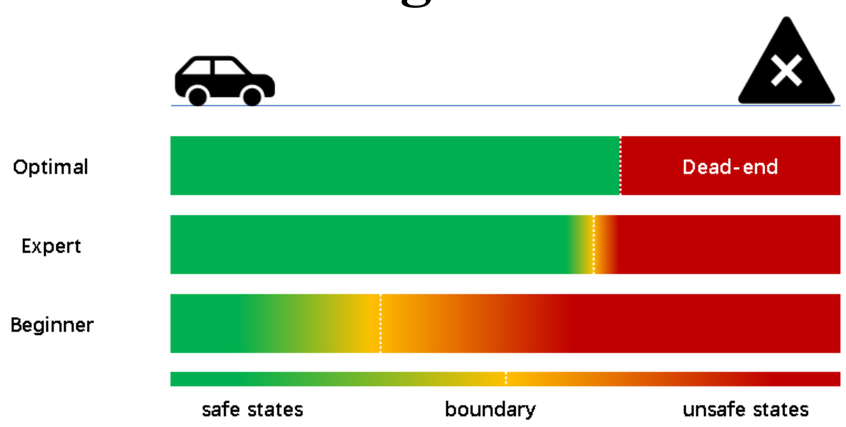
Xiao Zhang, Hai Zhang, Hongtu Zhou, Chang Huang, Di Zhang, Chen Ye*, Junqiao Zhao*

Safe Reinforcement Learning with Dead-ends Avoidance and Recovery, RA-L, 2023

Motivation

- Existing safe-based methods tend to adopt overly conservative policies.
- The reason is that they assess the safety of states and actions use the task policy.
- The epistemic bias cannot be corrected without violating safety.

Dead-ends Recognition



Measuring the safety of states **using the optimal safe policy,** which is trained to **minimize the cost and ignore the environmental reward**,

$$V_c^{\pi}(s) = \mathbb{E}_{\tau\sim\pi,P}[\sum_{t=0}^{\infty}\gamma_{safe}^{t}\mathcal{C}(s_t,a_t,s_{t+1})|s_0=s]$$

It's possible to **partition the dead-ends** in a deterministic Markov setting **using the state cost function corresponding to the optimal safety policy.**

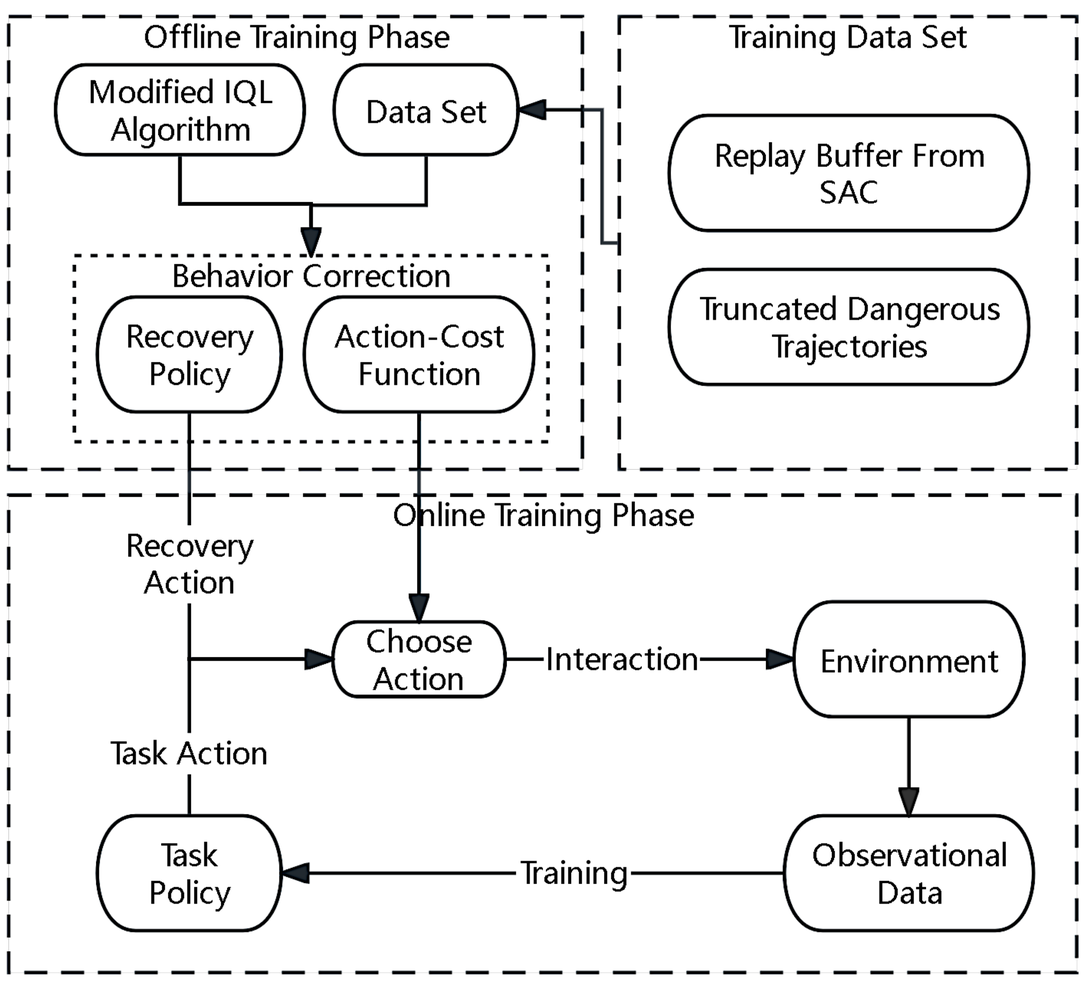
$$S_{dead} = \{s|s\in\mathcal{S}\ and\ \gamma_{safe}^{H-1} \le V_c^*(s) \le 1\}$$

Dead-ends Avoidance

$$A(s)_{safe} = \{a|a\in\mathcal{A}\ and\ Q_c^*(s,a) < \gamma_{safe}^{H}\}$$

$$a_t = \begin{cases} a^{\pi_{task}}, Q_c^{\bar{\pi}}(s_t, a^{\pi_{task}}) < \epsilon_{safe} \\ a^{\pi_{rec}}, otherwise \end{cases}$$

Algorithm



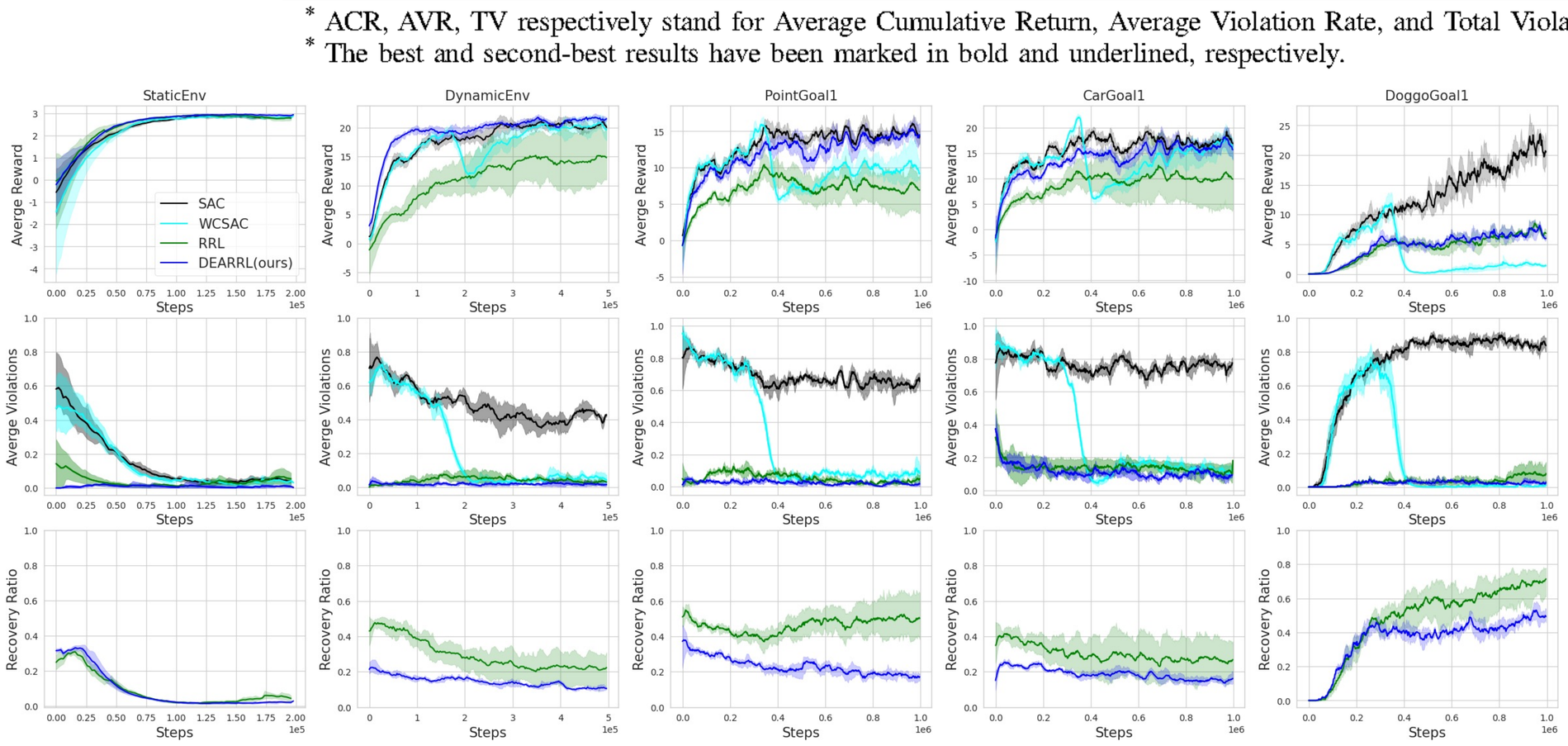
Theorem 1: $\Pi_c^{\pi_{com}}$ and $\Pi_c^{\pi_{rec}}$ are the accessible space for π to explore safely in RRL and DEARRL respectively, we have $\Pi_c^{\pi_{com}} \subseteq \Pi_c^{\pi^*_{rec}}$.

Results

TABLE I
RESULTS OF TRAINING AND TESTING EACH METHOD INDIVIDUALLY

Environments	SAC			WCSAC			RRL			DEARRL(ours)			IQL	
	ACR	AVR	TV	ACR	AVR	TV	ACR	AVR	TV	ACR	AVR	TV	ACR	AVR
StaticEnv	2.754	0.043	1642	**2.936**	**0.008**	1453	2.813	0.032	392	2.882	0.023	**119**	2.630	0.095
DynamicEnv	17.434	0.519	3330	19.004	0.069	1451	14.230	**0.006**	112	**20.784**	0.018	**86**	18.044	0.565
PointGoal1	12.609	0.753	12375	7.811	0.055	5739	7.203	0.094	476	**14.079**	**0.035**	**236**	12.740	0.76
CarGoal1	16.727	0.73	14850	**18.672**	**0.035**	6199	9.647	0.057	1109	15.430	0.091	**786**	11.357	0.94
DoggoGoal1	16.956	0.908	31450	1.055	**0.013**	10825	5.924	0.192	**623**	7.002	0.048	706	**18.889**	0.838

* ACR, AVR, TV respectively stand for Average Cumulative Return, Average Violation Rate, and Total Violations in training.
* The best and second-best results have been marked in bold and underlined, respectively.



Open-sourced!

https://github.com/tiev-tongji/dea-rrl