On the value of Interaction and Function approximation in Imitation Learning

Nived Rajaraman, *Yanjun Han, Lin F. Yang, Jingbo Liu, Jiantao Jiao, Kannan Ramchandran*

NeurIPS 2021

Challenges in RL

Rewards for practical RL problems are often hard to specify.

$$r(b_z^{(1)}, s^P, s^{B1}, s^{B2}) = \begin{cases} 1 & \text{if stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$r(b_z^{(1)}, s^P, s^{B1}, s^{B2}) = \begin{cases} 1 & \text{if stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0.25 & \text{if } \neg\text{stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \wedge \text{grasp}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$r(b_z^{(1)}, s^P, s^{B1}, s^{B2}) = \begin{cases} 1 & \text{if stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0.25 & \text{if } \neg\text{stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \wedge \text{grasp}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0.125 & \text{if } \neg(\text{stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \vee \text{grasp}(b_z^{(1)}, s^P, s^{B1}, s^{B2})) \wedge \text{reach}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

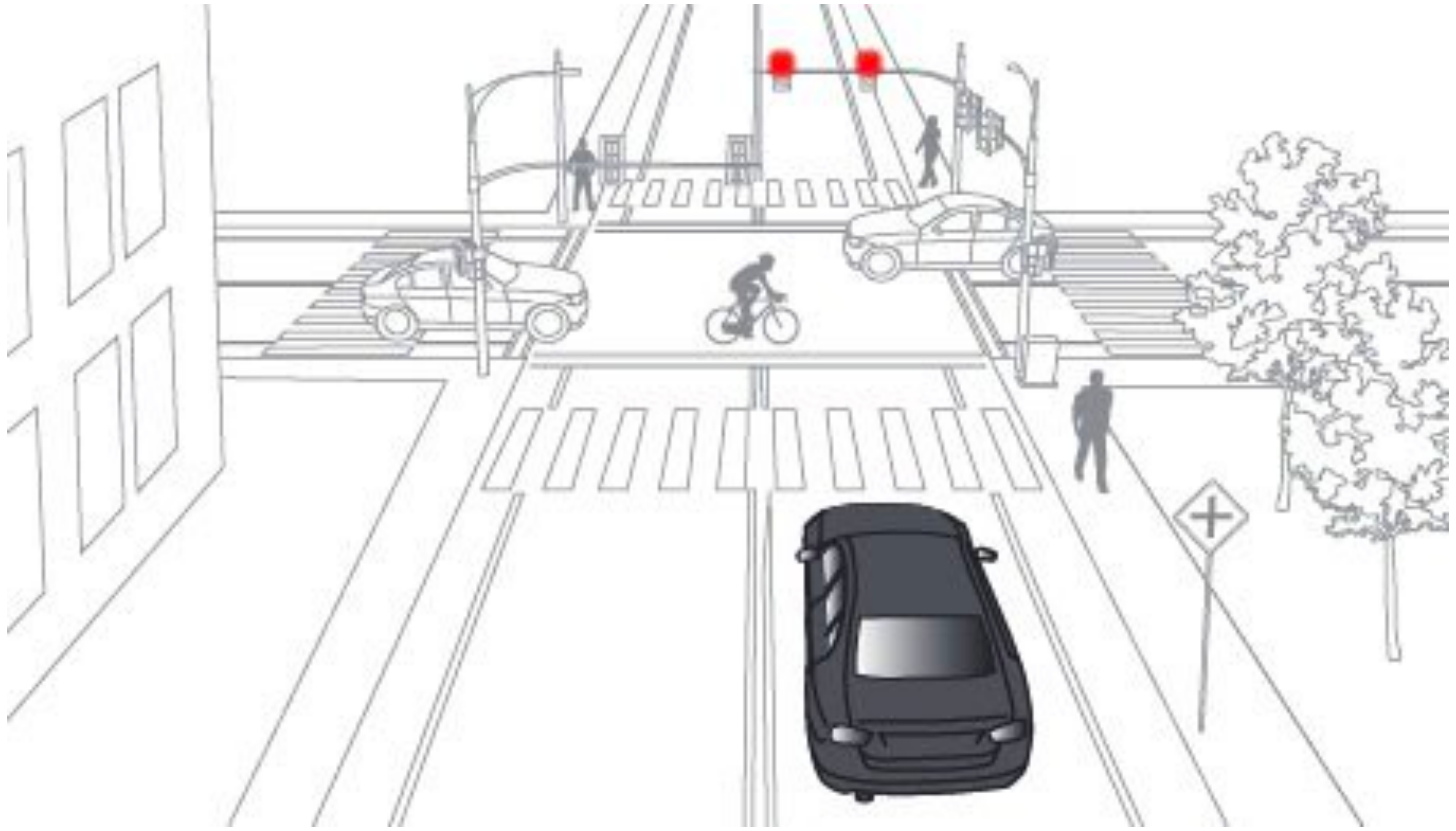
$$r(b_z^{(1)}, s^P, s^{B1}, s^{B2}) = \begin{cases} 1 & \text{if stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0.25 + 0.25 r_{S2}(s^{B1}, s^P) & \text{if } \neg\text{stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \wedge \text{grasp}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0.125 & \text{if } \neg(\text{stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \vee \text{grasp}(b_z^{(1)}, s^P, s^{B1}, s^{B2})) \wedge \text{reach}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0 + 0.125 r_{S1}(s^{B1}, s^P) & \text{otherwise} \end{cases}$$

Popov et al. 2017

Reward design must be consistent with counterfactual questions:
"What would an expert have done?"

Need to correctly balance **interpretability** and **sparsity.**

Imitation learning over reward engineering



Expert demonstrations

Learner

"Learning from demonstrations in the absence of reward feedback"

Image source: Gettyimages

Motivation

What are the theoretical limits of Imitation Learning (i) with interaction and (ii) in the presence of function approximation?

Notation:

$J(\pi)$: Expected total reward of policy π in an episode of length H.

Learner $\widehat{\pi}$ tries to minimize **Suboptimality** $\triangleq \mathbb{E}\left[J(\pi^*) - J(\widehat{\pi})\right]$, π^* is expert's policy

- Difference in expected reward of the expert and the learner policy.

Theoretical understanding of IL: Prior work

No interaction: *Learner is only provided a dataset of N expert demonstrations;*
Cannot interact with the MDP

Theorem [RYJR20]
In the ***no-interaction*** *and* ***tabular*** *setting,* ***Behavior Cloning*** *achieves,*

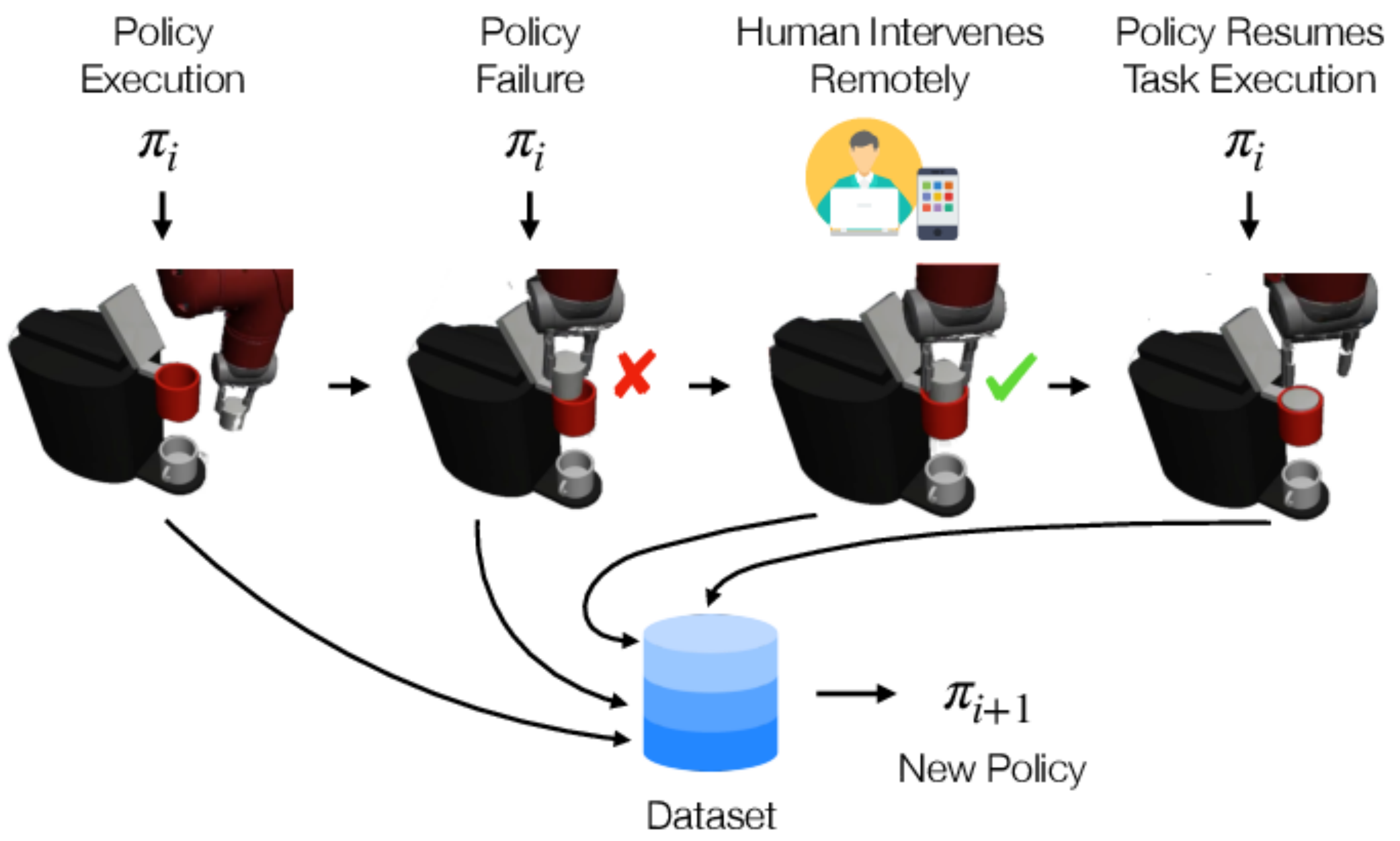
$$\text{Suboptimality} \lesssim \frac{SH^2 \log(N)}{N}$$

Best achievable (up to log-factors) by any algorithm.

Going beyond the no-interaction setting

Interactive expert: *Learner can interact with the environment N times and* ***query the expert policy at visited states***

Setting is closely related to human-in-the-loop RL



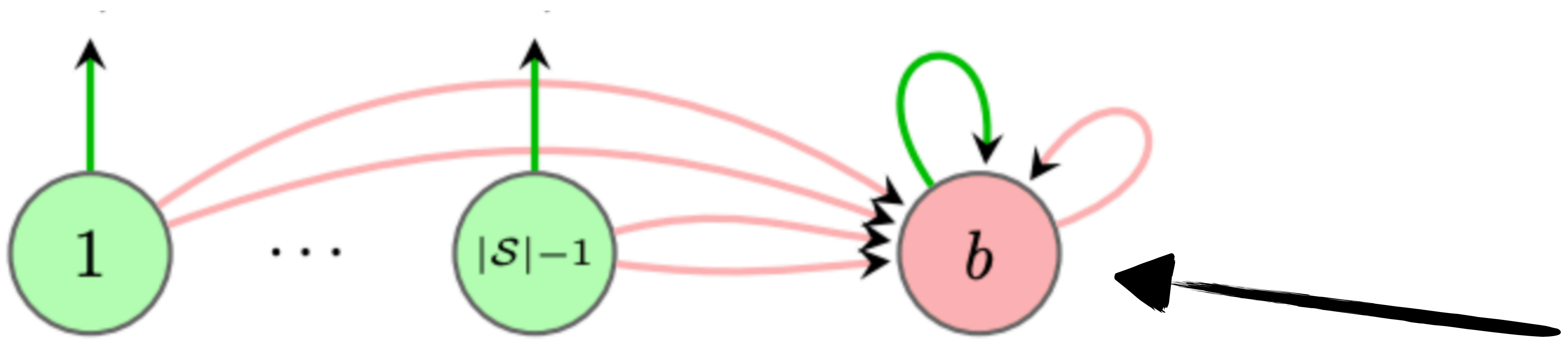
Mandlekar et al. 2020

IL with an interactive expert

Is it possible to improve the suboptimality of behavior cloning if the expert is **interactive**?

In the worst case, **no**.

For all algorithms even with an interactive expert, in the worst case,
Suboptimality $\gtrsim SH^2/N$ [RYJR20]



All learners get stuck at bad state

Hard instance: Reset cliff MDP

IL with an interactive expert

Is it possible to improve the suboptimality of behavior cloning if the expert is **interactive**?

μ-recoverability assumption [RB11]: For any state s, action a',

$$max_a Q_t^*(s, a) - Q_t^*(s, a') \leq \mu$$

Interpretation: *Expert knows how to "recover" after making a mistake at some time t and pays an expected cost of at most μ.*

IL with an interactive expert

Is it possible to improve the suboptimality of behavior cloning if the expert is **interactive**?

Theorem 1 [RHYLJR21]

*Under μ-recoverability, in the **interactive** and **tabular** setting, **DAGGER** (FTRL) achieves,*

$$\text{Suboptimality} \lesssim \frac{\mu S H \log(N)}{N}$$

Best achievable (up to log-factors) by any algorithm.

IL with function approximation

How do approaches such as BC and Mimic-MD [RYJR20] perform in the presence of function approximation?

IL with linear function approximation

Linear expert: *For every state* s, *the deterministic expert plays an action*

$$\pi_t^*(s) \in \text{argmax}_a \langle \theta_t, \phi_t(s, a) \rangle$$

$\phi_t(s, a) \in \mathbb{R}^d$ *is a known representation of state-actions*

Interpretation: *Expert policy is realized by a linear multi-class classifier*

Linear expert with no MDP interaction

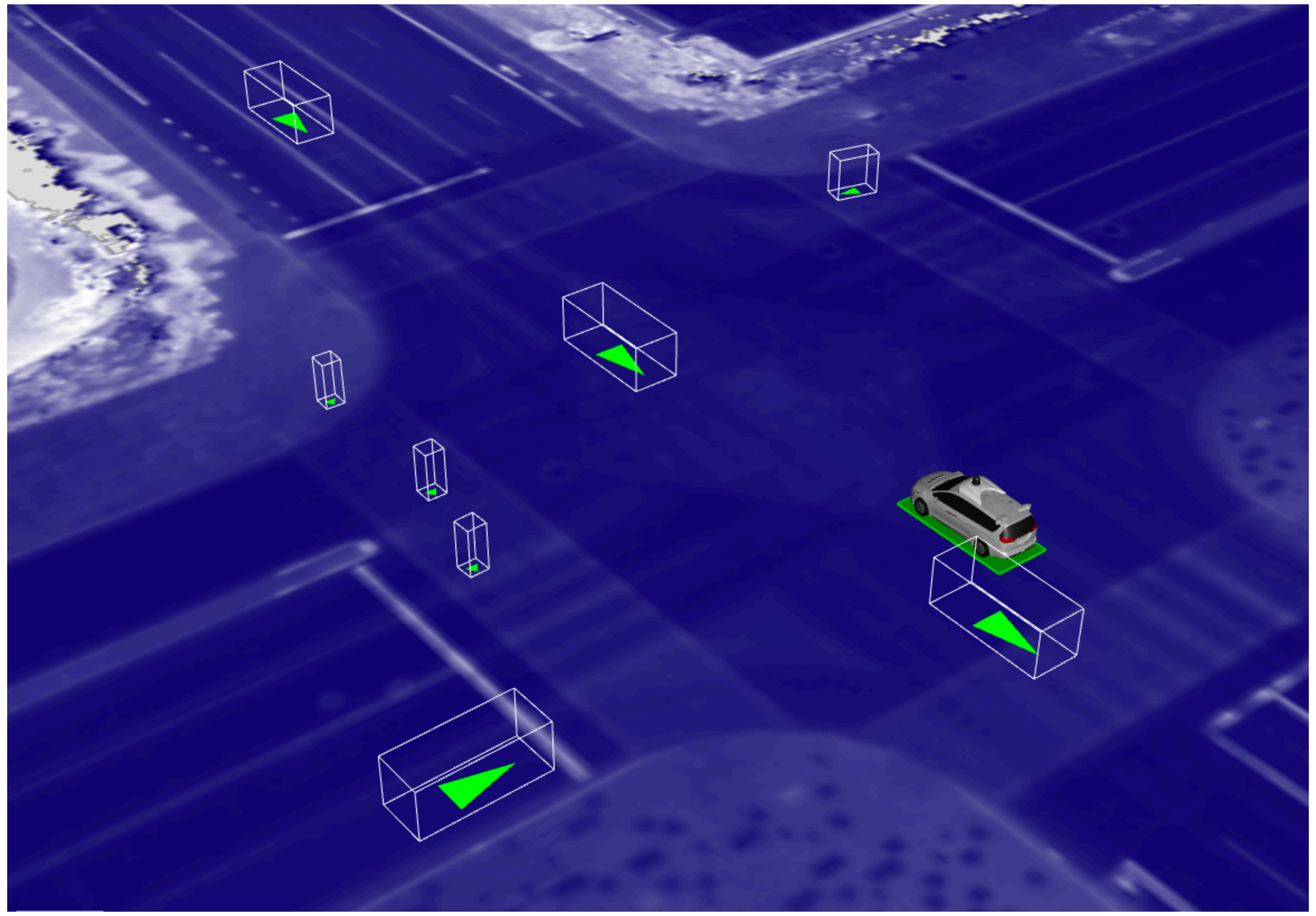
Theorem 2 [RHYLJR21]:
In the ***no-interaction*** *and* ***linear expert*** *setting,* ***Behavior Cloning*** *achieves,*

$$\text{Suboptimality} \lesssim \frac{dH^2 \log(N)}{N}$$

With $d = S$ recovers bounds in the tabular setting.

Linear expert with known transition

Known transition: *Learner is provided a dataset of N expert demonstrations;*
Knows the MDP transition

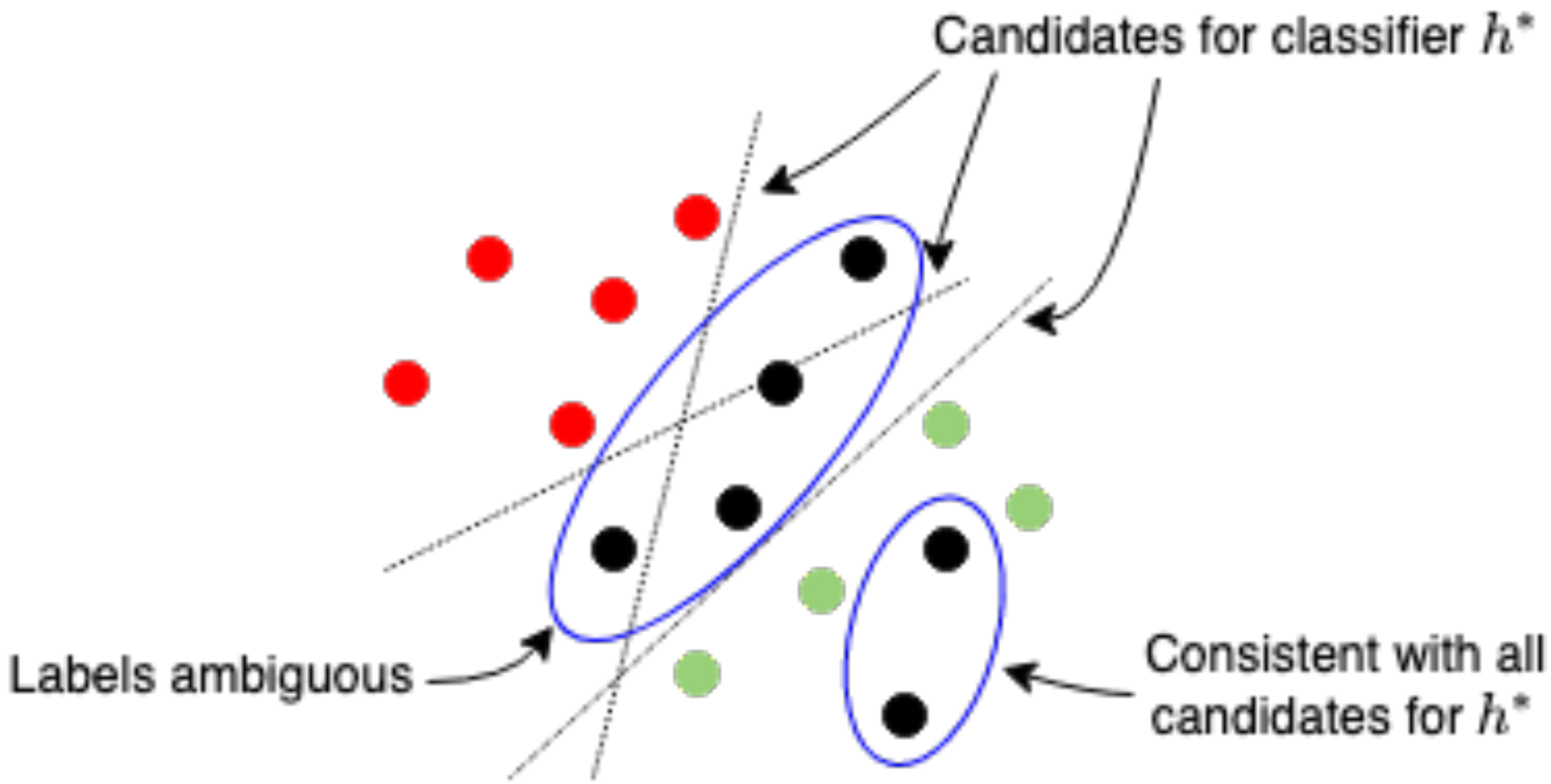


Interpretation: carrying out Imitation Learning in a simulation environment.

Image source: Waymo

Linear expert with known transition

Confidence set classification:
Consider classification over family of hypotheses, $\mathscr{H}$ from $\mathscr{X} \to \mathscr{Y}$.
From a dataset of examples D from a classifier $h*$ return the largest measure of points where $h*(x)$ is known without ambiguity.



Linear expert with known transition

Theorem 3 [RHYLJR21]:
For each t, consider the linear classifier $\pi_t^* : S \to A$.
Given a confidence set classifier with expected loss ℓ_t, there exists an IL algorithm such that,

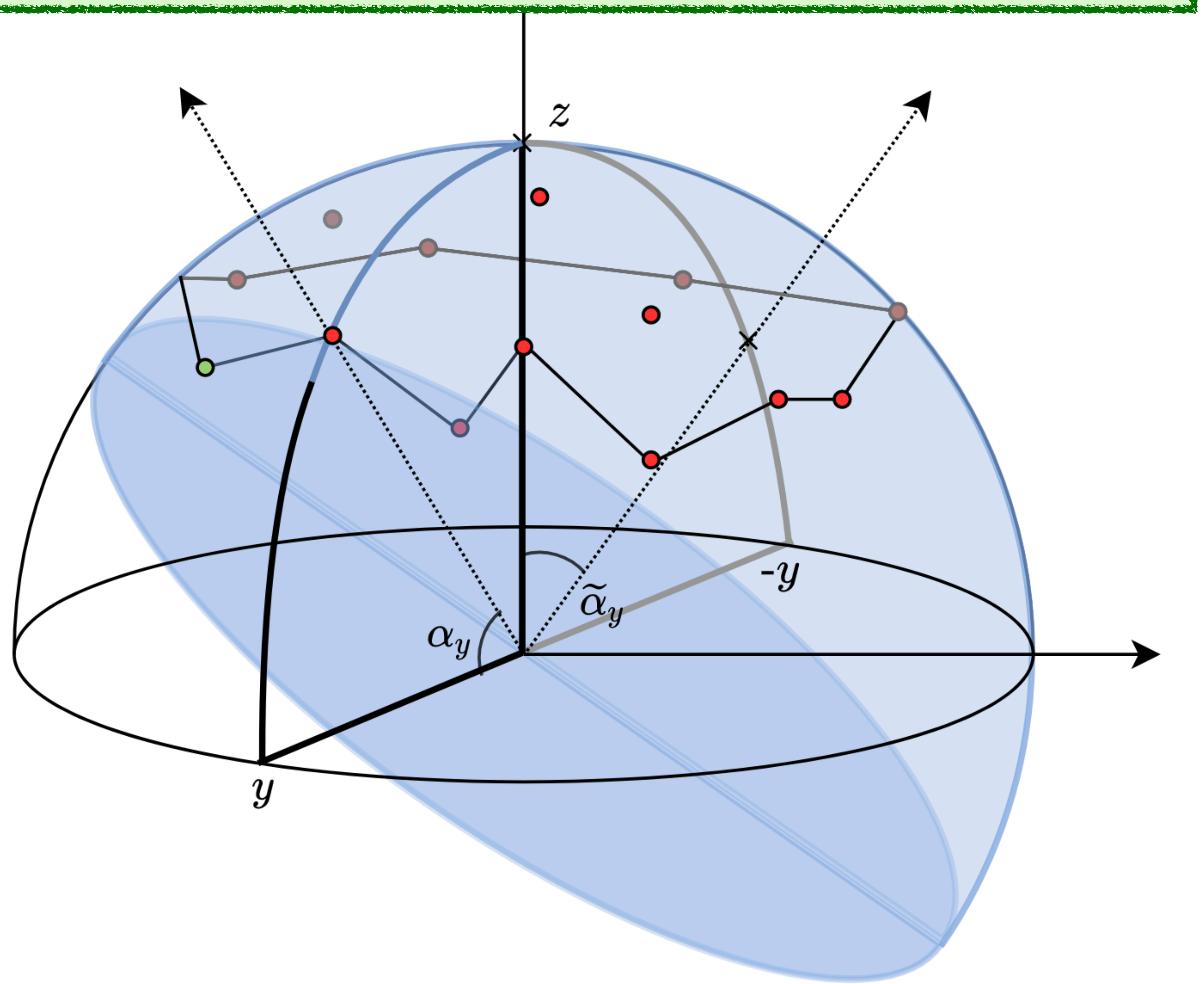
$$\text{Suboptimality} \lesssim H^{3/2} \sqrt{\frac{d}{N} \frac{\sum_{t=1}^{H} \ell_t}{H}}$$

Message: *Error compounding (H^2 dependence) can be broken if confidence set linear classification is possible to expected loss of $o_N(1)$.*

Linear expert with known transition

Theorem 4 [RHYLJR21]:
If distribution over inputs is uniform over the unit sphere $\mathbb{S}^{d-1}$, the minimax loss of confidence set linear classification is $\Theta(d^{3/2}/N)$.

Confidence set linear classification is sample efficient for the uniform distribution



Extending to general distributions?

On the value of Interaction and Function approximation in Imitation Learning

Nived Rajaraman, *Yanjun Han, Lin F. Yang, Jingbo Liu, Jiantao Jiao, Kannan Ramchandran*

NeurIPS 2021