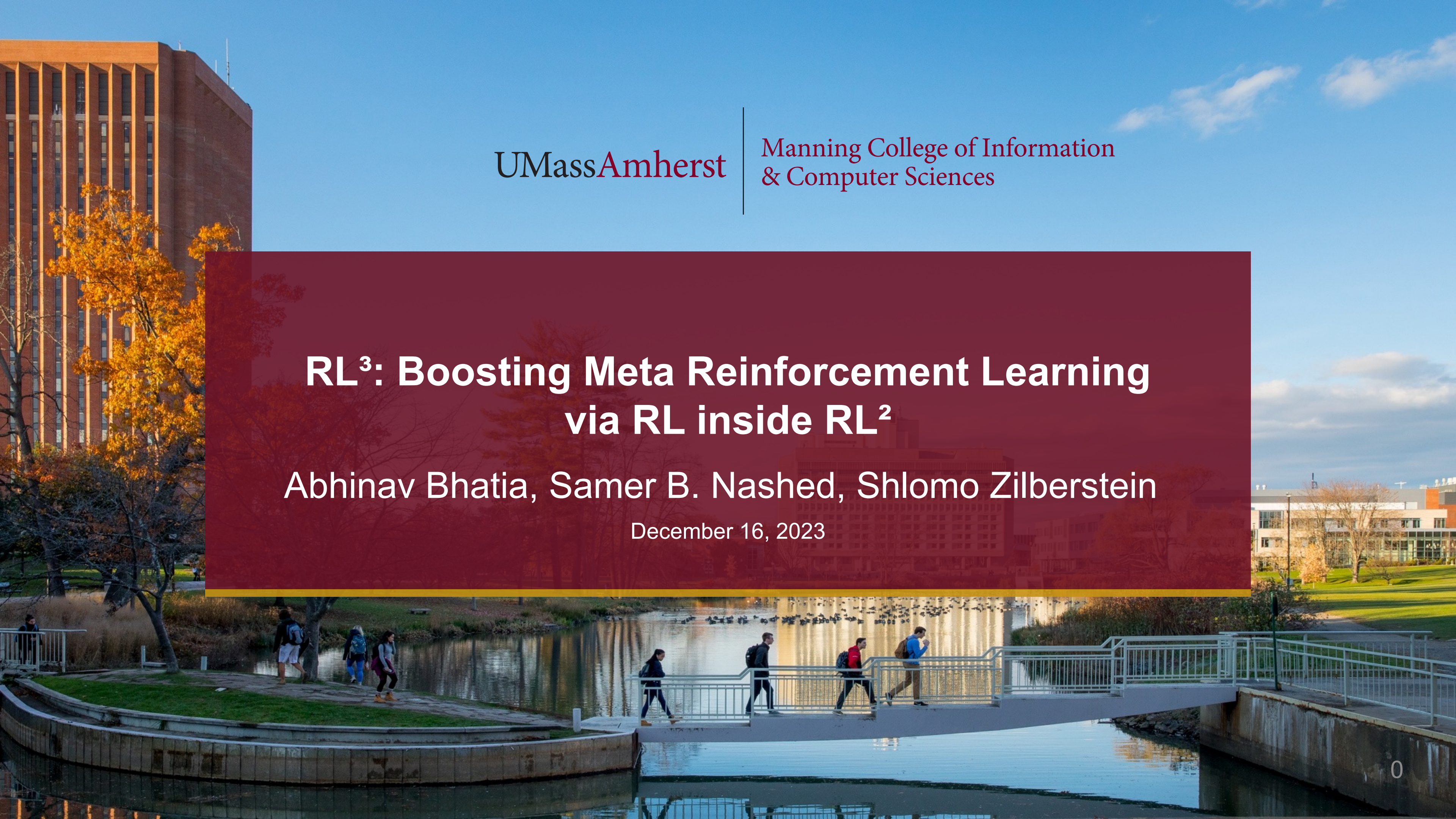


UMassAmherst | Manning College of Information & Computer Sciences

RL³: Boosting Meta Reinforcement Learning via RL inside RL²

Abhinav Bhatia, Samer B. Nashed, Shlomo Zilberstein

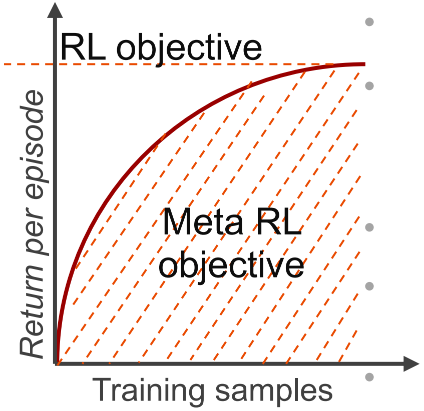
December 16, 2023

Overview

An RL algorithm: a mapping from experience data to actions.

Classic RL

- Given: an MDP
- Objective: learn a *state-to-action* mapping to maximize cumulative reward *per episode*.
- *Output: "Policy"*
- Classic RL involves value functions to distill data.
- Classic RL Pros & cons:
  - Data inefficient
  - General
  - Asymptotically optimal



Meta RL

- Given: a distribution of MDPs
- Objective: learn a *data-to-action* mapping to maximize cumulative reward *over entire interaction.*
- *"Meta-RL policy" or "Learned RL"*
- Learned RL *involves* a data-sequence model like an RNN.
- Learned RL Pros & cons:
  - Data-efficient (minimizes regret)
  - Poor OOD generalization
  - Poor long-context reasoning

RL³: Injects classic RL into Learned RL: Aids RNN with action-value estimates.

Meta Reinforcement Learning

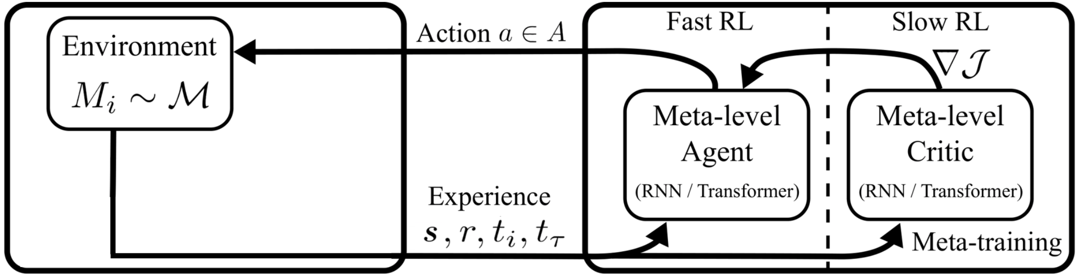
- Objective: Learn a data-to-action mapping to maximizes cumulative reward

$$\mathcal{J}(\theta) = \mathbb{E}_{M_i \sim \mathcal{M}} \left[\sum_{t=0}^{H} \gamma^t \mathbb{E}_{(s_t, a_t) \sim \rho_i^{\pi_\theta}} [R_i(s_t, a_t)] \right]$$

- As a meta-level Markov decision process:
  - Each meta-episode: sample a new MDP, or “task”, play for *H* interactions.
  - Optimal meta policy maximizes cumulative reward. ✅
  - Dynamics different across meta-episodes?
  - POMDP where hidden variable is the task identity. Also called *BAMDP*.
  - Beliefs over tasks capture history sufficiently.

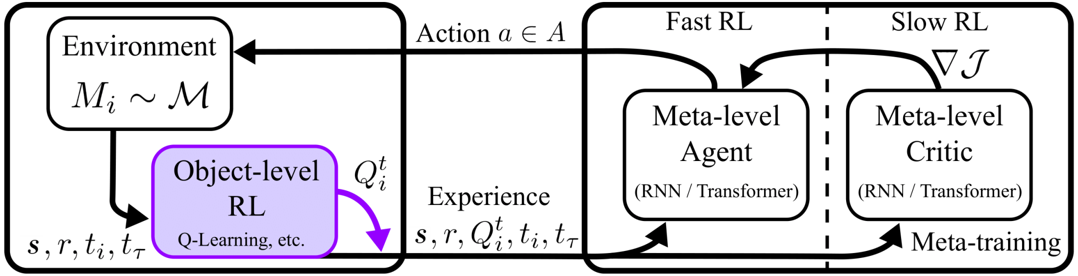
RL²: Fast RL using Slow RL (Duan et al. 2016)

- Meta-RL policy directly maps raw-data to actions using an RNN.
- Trained with standard “slow” deep RL.
- Note: Some approaches map data-to-beliefs first e.g., VeriBAD (Zintgraf et al., 2019)



RL³: Inject RL into RL²

- Insert RL subroutine: estimate Q*-values e.g., use Q-learning.
- Provide to meta-RL. Provide action-counts too.
- Meta-RL decides how to use.

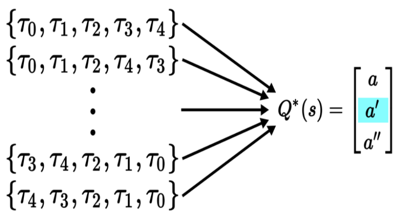


RL³: Inject RL into RL²: But Why?

Q-injection 💉 to improve OOD generalization and long-context reasoning?

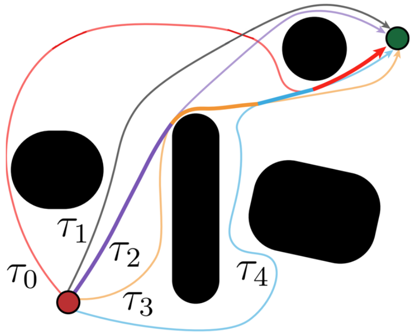
Inherent generality:

Key component in general-purpose RL



Summarization: Many-to-one mapping. Order is irrelevant.

Lossy, but "remembers" key details



Actionability: optimal policy given data.

Can ignore history, just exploit

Bottom line: Over time, data overwhelming, Q-estimates become more useful.

RL³: Inject RL into RL²: But Why?

Additional Reasons

Excellent task discriminators:

Rare for MDPs to have same Q-value function

Sufficient for Bayes optimal beliefs? Sometimes, yes.

For Bernoulli MAB, RL3 works without history.

Related to meta-value function:

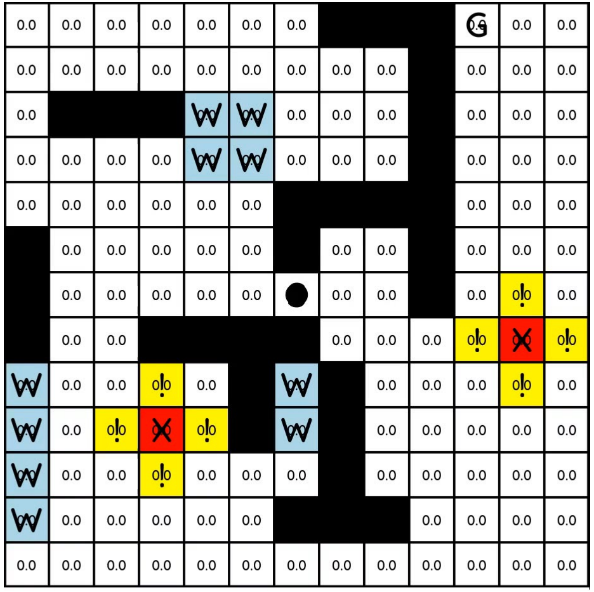
The Q* term appears in the meta-V* equation

$$\bar{V}^t(\bar{b}) = \arg\max_{a \in A}\Big[\sum_{M_i \in \mathcal{M}} \bar{b}(i) R_i(s,a) + \gamma \sum_{\bar{\omega} \in \bar{\Omega}} \bar{O}(\bar{\omega}|\bar{b},a) \sum_{M_i \in \mathcal{M}} \bar{b}'(i) \sum_{s' \in S} T_i(s,a,s')(Q_i^t(s') + \varepsilon_i(\tau))\Big]$$

UMassAmherst | Manning College of Information & Computer Sciences

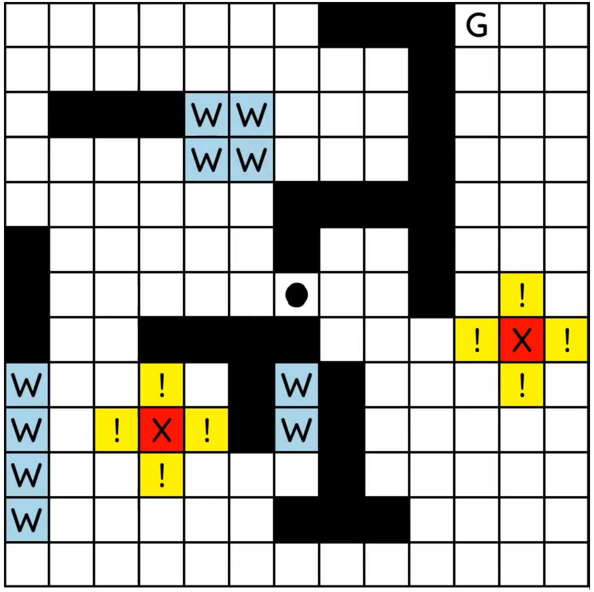
RL³ vs RL² - Gridworlds Results Demo

RL³



Action: ▯ Reward: 0.0

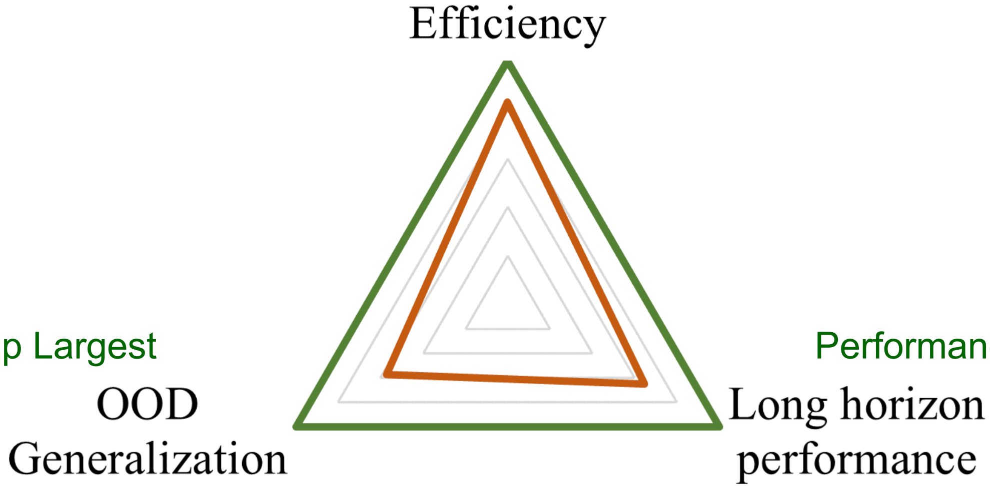
RL²



Action: ▯ Reward: 0.0

RL³ vs RL² - Gridworld Results

RL³ achieves lower regret.



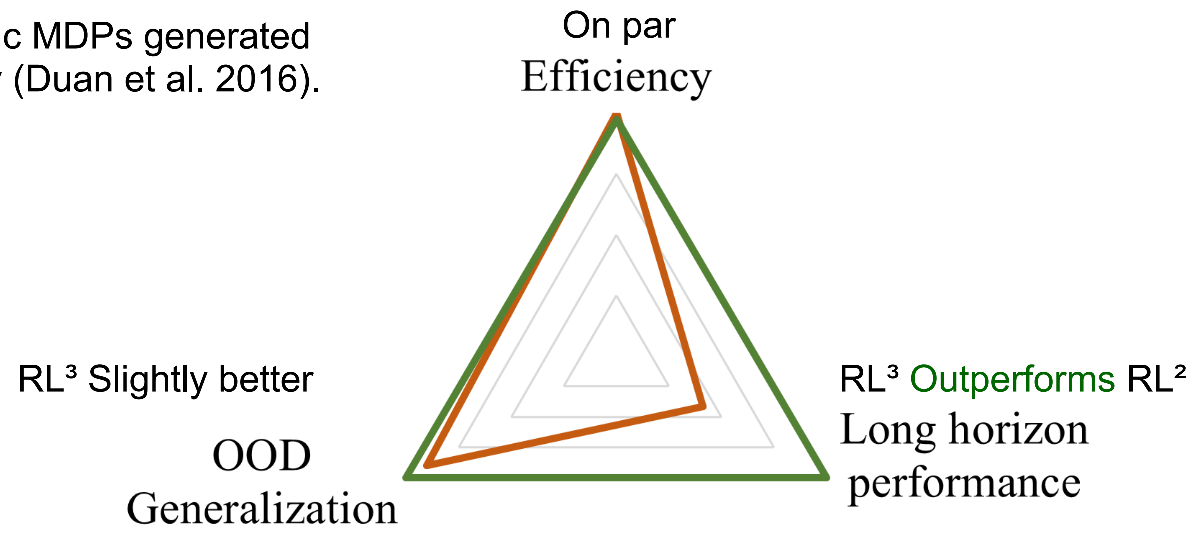
Gap Largest

Performance gap larger

RL³ with state-abstractions: *RL³-coarse:* 2x fast, 90% of RL³.

RL³ vs RL² - Random MDPs Results

- Stochastic MDPs generated randomly (Duan et al. 2016).



Conclusion

- We introduced RL^3, aiming to combine best of RL and RL^2 – to achieve good efficiency (minimize regret), better long-term reasoning, and better OOD generalization.
- Intuitions: Universality, summarization, actionability and with helps task identification. With time, data gets overwhelming, Q-estimates useful, almost sufficient.
- Key experimental takeaways:
  - RL^3 retains (and sometime improves) efficiency of RL^2 on all domains
  - RL^3 benefits with increase with horizon, distribution shift, and determinism
  - Injected Q-values can be imprecise, and still be useful.
- Future: extend this to continuous action space setting!