Understanding the utility of interaction in imitation learning

Chicheng Zhang

Joint work with Yichen Li (Amazon)

THE UNIVERSITY OF ARIZONA

Outline

- Introduction to Imitation Learning
- Preliminaries
- The Benefit of Interaction in the Realizable Setting
- Efficient Algorithms Beyond Realizable Setting
- Hybrid Imitation Learning

THE UNIVERSITY OF ARIZONA

Introduction to Imitation Learning

What is Imitation Learning (IL)?

Given: Demonstrations or demonstrator (expert)

Goal: Learn a good control policy

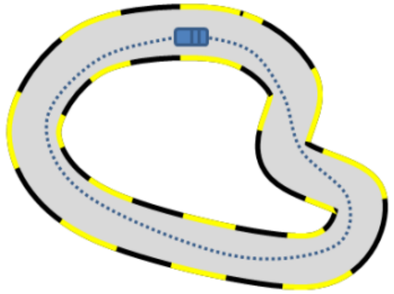
Applications: Autonomous driving Robot control Game playing

THE UNIVERSITY OF ARIZONA

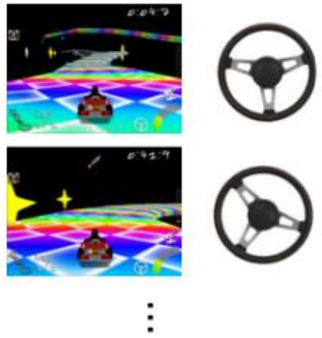
Example: Learning to Drive from Demonstrations

Training stage:

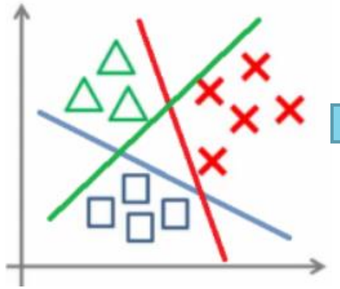
Expert demonstrations



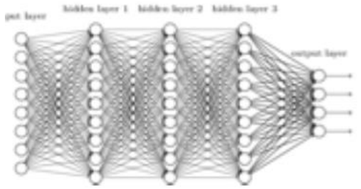
Dataset



Learning algorithm

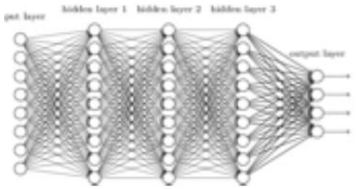


Policy $\hat{\pi}$

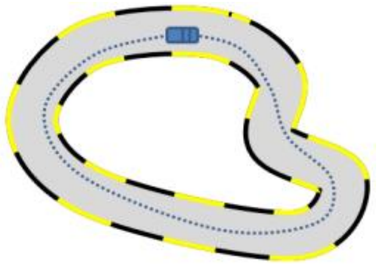


Test stage:

Policy $\hat{\pi}$



Perform control in environment



THE UNIVERSITY OF ARIZONA

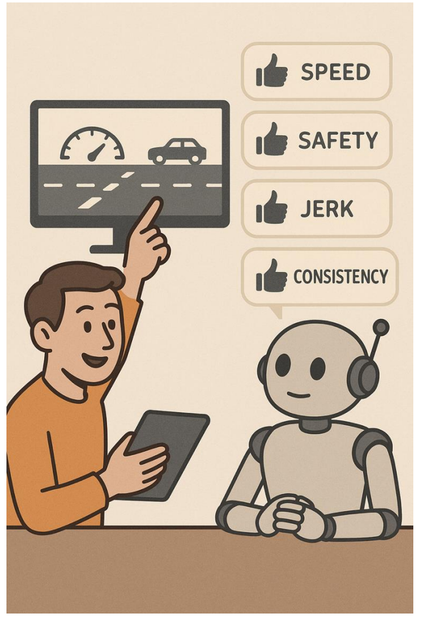
Pictures from Stephane Ross

Imitation Learning vs. Reinforcement Learning

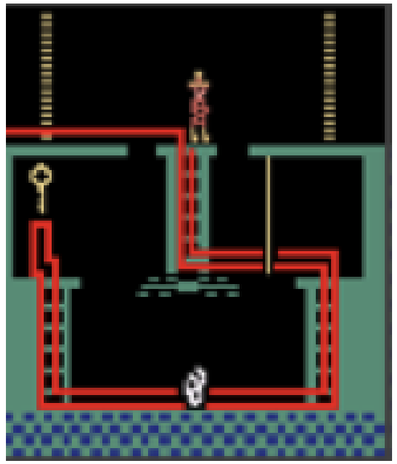
1. Designing reward functions in RL can be challenging.

Example: autonomous driving

But: easy for humans to demonstrate driving.



2. Sidesteps exploration challenges in reinforcement learning.



THE UNIVERSITY OF ARIZONA

Imitation Learning: Two Paradigms

Offline IL: IL with offline expert demonstrations only.

Example: Behavior Cloning [P88] solves offline IL with supervised learning.

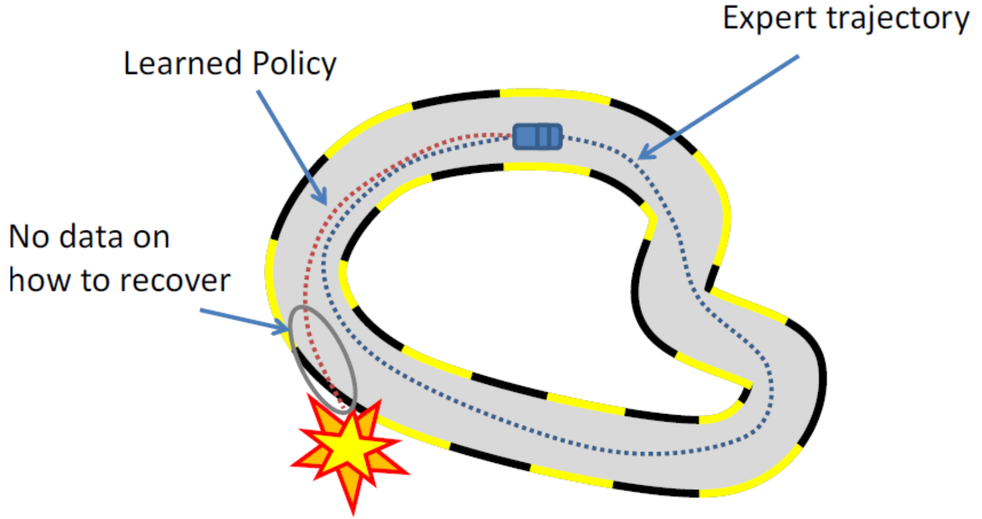
THE UNIVERSITY OF ARIZONA

[P88] Pomerleau, 1988, Alvinn: an autonomous land vehicle in a neural network

How is IL Different from Supervised Learning?

Supervised Learning: $\boldsymbol{D}_{\text{test}} = \boldsymbol{D}_{\text{train}}$

Covariate Shift (Compounding Error) in IL: $\boldsymbol{D}_{\text{test}} \neq \boldsymbol{D}_{\text{train}}$



Imperfect Trained Policy ⇒ Unseen States ⇒ Unable to Recover

Pomerleau (1988): "the network must not solely be shown examples of accurate driving, but also *how to recover* once a mistake has been made."

When Does Behavior Cloning Work?

Good scenario 1:
The learner exactly follows expert demonstrations.

Good scenario 2:
The offline dataset covers all relevant states to allow learning how to recover from mistakes.

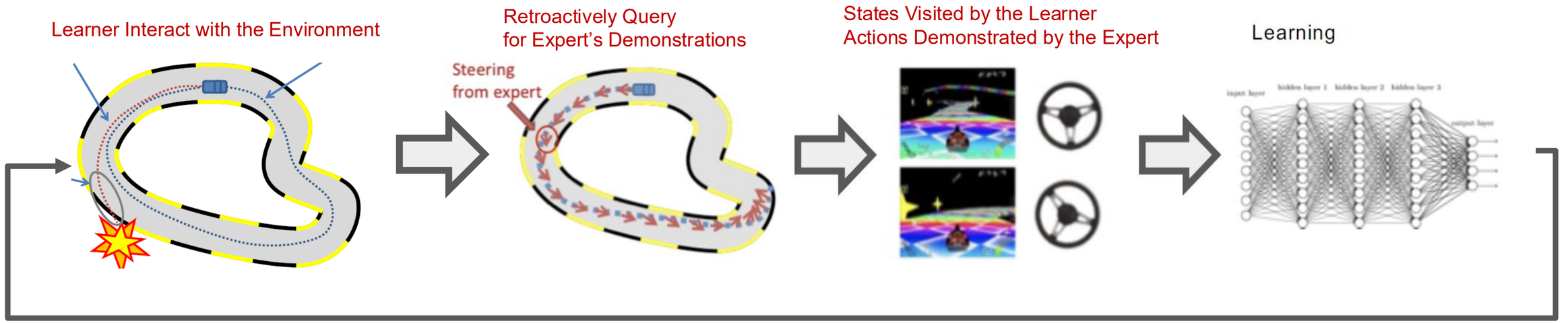
In practice: Given limited offline data, the learner can fail to follow demonstrations and fail to recover from mistakes.

THE UNIVERSITY OF ARIZONA

Imitation Learning: Two Paradigms

Interactive IL: IL with environment interactions and interactive expert annotations.

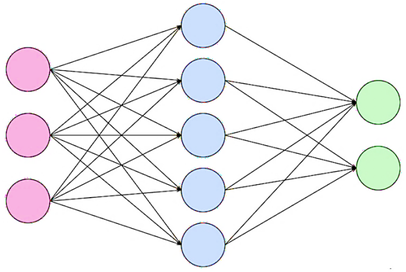
Example: DAgger (Data Aggregation) [RGB11]

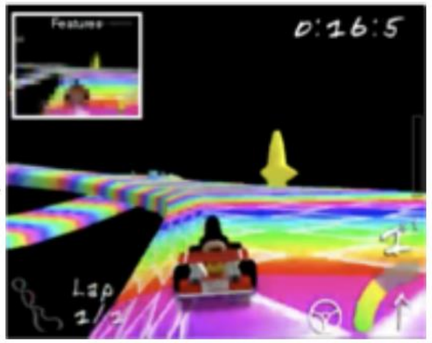


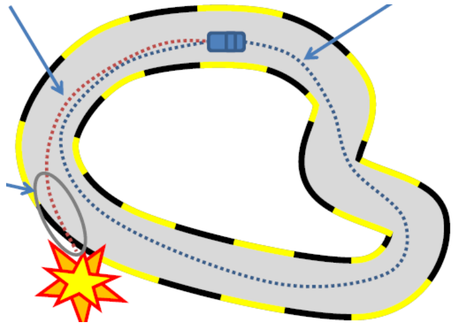
[RGB11] Ross, Gordon, and Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning."

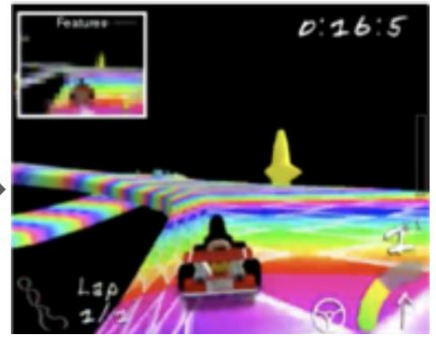
Why Interaction helps in Imitation Learning?

Interactive imitation learning mitigates the covariate shift issue by learning to recover

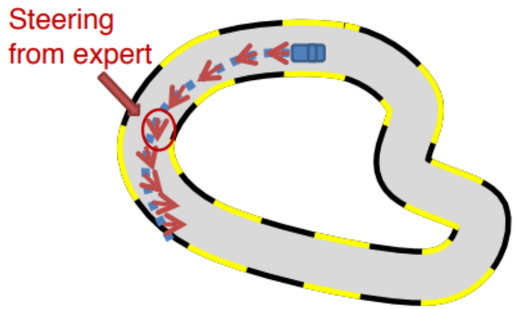






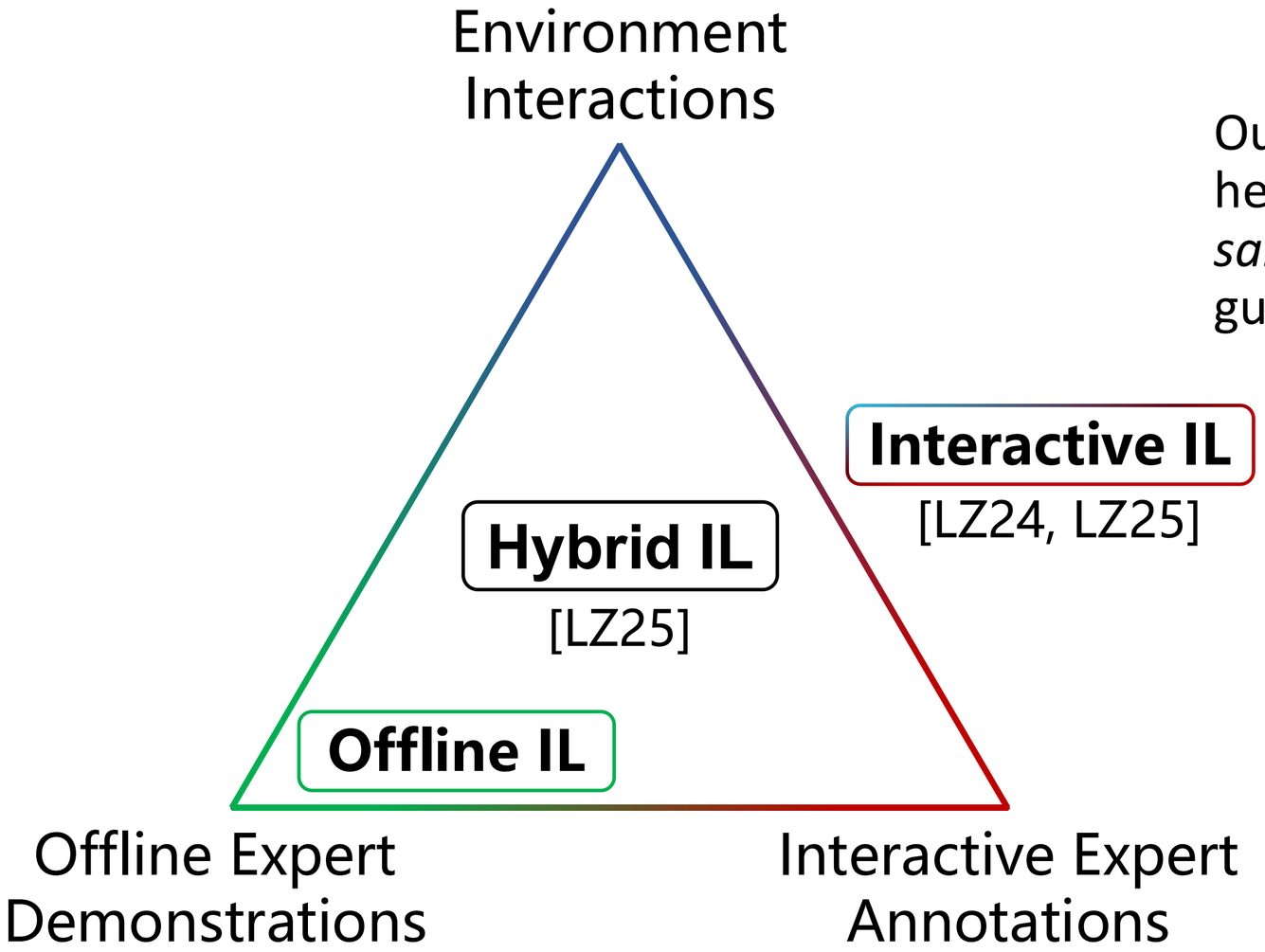


Imperfect Policy Unseen States Crash Query Expert Demonstrations



[RGB11] Ross, Gordon, and Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning."

Understanding the benefit of interaction in IL



Our perspective: understanding how interaction helps imitation learning in establishing better *sample-efficiency* and *policy suboptimality* guarantees

THE UNIVERSITY OF ARIZONA

[LZ24] Li, and Zhang. "Agnostic Interactive Imitation Learning: New Theory and Practical Algorithms." ICML 2024
[LZ25] Li, and Zhang. "Interactive and Hybrid Imitation Learning: Provably Beating Behavior Cloning." NeurIPS 2025

Preliminaries

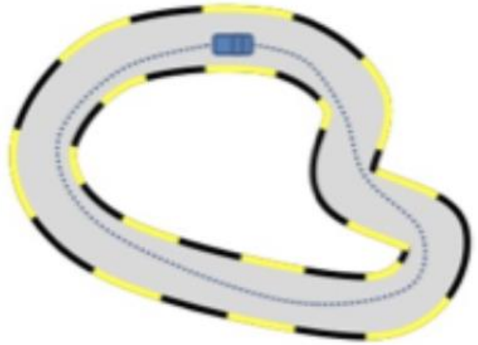
Basic Settings: Fixed-horizon Markov Decision Process (MDP)

Episodic MDP: $\mathcal{M} = (H, \mathcal{S}, \mathcal{A}, P, c)$

Episode Length: H

State Space: $\mathcal{S}$

Action Space: $\mathcal{A}$



Transition Dynamics: $P = \{P_t(\cdot \mid s, a) : (s, a) \in \mathcal{S} \times \mathcal{A}\}_{t=1}^{H}$

Cost Function: $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ (For Evaluation Only)

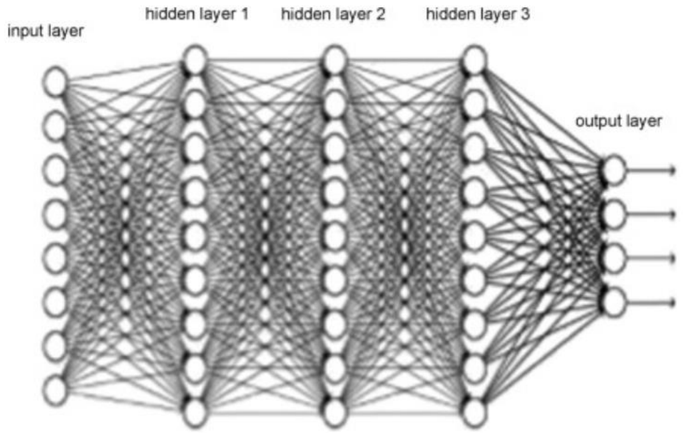
THE UNIVERSITY OF ARIZONA

Basic Settings: Policy Class

Stationary deterministic expert policy $\pi^E: \mathcal{S} \to \mathcal{A}$

Base policy class $\mathcal{B} \subset \{\mathcal{S} \to \mathcal{A}\}$ of size B (Finite but exponential in #parameters)

π^E is said to be *realizable* if $\pi^E \in \mathcal{B}$



THE UNIVERSITY OF ARIZONA

Basic Settings: Policy Rollout in MDP

Given policy $\boldsymbol{\pi}$, MDP $\mathcal{M}$,

For every time step $t = 1, .., H$:

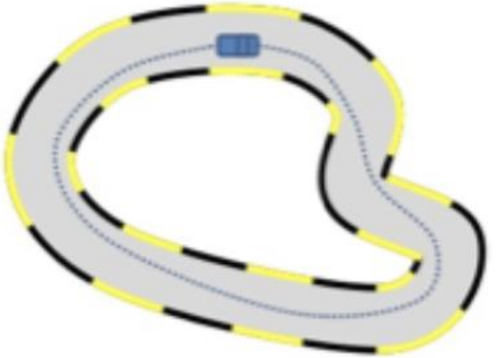
observes a state s_t

takes an action a_t by π

receives cost $c_t \coloneqq c(s_t, a_t)$ (Not Observable)

moves to $s_{t+1} \sim P_t(\cdot \,|s_t, a_t)$

$$\mathcal{M} = (H, \mathcal{S}, \mathcal{A}, P, c)$$



Total cost $\sum_{t=1}^{H} c_t$ ⇒ Expected cost $\mathrm{J}(\pi) \coloneqq \mathbb{E}_{\pi,\mathcal{M}}[\sum_{t=1}^{H} c_t]$

(For Evaluation Only)

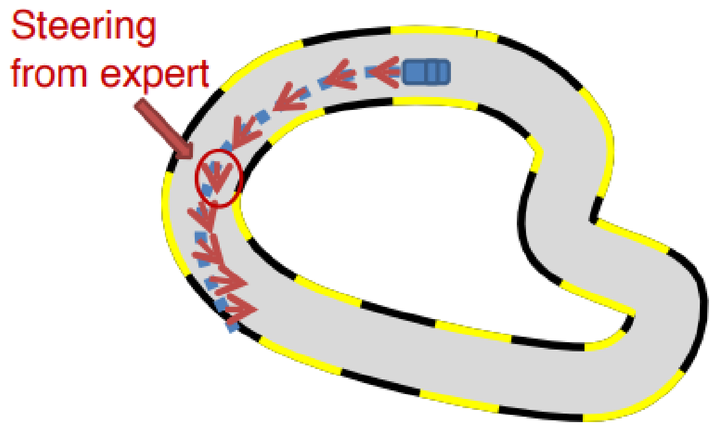
(Smaller the better)

The Benefit of Interaction in the Realizable Setting

Interactive Imitation Learning

Given:
Environment interactions & ability to query expert for demonstrations

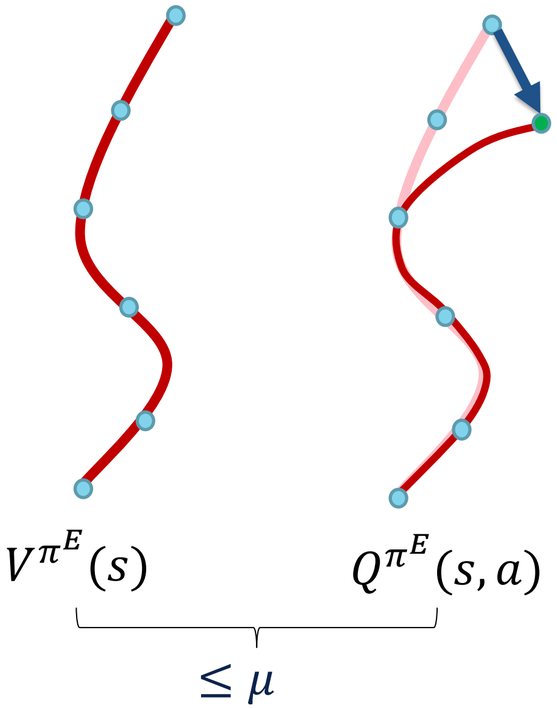
Goal:
Output policy competitive with the expert, using a few queries



THE UNIVERSITY OF ARIZONA

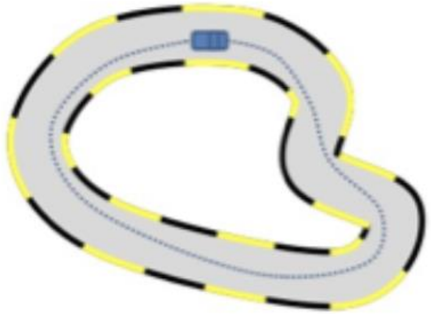
Key Assumption: μ-recoverability

The cost of recovery when deviating one-step from π^E



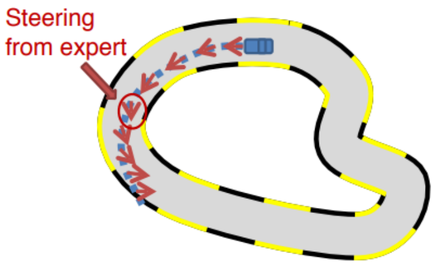
$\mu \ll H$ (favorable case)

$\mu \approx H$ (worst case)



DARPA Grand Challenge 2005 https://www.youtube.com/watch?v=Natbh3t6mHM

Sample Complexity of Imitation Learning: State of the Art



Deterministic Realizable Expert π^E
For $J(\hat{\pi}) - J(\pi^E) \leq \epsilon$:

Approach	#Full Trajectory Annotations
LogLossBC [FBM24] (Offline)	$H\frac{\log(B)}{\epsilon}$
LogLossDAgger [FBM24] (Interactive)	$\mu H\frac{\log(B)}{\epsilon}$
Lower Bound (among interactive algorithms)	$H\frac{\log(B)}{\epsilon}$

Typically: $1 \leq \mu \leq H$

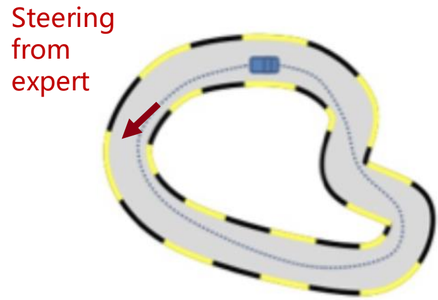
Interactivity does not seem to enjoy sample complexity benefits?

[FBM24] Foster, Block, and Misra. "Is Behavior Cloning All You Need?"

Sample Complexity of Imitation Learning: Our Results

Observation: annotating a length-H trajectory requires annotating H individual states

Approach	#State-wise annotations
LogLossBC [FBM24] (Offline)	$\frac{H^2 \log(B)}{\epsilon}$
Stagger (Ours, Interactive)	$\mu H \frac{\log(B)}{\epsilon}$



Typically: $\mathbf{1 \leq \mu \leq H}$

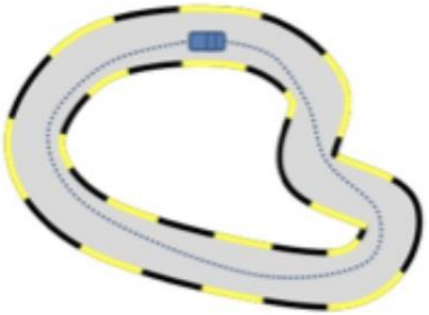
Interactivity enjoys sample complexity benefits, if we measure it using number of state-wise annotations!

[FBM24] Foster, Block, and Misra. "Is Behavior Cloning All You Need?"

Stagger (State-wise DAgger): Main idea

Find policy π that minimizes its Imitation Loss

$$L(\pi) \coloneqq \mathbb{E}_{s\sim \boldsymbol{d_\pi}}[I(\boldsymbol{\pi}(s) \neq \pi^E(s))]$$



$\boldsymbol{d_\pi}$**:** average state visitation distribution by policy π

Why this is a good objective: $\mathrm{J}(\pi) - \mathrm{J}(\pi^E) \leq \mu H \cdot L(\pi)$ [KL02]

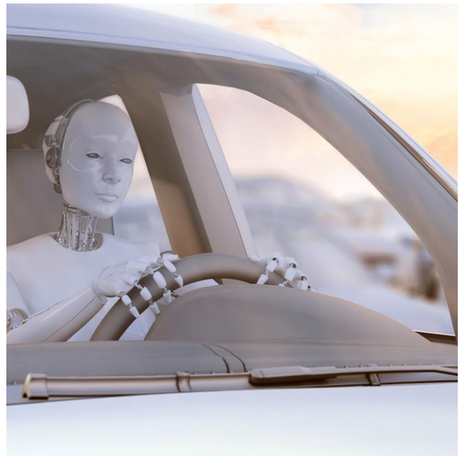
Not a supervised learning problem; but interaction enables optimizing $L(\pi)$ sample-efficiently!

[KL02] Kakade and Langford. "Approximately optimal approximate reinforcement learning."

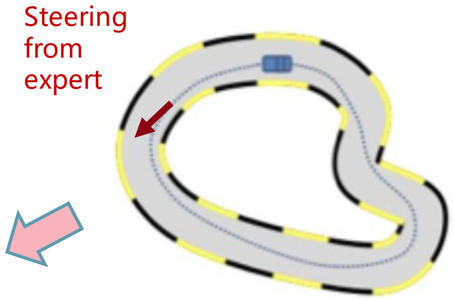
Using online learning to minimize imitation loss

Stagger reduces minimizing $L(\pi)$ to an online learning game [RGB11]:

Require:
$\mathcal{B} \subset \{\mathcal{S} \to \mathcal{A}\}$



Learned policy $\pi_n \in \mathcal{B}$



Approximation of $F_n(\pi) := \mathbb{E}_{s \sim d_{\pi_n}}[I(\pi(s) \neq \pi^E(s))]$

$F_n(\pi_n) = L(\pi_n)$

$\vdots$

Learner rolls out π_n, samples $s \sim d_{\pi_n}$, and queries π^E for annotation.

Return $\hat{\pi} \in \{\pi_n\}_1^N$

Key observation in Stagger: we only need to query one state for each iteration!

[RGB11] Ross, Gordon, and Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning."

The Guarantee of Stagger

Regret of online learning: $\mathrm{Reg}_N := \sum_{n=1}^{N} F_n(\pi_n) - \min_{\pi \in \mathcal{B}} \sum_{n=1}^{N} F_n(\pi)$

$\| $

$\sum_{n=1}^{N} L(\pi_n)$

Approximation Error
(0 when realizable)

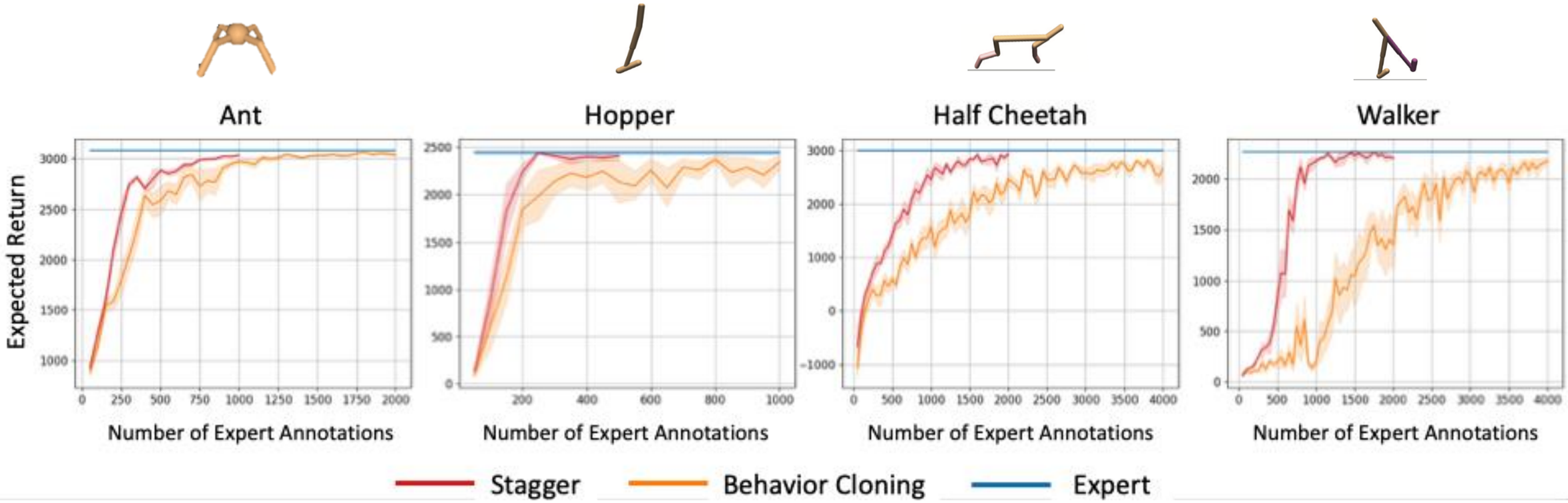
Regret minimization ⇨ Policy with low imitation loss

Theorem [LZ25]: Stagger admits an online learning regret of $\mathrm{Reg}_N = O(\log(B))$, and thus can learn a policy ϵ-suboptimal to the expert with $N = O\left(\frac{\mu H \log(B)}{\epsilon}\right)$ expert annotations.

[LZ25] Li, and Zhang. "Interactive and Hybrid Imitation Learning: Provably Beating Behavior Cloning."

Experimental Validation

Continuous control tasks from OpenAI Gym; $H = 1000$ steps.



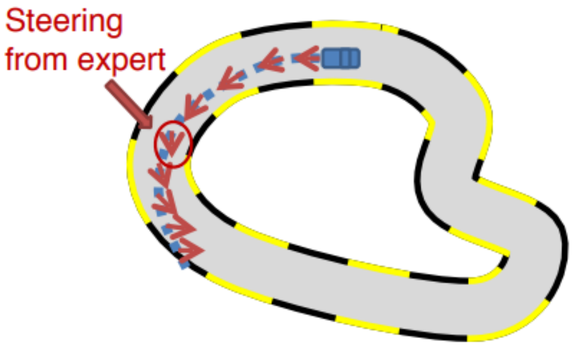
Stagger matches BC with 50% or fewer annotations.

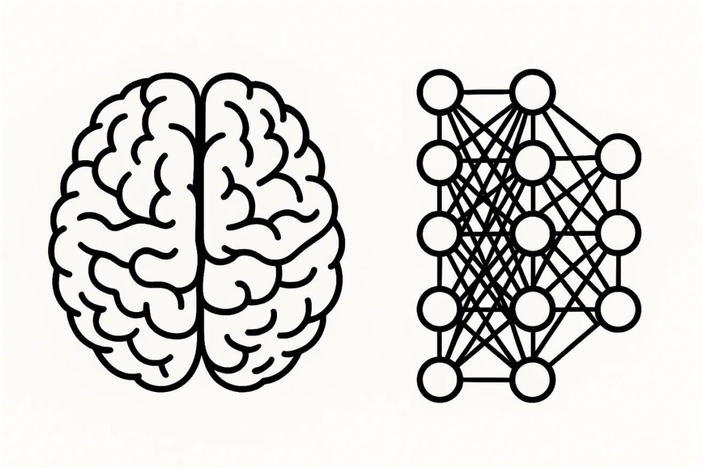
Efficient Algorithms Beyond Realizable Setting

Interactive IL with Nonrealizability

Given:
Environment interactions & ability to query expert for demonstrations

Goal:
Output policy competitive with the expert, using a few queries

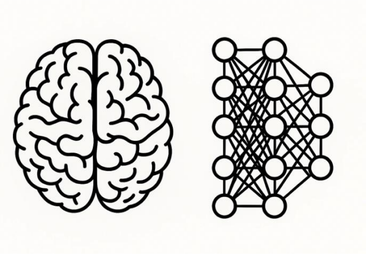




Nonrealizable setting: expert policy π^E may not lie in policy class $\mathcal{B}$

THE UNIVERSITY OF ARIZONA

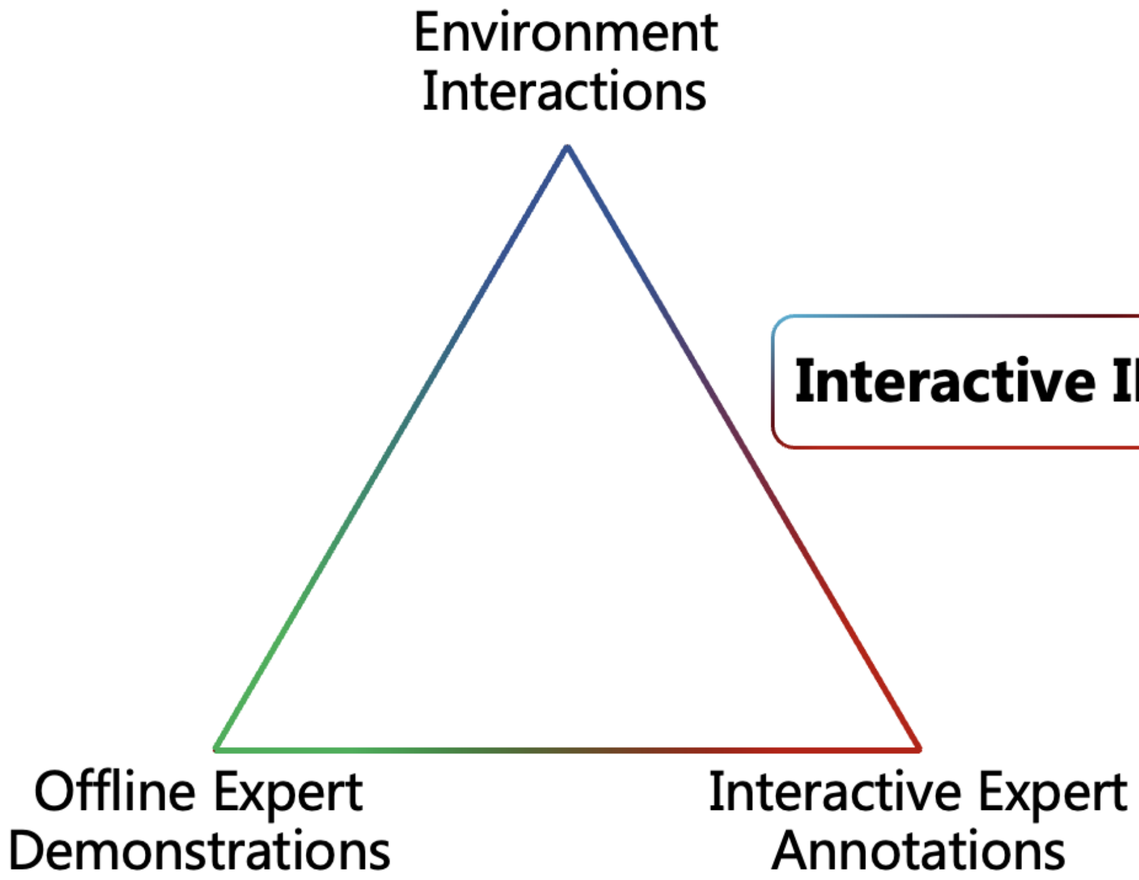
Bridging Theory and Practice



Can we design statistically and computationally efficient algorithms for interactive IL with a non-realizable expert?

THE UNIVERSITY OF ARIZONA

The DAgger Reduction Framework



Interactive IL ← **Online Learning**

Guarantee: Regret Minimization

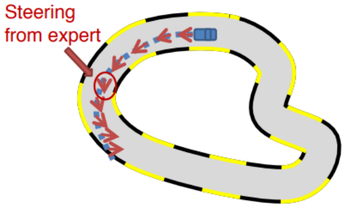
$$\mathrm{Reg}_N := \sum_{n=1}^{N} F_n(\pi_n) - \min_{\pi \in \mathcal{B}} \sum_{n=1}^{N} F_n(\pi)$$

Challenges in Using the DAgger Reduction Framework

$$\mathrm{Reg}_N := \sum_{n=1}^{N} F_n(\pi_n) - \min_{\pi \in \mathcal{B}} \sum_{n=1}^{N} F_n(\pi)$$

How to minimize Reg_N?

$\mathrm{E}_{s \sim d_{\pi_n}}[I(\pi(s) \neq \pi^E(s))]$



[RGB11] and subsequent works: Assume some parameterization of π, and optimize for a convex surrogate of $F_n(\pi)$

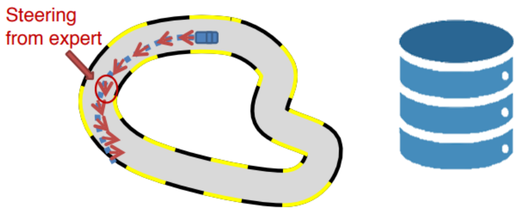
Limitations:

- In nonrealizable settings, convex surrogate may result in poor minimization of 0-1 error [BL+12]
- π may not have a parameterization amenable for optimization (e.g. decision trees)

[RGB11] Ross, Gordon, and Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning."

Computational Primitive: Offline classification oracle

Dataset $\boldsymbol{D}$



$$\operatorname*{argmin}_{\pi \in \mathcal{B}} \mathbb{E}_{s \sim D} \left[I(\pi(s) \neq \pi^{E}(s)) \right]$$

$\mathcal{O}$ can be implemented by standard off-the-shelf libraries

PyTorch

scikit learn

A fairly mild assumption: if offline learning cannot even be solved efficiently, we generally do not hope for efficient online learning

THE UNIVERSITY OF ARIZONA

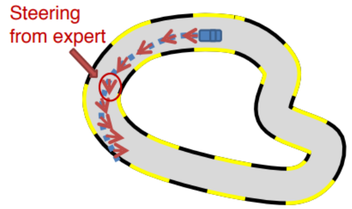
Our Contributions

We design computationally efficient algorithms that can perform regret minimization

$$\mathrm{Reg}_N := \sum_{n=1}^{N} F_n(\pi_n) - \min_{\pi \in \mathcal{B}} \sum_{n=1}^{N} F_n(\pi)$$

$$F_n(\pi_n) = \mathrm{E}_{s \sim d_{\pi_n}}[I(\pi(s) \neq \pi^E(s))]$$

for general policy classes $\mathcal{B}$



This problem is already well studied in the online learning literature?

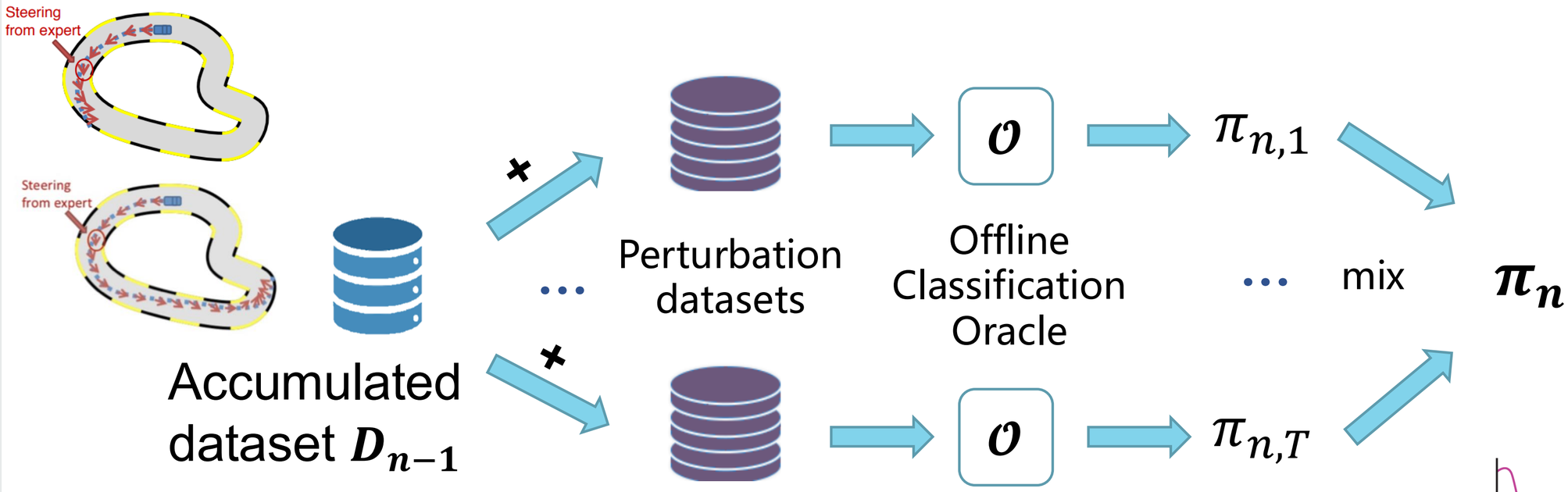
- Not exactly.. The function F_n can depend on the random realization of π_n
- Classical Algorithms, e.g., Follow the Perturbed Leader does not work here!

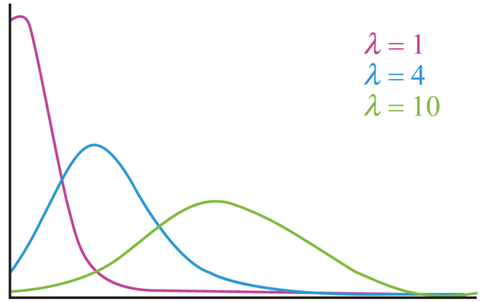
Accumulated dataset $\boldsymbol{D_{n-1}}$

Perturbation dataset

THE UNIVERSITY OF ARIZONA

Our Solution: Mixed Follow the Perturbed Leader (MFTPL)





The size of added perturbation datasets follows a Poisson distribution

[LZ24] Li, and Zhang. "Agnostic Interactive Imitation Learning: New Theory and Practical Algorithms."

Sample Complexity Comparison

$$J(\hat{\pi}) - J(\pi^E) \leq \text{ApproxErr} + \text{EstimErr}$$

Incomparable in general

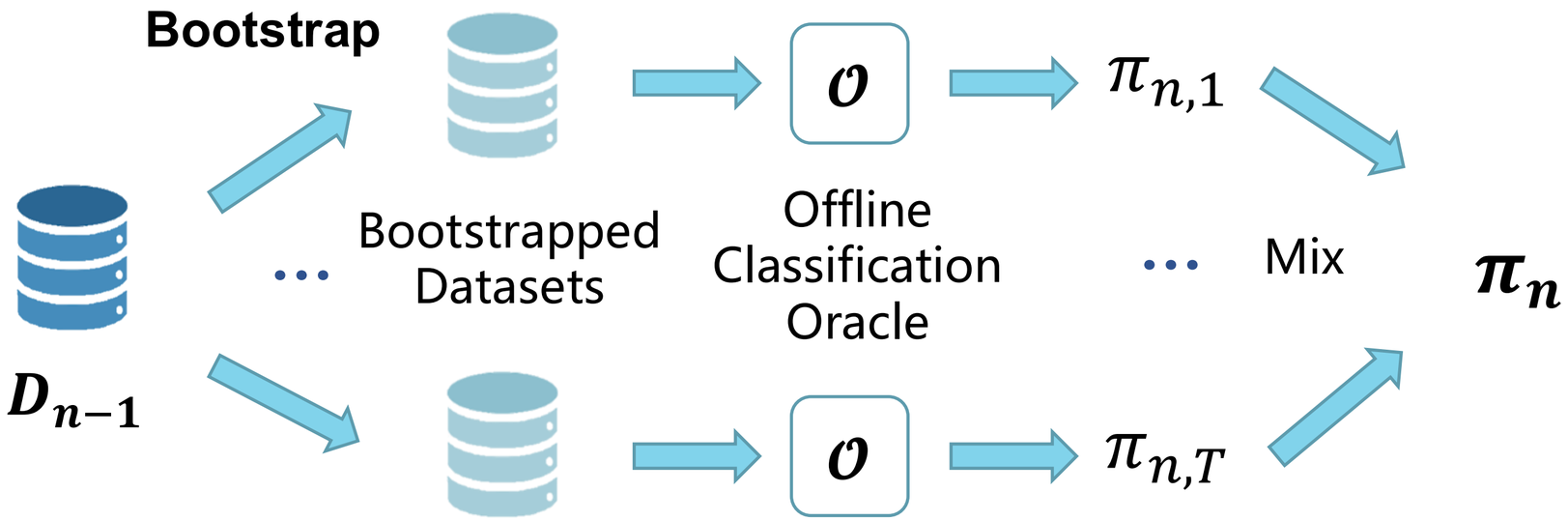
Related to Reg_N in online learning

Algorithm	For EstimErr to be $< \epsilon$
	#Expert's state-wise Annotations
MFTPL [LZ24]	$\mu^2 H^2/\epsilon^2$
Behavior Cloning	H^4/ϵ^2

THE UNIVERSITY OF ARIZONA

Practical Algorithm: Bootstrap-DAgger [LZ24]

Theoretically, the perturbation datasets need to come from an external source.
In practice, we replace it with a bootstrapping heuristic:



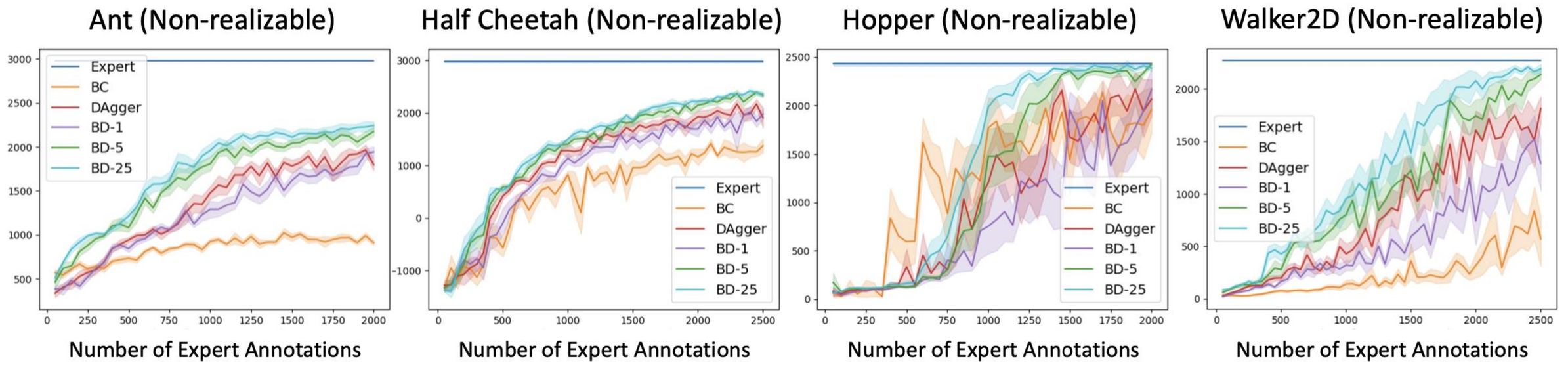
We choose $T = 1, 5, 25$ for evaluations

THE UNIVERSITY OF ARIZONA

[LZ24] Li, and Zhang. "Agnostic Interactive Imitation Learning: New Theory and Practical Algorithms."

Experiments with Bootstrap-DAgger

4 continuous control tasks from OpenAI Gym
Learner and expert use Multilayer Perceptron with different #hidden units



Bootstrap-DAgger (BD) outperforms DAgger and Behavior Cloning

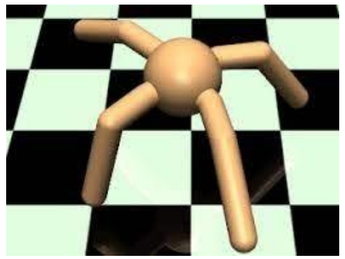
THE UNIVERSITY OF ARIZONA

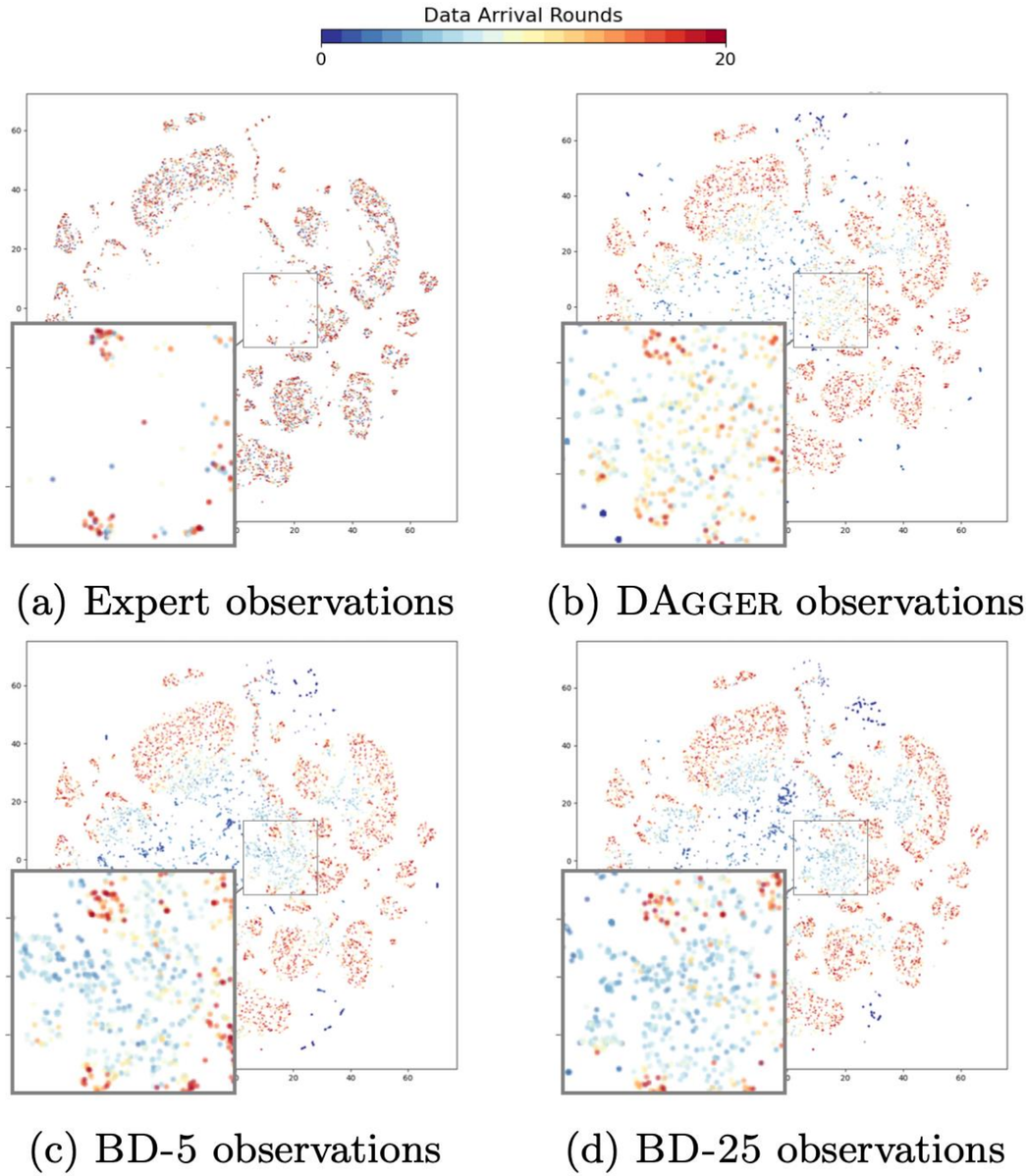
Bootstrap Dagger: Visualization of States Queried

t-SNE visualization of states queried by algorithms for Ant

Points in the zoomed-in area for BD-5 and BD-25 queried earlier than DAgger

⇒ more efficient exploration by BD in finding examples to learn to recover





(a) Expert observations (b) DAGGER observations

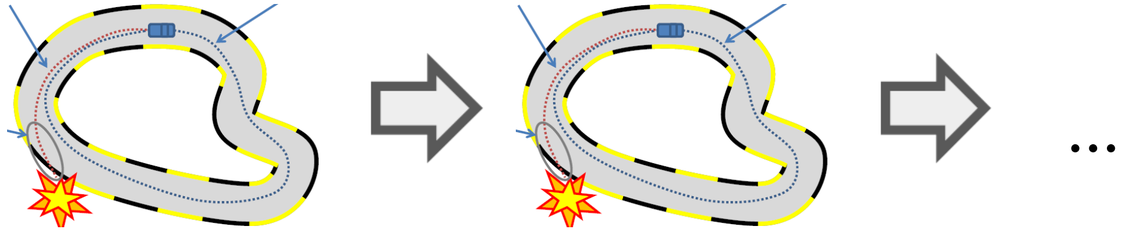
(c) BD-5 observations (d) BD-25 observations

Hybrid Imitation Learning: combining offline data and interactive queries

The Cold Start Problem of Interactive Imitation Learning

The Cold Start Problem [L20]:

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



...

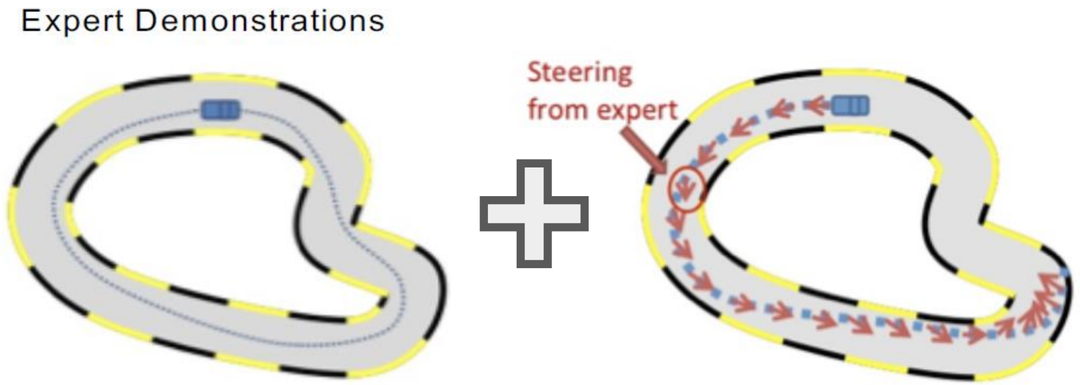
THE UNIVERSITY OF ARIZONA

[L20] Lin, et al. "Autoregressive knowledge distillation through imitation learning."

Combining Offline Data and Interactive Queries

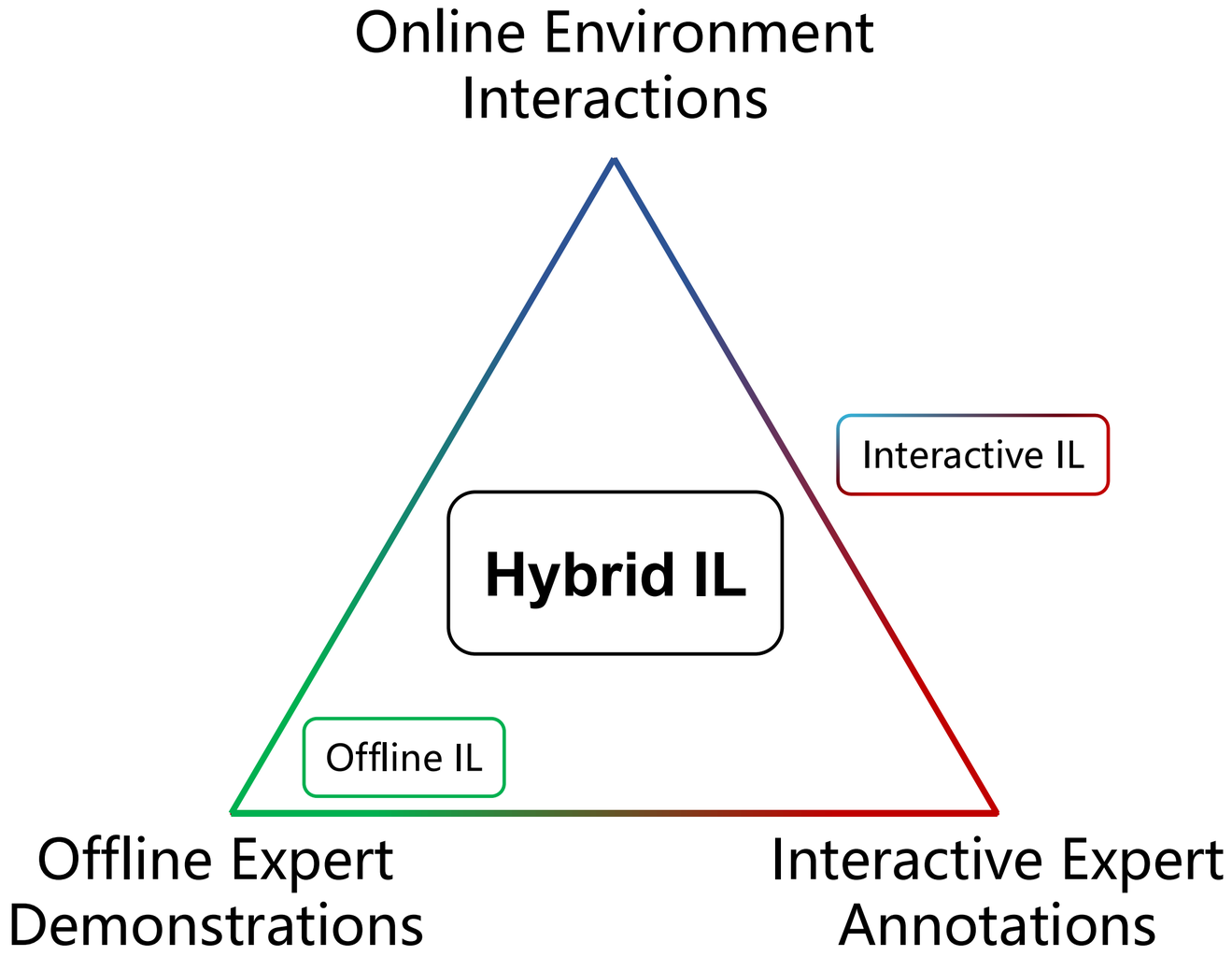
Many interactive IL applications do not start from scratch; they warm-start with offline demonstrations [B18, C22]

We propose to study Hybrid Imitation Learning: combining offline demonstrations and interactive expert annotations.



[B18] Bi, et al. "Navigation by imitation in a pedestrian-rich environment."
[C22] Jiaxun Cui, et al. "Coopernaut: End-to-end driving with cooperative perception for networked vehicles."
[LZ25] Li, and Zhang. "Interactive and Hybrid Imitation Learning: Provably Beating Behavior Cloning."

Hybrid Imitation Learning



THE UNIVERSITY OF ARIZONA

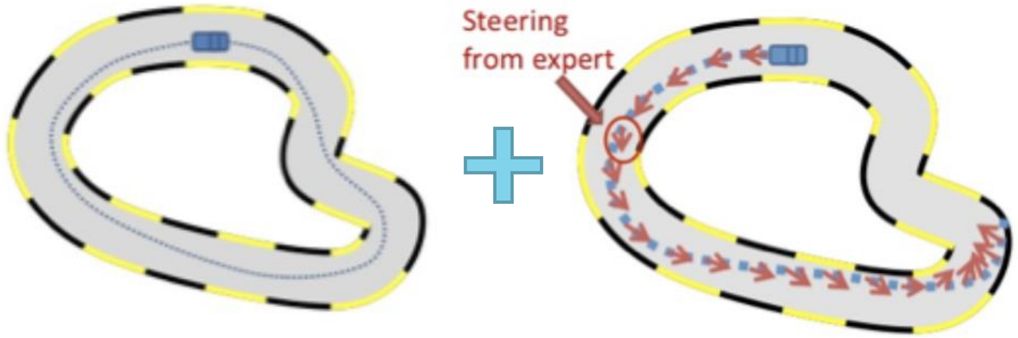
Hybrid Imitation Learning

Goal: Learn a policy that has performance competitive with the expert, with small total cost $\boldsymbol{HN_{\text{off}} + N_{\text{int}}}$.

A basic model: each state-wise annotation is charged a unit cost

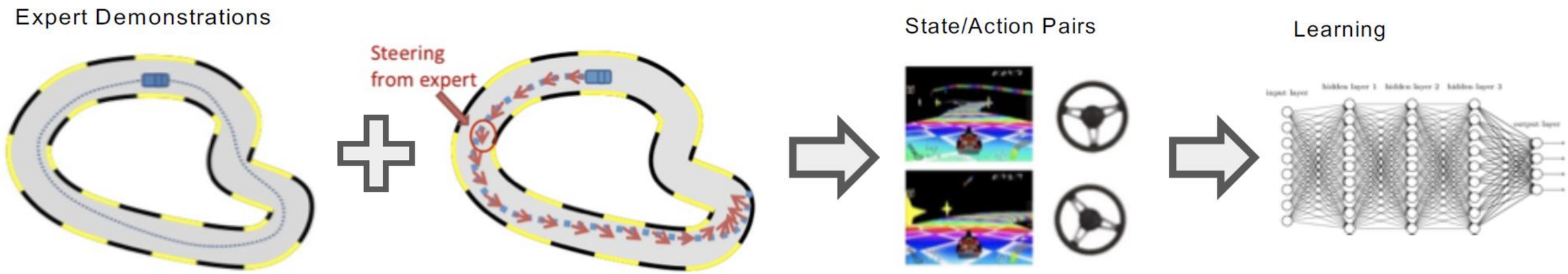
$\boldsymbol{N_{\text{off}}}$: number offline expert trajectories

$\boldsymbol{N_{\text{int}}}$: number of state-wise interactive queries to the expert



THE UNIVERSITY OF ARIZONA

Warm-Stagger: Warm-starting Stagger using offline data



Fall-back Guarantee [LZ25]: Under realizable expert, Warm-Stagger returns policy with cost worse than the expert for at most

$$\min\left(H^2\frac{\log(\boldsymbol{B})}{\boldsymbol{N_{\mathbf{off}}}}, \mu H\frac{\log(\boldsymbol{B})}{\boldsymbol{N_{\mathbf{int}}}}\right).$$

H**-factor worse than Offline Only**

Interactive Only

[LZ25] Li, and Zhang. "Interactive and Hybrid Imitation Learning: Provably Beating Behavior Cloning."

The Benefit of Hybrid Imitation Learning

Theorem For any large enough S, there is an MDP with S states and an expert policy, such that:

- With $N_{\text{off}} = O(S)$ expert trajectories, Behavior Cloning learns a policy no better than random guessing Total cost: $O(HS)$
- With $N_{\text{int}} = O(HS)$ state-wise interactions, Stagger learns a policy no better than random guessing Total cost: $O(HS)$
- With $N_{\text{off}} = O(S/H)$ and $N_{\text{on}} = O(1)$, Warm-Stagger learns a policy on par with the expert Total cost: $O(S)$

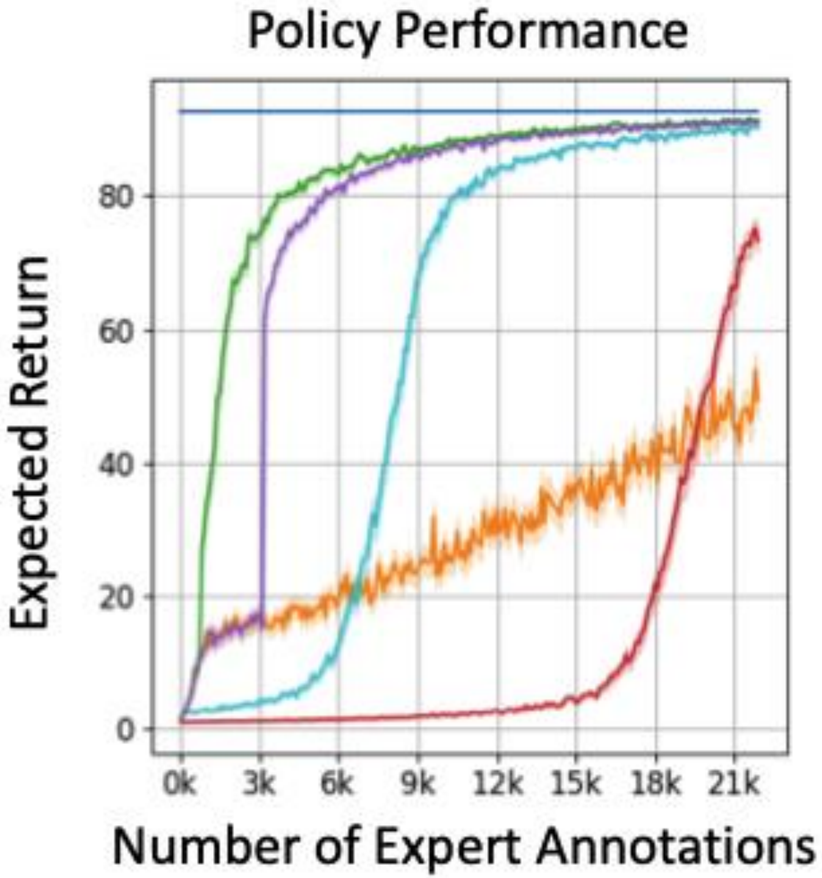
Implication: Warm-Stagger learns more cost-efficiently than using either source alone!

Experimental Validation

Toy MDP in the theorem

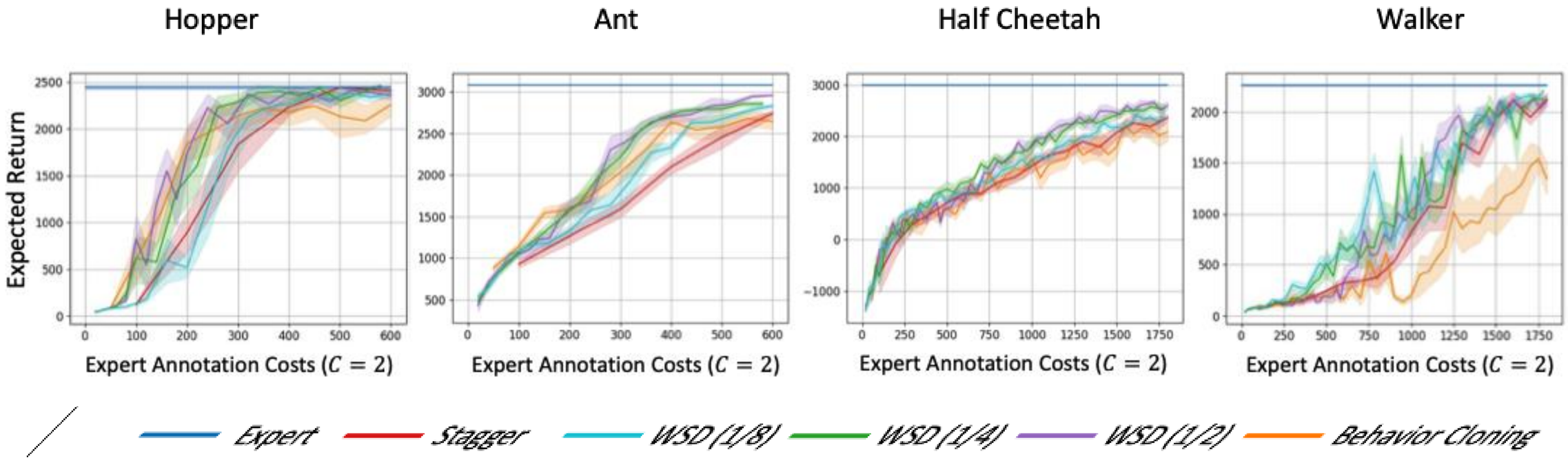
Stagger and Behavior Cloning
does not yet converge to optimal

Warm-Stagger with different number of offline expert demonstrations (200, 800, 3200) converge quickly



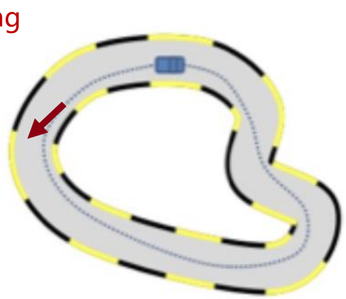
Experimental Validation

When each interactive state annotation is 2 times more expensive than offline state annotation, Warm-Stagger (WSD) has sometimes better cost-efficiency than both Stagger and Behavior Cloning



THE UNIVERSITY OF ARIZONA

Conclusions and Open Problems



We show the benefit of interaction in imitation learning when the sample costs are measured using state-wise annotation, and extend it to nonrealizable and hybrid settings

Open problems:

- More realistic cost models – e.g., demonstrating on 10 consecutive states should be much easier than annotating 10 totally separate frames
- Understanding imitation learning beyond the regret minimization framework?
- Are there general properties of MDP that allows Hybrid IL to beat learning using either data source alone?

THE UNIVERSITY OF ARIZONA

Thank You!