Exploring Long-Horizon Reasoning with Deep RL in Combinatorially Hard Tasks

Andrew C. Li [1,2] Pashootan Vaezipoor[1,2] Rodrigo Toro Icarte[2,4,5] Sheila A. McIlraith[1,2,3]

[1]University of Toronto [2]Vector Institute [3]Schwartz Reisman Institute [4]Pontificia Universidad Católica de Chile [5]Centro Nacional de Inteligencia Artificial

TL;DR We propose **MuJoCo** tasks based on **NP-hard optimization problems** (e.g. TSP) to challenge the **long-term reasoning ability** of RL agents. We find that state-of-the-art RL and hierarchical RL approaches perform poorly and motivate two new approaches based on their weaknesses.

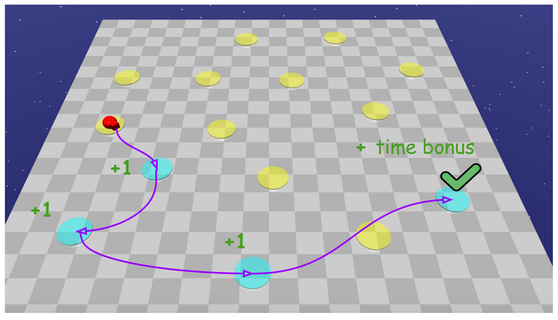
Overview

Motivation

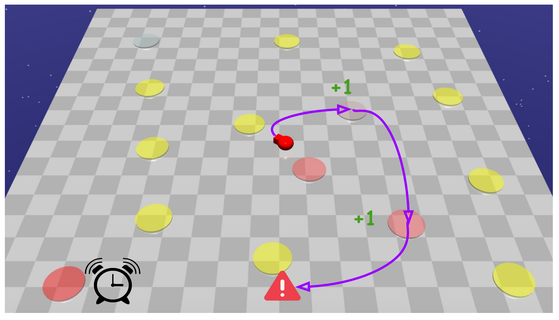
- Many real-world tasks involve high-level **combinatorial reasoning** and low-level **complex control** over long horizons.
- Standard benchmark tasks mostly involve simple high-level structure (e.g. reaching a goal location, opening a door).
- **Challenge:** Complex tasks often lead to **sparse rewards**.

Our tasks

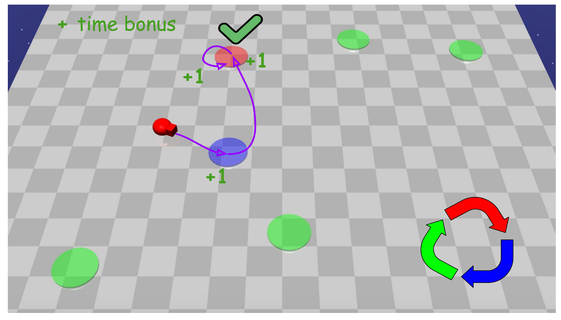
- Contain combinatorial structure.
- Require long-term reasoning for the best performance.
- Decompose into dense rewards — **no specialized exploration required!**



PointTSP: Visit all the zones as quickly as possible.



TimedTSP: Visit all the zones as quickly as possible without letting any unvisited zone reach its timeout.



ColourMatch: Make all zones the same colour as fast as possible. Visiting a zone cycles it to the next colour.

Can PPO reason over long horizons?

The paradox of discounting

- Discounting ($\gamma < 1$) leads to a myopic policy that **fails to consider long-term effects**.
- No discounting ($\gamma = 1$) is known to cause instability.

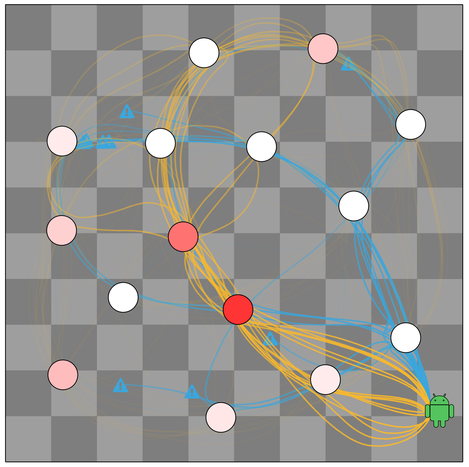
A simple fix for undiscounted ($\gamma = 1$) PPO

- **Hypothesis:** Value estimation is significantly harder with long horizons and $\gamma = 1$ due to increased variance.
- **Proposal (PPO$_{\mathcal{VD}}$):** Model the **mean and variance of the value function** rather than a point estimate.

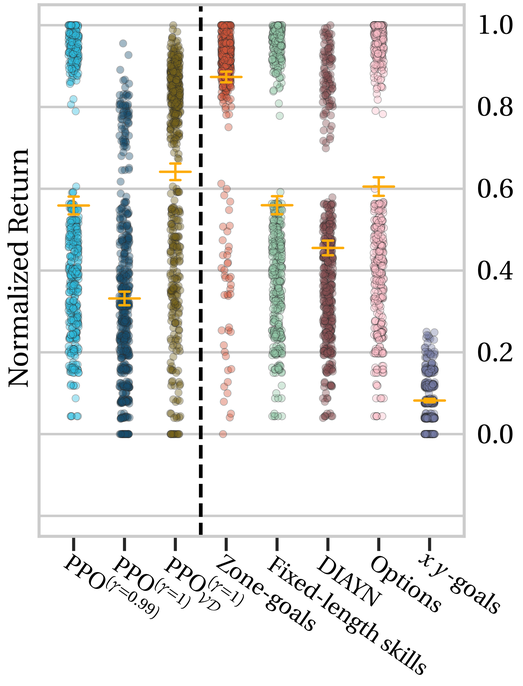
Result

- **PPO$_{\mathcal{VD}}^{(\gamma=1)}$ (our approach)** performed equal to or better than PPO at any discount factor.
- Discounting with PPO led to myopic behaviour.

TimedTSP



Undiscounted PPO$_{\mathcal{VD}}^{(\gamma=1)}$ (**ours**) immediately visits the two zones in danger of timing out. Discounted PPO$^{(\gamma=0.99)}$ myopically optimizes for dense reward and quickly fails.



Does hierarchy improve long-term reasoning?

Most work in learning hierarchy focuses on improving exploration under sparse rewards.

Motivating Problem

- Can HRL exploit high-level task structure to improve long-term reasoning in our **dense-reward** tasks?

Zone-goals (Ours)

- We design a domain-specific hierarchy for these tasks.
  - High-level policy selects the next zone to visit (trained via task rewards).
  - Low-level policy aims to navigate to the target zone (trained via shaped xy-rewards).

Result

- A handcrafted hierarchy (**Zone-goals**) significantly outperformed all other methods.
- State-of-the-art general-purpose HRL methods showed no improvement over flat PPO.
- Skill-based approaches were prone to collapsing into a single skill.