Model-Based Exploration in Monitored Markov Decision Processes

Alireza Kazemipour, Matthew E. Taylor, Michael Bowling

The 42nd Conference on Machine Learning (ICML 2025)

UNIVERSITY OF ALBERTA

The Intelligent Robot Learning Laboratory

amii

RL & AI

With feedback, Optimism Leads to Efficient Exploration

Core Problem: Can we use optimism in the absence of feedback?

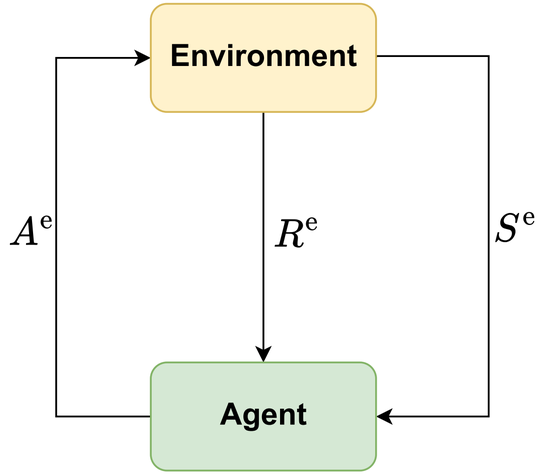
(a) (b) (c)

Without feedback, pessimism leads to worst-case Guarantee

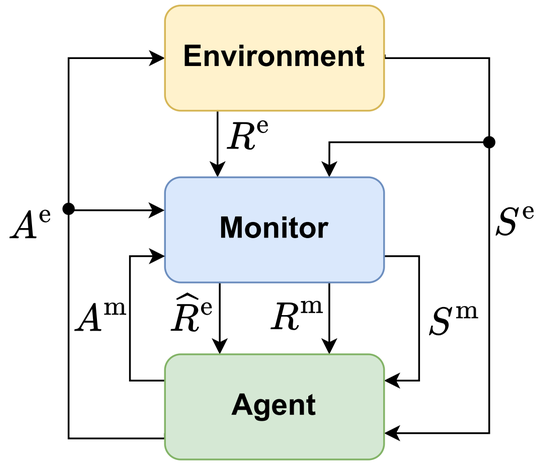
Monitored Markov Decision Processes (Mon-MDPs)

A framework where the reward can be unobservable

- The monitor is another MDP in addition to the environment
- The goal is to maximize the joint reward $\sum_{t=0}^{\infty} \gamma^t \left(R_t^e + R_t^m \right)$
- But the agent observes $\widehat{R}_t^e$ instead of the true environment reward R_t^e
- For some state-action $\widehat{R}_t^e = \perp$ (NaN) at all times t!



MDP



Mon-MDP

[“Monitored Markov Decision Processes”, AAMAS 2024]

Monitored Model-Based Interval Estimation with Exploration Bonus

Monitored MBIE-EB

- **Innovation** 1: Extending MBIE-EB to Mon-MDPs

$$\tilde{R}_{opt}(s,a) = \bar{R}^e\left(s^e,a^e\right) + \underbrace{\frac{\beta^e}{N(s^e,a^e)}}_{\text{bonus for } r^e} + \bar{R}^m\left(s^m,a^m\right) + \underbrace{\frac{\beta^m}{N(s^m,a^m)}}_{\text{bonus for } r^m} + \underbrace{\frac{\beta}{N(s,a)}}_{\text{bonus for } p}$$

- **Innovation 2**: Pessimism instead of the optimism

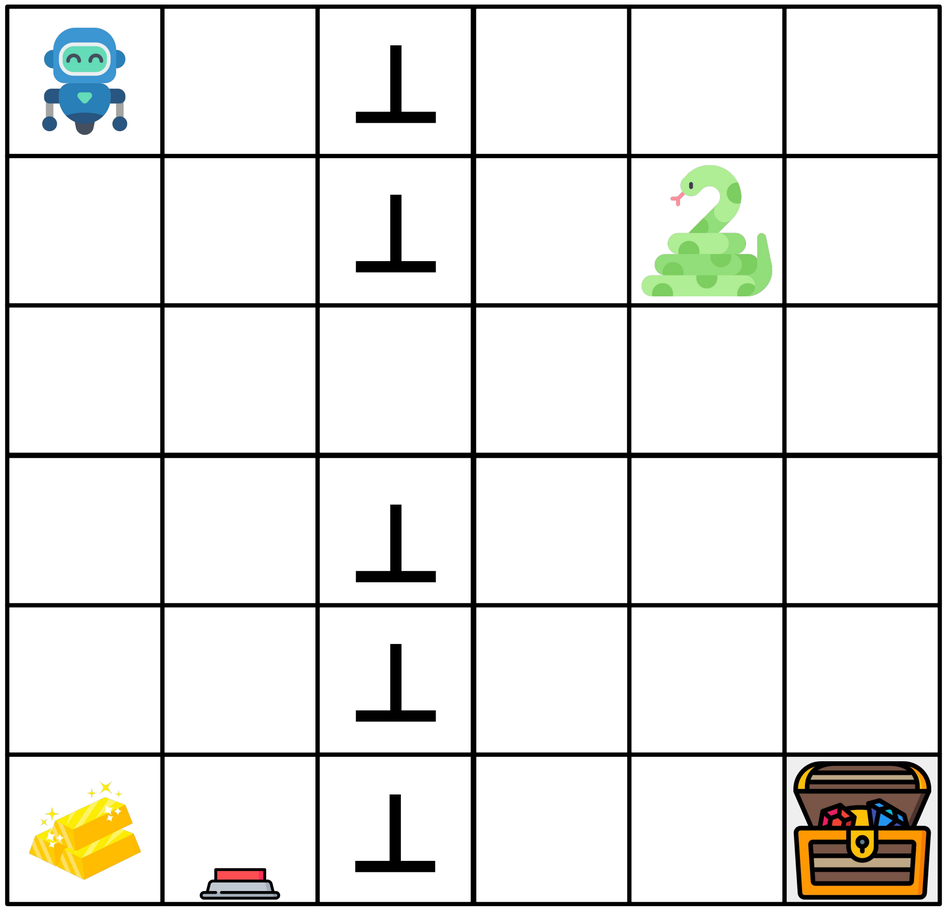
$$\tilde{R}_{opt}(s,a) = r^e_{\min} + \bar{R}^m\left(s^m,a^m\right) + \underbrace{\frac{\beta^m}{N(s^m,a^m)}}_{\text{bonus for } r^m} + \underbrace{\frac{\beta}{N(s,a)}}_{\text{bonus for } p}$$

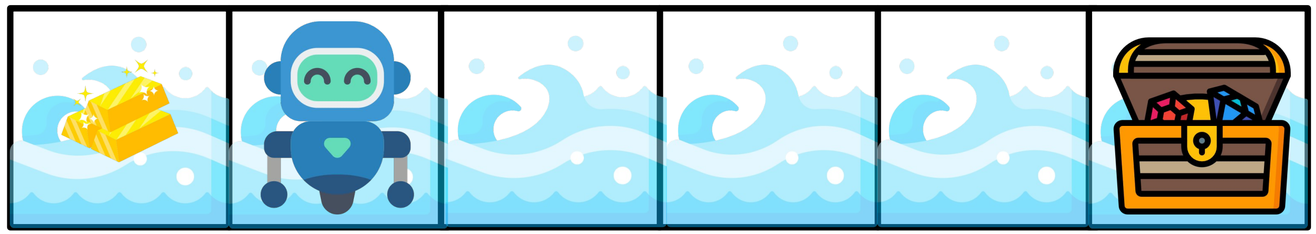
- **Innovation 3**: Explore to observe rewards

$$\tilde{R}_{obs}(s,a) = \text{KL-UCB}\left(0,N(s,a)\right)\mathbb{I}\left\{N\left(s^e,a^e\right)=0\right\} + \frac{\beta^{obs}}{\sqrt{N(s,a)}},$$

Results

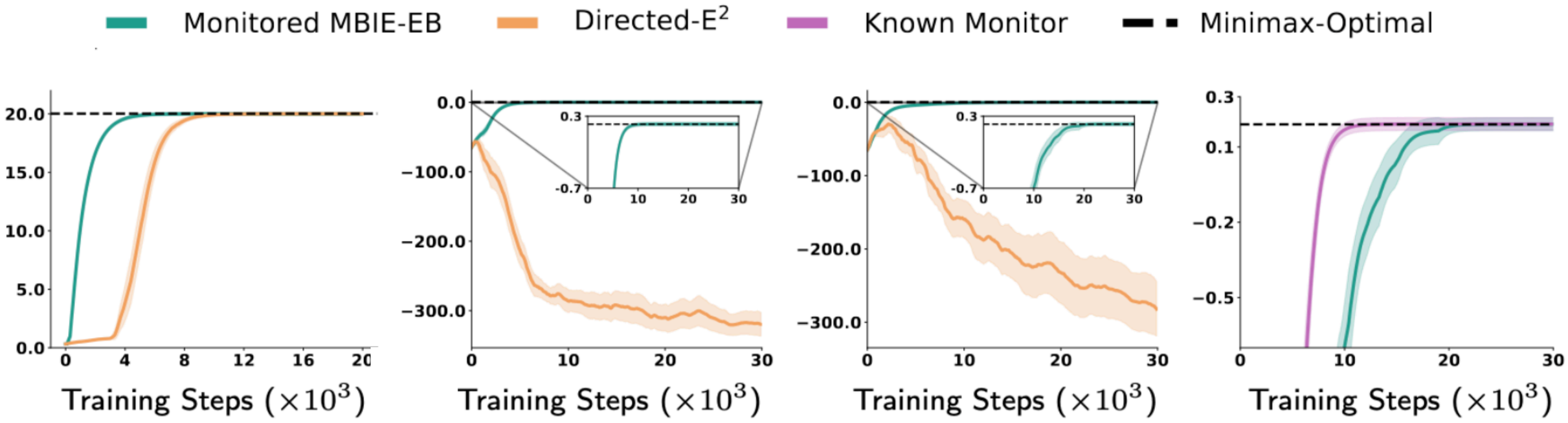
Environments





Results

Discounted Test Return



(a) River Swim (b) Bottleneck (100%) (c) Bottleneck (5%) (d) Bottleneck (5%)