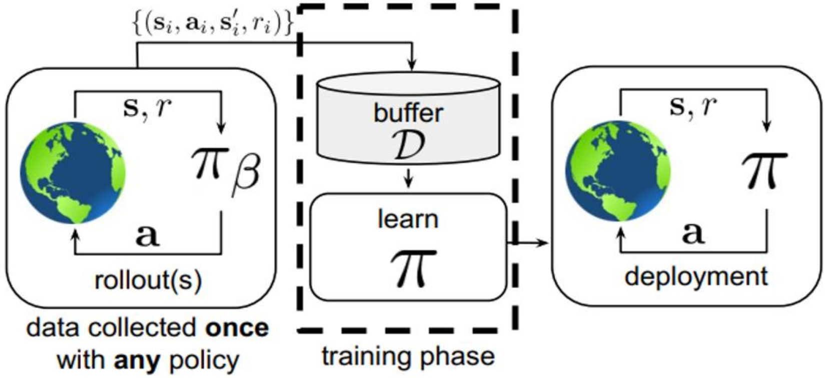
Context Shift Reduction for Offline Meta-Reinforcement Learning

Yunkai Gao, Rui Zhang, Jiaming Guo, Fan Wu, Qi Yi, Shaohui Peng, Siming Lan, Ruizhi Chen, Zidong Du, Xing Hu, Qi Guo, Ling Li, Yunji Chen

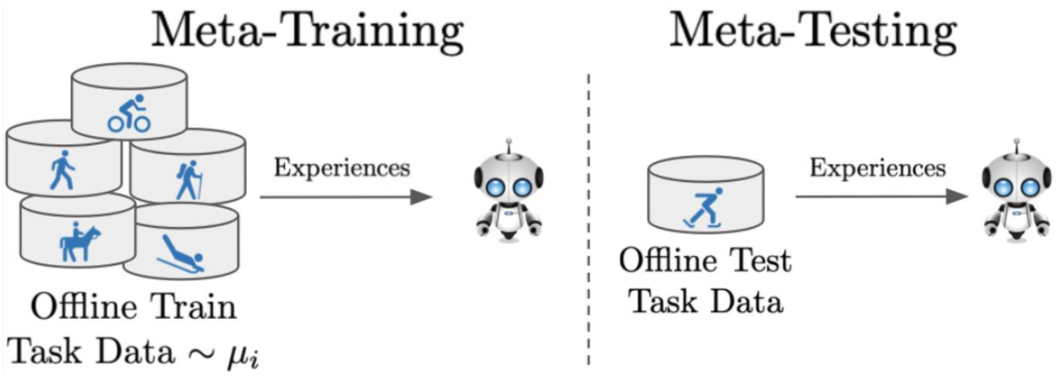
NeurIPS 2023

Background

- Offline Reinforcement Learning



- Offline Meta-Reinforcement Learning(OMRL):



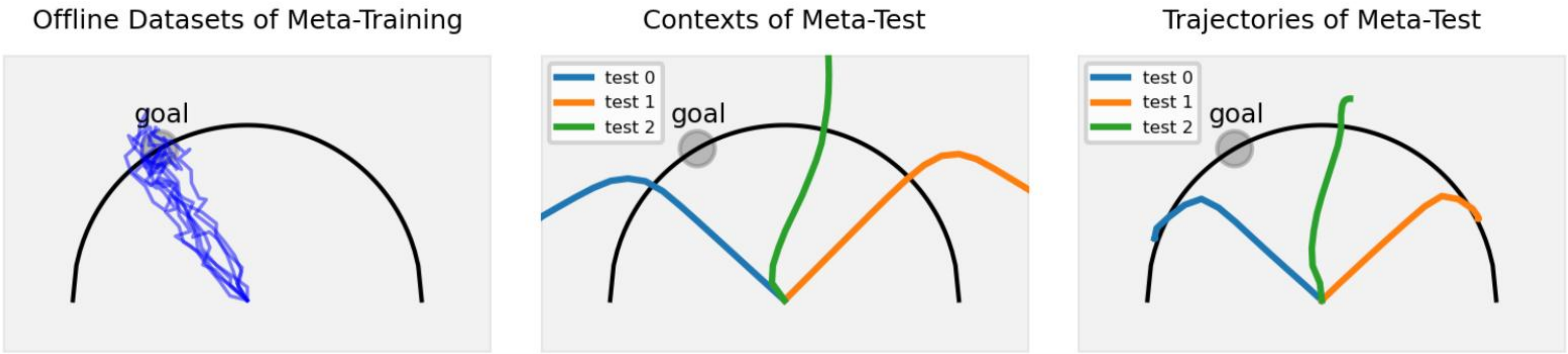
Problem

- context shift:
    - context from behavior policy during meta-traning
    - context from exploraion policy during meta-testing

  behavior policy ≠ exploraion policy

Env	Point-Robot		Half-Cheetah-Vel	
	context A	context B	context A	context B
FOCAL	**-4.4**±0.1	-14.9±1.1	**-45.7**±2.7	-69.5±9.6
OffPearl	**-5.1**±0.1	-17.8±1.5	**-123.0**±11.5	-162.8±28.8

Motivation



- Eliminate information about behavioral policy

- Weakening the impact of exploration policy during testing

Method

- Max-min Mutual Information Representation Learning:
  - maximize the MI with task (maxMI)

$$L_{maxMI}(\phi) = 1\{y_i = y_j\}\|z_i - z_j\|_2^2 + 1\{y_i \neq y_j\}\frac{\beta}{\|z_i - z_j\|_2^n + \epsilon}$$

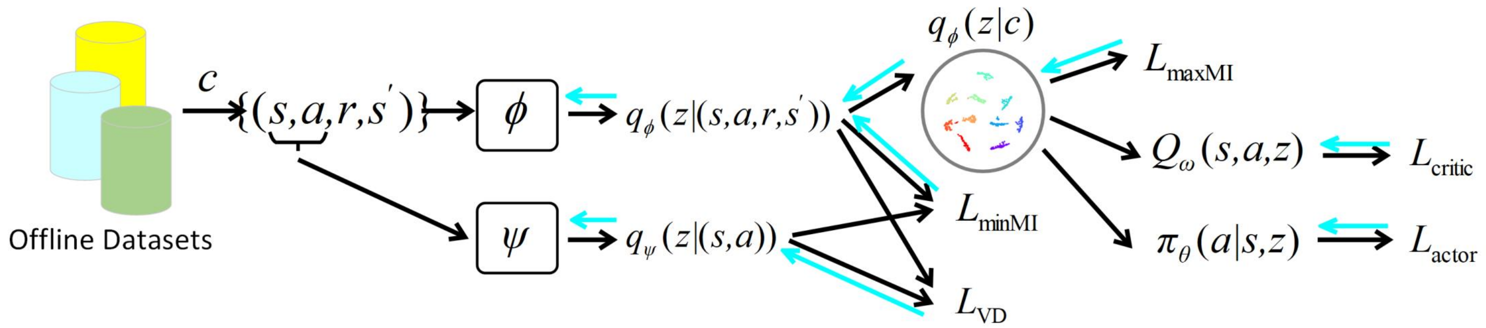
  - minimize the MI with behavior policy (minMI)

$$I_{CLUB}(z,(s,a)) = \mathbb{E}_i[\log p(z_i|(s_i,a_i)) - \mathbb{E}_j[\log p(z_j|(s_i,a_i))]].$$

$$L_{VD}(\psi) = -\mathbb{E}_{M\sim p(M)}\mathbb{E}_i[\log q_\psi(z_i|(s_i,a_i))]$$

$$L_{minMI}(\phi) = \mathbb{E}_{M\sim p(M)}\mathbb{E}_i[\log q_\psi(z_i|(s_i,a_i)) - \mathbb{E}_j[\log q_\psi(z_j|(s_i,a_i))]]$$

Meta-Training Phase



Method

- Common exploration strategy

$$z_0 \sim p(z) \longrightarrow \text{context } c \longrightarrow q_\phi(z|c)$$

- Non-prior Context Collection Strategy(Np)

explore independently and randomly at each step

$$\downarrow$$

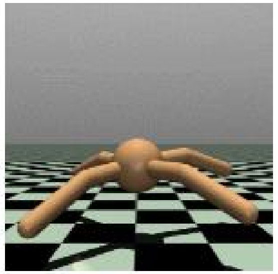
$$q_\phi(z|c)$$

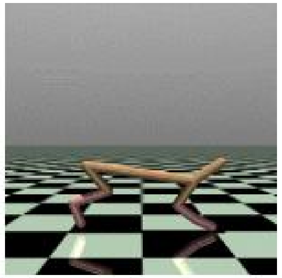
$$\downarrow \quad \uparrow$$

explore

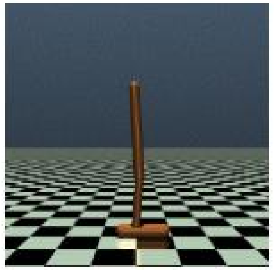
Experiments

- environments:
    - reward function change:
        - goal, velocity etc.





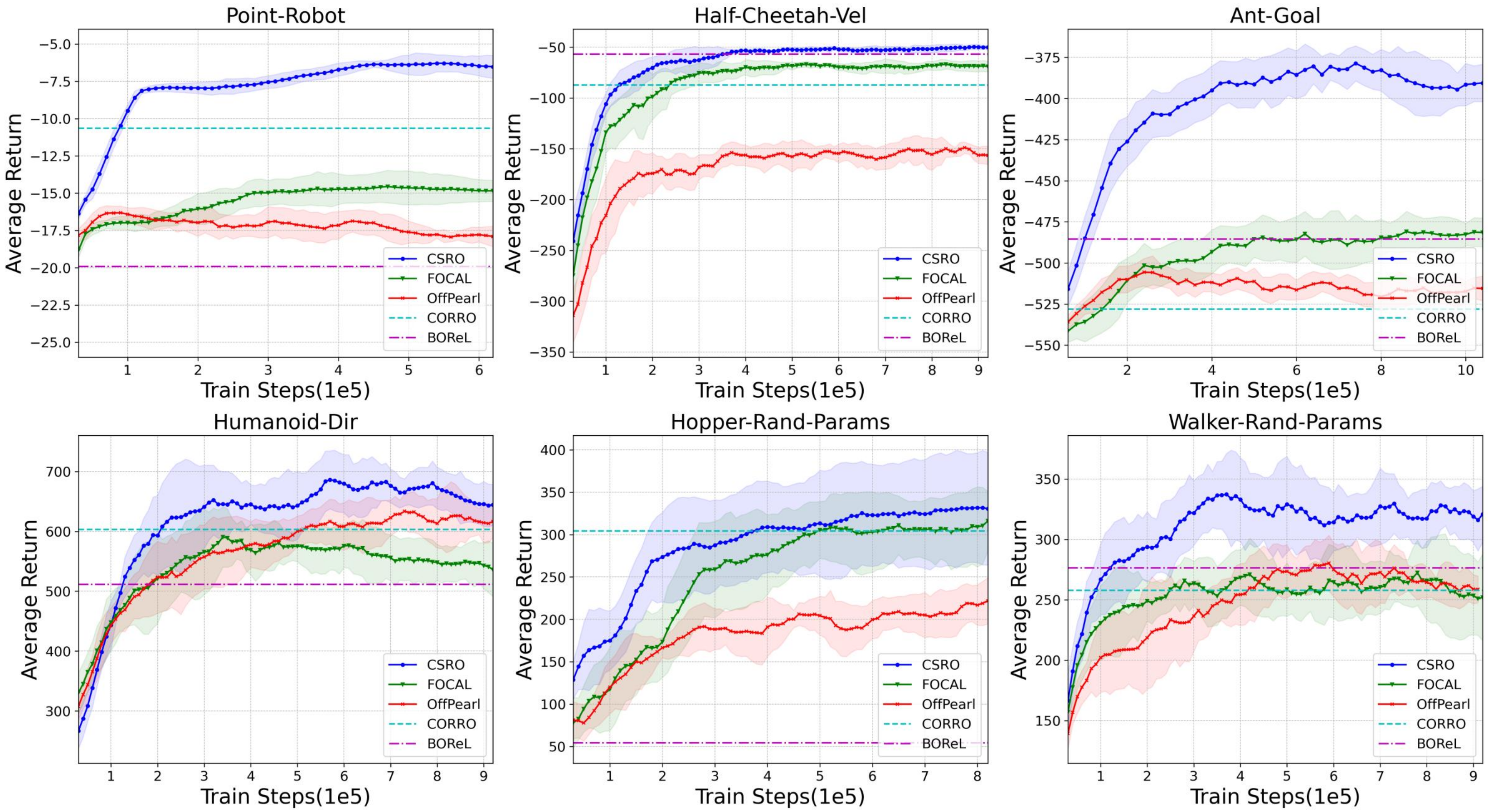
- dynamic function change:
    - mass, inertia, etc.



- datasets:
    - use SAC on each training task as behavior policy

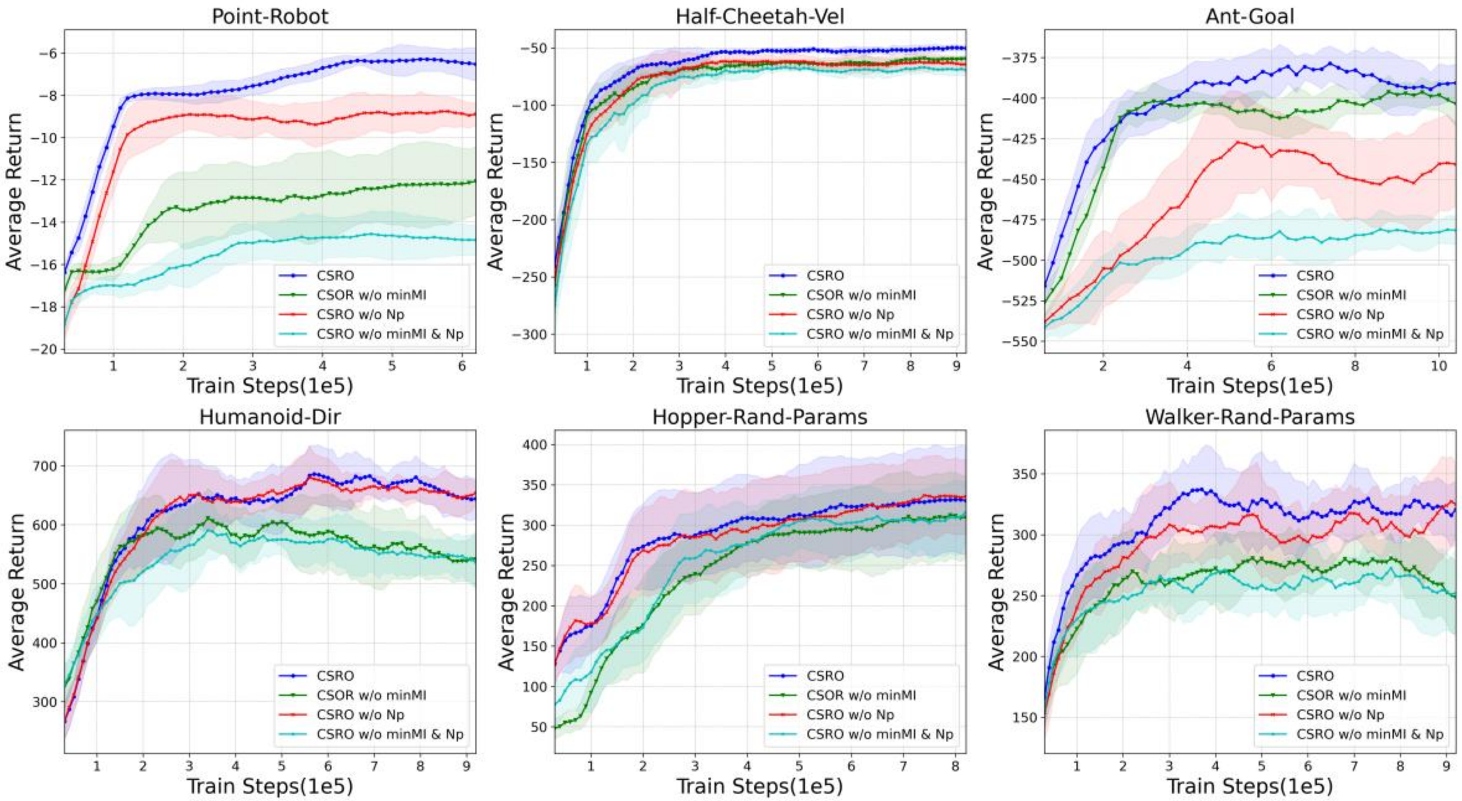
Experiments

- Main result: CSRO achieves the best performance



Experiments

- Ablation:
  - without minMI and Np components



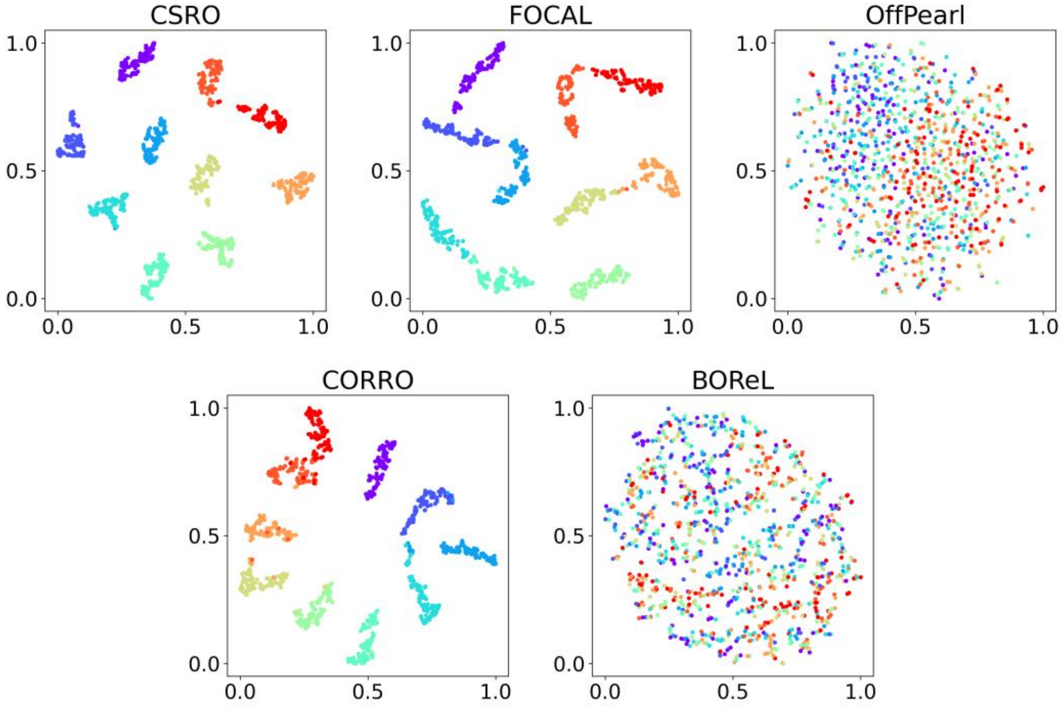
Experiments

- Ablation:
  - compare CSRO with other baselines without and with Np

Env	Point-Robot		Half-Cheetah-Vel		Walker-Rand-Params	
	w/ Np	w/o Np	w/ Np	w/o Np	w/ Np	w/o Np
CSRO	**-6.4**±0.8	-9.2±0.6	**-48.4**±3.9	-68.5±13.9	**344.2**±38.0	319.7±38.4
FOCAL	-11.8±1.6	-14.9±1.1	-60.9±5.7	-69.5±9.6	253.3±42.7	247.5±29.4
OffPearl	-17.0±1.6	-17.8±1.5	-133.7±18.9	-162.8±28.8	284.5±30.9	262.0±24.5
CORRO	-7.8±1.9	-10.5±3.0	-65.6±9.3	-92.1±23.2	312.5±46.6	275.2±73.9
BOReL	-21.6±3.9	-23.2±5.8	-90.1±28.3	-56.1±10.7	260.6±40.2	245.8±32.9

Experiments

- visualize the task representations



Thanks!