On the Power of Pre-training for Generalization in RL: Provable Benefits and Hardness

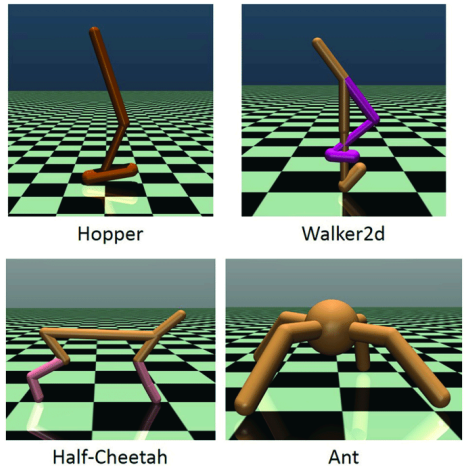
Haotian Ye[*], *Xiaoyu Chen*[*], *Liwei Wang, Simon S. Du*

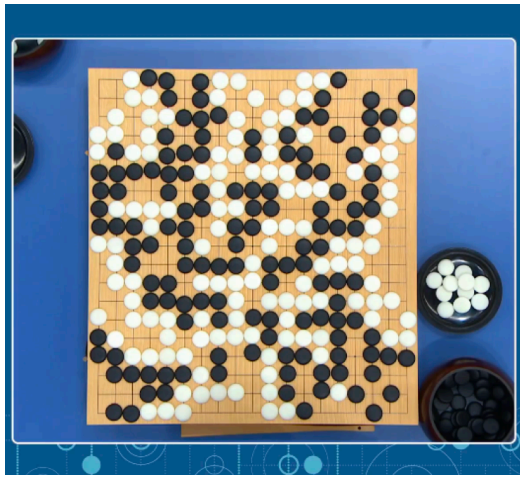
ICML Oral Presentation

Poster 6 (Exhibit Hall 1 #430)

Standard Reinforcement Learning

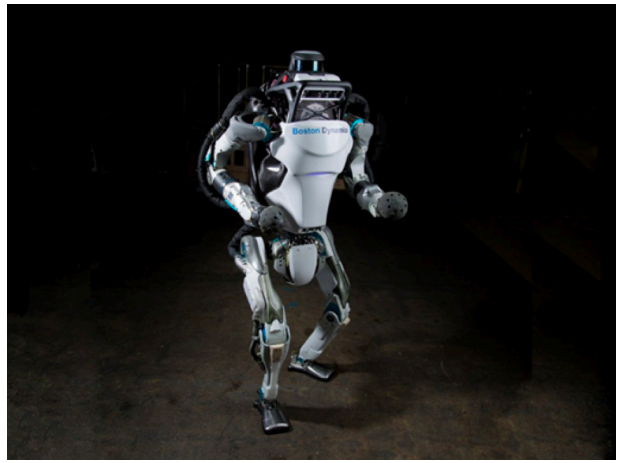
- Setting: An agent interacts with an unknown environment M.
- Goal: Maximize expected cumulative rewards.
- We consider the Markov Decision Processes environments, where the state space and the action space is finite.

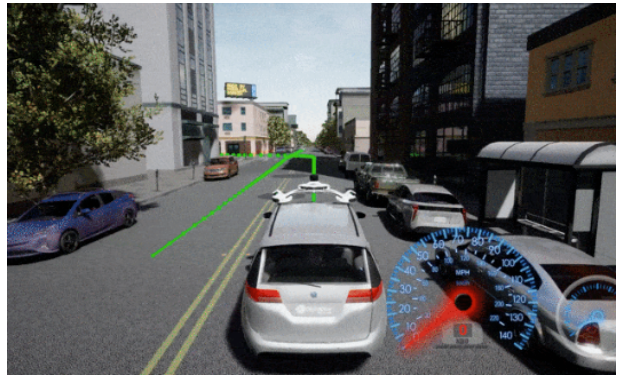




Beyond Standard RL: Generalization in RL

- Motivation: In practice, tests environments are *different* from training environments.
- We hope the agent can learn meaningful skills in the training stage and be robust to the variation during the test stage.
- Example: autonomous driving, robotics, health care …





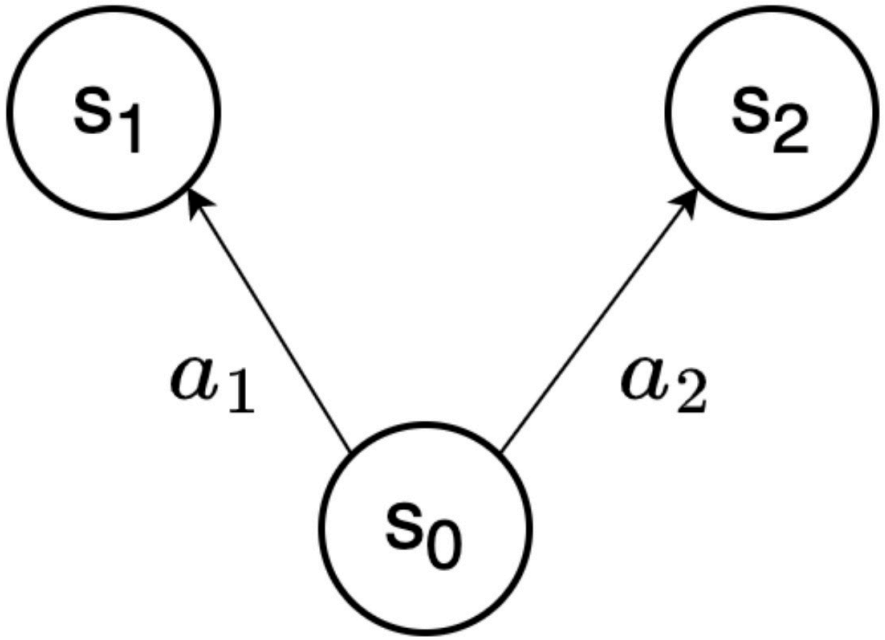
Generalization in RL v.s. Generalization in Supervised learning

Generalization in supervised learning	Generalization in reinforcement learning
Setting: data distribution $\mathbb{D}$, loss function l	**Setting**: MDP distribution $\mathbb{D}$
Training: collect a dataset $(x_i, y_i)_{i=1}^n$ sampled from $\mathbb{D}$, find $\hat{h}$ that performs well on the set	**Training**: collect a MDP set $(\mathcal{M}_i)_{i=1}^n$ sampled from $\mathbb{D}$, find $\hat{\pi}$ that performs well on the set
Testing: sample a test dataset, evaluate the performance of $\hat{h}$ on the set	**Testing**: sample a test MDP $\mathcal{M}$ from $\mathbb{D}$, interacts with the MDP according to policy $\hat{\pi}$
measure: optimality in expectation $\mathbb{E}_{\mathbb{D}}\left[l\left(\hat{h}(x,y)\right)\right] - \min_h \mathbb{E}_{\mathbb{D}}[l(h(x,y))]$	**measure**: optimality in instance $\mathbb{E}_{\mathcal{M}\sim\mathbb{D}}\left[\max_\pi V_{\mathcal{M}}^{\pi} - V_{\mathcal{M}}^{\hat{\pi}}\right]$

Instance optimality is required in RL generalization!

Impossibility of Instance Optimality

- Unfortunately, without interacting with the test MDP, it is impossible to pursue an instance optimal policy.
- Example: $\mathbb{D}$ is a Bernoulli distribution.
  - Both MDPs have 3 states and 2 actions.
  - In $\mathcal{M}_1$, $R(a_1) = 1, R(a_2) = 0$.
  - In $\mathcal{M}_2$, $R(a_1) = 0, R(a_2) = 1$.
- A better formulation is required!



Problem Formulation of RL Generalization (Informal)

- Measure: Compared to directly interact with $\mathcal{M}$, can pre-training help reduce Reg?
- Intuition: Training on $(\mathcal{M}_i)_{i=1}^n$ provides information of $\mathbb{D}$.
- Question: How much can the information obtained from pretraining help *reduce* the K episode regret suffered during the test stage?

Setting: MDP distribution $\mathbb{D}$

Pre-training: collect a MDP set $(\mathcal{M}_i)_{i=1}^n$ from $\mathbb{D}$, pretrained a model $\mathcal{F}$

Fine-tuning: sample a test MDP $\mathcal{M}$ from $\mathbb{D}$, fine-tune in the test MDP for K episodes

measure: $\mathrm{Reg} = \sum_{k=1}^{K} \left[\max_{\pi} V_{\mathcal{M}}^{\pi} - V_{\mathcal{M}}^{\mathcal{F}(\mathrm{history})} \right]$

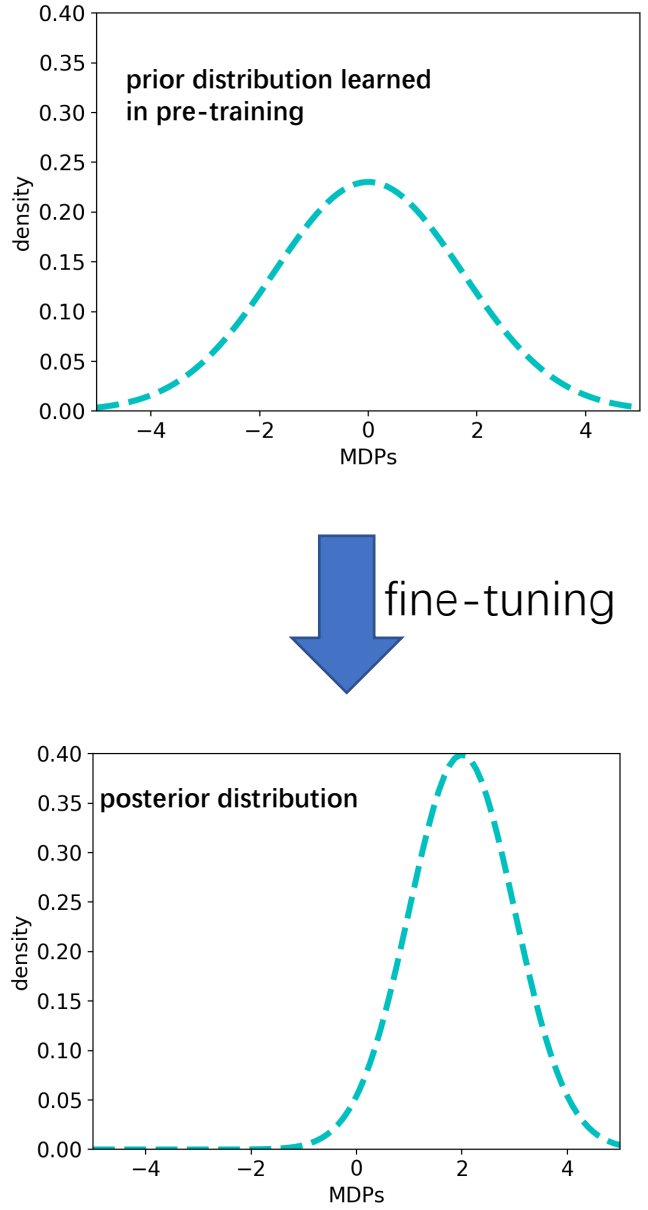
Negative Result (Informal)

- Theorem:

  Knowing the distribution is useless up to a constant, i.e.

$$\liminf_{K\to\infty} \frac{\mathrm{Reg}_K(\mathbb{D}, \mathcal{A}(\mathbb{D}, K))}{\mathrm{Reg}_K(\mathbb{D}, \hat{\mathcal{A}}(K))} \geq c_0.$$

- Intuition:
  - We can *at most* learn the entire distribution $\mathbb{D}$, but cannot know the test environment $\mathcal{M}$ *exactly*.
  - After more interactions with $\mathcal{M}$, the value of pretraining *decreases* relatively.
  - Consequently, the helpfulness of pretraining is bounded asymptotically.



Positive Result: Fine-tuning Algorithm

- On the contrary, the improvement in the non-asymptotic setting is achievable.
- Algorithm PCE: collect a minimum near-optimal policy set Π that generalizes to most of MDPs in $\mathbb{D}$.
- Advantages:
  - Reduce regret dependence from $|S|, |A|$ to $|\Pi|$, which depends on the complexity of $\mathbb{D}$.
  - Regret can still be small even when the MDP support is *infinite*.
  - The idea of finding a policy covering set is could be helpful in practice.

Algorithm PCE (Policy Collection-Elimination)

Pre-training: find a policy set $\widehat{\Pi}$ covering distribution $\mathbb{D}$

Fine-tuning: given a sampled MDP $\mathcal{M}^*$, find a policy in $\widehat{\Pi}$ for $\mathcal{M}^*$ by elimination

Summary & Takeaways

Allow fine-tuning	Setting	Optimality in instance	Optimality in expectation
✗	Non-asymptotic	Linear in K	$\mathrm{Poly}(S,A,H)\sqrt{K}$
✓	Asymptotic	$\mathrm{Poly}(S,A,H)\sqrt{K}$	-
✓	Non-asymptotic	$O(C(\mathbb{D})\sqrt{K})$	-

- When fine-tuning is not allowed, instance optimality is impossible in RL generalization.
- Even when fine-tuning is allowed, the regret can not be substantially reduced in the asymptotic setting.
- However, when K is fixed (non-asymptotic), the dependence of $|S|, |A|$ can be removed using the idea of policy covering.

Thanks!

ICML Oral Presentation

Poster 6 (Exhibit Hall 1 #430)