NEURAL INFORMATION PROCESSING SYSTEMS

When to Update Your Model: Constrained Model-based Reinforcement Learning

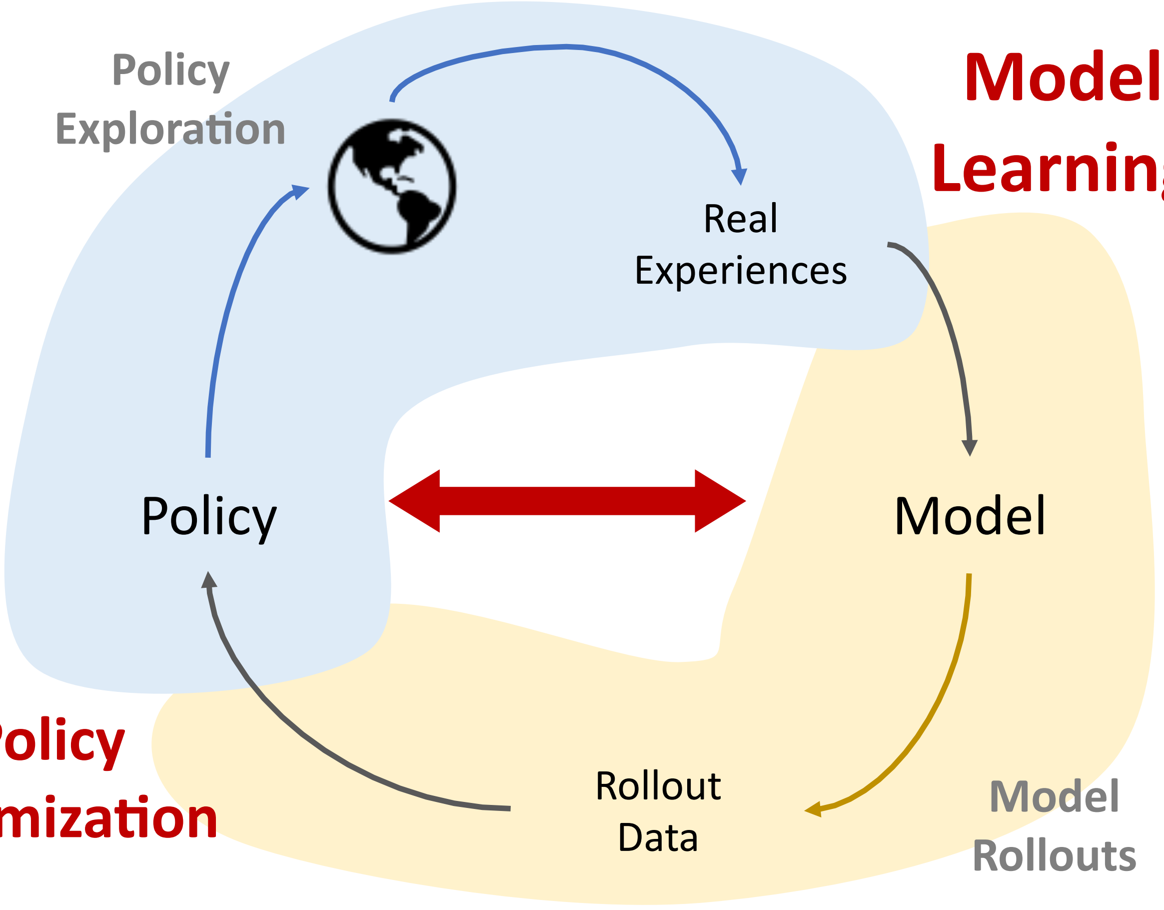
Tianying Ji, Yu Luo, Fuchun Sun, Mingxuan Jing, Fengxiang He, Wenbing Huang

中国科学院软件研究所
Institute of Software Chinese Academy of Sciences

JDT
京东科技

Setting : Dyna-style Model-based RL

■ Alternating between Two-stages : Model Learning & Policy Optimization

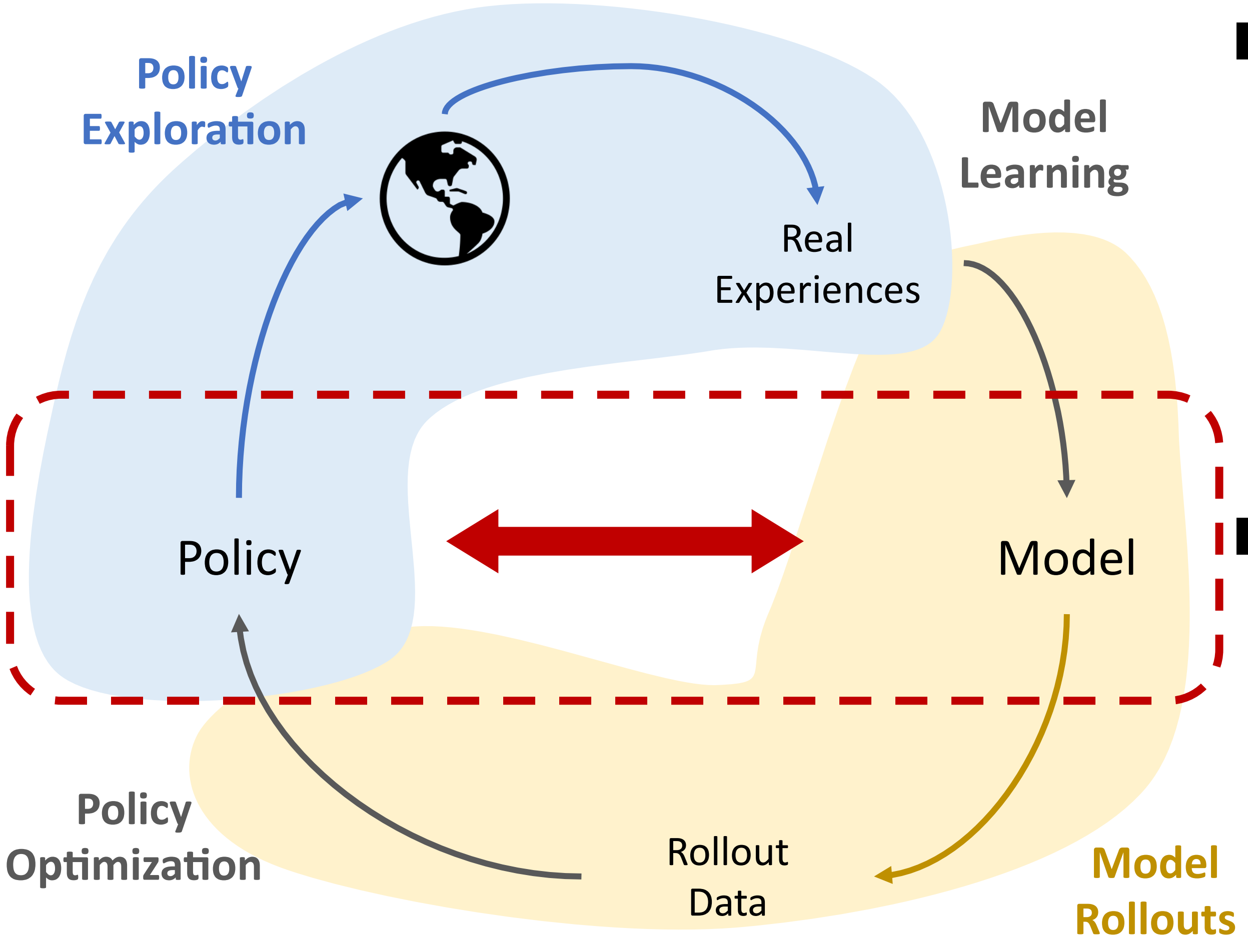


Model Learning | based on policy exploration

Policy Optimization | based on model rollouts

Entangled Nature of Model-based RL

❗ Complexity: Model ⇔ Policy (Chain Reaction)



- model quality → policy optimization

$M_1 \longrightarrow \ldots$

$\pi_{11}, \pi_{12}, \ldots, \pi_{1n}$

Env

✓

- policy exploration → model learning

$M_1 \xrightarrow{\pi_{11}, \pi_{12}, \ldots, \pi_{1n}} M_2$

?

Previous Monotonic Improvement Guarantee

■ **Local View** model quality → policy optimization

$M_1 \longrightarrow \dots$

ϵ_m Env

$\pi_{11}, \pi_{12}, \dots, \pi_{1n}$

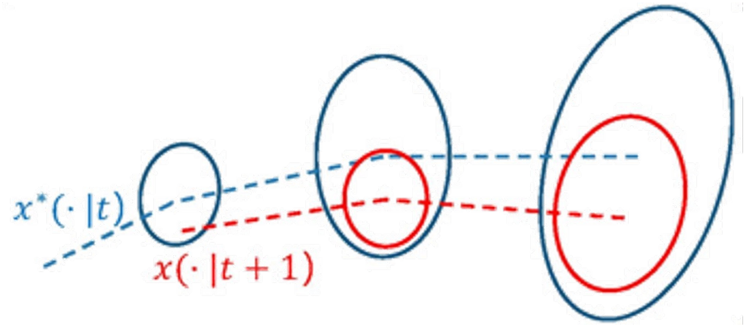
ϵ_π

▶ Discrepancy Bound Scheme

$$V^\pi(\mu) \geq V_M^\pi(\mu) - C(\epsilon_m, \epsilon_\pi)$$

✗ Weak Feasibility and Coarse Solution

Large ϵ_m → Large $C(\epsilon_m, \epsilon_\pi)$



✗ Model Quality = Validation Loss

$M_2 > M_1$

Supervised Learning

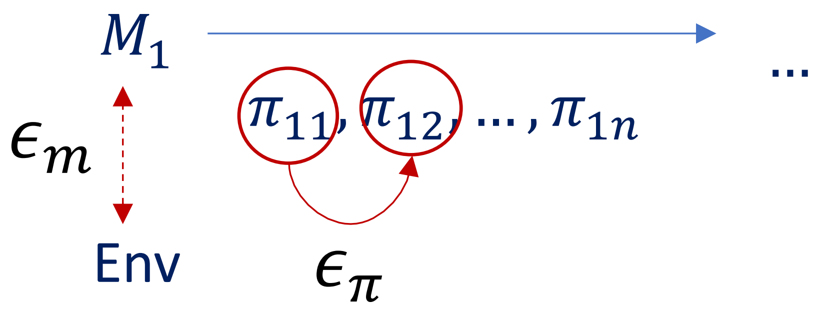
$\epsilon_{M_2} < \epsilon_{M_1}$

⭐ Derived policy achieves higher return

Previous Monotonic Improvement Guarantee

■ Local View

model quality → policy optimization



▶ Discrepancy Bound Scheme

$$V^{\pi}(\mu) \geq V_M^{\pi}(\mu) - C(\epsilon_m, \epsilon_\pi)$$

? How does the policy affect model updating?

? What is an indeed better model in MBRL?

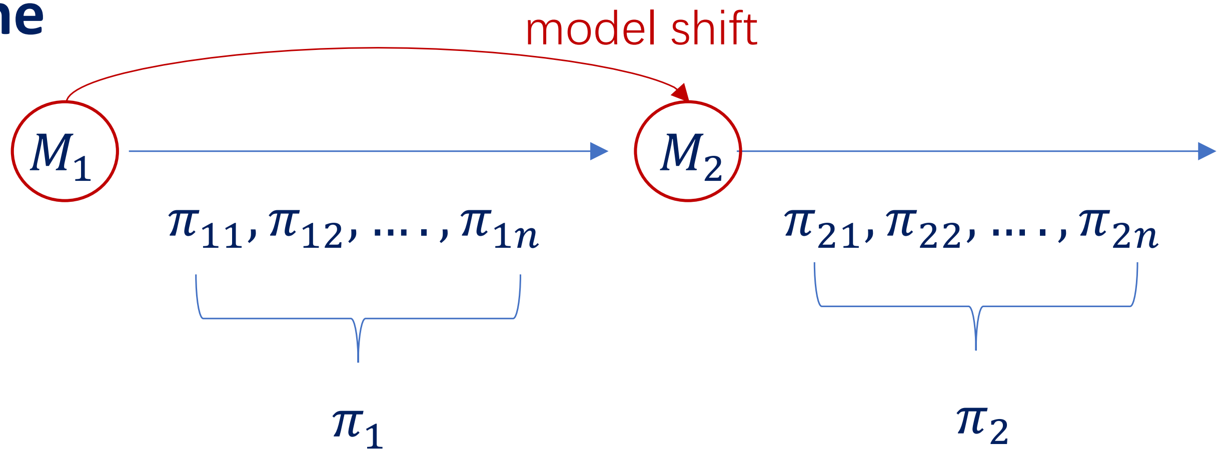
? Can model-based RL algorithms be guranteed to improve the policy monotonically when considering model shifts?

Towards Better Monotonic Improvement Guarantee

■ Global View – What we really want !

⭐ Performance Difference Bound Scheme

$$V^{\pi_2|M_2}(\mu) - V^{\pi_1|M_1}(\mu) \geq C.$$



Advantages

- ✓ guarantee monotonicity across models
- ✓ modeling the entangled nature
- ✓ a novel measurement of model quality

Lower-bound optimization with model shift constraints

⭐ Monotonicity Improvement → Jointly Constrained Optimization Problem

Performance Difference Bound for Model-based RL

$$V^{\pi_2|M_2} - V^{\pi_1|M_1} \geq \boxed{\kappa \cdot (\epsilon_{M_1}^{\pi_1} - \epsilon_{M_2}^{\pi_2})} + V_{M_2}^{*} - V_{M_1}^{*} - \epsilon_{opt}.$$

Performance Difference

Inconsistency Gap

Ceiling Performance Gap

Better Model $M_2 > M_1$

▶ Lower inconsistency with Env

▶ Higher ceiling performance

Ceiling Return Gap under Model Shift

$$V_{M_2}^{*} - V_{M_1}^{*} \geq -\frac{\gamma}{1-\gamma} L \cdot \sup_{\pi \in \Pi} \mathbb{E}_{s,a \sim d_{M_2}^{\pi}} \Big[|P_{M_2}(\cdot|s,a) - P_{M_1}(\cdot|s,a)| \Big]$$

Sharp model shift may corrupts monotonicity

→ Introduce model shift constraints

Lower-bound optimization with model shift constraints

⭐ Monotonicity Improvement → Jointly Constrained Optimization Problem

Refined Bound with Constraint

$$
\begin{aligned}
V^{\pi_2|M_2} - V^{\pi_1|M_1} \geq \kappa \cdot \Big\{ & \mathbb{E}_{s,a\sim d^{\pi_1}} \mathcal{D}_{\mathrm{TV}}\left[P(\cdot|s,a) \| P_{M_1}(\cdot|s,a)\right] \\
& - \mathbb{E}_{s,a\sim d^{\pi_2}} \mathcal{D}_{\mathrm{TV}}\left[P(\cdot|s,a) \| P_{M_2}(\cdot|s,a)\right] \Big\} - \frac{\gamma}{1-\gamma} L \cdot (2\sigma_{M_1,M_2}) - \epsilon_{opt}, \\
s.t. \quad & \mathcal{D}_{\mathrm{TV}}(P_{M_2}(\cdot|s,a) \| P_{M_1}(\cdot|s,a)) \leq \sigma_{M_1,M_2}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.
\end{aligned}
$$

Achieve Non-Negative Lower Bound

Constrained Lower-Bound Optimization Problem

$$
\begin{aligned}
& \min_{\substack{M_2 \in \mathcal{M} \\ \pi_2 \in \Pi}} \ \mathop{\mathbb{E}}_{s,a\sim d^{\pi_2}} \left[\sum_{s' \in \mathcal{S}} \left| P(s'|s,a) - P_{M_2}(s'|s,a) \right| \right], \\
& s.t. \quad \sup_{s\in\mathcal{S}, a\in\mathcal{A}} \mathcal{D}_{\mathrm{TV}}(P_{M_1}(\cdot|s,a) \| P_{M_2}(\cdot|s,a)) \leq \sigma_{M_1,M_2}.
\end{aligned}
$$

Feasible example for constrained optimization problem

⭐ To derive a feasible solution under the generative model setting

Assumption: Generative Model

$$\forall s' \in \mathcal{S}, \quad \hat{P}(s'|s,a) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{s^i_{s,a} = s'\}$$

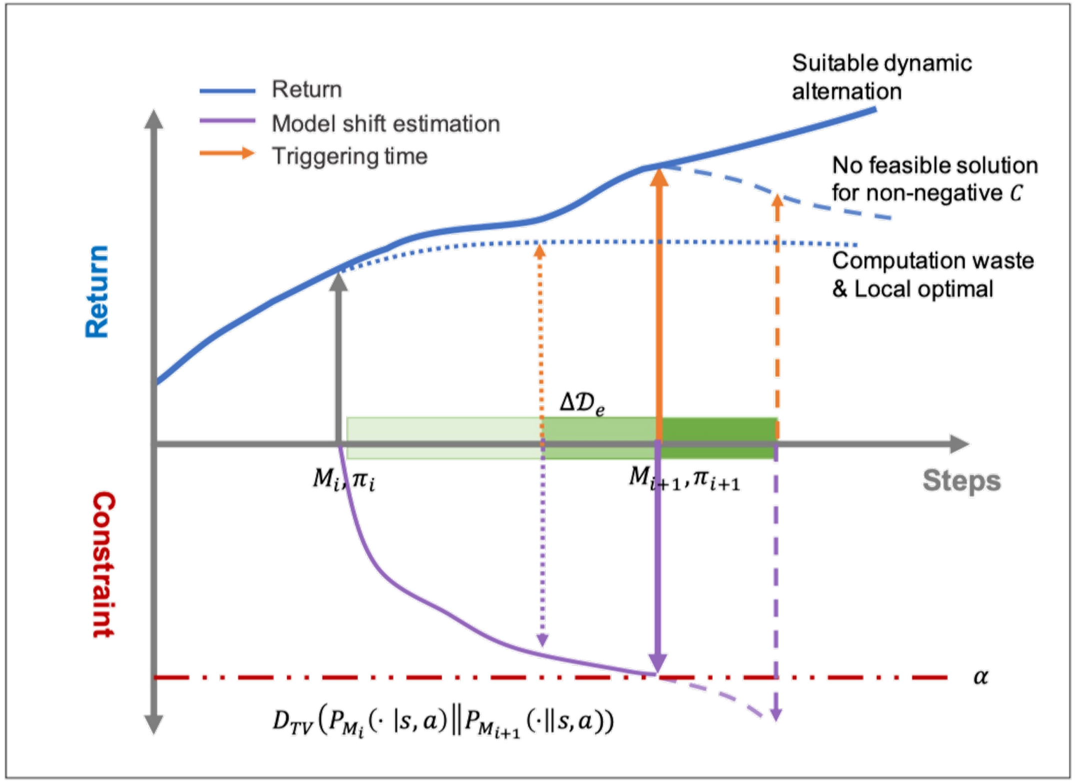
Conclusion:

$$\boxed{k} = \frac{2}{\epsilon^2} \log \frac{2^{vol(\mathcal{S})} - 2}{\xi} - N.$$

$$\epsilon = \delta_{M_1}(\cdot|s,a) - \frac{(1-\gamma)L}{R} \cdot (2\sigma_{M_1,M_2}) - \frac{(1-\gamma)^2}{R\gamma} \cdot \epsilon_{opt}$$

✻ A dynamic adjustment between model learning and policy interaction may beneift the performance monotonicity.

Practical Algorithm: CMLO



- Objective minimization - how to train the model
  - Model ensemble
  - optimize the Negative Log Likelihood
- Constraint estimation

$$\mathcal{D}_{\mathrm{TV}}(P_{M_1}(\cdot|s,a)\|P_{M_2}(\cdot|s,a)) = \frac{1}{2}\boxed{\sum_{s'\in\mathcal{S}}}\boxed{\left[|P_{M_1}(s'|s,a) - P_{M_2}(s'|s,a)|\right]}$$

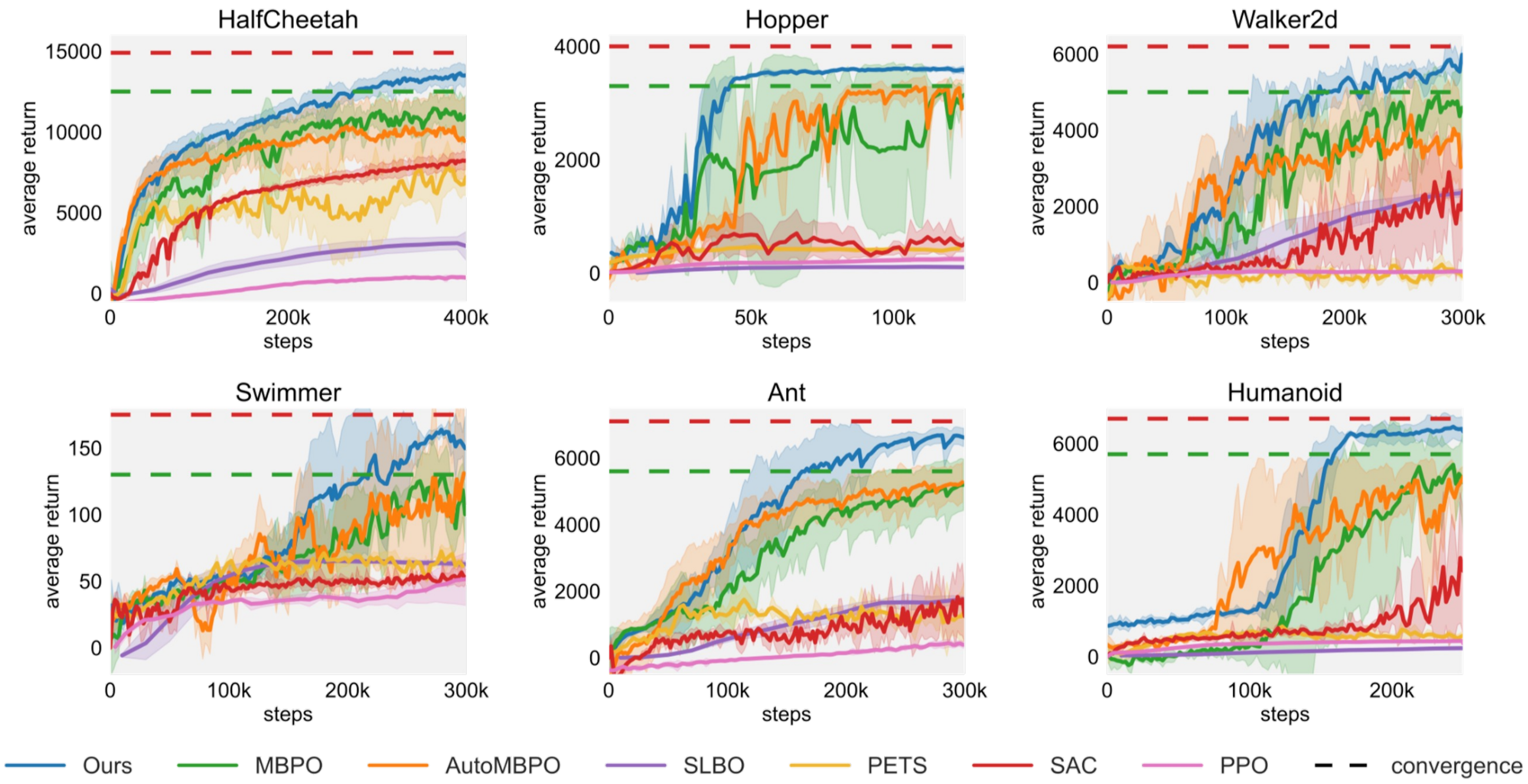
Policy coverage $vol(\mathcal{S}_\mathcal{D})$

Prediction bias $\mathcal{L}(\Delta\mathcal{D})$

- **Event-triggered mechanism** - when to train the model

$$\sum_{i=0}^{[\tau/F]} \log\left(\frac{vol(\mathcal{S}_{\mathcal{D}_t \cup \Delta\mathcal{D}(Fi)})}{vol(\mathcal{S}_{\mathcal{D}_t})} \cdot \mathcal{L}(\Delta\mathcal{D}(Fi)) + \beta\right) \geq \alpha$$

Evaluation on MuJoCo benchmarks



✓ Faster convergence speed
✓ Better eventual performance

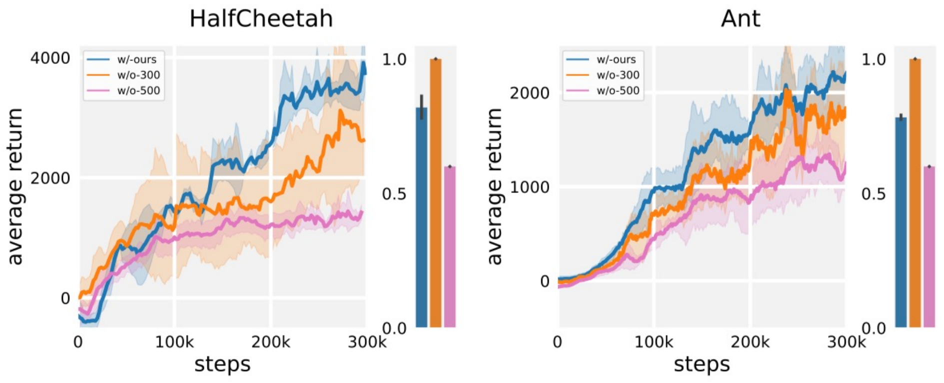
Additional experiments on the generalizability

Policy Optimization Oracle

$$V_M^*(\mu) - \epsilon_{opt} \leq V_M^\pi(\mu) \leq V_M^*(\mu)$$

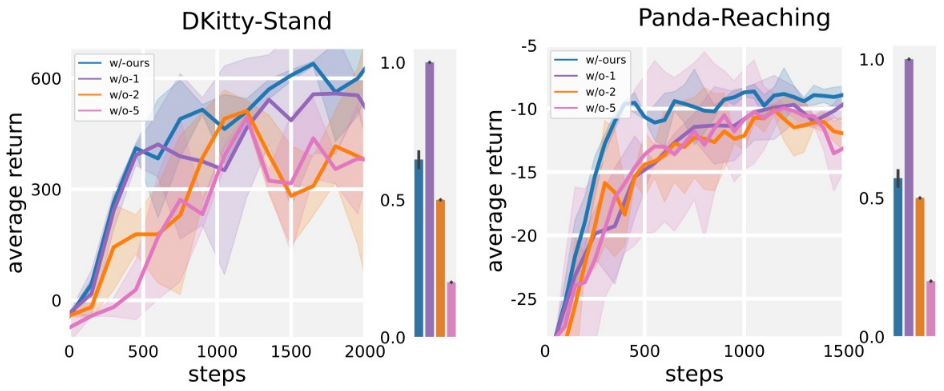
- ✓ Compatible with many local view guarantees
- ✓ Allows for many policy optimization methods

✓ Under Dyna-style



(a) policy optimization oracle: TRPO

✓ Jumping off Dyna-style



(b) policy optimization oracle: iLQR

Summary

- We propose a novel and general theoretical scheme for a non-decreasing performance guarantee of MBRL
- Follow-up derivations reveal previously neglected entanglement nature
- Empirical results verify both the effectiveness and generalizability

⭐ Thanks for listening !