What would DAgger Return?

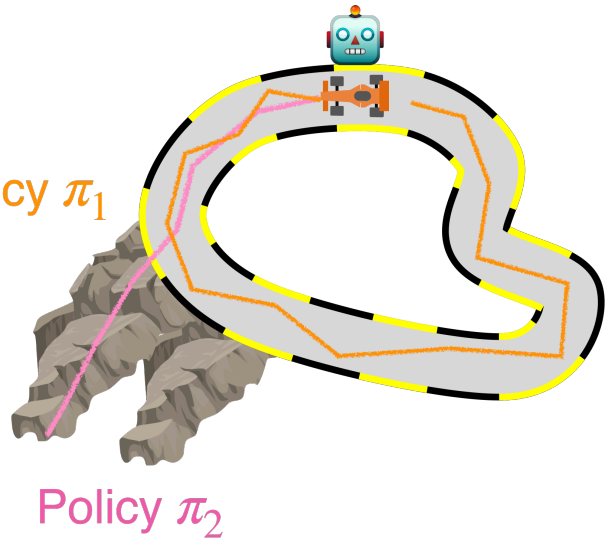
Round 1 — Round 2 — ...

π_1 ..... π_2 — π_2 — π_2

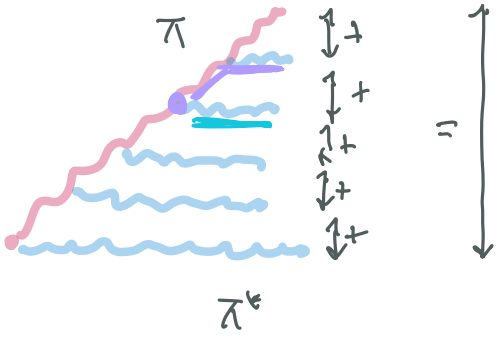
$\ell_1(\pi_1)$ ..... $\ell_2(\pi_2)$ — $\ell_2(\pi_2)$ — $\ell_2(\pi_2)$

$\ell_i(\pi)$

$= \mathop{\mathbb{E}}_{s \sim d_{\pi_i}} 1(\pi(s) \neq a^*)$



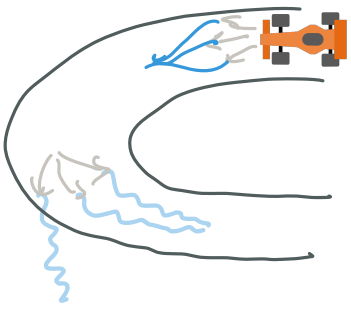
Performance Difference Lemma



$$J(\pi) - J(\pi^*) = \sum_{t=1}^{T} \underset{s_t \sim d_\pi^t}{E} \left[Q^{\pi^*}(s_t, \pi(s_t)) - Q^{\pi^*}(s_t, \pi^*(s_t)) \right]$$

$$= \sum_{t=1}^{T} \underset{s_t \sim d_\pi^t}{E} \left[A^{\pi^*}(s_t, \pi(s_t)) \right]$$

On-policy

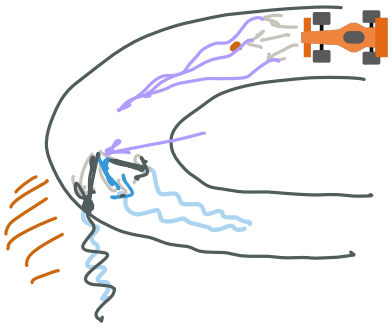


NOT ALL ERRORS ARE EQUAL

PDL

DAGGER

$A^{\pi^*}(s, \pi(s))$ $\quad O(\varepsilon T)$ $\quad \mathbb{I}(a^* \neq \pi(s))$



$$\begin{bmatrix} 10000.0 \\ 0 \\ \varepsilon \end{bmatrix} \qquad \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

$$A^{\pi^*}(s, a_1) = \underbrace{Q^{\pi^*}(s, a_1)}_{10000.0} - \underbrace{Q^{\pi^*}(s, a_2)}_{0.0}$$

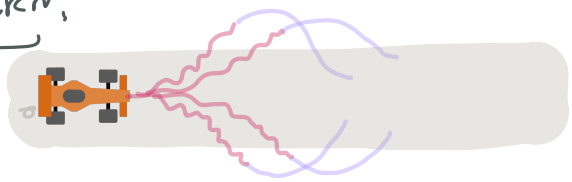
$$A^{\pi^*}(s, \pi(s)) \leq \underbrace{\mathbb{I}(a^* \neq \pi(s))}_{0-1} \cdot \underbrace{\max_a A^{\pi^*}(s, a)}_{\mu}$$

"RECOVERABILITY"

$$O(\varepsilon T) \rightarrow O(\varepsilon T \cdot \mu)$$

$$\approx O(\varepsilon T^2)$$

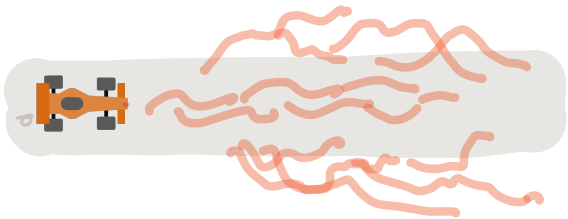
WHAT POLICY WOULD HG-DAGGER RETURN?



POLICY 1:

GOOD: STAYIN IN THE TRACK

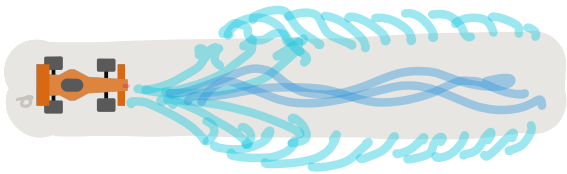
OK: RECOVERY



POLICY 2:

BAD: STAYING IN TRACK

AWESOME: RECOVERY



BEST IN HINDSIGHT

$N \to \infty$