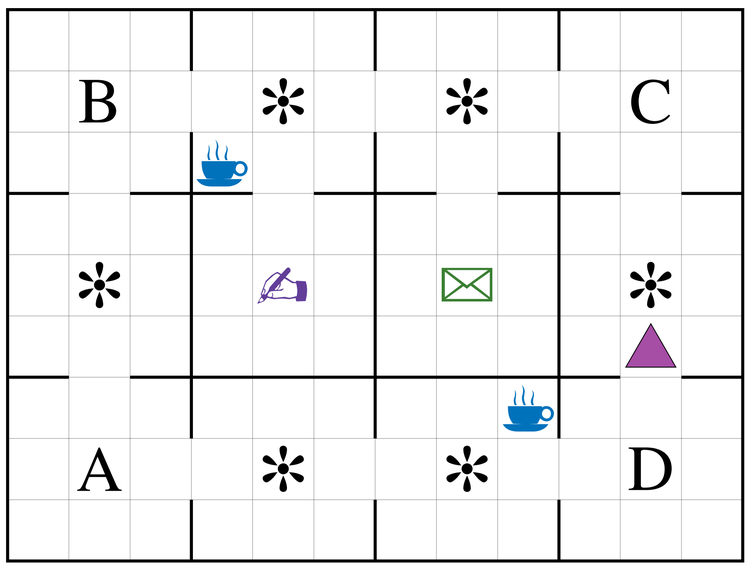
Symbolic Plans as High-Level Instructions for Reinforcement Learning

León Illanes Xi Yan Rodrigo Toro Icarte Sheila A. McIlraith

Department of Computer Science, University of Toronto Vector Institute

Motivation: Tell an RL agent to solve a specific task.



Symbol	Meaning
▲	Robot
✻	Furniture
☕	Coffee machine
✉	Mail room
✍	Office
A, B, C, D	Marked locations

Task examples

T1. Deliver mail to the office
T2. Deliver coffee and mail to the office
T3. Visit A, B, C, and D (in any order)

Q: How do we specify a new task?

A: ~~An expert designs a reward function.~~

We specify a goal for an abstract, high-level model.

Propositions: have-mail/coffee, delivered-mail/coffee, visited-A/B/C/D

Actions: get-mail/coffee, deliver-mail/coffee, visit-A/B/C/D

get-coffee:
pre: (none)
eff: have-coffee

deliver-coffee:
pre: have-coffee
eff: delivered-coffee, **not** have-coffee

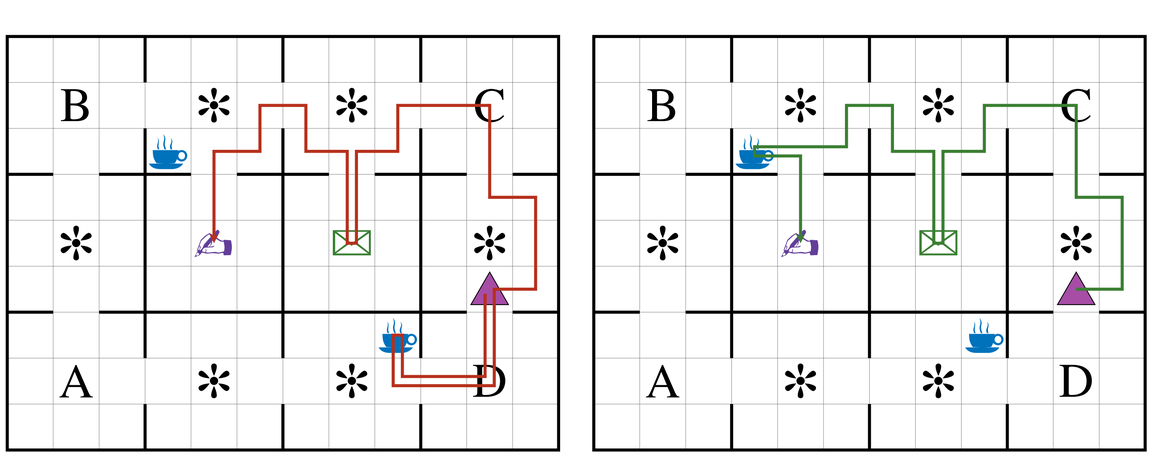
G1. { delivered-mail }
G2. { delivered-mail, delivered-coffee }
G3. { visited-A, visited-B, visited-C, visited-D }

We can use plans to guide the RL algorithms!

T1. ⟨get-coffee, deliver-coffee⟩
T2. ⟨get-coffee, get-mail, deliver-coffee, deliver-mail⟩
T3. ⟨go-to-A, go-to-B, go-to-C, go-to-D⟩

- This is **hierarchical RL** with a fixed high-level policy
  - Every action maps to an **option**
  - Many actions can map to a single option
    - e.g., deliver-mail and deliver-coffee

Blindly following the plans is suboptimal...

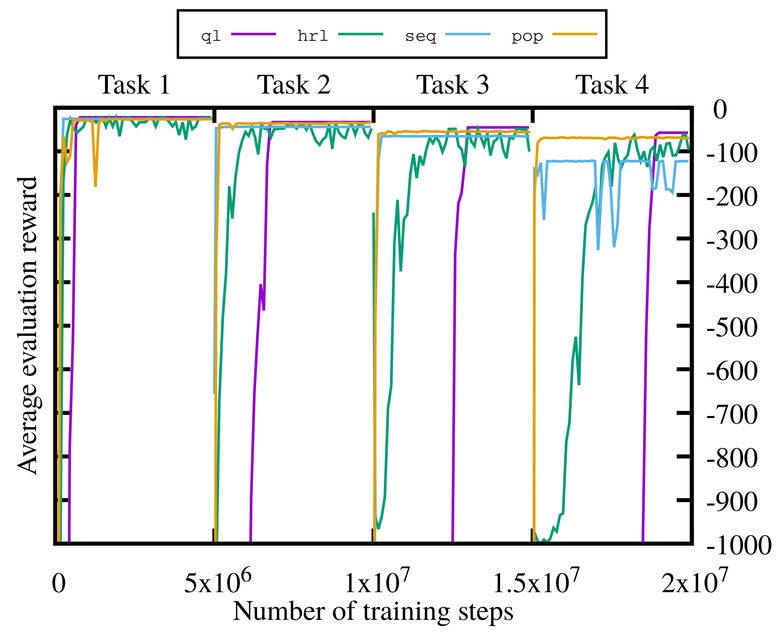


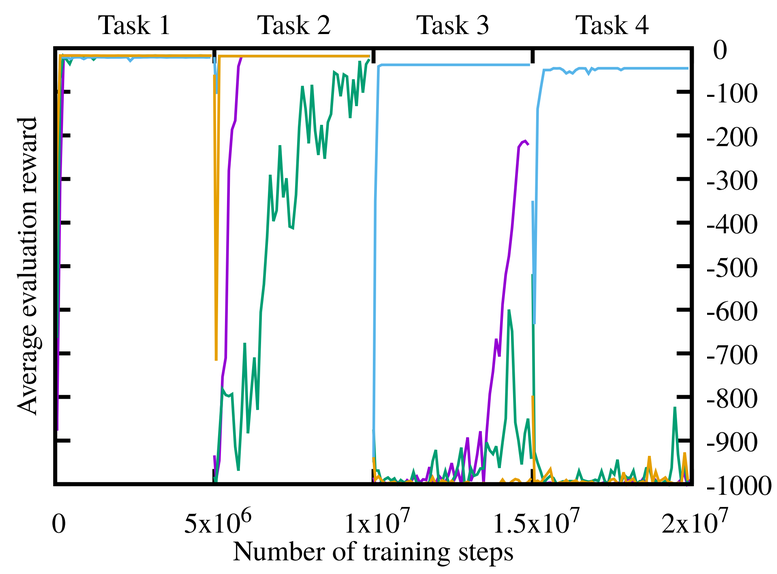
Relax the ordering constraints!

- Use **partial-order plans** instead of sequential ones
- The high-level policy is no longer fixed, but very restricted
  - The agent also has to learn how to order the plans

Experiments

We ran experiments in two different environments. In each, we train agents to solve four tasks order of complexity. The figures below show the rewards obtained by standard q-learning (ql), hierarchical RL using the options identified in the high-level models (hrl), and our approaches using sequential (seq) and partial-order (pop) plans. Our approaches find good solutions much faster than the alternatives, and effectively transfer learned information from one task to the next.





We have additional experiments that evaluate an execution monitoring approach that may help when models are incorrect.