Disagreement-Regularized Imitation Learning

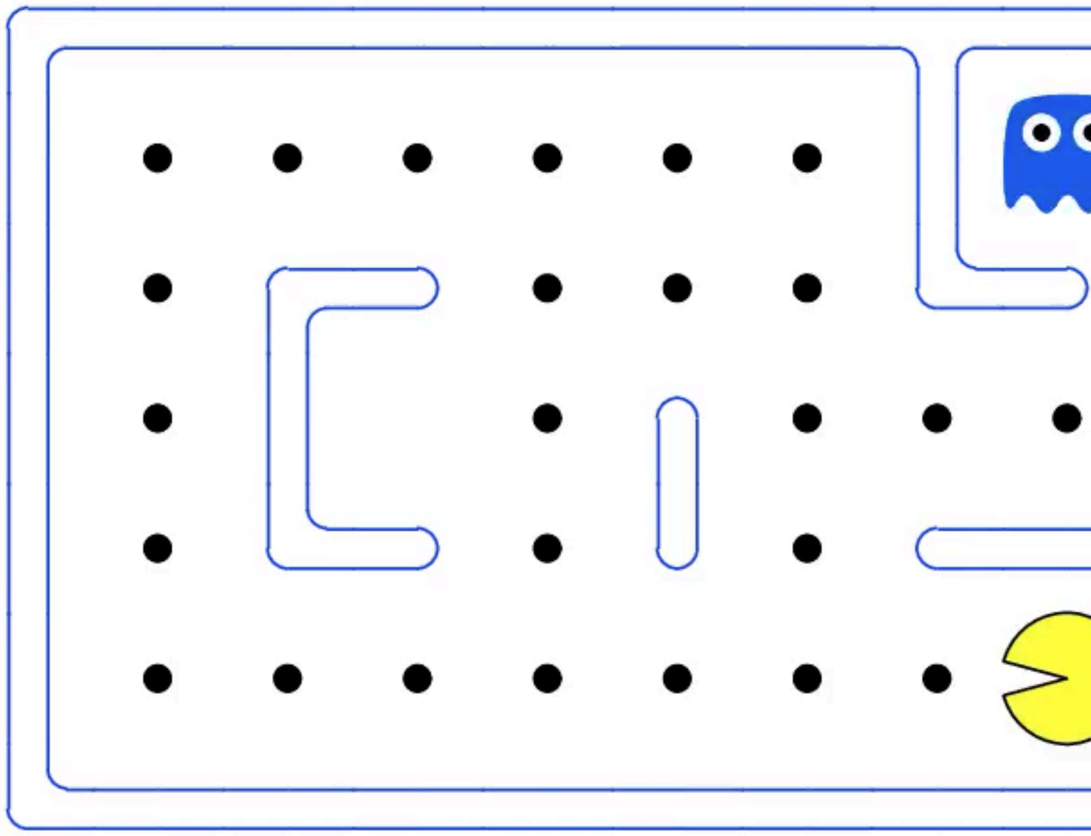
Kianté Brantley,[1] Wen Sun,[2] Mikael Henaff [2]

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

Imitation Learning

Expert Demonstrator

- state
- actions   up, down, left, right



Training set: D = {(state, actions)} from expert π^*

Goal: learn agent π_θ(s) -> a

Imitation Learning
using Behavior Cloning

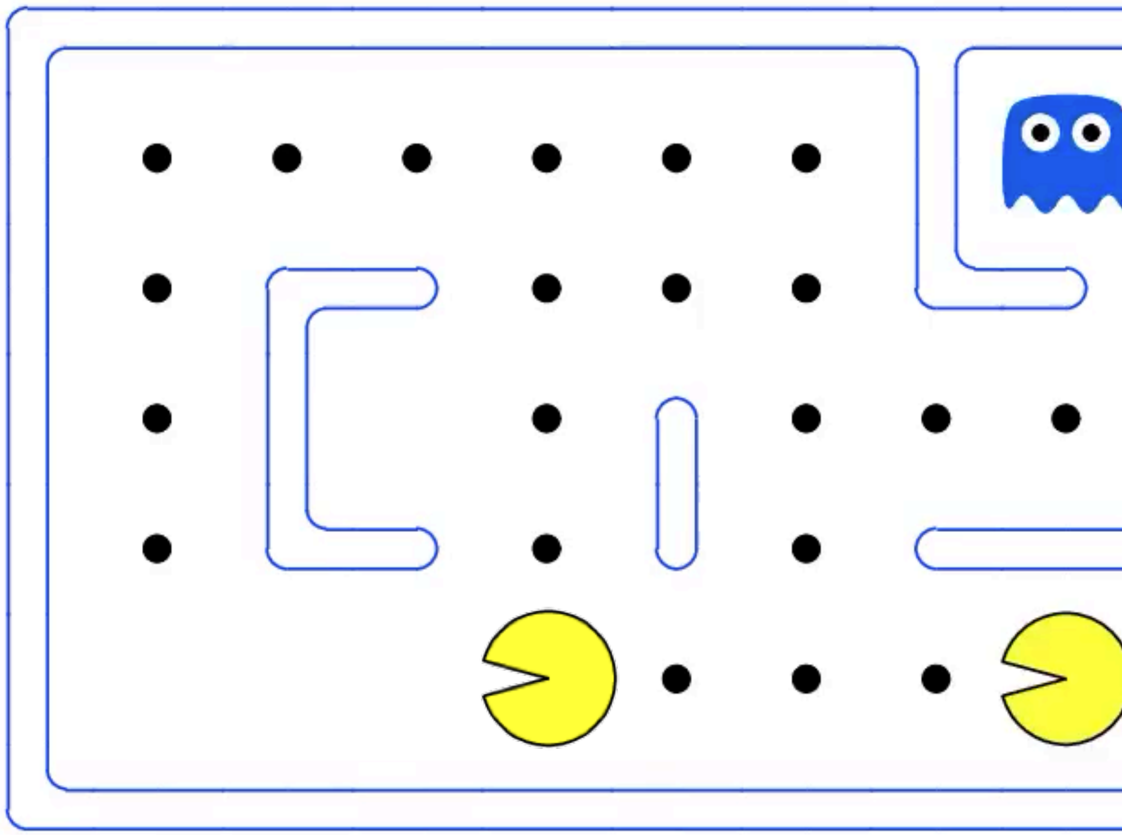
$$J_{BC}(\pi) = \mathbf{E}_{s \sim d_{\pi^*}}\left[\ell(\pi_\theta(s), \pi^*(s))\right]$$

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]

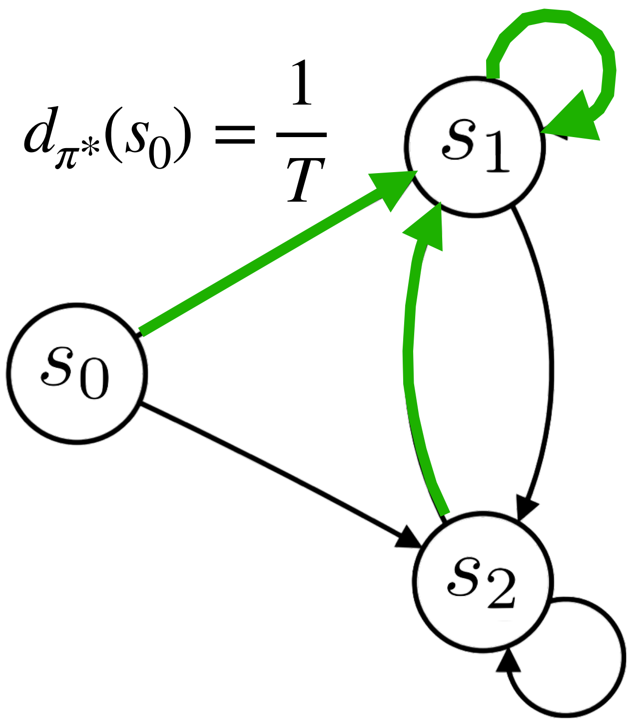
Formalizing

the compounding error problem

Given an expert policy: π^*

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

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



Behavior Cloning Loss:

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

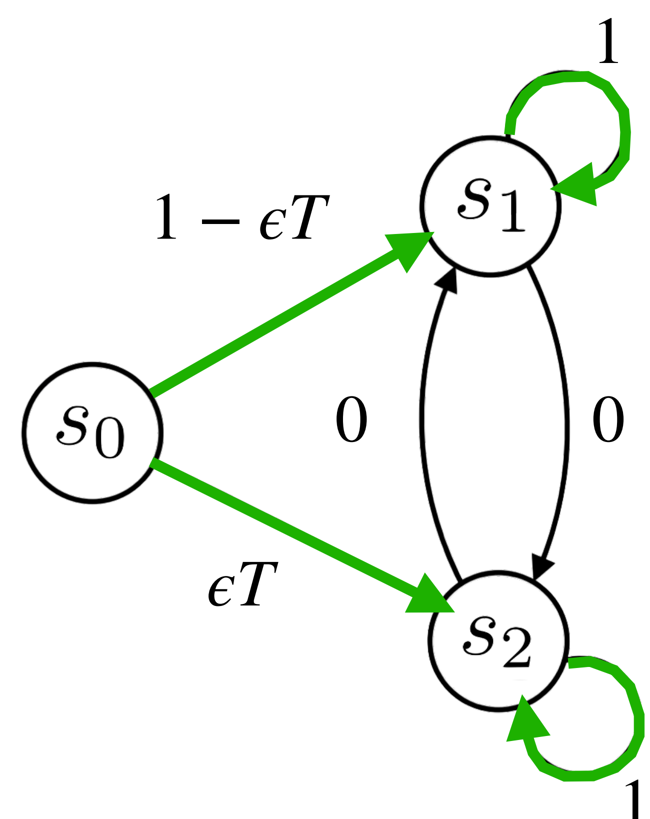
(loss is small)

Behavior Cloning Regret:

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

(quadratic regret)

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



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

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

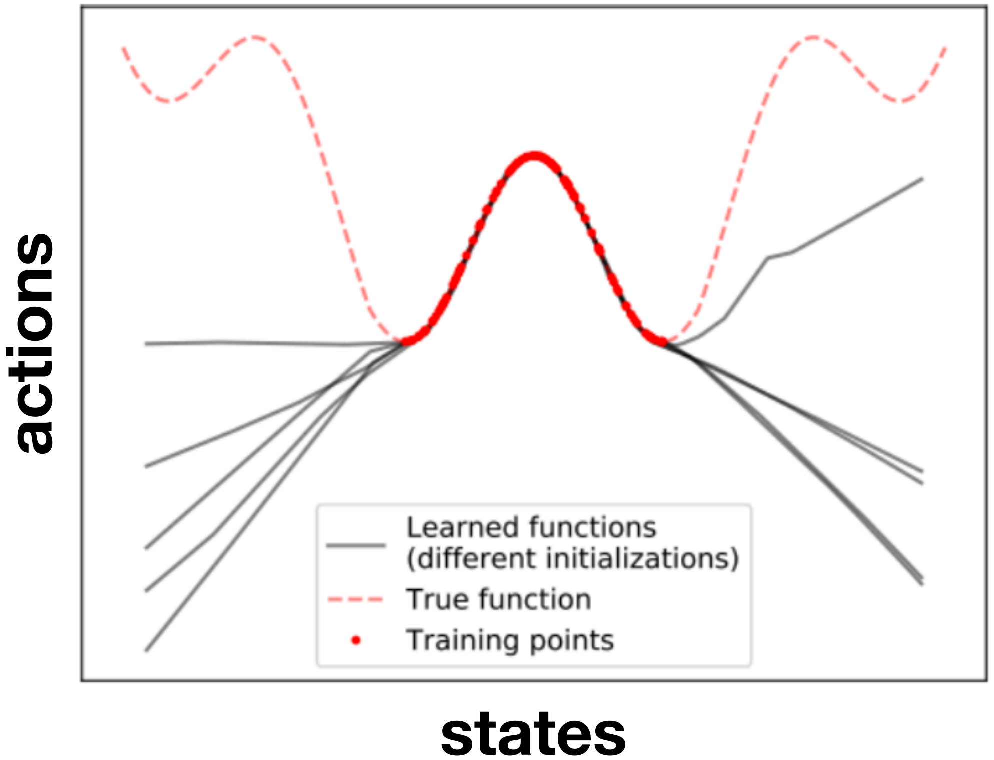
Our Approach

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 polices $\Pi_E = \{\pi_1, \ldots, \pi_E\}$ **on demonstration data** D

Uncertainty Cost: $C_{\mathrm{U}}(s, a) = \mathrm{Var}_{\pi \sim \Pi_{\mathrm{E}}}(\pi(a \mid s))$

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 π **using demonstration data** D

for $i = 1$ **to** $\ldots$ **do**

- **Perform one gradient update to minimze** $J_{BC}(\pi)$ **using minibacth from** D
- **Perform one step of policy gradient to minimize** $\mathbf{E}_{s \sim d_\pi, a \sim \pi(\cdot|s)}\left[C_{\mathrm{U}}(s, a)\right]$

end for

Our Approach
DRIL (Analysis)

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

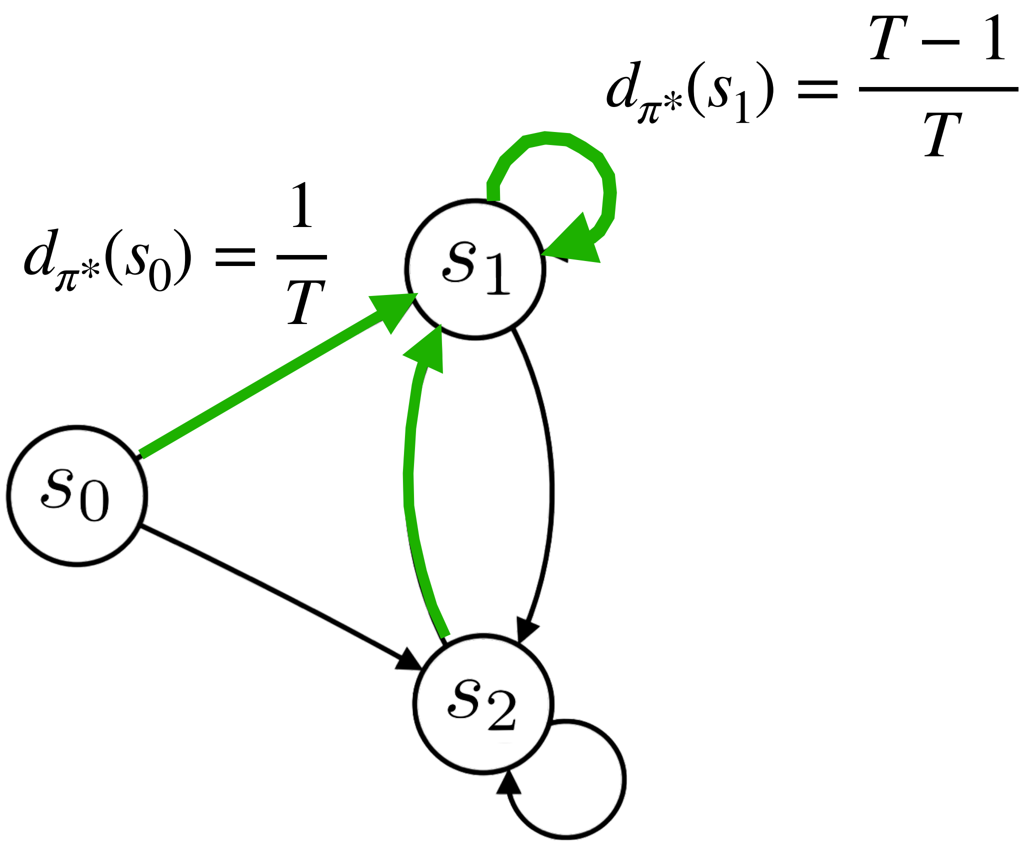
Assumption 1: (Realizability) $\pi^* \in \Pi$

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: π^*



Behavior Cloning Regret:

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

(quadratic regret)

DRIL Regret $\mathcal{O}(\epsilon\kappa T)$

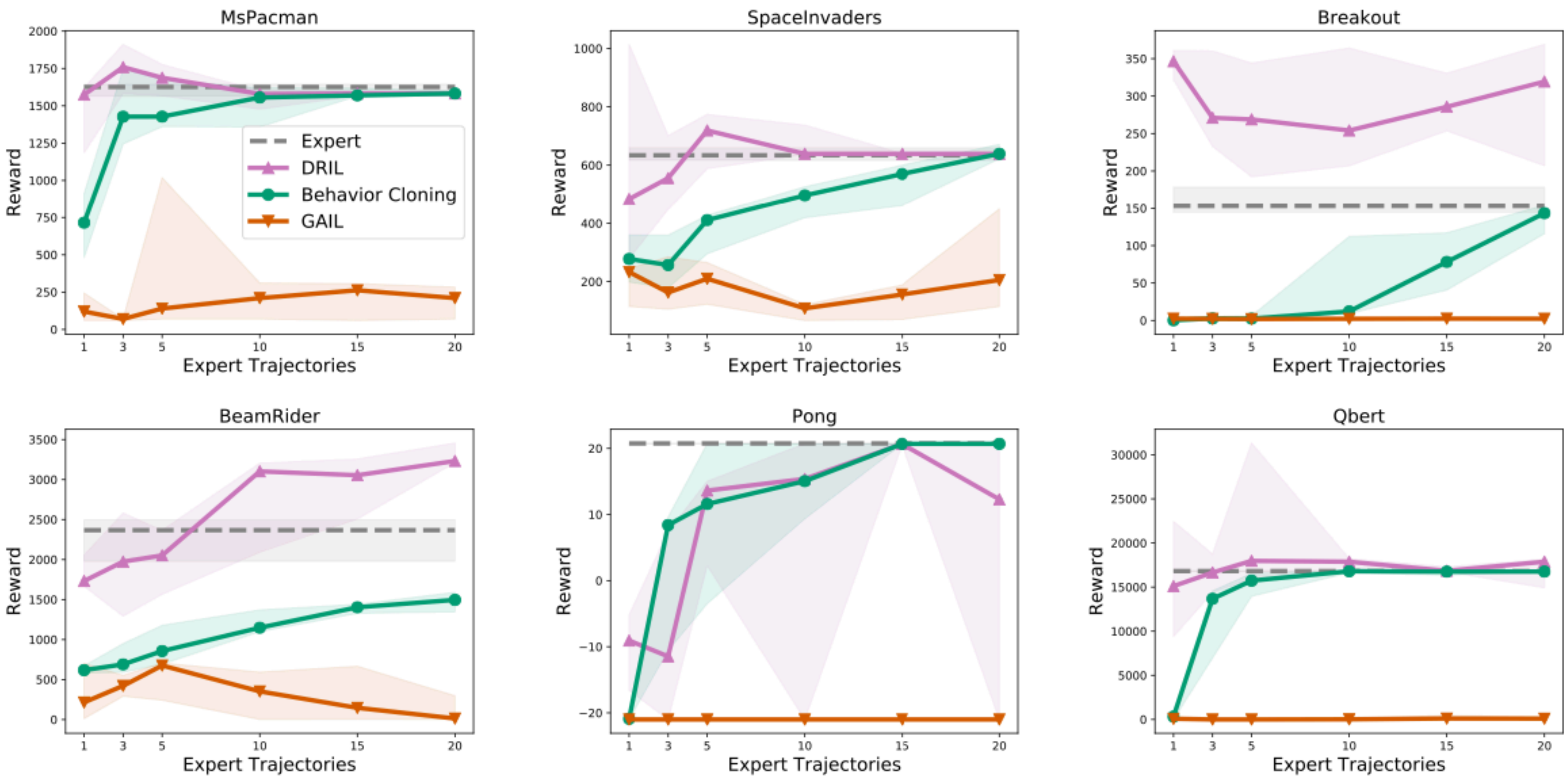
DRIL Regret:

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

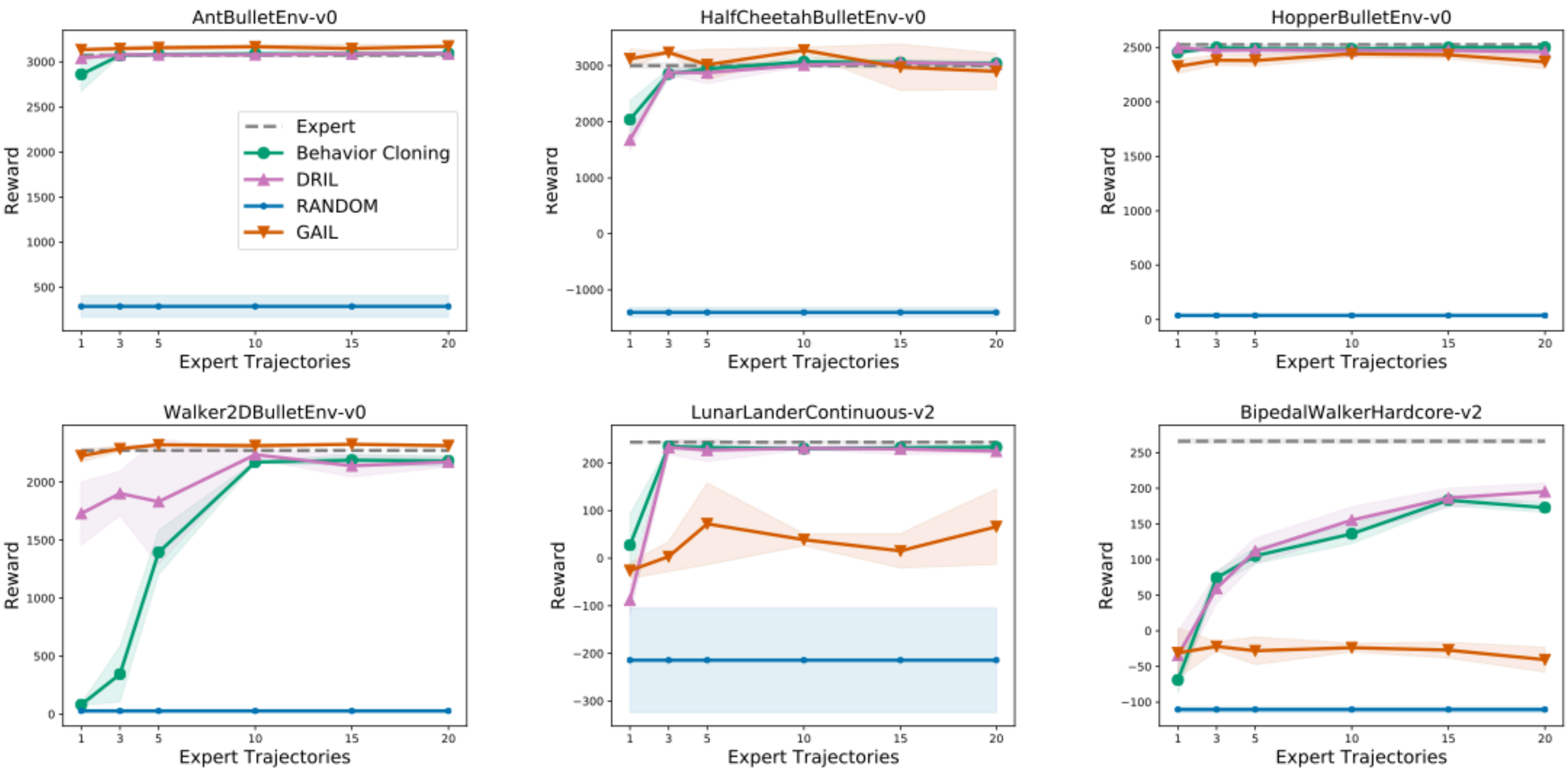
$$\kappa = 1 + \mathcal{O}\left(\frac{1}{\sqrt{|\text{ensemble}|}}\right)$$

(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