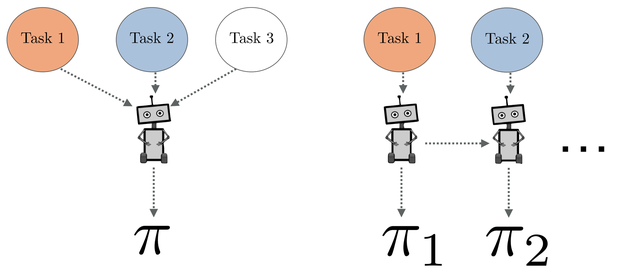
Policy and Value Transfer in Lifelong Reinforcement Learning

Yuu Jinnai, David Abel, Yue Guo, George Konidaris, Michael L. Littman
{yuu_jinnai,david_abel,yue_guo,george_konidaris,michael_littman}@brown.edu
Department of Computer Science, Brown University

Goal

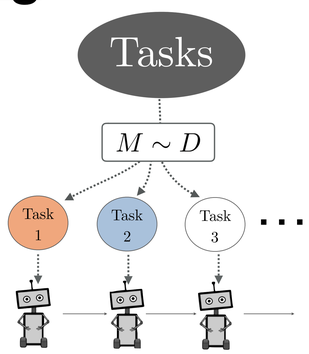
Understand knowledge transfer in lifelong RL.



1. Optimal Fixed Policy
2. Informed Transfer for RL

Lifelong Reinforcement Learning *[1]*

```
repeat:
1. Sample a task from a
   distribution: M ~ D.
2. Solve the sampled
   task M.
```



Objective 1: Find the policy that maximizes the expected performance over the distribution:

$$\arg\max_{\pi \in \Pi} \mathop{\mathbb{E}}_{M \sim D} [V_M^{\pi}(s_0)]$$

Performance of π on task M.

Jumpstart Policies

	Π	$\mathcal{R} \sim D$	$G \sim D$
(Deterministic)	$\Pi_d : \mathcal{S} \mapsto \mathcal{A}$ *[2]*	Avg. MDP	Avg. V *[3]*
(Stochastic)	$\Pi_s : \mathcal{S} \mapsto \Pr(\mathcal{A})$	Avg. MDP	Avg. V
(Any policy)	$\Pi_b : \mathcal{S} \times \Pr(\mathscr{M}) \mapsto \mathcal{A}$	Belief MDP	Belief MDP *[4]*

Action Prior *[5]*

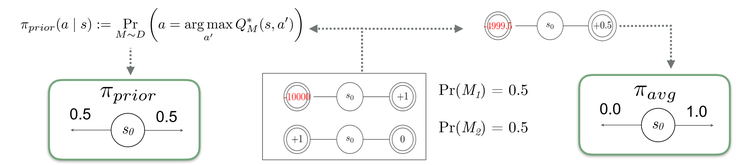
Probability the action is optimal:

$$\pi_{prior}(a \mid s) := \Pr_{M \sim D}\left(a = \arg\max_{a'} Q_M^*(s, a')\right)$$

π_{prior}: 0.5 — s_0 — 0.5

Average MDP

Choose optimal action in averaged MDP:



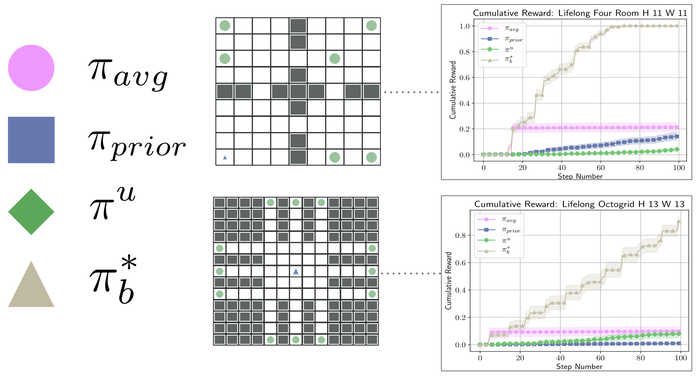
Theorem 1. *For a distribution of MDPs with R ~ D, Average MDP is an optimal fixed policy.* *[2]*

Theorem 2. *For a distribution of MDPs with R ~ D, the performance of Average MDP has a lower bound:*

Expected performance of Avg. MDP policy

$$\mathop{\mathbb{E}}_{M \sim D}\left[V_M^{\pi^*_{avg}}(s)\right] \geq \max_{M \in \mathscr{M}} \Pr(M) V_M^*(s)$$

Jumpstart Experiments



Informed Transfer for RL

PAC-MDP *[6]*

Sample complexity of many PAC-MDP algorithms depends on the of overestimation of the initial value function:

$$\tilde{O}\left(\sum_{(s,a) \in \mathcal{S} \times A} \max\{Q_{init}(s,a) - V^*(s), 0\}\right)$$

(Objective) Minimize the **overestimate**

(Constraint) Do not **underestimate:** $Q_{init}(s,a) \geq Q^*(s,a)$

(Our Solution) Initialize with: $Q_{max}(s,a) := \max_{M \in \mathscr{M}} Q_M^*(s,a)$

Proposed Algortihm: MaxQInit

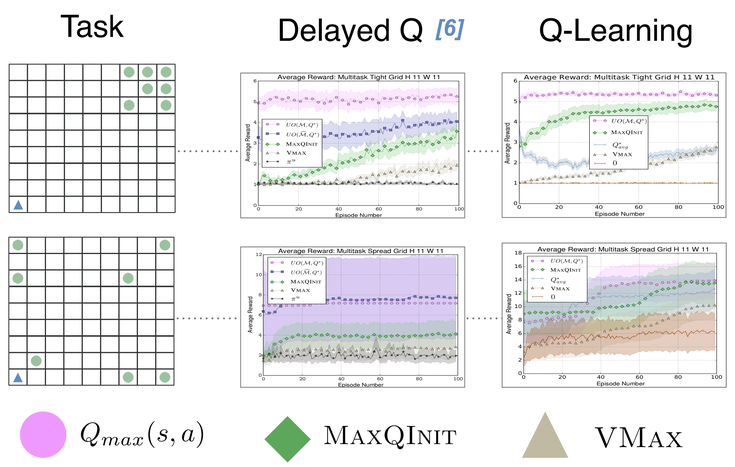
Approximate $Q_{max}(s,a)$ from *n* sampled MDPs:

$$\textsc{MaxQInit}(s,a) := \max_{M \in \{M_1, \ldots, M_n\}} \widehat{Q}_M(s,a)$$

Theorem 3. *For a given δ, after* $n \geq \frac{\ln(\delta)}{\ln(1 - p_{min})}$ *sampled MDPs, MaxQInit will retain optimism with probability 1 - δ:*

$$\forall_{s,a} : \textsc{MaxQInit}(s,a) \geq \max_{M \in \mathscr{M}} Q_M^*(s,a)$$

Transfer Experiments



References:
[1]: Brunskill & Li. Pac-inspired option discovery in lifelong reinforcement learning. 2014. *[2]: Ramachandran & Amir. Bayesian inverse reinforcement learning. 2007.* *[3]: Singh, Jakkola, Jordan. Learning without state-estimation in partially observable markovian decision processes 2003* *[4]: Astrom. Optimal control of markov processes with incomplete state information 1965* *[5]: Rossman & Ramamoorthy. What good are actions? 2012.* *[6]: Strehl et al.. Reinforcement learning in finite MDPs: PAC analysis. 2009.*