Policy regularized Offline safe RL with Preference aligned sampling

Speaker: Cheng Tang

Department of Mechanical Engineering

Carnegie Mellon University

Contents

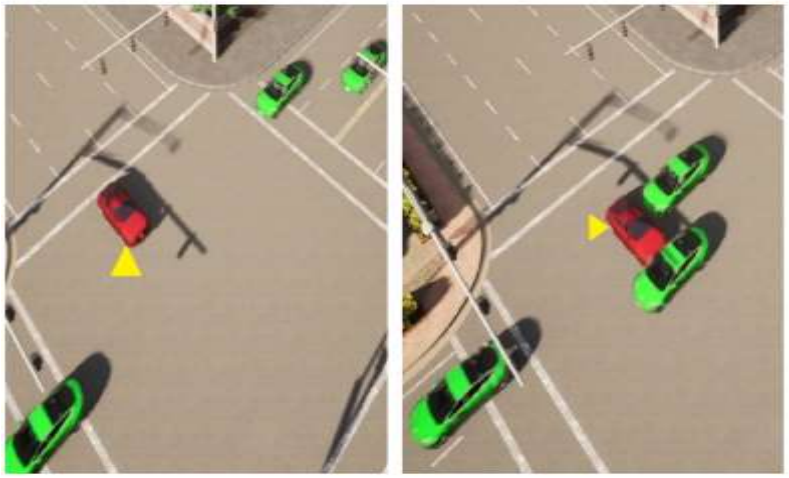
Carnegie Mellon University

Introduction

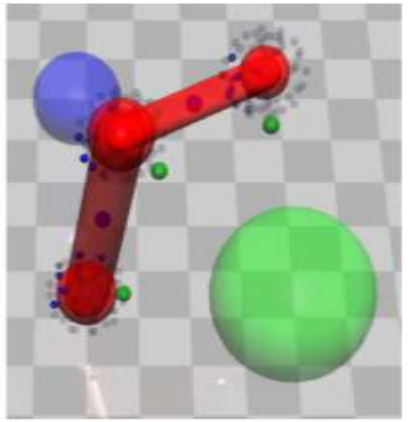
Carnegie Mellon University

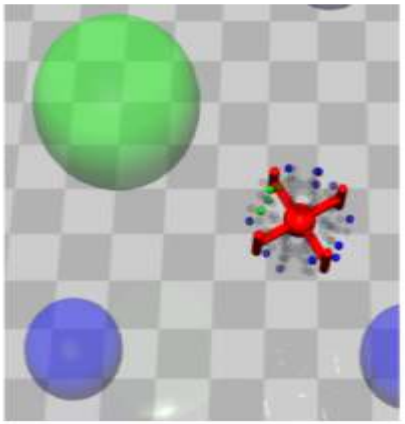
■ Application in Safe Reinforcement Learning

Autonomous Driving tasks

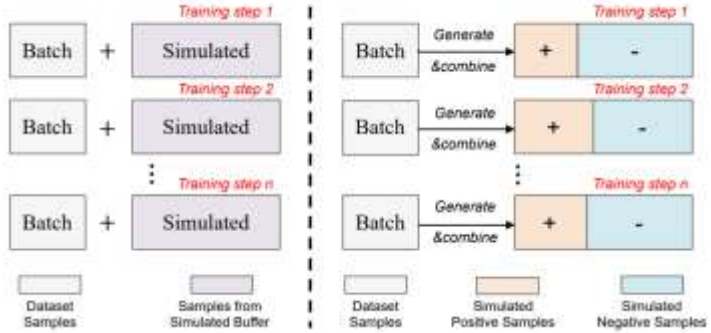


Robotic tasks





Control



[1] Zhili Zhang, Songyang Han, Jiangwei Wang, and Fei Miao. Spatial-temporal-aware safe multiagent reinforcement learning of connected autonomous vehicles in challenging scenarios. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 5574–5580. IEEE, 2023.

[2] Weiye Zhao, Rui Chen, Yifan Sun, Ruixuan Liu, Tianhao Wei, and Changliu Liu. Guard: A safe reinforcement learning benchmark. arXiv preprint arXiv:2305.13681, 2023.

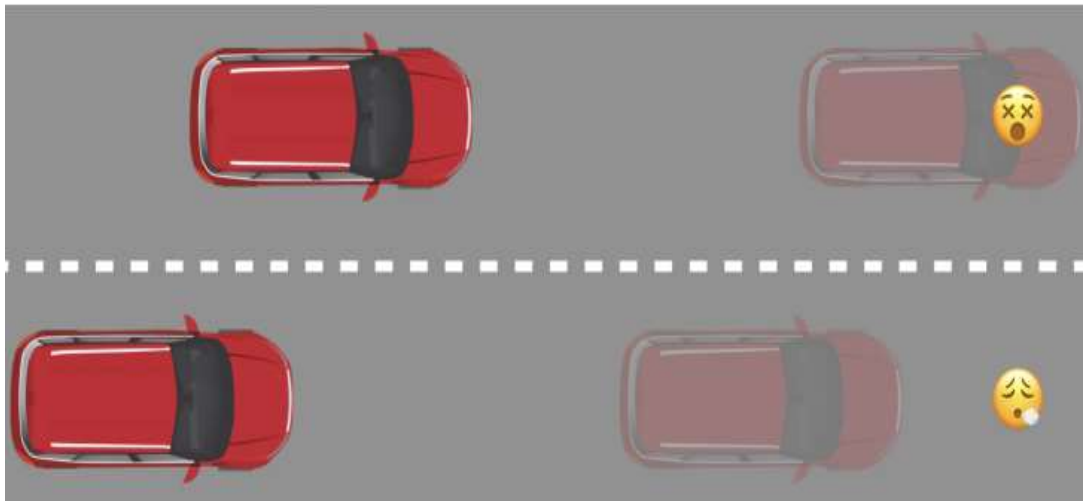
[3] Xianyuan Zhan, Haoran Xu, Yue Zhang, Xiangyu Zhu, Honglei Yin, and Yu Zheng. Deepthermal: Combustion optimization for thermal power generating units using offline reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 4680–4688, 2022.

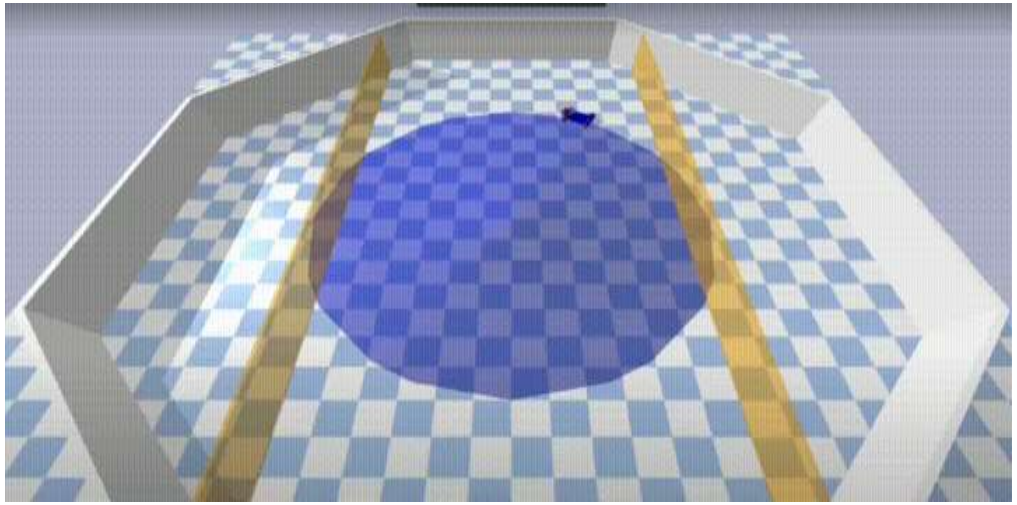
Introduction

Carnegie Mellon University

Safety Constrains in Reinforcement learning

- High velocity constrain
- Low velocity constrain
- Boundary constrain





Contents

Carnegie Mellon University

Problem formulation

Carnegie Mellon University

■ Multi-constrain Safe Reinforcement Learning

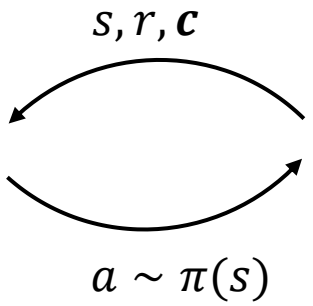
- Constrained Markov Decision Process(CMDP): $M = (S, A, P, r, \boldsymbol{c}, \gamma, \mu_0)$,
- State s, action a, transition kernel P, initial distribution μ_0
- reward $r: S \times A \rightarrow \mathrm{R}$, cost $\boldsymbol{c}: S \times A \rightarrow \mathrm{R}^N$, cost dimension N
- Value function: $V_r^\pi(\mu_0) = \mathbb{E}_{\tau\sim\pi} \sum_t \gamma^t r_t$, $V_{c_i}^\pi(\mu_0) = \mathbb{E}_{\tau\sim\pi} \sum_t \gamma^t c_i^t$, $i = 1, 2, 3 \ldots N$
- Safe RL problem:

$$\pi^* = \arg\max_\pi V_r^\pi \qquad s.t. \quad \boldsymbol{V_c^\pi} \leq \boldsymbol{\epsilon}$$

Threshold $\boldsymbol{\epsilon}$: max-tolerate cost vector for one trajectory

- behavior policy of the dataset and optimal policy

$$\mathbb{E}_{\tau\sim\pi^*}[R(\tau)] - \mathbb{E}_{\tau\sim\pi_r^*}[R(\tau)] \leq wL(\pi^*, \pi_B).$$

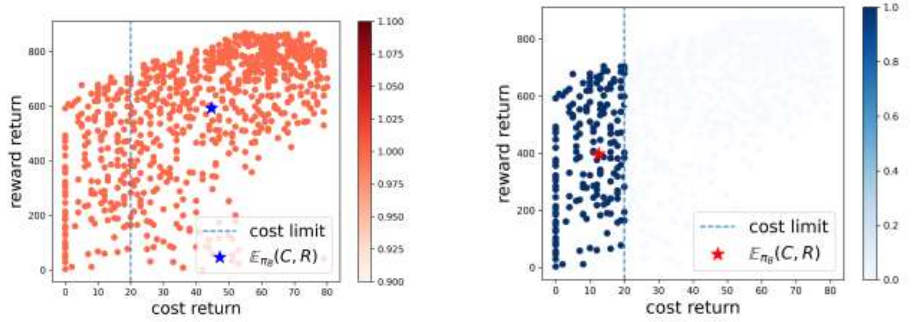


Introduction

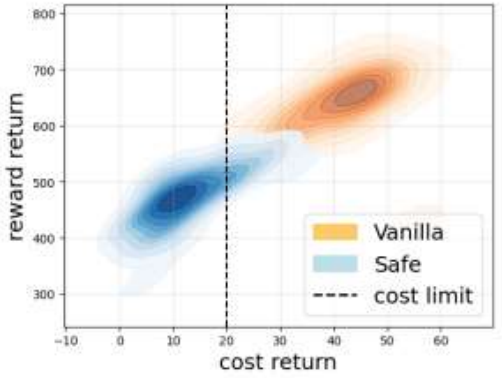
Carnegie Mellon University

Safety in Offline Multi-constrain Reinforcement Learning

- **Challenges in Offline multitask Reinforcement learning :**
  - **Distribution shift:** Limited Dataset
- **Challenges in Safety in Robot Reinforcement Learning:**
  - **Unfeasible behavior policy:** policy tend to ignore safety constrains when optimizing reward
  - **Conservative behavior policy:** policy tend to be overly conservative



(a) Unfeasible (b) Conservative



(c) Behavior policy π_B

Contents

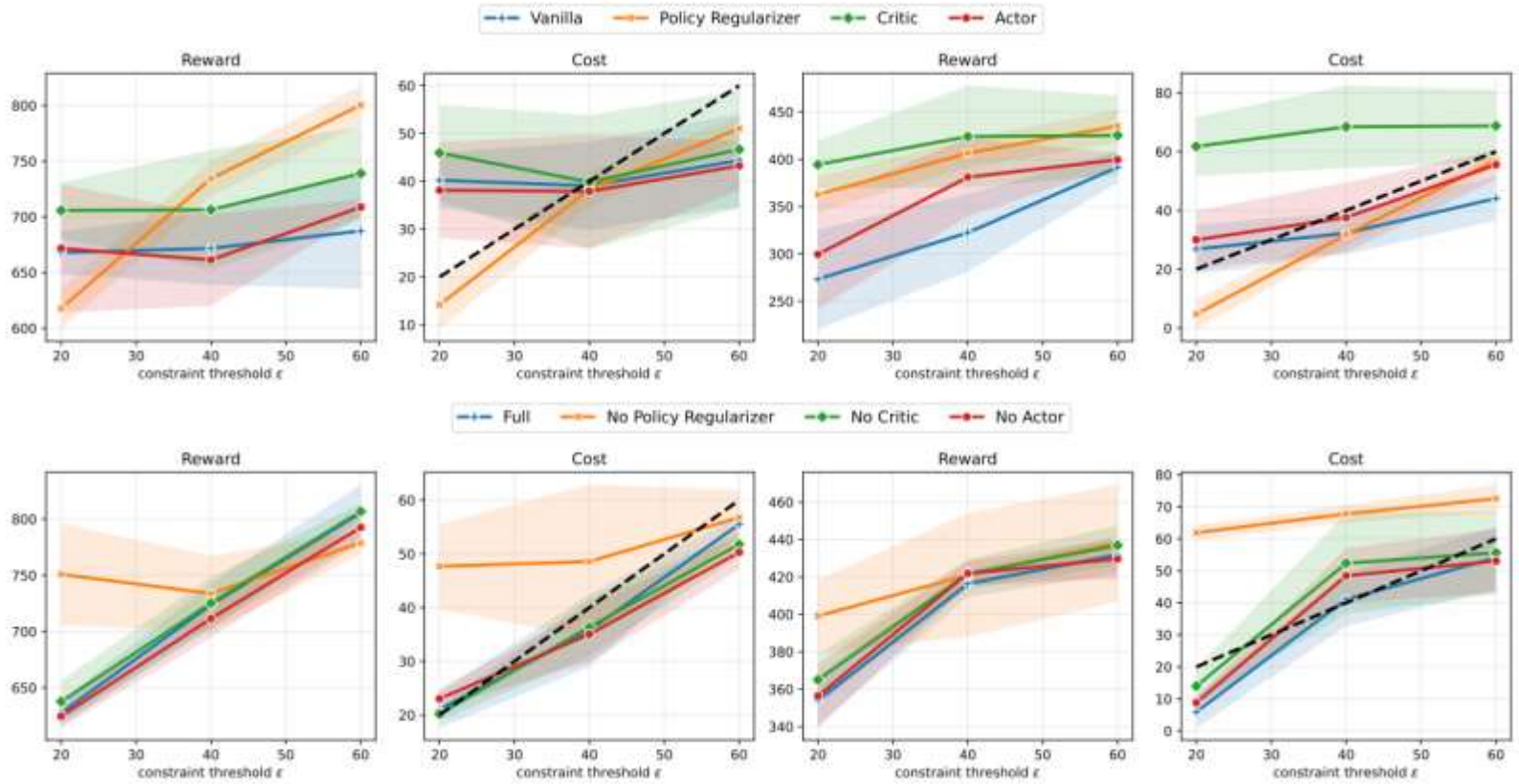
Carnegie Mellon University

Problem Identification

Carnegie Mellon University

■ **How does suboptimality of behavior policy affect policy-regularized safe offline RL algorithms?**

- **Three components of policy-regularized safe algorithm: Policy Regularizer, Critics and Actor**
- **the Policy Regularizer pushes the learned policy to be close to the behavior policy**
- **Critics and Actor show little variation under different dataset with same support.**

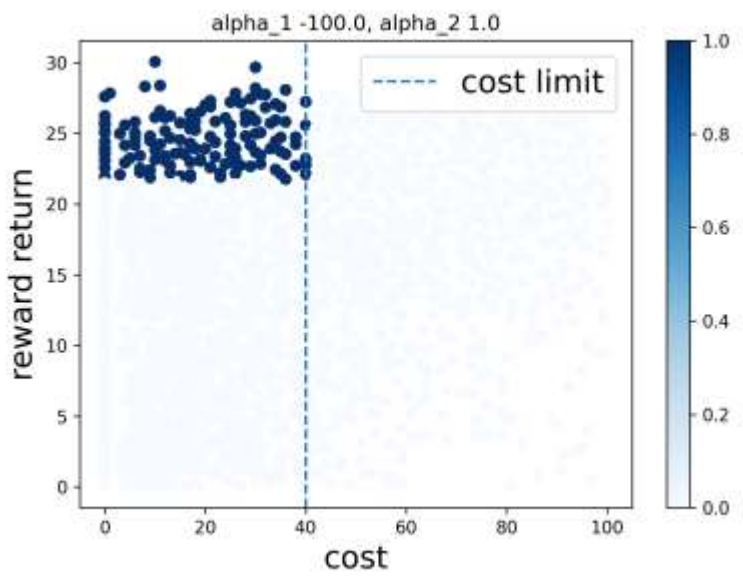


Methods

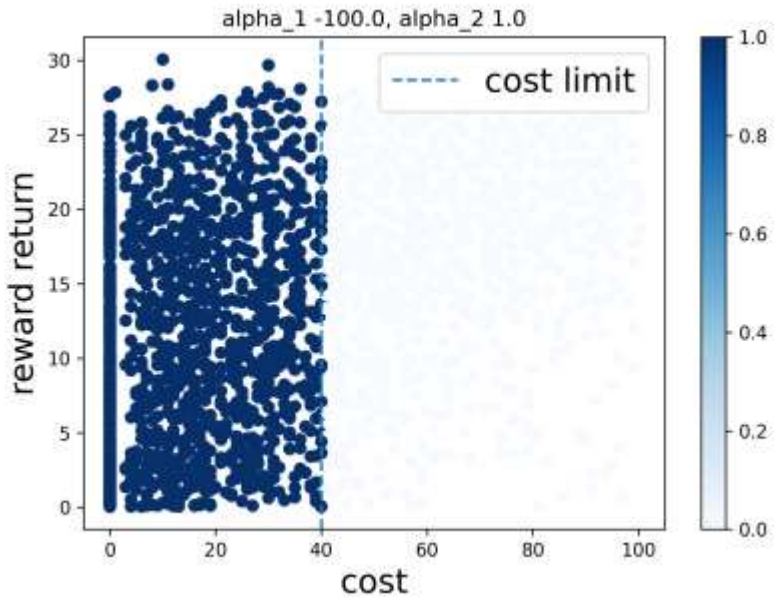
Carnegie Mellon University

■ Dataset resampling (achieving a better behavior policy)

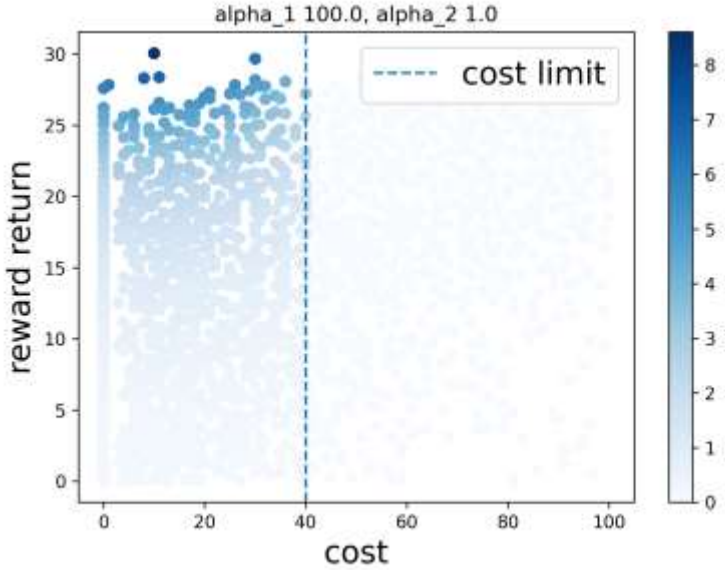
Safe top 10% dataset(too tempting)



Safe dataset(too conservative)



Ours(DIAM)



Methods

Carnegie Mellon University

- **Unbiased estimation of $E_{\tau\sim\pi_B}$**
  - **Suppose the trajectory τ_i is sampled with weight w_i, then the weighted state-action distribution**

$$d_W(s,a) = \sum_{i=1}^{N} w_i d_{\pi_B}(s)\pi_B(a|s)$$

  - **with the assumption that the discount factor of D_W is less than 1, the trajectory-wise expected return and cost of the weighted behavior policy can be bounded by the Hoeffding's inequality:**

$$\mathbb{P}\left[\left|\mathbb{E}_{\tau\sim\pi_W}[R(\tau)] - \sum_{k=1}^{N} w_i R(\tau_k)\right| \geq \epsilon\right] \leq 2\exp\left(\frac{2\epsilon^2}{\delta_R^2 \sum_{k=1}^{N} w_k^2}\right),$$

$$\mathbb{P}\left[\left|\mathbb{E}_{\tau\sim\pi_W}[C_i(\tau)] - \sum_{i=1}^{N} w_k C_i(\tau_k)\right| \geq \epsilon\right] \leq 2\exp\left(\frac{2\epsilon^2}{\delta_{C_i}^2 \sum_{k=1}^{N} w_k^2}\right),$$

Unbiased

Methods

Carnegie Mellon University

■ Optimize Weighted Behavior Policy π_B

- **Optimization objective** $\max_W \mathbb{E}_{\tau\sim\pi_W}[R(\tau)], \quad \text{s.t.} \quad \mathbb{E}_{\tau\sim\pi_W}[C_i(\tau)] \leq \kappa_i, i = 1, 2, 3...N.$

- **By the Unbiased estimation of $E_{\tau\sim\pi_B}$, we formalize the minimization problem as:**

$$\max_W \sum_{k=1}^{N} w_i R(\tau_k) + \beta H(W),$$

$H(\cdot)$ is information entropy that acts as a regularization, i . e, $H(W) = -\sum w_i log w_i$.

- **We hope the weighted sampled dataset can achieve at least as conservative as the behavior policy**

$$\frac{\kappa_i - \mathbb{E}_{\tau\sim\pi_W}[C_i(\tau)]}{\sqrt{\text{Var}_{\tau\sim\pi_W}[C_i(\tau)]}} \geq \frac{\kappa_i - \mathbb{E}_{\tau\sim\pi}[C_i(\tau)]}{\sqrt{\text{Var}_{\tau\sim\pi}[C_i(\tau)]}}, \forall i \in [N].$$

Contents

Carnegie Mellon University

Experiment

Carnegie Mellon University

■ We want to explore:

- **How does DIAM perform compared to other offline safe RL baselines ?**
- **How does DIAM benefit policy-regularized safe reinforcement learning method in different cost thresholds?**
- **How does DIAM applied to different policy-regularized offline safe RL algorithms?**

■ Baselines

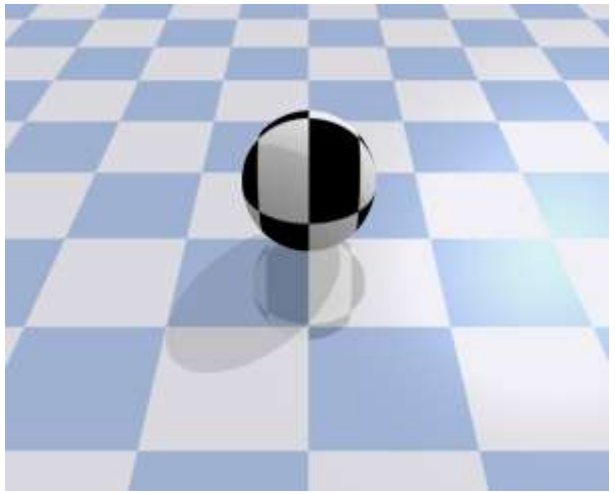
- **Model-centric approach**: (1) Imitation Learning: **BC** (Behavior cloning), **BC-safe** (Behavior cloning with safe dataset) (2) Q-Learning based methods: **BCQ-Lag**, **BEAR-Lag**, and **CPQ**; (3) Distribution Correction Estimation: **COptiDICE** (4) Sequential modeling: **CDT**.
- **Data-centric approach:** (1) **Vanilla-BCQ-Lag**, (2) **Safe-BCQ-Lag**, (3) **Safe-10-BCQ-Lag**, and (4) **Top-10-BCQ-Lag**.

Experiment

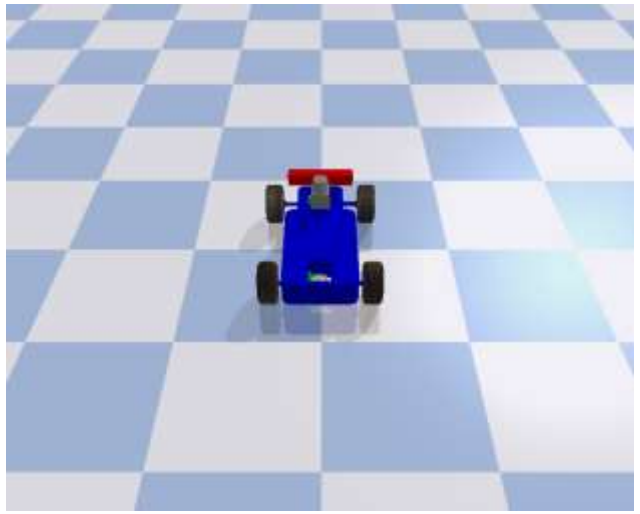
Carnegie Mellon University

Experiment task settings

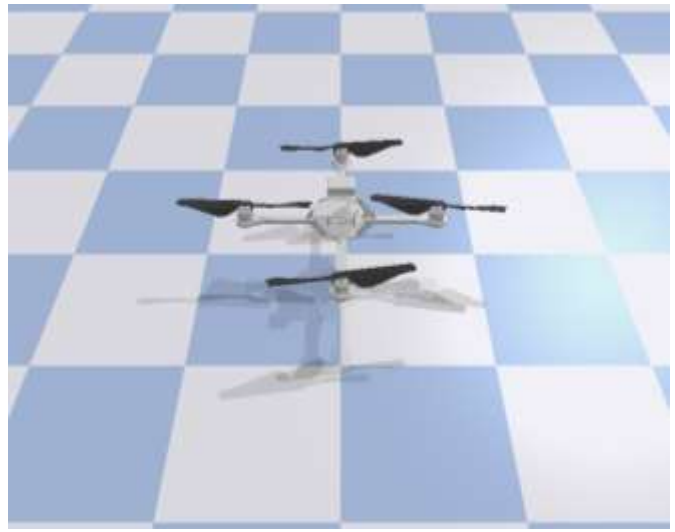
- Safe RL platform Bullet-Safety-Gym [1] with single and multi-constrains
- Diverse robotic dynamics.



Ball agent



Car agent



Drone agent

Experiment

Carnegie Mellon University

■ Q1: how is our method compared to other baselines? (Model-centric approach)

Algorithm	Stats	BallCircle	CarCircle	DroneCircle	BallRun	CarRun	DroneRun
BC	reward ↑	0.76 ± 0.03	0.50 ± 0.04	0.84 ± 0.05	0.73 ± 0.04	**0.98 ± 0.00**	0.42 ± 0.18
	cost ↓	2.24 ± 0.17	2.66 ± 0.88	3.19 ± 0.56	3.02 ± 0.16	**0.72 ± 0.78**	1.33 ± 1.23
BC-Safe	reward ↑	0.55 ± 0.03	0.35 ± 0.11	**0.62 ± 0.01**	**0.20 ± 0.01**	**0.98 ± 0.00**	**0.57 ± 0.02**
	cost ↓	1.08 ± 0.27	1.01 ± 0.28	**0.42 ± 0.17**	**0.89 ± 0.17**	**0.01 ± 0.00**	**0.27 ± 0.38**
BCQ-Lag	reward ↑	0.71 ± 0.04	0.59 ± 0.03	0.57 ± 0.87	0.04 ± 0.09	**0.91 ± 0.04**	0.67 ± 0.05
	cost ↓	1.96 ± 0.26	1.78 ± 0.14	3.57 ± 0.49	1.97 ± 1.50	**0.00 ± 0.00**	5.28 ± 0.31
BEAR-Lag	reward ↑	0.80 ± 0.04	0.85 ± 0.05	0.87 ± 0.03	0.56 ± 0.42	0.62 ± 0.29	0.16 ± 0.14
	cost ↓	2.49 ± 0.26	3.08 ± 0.88	3.61 ± 0.26	3.08 ± 2.18	3.62 ± 2.85	4.11 ± 2.42
CPQ	reward ↑	**0.62 ± 0.03**	**0.65 ± 0.03**	**0.01 ± 0.02**	0.37 ± 0.03	0.98 ± 0.01	0.34 ± 0.01
	cost ↓	**0.78 ± 0.14**	**0.45 ± 0.63**	**0.64 ± 0.36**	3.08 ± 0.66	1.69 ± 1.57	1.08 ± 0.79
COptiDICE	reward ↑	0.72 ± 0.01	0.44 ± 0.05	**0.42 ± 0.01**	0.63 ± 0.03	**0.95 ± 0.01**	0.67 ± 0.03
	cost ↓	2.30 ± 0.11	3.45 ± 0.56	**0.85 ± 0.34**	3.12 ± 0.15	**0.00 ± 0.00**	3.75 ± 0.16
CDT	reward ↑	0.70 ± 0.00	**0.72 ± 0.02**	0.64 ± 0.00	**0.30 ± 0.00**	1.00 ± 0.00	**0.60 ± 0.00**
	cost ↓	1.06 ± 0.04	**0.82 ± 0.01**	1.07 ± 0.04	**0.28 ± 0.40**	1.01 ± 0.20	**0.33 ± 0.13**
DIAM	reward ↑	**0.71 ± 0.01**	**0.66 ± 0.02**	**0.64 ± 0.01**	**0.30 ± 0.11**	**0.96 ± 0.02**	**0.34 ± 0.04**
	cost ↓	**0.98 ± 0.12**	**0.30 ± 0.25**	**0.45 ± 0.17**	**0.57 ± 0.46**	**0.45 ± 0.63**	**0.69 ± 0.55**

Gray: unsafe agent **Bold**: safe agent **Blue:** safe agent with highest reward

- **DIAM** is able to balance well the demand for safety and reward, learning a safe and relatively rewarding behavior

Experiment

Carnegie Mellon University

■ Q1: how is our method compared to other baselines? (Data-centric approach)

Tasks	Stats	Safe-Top-10	Safe	Vanilla	Top-10	DIAM
BallCircle	reward ↑	0.85 ± 0.02	0.68 ± 0.04	0.75± 0.03	0.89 ± 0.01	**0.77 ± 0.02**
	cost ↓	1.06 ± 0.03	0.48 ± 0.06	1.08 ± 0.14	1.35 ± 0.04	**0.96 ± 0.04**
	high vel cost ↓	0.84 ± 0.11	1.18 ± 0.27	1.05 ± 0.12	0.77 ± 0.30	**0.88 ± 0.07**
	low vel cost ↓	1.17 ± 0.02	0.92 ± 0.05	0.95 ± 0.04	1.06 ± 0.01	**0.84 ± 0.04**
CarCircle	reward↑	0.73 ± 0.02	**0.56 ± 0.04**	0.92± 0.03	0.79 ± 0.02	**0.67 ± 0.01**
	cost ↓	1.82 ± 0.59	**0.39 ± 0.09**	2.35 ± 0.04	1.29 ± 0.34	**0.91 ± 0.16**
	high vel cost ↓	1.04 ± 0.63	**0.61 ± 0.23**	0.25 ± 1.12	0.76 ± 0.48	**0.89 ± 0.13**
	low vel cost ↓	0.85 ± 0.01	**0.87 ± 0.03**	0.91 ± 0.02	0.96 ± 0.06	**0.86 ± 0.03**
DroneCircle	reward ↑	0.57 ± 0.03	**0.63 ± 0.01**	0.71 ± 0.02	0.72 ± 0.09	**0.63 ± 0.01**
	cost ↓	1.38 ± 0.40	**0.62 ± 0.34**	1.38 ± 0.38	3.84 ± 0.22	**0.50 ± 0.38**
	high vel cost ↓	0.94 ± 0.43	**0.45 ± 0.33**	0.86 ± 0.50	1.07 ± 1.49	**0.46 ± 0.33**
	low vel cost ↓	0.53 ± 0.04	**0.49 ± 0.01**	0.48 ± 0.03	0.54 ± 0.05	**0.48 ± 0.04**

Gray: unsafe agent **Bold**: safe agent **Blue:** safe agent with highest reward

- We also set up two additional constrains, **High Velocity Constrain** and **Low Velocity Constrain**
- DIAM samples with certain weights and achieves **high reward** without violating the **safety constraint**, which shows strength in even difficult task settings.

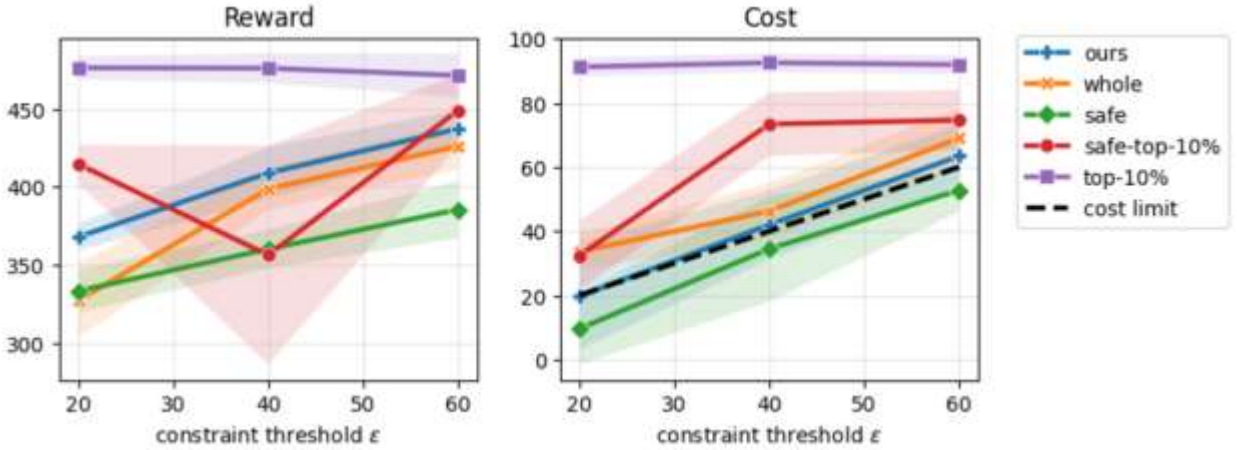
Experiment

Carnegie Mellon University

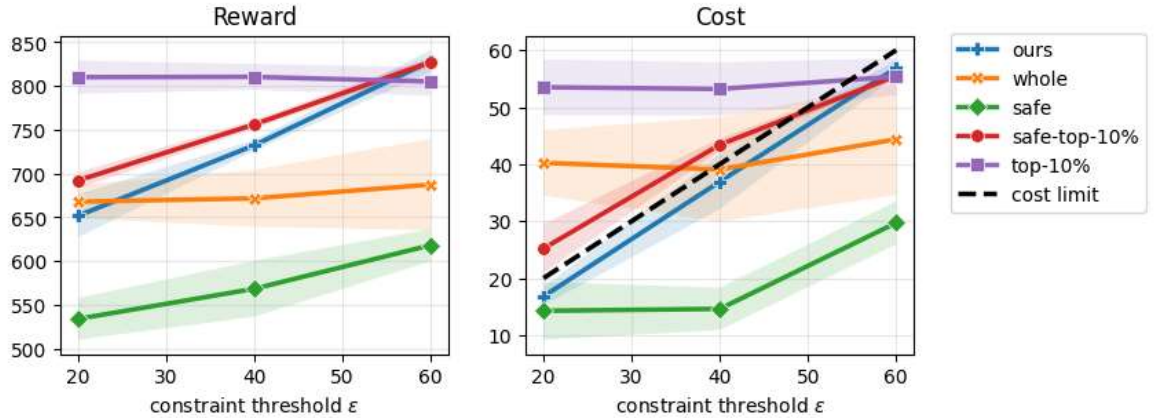
- **Q2: How does DIAM benefit policy-regularized safe reinforcement learning method in different cost thresholds?**

  - In order to answer this question, we conduct research on **Ball Circle(left)** and **Car Circle(right), Drone Circle(not shown),** where we choose three most commonly used cost threshold **20, 40, 60**.

Ball Circle



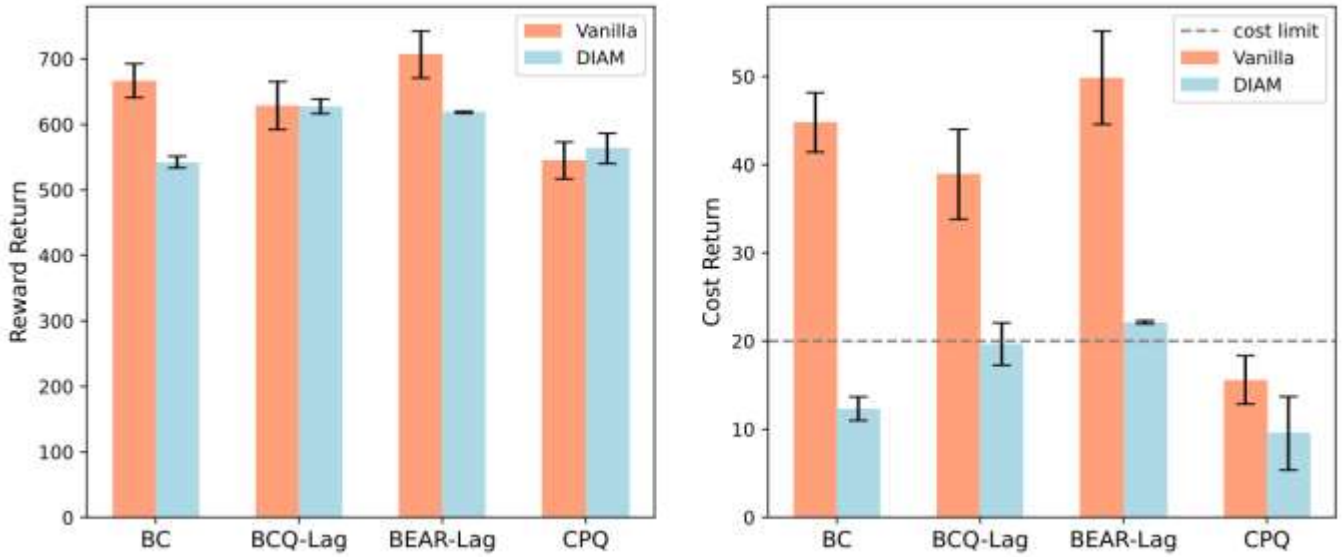
Car Circle



Experiment

Carnegie Mellon University

■ Q3: How does DIAM applied to different policy-regularized offline safe RL algorithms?



- **BC** shows highly tempting property due to the suboptimality, of the behavior policy, while **DIAM** shows ability to form a safe and rewarding behavior
- The **BCQ-Lag** and **BEAR-Lag** are both compatible with **DIAM**
- Due to the pessimistic estimation methods inducing conservative policy, **CPQ** will be more conservative after **DIAM** hence show little improvement of the performance.

Contents

Future work

Carnegie Mellon University

- **Q: Can sampling method applied to other safe offline RL method like CDT?**
  - **No!** Because our sampling method is **trajectory-wise**
- **Q: Can we provide our sampling method with theoretical guarantee?**
  - **Hard to say.** Existing works have not provided any framework yet.
- **Q: Can sampling method deal with the problems w.r.t multi-constraint?**
  - **Just empirically work,** we don't solve the problem intrinsically
- The above limitations encourage me to focus on deal with multi-constraint in a more intrinsic way, that is, to train a model to achieve the **Pareto Frontier** in the multi-constraint problem

Propose method: CDT + gradient surgery

Carnegie Mellon University

■ Constraint Decision Transformer (CDT) and gradient surgery (PC-Grad)

- Gradient surgery works but not obvious!

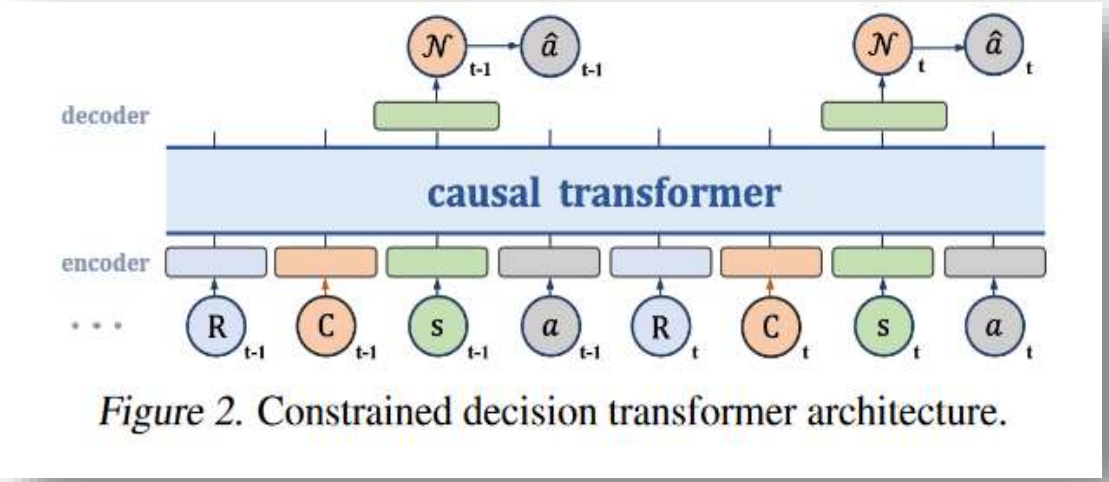


Figure 2. Constrained decision transformer architecture.

$HV(with\ grad) = 0.4462 > HV(without\ grad) = 0.4278$