



Lanqing Li^{1,2*}, Hai Zhang^{3*}, Xinyu Zhang⁴, Shatong Zhu³, Yang Yu², Junqiao Zhao³, Pheng-Ann Heng²

Zhejiang Lab¹, The Chinese University of Hong Kong², Tongji University³, Stony Brook University⁴



之江实验室
ZHEJIANG LAB



NEURAL INFORMATION
PROCESSING SYSTEMS

Why Offline Meta-RL (OMRL)?

Offline RL

Meta-RL

Safety	✓	Safety	✗
Cost	✓	Cost	✗
Adaptation	✗	Adaptation	✓
Generalization	✗	Generalization	✓

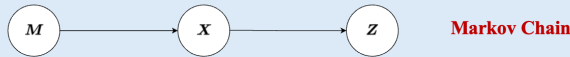
Problem Setup

Context-based OMRL (COMRL) seeks an optimal universal policy conditioning on a task representation \mathbf{z}^i for any task/MDP M^i :

$$\pi(\mathbf{a}|\mathbf{s}, \mathbf{z}^i) = \arg \max \sum_{t=0}^{H-1} \gamma^t \mathbb{E}_{\mathbf{s}_t \sim \mu_{\pi}^i(\mathbf{s}), \mathbf{a}_t \sim \pi} [R^i(\mathbf{s}_t, \mathbf{a}_t)], \forall M^i$$

Task Representation Learning in COMRL

Definition 1 Given an input context variable $\mathbf{X} \in \mathcal{X}$ and its associated task/MDP random variable $M \in \mathcal{M}$, task representation learning in COMRL aims to find a sufficient statistics \mathbf{Z} of \mathbf{X} with respect M .



Pre-existing algorithms propose seemingly disconnected objectives for task representation learning in COMRL:

• FOCAL [ICLR 2021]

$$\mathcal{L}_{\text{FOCAL}} = \min_{\phi} \mathbb{E}_{i,j} \left\{ \mathbb{1}\{i=j\} \|\mathbf{z}^i - \mathbf{z}^j\|_2^2 + \mathbb{1}\{i \neq j\} \frac{\beta}{\|\mathbf{z}^i - \mathbf{z}^j\|_2^2 + \epsilon} \right\}$$

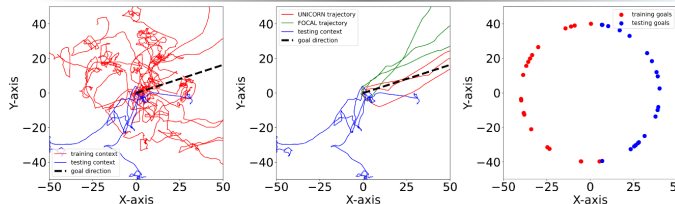
• CORRO [ICML 2022]

$$\mathcal{L}_{\text{CORRO}} = \min_{\phi} \mathbb{E}_{\mathbf{x}, \mathbf{z}} \left[-\log \left(\frac{h(\mathbf{x}, \mathbf{z})}{\sum_{M^* \in \mathcal{M}} h(\mathbf{x}^*, \mathbf{z})} \right) \right]$$

• CSRO [NeurIPS 2023]

$$\mathcal{L}_{\text{CSRO}} = \min_{\phi} \{ \mathcal{L}_{\text{FOCAL}} + \lambda \mathbb{E}_i [\log q_{\phi}(\mathbf{z}_i | \mathbf{s}_i, \mathbf{a}_i) - \mathbb{E}_j [\log q_{\phi}(\mathbf{z}_j | \mathbf{s}_j, \mathbf{a}_j)]] \}$$

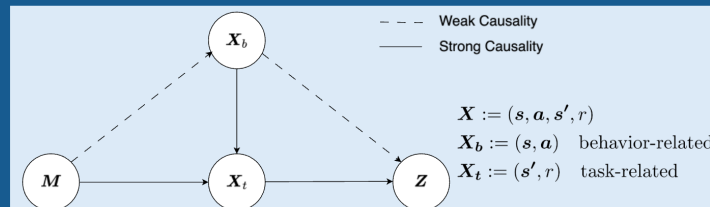
Challenges



Context shift of COMRL. Since the offline training data are *static*, the agent could encounter severe context shift in state-action distribution (left) or task distribution (right) at test time.

UNICORN: A Unified Framework

Decomposition of Input Data by Causality



Theorem 1 (Central Theorem). Let \equiv denote equality up to a constant, then

$$\underbrace{I(\mathbf{Z}; \mathbf{X}_t | \mathbf{X}_b)}_{\text{primary causality}} \leq I(\mathbf{Z}; M) \leq I(\mathbf{Z}; \mathbf{X}_t | \mathbf{X}_b) + I(\mathbf{Z}; \mathbf{X}_b) = \underbrace{I(\mathbf{Z}; \mathbf{X})}_{\text{primary + lesser causality}}$$

holds up to a constant, where

- $\mathcal{L}_{\text{FOCAL}} \equiv -I(\mathbf{Z}; \mathbf{X})$.
- $\mathcal{L}_{\text{CORRO}} \equiv -I(\mathbf{Z}; \mathbf{X}_t | \mathbf{X}_b)$.
- $\mathcal{L}_{\text{CSRO}} \geq -((1-\lambda)I(\mathbf{Z}; \mathbf{X}) + \lambda I(\mathbf{Z}; \mathbf{X}_t | \mathbf{X}_b))$.

Take-away Message

$I(\mathbf{Z}; M)$ provides a unified learning objective and is **robust** to context shift, by trading off the primary and lesser causalities of COMRL.

Results

The central theorem offers ample implementation choices for $I(\mathbf{Z}; M)$. This paper investigates 2 examples:

• Supervised UNICORN:

$$\mathcal{L}_{\text{UNICORN-SUP}} = -\mathbb{E}_{\mathbf{x}, \mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x})} \left[\sum_{j=1}^{n_M} \mathbb{1}(M^j = M) \log p_{\theta}(M^j | \mathbf{z}) \right]$$

• Self-supervised UNICORN:

$$\mathcal{L}_{\text{UNICORN-SS}} = -\mathbb{E}_{\mathbf{x}_t, \mathbf{x}_b, \mathbf{z} \sim q_{\phi}(\mathbf{z}|\mathbf{x}_t, \mathbf{x}_b)} [\log p_{\theta}(\mathbf{x}_t | \mathbf{z}, \mathbf{x}_b)] + \frac{\alpha}{1-\alpha} \mathcal{L}_{\text{FOCAL}}$$

Experiments

The proposed implementations achieves **competitive** in-distribution performance and **remarkable** out-of-distribution generalization across a wide range of RL domains, OOD settings, data qualities and model architectures.

Table 2: Average testing returns of UNICORN against baselines on datasets collected by IID and OOD behavior policies. Each result is averaged by 6 random seeds. The best is **bolded** and the second best is underlined.

Algorithm	HalfCheetah-Dir		HalfCheetah-Vel		Ant-Dir		Hopper-Param		Walker-Param		Reach	
	IID	OOD	IID	OOD	IID	OOD	IID	OOD	IID	OOD	IID	OOD
UNICORN-SS	1307±26	1296±24	-22±1	-94±5	<u>267±14</u>	236±18	316±6	304±11	407±46	2775±241	2694±183	
UNICORN-SUP	1296±20	1130±76	-25±1	-91±5	239±4	239±16	112±4	302±12	322±28	312±39	2681±111	2641±40
CSRO	1180±228	458±233	-28±1	-102±5	276±19	233±12	310±6	301±10	310±58	279±65	2720±235	2801±82
CORRO	704±450	245±146	-37±1	-112±2	148±13	120±12	283±8	272±15	277±38	313±48	2468±175	2322±227
FOCAL	1186±272	861±253	-22±1	-97±2	217±29	173±24	302±4	297±13	308±98	286±91	2424±256	2316±303
Supervised	962±356	782±429	-24±1	-104±1	238±39	202±38	306±10	294±8	256±60	210±28	2489±248	2285±205
MACAW	1155±10	450±6	-56±2	-108±1	26±3	0±0	218±6	205±2	141±9	130±5	2431±157	1728±79
Prompt-DT	1176±40	-25±9	-118±6	-249±21	140	0±0	234±5	202±5	185±9	156±17	2165±85	1896±111

Table 3: UNICORN vs. baselines on Ant-Dir datasets of various qualities. Each result is averaged by 6 random seeds. The best is **bolded** and the second best is underlined.

Algorithm	Random		Medium		Expert	
	IID	OOD	IID	OOD	IID	OOD
UNICORN-SS	81±18	62±6	220±23	243±10	279±10	262±13
UNICORN-SUP	75±15	60±5	160±11	170±3	247±15	220±19
CSRO	2±3	0±1	166±10	198±17	252±39	202±45
CORRO	1±1	0±0	8±5	-7±2	-4±10	-14±9
FOCAL	67±26	44±10	171±84	187±86	229±42	246±20
Supervised	65±8	47±12	149±30	110±80	249±33	215±60
MACAW	3±1	0±0	28±2	1±1	88±43	1±1
Prompt-DT	1±0	0±0	2±4	0±1	78±15	1±2

Table 4: DT implementation of COMRL on HalfCheetah-Dir and Hopper-Param. Each result is averaged by 6 random seeds.

Algorithm	HalfCheetah-Dir		Hopper-Param	
	IID	OOD	IID	OOD
UNICORN-SS	1307±26	1296±24	316±6	304±11
UNICORN-SS-DT	1233±10	1186±43	304±4	291±4
UNICORN-SUP-DT	1227±21	1065±77	308±6	297±2
FOCAL-DT	1209±33	652±36	293±4	284±5
Prompt-DT	1177±40	-25±9	234±5	203±5

ArXiv



Code

