The Virtues of Laziness in Model-based RL: A Unified Objective and Algorithms

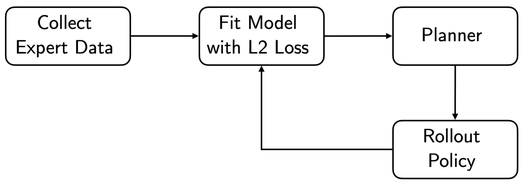
Sanjiban Choudhury, Assistant Professor, Cornell University

Joint work with Anirudh Vemula, Yuda Song, Aarti Singh, Drew Bagnell. To appear in ICML'23.

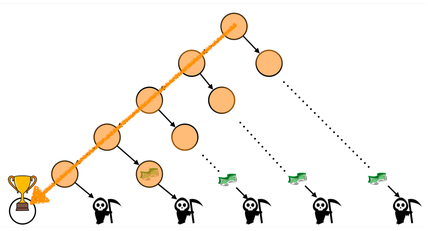
Fundamental Challenges in Model-Based RL

Model Learning with Planner in Loop

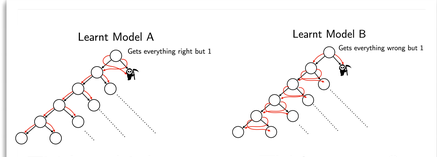
(Ross & Bagnell, 2012)



Challenge 1: Planning is exp(T)!



Challenge 2: Mismatched Objective

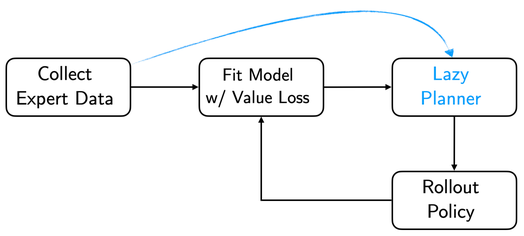


TV loss high TV hoss ≈ 0

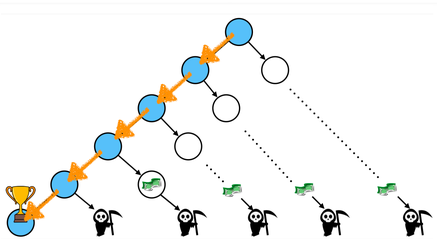
Low TV Loss $\neq$ Better Performance!

Key Insight: Don't explore. Be Lazy. Do well on Expert States.

Lazy Model-based Policy Search (LAMPS)

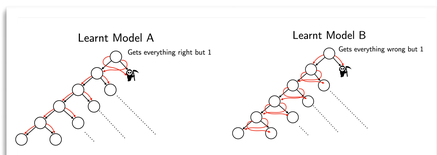


Solution 1: Restart from expert states to be Poly(T)



Solution 2: Match value moments

$E_{s' \sim \hat{M}} V(s') = E_{s'' \sim M^*} V(s'')$



Value loss = 0 Value loss high

New Lemma: Performance Difference via Advantage in Model

$$J_{M^*}(\pi^*) - J_{M^*}(\hat{\pi})$$

$$= \mathbb{E}_{s^* \sim \pi^*} \left[A^{\hat{\pi}}(s^*, a^*) \right]$$

Advantage of expert in model

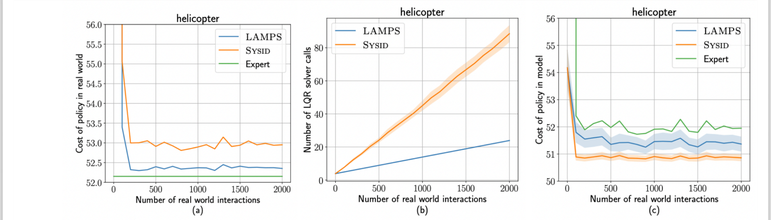
$$+ T \mathbb{E}_{s,a \sim \pi^*} \left[E_{s' \sim \hat{M}} V^{\hat{\pi}}(s') - E_{s'' \sim M^*} V^{\hat{\pi}}(s'') \right]$$

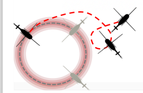
Value matching on expert states

$$+ T \mathbb{E}_{s,a \sim \hat{\pi}} \left[E_{s' \sim \hat{M}} V^{\hat{\pi}}(s') - E_{s'' \sim M^*} V^{\hat{\pi}}(s'') \right]$$

Value matching on learner states

LAMPS finds a better policy with fewer samples + fewer computation





SysID: Use planner (iLQR)

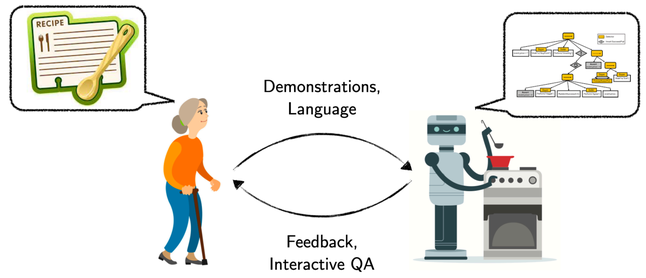
LAMPS: Use PSDP (LQR on expert traj)

About Me

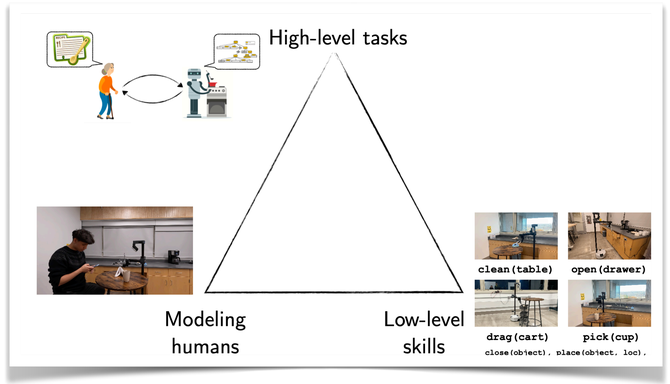
Cornell Bowers CIS Computer Science

PORTAL

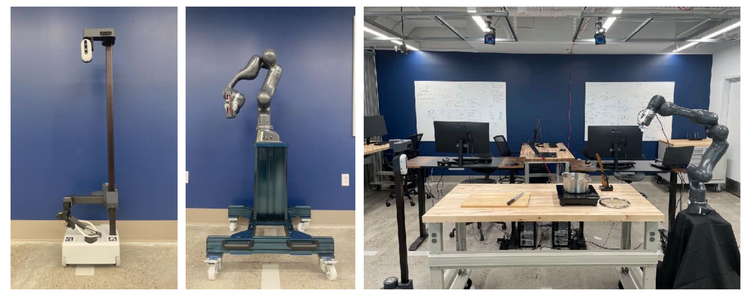
How can we program robots via natural interactions?



Three main research thrusts



Robots and applications



2023 FRR & NRI Principal Investigators' Meeting

May 2-3, 2023