Disagreement-Regularized Imitation Learning

Kianté Brantley[1], Wen Sun[2], Mikael Henaff[2]

[1]*University of Maryland*, [2]*Microsoft Research*

Main ideas

Using disagreement among an ensemble of pre-trained polices to reduce the *compounding error* problem in Imitation Learning

We seek an algorithmic scheme that:

- mimics the expert within its distribution
- returns to the expert's distribution if it deviates

Our approach:

- uses ensemble uncertainty as reward function
- can use any policy gradient algorithm
- has linear regret in certain settings
- simple and practically robust

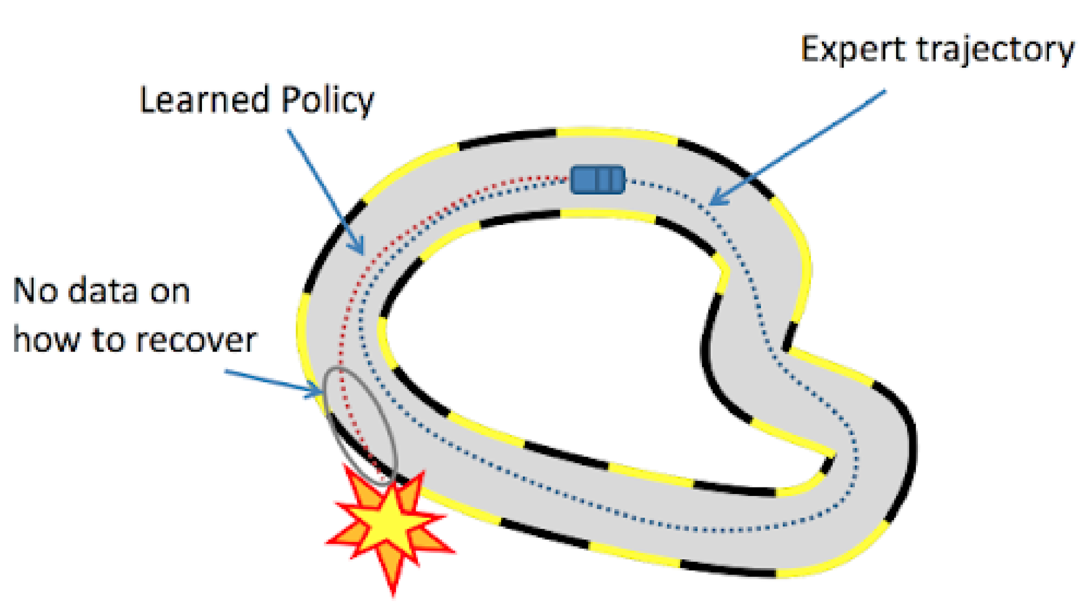
Compounding error problem

Behavior cloning treats imitation learning as a supervised learning problem.

$$J_{\mathrm{BC}}(\pi) = \mathbb{E}_{s\sim d_{\pi^\star}}[\|\pi^\star(\cdot|s) - \pi(\cdot|s)\|]$$

($d_{\pi^\star}$ is computed from demonstration data)

But doing this the model may suffer from the *cascading error problem*



this can be formalized with the *quadratic regret bound* where there exist problems when

$$J_{\mathrm{BC}}(\pi) = \epsilon \text{ and } \mathrm{Regret} = \Omega(\epsilon T^2)$$

(Ross and Bagnell, *AISTATS* 2010)

Our approach:

- **Our objective has two parts**

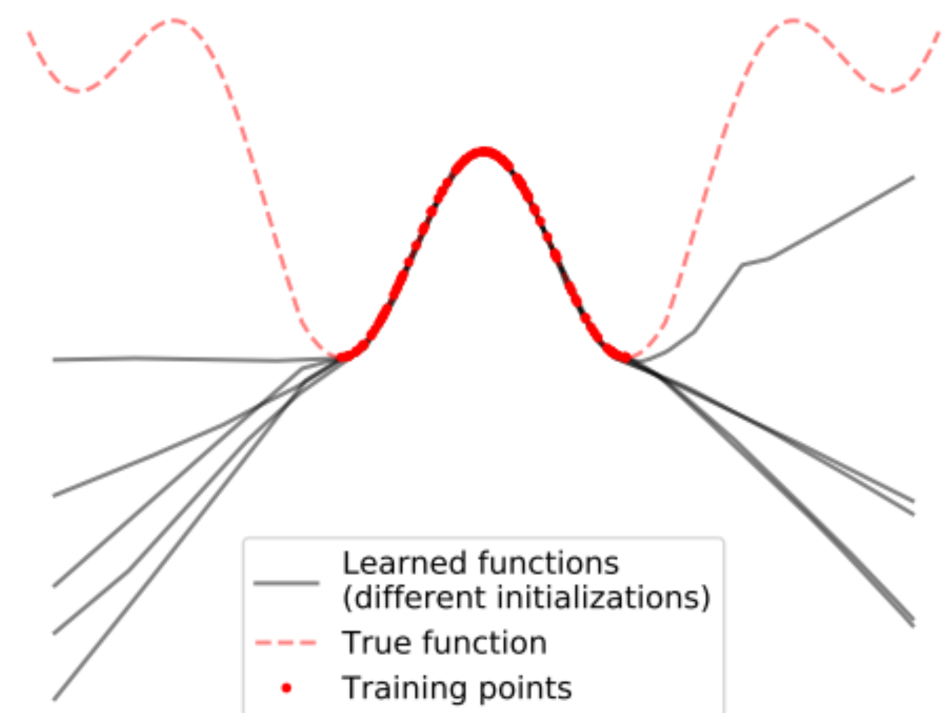
$$J_{\mathrm{alg}}(\pi) = \underbrace{\mathbb{E}_{s\sim d_{\pi^\star}}[\|\pi^\star(\cdot|s) - \pi(\cdot|s)\|]}_{J_{\mathrm{BC}}(\pi)} + \underbrace{\mathbb{E}_{s\sim d_\pi, a\sim\pi(\cdot|s)}[C_{\mathrm{U}}(s,a)]}_{J_{\mathrm{U}}(\pi)}$$

$J_{\mathrm{BC}}(\pi)$ is the *supervised behavior cloning cost*
(mimics the expert within its distribution)

$J_{\mathrm{U}}(\pi)$ is an *uncertainty cost*
(returns to the expert's distribution if it deviates)

$$C_{\mathrm{U}}(s,a) = \mathrm{Var}_{\pi\sim\Pi_{\mathrm{E}}}(\pi(a|s))$$

where Π_{E} is an *ensemble of policies* trained on the demonstration data



Key insight: ensemble *variance is high* where data is sparse and *variance is low* where data is dense

Our algorithm: DRIL

(DRIL: Disagreement-Regularized Imitation Learning)

Input π^* demonstration data

train π and Π_{E} using data

For $t = 1, \ldots$

- Perform supervised update to minimize $J_{\mathrm{BC}}(\pi)$ using $\mathcal{D}$
- Perform step of policy gradient using $C_{\mathrm{U}}^{\mathrm{clip}}(s,a)$

End For

Guarantees and further details

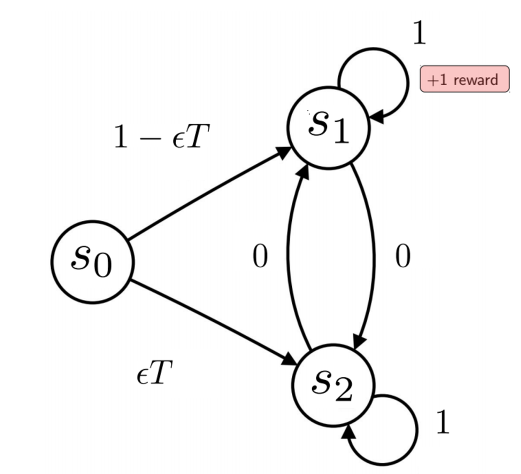
Regret Gurantee:

$J_{\mathrm{alg}}(\pi)$ has regret $\mathcal{O}(\kappa\epsilon T)$

we define κ as:

$$\kappa = \min_{\mathcal{U}\subseteq\mathcal{S}} \frac{\alpha(\mathcal{U})}{\beta(\mathcal{U})}$$

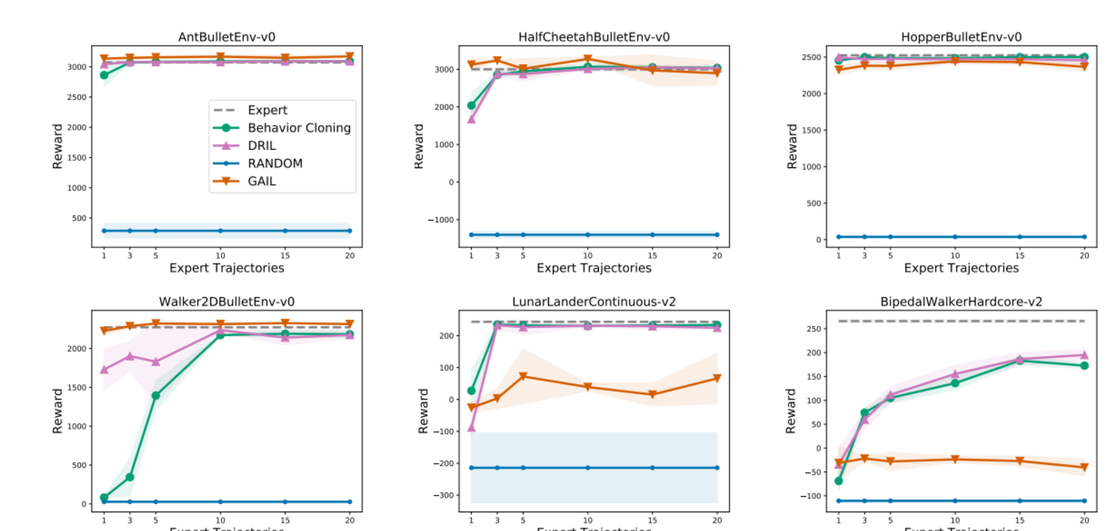
where $\alpha(\mathcal{U})$ is *concentrability* inside of $\mathcal{U}$ and $\beta(\mathcal{U})$ is *minimum variance of the ensemble* outside of $\mathcal{U}$



we can show that behavior cloning has quadratic regret on this problem and dril has linear regret

Experiments

Continuous Control



Atari