Cost-aware Offline Safe Meta Reinforcement Learning with Robust In-Distribution Online Task Adaptation

Guan Cong, Ruiqi Xue, Ziqian Zhang, Lihe Li, Yi-Chen Li, Lei Yuan, Yang Yu*
Nanjing University, Polixir Technologies
accepted by AAMAS2024 oral

Background & Motivation

Real-world tasks offten come with **safety concerns**. Directly interacting with the real-world environment for safety-concerned trial-and-error incurs various costly actions, thus **offline safe RL** matters.

Real-world datasets may contain data generated under **diverse safety constraints**, and different constraints may **conflit** with each other, thus leveraging **context-based meta RL for task identification** is critical.

A safety-concerned environment is modeled as a Constrained Markov Decision Process (CMDP):

$$\mathcal{M} = \langle S, A, P, R, C, \gamma, b \rangle$$

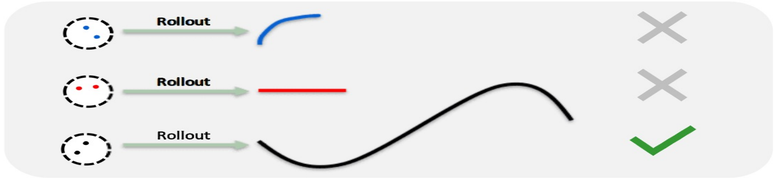
Cost-aware context encoder learning:

- **Distance metric learning loss** for distinguishing tasks.
- **Cost contrastive loss** using cost model relabeling.
- **Cost decoding loss** to ensure informative expression.

Safe in-distribution online adaptation:

- Selecting in-distribution trajectories as contexts-**safe truncated in-distribution score**

$$\mathrm{SIDS}_{\mathrm{trunc}}(\tau) = \begin{cases} R(\tau_{1:t_{tc}}) \;\; C(\tau_{1:t_{tc}}) \le b + \epsilon_1, R(\tau_{1:t_{tc}}) \ge \delta \\ -C(\tau_{1:t_{tc}}) \;\; C(\tau_{1:t_{tc}}) > b + \epsilon_1, R(\tau_{1:t_{tc}}) \ge \delta \\ -K \;\; R(\tau_{1:t_{tc}}) < \delta \end{cases}$$



Performance on different benchmarks

Task		COSTA		FOCAL		FOCAL_oracle		IDAQ		CORRO		PEARL		Vanilla	
		reward↑	cost↓	reward↑	cost↓	reward↑	cost↓	reward↑	cost↓	reward↑	cost↓	reward↑	cost↓	reward↑	cost↓
AntDir	Training	1071±33.75	0.40±0.05	1097±28.30	0.39±0.10	/	/	/	/	1064±17.65	0.56±0.45	1048±52.57	0.66±0.29	1529±83.85	2.59±0.15
	Adaptation	1071±35.70	0.40±0.06	1010±140.38	4.00±1.90	1103±28.08	0.44±0.05	757±67.16	4.51±0.70	870±197.22	5.04±0.35	771±32.35	5.50±0.14	/	/
CheetahVel	Training	354±31.70	0.89±0.12	254±189.34	0.65±0.37	/	/	/	/	329±56.61	1.74±0.89	169±177.03	1.70±1.41	136±195.84	1.73±1.35
	Adaptation	348±40.69	0.84±0.21	137±28.64	0.17±0.12	245±170.51	0.40±0.18	190±27.18	0.14±0.06	319±101.15	2.75±1.05	138±188.25	1.60±1.69	/	/
AntWalk	Training	495±17.98	0.12±0.17	506±5.02	0.06±0.04	/	/	/	/	496±8.32	0.10±0.11	528±12.03	7.07±0.73	523±16.61	7.93±0.26
	Adaptation	490±15.62	0.09±0.12	537±17.07	5.42±0.11	502±7.48	0.11±0.07	415±50.91	3.64±3.28	511±34.32	6.54±2.50	525±9.89	5.46±2.95	/	/
CheetahWalk	Training	260±72.97	0.00±0.00	234±103.93	1.19±2.38	/	/	/	/	263±175.02	6.64±6.43	216±126.89	0.40±0.80	502±108.14	14.20±0.02
	Adaptation	248±66.97	0.00±0.00	296±120.04	8.51±6.95	225±117.70	0.72±1.44	298±105.74	2.84±5.68	244±159.75	13.06±2.30	162±42.27	11.37±5.68	/	/
AntCircle	Training	1181±219.65	0.55±0.36	1019±351.95	1.01±0.66	/	/	/	/	1232±418.41	1.05±0.70	982±287.94	0.67±0.24	1869±812.58	3.21±0.53
	Adaptation	1147±118.63	0.58±0.35	727±278.45	2.78±1.11	923±396.39	0.92±0.90	730±107.51	2.02±1.1	1001±518.52	2.24±1.39	1217±684.25	2.64±0.86	/	/
AntGoal	Training	819±14.80	0.64±0.10	826±15.54	0.57±0.17	/	/	/	/	824±8.55	0.69±0.14	810±6.48	0.58±0.23	892±12.01	1.14±0.13
	Adaptation	827±20.86	0.75±0.21	851±18.92	2.28±1.02	824±19.33	0.60±0.17	852±13.26	2.33±1.15	810±40.78	2.9±0.67	853±24.28	1.57±0.74	/	/
Average	Training	697	0.43	656	0.65	/	/	/	/	701	1.80	626	1.85	909	5.13
	Adaptation	689	0.45	593	3.86	637	0.53	583	2.58	599	5.42	611	4.66	/	/

Visualization, Generalization and Transfer

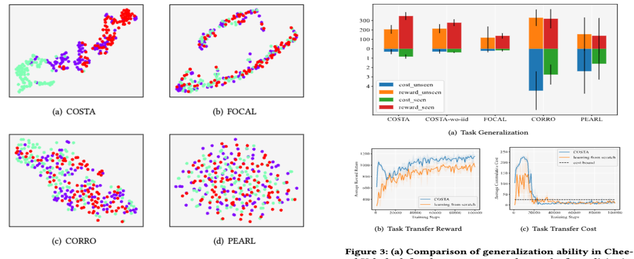
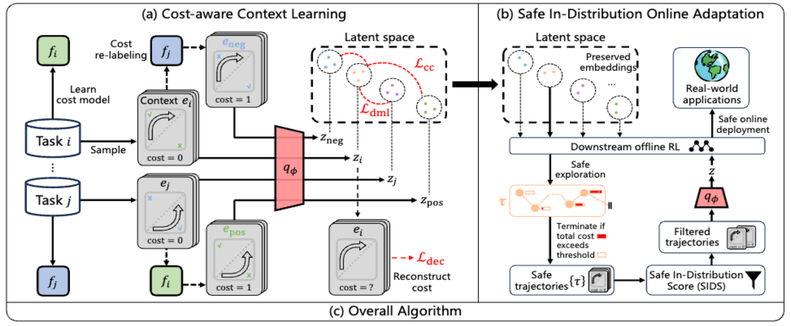


Figure 4: The 2D projection of the learned task representation space in AntDir using t-SNE. Here, all contexts from different tasks are collected using a same behavioral policy.

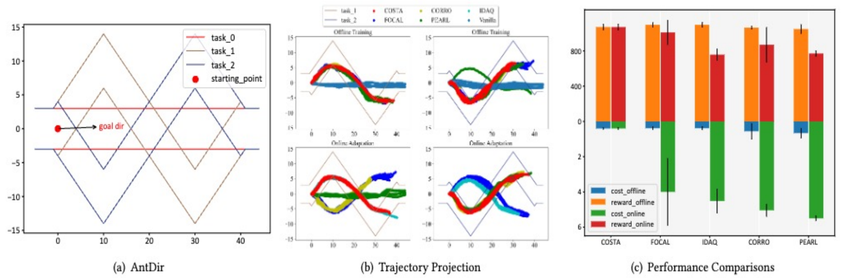
Figure 3: (a) Comparison of generalization ability in CheetahVel, the left column represents the results for policies in unseen tasks, while the right column represents the results for policies in seen tasks. (b), (c) Training results of being transferred to unseen tasks in AntDir.

Method



Experiments

Motivated Example



Ablation Studies

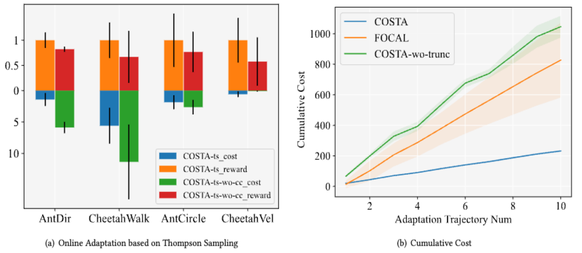


Figure 6: (a) Ablation study on the performance of COSTA with or without cost contrastive loss during online adaptation based on Thompson sampling. The rewards are normalized by COSTA-ts's reward. (b) Ablation study on cumulative cost during the whole process of online adaptation in AntDir.