Don't Trade Off Safety: Diffusion Regularization for Constrained Offline RL

Junyu Guo[1], Zhi Zheng[1], Donghao Ying[1], Ming Jin[2],
Shangding Gu[1], Costas Spanos[1], Javad Lavaei[1]

[1]University of California, Berkeley [2]Virginia Tech

NeurIPS 2025

Poster Presentation

Overview

Motivation

Challenges in Offline Safe RL:

- Distributional shift
- Safe policy training often comes at the cost of performance

Key Question:

How can we balance reward maximization and constraint satisfaction without risking out-of-distribution actions or unsafe behavior when no additional data can be collected?

Problem Formulation: MDP Setup

Constrained Markov Decision Process (CMDP)

MDP: $(\mathcal{S}, \mathcal{A}, P, r, c, \gamma)$

- $\mathcal{S}$: state space, $\mathcal{A}$: action space
- $P : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$: transition kernel
- r: reward function, c: cost function
- Policy: $\pi : \mathcal{S} \to \Delta(\mathcal{A})$

Value Functions

Under policy π, the value function for $\diamond \in \{r, c\}$ is:

$$V_{\diamond}^{\pi}(\rho) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \diamond (s_t, a_t) \,\middle|\, s_0 \sim \rho\right]$$

Problem Formulation: Optimization

Objective: Find policy $\pi \in \Pi$ that maximizes rewards $V_r^\pi(\rho)$ while keeping cost $V_c^\pi(\rho)$ within budget l.

Offline RL Setting: Fixed dataset collected by behavioral policy π_b:

$$\mathcal{D}^\mu = \{(s_i, a_i, r_i, s_i', c_i)\}_{i=1}^N$$

Constrained Optimization Problem

$$\max_{\pi \in \Pi} \mathbb{E}[V_r^\pi(\rho)] \quad \text{subject to} \quad \begin{cases} D_{\mathrm{KL}}(\pi \| \pi_b) \leq \epsilon & \text{(behavioral constraint)} \\ \mathbb{E}[V_c^\pi(\rho)] \leq l & \text{(safety constraint)} \end{cases} \tag{1}$$

Key constraints: Stay close to behavioral policy & satisfy safety requirements

Existing Works: Primal-Dual Approach

Lagrangian Formulation

Most prior work reformulates the constrained problem as:

$$\max_{\pi \in \Pi} \mathbb{E}[V_r^\pi(\rho)] - \lambda \left(\mathbb{E}[V_c^\pi(\rho)] - l\right) \quad \text{s.t.} \quad D_{\mathsf{KL}}(\pi \| \pi_b) \leq \epsilon \tag{2}$$

where $\lambda \geq 0$ is the Lagrange multiplier.

Key Mechanism

- When safety constraint is **violated**: λ increases $\rightarrow$ greater penalty on cost
- When safety constraint is **satisfied**: λ decreases $\rightarrow$ emphasis on reward
- **Trade-off**: Balancing reward and safety through dynamic λ

Step 1: Diffusion-Based Behavioral Cloning

Goal

Learn the behavioral policy $\pi_b(a \mid s)$ from offline dataset using diffusion models.

Diffusion Training Objective

$$\mathcal{L} = \mathbb{E}_{(s,a)\sim\mathcal{D}^\mu}\mathbb{E}_{t\sim\text{Unif}(0,T)}\left[w(t)\|\epsilon_\psi(a_t, t \mid s) - \epsilon_t\|^2\right] \tag{3}$$

Noising process: $a_t = \sqrt{\bar{\alpha}_t}a + \sqrt{\bar{\beta}_t}z$, where $z \sim \mathcal{N}(0, I)$

- $\bar{\alpha}_t$, $\bar{\beta}_t$: pre-selected diffusion parameters
- ϵ_ψ: learned noise prediction network
- Captures the full distribution of behavioral actions

Step 2: Diffusion Regularization

Score Function Approximation

$$\nabla_a \log p(a_t \mid s) \approx s_\psi(a_t, t \mid s) = -\frac{1}{\sqrt{\bar{\beta}_t}} \epsilon_\psi(a_t, t \mid s) \tag{4}$$

Policy parameterization: Gaussian policy $\pi_\theta(a \mid s) = \mathcal{N}(a; m_\theta(s), \Sigma_\theta(s))$

KL Divergence Gradient

$$\nabla_\theta D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s) \| \mu_\psi(\cdot \mid s)) = \mathbb{E}_{z \sim \mathcal{N}(0,I)} \Big[-\nabla_a \log \mu_\psi(m_\theta(s) + \Sigma_\theta^{1/2} z \mid s) \cdot \nabla_\theta (m_\theta(s) + \Sigma_\theta^{1/2} z) \Big] - \frac{1}{2} \nabla_\theta \log(\det \Sigma_\theta(s)) \tag{5}$$

Step 3: Safe Policy Adaptation

Dual Optimization Objectives

$$\textbf{Reward:} \quad \max_{\pi \in \Pi} \mathbb{E}_{s \sim \mathcal{D}^\mu} \left[V_r^\pi(s) - \frac{1}{\beta} D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi_b(\cdot|s)) \right] \tag{6}$$

$$\textbf{Cost:} \quad \max_{\pi \in \Pi} \mathbb{E}_{s \sim \mathcal{D}^\mu} \left[-(V_c^\pi(s) - l) - \frac{1}{\beta} D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi_b(\cdot|s)) \right] \tag{7}$$

Standard Gradient Updates:

- Reward optimization: $\theta \leftarrow \theta + \eta g_r$
- Cost optimization: $\theta \leftarrow \theta + \eta g_c$

Challenge: Reward and cost objectives may conflict!

Gradient Conflict Detection

Key Idea: Combine both gradients as $\beta_r g_r + \beta_c g_c$ to balance reward and safety simultaneously.

Measuring Gradient Alignment

Use the angle between gradients to detect conflicts:

$$\phi := \cos^{-1}\left(\frac{\langle g_r, g_c \rangle}{\|g_r\|\|g_c\|}\right)$$

- $\phi < 90°$: Gradients **aligned** — reward and cost improvements agree
- $\phi \geq 90°$: Gradients **conflicting** — improving one hurts the other

Solution: Apply different gradient manipulation strategies based on ϕ

Gradient Manipulation Mechanism

Final gradient:

$$g = \begin{cases} \frac{g_r + g_c}{2}, & \phi \in (0°, 90°) \\ \frac{g_r^+ + g_c^+}{2}, & \phi \in [90°, 180°] \end{cases} \tag{8}$$

When aligned ($\phi < 90°$):
Simply average the gradients

When conflicting ($\phi \geq 90°$):
Use projections to remove interference:

- $g_r^+ = g_r - \frac{g_r \cdot g_c}{\|g_c\|^2} g_c$
- $g_c^+ = g_c - \frac{g_c \cdot g_r}{\|g_r\|^2} g_r$

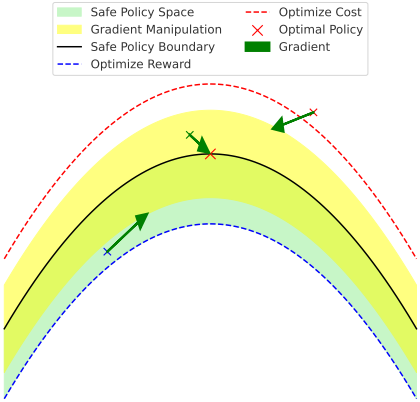


Figure: Gradient strategies

DRCORL Algorithm

Algorithm 1 DRCORL: Diffusion Regularized Constrained Offline RL

Input: Dataset $\mathcal{D}^{\mu}$, slack bounds h^{+}, h^{-}
Phase 1: Pretraining
  Pretrain diffusion model ϵ_{ψ} for behavioral policy
  Pretrain critics Q_r^{π}, Q_c^{π} via behavioral cloning
Phase 2: Safe Policy Adaptation
for each gradient step $t = 1, 2, \ldots, T$ **do**
  Sample mini-batch $\mathcal{B}$ from $\mathcal{D}^{\mu}$
  Update critics for reward and cost
  $g \leftarrow \textsc{SafeAdaptation}(g_r, g_c)$ *// Gradient manipulation*
  $\theta \leftarrow \theta + \eta g$
end for
Output: Policy π_{θ}

Theoretical Guarantee

Theorem (Performance Bound)

Let $\tilde{\pi}$ be the weighted policy after T iterations with dataset size $|\mathcal{D}^{\mu}| = \mathcal{O}\left(\frac{C \ln(|\mathcal{F}|/\delta)}{\epsilon_{off}(1-\gamma)^4}\right)$. Then with probability $\geq 1-\delta$:

Optimality gap:

$$V_r^{\pi^*}(\rho) - \mathbb{E}[V_r^{\tilde{\pi}}(\rho)] \leq \mathcal{O}(\epsilon_{off}) + \mathcal{O}\left(\sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 T}}\right) \tag{9}$$

Constraint violation:

$$\mathbb{E}[V_c^{\tilde{\pi}}(\rho)] - I \leq \mathcal{O}(\epsilon_{off}) + \mathcal{O}\left(\sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 T}}\right) \tag{10}$$

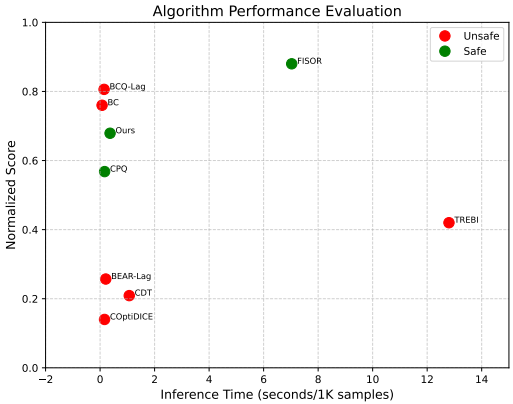
ϵ_{off}: offline approximation error; $\mathcal{F}$: critic function class

Experimental Results

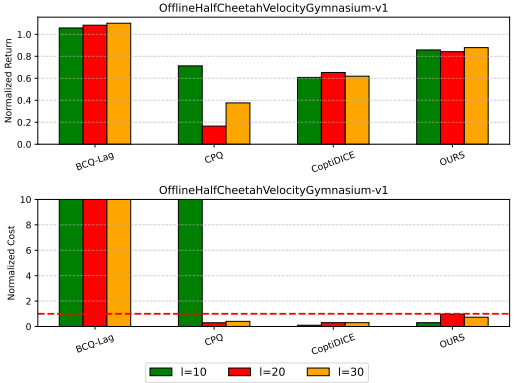
Table: Normalized return (↑: higher is better) and cost (↓: lower is better; threshold at 1)

	BC-Safe		BEARL		BCQ-Lag		CPQ		COptiDICE		CDT		CAPS		CCAC		Ours	
Task	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓
CarGoal1	**0.38**	**0.46**	0.71	4.45	0.46	3.27	0.68	3.73	0.48	2.31	0.65	3.75	0.40	1.35	0.84	6.52	**0.91**	**0.00**
CarGoal2	**0.25**	**0.82**	0.46	10.98	0.29	3.46	0.27	28.24	0.18	2.28	**0.03**	**0.00**	0.12	2.20	0.96	6.97	**0.79**	**0.60**
PointButton1	0.17	1.37	0.35	6.71	0.17	4.00	0.56	11.63	0.09	3.55	0.62	13.05	0.16	3.66	0.71	4.27	**0.81**	**0.40**
PointCircle1	0.88	3.79	-0.33	17.84	0.45	8.13	0.23	6.77	0.85	18.30	**0.51**	**0.00**	**0.50**	**0.14**	0.62	7.58	**0.53**	**0.40**
PointGoal1	**0.53**	**0.88**	0.76	3.46	0.59	5.06	**0.41**	**0.69**	0.52	5.26	0.68	4.35	**0.20**	**0.53**	0.77	5.20	**0.88**	**0.00**
PointGoal2	0.60	3.15	0.80	12.33	0.65	10.80	0.34	5.10	0.38	4.62	0.32	1.45	0.23	1.91	0.80	4.88	**0.82**	**0.00**
AntVel	**0.87**	**0.52**	**-1.01**	**0.00**	0.99	8.39	**-1.01**	**0.00**	0.99	11.41	**0.91**	**0.97**	**0.81**	**0.36**	**0.71**	**0.39**	**0.88**	**0.89**
HalfCheetahVel	**0.94**	**0.71**	0.95	104.45	1.06	63.94	0.71	14.70	**0.61**	**0.00**	**0.96**	**0.61**	**0.88**	**0.22**	**0.85**	**0.93**	**0.86**	**0.00**
HopperVel	0.21	1.50	0.26	124.05	0.81	14.91	**0.57**	**0.00**	0.14	7.83	0.21	1.12	**0.83**	**0.00**	**0.11**	**0.68**	**0.68**	**0.79**
Walker2dVel	**0.78**	**0.08**	**0.76**	**0.80**	**0.80**	**0.07**	**0.08**	**0.95**	0.12	2.34	0.73	1.95	**0.78**	**0.00**	**0.21**	**0.26**	**0.74**	**0.30**
CarRun	**0.97**	**0.10**	0.49	7.43	**0.95**	**0.00**	**0.92**	**0.05**	**0.93**	**0.00**	0.99	1.10	**0.98**	**0.86**	**0.93**	**0.05**	**0.99**	**0.30**

Ablation



(a)



(b)

Figure: (a) Computational Efficiency vs. Performance Trade-off. (b) Normalized return and cost under varying cost limits ($l = 10, 20, 30$).

Conclusion

DRCORL: Don't Trade Off Safety

Technical Innovation

Diffusion Regularization

- Accurate behavioral modeling
- Flexible constraint handling

Gradient Manipulation

- Conflict-aware optimization
- No reward-safety trade-off

Impact

Theory

- Convergence guarantees
- Sample complexity bounds

Practice

- SOTA performance
- Consistent safety compliance

Thank you!