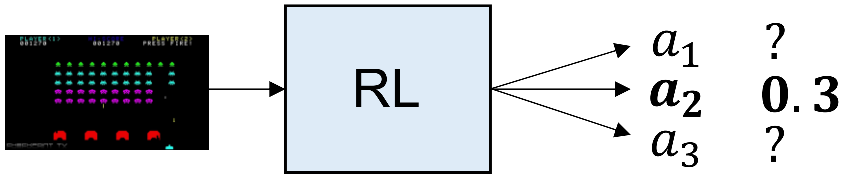
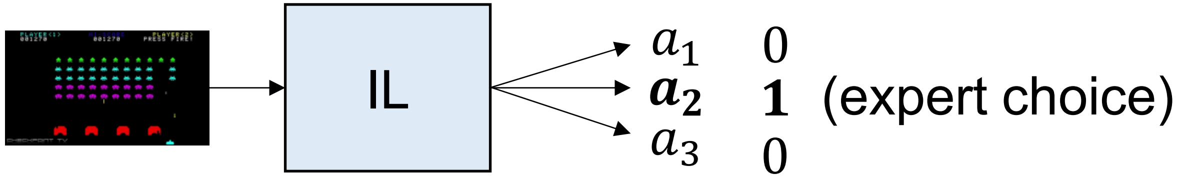
Imitation Learning

Chen-Yu Wei

Imitation Learning ∈ Supervised Learning





π^* : expert's policy

$\pi^*(\cdot|s_1)$ $P(\cdot|s_1, a_1^*)$

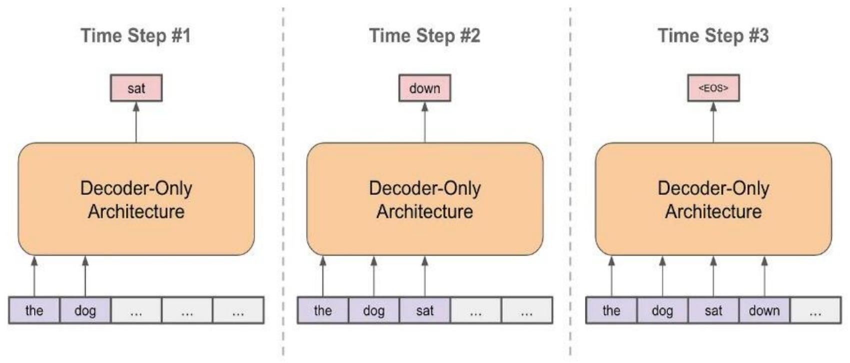
Offline IL: learn from static data generated by the expert $\{(s_1, a_1^\star, s_2, a_2^\star, \ldots, s_H, a_H^\star)\}$

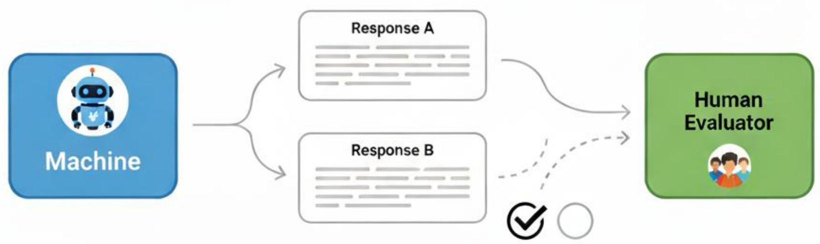
Online IL: may interact with MDP and query the expert $\{(s_1, a_1, a_1^\star, s_2, a_2, a_2^\star, \ldots, s_H, a_H, a_H^\star)\}$

Goal: output a policy $\hat{\pi}$ such that $V^{\pi^\star}(\rho) - V^{\hat{\pi}}(\rho)$ is small

Examples

- Language models

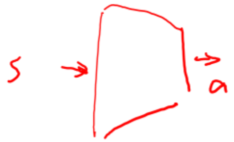




RLHF

- Robotics

Types



$\pi(a|s) \leftrightarrow \pi^*(a|s)$

- Direct Imitation: directly learn policy to imitate the expert
  - Behavior cloning
  - DAgger
  - Direct preference optimization (preference feedback)
- Occupancy matching
  - DICE
  - GAIL

$d^{\pi}(s,a) \leftrightarrow d^{\pi^*}(s,a)$

$(s_1, a_1^*, s_2, a_2^*, \ldots)$

- Inverse RL: learn an MDP (or just reward function) from expert, and perform RL on it
  - Adversarial IRL (paper)
  - MaxEnt IRL (paper)
  - RLHF (preference feedback)

$\boxed{\text{resp}} \to \text{reward}$

Behavior Cloning: Reduction to Classification

$(s_1, a_1^\star, s_2, a_2^\star, \ldots)$ are from expert

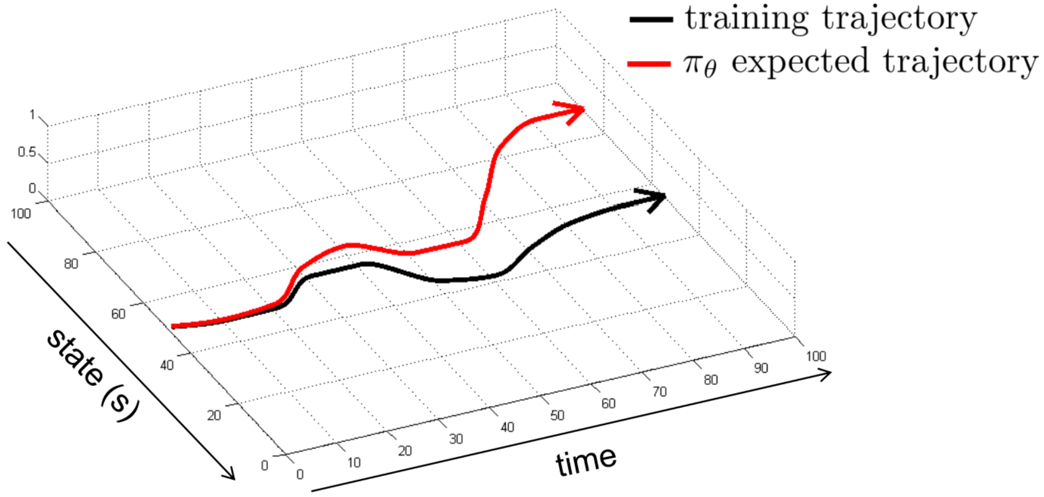
Relate $V^{\pi^\star}(\rho) - V^{\hat{\pi}}(\rho)$ to $\mathbb{E}^{\pi^\star}\left[\frac{1}{H}\sum_{h=1}^{H} \mathbb{I}\{\hat{\pi}(s_h) \neq a_h^\star\}\right]$

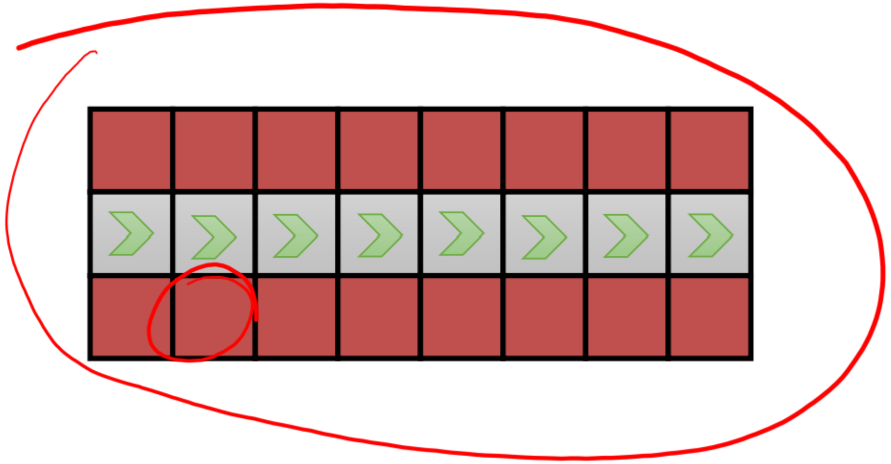
$$V^{\pi^*}(\rho) - V^{\hat{\pi}}(\rho) = \sum_h \sum_{s,a} d^{\pi^*}(s)\left(\pi^*(a|s) - \pi(a|s)\right) \overbrace{Q^{\pi}(s,a)}^{\leq H}$$

$$\leq 2H \sum_h \sum_s d^{\pi^*}(s)\, \mathbb{I}\{\pi^*(s) \neq \pi(s)\}$$

$$\leq O(H^2) \cdot \underbrace{\frac{1}{H}\sum_h d^{\pi^*}(s)\,\mathbb{I}(\pi^*(s) \neq \pi(s))}_{\text{indicator loss}}$$

Behavior Cloning: Reduction to Classification





$$\text{total loss} = \sum_h \mathbb{1}(\text{off-track at step } h)$$

$$\begin{aligned}
\mathbb{E} &= \varepsilon \times \mathbb{1}\{\text{go off-track at 1st step}\} \times H && \approx \varepsilon H \\
&+ (1-\varepsilon)\varepsilon\, \mathbb{1}\{ \sim\sim\sim \text{ 2nd step}\} \times (H-1) && \approx \varepsilon H \\
&+ (1-\varepsilon)^2 \varepsilon \quad \text{———} \quad \times (H-2) \\
&\ \ \vdots
\end{aligned} \quad \varepsilon H^2$$

Issue: **distribution shift**

$$\frac{1}{H} \sum_h \mathbb{1}(\pi_h(s_h) \neq \pi(s_h)) \leq \varepsilon$$

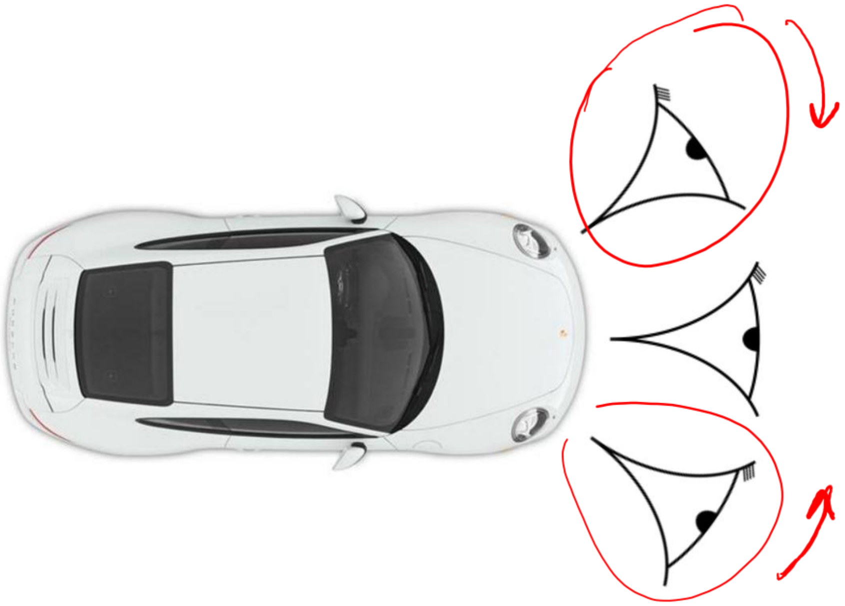
Behavior Cloning: Reduction to Classification

The bound might be pessimistic

- Single mistake may not lead to catastrophic failure

Solution

- Data augmentation



Bojarsky et al. End to End Learning for Self-Driving Cars. 2016

Solution: Interact with Expert (Online IL)

DAgger: **D**ataset **Agg**regation

goal: collect training data from $p_{\pi_\theta}(\mathbf{o}_t)$ instead of $p_{\text{data}}(\mathbf{o}_t)$

how? just run $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$

but need labels $\mathbf{a}_t$!

1. train $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ from human data $\mathcal{D} = \{\mathbf{o}_1, \mathbf{a}_1, \ldots, \mathbf{o}_N, \mathbf{a}_N\}$
2. run $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ to get dataset $\mathcal{D}_\pi = \{\mathbf{o}_1, \ldots, \mathbf{o}_M\}$
3. Ask human to label $\mathcal{D}_\pi$ with actions $\mathbf{a}_t$
4. Aggregate: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$