Expert

"Advantage"

$$A^{\pi^*}(s_t, \pi)$$

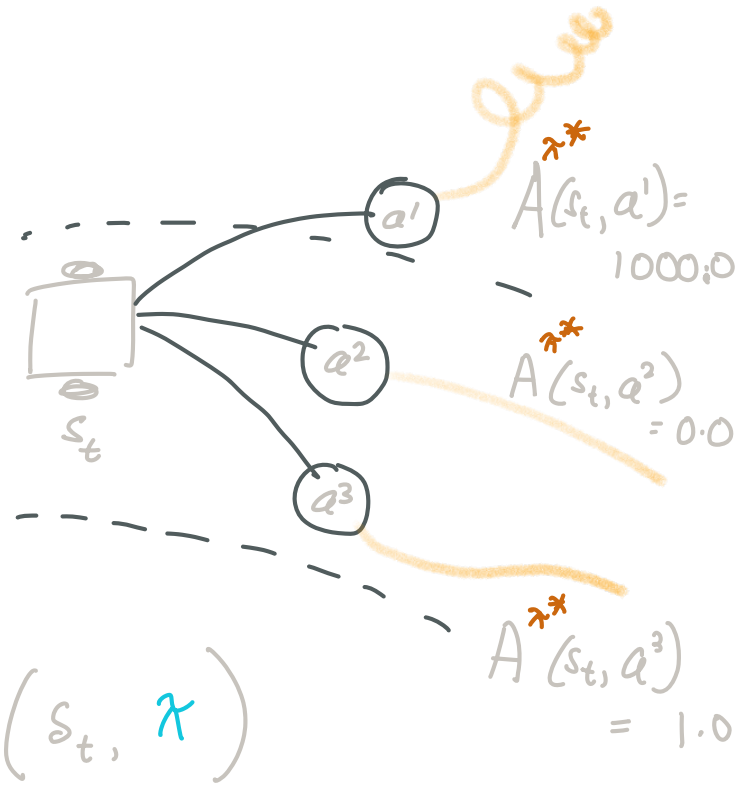
INTUITION

$$\min_{\pi} \sum_{t=0}^{T-1} E_{s_t \sim d_t^{\pi}} A(s_t, \pi)$$

π ⟷ DISTRIBUTION

$A(s_t, \pi)$ ⟷ LOSS

STATES VISITED BY π

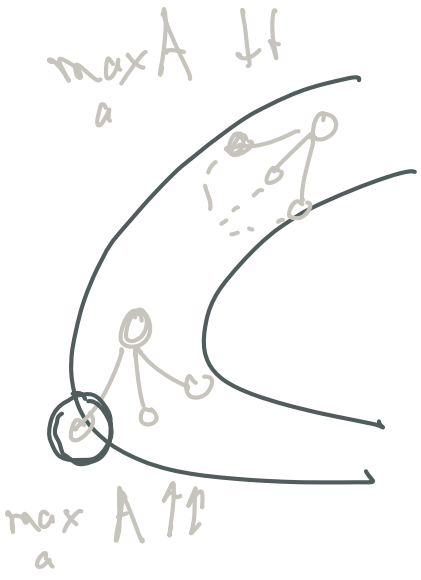


$$A^{\pi^*}(s_t, \pi) = \begin{bmatrix} \pi(a^1) \\ \pi(a^2) \\ \pi(a^3) \end{bmatrix} \cdot \begin{bmatrix} A^{\pi^*}(a^1) \\ A^{\pi^*}(a^2) \\ A^{\pi^*}(a^3) \end{bmatrix} \begin{matrix} 1000.0 \\ 0 \\ 1.0 \end{matrix}$$

COST SENSITIVE CLASSIFICATION

$$\min_{\pi} \sum_{t=0}^{T-1} E_{s_t \sim d_t^{\pi^*}} A(s_t, \pi)$$

$$J(\pi) - J(\pi^*) = \sum_{t=0}^{} \mathop{E}_{s_t \sim d_\pi^t} A(s_t, \pi)$$



$$\leq \sum_{t=0}^{T-1} \mathop{E}_{s_t \sim d_\pi^t} \left[\max_a A^{\pi^*}(s_t, a) \right] \cdot \mathbb{1}\left(\pi(s_t) \neq \pi^*(s_t)\right)$$

$$\leq \left[\max_{s,a} A^{\pi^*}(s, a) \right] \sum_{t=0}^{T} \mathop{E}_{s_t \sim d_a^t} \mathbb{1}\left(a(s_t) \neq \pi^*(s_t)\right)$$

M

$$\leq O(\varepsilon T \cdot M)$$

Applications

① Fundamentals: Policy Iteration

$$\pi^+ \equiv \arg\max_{\pi} Q^{\pi^-}(s, \pi) \quad \forall s \Rightarrow$$ Does this monotonically improve?

$$J(\pi^+) - J(\pi^-) = \sum_{t=0}^{T-1} \underset{s_t, d_t^{\pi^+}}{E} A^{\pi^-}(s_t, \pi^+) \geq 0 \rightarrow \text{Yes!}$$

② Algorithms in Imitation Learning / Reinforcement Learning

$$\max_{\pi} J(\pi) - J(\pi^*) \equiv \max_{\pi} \sum_{t=0}^{T-1} \underset{s_t, d_t^{\pi}}{E} A^{\pi^*}(s, \pi)$$

Rollin learner

Aggrevate (Ross & Bagnell '14)

(3) Other Lemmas: Simulation Lemma

$$J_M(\pi) - J_{M'}(\pi)$$

"Performance in model M"

"Perform in model M'"

Proof

$$J(\pi) - J(\pi') = \mathop{E}_{s_0} Q^{\pi}(s_0, \pi) - \mathop{E}_{s_0} Q^{\pi'}(s_0, \pi')$$

$$= \mathop{E}_{s_0} Q^{\pi}(s_0, \pi) - \mathop{E}_{s_0} Q^{\pi'}(s_0, \pi) + \mathop{E}_{s_0} Q^{\pi'}(s_0, \pi) - \mathop{E}_{s_0} Q^{\pi'}(s_0, \pi')$$

$$= \cancel{\mathop{E}_{s_0, a_0 \sim \pi} r(s_0, a_0)} - \cancel{\mathop{E}_{s_0, a_0 \sim \pi} r(s_0, a_0)} + \mathop{E}_{s_0} A^{\pi'}(s_0, \pi)$$

$$+ \underbrace{\mathop{E}_{s_1 \sim d_1^{\pi}} Q^{\pi}(s_1, \pi) - \mathop{E}_{s_1 \sim d_1^{\pi}} Q^{\pi'}(s_1, \pi')}_{\text{Recursively expand this}}$$

$$= \sum_{t=0}^{T-1} \mathop{E}_{s_t \sim d_t^{\pi}} A^{\pi'}(s_t, \pi)$$