Scrutinize What We Ignore: Reining In Task Representation Shift Of Context-Based Offline Meta Reinforcement Learning

Hai Zhang[1], Boyuan Zheng[1], Tianying Ji[2], Jinhang Liu[1], Anqi Guo[1], Junqiao Zhao[1], Lanqing Li[3,4]

[1] Tongji University, [2] Tsinghua University, [3] The Chinese University of HongKong, [4] Zhejiang Lab

ICLR

Motivation

- Monotonic Performance Improvement Guarantee for COMRL.

Model-Free Case

- REINFORCE → TRPO/PPO

Let $\alpha = D_{TV}^{max}(\pi_{old}, \pi_{new})$. Then the following bound holds:

$$\eta(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{1}$$

where $\epsilon = \max_{s,a} |A_\pi(s,a)|$.

1. TRPO: Hyper-parameter δ to constrain the variation of the policy.
2. PPO: Clipped surrogate objective or adaptive KL penalty coefficient to constrain the variation of the policy.

Model-Based Case

- MBPO → CMLO/USB-PO

Let policy π_i denotes the ϵ_{opt} optimal policy under the dynamic model M_i and σ_{M_1,M_2} be the constraint threshold for M_1 and M_2. Then the following bound holds:

$$\begin{aligned} V^{\pi_2|M_2} - V^{\pi_1|M_1} \geq & \kappa(\mathbb{E}[D_{TV}[P(\cdot|s,a)||P_{M_1}(\cdot|s,a)]] \\ & - \mathbb{E}[D_{TV}[P(\cdot|s,a)||P_{M_2}(\cdot|s,a)]]) \\ & - \frac{\gamma}{1-\gamma}L(2\sigma_{M_1,M_2}) - \epsilon_{opt} \end{aligned} \tag{2}$$

where $\sigma_{M_1,M_2} = \max_{s,a} D_{TV}(P_{M_1}(s,a)||P_{M_2}(s,a))$

1. CMLO: Hyper-parameter α to constrain the variation of the model.
2. USB-PO: Automatically fine-tune the variation of the model.

What About The Multi-Task/COMRL Setting?

Performance Improvement Guarantee For Previous COMRL Endeavors

- Our Previous UNICORN Framework:

Denote X_b and X_t are the behavior-related (s,a)-component and task-related (s',r)-component of the context X, with $X = (X_b, X_t)$.

$$I(Z;X_t|X_b) \leq I(Z;M) \leq I(Z;X) \tag{3}$$

1. $\mathcal{L}_{FOCAL} \equiv -I(Z;X) = -I(Z;X_t|X_b) - I(Z;X_b)$
2. $\mathcal{L}_{CORRO} \equiv -I(Z;X_t|X_b)$
3. $\mathcal{L}_{CSRO} \geq (\lambda-1)I(Z;X) - \lambda I(Z;X_t|X_b)$

- Return Discrepancy in COMRL:

$$\begin{aligned} |J^*(\theta) - J(\theta)| \leq & \frac{2R_{max}L_z}{(1-\gamma)^2}\mathbb{E}_{m,x}(|Z(\cdot|x;\phi) \\ & - Z(\cdot|x;\phi^{mutual})| + |Z(\cdot|x;\phi^{mutual}) - Z(\cdot|x;\phi^*)|) \end{aligned} \tag{4}$$

1. Adopt maximizing $I(Z;M)$ → Minimizing $|Z(\cdot|x;\phi) - Z(\cdot|x;\phi^{mutual})|$.
2. Adopt standard offline RL algorithms → Maximizing $J^*(\theta)$.

Monotonic Performance Improvement Guarantee For COMRL

- Monotonic Performance Improvement Condition For Previous Framework

$$\begin{aligned} \epsilon_{12}^* \triangleq & J^*(\theta_2) - J^*(\theta_1) \\ \geq & \frac{4R_{max}L_z}{(1-\gamma)^2}\mathbb{E}_{m,x}(|Z(\cdot|x;\phi) - Z(\cdot|x;\phi^*)|) \end{aligned} \tag{5}$$

- Monotonic Performance Improvement Guarantee With Variation Of Task Representation

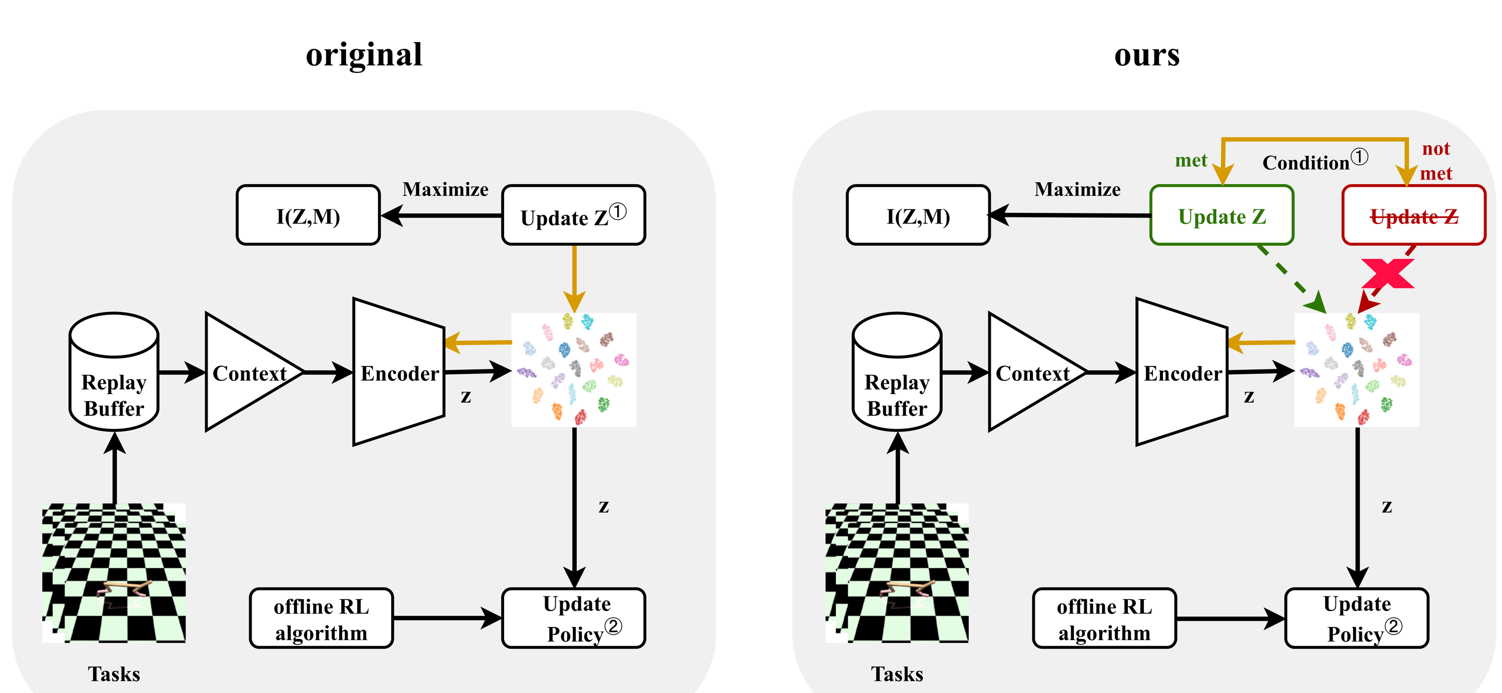
$$\begin{aligned} \epsilon_{12}^* - & \frac{2R_{max}L_z}{(1-\gamma)^2}\mathbb{E}_{m,x}[2|Z(\cdot|x;\phi_2) - Z(\cdot|x;\phi^*)| \\ & + |Z(\cdot|x;\phi_2) - Z(\cdot|x;\phi_1)|] \geq 0 \end{aligned} \tag{6}$$

- How To Achieve Monotonic Performance Improvement Guarantee

Given that the context encoder has already been trained by maximizing $I(Z;M)$ to some extent. Update the context encoder via maximizing $I(Z;M)$ from at least extra k samples, where:

$$k = \frac{1}{\kappa^2}\left(\sqrt{2\log\frac{2^{|Z|}-2}{\xi}} + \sqrt{\alpha}\right)^2 \tag{7}$$

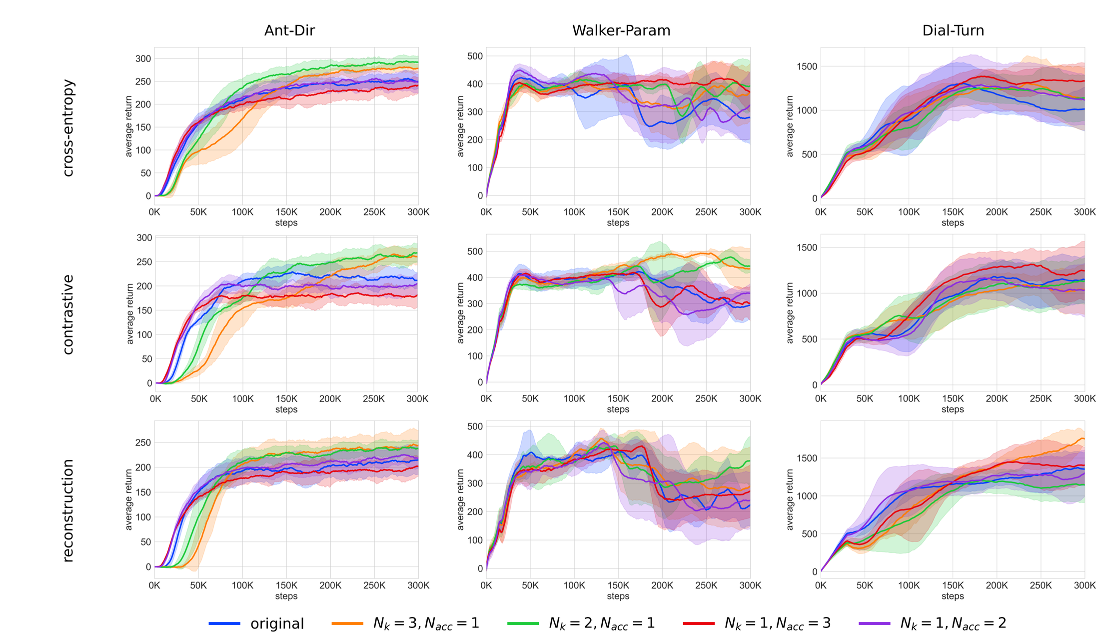
Task Representation Shift Determines The Monotonic Performance Improvement Guarantee!



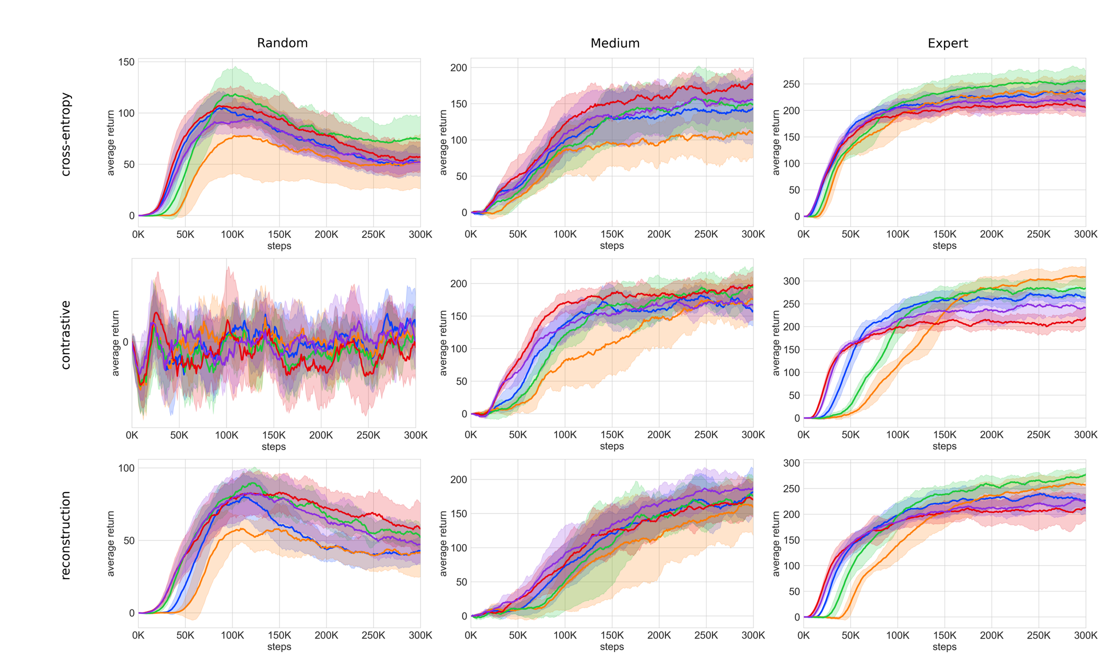
Alternating update process: ①->②->①->②->...->①->②

Promising Experimental Results

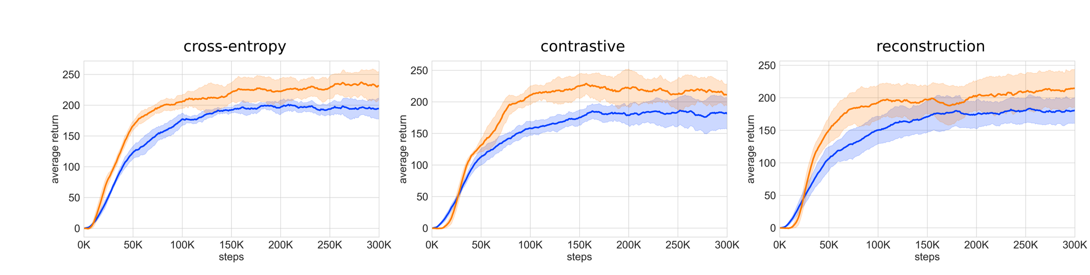
- MuJoCo And MetaWorld Benchmarks



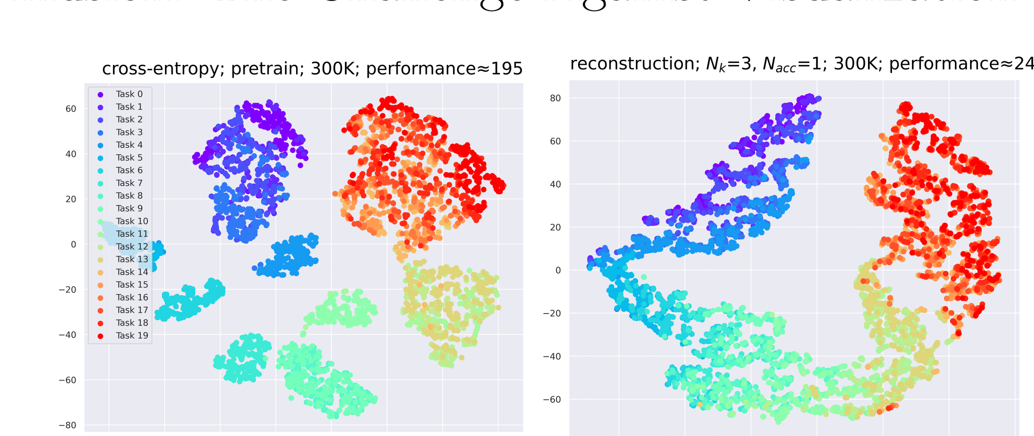
- Different Data Qualities



- Training From Scratch v.s. Pre-training



- Illusion: The Challenge Against Visualization



Personal Web

Code