Interactive and Hybrid Imitation Learning: Provably Beating Behavior Cloning

Yichen Li and Chicheng Zhang
University of Arizona

Imitation Learning (IL)

Given: Expert Demonstrations.
Goal: Learn good policy for sequential decision making.

Our Contributions:

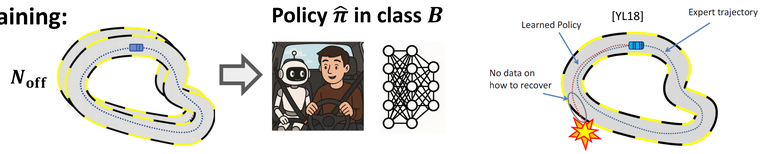
First to highlight the benefit of state-wise interactive annotation and hybrid feedback in imitation learning.

When annotation cost is measured per state, interactive IL algorithms can provably outperform Behavior Cloning (BC). (1) We show **Stagger**, a one-sample-per-round variant of DAgger, beats BC in low-recovery-cost settings; (2) We propose **Hybrid IL**, combining offline demonstrations with interactive annotations, and introduce **Warm-Stagger** (WS), which achieves lower annotation cost compared to BC and Stagger in a toy MDP motivated by practical applications; (3) Experiments on continuous-control tasks show that both interactive and hybrid methods outperform BC.

Behavior Cloning and Covariate Shift

Imperfect Trained Policy -> Unseen States -> Inability to Recover

Training: N_{off} **Policy $\hat{\pi}$ in class B**



Test: Deploy $\hat{\pi}$ in Episodic MDP $\mathcal{M}$ with length H.

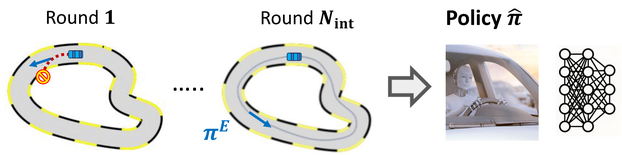
$J(\hat{\pi})$

Interactive IL and State-wise DAgger (Stagger)

Learner adaptively queries expert for demonstrations [RGB11].

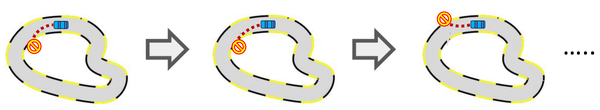
Training: Round $\mathbf{1}$ Round N_{int} **Policy $\hat{\pi}$**

Expert π^E 1 State-wise Annotation With cost $C \geq 1$



The Cold Start Problem in Interactive IL

Early Crashes -> Fail to Explore -> Limited Data Coverage -> Slow Learning



Hybrid IL

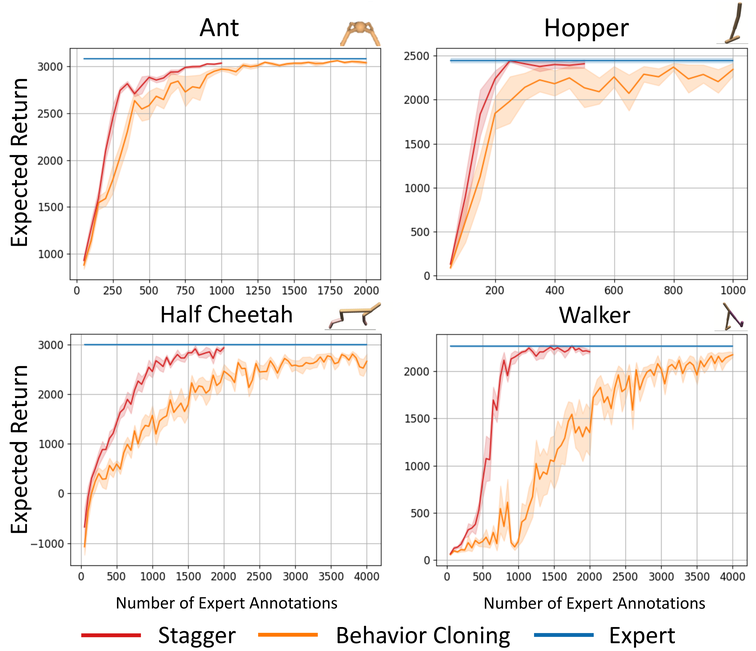
Learner has N_{off} offline expert trajectories and interactive state-wise annotations up to N_{int} times. Each offline (state, action) pair costs 1, and each interactive query costs $C \geq 1$.

Sample Efficiency and Annotation Cost

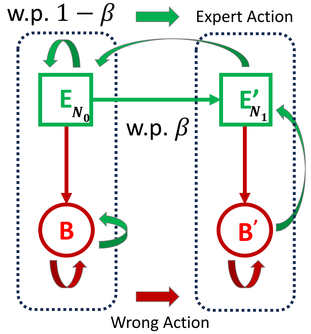
μ: recoverability

Realizable Expert	Policy Suboptimality	Annotation Cost	Type
Behavior Cloning (BC) [FBM24]	$\frac{H\log(B)}{N_{\text{off}}}$	HN_{off}	Offline Trajectories
Stagger (ours)	$\frac{\mu H\log(B)}{N_{\text{int}}}$	CN_{int}	Interactive State-wise
Warm-Stagger (WS) (ours)	$\min\left(H^2\frac{\log(B)}{N_{\text{off}}}, \mu H\frac{\log(B)}{N_{\text{int}}}\right)$	$HN_{\text{off}} + CN_{\text{int}}$	Hybrid

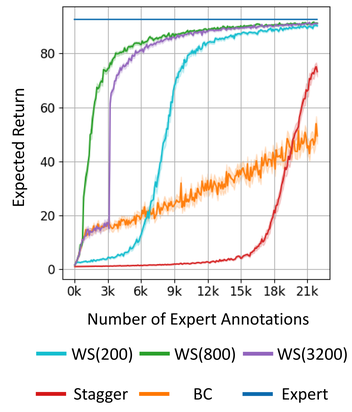
Continuous Control Experiment with MLPs



The Benefit of Hybrid Imitation Learning

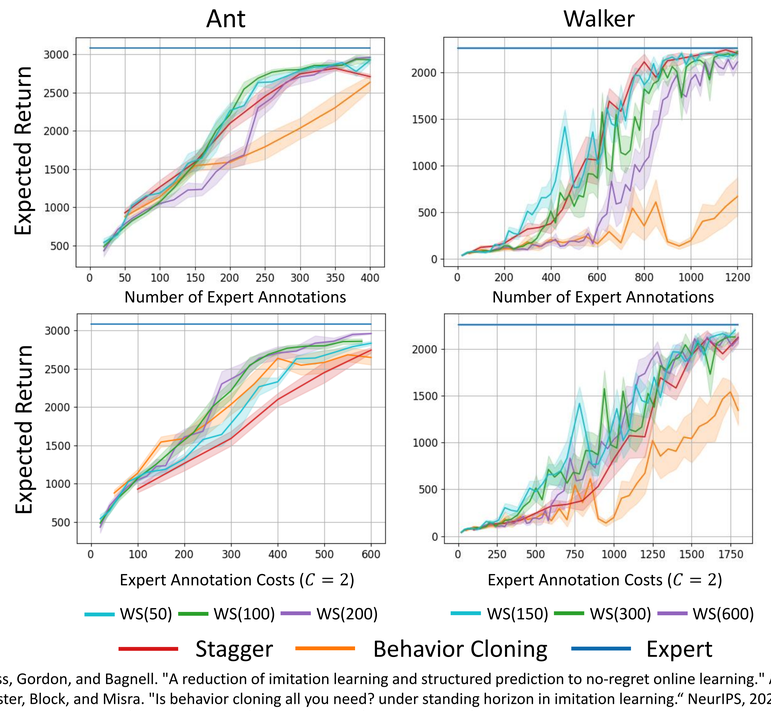


Cold Start **Covariate Shift**



Theorem: For the above MDP, the following happens:
BC with $N_{\text{off}} = \Omega(S)$ offline trajectories is $\Omega(H)$ suboptimal.
Stagger with $N_{\text{int}} = \Omega(HS)$ interactive annotations is $\Omega(H)$ suboptimal.
Warm-Stagger with $N_{\text{off}} = O\left(\frac{S}{H}\right)$, $N_{\text{int}} = O(1)$ achieves expert's performance.

Continuous Control with Different Annotation Costs



[RGB11] Ross, Gordon, and Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." AISTATS, 2011.
[FBM24] Foster, Block, and Misra. "Is behavior cloning all you need? under standing horizon in imitation learning." NeurIPS, 2024.
[YL18] Yisong Yue, Hoang M Le https://sites.google.com/view/icml2018-imitation-learning