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!

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$_{VD}$): Model the mean and variance of the value function rather than a point estimate.

Result

- PPO$_{VD}$ (our approach) performed equal to or better than PPO at any discount factor.
- Discounting with PPO led to myopic behaviour.

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.