Disagreement-Regularized Imitation Learning

Kianté Brantley,\textsuperscript{1} Wen Sun,\textsuperscript{2} Mikael Henaff \textsuperscript{2}

\textsuperscript{1} University of Maryland, \textsuperscript{2} Microsoft Research
Imitation Learning

Expert Demonstrator

- state
- actions    up, down, left, right

Training set: $D = \{(\text{state, actions})\}$ from expert $\pi^*$

Goal: learn agent $\pi_\theta(s) \rightarrow a$
Imitation Learning using Behavior Cloning

$$J_{BC}(\pi) = \mathbb{E}_{s \sim d_{\pi^*}}[\mathcal{L}(\pi_\theta(s), \pi^*(s))]$$

**Problem:**
- Assumptions underlying supervised learning no longer hold
- Compounding error problem
- Can we design an agent that can deal with the compounding error problem without needing more demonstrations?

[ALVINN: An Autonomous Land Vehicle in a Neural Network, Dean Pomerleau Neurips 1989]
[An Invitation to Imitation - Semantic Scholar, Bagnell]
Formalizing the compounding error problem

Given an expert policy: $\pi^*$

Consider a policy: $\hat{\pi}$

$\hat{\pi}(s_1) = \frac{T - 1}{T}$

$\hat{\pi}(s_0) = \frac{1}{T}$

Behavior Cloning Loss:

$J_{BC}(\pi) = \epsilon$

(loss is small)

Behavior Cloning Regret:

Regret($\hat{\pi}$) = $\mathcal{O}(\epsilon T^2)$

(quadratic regret)

[Efficient Reductions for Imitation Learning, Ross & Bagnell, AISTATS 2010]

[Lower bounds for reductions, Matti Kaariainen, Atomic Learning Workshop 2006]
**Our Approach**

**DRIL**

**Motivation:**
1. Mimic expert within the expert distribution
2. Stay within the expert distribution

\[ J_{DRIL}(\pi) = J_{BC}(\pi) + J_U(\pi) \]

Train ensemble of policies \( \Pi_E = \{\pi_1, \ldots, \pi_E\} \) on demonstration data \( D \)

**Uncertainty Cost:**
\[ C_U(s, a) = \text{Var}_{\pi \sim \Pi_E}(\pi(a|s)) \]

**DRIL** cost can be optimized using any RL algorithm
Our Approach

DRIL (Final Algorithm)

Input: Expert Demonstration data $D = \{(s_i, a_i)\}_{i=1}^{N}$

Train Policy Ensemble $\Pi_E = \{\pi_1, \ldots, \pi_E\}$ using demonstration data $D$

Train policy behavior cloning $\pi$ using demonstration data $D$

for $i = 1$ to ... do
  - Perform one gradient update to minimize $J_{BC}(\pi)$ using minibatch from $D$
  - Perform one step of policy gradient to minimize $\mathbf{E}_{s \sim d, a \sim \pi(\cdot|s)}[C_U(s, a)]$
end for
Our Approach

DRIL (Analysis)

Theorem (informal): $J_{DRIL}(\pi)$ has regret $\mathcal{O}(\epsilon \kappa T)$

Assumption 1: (Realizability) $\pi^* \in \Pi$

Assumption 2: (Optimization Oracle) $J(\hat{\pi}) \leq \text{argmin}_{\pi \in \Pi} J(\pi) + \epsilon$

Assumption 3: (Smoothness on true Q-Function) $Q^{\pi^*}(s, a) - Q^{\pi^*}(s, \pi^*) \leq u$
Revisiting the compounding error problem

Given an expert policy: $\pi^*$

$$d_{\pi^*}(s_0) = \frac{1}{T}$$

$$d_{\pi^*}(s_1) = \frac{T - 1}{T}$$

Behavior Cloning Regret:

$$\text{Regret}(\hat{\pi}) = \mathcal{O}(\epsilon T^2)$$

(linear regret)

DRIL Regret:

$$\text{Regret}(\hat{\pi}) = \mathcal{O}(\epsilon \kappa T)$$

$$\kappa = \frac{1}{\sqrt{|\text{ensemble}|}}$$

(linear regret)
Experiments: (Atari)
Experiments: (Continuous Control)
Summary:

- Compounding error problem has been a fundamental issue in imitation learning
- Provide a new algorithm which uses uncertainty as an additional learning signal
- Theoretical guarantees in some settings
- Simple and Robust