Efficient Algorithms for Inverse Reinforcement Learning

Gokul Swamy

Recap from Last Time

1. Why do we need interaction in imitation learning?

***A**: to be able to tell that we've made a mistake that compounds.*

2. What else do we need to tell which mistakes matter?

***A**: information about the **set** of rewards we could be judged on.*

3. How do we learn a policy that recovers from mistakes that matter if we don't know what the reward function is?

***A**: Find the policy that is the least distinguishable from the expert's under **any** reward function in the moment set $\mathscr{R}$.*

Outline for Today

1. What makes inverse RL sample-inefficient?

2. Are best responses required for solving the IRL game?

3. What algorithms can we use in our new reduction?

Outline for Today

1. What makes inverse RL sample-inefficient?

***A**: Repeatedly solving a global exploration (RL) problem.*

2. Are best responses required for solving the IRL game?

3. What algorithms can we use in our new reduction?

Recap: Inverse RL as Game-Solving

$$\max_{\pi \in \Pi} \min_{f \in \mathscr{R}} J(\pi, f) - J(\pi_E, f)$$

where $J(\pi, f) \triangleq \mathbb{E}_{\xi \sim \pi}\left[\sum_{h}^{H} f(s_h, a_h)\right].$

Recap: Two Flavors of IRL Algorithms

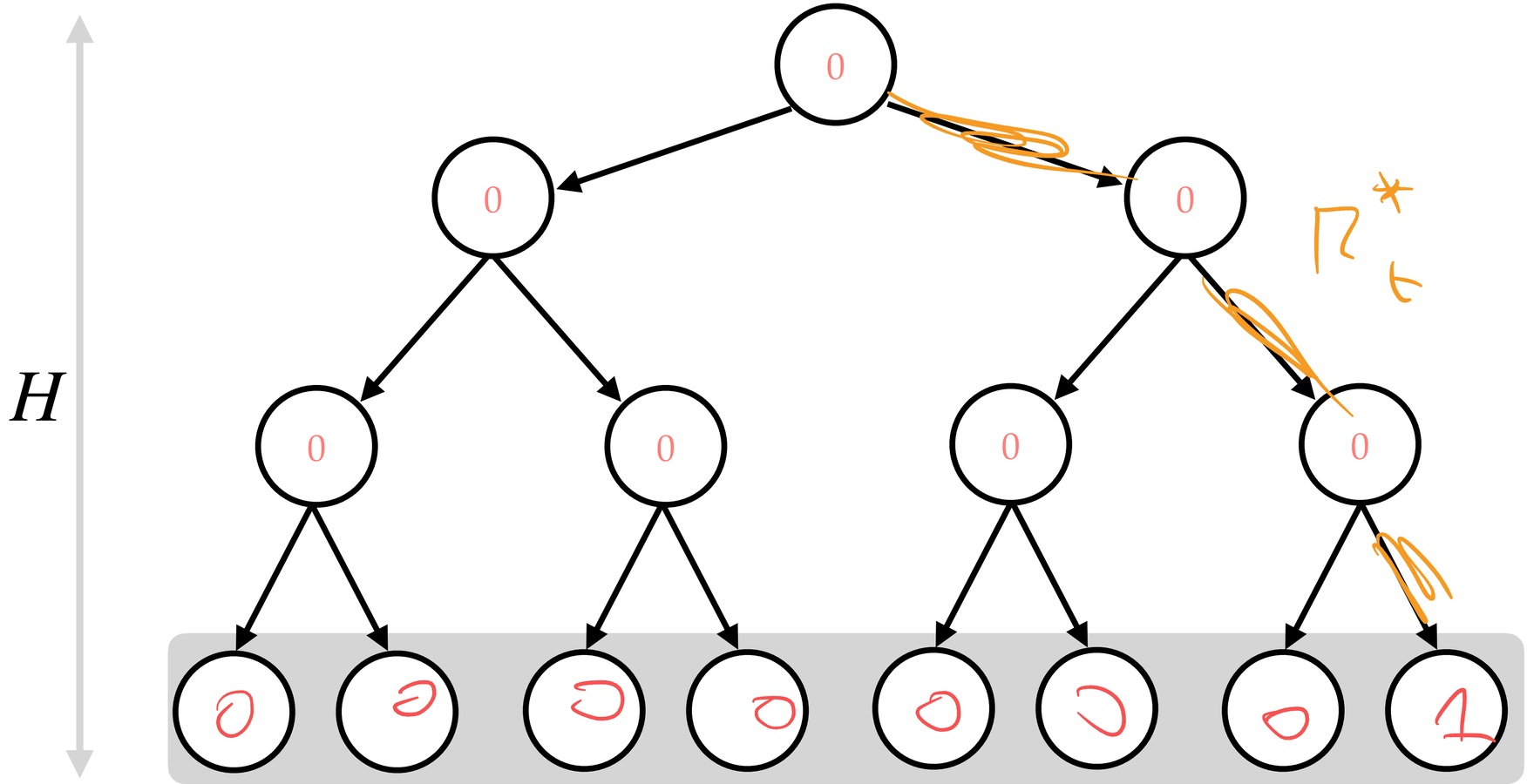
$$\max_{\pi \in \Pi} \min_{f \in \mathcal{R}} J(\pi, f) - J(\pi_E, f)$$

	Dual	*Primal*
Policy Update	**BR:** RL	**NR:** GD
Reward Update	**NR:** GD	**NR:** GD

Exploration makes (I)RL Inefficient

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

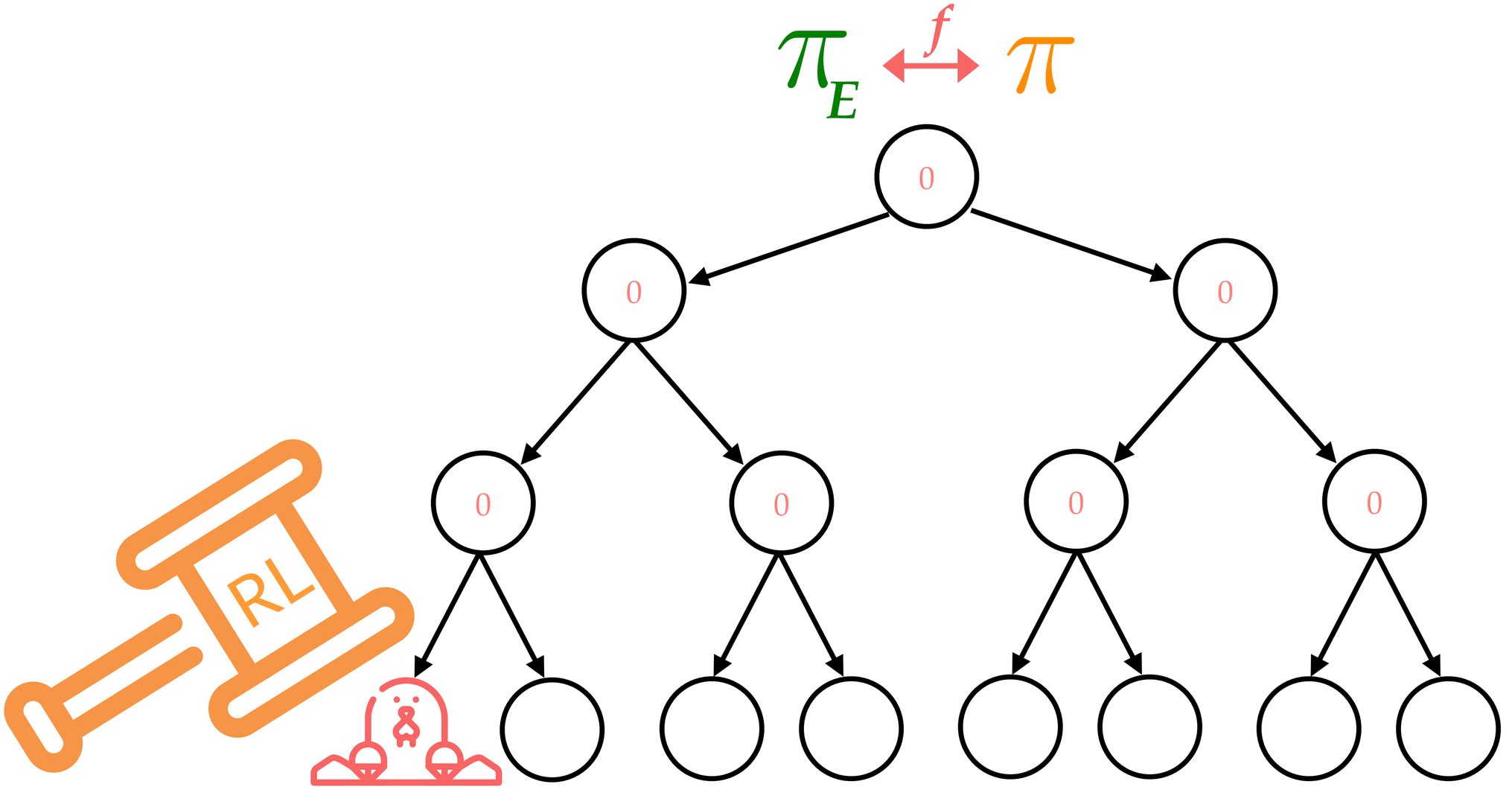
$\mathcal{R} = \{$ indicator of a leaf node $\}$



r^*

$O(2^H)$

*We're Playing **Adversarial** Whack-a-Mole!*



We've reduced the "easier" problem of IL to the "harder" problem of RL

$$\max_{\pi \in \Pi} \min_{f \in \mathcal{R}} J(\pi, f) - J(\pi_E, f)$$

Inefficient

	Dual	*Primal*
Policy Update	**BR**: RL	**NR**: GD
Reward Update	**NR**: GD	**NR**: GD

Primal algorithms also need to explore as we can force *any* no-regret algorithm to compute $\pi^\star$ by playing the same f!

Outline for Today

1. What makes inverse RL sample-inefficient?

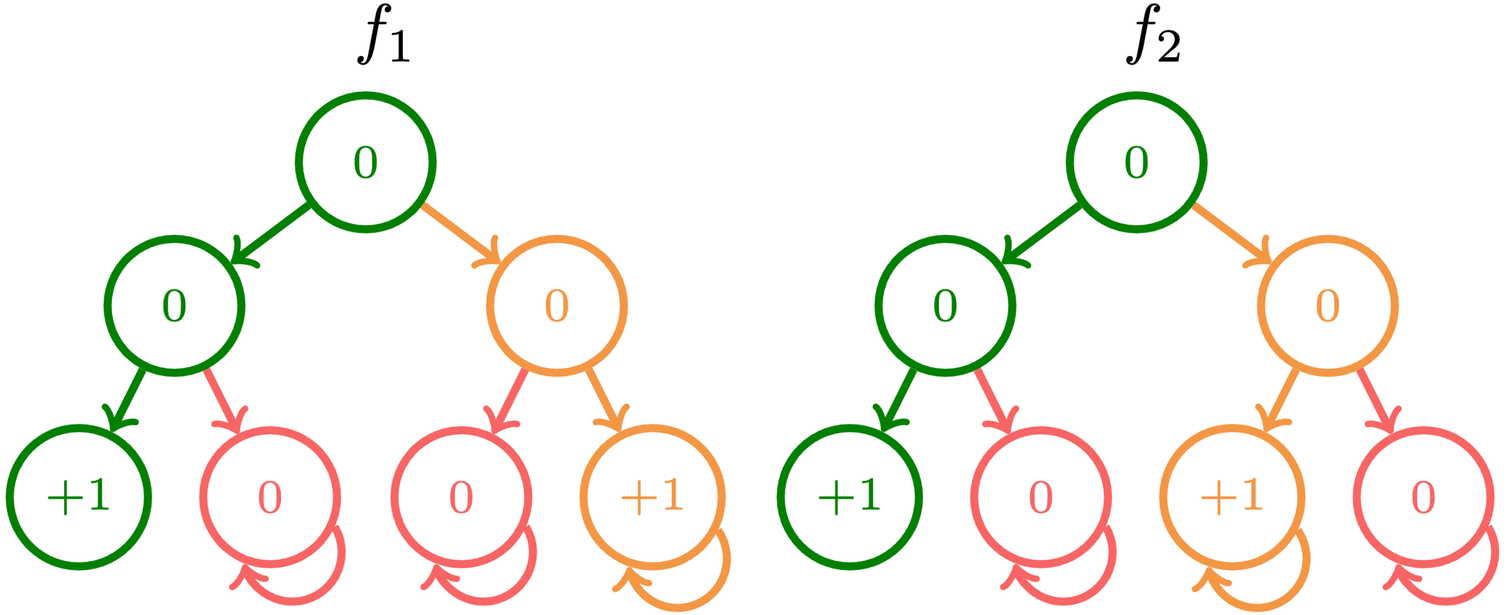
***A**: Repeatedly solving a global exploration (RL) problem.*

2. Are best responses required for solving the IRL game?

***A**: Actually, competing with the expert is "all you need".*

3. What algorithms can we use in our new reduction?

π_E need not be optimal for f_t ...



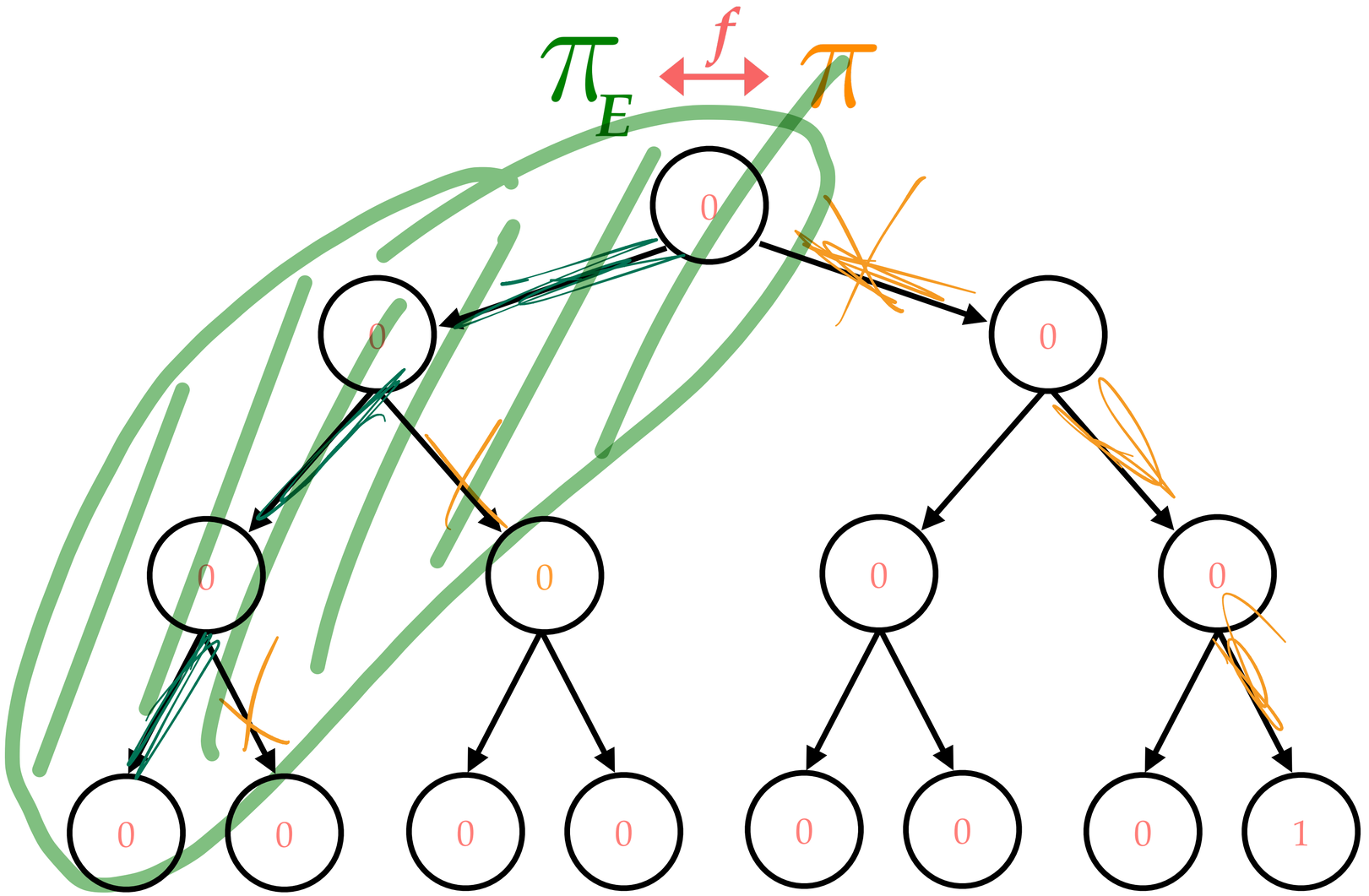
$\pi_1 = \text{argmax}_{\pi \in \Pi} J(\pi, f_1)$ $\pi_2 = \text{argmax}_{\pi \in \Pi} J(\pi, f_2)$

compete

... so $\pi_t^\star$ might be far from π_E ...

... which means it can't be the policy we want!

🔑 Idea: save compute by competing with π_E, not $\pi_t^\star$!



Reducing IRL to Expert-Competitive RL

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

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

🔑 *Idea: We never need to compute a best response to an f_t!*

Reducing IRL to Expert-Competitive RL

$\mathbb{A}_f$ is a ***no-regret reward selection algorithm*** if when given a sequence of policies $\pi_{1:t}$, it produces iterates $f_{t+1} = \mathbb{A}_f(\pi_{1:t})$ such that

$$\sum_{t=1}^{T} J(\pi_t, f_t) - J(\pi_E, f_t) - \min_{f^\star \in \mathcal{F}_r} \sum_{t=1}^{T} J(\pi_t, f^\star) - J(\pi_E, f^\star) \leq H\mathrm{Reg}_f(T),$$

$$\text{with } \lim_{T\to\infty} \frac{\mathrm{Reg}_f(T)}{T} = 0.$$

Reducing IRL to Expert-Competitive RL

linearity of expectation

$$J(\pi_E, r) - J(\bar{\pi}, r) = \frac{1}{T}\sum_{t=1}^{T} J(\pi_E, r) - J(\pi_t, r)$$

reward realizability

$$\leq \max_{f^\star \in \mathcal{F}_r} \frac{1}{T}\sum_{t=1}^{T} J(\pi_E, f^\star) - J(\pi_t, f^\star)$$

defn. of regret

$$\leq \frac{1}{T}\sum_{t=1}^{T} J(\pi_E, f_t) - J(\pi_t, f_t) + \frac{\mathrm{Reg}_f(T)}{T} H$$

$$\leq \frac{\mathrm{Reg}_\pi(T)}{T} + \frac{\mathrm{Reg}_f(T)}{T} H\,.$$

Outline for Today

1. What makes inverse RL sample-inefficient?

***A**: Repeatedly solving a global exploration (RL) problem.*

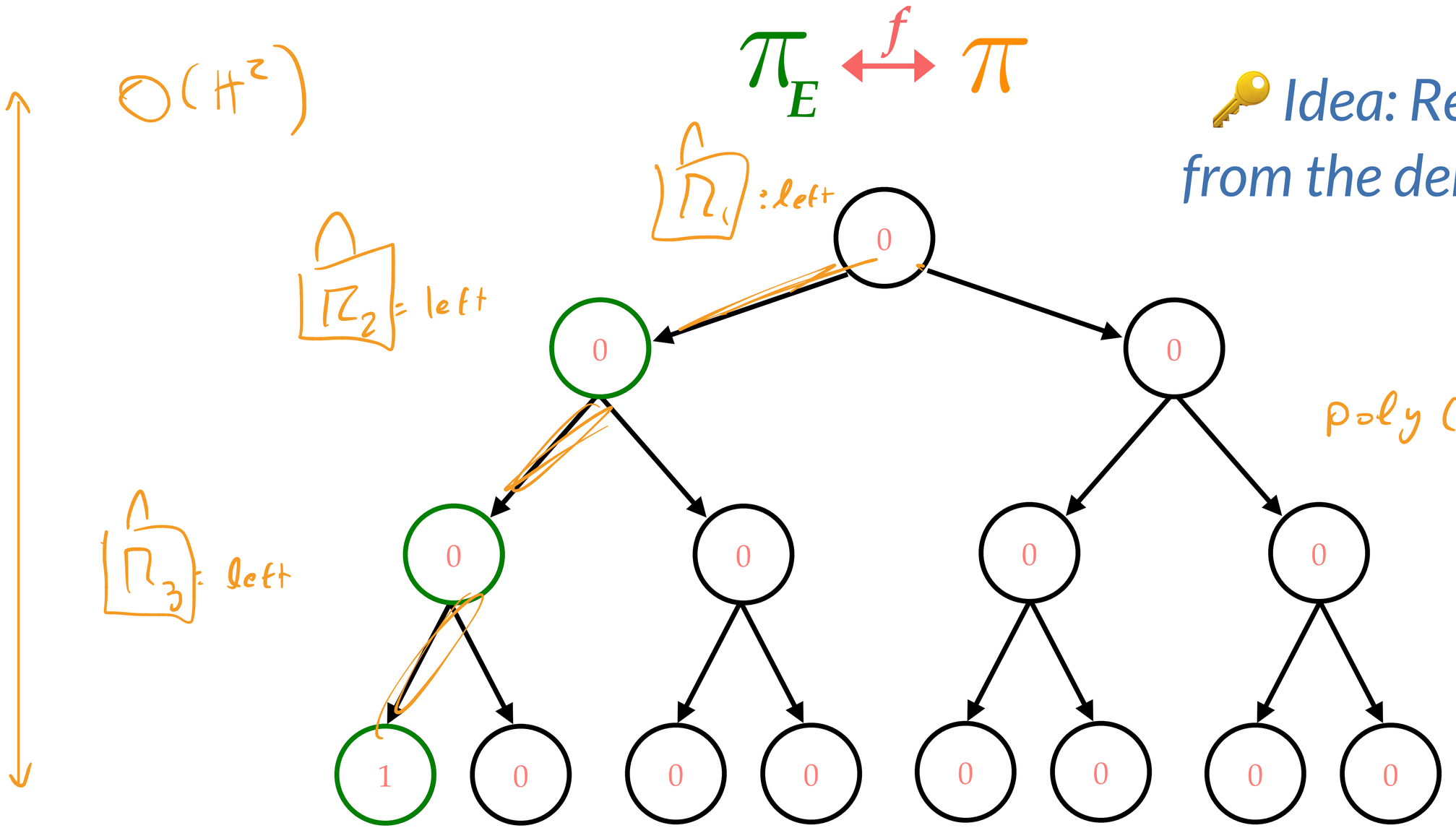
2. Are best responses required for solving the IRL game?

***A**: Actually, competing with the expert is "all you need".*

3. What algorithms can we use in our new reduction?

***A**: A wide variety of sample-efficient "local search" algorithms.*

Expert-Competitive RL via PSDP



🔑 *Idea: Reset to states from the demonstrations!*

Expert-Competitive RL via PSDP

$$\pi_H = \arg\max_{\pi \in \Pi} \mathbb{E}_{s_H \sim \pi_E}[\mathbb{E}_{a_H \sim \pi}[f_t(s_H, a_H)]]$$

For $h \in [H-1, H-2, \ldots, 1]$:

$$\pi_h = \arg\max_{\pi \in \Pi} \mathbb{E}_{s_h \sim \pi_E}[\mathbb{E}_{a_h \sim \pi}[f_t(s_h, a_h) + \mathbb{E}_{T(s_h, a_h)}[V_t^{\pi_{h+1:H}}(s_{h+1})]]]$$

Lemma: Assume that at each time-step $h \in [H]$, we perform policy optimization up to ε-optimality:

$$\mathbb{E}_{s_h, a_h \sim \pi_E}[Q_t^{\pi_{h+1:H}}(s_h, a_h) - \mathbb{E}_{a \sim \pi_h(s_h)}[Q_t^{\pi_{h+1:H}}(s_h, a)]] \leq \varepsilon H$$

Then,

$$J(\pi_E, f_t) - J(\pi_{1:H}, f_t) \leq \mathcal{O}(\varepsilon H^2)$$

Proof: We proceed via the PDL (shocking, I know):

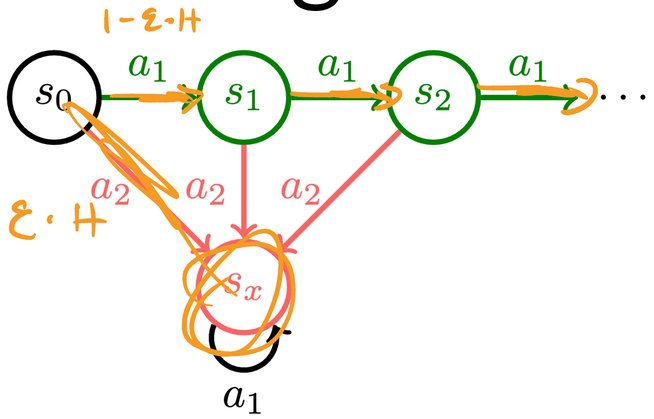
$$J(\pi_E, f_t) - J(\pi_{1:H}, f_t) = \sum_h^H \mathbb{E}_{s_h, a_h \sim \pi_E} \left[Q^{\pi_{h+1:H}}(s_h, a_h) - \mathbb{E}_{a \sim \pi_h(s_h)}[Q^{\pi_{h+1:H}}(s_h, a_h)] \right]$$

$$\leq \sum_h^H \varepsilon(H - h)$$

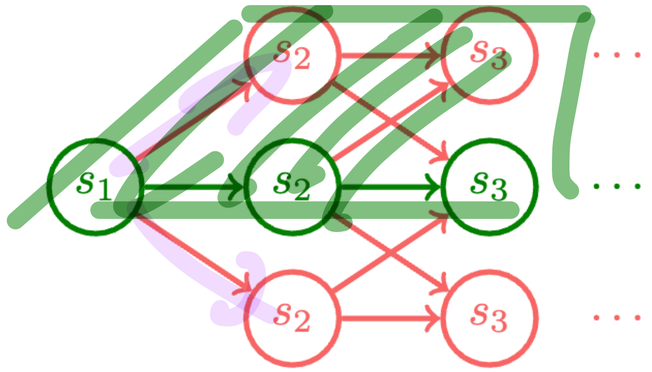
$$\leq \mathcal{O}(\varepsilon H^2)$$

$$\Rightarrow \mathsf{Reg}_\pi(T) \leq \varepsilon H^2 T$$

Aside: Compounding Errors in PSDP



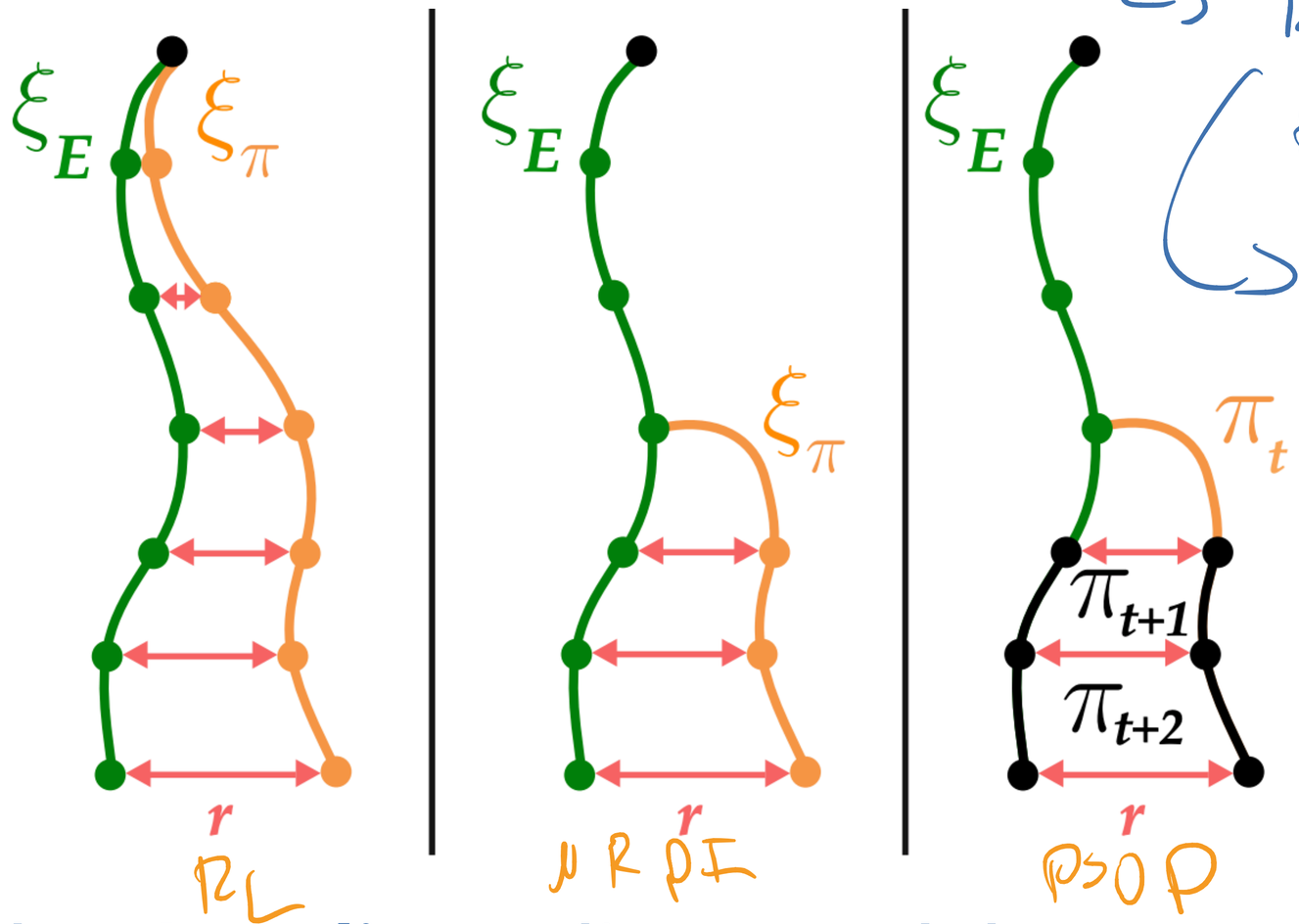
Unavoidable in general on cliff-like (irrecoverable) problems.



Interaction can still help us figure out which mistakes compound.

(Can interpolate with or anneal towards ρ_0 in practice)

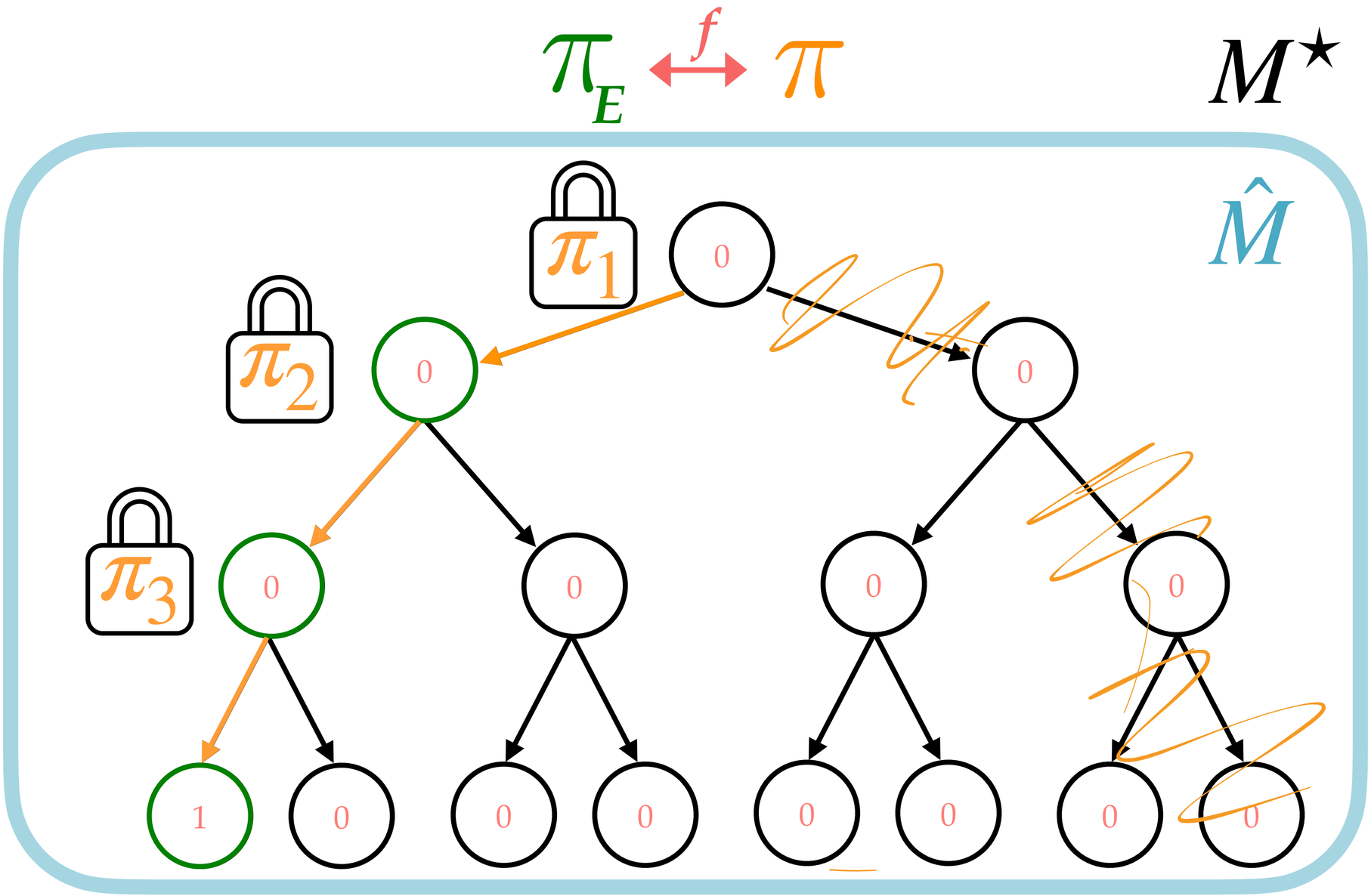
poly(H) Algorithms for IRL!



↳ RL w/ resets to expert states
↳ PPO, SAC)

🔑 *Idea: **Localize** policy search by resetting to states from the demonstrations!*

Aside: Use (Hybrid) Model for Resets In Practice



Aside: Can Also Use HyQ w/o Simulator Access

Ls off-policy RL (e.g. SAC) w/ hybrid replay buffer

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

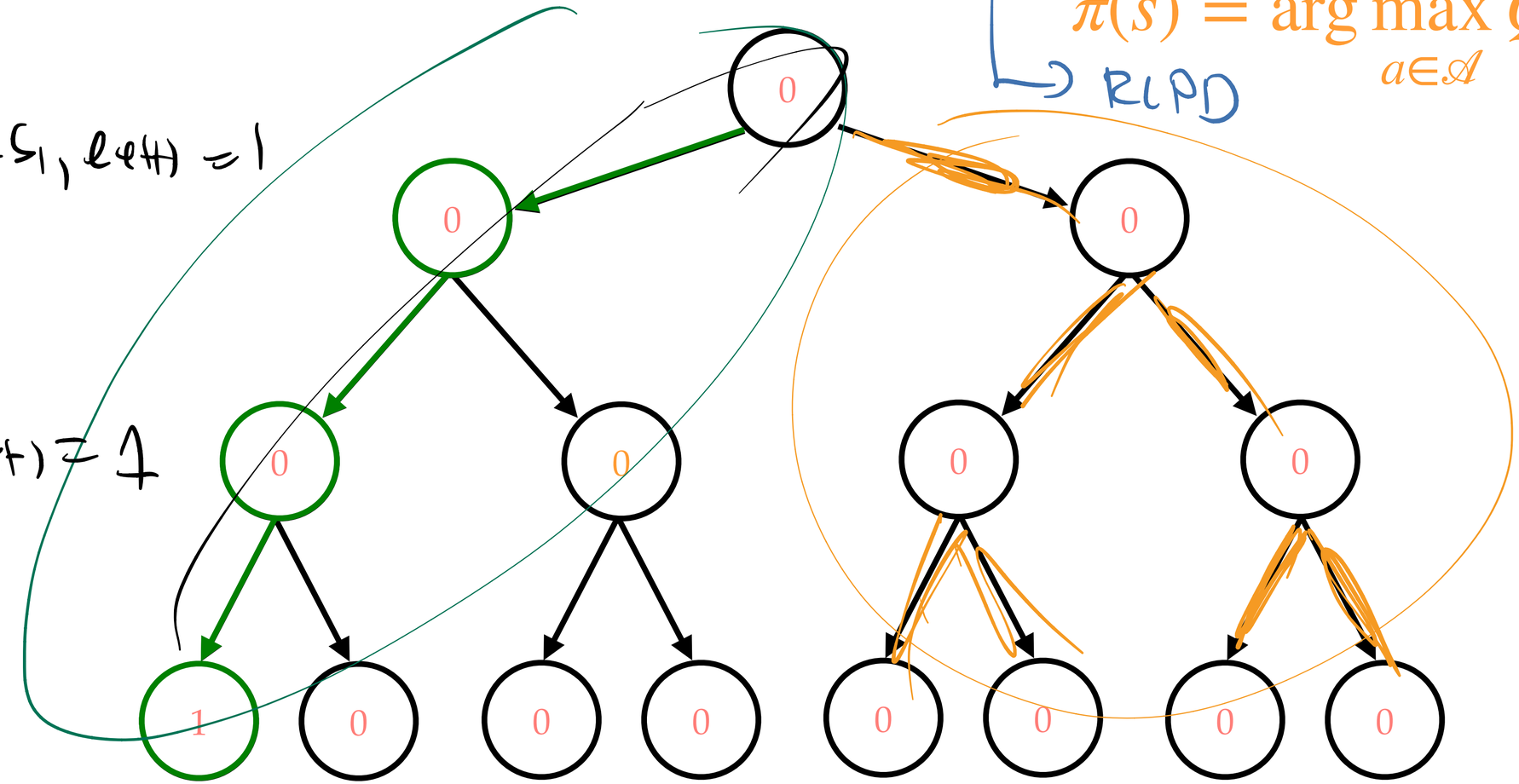
$\forall s \in \mathcal{S}, a \in \mathcal{A}, \hat{Q}(s,a) = 0$

$\pi(s) = \arg\max_{a \in \mathcal{A}} \hat{Q}(s,a)$

→ RLPD

$Q^*(s_1, left) = 1$

$Q^*(s_1, left) = 1$



Outline for Today

1. What makes inverse RL sample-inefficient?

***A**: Repeatedly solving a global exploration (RL) problem.*

2. Are best responses required for solving the IRL game?

***A**: Actually, competing with the expert is "all you need".*

3. What algorithms can we use in our new reduction?

***A**: A wide variety of sample-efficient "local search" algorithms.*

↳ MMPI, PSDP, HyQ, Agnostic SysID

If ⏱: PSDP Competes Against Policies Covered by μ

$$\pi_H = \arg\max_{\pi \in \Pi} \mathbb{E}_{s_H \sim \mu}[\mathbb{E}_{a_H \sim \pi}[f_t(s_H, a_H)]]$$

For $h \in [H-1, H-2, \ldots, 1]$:

$$\pi_h = \arg\max_{\pi \in \Pi} \mathbb{E}_{s_h \sim \mu}[\mathbb{E}_{a_h \sim \pi}[f_t(s_h, a_h) + \mathbb{E}_{T(s_h, a_h)}[V_t^{\pi_{h+1:H}}(s_{h+1})]]]$$

Lemma: Assume that at each time-step $h \in [H]$, we perform policy optimization up to ε-optimality:

$$\mathbb{E}_{s_h \sim \mu}[\mathbb{E}_{a \sim \pi'_h(s_h)}[Q_t^{\pi_{h+1:H}}(s_h, a)] - \mathbb{E}_{a \sim \pi_h(s_h)}[Q_t^{\pi_{h+1:H}}(s_h, a)]] \leq \varepsilon H$$

Then,

$$J(\pi'_{1:H}, f_t) - J(\pi_{1:H}, f_t) \leq \mathcal{O}((\varepsilon + \mathbb{D}_{TV}(\mu, \rho_{\pi'})) \cdot H^2)$$

Proof: We proceed via the PDL:

$$J(\pi'_{1:H}, f_t) - J(\pi_{1:H}, f_t) = \sum_h^H \mathbb{E}_{s_h, a_h \sim \pi'_{1:h}} \left[Q^{\pi_{h+1:H}}(s_h, a_h) - \mathbb{E}_{a \sim \pi_h(s_h)}[Q^{\pi_{h+1:H}}(s_h, a_h)] \right]$$

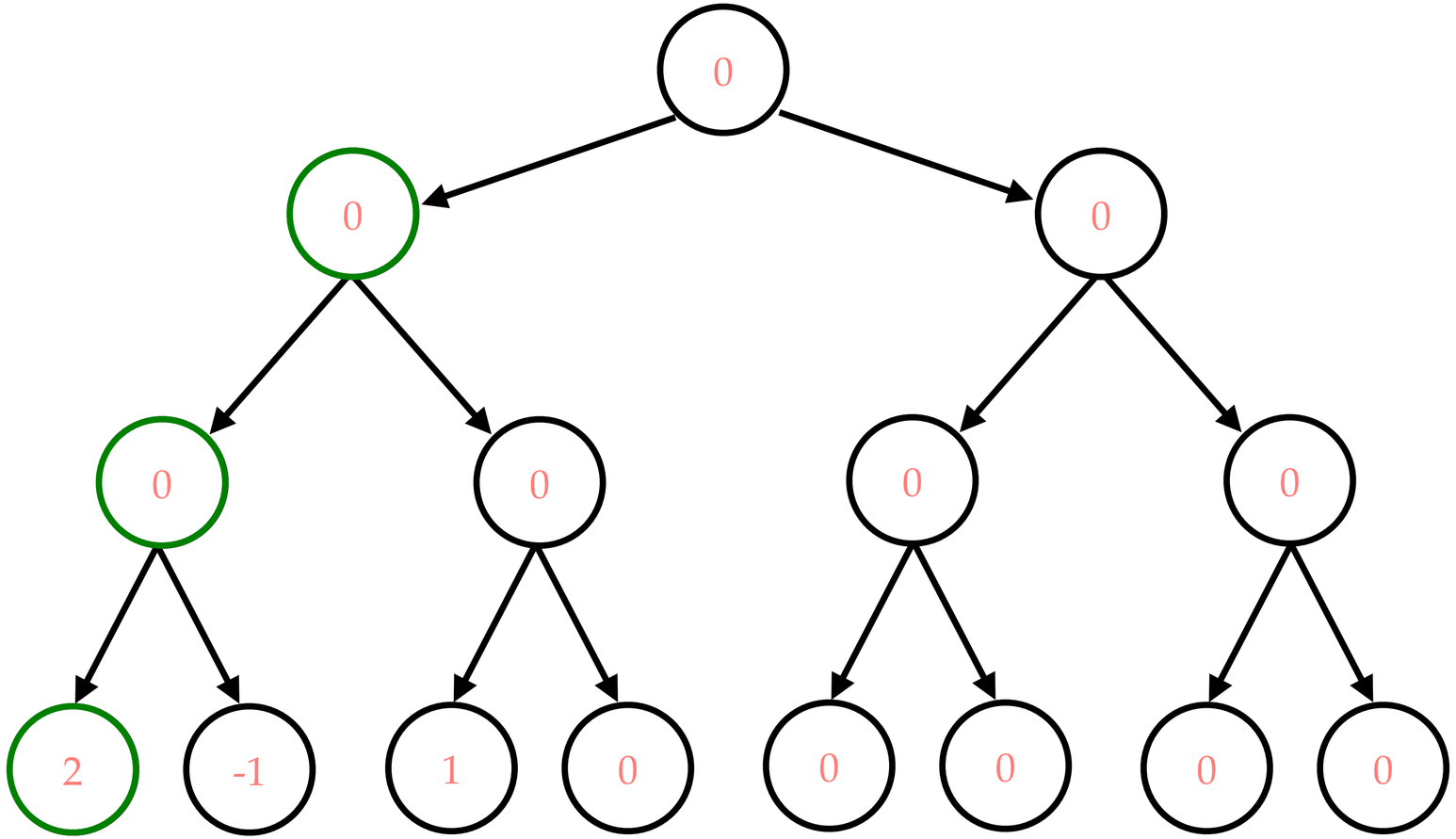
$$\leq \sum_h^H \mathbb{E}_{s_h, a_h \sim \mu} \left[Q^{\pi_{h+1:H}}(s_h, a_h) - \mathbb{E}_{a \sim \pi_h(s_h)}[Q^{\pi_{h+1:H}}(s_h, a_h)] \right]$$

$$+ H \cdot \mathbb{D}_{TV}(\mu_h, \rho_h^{\pi'})$$

$$\leq \sum_h^H (\varepsilon + \mathbb{D}_{TV}(\mu_h, \rho_h^{\pi'})) \cdot 2 \cdot (H - h)$$

$$\leq (\varepsilon + \mathbb{D}_{TV}(\mu, \rho_{\pi'})) \cdot H^2$$

If ⏱️: What μ should we reset to when $\pi_E \notin \Pi$?



Widen the baseline distribution μ to cover $\pi^\star \in \Pi$,
potentially by using suboptimal / offline data!