Efficient Imitation Under Misspecification

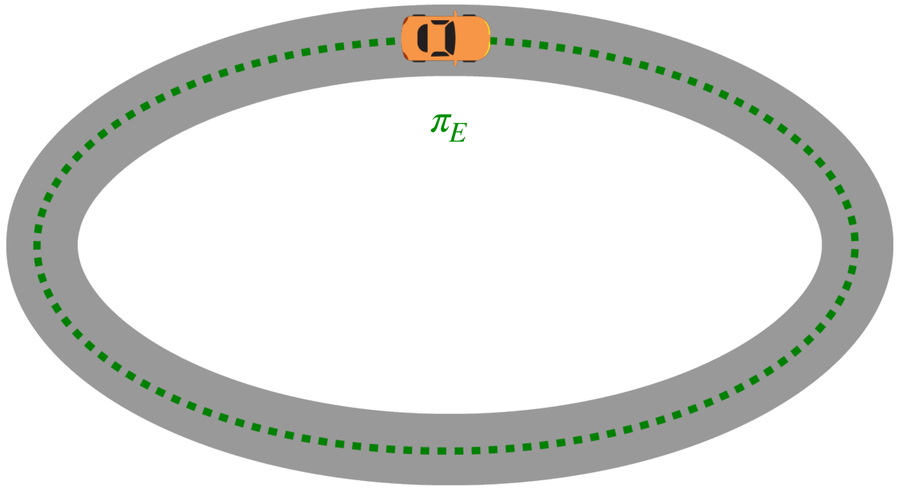
Nicolas Espinosa Dice

Joint work with Sanjiban Choudhury, Wen Sun, and Gokul Swamy

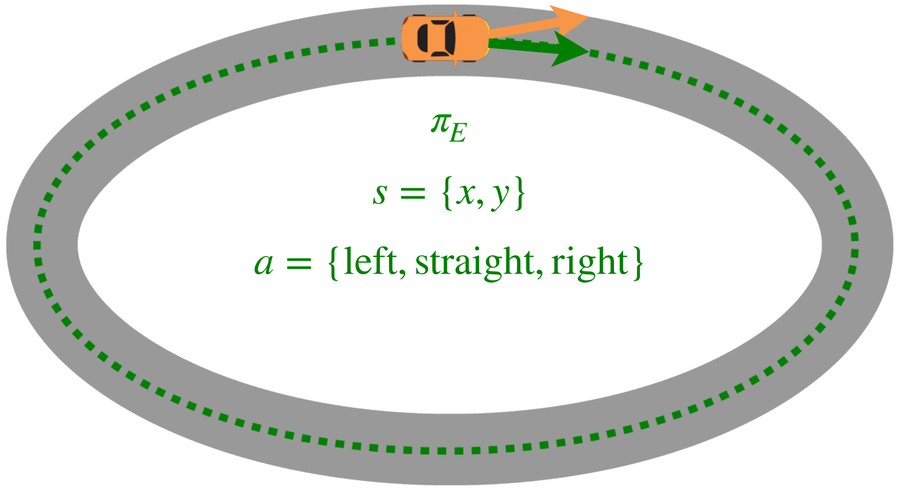



Cornell Bowers C·IS
College of Computing and Information Science

How do we learn from expert demonstrations?



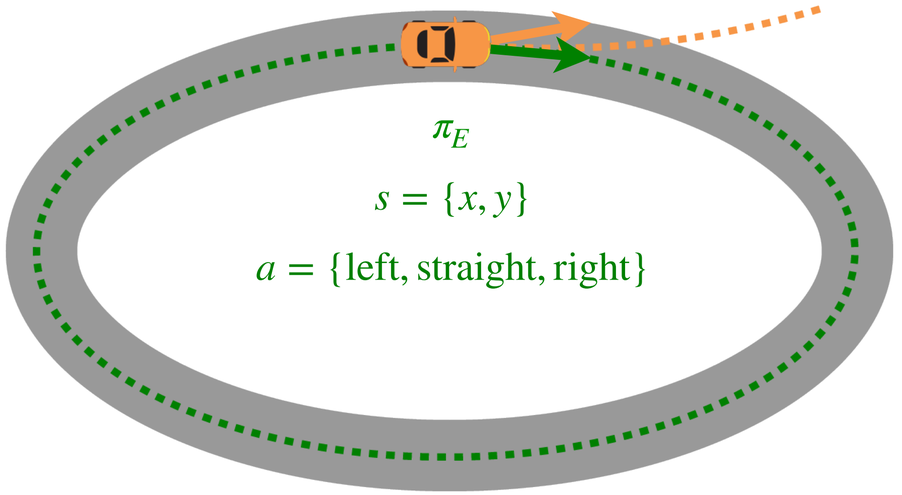
How do we learn from expert demonstrations?



$\{s_1, \ldots, s_n\} \mapsto \{a_1, \ldots, a_n\}$

Behavioral Cloning (BC) → Stymied by *compounding errors*

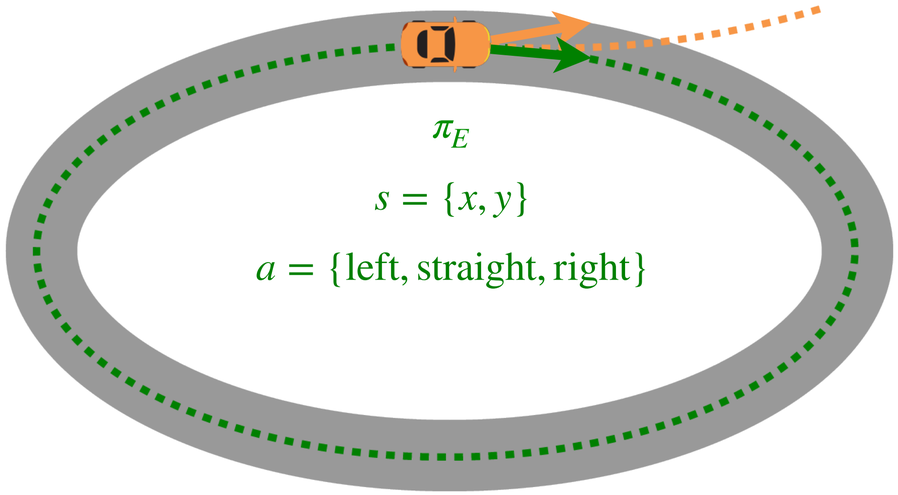
How do we learn from expert demonstrations?



$\{s_1, \ldots, s_n\} \mapsto \{a_1, \ldots, a_n\}$

Behavioral Cloning (BC) → Stymied by *compounding errors*

How do we learn from expert demonstrations?



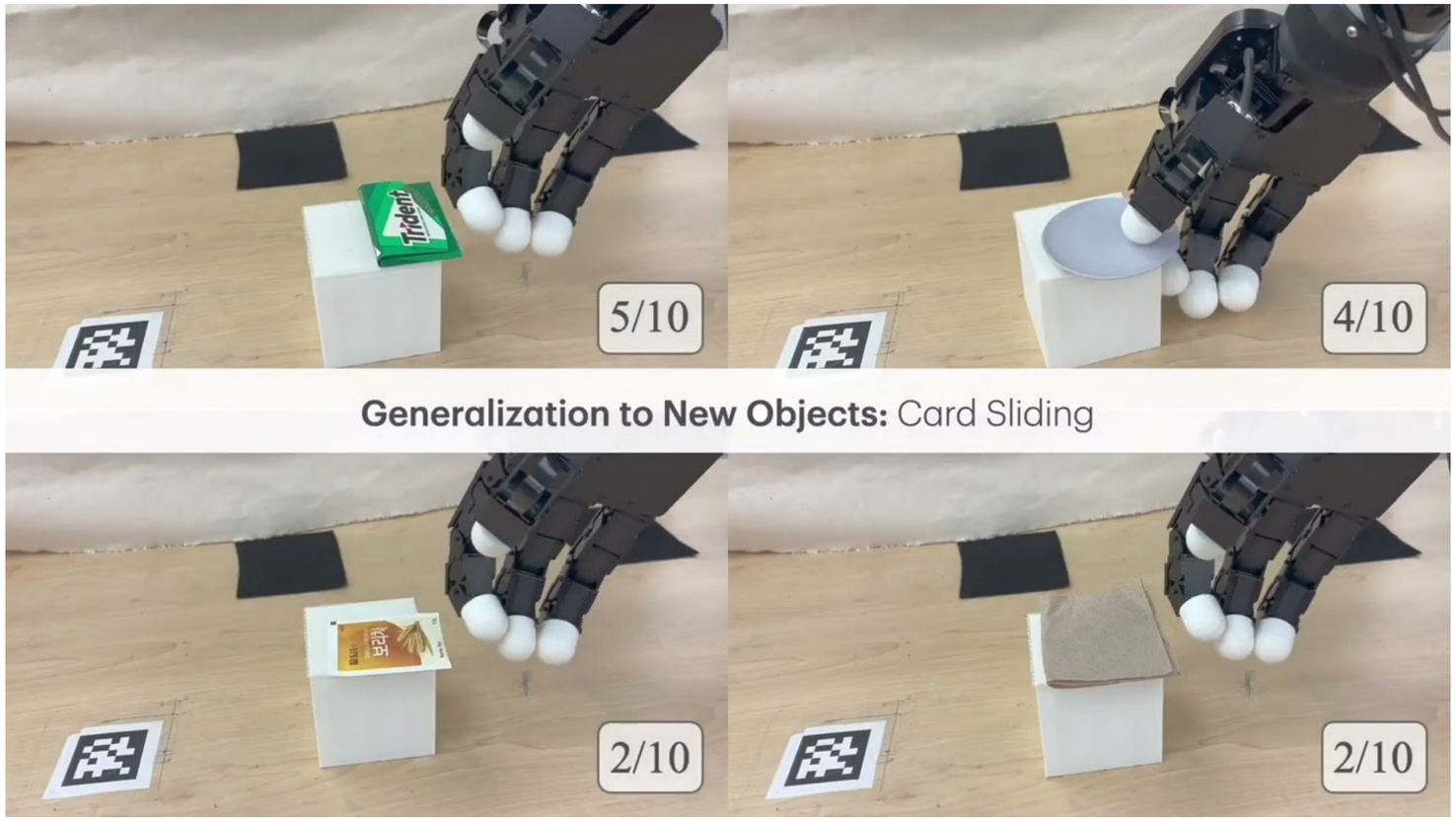
$\{s_1, \ldots, s_n\} \mapsto \{a_1, \ldots, a_n\}$

Behavioral Cloning (BC) ⟶ Stymied by *compounding errors*

Exacerbated by *misspecification*:
When the learner cannot perfectly imitate the expert
$(\pi_E \notin \Pi)$

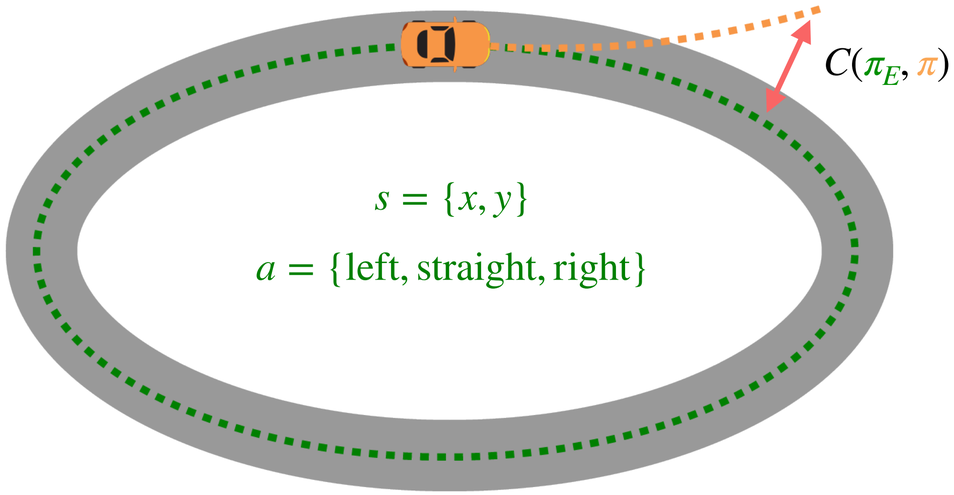
In practice, misspecification is common

Embodiment mismatch



(Guzey *et al.*, 2024)

How do we learn to ***recover from mistakes****?*

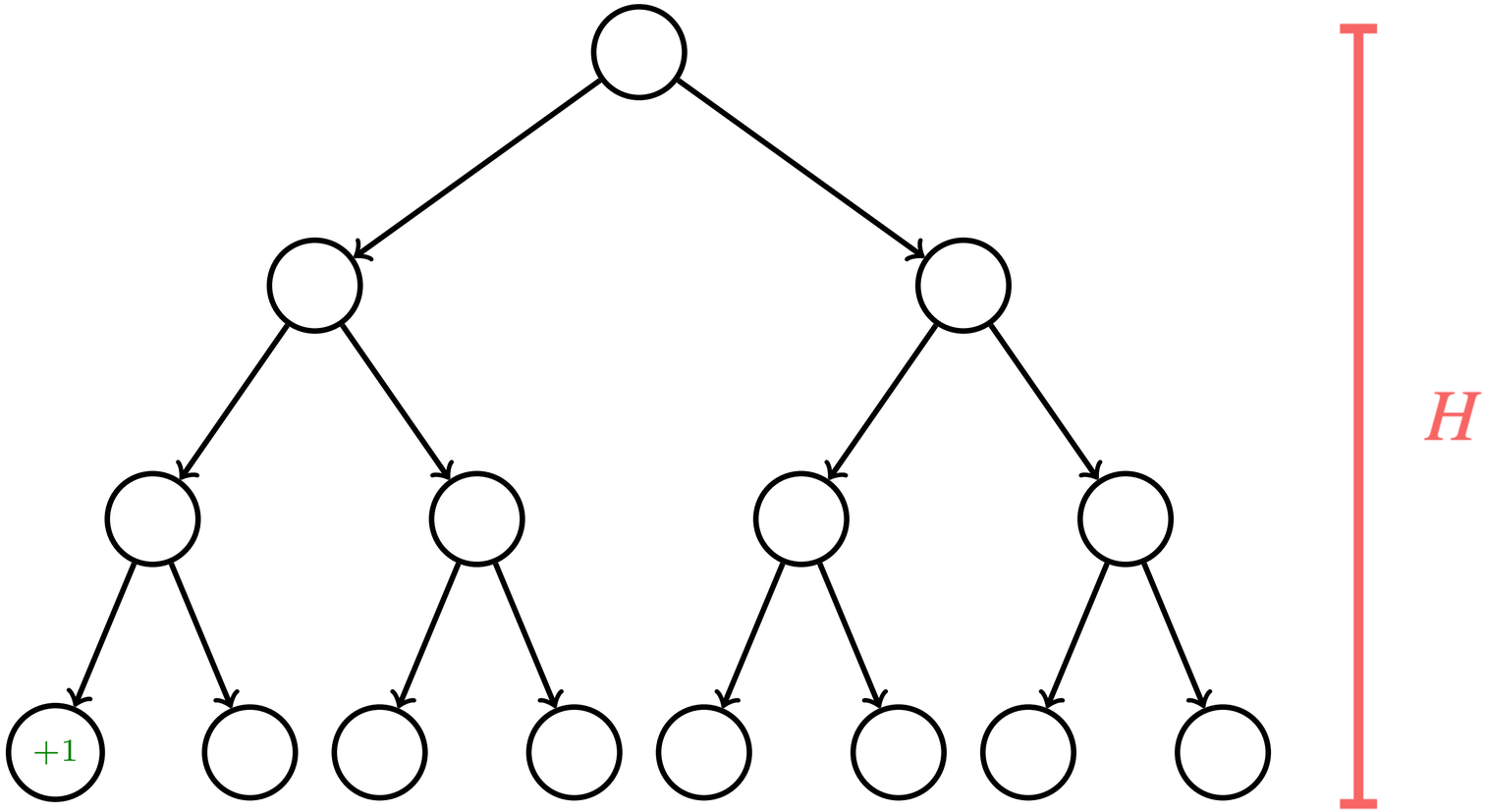


$\{s_1, \ldots, s_n\} \mapsto \{a_1, \ldots, a_n\}$

Behavioral Cloning (BC) → Stymied by *compounding errors*

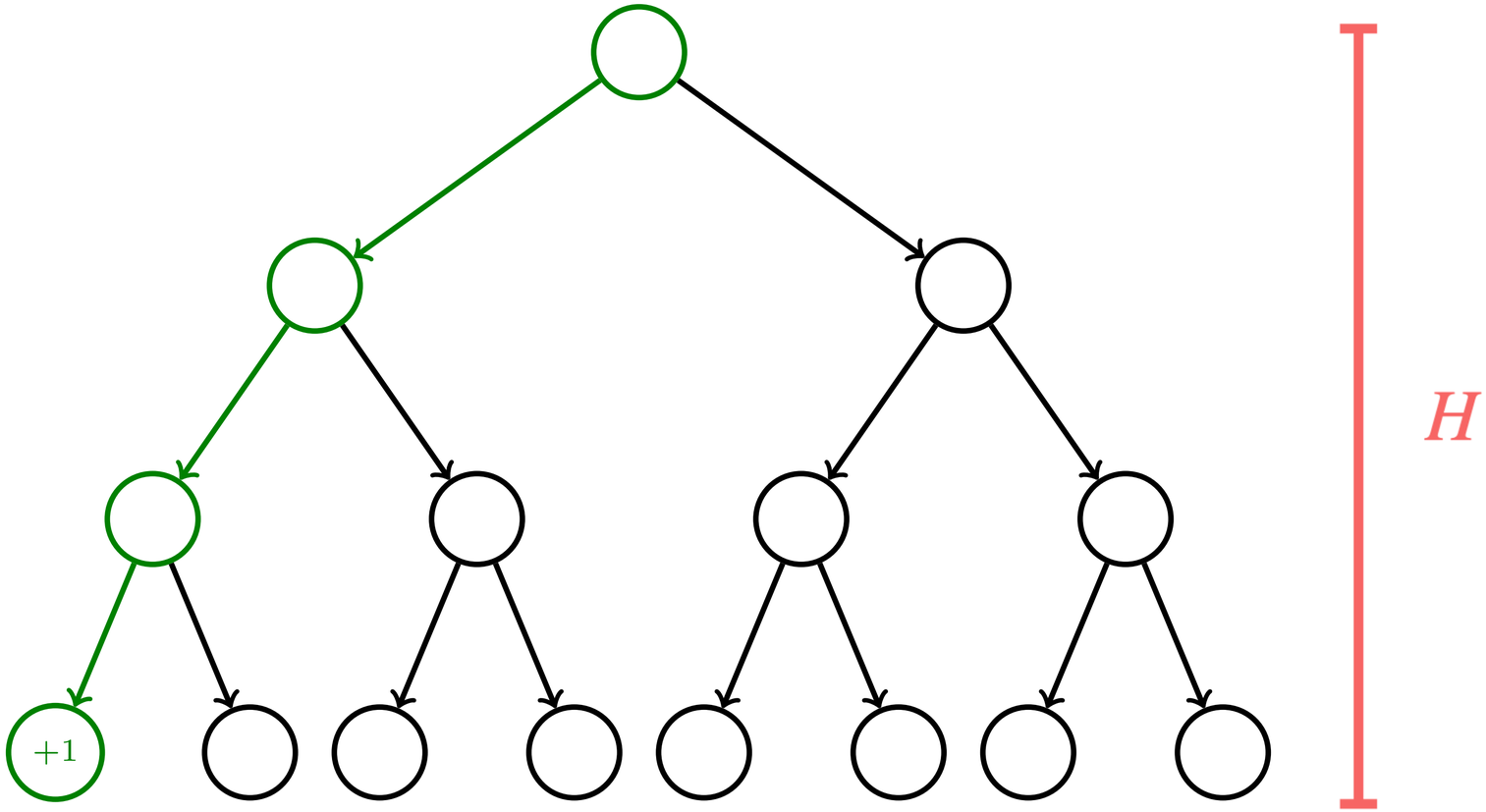
Inverse Reinforcement Learning (IRL)

Challenge: IRL is computationally inefficient



Challenge: IRL is computationally inefficient

Solution: Local search is sufficient in ***well-specified setting*** ($\pi_E \in \Pi$)



(Swamy et al. 2023)

Key Contribution 1

Local search is sufficient for learning to recover from mistakes

in the misspecified setting

Key Contribution 1

Local search is sufficient for learning to recover from mistakes in the misspecified setting

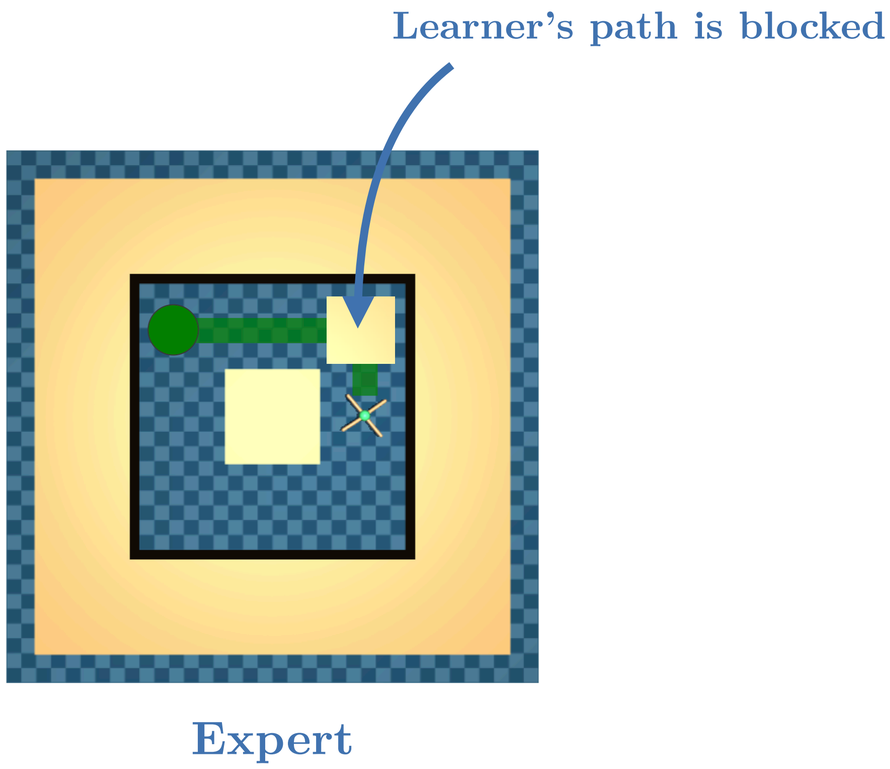
We generalize ***policy completeness*** *to the IL setting*

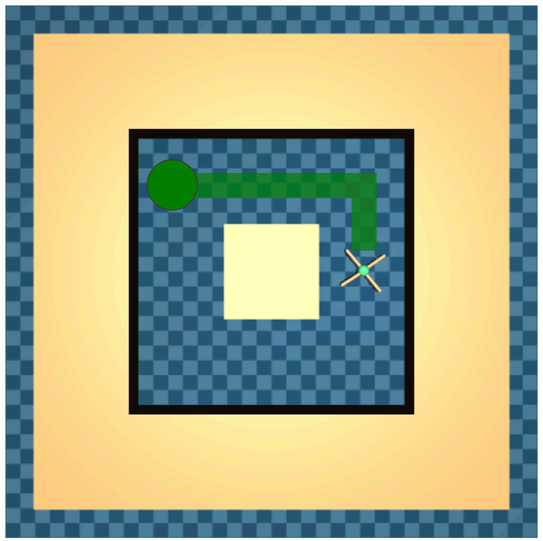
(1) **Reward-agnostic policy completeness**

Theorem 1 (Informal): *Under (1), efficient IRL avoids compounding errors in misspecified settings*

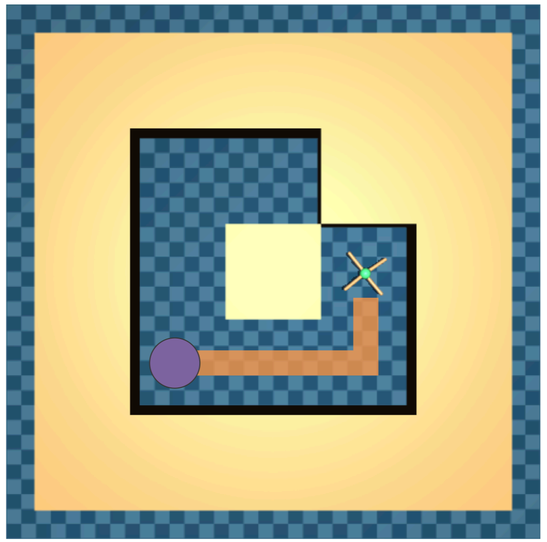
Key Contribution 2

*We analyze where **local search** should be performed*

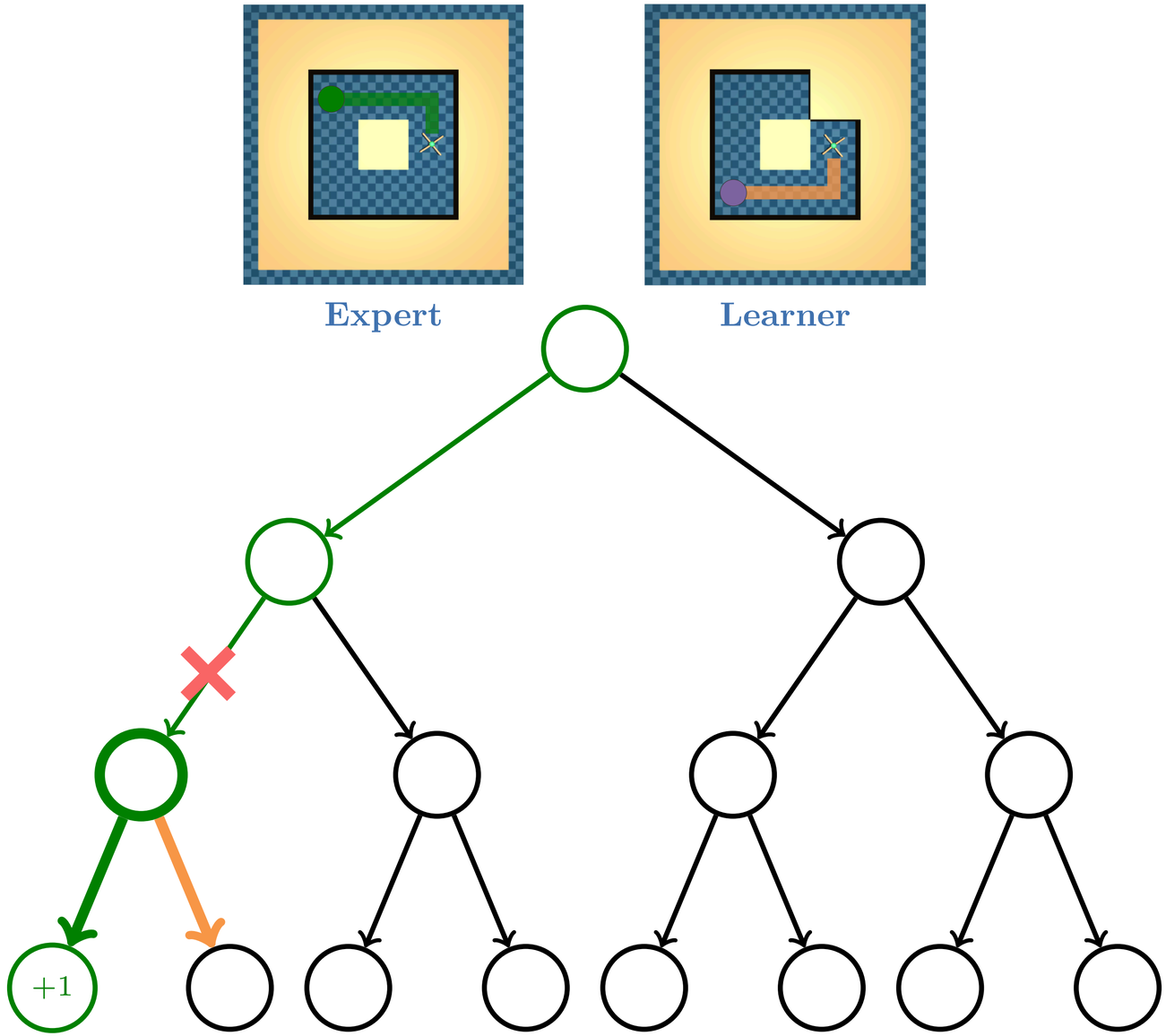

Learner's path is blocked
Expert



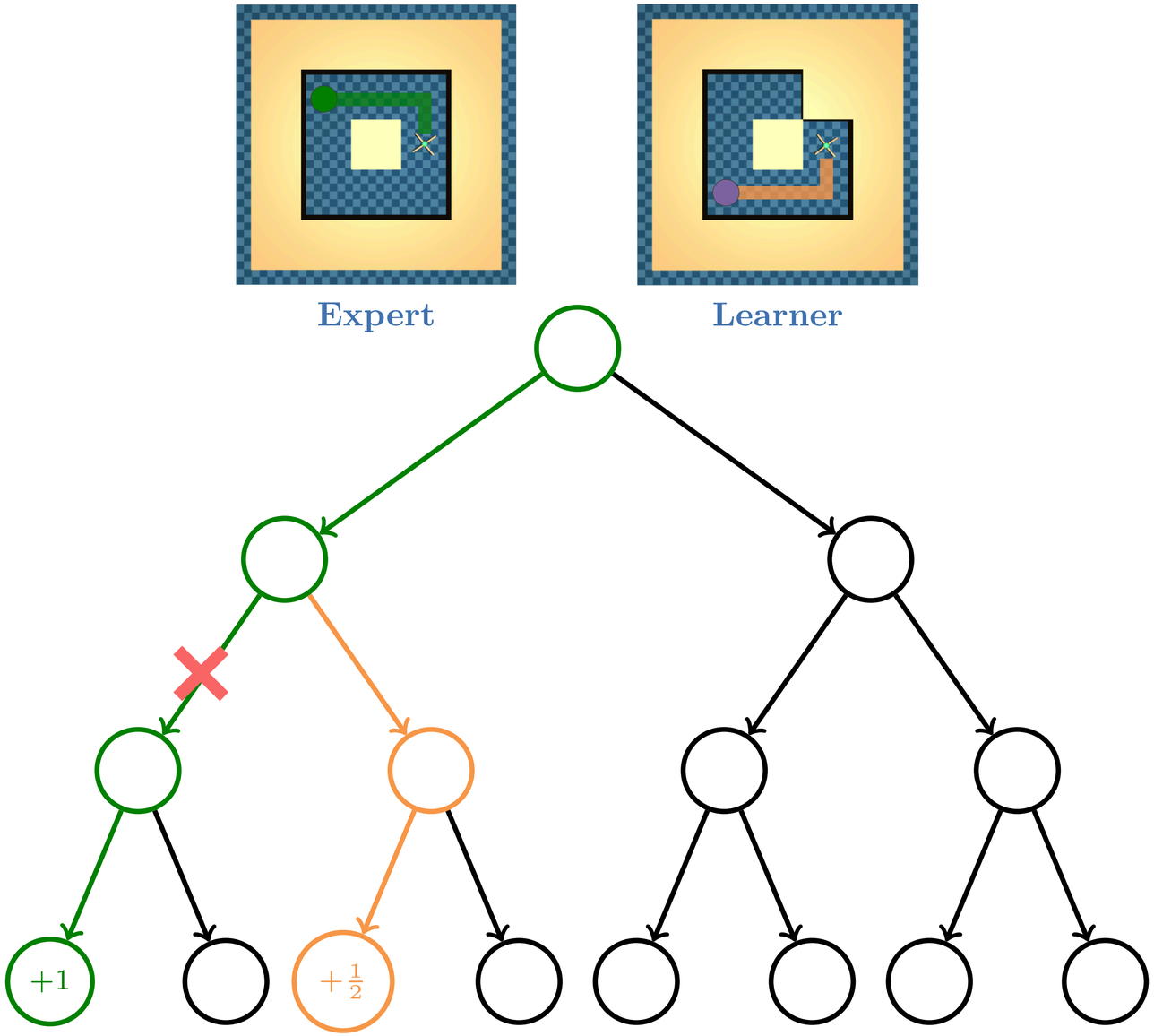
Expert



Learner



$\pi_E \notin \Pi$: optimal *(unrealizable)* policy

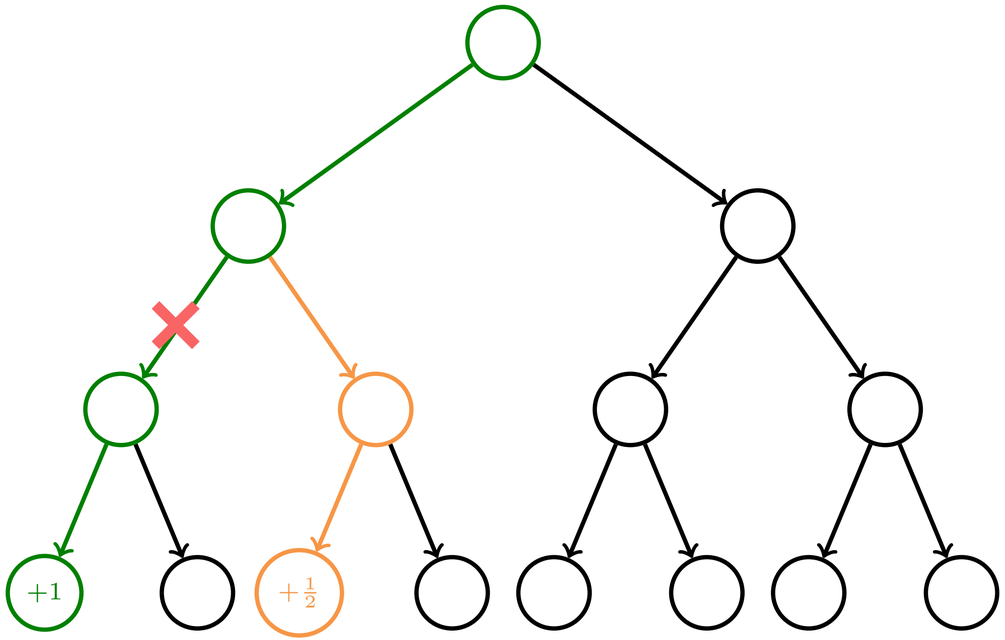


$\pi_E \notin \Pi$: optimal *(unrealizable)* policy

$\pi^\star \in \Pi$: optimal *realizable* policy

Key Contribution 2

*We analyze where **local search** should be performed*

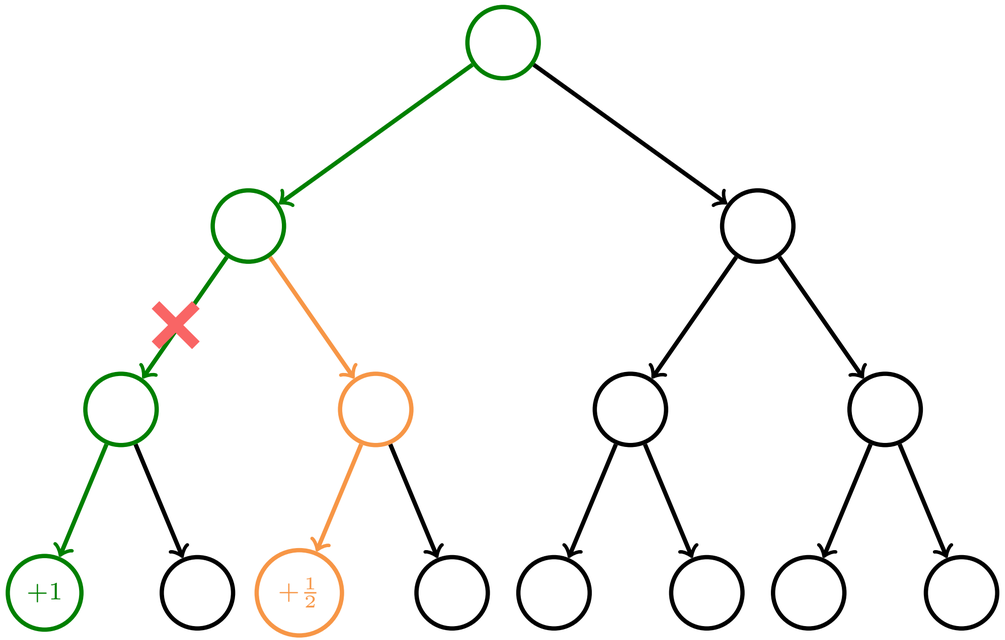


$\pi_E \notin \Pi$: optimal *(unrealizable)* policy

$\pi^\star \in \Pi$: optimal *realizable* policy

Key Contribution 2

*We analyze where **local search** should be performed*



$\pi_E \notin \Pi$: optimal *(unrealizable)* policy

$\pi^\star \in \Pi$: optimal *realizable* policy

Reset distribution

Specifies the policies

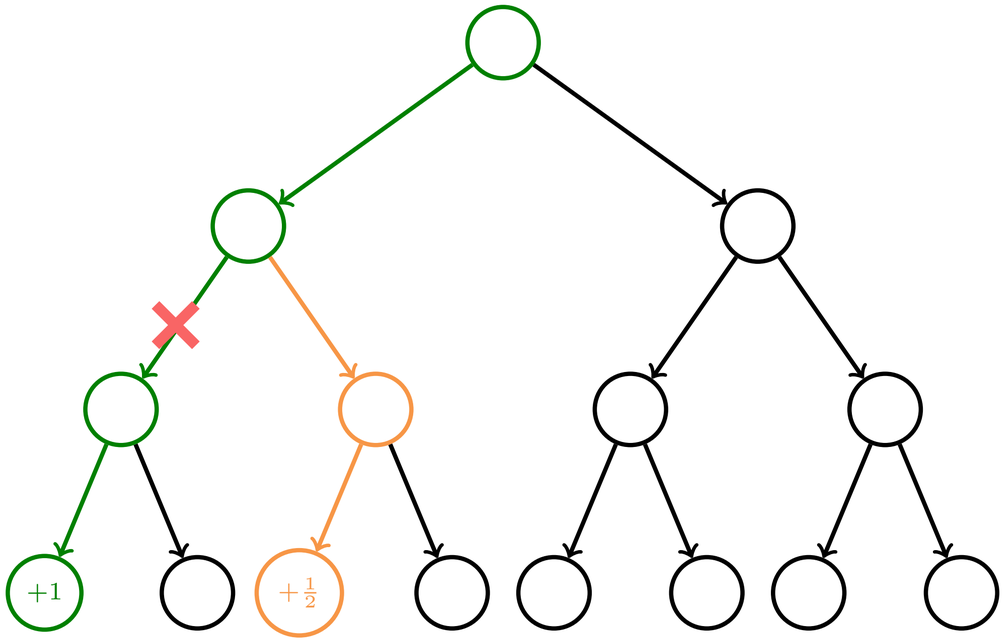
learner "competes" against

Key Contribution 2

We analyze where ***local search*** *should be performed*

↓

Reset distribution should cover $\pi^\star$



$\pi_E \notin \Pi$: optimal *(unrealizable)* policy

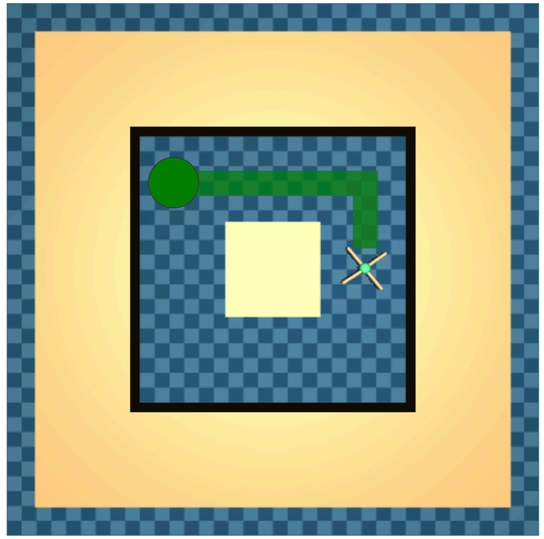
$\pi^\star \in \Pi$: optimal *realizable* policy

Reset distribution

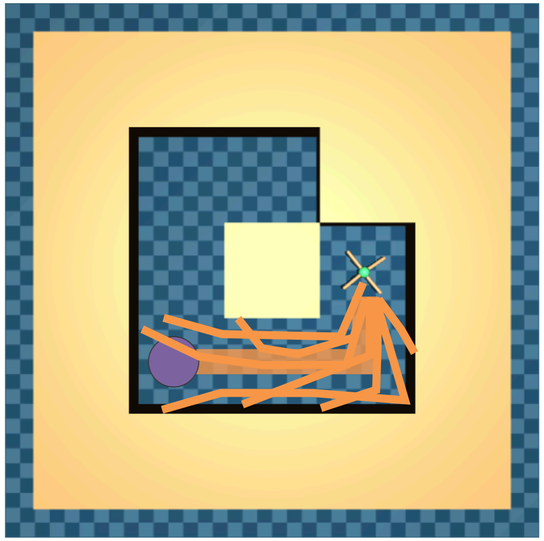
Specifies the policies learner "competes" against

How do use $\pi^\star$ as the reset distribution?

Solution: Use offline data (e.g. sub-optimal demos, self-play data, internet data)



Expert Policy

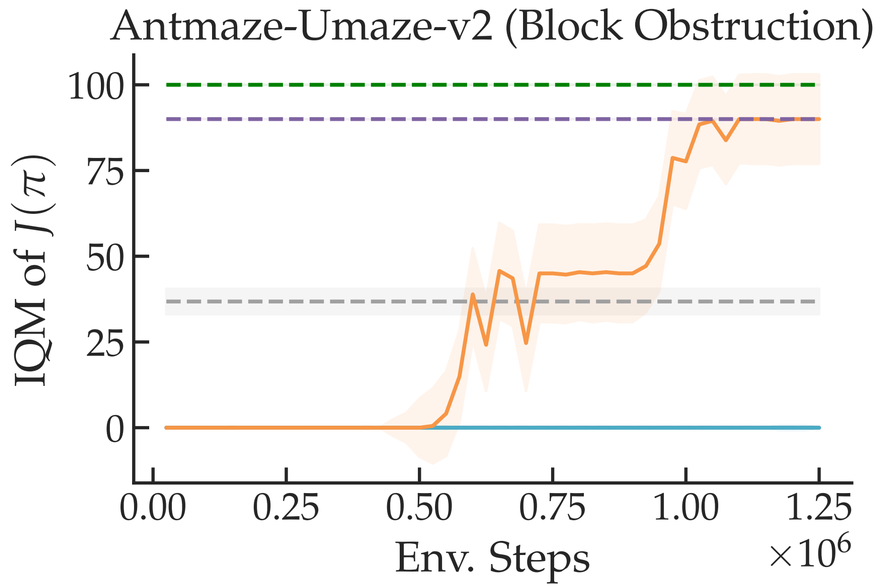


Realizable Policy

Augment our reset distribution with offline data

✔ Resets to *realizable offline data* outperform expert resets

In misspecified settings



BC(π_E) BC($\pi_E + \pi_B$) MM FILTER, $\alpha = 1.0$ GUITAR, $\alpha = 1.0$ $\pi^\star$ π_E

Resets to starting state

Resets to expert data

Resets to offline data

Efficient Imitation Under Misspecification

Nicolas Espinosa Dice

Joint work with Sanjiban Choudhury, Wen Sun, and Gokul Swamy




Cornell Bowers C·IS
College of Computing and Information Science