Oracle Efficient Reinforcement Learning for Max Value Ensembles

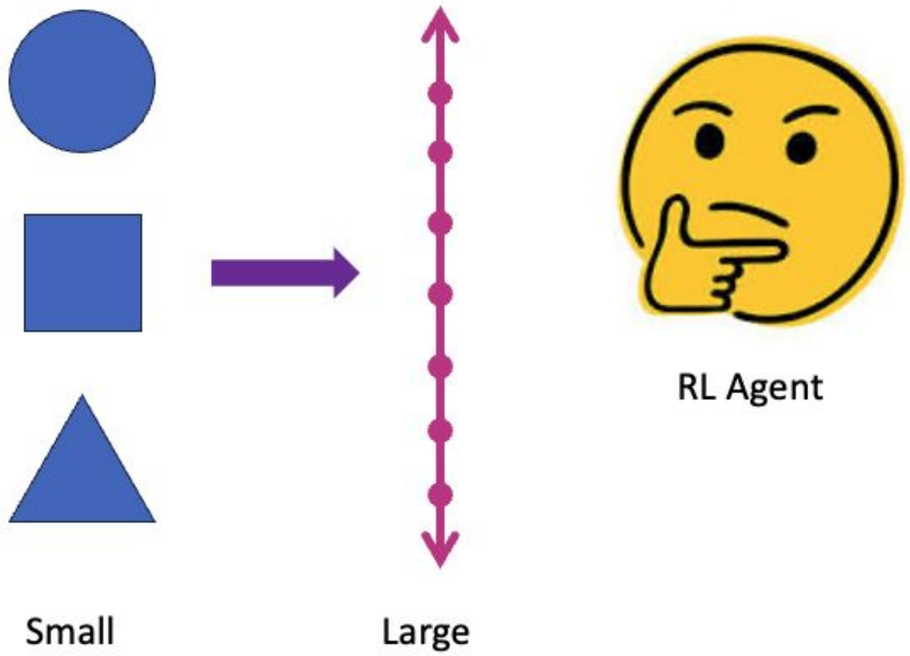
University of Pennsylvania

Marcel Hussing, Michael Kearns, Aaron Roth, Sikata Sengupta, Jessica Sorrell

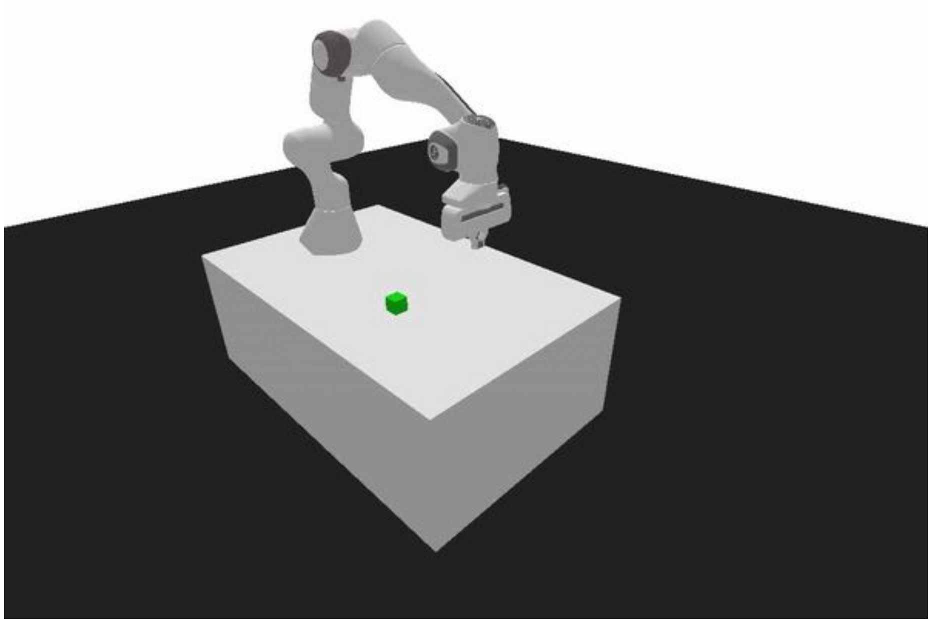




Motivation



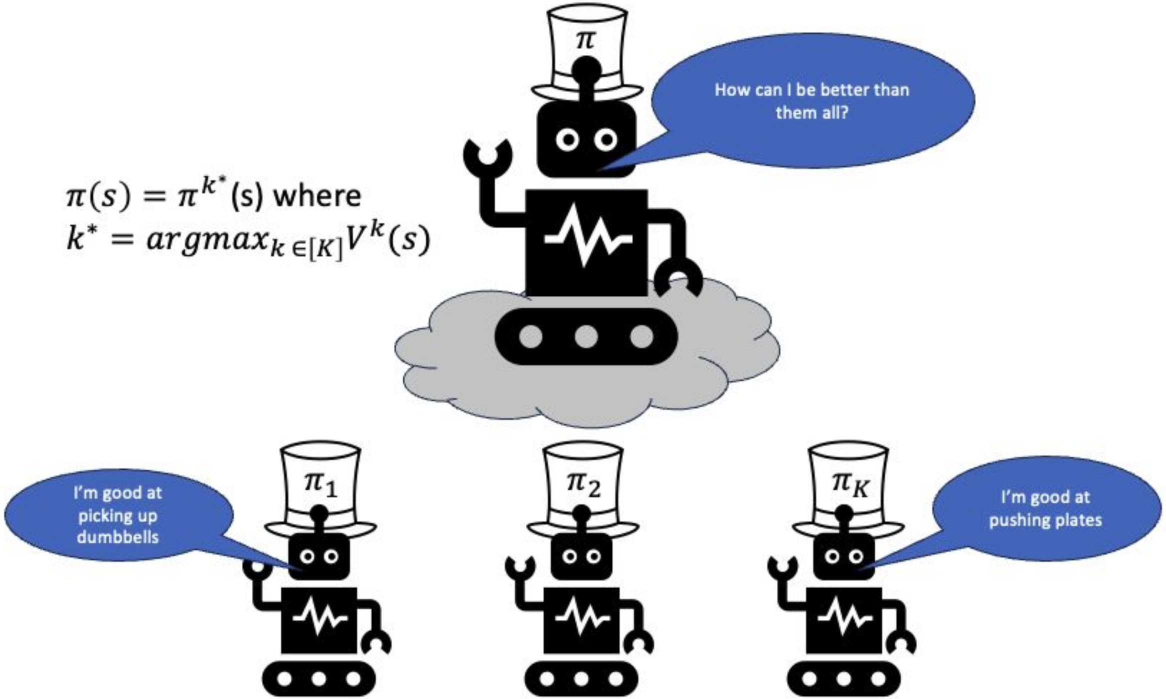
Motivation



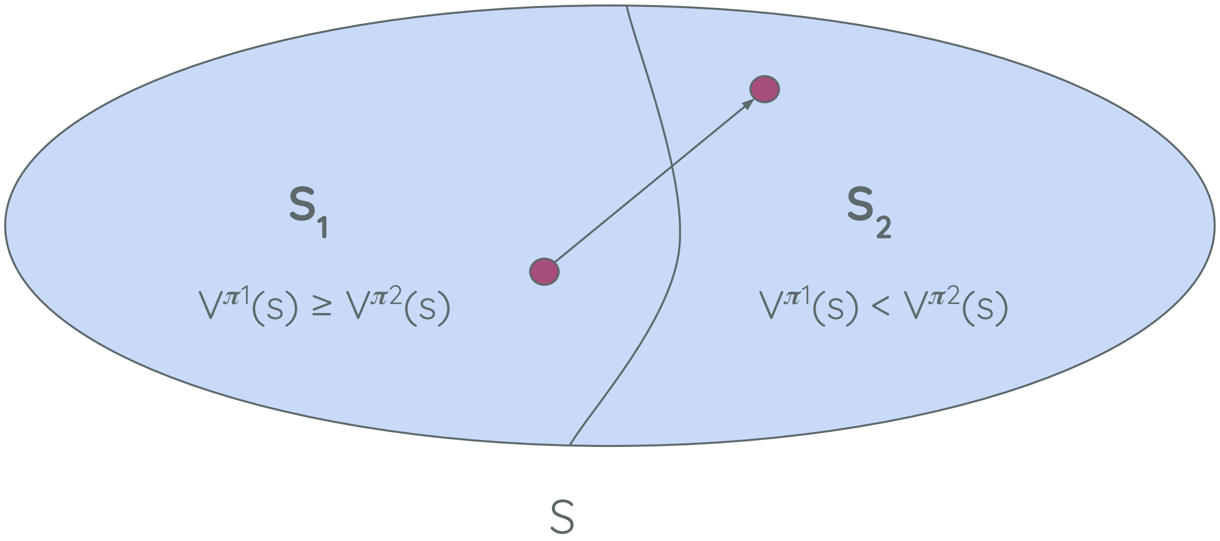
https://uwaterloo.ca/scholar/ajlobbez/reinforcement-learning-robotic-control

Warm Up (Known Value Functions)

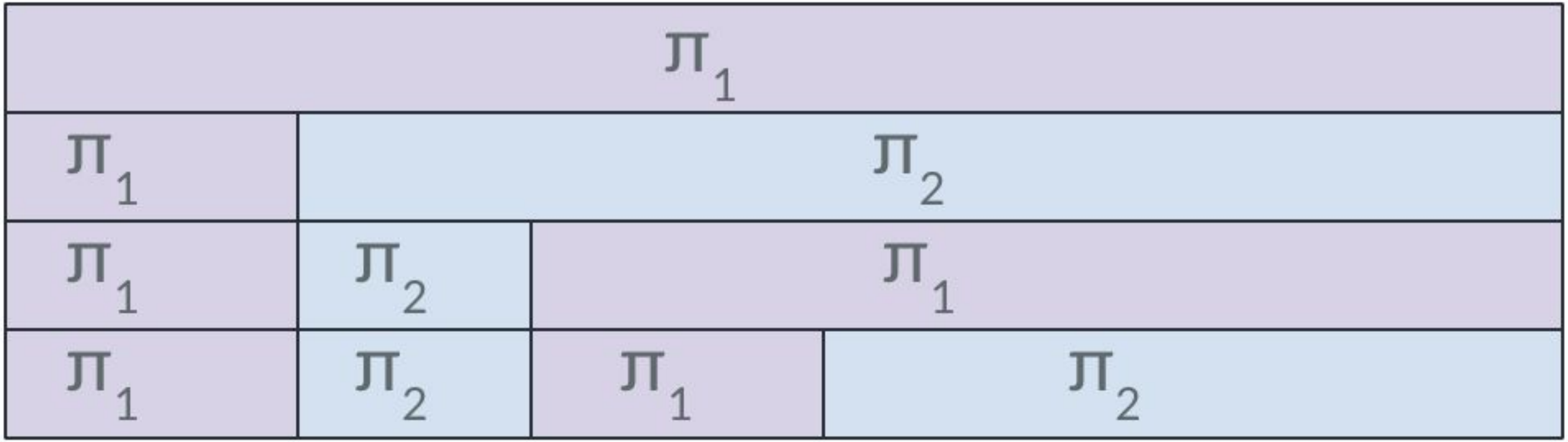
$\pi(s) = \pi^{k^*}(\text{s})$ where
$k^* = argmax_{k \in [K]} V^k(s)$



Intuition



Intuition



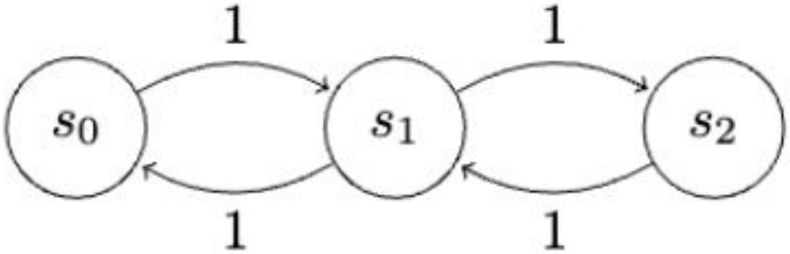
“The Problem of RL” → Batch ERM

ERM Oracle Access

- Regression oracle access insufficient to learn optimal policy [Golowich, Moitra, Rohatgi, 2024, *Exploration is Harder than Prediction...*]

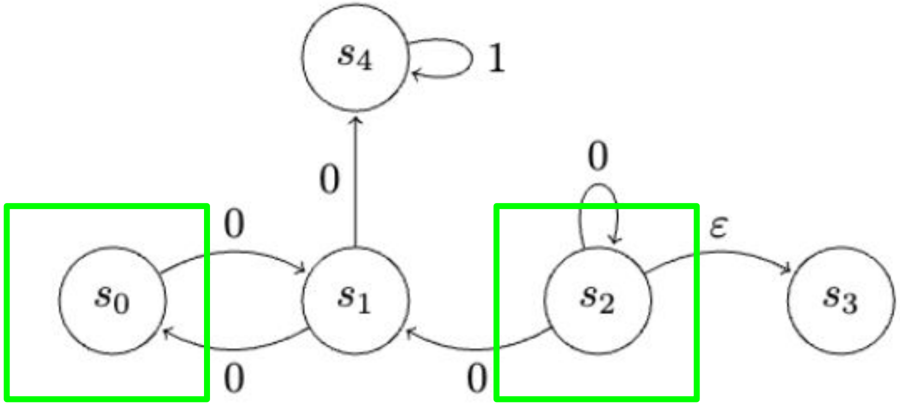
$$\mathrm{E}_{s\sim\mu}[(\hat{V}_h^{\pi}(s) - V_h^{\pi}(s))\hat{\ }2] \leq \alpha$$

Observation 1



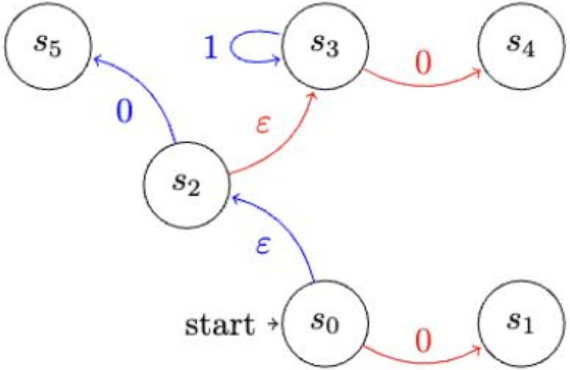
(a) MDP in which two policies going either only left or right obtain low return but max-following them would be optimal.

Observation 2



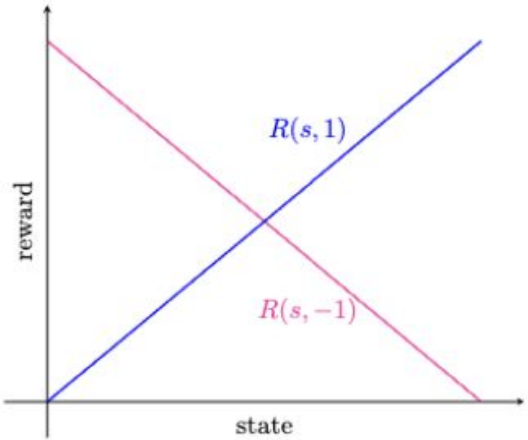
(b) MDP with $\mathcal{A} = \{\mathsf{right}, \mathsf{left}, \mathsf{up}\}$ where starting from s_2, max-following is far worse than optimal and starting from s_0, different max-following policies have different values (depending on tie-breaking).

Observation 3



(a) MDP where small value approximation errors at s_0 hinder max-following. Arrows representing transition dynamics are color-coded red to indicate actions taken by π^0 and blue to indicate actions taken by π^1.

Observation 4



(b) MDP where the max-following value function is piecewise linear, but constituent policy's values are affine functions of the state for fixed actions.

Approximate Tie-Breaking Policy Class

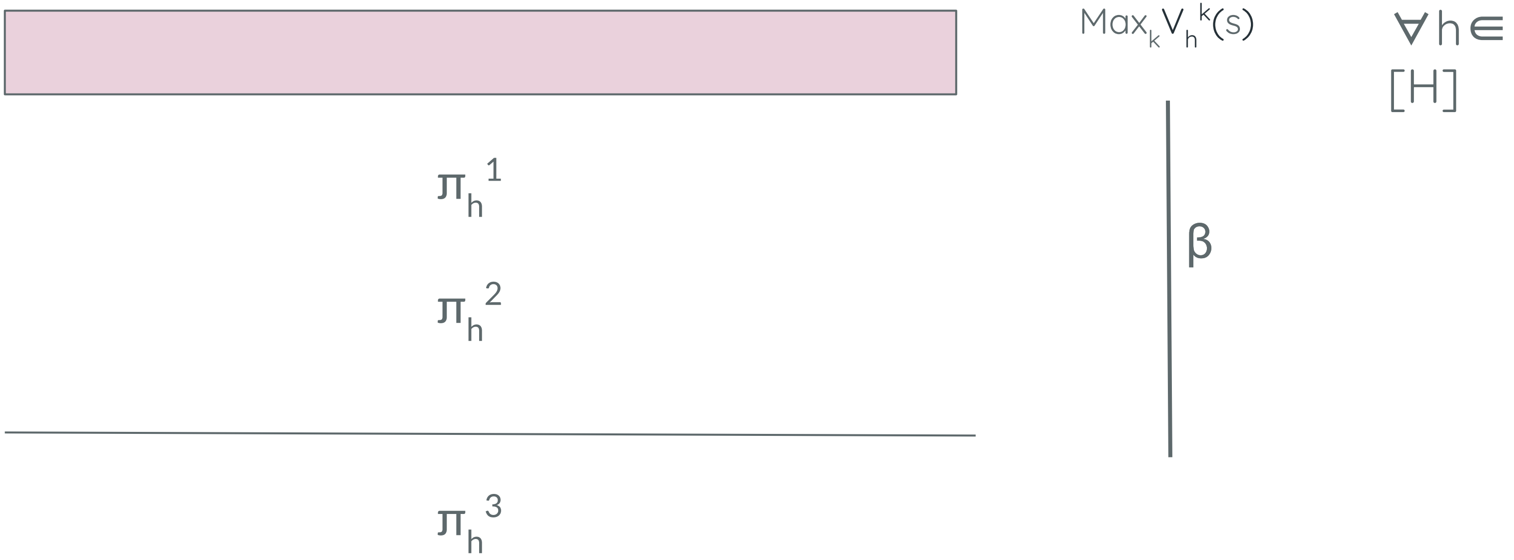
Approximate max-following policies: We define a set of β-good policies at state $s \in \mathcal{S}$ and time $h \in [H]$, selected from a set Π^k, as follows.

$$T_{\beta,h}(s) = \{\pi \in \Pi^k : V_h^\pi(s) \geq \max_{k \in [K]} V_h^k(s) - \beta\}.$$

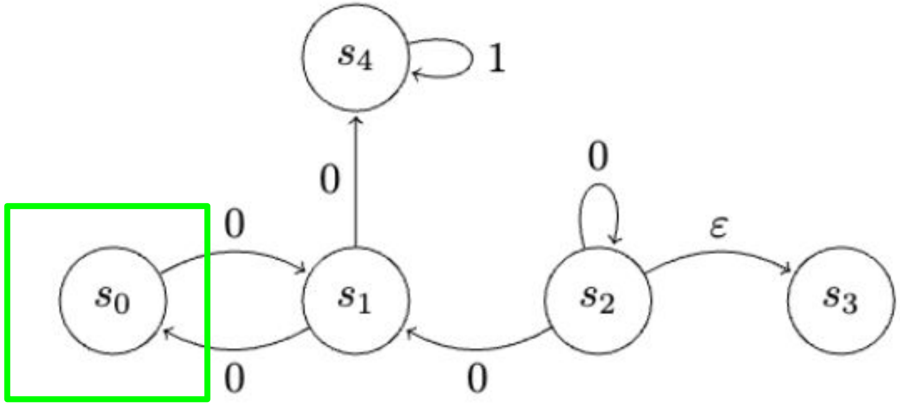
Then we define the set of approximate max-following policies for Π^k to be

$$\Pi_\beta^{k^*} = \{\pi : \forall h \in [H], \forall s \in \mathcal{S}, \pi_h(s) = \pi_h^t(s) \text{ for some } \pi^t \in T_{\beta,h}(s)\}.$$

Approximate Tie-Breaking Policy Class (Π_β^{k*})



Observation 2



(b) MDP with $\mathcal{A} = \{\mathsf{right}, \mathsf{left}, \mathsf{up}\}$ where starting from s_2, max-following is far worse than optimal and starting from s_0, different max-following policies have different values (depending on tie-breaking).

Algorithm

Algorithm 1 MaxIteration$_\alpha^{\mathcal{M}}(\Pi^k)$

for $h \in [H]$ **do**
    for $k \in [K]$ **do**
        let μ_h be the distribution sampled by executing the following procedure:
            sample a starting state $s_0 \sim \mu_0$
            for $i \in [h]$ **do**
                $s_{i+1} \sim P(\,\cdot \mid s_i, \pi^{\mathrm{argmax}_k \hat{V}_i^k(s_i)}(s_i))$
            end for
            output s_h
        $\hat{V}_h^k \leftarrow \mathcal{O}_\alpha^k(\mu_h, h)$
    end for
end for
return policy $\hat{\pi} = \{\hat{\pi}_h\}_{h \in [H]}$ where $\hat{\pi}_h(s) = \pi^{\mathrm{argmax}_{k \in [K]} \hat{V}_h^k(s)}(s)$

Theoretical Results

Value of π' from *MaxIteration*

$O(\varepsilon)$ {

Value of worst π from Π_β^{k*}

$O(\varepsilon)$ {

Values of base policy class Π^k

Theorem 3.1

[**Theorem 3.1: MaxIteration provides algorithm competitive with worst-case of benchmark class**] For any $\varepsilon \in (0, 1]$, any MDP $\mathcal{M}$ with starting state distribution μ_0, any episode length H, and any K policies Π^k defined on $\mathcal{M}$, let $\alpha \in \Theta(\frac{\varepsilon^3}{KH^4})$ and $\beta \in \Theta(\frac{\varepsilon}{H})$. Then $\textbf{MaxIteration}_{\alpha}^{\mathcal{M}}(\Pi^k)$ makes $O(HK)$ oracle queries and outputs $\hat{\pi}$ such that

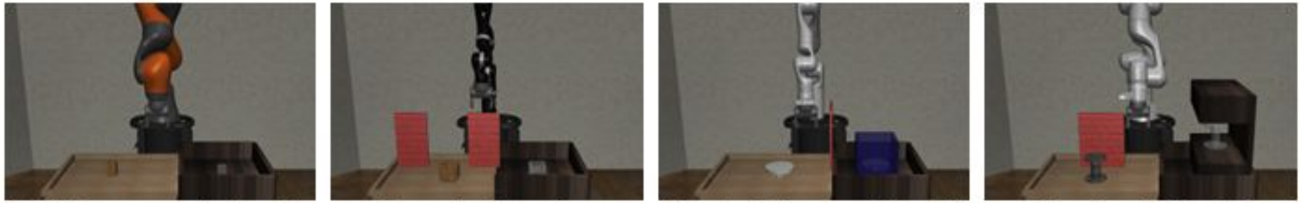
$$\mathop{\mathbb{E}}_{s_0 \sim \mu_0}\left[V^{\hat{\pi}}(s_0)\right] \geq \min_{\pi \in \Pi_{\beta}^{k*}} \mathop{\mathbb{E}}_{s_0 \sim \mu_0}\left[V^{\pi}(s_0)\right] - O(\varepsilon).$$

Lemma 4.1

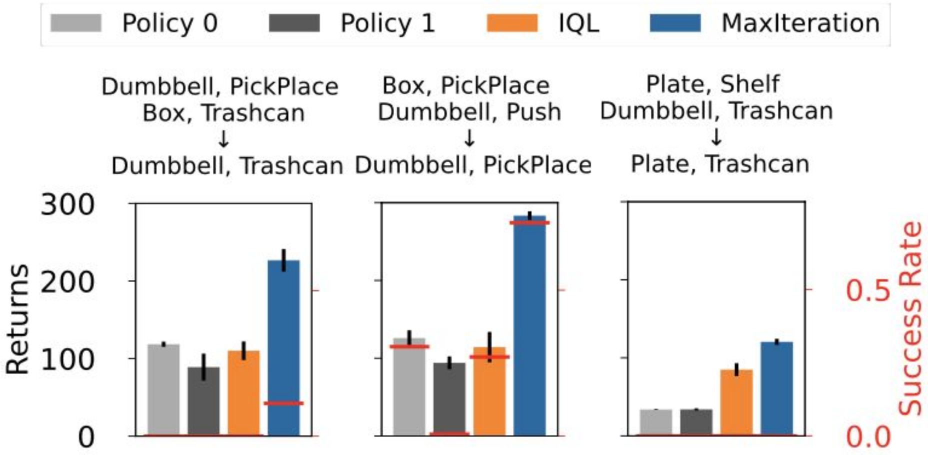
[**Lemma 4.1: Worst approximate max-following policy competes with best fixed policy**] For any $\varepsilon \in (0, 1]$ and any episode length H, let $\beta \in \Theta(\frac{\varepsilon}{H})$. Then for any MDP $\mathcal{M}$ with starting state distribution μ_0, and any K policies Π^k defined on $\mathcal{M}$,

$$\min_{\pi \in \Pi_{\beta}^{k*}} \mathbb{E}_{s_0 \sim \mu_0} \left[V^{\hat{\pi}}(s_0) \right] \geq \max_{k \in [K]} \mathbb{E}_{s_0 \sim \mu_0} \left[V^{k}(s_0) \right] - O(\varepsilon).$$

Experiments



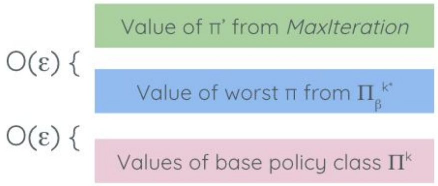
⟨IIWA, box, no_obstacle, pick-and-place⟩ ⟨Jaco, hollow_box, object_door, push⟩ ⟨Gen3, plate, goal_wall, trash_can⟩ ⟨Panda, dumbbell, object_wall, shelf⟩

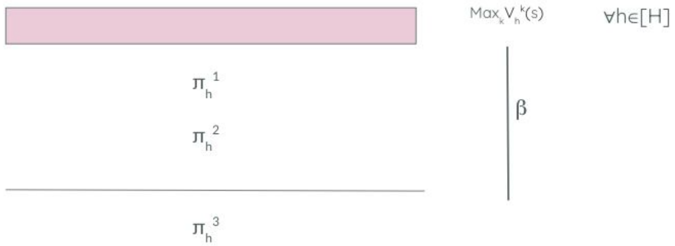


Recap

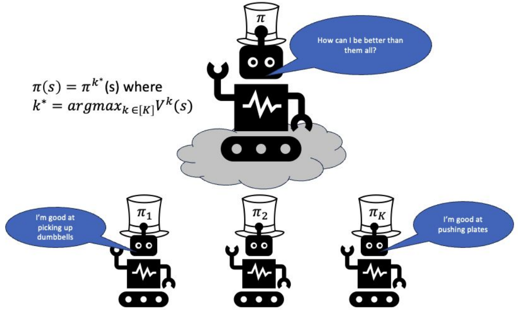
$$E_{s\sim\mu}[(\hat{V}_h^\pi(s) - V_h^\pi(s))\text{^}2] \leq \alpha$$

1: Max-Iteration Algorithm (oracle-efficient)





2: Apx Max-Following Policy Class



3: Superior to base policy class (w.h.p.)

References

1. Ching-An Cheng, Andrey Kolobov, and Alekh Agarwal. Policy improvement via imitation of multiple oracles. *Advances in Neural Information Processing Systems*, 33:5587–5598, 2020.
2. Xuefeng Liu, Takuma Yoneda, Chaoqi Wang, Matthew Walter, and Yuxin Chen. Active policy improvement from multiple black-box oracles. In *International Conference on Machine Learning*, pages 22320–22337. PMLR, 2023.
3. André Barreto, Shaobo Hou, Diana Borsa, David Silver, and Doina Precup. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117 (48):30079–30087, 2020. doi: 10.1073/pnas.1907370117.
4. Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Exploration is harder than prediction: Cryptographically separating reinforcement learning from supervised learning. *arXiv preprint arXiv:2404.03774*, 2024.
5. Nataly Brukhim, Elad Hazan, and Karan Singh. A boosting approach to reinforcement learning. *Advances in Neural Information Processing Systems*, 35:33806–33817, 2022.
6. Jorge A. Mendez, Marcel Hussing, Meghna Gummadi, and Eric Eaton. Composuite: A compositional reinforcement learning benchmark. In *1st Conference on Lifelong Learning Agents*, 2022.

Thank You!