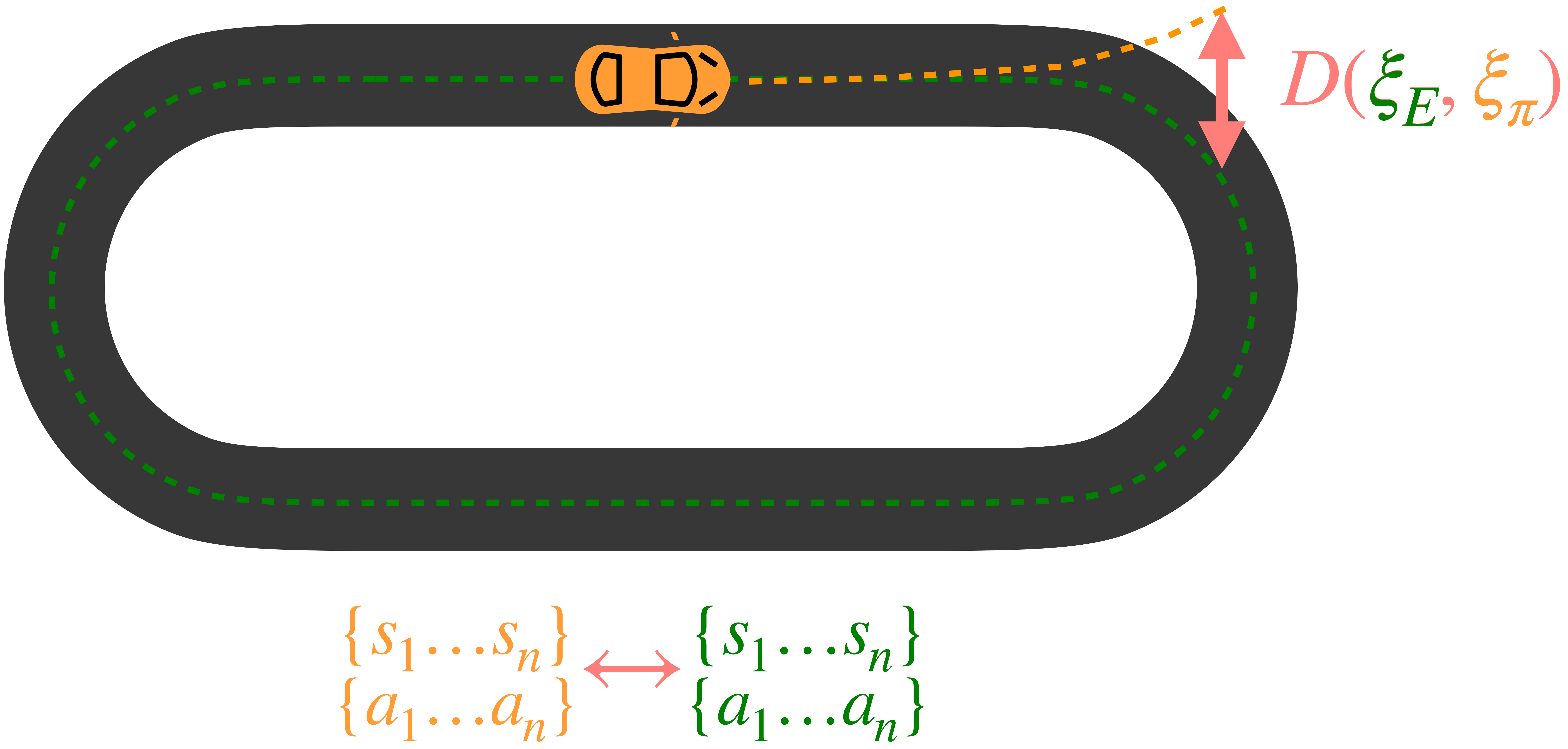
Hybrid Inverse Reinforcement Learning

Juntao Ren, Gokul Swamy*, Steven Wu, Drew Bagnell, Sanjiban Choudhury*

Inverse Reinforcement Learning for Imitation

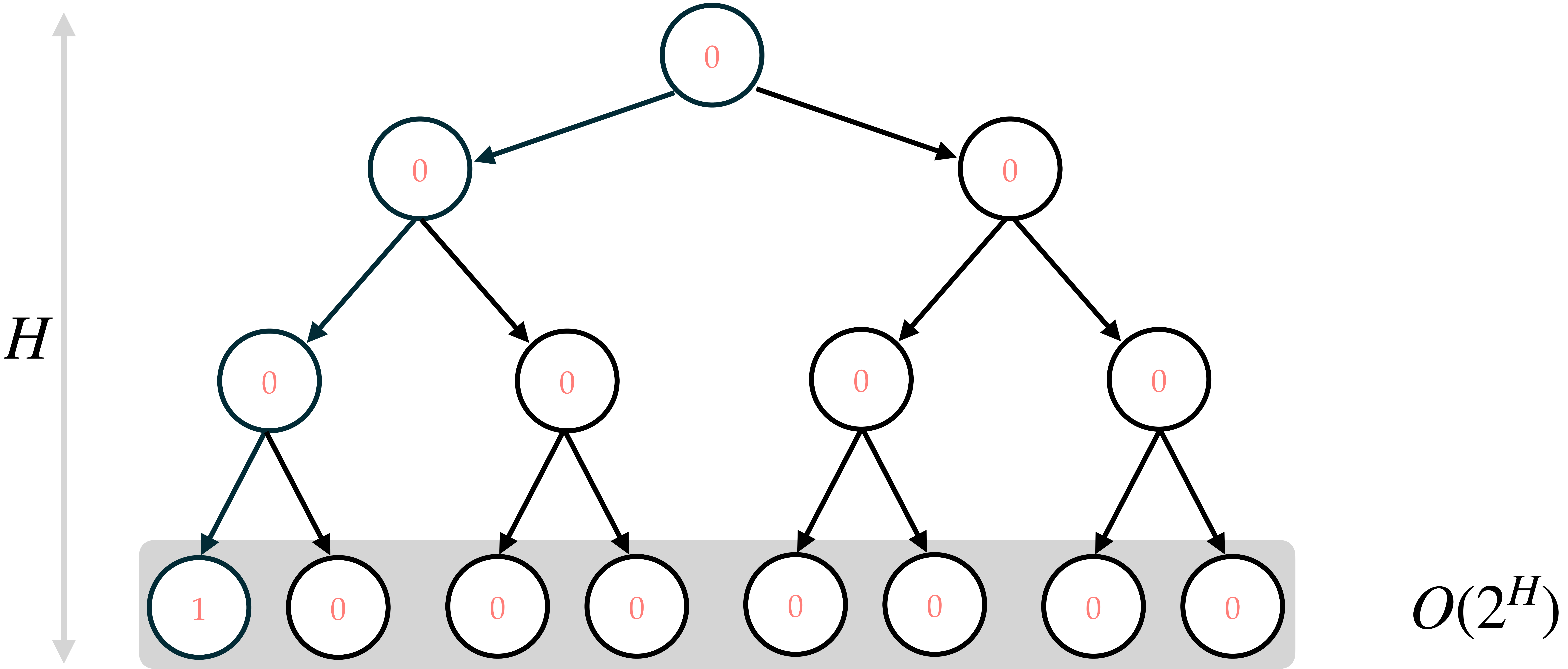


Requires repeatedly solving a hard exploration problem.

Robust to compounding errors …

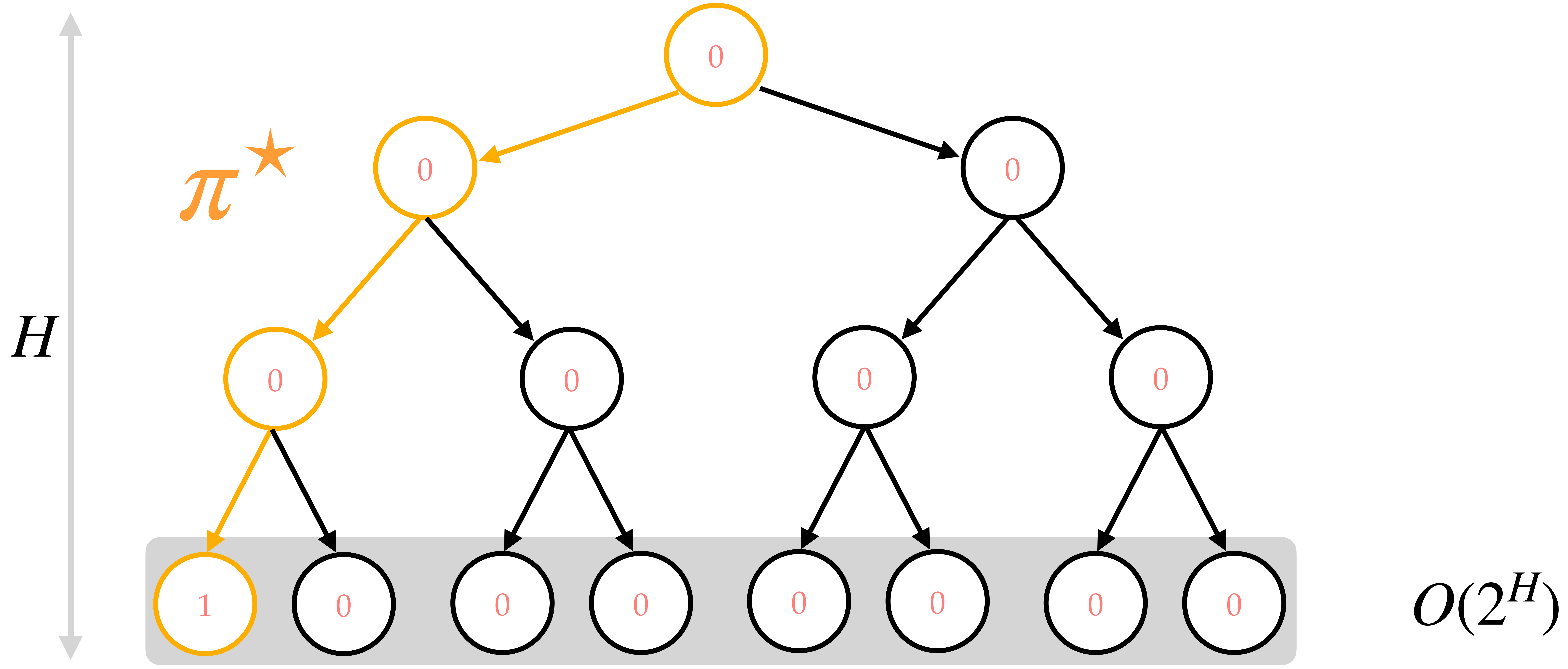
Exploration makes IRL Inefficient

$$\pi_E \overset{f}{\longleftrightarrow} \pi$$



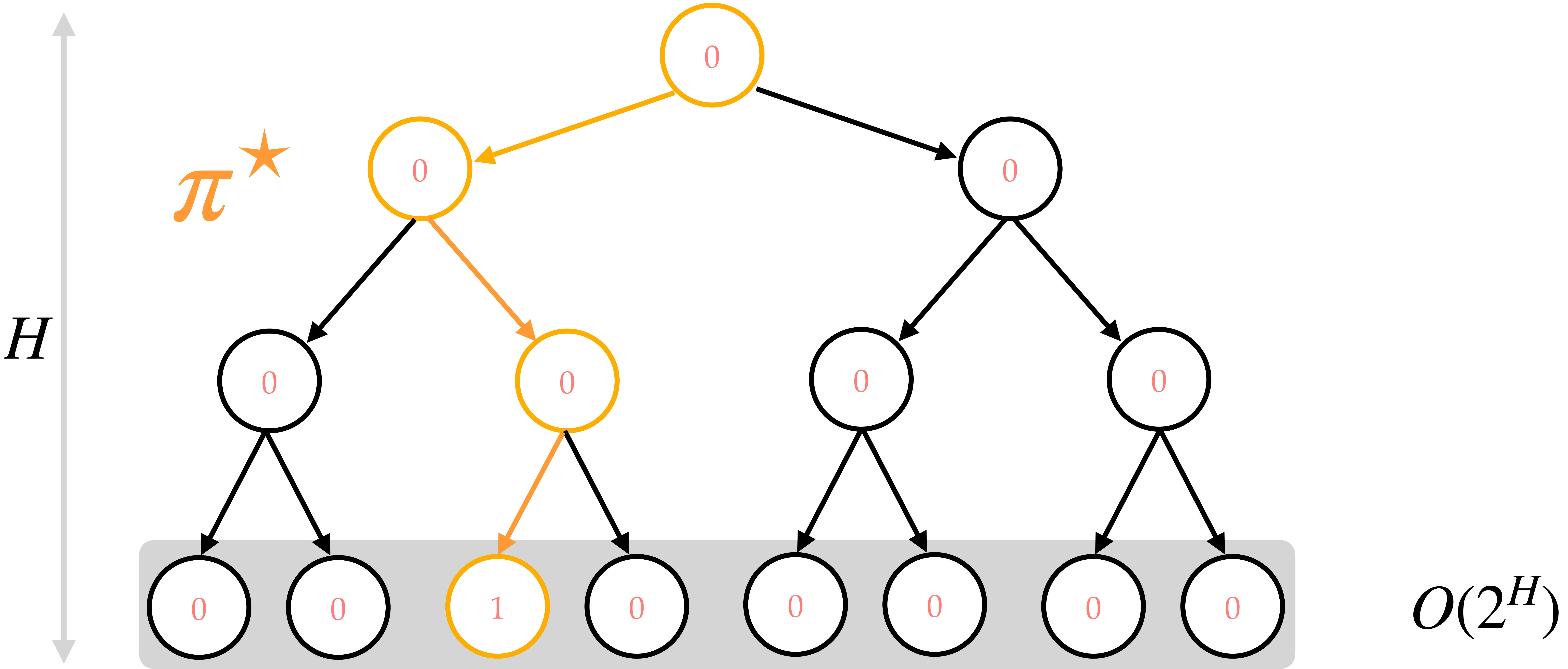
Exploration makes IRL Inefficient

$$\pi_E \overset{f}{\longleftrightarrow} \pi$$



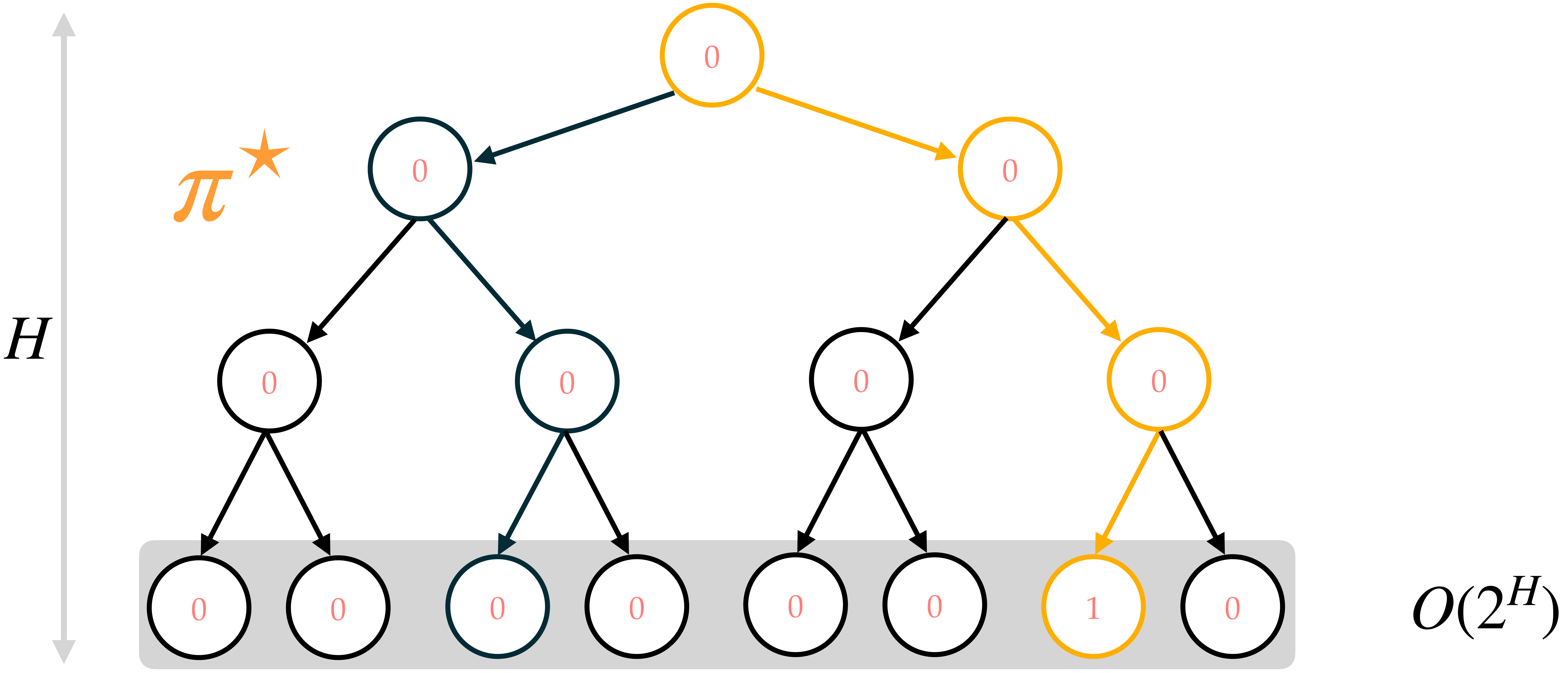
Exploration makes IRL Inefficient

$$\pi_E \overset{f}{\longleftrightarrow} \pi$$



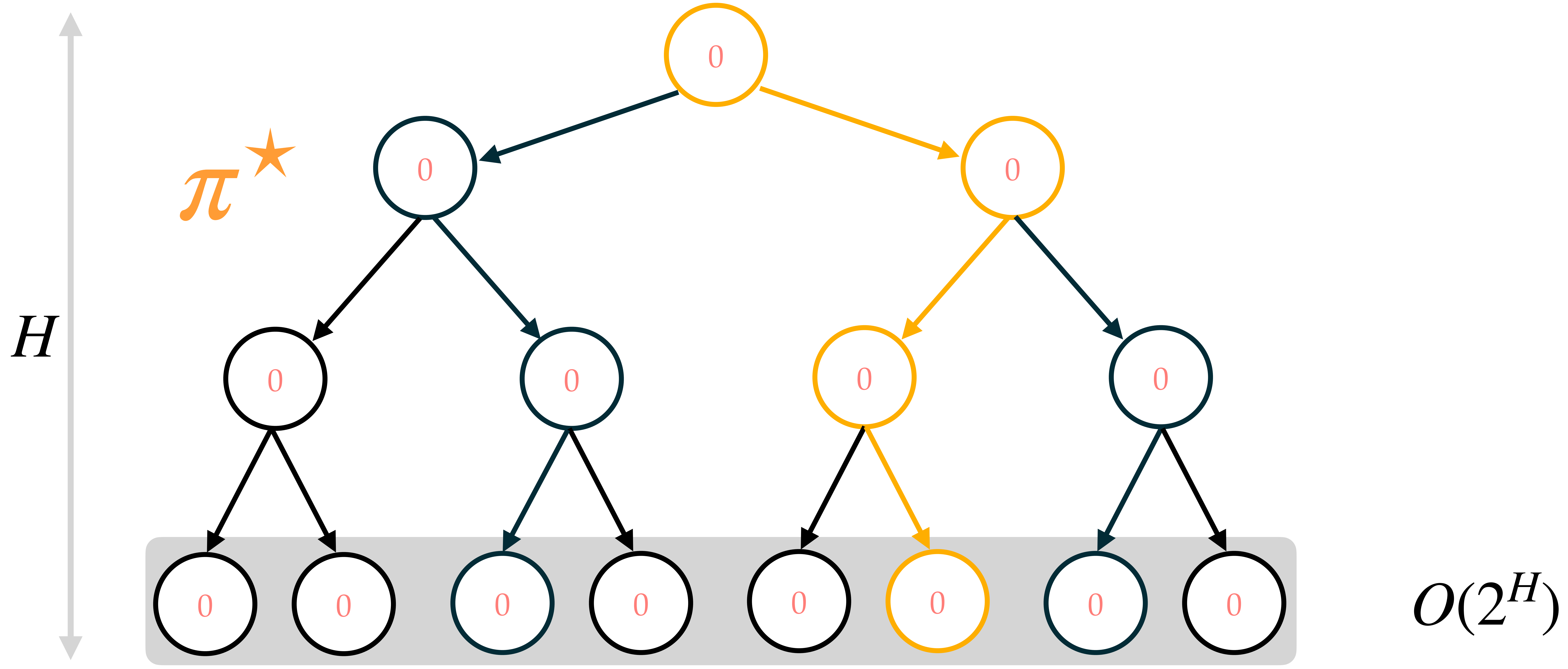
Exploration makes IRL Inefficient

$$\pi_E \overset{f}{\longleftrightarrow} \pi$$



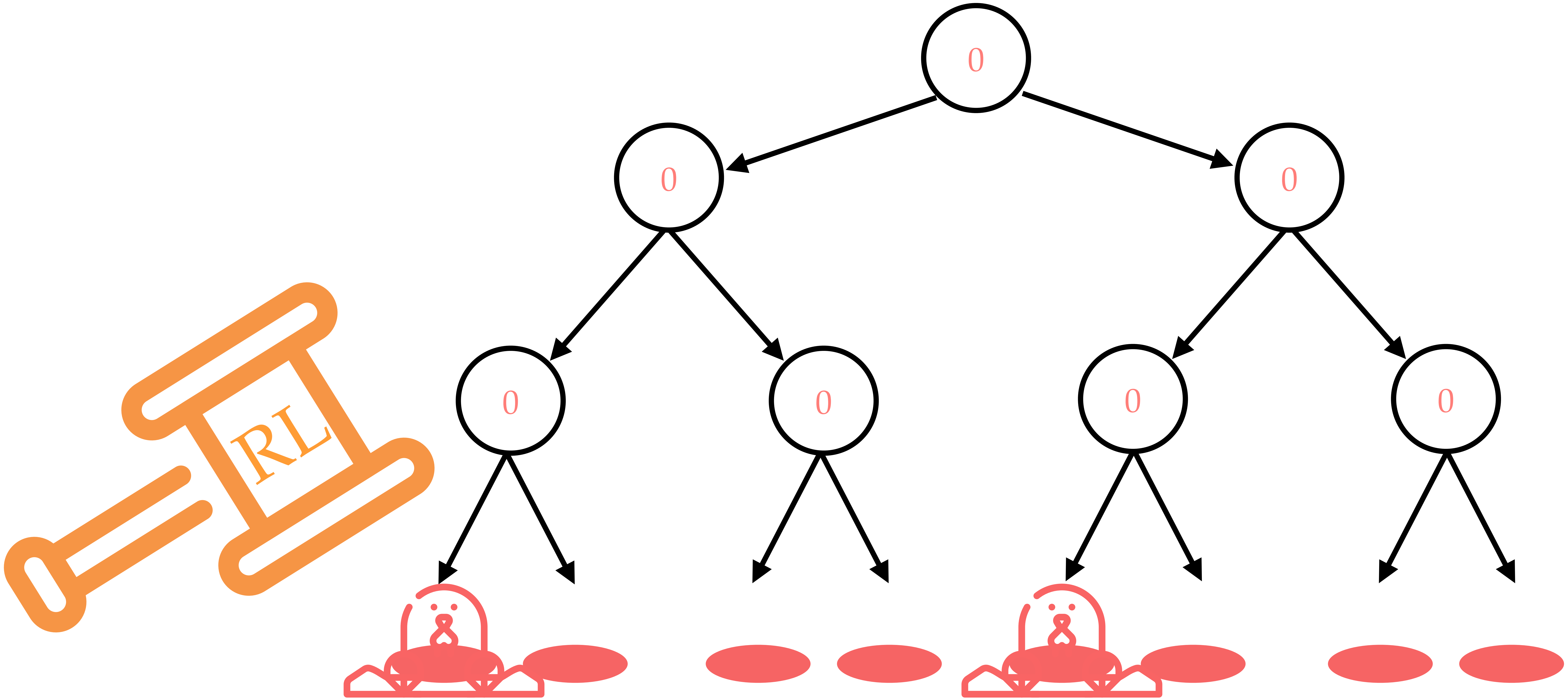
Exploration makes IRL Inefficient

$$\pi_E \overset{f}{\longleftrightarrow} \pi$$



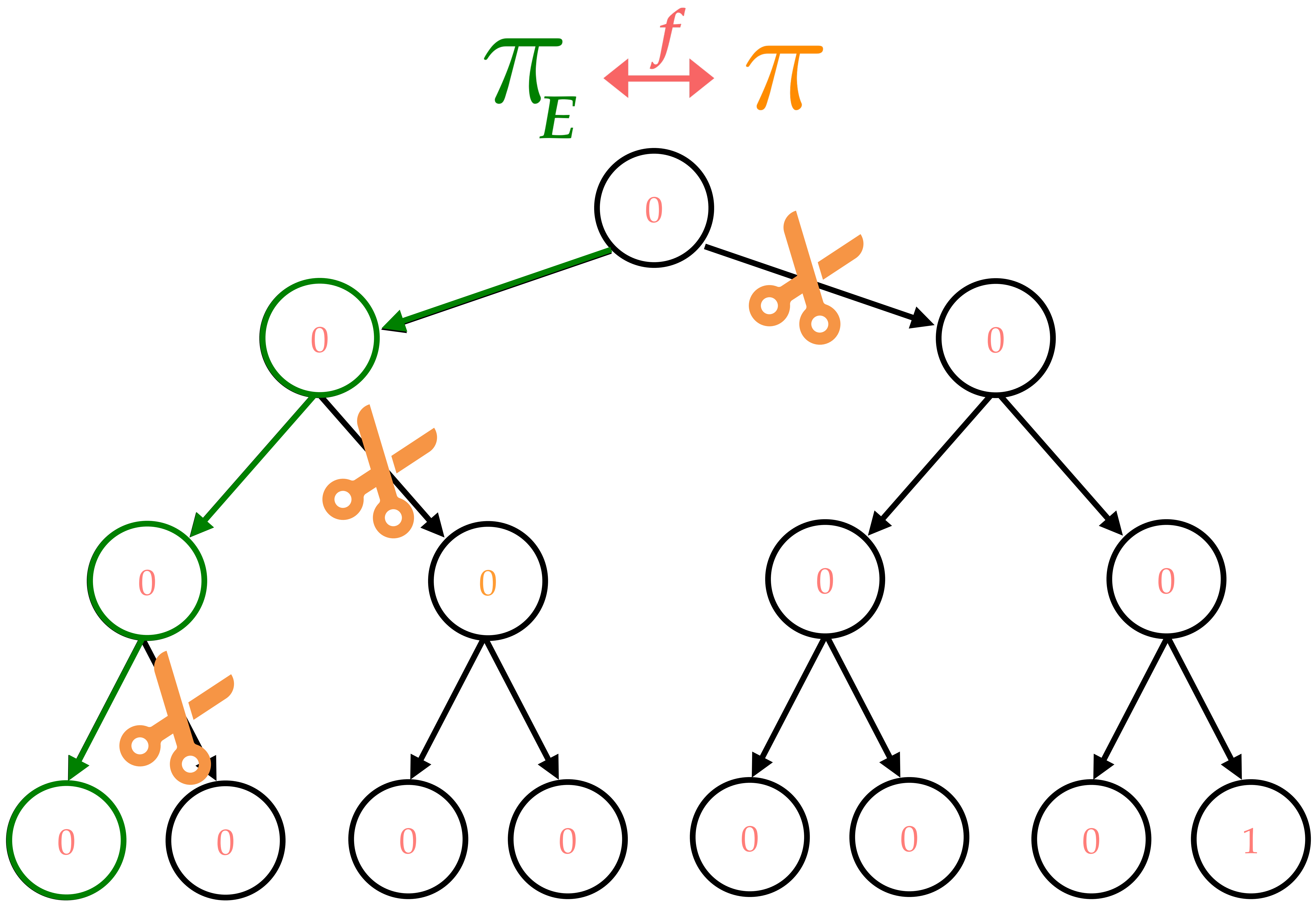
We're playing adversarial whack-a-mole with an RL Hammer

$$\pi_E \overset{f}{\leftrightarrow} \pi$$



🔑 ***Question**: How do we reduce the amount of exploration performed in inverse RL?*

🔑 ***Idea:** We don't need to compute a best response via RL, just compete with the expert!*



A Unifying Mathematical Framework for Efficient IRL

$\mathsf{ERROr}\{\mathsf{Reg}_\pi(T)\}$: A policy-selection algorithm $\mathbb{A}_\pi$ satisfies the $\mathsf{Reg}_\pi(T)$ expert-relative regret guarantee if given any sequence of reward functions $f_{1:T}$, it produces a sequence of policies $\pi_{t+1} = \mathbb{A}_\pi(f_{1:t})$ such that

$$\sum_{t=1}^{T} J(\pi_E, f_t) - J(\pi_t, f_t) \leq \mathsf{Reg}_\pi(T)\,.$$

Notice that we never need to compute a best response to an f_t*!*

A Unifying Mathematical Framework for Efficient IRL

$$
\begin{aligned}
J(\pi_E, r) - J(\bar{\pi}, r) &= \frac{1}{T}\sum_{t=1}^{T} J(\pi_E, r) - J(\pi_t, r) \\
&\leq \max_{f^\star \in \mathcal{F}_r} \frac{1}{T}\sum_{t=1}^{T} J(\pi_E, f^\star) - J(\pi_t, f^\star) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} J(\pi_E, f_t) - J(\pi_t, f_t) + \frac{\mathrm{Reg}_f(T)}{T} H \\
&\leq \frac{\mathrm{Reg}_\pi(T)}{T} + \frac{\mathrm{Reg}_f(T)}{T} H .
\end{aligned}
$$

[RS+ ’24]

Q: *What algorithms satisfy the* ERROr *property?*

A1*: Expert Resets*

`FILTER`

A2*: Hybrid Training*

`HyPE`

A3*: Hybrid Model Fitting*

`HyPER`

***Q**: What algorithms satisfy the* ERROr *property?*

***A1**: Expert Resets*

`FILTER`

***A2**: Hybrid Training*

`HyPE`

***A3**: Hybrid Model Fitting*

`HyPER`

[S+'23, B+'03]

***Q**: What algorithms satisfy the* ERROr *property?*

***A1**: Expert Resets*

`FILTER`

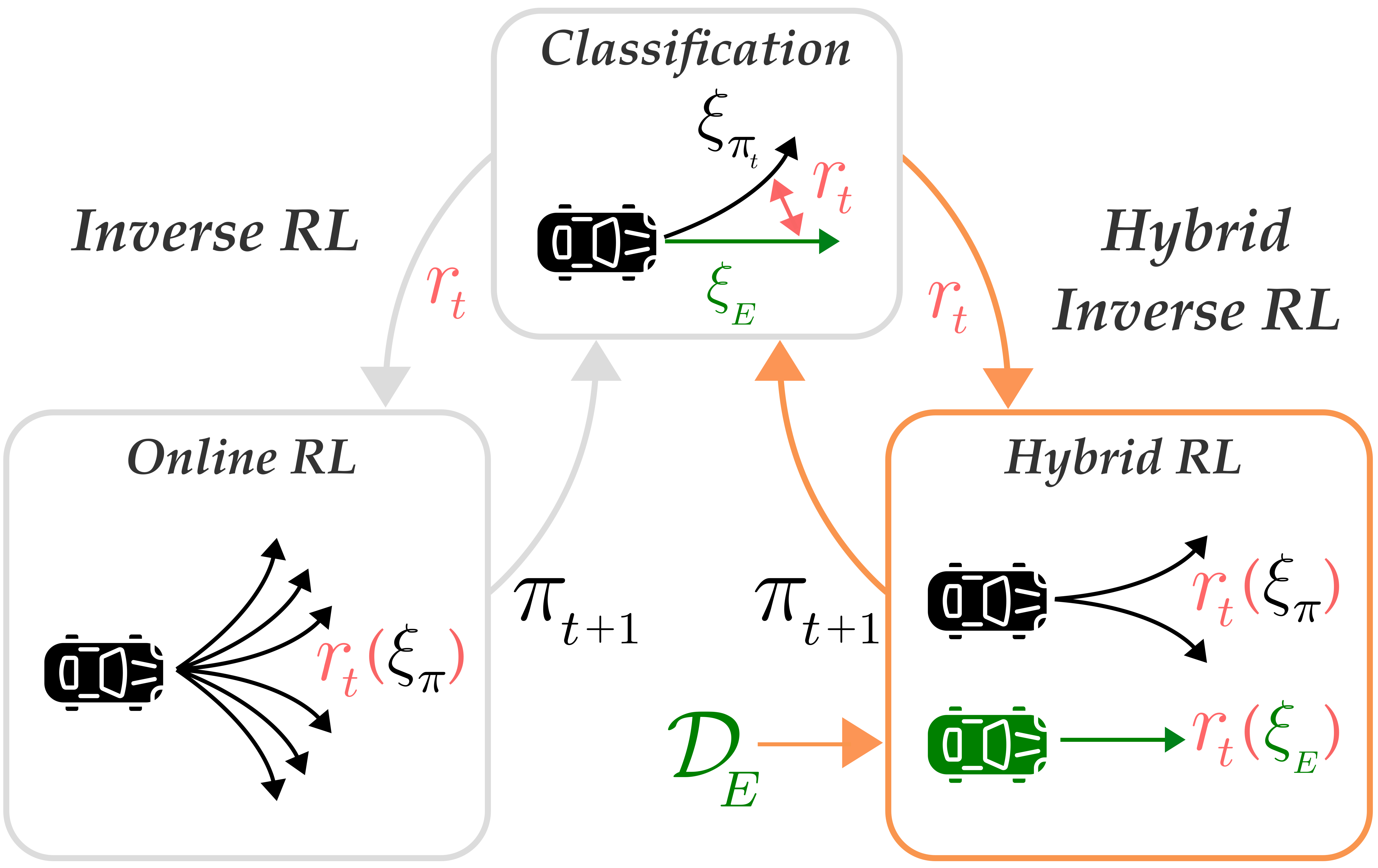
***A2**: Hybrid Training*

`HyPE`

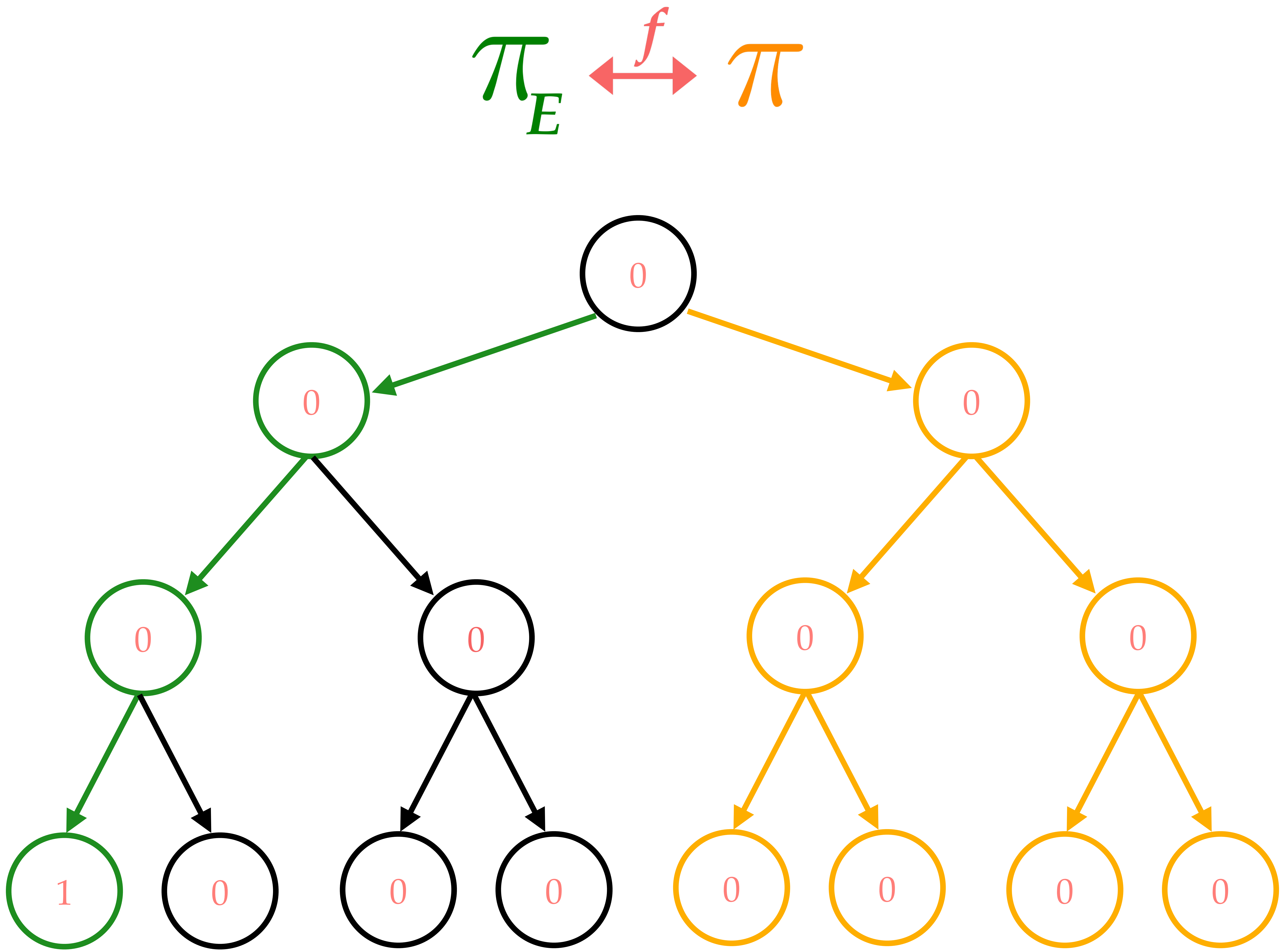
***A3**: Hybrid Model Fitting*

`HyPER`

Speeding up IRL with Hybrid Training



Speeding up IRL with Hybrid Training



[So+'22]

***Q**: What algorithms satisfy the* ERROr *property?*

***A1**: Expert Resets*

`FILTER`

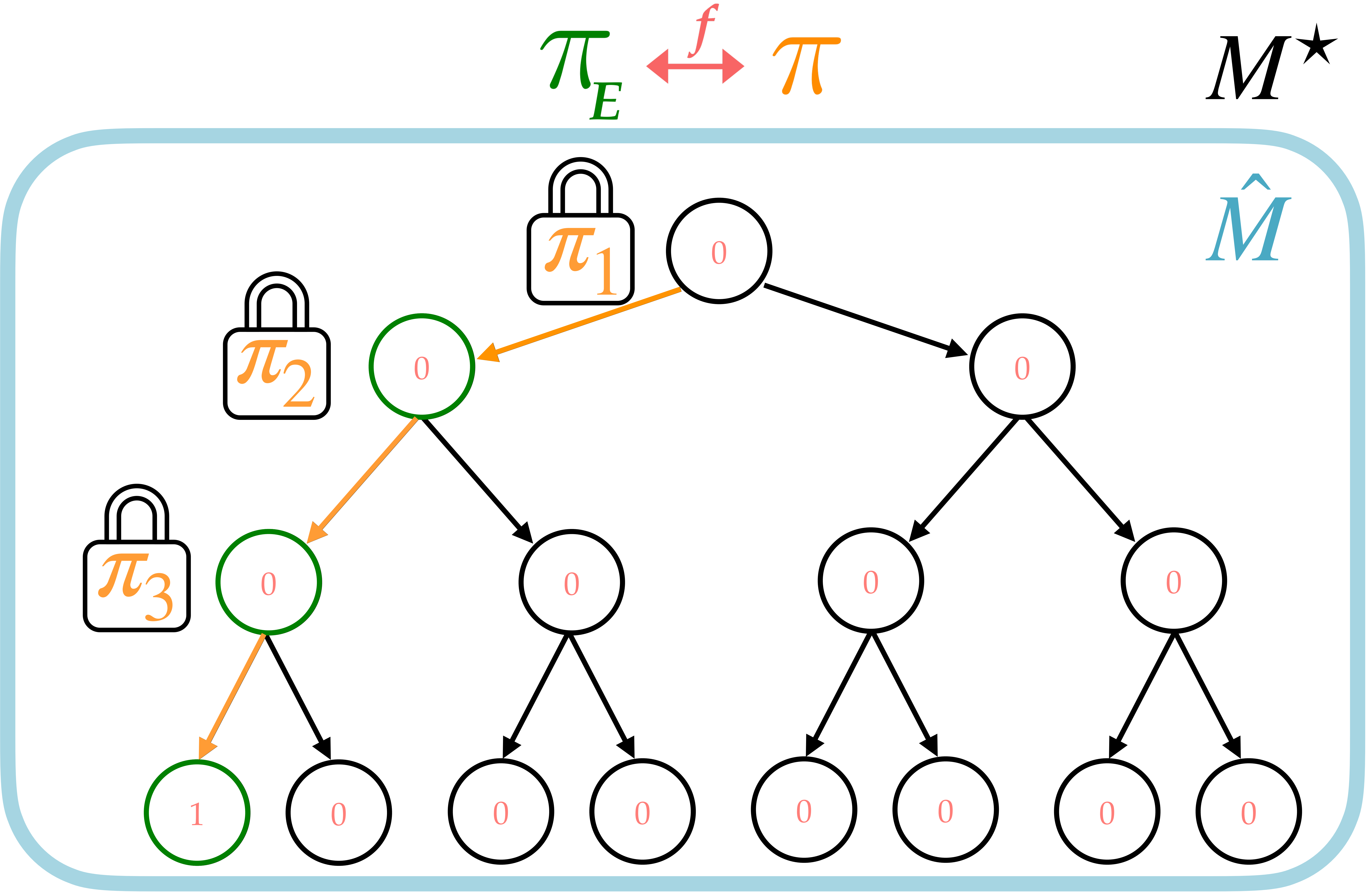
***A2**: Hybrid Training*

`HyPE`

***A3**: Hybrid Model Fitting*

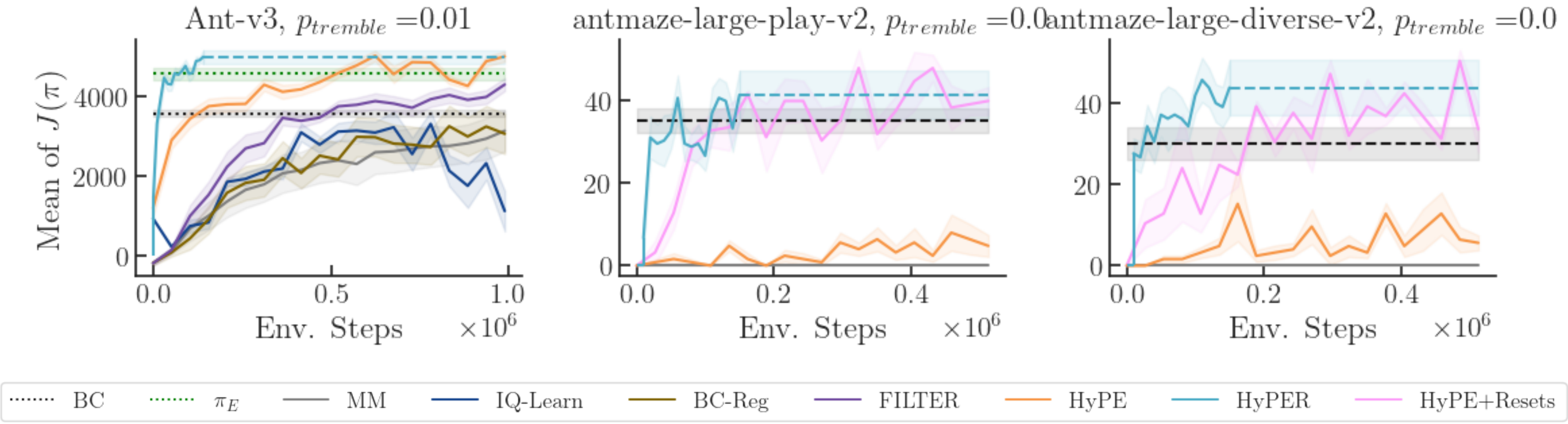
`HyPER`

Speeding up IRL with Hybrid Model Fitting



[V+'23]

Hybrid Model Fitting Speeds Up IRL (even more)



Thanks!

Paper

Code

gswamy@cmu.edu, jlr429@cornell.edu