RL^3: Boosting Meta RL via RL inside RL^2

UMassAmherst | Manning College of Information & Computer Sciences

Abhinav Bhatia [1] Samer B. Nashed [1,2,3] Shlomo Zilberstein [1]

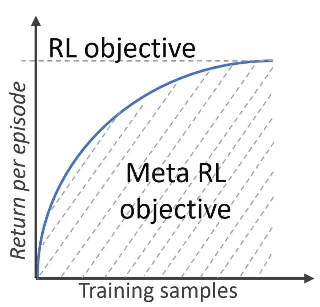
[1]University of Massachusetts Amherst [2]Université de Montréal [3]Mila - Quebec Artificial Intelligence Institute

Reinforcement Learning Conference

TL;DR: Supplementing Meta-RL inputs with Q-values learned online improves long-horizon performance and OOD generalization.

Meta-Reinforcement Learning

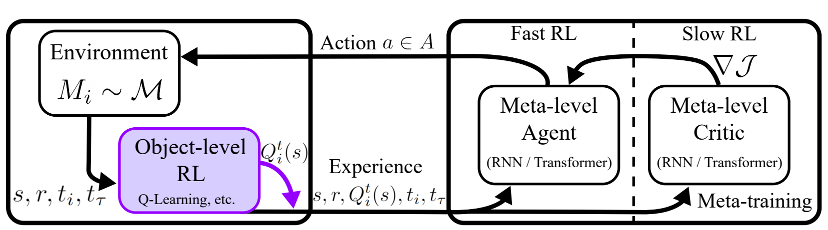
	RL	Meta-RL
Objective	Maximize return *within an episode*.	Maximize return over a *meta-episode* (whole interaction window) via online adaptation.
Scope	Single MDP (asymptotically optimal).	Family of MDPs (or *tasks*) from a known distribution.
Uses experience	To learn value functions.	To map history → action with sequence models like RNN/Transformer (e.g., RL^2), and/or incremental task-inference as an intermediate step.



- **Meta-RL Challenges:** i) OOD generalization and ii) long horizons (truncated gradients / long-range credit assignment / compounding inference errors).

RL^3: RL inside RL^2

- **RL^2 input:** experience sequence (s, a, r), into a Transformer (our impl.).
- **RL^3 input:** experience sequence **+ online $Q_t(s)$** (value estimates *for the current MDP*).



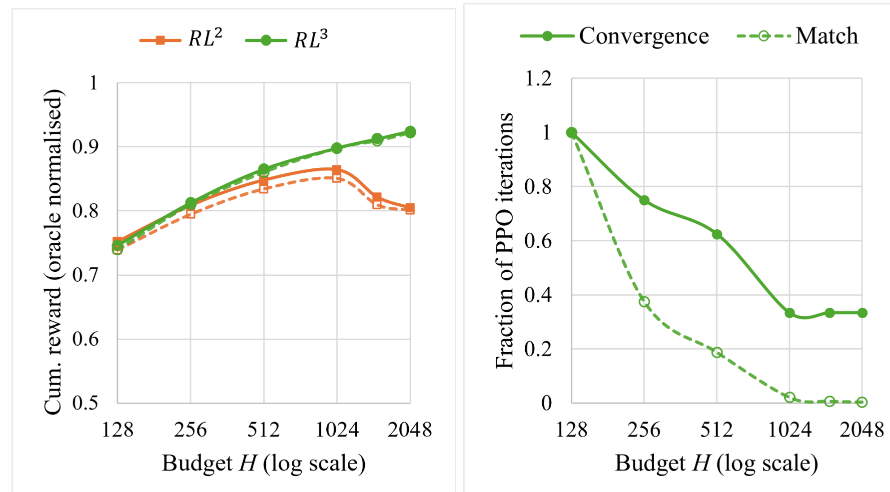
Overview: purple = additions to RL^2.

- **Insight:** Online Q_t adds **task-agnostic inductive bias** and an **actionable summary of experience** that *improves* as data accumulates, easing the sequence model's long-range credit assignment burden.
- **Outcome:** Improved OOD generalization and higher long-horizon returns.

	RL	RL^2	RL^3
Short-Term Efficiency	✗	✓	✓
Long-Term Performance	✓	✗	✓
OOD Generalization	✓	✗	✓
	(General Purpose)		(Improved)

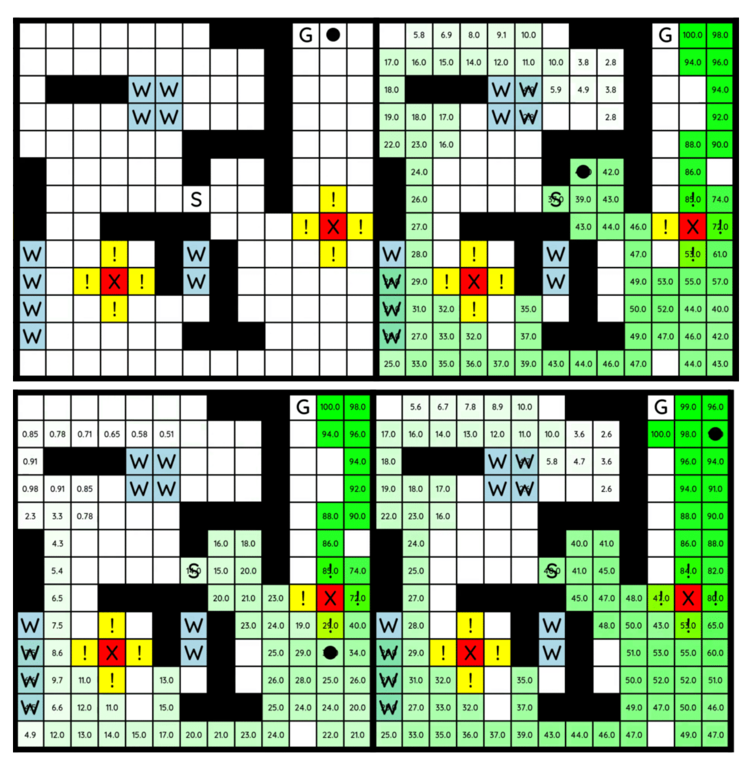
Results

Random MDPs: Stochastic MDPs drawn from a fixed distribution; parameters are varied for OOD tests. **RL^3 preserves RL's asymptotic scaling, maintains strong OOD performance**, and **meta-training is drastically more efficient**.



Left: Return vs. interaction budget H. The **gap widens as H increases** $128 \to 2048$ (solid: in-dist., dashed: OOD). *Right:* RL^3 meta-training converges or matches RL^2 with a fraction of the samples.

Gridworlds: Procedurally generated grids with obstacles (black), goals (G), hazards ('X', '!'), and slippery tiles ('W'). **RL^3 averages +50% return vs. RL^2**; and **+80% under OOD shifts**, where we varied obstacle density, stochasticity, number of hazard tiles, etc. Interestingly, even with value estimates on a **2× coarser state abstractions, return drops only ≈10% while compute is ≈2× faster**. Finally, **Meta-training is 30% more sample-efficient**.



Example meta-episode. After exploration in the first episode (top left), the RL^3 agent uses the estimated value function (text and green color gradients inside tiles) in future episodes. Demonstrably, **RL^3 avoids relying only on the RNN/transformer to plan the shortest path.**

Why RL^3 Works (Intuition)

RL^3 exploits properties of the action-value function $Q(s, a)$:

- **Inductive bias:** Q is a ubiquitous RL signal; adding it provides task-agnostic structure to the meta-learner.
- **Compression:** summarizes arbitrarily long, out-of-order experience into a fixed-size representation.
- **Actionable:** a policy greedy w.r.t. Q approaches optimal as $Q \to Q^\star$, shortening long-range credit assignment.
- **Refinement:** DP/TD updates contract toward $Q^\star$, so estimates typically improve with more interaction.
- **Task identification:** the Q-landscape is task-specific, helping disambiguate the current MDP.
- **Separation of concerns:** the RL module estimates/updates Q; the meta-policy explores and decides when/how to trust Q_t (at convergence, meta- and task-level value align).

Summary

RL^3 augments **RL^2** with **online Q-values** for the current task, adding a **task-agnostic inductive bias** and a **compact, actionable history summary**. This reduces **long-horizon credit assignment** load on the sequence model, so the meta-learner **adapts more effectively over long-horizons** and **generalizes OOD**; the idea extends to function-approximate Q in richer domains.

Additional Information

Find the link to the paper using the QR code. For additional information contact abhinavb-hati@umass.edu

Acknowledgments

This work was supported in part by the National Science Foundation grant numbers 1954782, 2205153, and 2321786.

abhinavbhati@umass.edu